Toward Understanding Visual Perception in Machines with Human Psychophysics

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt
von
Judith Borowski
aus München

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	24.10.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Matthias Bethge
2. Berichterstatter:	Prof. Felix A. Wichmann, DPhil

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel

"Toward Understanding
Visual Perception in Machines
with Human Psychophysics"

selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, den _____ _____

Datum/Date Unterschrift/Signature

Summary

Over the last several years, Deep Learning algorithms have become more and more powerful. As such, they are being deployed in increasingly many areas including ones that can directly affect human lives. At the same time, regulations like the GDPR or the AI Act are putting the request and need to better understand these artificial algorithms on legal grounds. How do these algorithms come to their decisions? What limits do they have? And what assumptions do they make?

This thesis presents three publications that deepen our understanding of deep convolutional neural networks (DNNs) for visual perception of static images. While all of them leverage human psychophysics, they do so in two different ways: either via direct comparison between human and DNN behavioral data or via an evaluation of the helpfulness of an explainability method. Besides insights on DNNs, these works emphasize good practices: For comparison studies, we propose a checklist on how to design, conduct and interpret experiments between different systems. And for explainability methods, our evaluations exemplify that quantitatively testing widely spread intuitions can help put their benefits in a realistic perspective.

In the first publication, we test how similar DNNs are to the human visual system, and more specifically its capabilities and information processing. Our experiments reveal that DNNs (1) can detect closed contours, (2) perform well on an abstract visual reasoning task and (3) correctly classify small image crops. On a methodological level, these experiments illustrate that (1) human bias can influence our interpretation of findings, (2) distinguishing necessary and sufficient mechanisms can be challenging, and (3) the degree of aligning experimental conditions between systems can alter the outcome.

In the second and third publications, we evaluate how helpful humans find the explainability method feature visualization. The purpose of this tool is to grant insights into the features of a DNN. To measure the general informativeness and causal understanding supported via feature visualizations, we test participants on two different psychophysical tasks. Our data unveil that humans can indeed understand the inner DNN semantics based on this explainability tool. However, other visualizations such as natural data set samples also provide useful, and sometimes even *more* useful, information. On a methodological level, our work illustrates that human evaluations can adjust our expectations toward explainability methods and that different claims have to match the experiment.

Zusammenfassung

In den letzten Jahren sind Deep Learning Algorithmen immer leistungsfähiger geworden. Daher werden sie in immer mehr Bereichen eingesetzt, und zwar auch in solchen, die das Leben der Menschen direkt beeinflussen können. Gleichzeitig fordern die EU Datenschutz-Grundverordnung oder der AI Act und formulieren es als notwendig, diese künstlichen Algorithmen besser verstehen zu können. Wie kommen diese Algorithmen zu ihren Entscheidungen? Wo erreichen sie ihre Grenzen? Und welche Annahmen treffen sie?

In dieser Doktorarbeit werden drei Veröffentlichungen vorgestellt, die unser Verständnis von deep convolutional neural networks (DNNs, auf Deutsch etwa tiefe faltende neuronale Netze) für visuelle Wahrnehmung statischer Bilder vertiefen. In allen dreien wird menschliche Psychophysik genutzt, allerdings auf zwei unterschiedliche Arten: entweder durch den direkten Vergleich zwischen menschlichen und DNN-Verhaltensdaten oder durch die Bewertung der Nützlichkeit einer Erklärungsmethode. Neben den Erkenntnissen über DNNs betonen unsere Arbeiten bewährte Vorgehensweisen: Für Vergleichsstudien schlagen wir eine Checkliste für die Planung, Durchführung und Interpretation von Experimenten zwischen verschiedenen Systemen vor. Und für Erklärungsmethoden zeigen unsere Evaluierungen, dass die quantitative Prüfung weit verbreiteter Intuitionen dazu beitragen kann, deren Nutzen in eine realistische Perspektive zu rücken.

In der ersten Veröffentlichung testen wir, inwieweit DNNs dem menschlichen visuellen System ähneln, genauer gesagt dessen Fähigkeiten und Informationsverarbeitung. Unsere Experimente zeigen, dass DNNs (1) geschlossene Konturen erkennen können, (2) bei einer abstrakten visuellen Logikaufgabe gut abschneiden und (3) kleine Bildausschnitte korrekt klassifizieren. Auf methodischer Ebene zeigen diese Experimente, dass (1) menschliche Voreingenommenheit unsere Interpretation der Ergebnisse beeinflussen kann, (2) die Unterscheidung zwischen notwendigen und hinreichenden Mechanismen schwierig sein kann und (3) das Ausmaß der Angleichung der experimentellen Bedingungen zwischen den Systemen das Ergebnis verändern kann.

In der zweiten und dritten Veröffentlichung messen wir, wie hilfreich Menschen die Erklärungsmethode feature visualization (auf Deutsch etwa Merkmalsvisualisierung) finden. Das Ziel dieses Werkzeugs ist es, Einblicke in die Eigenschaften eines DNN zu gewähren. Um den allgemeinen Informationsgrad und das kausale Verständnis zu messen, das durch feature visualizations unterstützt wird, testen wir Teilnehmende in zwei verschiedenen Psychophysik-Aufgaben. Unsere Daten zeigen, dass Menschen tatsächlich die innere Semantik von DNNs anhand dieser Erklärungsinstrumentes verstehen können. Allerdings liefern auch andere Visualisierungen, wie z.B. natürliche Bilder des Datensatzes, nützliche und manchmal sogar noch nützlichere Informationen. Auf methodischer Ebene zeigt unsere Arbeit, dass menschliche Evaluierungen unsere Erwartungen an Erklärungsmethoden anpassen können und dass verschiedene Folgerungen dem Experiment entsprechen müssen.

Contents

1 Introduction

Machine algorithms are becoming more and more powerful and they are being deployed in more and more areas. An Artificial Intelligence (AI) system supporting self-driving cars (Grigorescu et al., 2020), healthcare applications (Miotto et al., 2018) or facial recognition (Schroff et al., 2015) is no novelty anymore. Many of these are high-stake situations, and decisions can have consequential effects on people's lives.

Given the broad deployment of AI algorithms under often critical circumstances, we want to understand these algorithms. How do they come to their decisions? What limits do they have? And what assumptions do they make? The need for transparency is becoming ever more pressing and is even strengthened by law: Since 2018, the European Union's (EU) General Data Protection Regulation (GDPR) grants users a "right to explanation" under certain conditions (Goodman and Flaxman, 2017). And since 2021, the EU has been discussing the AI Act (Parliament, 2021), the first major regulation on AI in the world, which requires transparency for e.g. technologies that interact with humans.

While there are many different ways to understand machine algorithms, this thesis presents two such approaches in three publications for modern machine learning systems of visual perception. To guide the reader, they are organized along decreasing granularity: First, comparison studies with humans on a behavioral level reveal both abilities and limits as well as insights into potential features and mechanisms of artificial visual systems (see Section 2.1). Then, humans' understanding of their internal information processing is tested based on an explainability method in two ways (see Sections 2.2 and 2.3).

Overall, this thesis is structured as follows: The current chapter provides background on vision research, modern machine learning algorithms and explains open questions that we address in our publications. Next, Chapter 2 summarizes our specific contributions and Chapter 3 discusses them in a bigger context. Finally, Chapter 4 gives an outlook on future research directions.

1.1 Human visual perception is a feat

Human visual perception is a feat: Our brain transforms physical visual information, which hits the retina, to a meaningful representation. Thereby, we not only understand color, depth, and distance, but we also recognize patterns as well as detect, identify and classify objects. The latter are then in turn further processed to e.g. take actions. All of this happens effortlessly and situations like sidestepping someone on a crowded sidewalk, even when it's raining or snowing, or finding one's keys in the middle of a plurality of items on a table require little thinking. Another great example of the numerous processes and factors at play in visual perception is the photograph below.

Most likely, the picture is perceived as funny — after all, the former American President Barack Obama is playing a trick on another official man. To recognize this, the viewer's brain processes a *lot* of information, e.g. it identified several people in a room as well as their facial expressions and gaze directions, while discerning

Figure 1: **A funny picture?!** Our brain processes a lot of information to comprehend the scene shown in the picture. (The photograph, which is in the public domain, is taken from (Souza, 2010) with "Courtesy Barack Obama Presidential Library". The analysis is inspired by (Karpathy, 2012).)

their reflections in mirrors. Crucially, a viewer comprehends that the person on the scale has different information than the other people: While most people in the scene know that Obama is pressing the scale down with his foot, the man on the scale is unaware of what is happening behind him. Finally, general knowledge of, for example, weight being a delicate topic or Obama being a high-profile man help to assess the situation. Altogether, this information allows a viewer to understand the situation in the photograph, and the number of different pieces of information as well as their nature exemplify what a great feat human visual perception is.

Even though it is debated whether vision is really the most important and most complex sensory modality (Kandel et al., 2000; Goldstein, 2010; Gerrig et al., 2015; Hutmacher, 2019), the body of literature about it is huge (Katz, 1989; Gallace and Spence, 2009; Hutmacher, 2019). As a matter of fact, many more studies investigate how we *visually* perceive the world as opposed to how we experience it via e.g. auditory or gustatory inputs (Katz, 1989; Gallace and Spence, 2009; Hutmacher, 2019).

In this thesis, human visual perception serves as a reference point or evaluation means. The main goal, however, is to better understand visual perception in *machines*.

1.2 Visual perception in machines: a brief history

Implementing vision in artificial systems has been attempted for many decades. While one goal, which is typically attributed to the Computer Vision community, is to engineer algorithms as powerful as possible, other efforts of artificially implementing vision try to more closely mirror biological systems. Below, a brief history of machine

algorithms for vision is given, interspersed with important milestones from biological vision science.

One of the earliest precursors of today's modern neural network machine learning algorithms was invented in 1943: With the goal of understanding how the brain could produce highly complex patterns, McCulloch and Pitts (1943) created a basic model of a brain cell. The authors showed that combinations of this unit realized as a thresholded sum of excitatory and inhibitory inputs could produce logical functions such as AND, OR, or NOT. However, this early version of a neural network did not yet learn. This changed in 1958, when Rosenblatt (1958) introduced the "Perceptron", a single-layer neural network which could classify images into two classes. However, the realization of Minsky and Papert (1969) that a Perceptron could only successfully solve linearly separable problems caused an "AI winter" and the community shifted its focus toward rule-based algorithms.

In consecutive years, a huge break-through was achieved in biological vision. In contrast, the Computer Vision community realized that the problem of artificially implementing vision was harder than expected. Hubel and Wiesel (1962) discovered two major cell types in primary visual cortex of cats in 1962: Simple cells respond to bars at specific locations, whereas complex cells respond to bars in several different, nearby locations. In contrast, the cognition and computer scientist Marvin Minsky had completely unrealistic expectations from a summer project in 1966: He asked his undergraduate student Gerald Jay Sussman to "spend the summer linking a camera to a computer and getting the computer to describe what it saw" (Szeliski, 2010)[1]. Poggio (1981) summarized that "Until the early 1970s the field of Computer Science and Artificial Intelligence failed to realize that problems in vision are difficult."

During the second wave of "Deep Learning" around 1980/90, algorithmic foundations were laid out for later success. The "Neocognitron" (Fukushima, 1980) included the idea of *location-invariant feature extractors*, which permitted a certain invariance to translation. At the same time, the Neocognitron was the first functioning model that incorporated the neurophysiological findings of Hubel and Wiesel (1962) and consisted of several layers of simple and complex cells. According to Schmidhuber (2015), it may be the "first artificial [neural net] that deserved the attribute *deep*". While the weights of the Neocognitron were set in an unsupervised way, the introduction of the *backpropagation* algorithm in 1986 by Rumelhart et al. (1986) meant that neural networks consisting of multiple stacked layers could be trained. A few years later, LeCun et al. (1989) demonstrated the success of combining backpropagation and the previously introduced idea of translation invariance, which was now implemented by *convolutions*, i.e. sliding windows of weights applied to all image patches: Handwritten digits could be classified.

In the field oriented toward modeling biological vision, many hierarchical models followed the Neocognitron Riesenhuber and Poggio (2000). One of the most famous ones is the "HMAX"-model. It added the *max*-operation, achieved robustness to image variations such as position, scale and translation, and was consistent with

[1]Tracing down this quote is not straight-forward. According to Szeliski (2010), the original vision memo "was authored by Papert (1966) and involved a whole cohort of students".

physiological data (Riesenhuber and Poggio, 1999; Tarr, 1999). As hierarchical models could be applied to the same inputs as in biological experiments, direct comparisons were performed (Lindsay, 2021): Experiments suggested that these models would correspond to the first $100 - 150$ msec of primate visual processing (Serre et al., 2007a) and that artificial responses would match neural ones (Cadieu et al., 2007).

In the 21st century, more computational power (particularly via graphics processing units) (Krizhevsky et al., 2012), larger data sets available from the web (Deng et al., 2009; Russakovsky et al., 2015; Geiger et al., 2013; Everingham et al., 2010), as well as the milestones achieved during the second AI wave allowed neural networks to become more and more powerful. The focus shifted away from hand-crafted features and toward *end-to-end learning* in deeper networks, or as LeCun et al. (1998) put it: "better pattern recognition systems can be built by relying more on automatic learning and less on hand-designed heuristics." Deep neural networks (DNNs[2]) are what powers the major AI advances of the last decade (LeCun et al., 2015; Mitchell, 2021b).

DNNs revolutionized Computer Vision and vision modeling. Today, modern DNNs cannot only, for example, detect cancer in ultrasound data (Jush et al., 2020), classify fauna (Seeland and Mäder, 2021) and other images (He et al., 2015, 2016), or drive a car (Ren et al., 2021; Janai et al., 2020), but also — though these tend to be based on other architectures than convolutions — recognize speech (Hinton et al., 2012a) or translate between languages (Sutskever et al., 2014). Competitions like the ImageNet Large Scale Visual Recognition Challenge were key drivers for even better DNNs (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2015, 2016; Szegedy et al., 2015) and many innovations in e.g. training methods (He et al., 2016; Duchi et al., 2011; Srivastava et al., 2014; Hinton et al., 2012b; Kingma and Ba, 2014) made DNNs even more powerful. In fact, they were found to *surpass* human performance on very special tasks such as image classification (He et al., 2015; Kheradpisheh et al., 2016).

With Deep Learning, scientists have successful models for the first time that can actually do a task. Only years ago, this situation seemed decades away (Wichmann et al., 2017). And what is more, the feat of object recognition emerges from training end-to-end, i.e. without modeling the visual system step by step (Lonnqvist et al., 2020).

Similar to the biological sensory modality, Computer Vision is the most or among the most researched areas in Computer Science / Machine Learning according to conference rankings (research.com, 2022; Tang, 2022), the impact of publications (scholar.google.com, 2022) and the number of publications uploaded to a popular pre-print server (arXiv, 2020).

Having sketched a brief history of visual perception in artificial systems, the focus is now turned toward *understanding* them.

[2]In this thesis, *DNN* always refers to deep *convolutional* neural network, unless otherwise stated. In the original publications 2 and 3, the term "CNN" is used.

1.3 Why do we not understand DNNs?

Seemingly paradoxically, we — i.e. Deep Learning scientists or practitioners — claim we would not understand DNNs. This often surprises the general public. After all, we are the experts who *ourselves* program these algorithms. Should it not consequently be absolutely logical that we know what is going on? The answer is yes and no.

What we do completely understand are the mathematical operations that the DNN components execute and that we train it with. Further, we have a reasonably good understanding of helpful engineering practices and we can even inspect all the resulting parameters and the learned representations that DNNs form. However, what is not so clear is how the "processing in these networks truly works" (Lillicrap and Kording, 2019), i.e. why the image of a cat is classified as a cat and not a dog. In other words, why those sometimes even human-like performing properties emerge remains hard to grasp.

While this may still sound abstract, the following parallel to biology can clarify that even understanding all constituents does not automatically mean understanding how behavior emerges. For example, since 1998, the nervous system as well as the genes of the nematode C. elegans have been completely decoded and sequenced. Nonetheless, what has not yet been understood is how this biological hardware connects to the worm's behavior (Mausfeld, 2003; Geirhos et al., 2020b). Another, even though slightly less analogous, example can be found in our human body: Despite having understood a fair bit of how chemical and physical processes work, scientists are still exploring different hypotheses regarding how anesthesia drugs are successful (Wang et al., 2020a; Jiang-Xie et al., 2019).

Shifting the focus back to DNNs, their reputation as "black boxes" (Guidotti et al., 2018; FEL et al., 2021; Adadi and Berrada, 2018; Rudin, 2019) becomes intuitive in the context of the described obscurity around how these algorithms produce their behavior. Clearly, it is important to better understand how their successes and failures come about.

1.4 What are the benefits of understanding machines?

While counteracting the lack of transparency is a big motivator for working toward better understanding machines, so are the *opportunities* that arise with a better understanding: For example, debugging (Koh and Liang, 2017; Elton, 2022) and improving models (Koh and Liang, 2017) would become easier. Also, steps toward increasing fairness (Taylor and Taylor, 2021), toward meeting legal requirements like in the GDPR (Goodman and Flaxman, 2017) and toward engendering trust (Kim, 2015; Swartout, 1983; Elton, 2022) would be accomplished. Furthermore, we would be able to guarantee that DNNs work as expected and that they can be reliably and responsibly deployed (Lakkaraju et al., 2020).

Put another way, these opportunities correspond to meeting the needs of the many different stakeholders involved in machine decisions. When broadening the horizon beyond vision applications, this can include and is not limited to end users (e.g. loan

applicants), decision makers (e.g. doctors, judges), researchers and engineers, as well as people from regulatory agencies (e.g. European Union) (Lakkaraju et al., 2020).

1.5 How can we understand DNNs?

The next question is: How can we understand DNNs? Luckily, there are myriad different ways. Here, two approaches that are relevant for the publications in Chapter 2 are highlighted. I present them along decreasing granularity from a coarse to a more fine-grained level of understanding.

1.5.1 Comparisons with humans

An obvious reference point for DNNs are humans. After all, vision is a feat that we are really good at. As explained in Section 1.1, we can not only robustly recognize the world around us but also reason about these inputs.

Comparing machines to humans reveals how well they mirror our visual system. As such, many parallels, but also differences exist. For example, both systems contain many hierarchically structured neurons. In contrast, e.g. the number of tasks that the machine or human visual system perform is different: Humans process information in many different ways and can utilize this knowledge for various purposes (see Figure 1). On the contrary, DNNs are typically trained to perform only *one* specific task.

In the following, we focus on similarities and differences on both a "functional" and an "algorithmic" level[3]. The former means that correspondences between inputs and outputs are compared — or in other words, *behavioral* reactions of DNNs and humans are evaluated. The latter, i.e. the algorithmic level, refers to the decision-making *strategy*. To this end, the processes and representations employed to solve the problem are investigated. Interesting questions for these two levels include but are not limited to e.g. what human abilities are replicable in DNNs? And are machines really starting to match human capabilities? Do they come to their decisions in similar ways as humans?

Functional and algorithmic machine-to-human comparisons leverage and build on extensive knowledge in psychophysics. The establishment of this discipline is attributed to the experimental physicist Gustav Theodor Fechner in 1860 (Fechner, 1860). Rigorously studying the mind and quantifying human behavior has granted innumerous valuable insights into complex visual systems and represents a foundation against which machines can be compared. For example, principles like exploring the entire psychometric function, i.e. measuring reactions to various levels of the stimulus, were introduced (Green, 1960) — and also transferred to DNNs (Wichmann et al., 2017; Geirhos et al., 2018b; Webster et al., 2018).

A growing number of studies is asking how similar DNNs are becoming to humans. One of the hallmarks is performance on object classification. In 2015/6, DNNs surpassed humans (Kheradpisheh et al., 2016; He et al., 2015). Other works test, for example, classical psychophysical principles such as closure from Gestalt

[3]David Marr introduced similar levels of descriptions and explanations (Poggio, 1981) and they are further discussed in Section 3.3.

theory (Kim et al., 2019, 2021a) or phenomena such as illusions (Gomez-Villa et al., 2019; Watanabe et al., 2018; Ward, 2019; Mély et al., 2018; Benjamin et al., 2019; Sun and Dekel, 2021; Baker et al., 2018a) and crowding (Volokitin et al., 2017; Doerig et al., 2020; Lonnqvist et al., 2020). Moreover, DNN capabilities on e.g. abstract visual reasoning (Barrett et al., 2018; Yan and Zhou, 2017; Zhang et al., 2019; Villalobos et al., 2020), intuitive physics understanding (Zhang et al., 2016), and gaze behavior (Kümmerer et al., 2014; Linardos et al., 2021; Kummerer et al., 2017) are investigated. While this list is far from complete, such studies improve our understanding of DNNs.

However, comparison findings between machines and humans are not always in agreement. For example, on the one hand, DNNs are claimed to "have potential to explain many aspects of human cognition" (Jozwik et al., 2017). On the other hand, though, Mitchell (2021a) states that "no current AI system is anywhere close to a capability of forming humanlike abstractions or analogies." In terms of specific phenomena, findings vary regarding how differently or similarly the two systems process e.g. illusions (Gomez-Villa et al., 2019; Watanabe et al., 2018; Ward, 2019; Mély et al., 2018; Benjamin et al., 2019; Sun and Dekel, 2021; Baker et al., 2018a) or adversarial examples (Zhou and Firestone, 2019; Dujmović et al., 2020). The latter are subtly altered images that are misclassified by DNNs but whose changes are typically not perceived by humans (Szegedy et al., 2013). Another example of varying results concerns what role texture- and shape-features play for DNNs (Geirhos et al., 2018a; Baker et al., 2018b; Kubilius et al., 2016; Hermann et al., 2020; Feinman and Lake, 2018; Ritter et al., 2017).

Comparing humans and machines can be difficult. This is not only illustrated by the previously mentioned diverging results but also other fundamental differences between the systems. For example, humans learn throughout their entire lives, whereas supervised algorithms are trained on one single data set. In comparison studies, such differences have to be addressed and how to best do so is not always straight-forward.

To adequately perform comparison studies between humans and machines, good practices have been developing. For example, to facilitate evaluating forward processing such as in a simple categorization task, fast stimulus presentation is recommended for humans[4] (Tang et al., 2018). The reason is that the biological process is believed to happen in under 150 msec (Thorpe et al., 1996; Serre et al., 2007b; DiCarlo et al., 2012). And hence, a presentation time in this order most adequately mirrors the forward pass in a feedforward DNN. Another aspect is that the increasing similarities between DNN and human psychophysical behavior demand more and more *challenging* experiments (Wichmann et al., 2017). After all, DNNs already surpassed human on simple tasks such as object categorization in 2015/6 (Kheradpisheh et al., 2016; He et al., 2015). As such, stimulus manipulations to increase the difficulty of standard input — as e.g. noise augmentations (Wichmann et al., 2017; Geirhos et al., 2018a) — have become a popular choice. On a more high-level note, Buckner (2019) warns of our human bias in comparison studies. Especially, the tendency of attribut-

[4]Tang et al. (2018) also recommend backward masking, however opinions about this practice diverge.

ing human-like characteristics to machines can be misleading. While the list of good practices is much longer, making good experimental choices is not always easy.

In summary, various differences between humans and machines can complicate comparison studies and open up challenges in experiments. While they are certainly a promising approach, how to adequately compare machines and humans is a tricky endeavor (see Section 2.1 which corresponds to publication one (Funke et al., 2021)).

1.5.2 Explainability method: feature visualization

Having looked at how modern artificial visual systems can be understood when comparing them to the human visual systems, the focus is now shifted toward an explainability method. They typically grant insight at a more fine-grained level of understanding: Loosely speaking, their goal is to make a machine's decision-making process more transparent and interpretable for humans. Referring back to the previously mentioned levels of understanding, explainability methods usually correspond to the algorithmic level, i.e. the one of processes and representations.

The beginnings of explanations for expert systems are associated with rule-based systems (Biran and Cotton, 2017) and the need for this was discussed as early as in the $1970's$ (Shortliffe and Buchanan, 1975). Today, explainable AI — often also called "XAI", "interpretability", or, though less frequently, "intelligibility" (Weld and Bansal, 2019) or "transparency" (Weller, 2019) — is a whole subfield in Machine Learning, and the community has been rapidly growing. For extensive information on XAI, the reader is referred to e.g. Gilpin et al. (2018) for an overview, Minh et al. (2021) for a comprehensive review, Guidotti et al. (2018) for a rather method-oriented survey, Carvalho et al. (2019) for a survey with a focus on methods and metrics, Molnar (2020) for a book and Lakkaraju et al. (2020) for a tutorial.

In this section, the explainability method "feature visualizations" is presented in detail as it is the technique of study in Sections 2.2 and 2.3 (corresponding to publications two (Borowski et al., 2021) and three (Zimmermann et al., 2021)).

Feature visualizations are synthetic images of the features which a deep convolutional neural network learns[5]. This means that they are intended to grant insight into the semantic properties of DNN units. In the literature, the method is sometimes also called "activation maximization" (Erhan et al., 2009; Nguyen et al., 2016a), and the images are also referred to as "most exciting inputs" (Walker et al., 2019).

First introduced by Erhan et al. (2009), the main idea is to iteratively update the pixels of a synthetic image via gradient ascent such that the activation of the network's unit in question becomes maximal. The resulting images often show interpretable (parts of) objects or geometrical structure (see Figure 2 left column). Given their generation procedure, synthetic feature visualizations are believed to isolate and highlight exactly what "causes" a unit's response (Olah et al., 2017; Schubert et al., 2021). This is in contrast to strongly activating natural data set samples, where any feature is unavoidably accompanied by many other image parts, and therefore under-

[5]Note that the method and the result of its generation process have the same name.

standing the truly underlying network feature can be challenging (see Figure 2 right column).

In the larger landscape of explainability methods, feature visualizations for DNNs belong to *post-hoc*, *model-specific* techniques and they reveal both *local and global* insights. Feature visualization's post-hoc explanation nature stands in contrast to *intrinsically interpretable models*, which are simpler though inherently interpretable models. The model-specific aspect means that they can be applied to the class of DNNs. In other words, feature visualizations are not model-agnostic and could not be applied to any model class. Moreover, feature visualizations are considered to grant local insight (Lakkaraju et al., 2020) as they use single data points. Because these data points simultaneously grant a wider impression of the whole decision-making process and the goal is not to obtain detailed insight for one single data point, the method can also be attributed a global notion.

Coming back to the generation procedure, the described vanilla optimization process for feature visualizations is usually augmented with a regularization mechanism. This is necessary because pure gradient ascent unfortunately yields images with high frequency artifacts. In the literature, three main approaches have evolved (Olah et al., 2017): For one, they directly target high frequency noise (Nguyen et al., 2015; Øygard, 2016; Tyka, 2016), e.g. by constraining the variance between neighboring pixels (Mahendran and Vedaldi, 2015). Alternatively, they introduce stochastic transformations such as jitter, rotation or scaling before updating the image (Mordvintsev et al., 2015; Tyka, 2016; Øygard, 2016). And as a third option, a (learned) prior can be integrated such that the final synthetic images look more photorealistic (Mordvintsev et al., 2015; Nguyen et al., 2016a, 2017; Wei et al., 2015). While no regularization mechanism has become a clear favorite, ways to improve the appearance of feature visualizations continue being explored.

Figure 2: **Feature visualizations**: This explainability method shows what features a DNN learned. Strongly activating data set samples can also reveal insights. Feature visualizations are produced with code from Olah et al. (2017) and data set samples are taken from ImageNet (Deng et al., 2009; Russakovsky et al., 2015).

A great advantage of feature visualizations is their flexibility — though as often, this also entails challenging decisions. As such, not only the final output units, but units from *any* layer can be visualized. In fact, that "unit" can be determined flexibly, too: Like in DeepDream (Mordvintsev et al., 2015), a whole layer can be subject to optimization, or just a channel or even only a single neuron. What is more, these units can be combined and jointly visualized. While this flexibility is an immense opportunity, finding suitable units or combinations thereof can be quite challenging (Olah et al., 2018), especially considering the

fact that *different* combinations of units can have strong responses to *similar* images (Szegedy et al., 2013). When further taking into account that even a *single* unit can respond to *different* features (Olah et al., 2017; Nguyen et al., 2016b; Olah et al., 2020b; Fong and Vedaldi, 2018), the complexity of selecting appropriate units becomes clear. To specifically account for the different features of these so-called "polysemantic" (Olah et al., 2020b) units, different strategies have been explored: Wei et al. (2015); Nguyen et al. (2016b, 2017) use diverse starting images for the generation process, and Olah et al. (2017) add a diversity term to the optimization objective such that the visualizations will differ from each other.

The advantages of feature visualizations and their potential for better understanding DNNs are reflected in the extensive use of the method: A large amount of research has gone into understanding the representations of the DNN InceptionV1 (Szegedy et al., 2015), which is also known as GoogLeNet, (Olah et al., 2017, 2020a,b,c; Cammarata et al., 2020, 2021; Schubert et al., 2021; Voss et al., 2021a,b; Petrov et al., 2021; Mordvintsev et al., 2015; Nguyen et al., 2016a). Researchers around Chris Olah focused on this model because they found it to be "unusually semantically meaningful" (Olah et al., 2018), though this finding is not consistent in the literature (e.g. Bau et al., 2017). Other studies investigated other networks (Nguyen et al., 2016a, 2017; Cadena et al., 2018; Tsipras et al., 2019; Engstrom et al., 2019a; Gonthier et al., 2020; Goh et al., 2021; Mahendran and Vedaldi, 2015; OpenAI, 2020), combined feature visualizations with different explainability methods (Olah et al., 2018; Carter et al., 2019; Addepalli et al., 2020; Hohman et al., 2019a) or built interactive tools with it (Wong et al., 2021; OpenAI, 2020; Sietzen et al., 2021). What is more, feature visualizations were successfully deployed in biological systems: After producing them in silico, these synthetic images elicited strong activations in mouse (Walker et al., 2019) and macaque visual cortex (Bashivan et al., 2019; Ponce et al., 2019).

Despite its popularity, feature visualization also has some downsides. For example, many synthetic images are difficult to interpret and do not convey a human-understandable concept. While this detachedness from human concepts reflects the DNN's features, an observer's interpretation is likely to be biased toward looking for familiar features. Considering the existence of adversarial examples, though, there is no guarantee that humans *could* recognize all features. Further, taking into account the finding of Fong and Vedaldi (2018) that in order to find semantically meaningful features, multiple units may be *required*, the challenge of combining DNN units sensibly comes back into the game. Simply considering the sheer number of DNN units and possible combinations thereof push obtaining a global understanding of a DNN out of reach. When shifting the focus to the often well-interpretable, "hand-picked" (Olah et al., 2017) synthetic images shown in publications, an open question is how representative and important they are for the network as a whole (Kriegeskorte, 2015).

Given these pros and cons of feature visualizations, one way to advance this debate is to evaluate how helpful the method is for humans. No matter how appealing an explainability method may seem, it is of practically no value, if it does not succeed in conveying an understandable explanation of the DNN to the target person. Measuring how well an explainability method fulfills its purpose is therefore crucial. Further,

it can help assess whether development progress is going in a good direction and whether the drawn conclusions from it are correct (Leavitt and Morcos, 2020).

"[H]uman-subject evaluation is not an easy task" (Doshi-Velez and Kim, 2017) and requires critical design choices. For example, "[t]he claim of the research should match the type of the evaluation" (Doshi-Velez and Kim, 2017). By now, various tasks have been conceived to target different questions regarding the helpfulness of an explainability method: They range from e.g. verifying a suggested response (Lage et al., 2019), declaring a preferred explanation out of two (Doshi-Velez and Kim, 2017), simulating a model's response for a new input given an explanation (Doshi-Velez and Kim, 2017; Lage et al., 2019) to creating a counterfactual given an input, an output and an explanation (Doshi-Velez and Kim, 2017). Other design choices concern but are not limited to the following aspects: Should experts or laypeople be inquired? Does the experiments take place in a well-controlled psychophysical lab or can the trials be posted on an online crowd-sourcing platform? How general or specific is the experiment to an explainability method, an algorithm and the data set?

For the method feature visualization, one human study has been performed — however, it does not address evaluating the helpfulness of feature visualizations. Instead, it tests the hypothesis whether one specific unit is really a curve detector. To this end, the first author Nick Cammarata decides whether more than 800 images correspond to the curve depicted in the feature visualization, or rather to an imperfect curve, an unrelated feature, or an opposing curve. Analyzing these labels reveals that the four concepts are roughly in line with the expected activation magnitude. This means that the unit usually fires most strongly for curves, less so for imperfect curves or unrelated features and least for opposing curves. As the authors further evaluate that most strong activations indeed occur for curves or imperfect curves, they conclude that their hypothesis is confirmed: The investigated unit is a curve detector. In the bigger picture, this human study is part of a deep-dive article on curve detectors (Cammarata et al., 2020) with the method by Olah et al. (2017). While it represents the first human study on feature visualizations and complements a suit of pure computational analyses, it is unrelated to evaluating the helpfulness of this method.

Despite the large body of work around feature visualizations, the informativeness of this explainability method has never been evaluated by humans. As such, it is unclear how well humans understand DNNs based on feature visualizations. Quantitatively measuring this would represent a big step forward.

1.5.3 Schematic overview of publications

Following a decreasing granularity level, approaches toward understanding visual perception in machines as well as their insights are presented. Specifically, comparison studies between humans and machines shed light from a coarse perspective, and an explainability method reveals more fine-grained information about DNNs. As such, the first publication addresses the problem described in Section 1.5.1 of how to adequately

compare humans and machines, and the second and third ones advance the dialogue around the informativeness of feature visualizations described in Section 1.5.2.

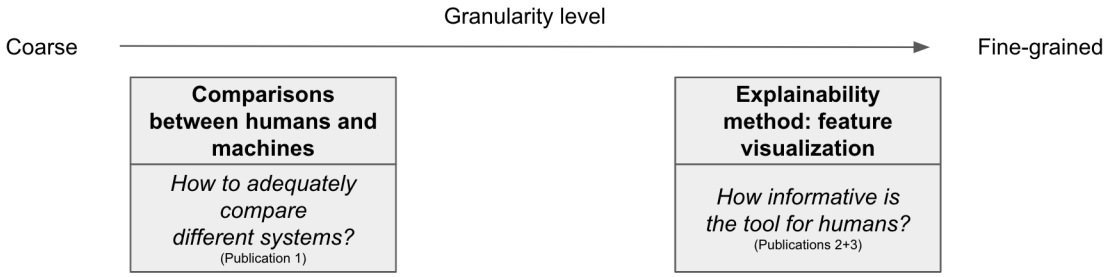


Figure 3: **Schematic overview of publications presented in this thesis**: Along decreasing granularity, comparison studies between human and machine behavioral data as well as human evaluations of the explainability method feature visualization are presented.

2 Papers

This chapter summarizes the main scientific contributions published as journal or conference papers that resulted from joint research during my PhD. In each of the following sections, I present one paper. Specifically, I explain what motivates the study as well as what main experiments we conduct, and what results we find. Finally, I discuss our work. This means that I not only set it into context of this thesis's story line, but I also highlight closely related literature and outline interesting future directions. Importantly, the discussions in this section go beyond the content of our publications.

On the content-level, each publication serves the overall goal of improving our understanding of visual perception in machine algorithms. In this thesis, I arrange the three papers by decreasing granularity: The first one addresses comparisons between machine and human visual perception on a behavioral level. By comparing performances of the two systems and varying a number of factors, we learn about DNNs' generalization abilities and limits as well as their inner workings. With the goal of understanding modern machine learning algorithms at an even more fine-grained level, the focus is next shifted toward an explainability method in the second and third paper. More specifically, this method is intended to grant insights into the *features* that a DNN learns. In our publications, we quantitatively evaluate it. Altogether, these experiments unveil how our understanding of DNNs' internal information processing improves based on an explainability tool.

In this chapter, all figures are taken from our original publications, and some, namely Figures 4-7, are slightly adapted. The complete papers themselves as well as the contributions from each author can be found in the appendix.

2.1 Five points to check when comparing visual perception in humans and machines

Christina Maria Funke*, Judy Borowski*, Karolina Stosio, Wieland Brendel‡, Thomas S.A. Wallis‡, Matthias Bethge‡. Journal of Vision, 2021.

2.1.1 Motivation

Comparing apples and oranges?![6] This idiom is commonly used to call out fundamental differences in comparisons. While it is generally good advice to avoid such scenarios, using them can sometimes be fruitful. In fact, comparison studies are frequently done in science. For example, when a system resembles a black box — like a modern machine learning algorithm —, it can be useful to look at model systems that we can interrogate — such as humans. And conversely, these artificial systems can be useful to model the human brain. Here, probing, taking apart and stitching back together is much easier. Taken together, the motivation for comparison studies is that new findings in one system can advance the understanding of the other system.

As for apples and oranges, comparison studies are not straightforward. Fields like comparative psychology have a long history of investigating animals to learn more about humans. Along the way, many tools and good practices were developed. They permit adequately comparing different systems as well as drawing robust conclusions. Unfortunately, these tools and good practices are not always applied in comparisons of modern machine learning algorithms. As a consequence, studies may come to rather fragile conclusions.

In this paper, we present both a checklist for comparison studies between two different systems as well as case studies in which we apply the checklist's points. In the latter, we often choose different assumptions and experimental designs than previous publications. As a result, we come to different conclusions regarding how DNNs work. These results influence our understanding of DNNs. On a meta-level, our experiments underline how tricky comparison studies are and that following good practices such as summarized in our checklist can be helpful. Taken together, this publication's contributions are two-fold: On the one hand, we present a checklist and investigate case studies with it. On the other hand, we discuss the impact of our new results and how they alter our understanding of DNNs.

2.1.2 Checklist for comparison studies

At first, we propose a checklist for comparison studies. It consists of five points on how to design, conduct, and interpret experiments that compare DNNs and humans. Below, a short explanation is given for each point. For the sake of brevity, the reader is referred to the original publication for examples. In terms of terminology, we often use "mechanism" as an umbrella term. By it, we refer to e.g. DNN architecture (e.g. feedback or lateral connections), learning schemes or the nature of representations.

[6] The first two paragraphs of this subsection are heavily inspired by earlier drafts that the authors wrote for the digital magazine on AI *The Gradient*.

i. **Isolate functional or implementational properties**: In order to understand the mechanism under investigation, experimental circumstances should be set such that the mechanism's effect will show as clearly as possible.

ii. **Align experimental conditions for both systems**: Despite unavoidable differences between different systems, experimental conditions should be aligned as much and as fairly as possible. Any remaining differences should be made explicit.

iii. **Differentiate between necessary and sufficient mechanisms**: Many mechanisms can lead to the same behavior. Therefore, stating clearly whether an experiment revealed if the investigated mechanism is necessary or just sufficient can help to create realistic expectations with respect to said mechanism.

iv. **Test generalization of mechanism**: To report adequately in which scenarios a certain mechanism can be deployed, testing its performance on various generalization data sets is essential.

v. **Resist human bias**: While we cannot remove our human bias, we should be aware of this potential influence. When designing and interpreting experiments, we should attempt to take on an objective perspective as much as possible.

2.1.3 Comparison case studies: experiments, results and discussions

Having suggested a five-point checklist for comparisons between different systems, we next apply these ideas in three case studies. The latter all aim to improve our understanding of visual perception in machines. Case study I is a thorough investigation of a certain stimulus type and illustrates how four of the five checklist points are put into action. Case studies II and III apply checklist points to experiments of previous publications. To this end, we run our own experiments, and come to different conclusions than the previous papers. On a meta-level, our case studies illustrate the following: (1) Human bias can influence the interpretation of experiments, (2) isolating and differentiating between necessary and sufficient mechanisms can be challenging, and (3) aligning experimental conditions is important to draw meaningful conclusions.

2.1.3.1 Comparison case study I: "Closed contour detection"

Motivation

In this first comparison case study, we want to test how well DNNs are able to distinguish open and closed contours. This task is known to be easy for humans. In fact, closed contours are believed to play an important role for the human visual system (Koffka, 1935; Elder and Zucker, 1993; Kovacs and Julesz, 1993; Tversky et al., 2004; Ringach and Shapley, 1996; Wertheimer, 1923). Their recognition is assumed to rely on a process called "global integration" which makes use of *global* information (Levi et al., 2007; Loffler et al., 2003; Mathes and Fahle, 2007). For DNNs, we hypothesize closed contour detection to be challenging. The reason is that their information processing was shown to heavily rely on local information (Geirhos et al., 2018a; Brendel and Bethge, 2018).

Experiments and results

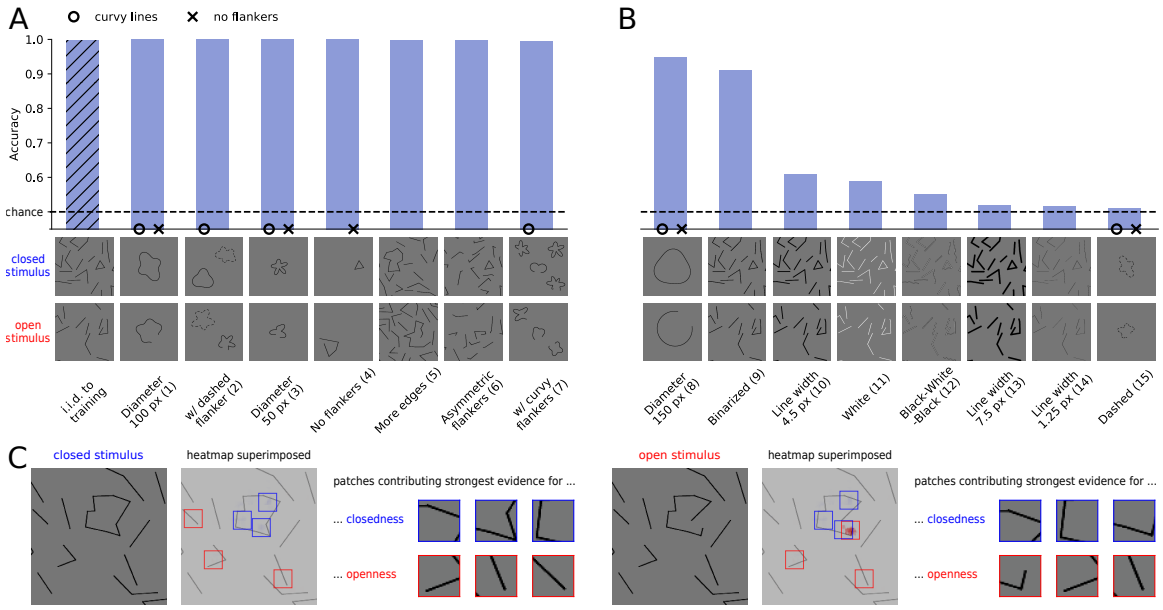


Figure 4: **Comparison case study I "Closed contour detection"**: As humans can easily detect closed contours, we test a DNN on this task. We find that our ResNet-50 separates closed and open contours surprisingly well (**A**, i.i.id to training). However, interpreting this result as if DNNs understood the human concept of closedness would be overhasty (point v): In generalization tests to stimulus variations (point iv), our DNN achieves high performance in only about half of them (**A** vs. **B**). This indicates limits of the investigated mechanism and differences to humans — their performance is expected to be consistently high. Finally, examining the heatmaps of a locally constrained DNN reveals insights into an alternative decision-making strategy (**C**).

Our experiments reveal that a DNN *can* successfully detect closed contours. Specifically, we train and test a ResNet-50 (He et al., 2016) on our custom data set. Its stimuli contain one main, i.e. open or closed, contour as well as multiple flankers (see Figure 4A i.i.d. to training). The network's performance is surprising: It achieves almost perfect accuracy. Relating this result to human psychophysical data, the two systems show similar behavior. As an interpretation, it would be enticing to conclude that DNNs would behave human-like. However, this would be overhasty (point v). To better understand the degree of similarity, we examine our model's performance in different generalization scenarios.

Generalization tests show that performance of our ResNet-50 without fine-tuning remains high only for about half of the additionally tested stimulus variations (see Figure 4A+B). This result contrasts our expectation of high human performance throughout stimulus modifications. Therefore, these generalization tests (point iv) indicate limits of our specific network and training procedure. Taking a step back, this helps moderate potential human bias in interpreting the previous result (point v).

In an additional experiment, we find that a purely local decision-making strategy suffices for a DNN to detect closed contours well. The motivation for this experi-

ment is to test an alternative mechanism to the human global integration process. Specifically, we evaluate a constrained model that has access to local features only: BagNet-33 (Brendel and Bethge, 2018) (point i). Our data reveals that this DNN also achieves good performance. Consequently, we derive that simple local information is sufficient for closed contour detection (point iii). Connecting this finding back to the human visual system, we reason that that biological global process is not necessary.

Via an in-depth analysis, we identify which local features BagNet-33 relies on most. To this end, we analyze the model's heatmaps (see Figure 4C). They indicate that a lot of evidence for open contours is attributed to image patches containing a short line connected to an open line.

Finally, it can of course not be assumed that our BagNet-33 model and our ResNet-50 model would utilize the same decision-making strategy. Generalization tests of our artificially locally constrained BagNet-33 reveal different performances than of ResNet-50. Thus, the two models seem to rely on different information processing strategies.

Discussion

In this first case study, our experiments unveil that modern machine learning algorithms *can* detect closed contours. However, they achieve high performance only on some and not all tested data set variations (point iv). Further, we discover that a DNN can base its decisions solely on local — as opposed to global — features (point i). Setting these results into context with humans, we infer that both the human visual system's global integration process as well as this DNN's local processing strategy are only *sufficient* mechanisms (point iii). The application of these checklist points illustrates how thorough experiments can help resist human bias when interpreting results (point v).

Zooming out, the results from this case study add to the bigger picture of which human psychological phenomena DNNs can or cannot mirror. As such, there is a big body of literature exploring DNN behavior on e.g. illusions (Gomez-Villa et al., 2019; Watanabe et al., 2018; Ward, 2019; Mély et al., 2018; Benjamin et al., 2019; Sun and Dekel, 2021; Baker et al., 2018a), crowding (Volokitin et al., 2017; Doerig et al., 2020; Lonnqvist et al., 2020) and other phenomena (see Section 3.3.1). Our presented results complement that DNNs — similar to humans — *can* solve tasks on closed vs. open contour stimuli. However, the extent of this ability is different to humans. Also, DNNs perceive these stimuli differently and other decision-making strategies than the human global integration procedure seem to be at play.

Related work

The importance of closed contours for the human visual system is mirrored by the Gestalt principle "closure". It describes that the human brain can fill in missing parts to create a whole. As such, stimuli can create the illusion of containing a closed shape, even though they physically do not.

Concurrent work investigates this closure phenomenon in DNNs (Kim et al., 2019, 2021a; Pang et al., 2021). To this end, they specifically employ illusory Kanizsa triangles (Kim et al., 2019, 2021a) and squares (Pang et al., 2021). These are stimuli

where the corners are either solid or implied with Pacman-shapes, and the connecting edges are removed. The authors find that at least some of the investigated DNNs can indeed detect closure on at least some of the tested stimuli. Overall, this result is similar to ours. Further, and again similar to our experiments, these works pre-train their DNNs on natural images before fine-tuning. While all authors report this to be important, they present opposing evidence regarding what kind of architecture permits the best performance: Kim et al. (2019, 2021a) state that convolutions are essential, whereas Pang et al. (2021) state that recurrency is the decisive factor. Finally, Kim et al. (2019, 2021a) go one step further and suggest that "Gestalt laws need not be considered as primitive assumptions underlying perception, but rather, that the laws themselves may arise from a more fundamental principle: adaptation to statistics of the environment."

In a different vein of work, Khan et al. (2020) investigate the advantages that mimicking biological properties provide for DNNs on contour integration. Specifically, they demonstrate that a "neuroanatomically grounded" model does not only successfully identify disconnected contours, but that it also exhibits realistic neurophysiological and behavioral properties. This latter aspect is in contrast to a pure feedforward DNN: While such a model achieves reasonable performance, too, it is "largely inconsistent with neurophysiological data" (Khan et al., 2020). In the bigger picture, this work adds to successful examples of taking inspiration from biology to improve DNNs.

Future directions

In order to further deepen our understanding of how artificial algorithms process closed contours, our work can be extended in several ways.

For example, the local processing mechanism of BagNet-33 can be subject to further investigation. Specifically, the statistical subtleties that this algorithm exploits can reveal new insights. To this end, a first step is sorting image patches according to their contributing evidence and analyzing repeating patterns. Going further, the data set can be manipulated and the stimuli's most important features can be removed. In the current version of the data set, this would entail changing the images to not contain a short line with an open ending anymore (see Figure 4C). With new edge lengths, answering questions like the following would be interesting: What "next-best" kind of features will BagNet-33 base its decisions on? Will they reveal new statistical biases in the data set? Zooming out, can an iterative process of improving the stimuli be started? And where would this lead given that a neural net with at least one hidden layer is already a universal approximator (Cybenko, 1989)?

A second direction toward better understanding the processing of closed contours in DNNs is to investigate why our two algorithms perform much more poorly on certain stimulus variations. As such, accuracy drops to (almost) chance level for different color and line-widths. A similar effect for these kinds of pixel-level factors is found for other kinds of stimuli in the literature (e.g. Puebla and Bowers, 2021). As this aspect seems unrelated to the opposing global vs. local decision-making strategies, investigating it may reveal other useful mechanisms helpful in detecting closed contours.

Summary
In the bigger picture of understanding DNNs, the experiments of this first case study provide evidence that DNNs *can* detect closed contours. However, the extent of this ability and the way information is processed are different to humans.

Zooming out, the techniques utilized in this comparison study represent complementary ways to investigate DNNs. They include but are by far not limited to measuring performance in generalization scenarios (point iv) and scrutinizing the decision-making process with a specifically designed network (point i). Findings from such investigations can point to the machine learning model's limits as well as its learned strategies. In general, such thorough examination can help us better understand differences between DNNs and humans. On the whole, this mitigates our human bias (point v) and improves our understanding of visual perception in machines.

2.1.3.2 Comparison case study II "Synthetic Visual Reasoning Test"

Motivation

Understanding how well machine systems can reason about abstract visual relations is of broad interest (Barrett et al., 2018; Yan and Zhou, 2017; Zhang et al., 2019; Villalobos et al., 2020). For humans, this ability is a hallmark of their intelligence (Barrett et al., 2018).

Previous research demonstrates that pure feedforward DNNs fall short in one of two task categories from a popular data set (Stabinger et al., 2016; Kim et al., 2018b). This "Synthetic Visual Reasoning Test" (SVRT) consists of 23 tasks and was developed precisely for comparing humans and machines (Fleuret et al., 2011). The tasks require either a decision based on whether shapes in two stimuli are identical (same-different tasks) or a decision based on whether spatial relations between shapes in two stimuli are similar (spatial tasks; for examples see Figure 5A). Fleuret et al. (2011) found that humans understand the underlying rules quickly, usually after a few examples. In contrast, two other studies (Stabinger et al., 2016; Kim et al., 2018b) tested DNNs (GoogLeNet (Szegedy et al., 2015) and $2-6$ layer nets, respectively), and detected high performance only on spatial but not same-different tasks (see Figure 5B, left side). Furthermore, Kim et al. (2018b) observed that *learning* same-different tasks is more difficult. The latter authors interpreted their results as "feedforward neural networks' fundamental inability to efficiently and robustly learn visual relations". Other works (Serre, 2019; Schofield et al., 2018) more broadly hypothesize that DNNs might need feedback mechanisms to be able to learn same-different tasks.

Experiments and results

In our own experiment, we find evidence that the task category, that was previously thought to be difficult, represents no inherent limitation for modern, pure feedforward DNNs. Specifically, our ResNet-50 achieves $> 90\,\%$ accuracy on all tasks (see Figure 5B). Besides showing that pure feedforward DNNs *can* succeed on same-different tasks, this result can further be interpreted as follows: Feedback mechanisms are not necessary for same-different tasks (point iii). Nonetheless, when considering that

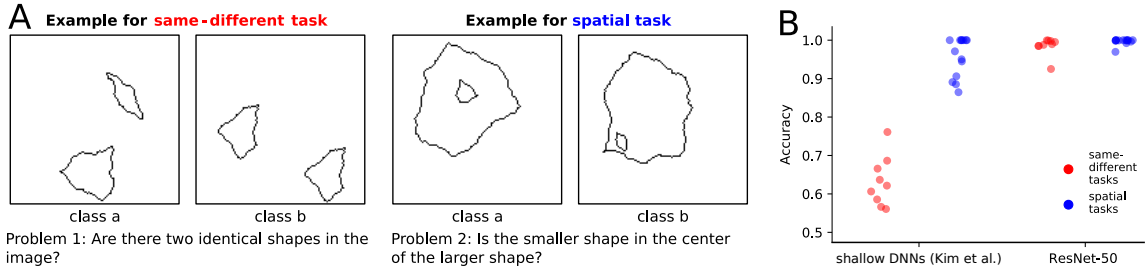


Figure 5: **Comparison case study II "Synthetic Visual Reasoning Test"**: Knowing humans can easily comprehend abstract visual relations, the data set "Synthetic Visual Reasoning Test" was developed (Fleuret et al., 2011) to evaluate both human and artificial machine perception. It consists of tasks targeting the evaluation of whether shapes are identical (same-different tasks), and tasks requiring a decision regarding the spatial relation of shapes (spatial tasks; for examples tasks see **A**). Previous studies (Kim et al., 2018b; Stabinger et al., 2016) showed poor performance for shallow, pure feedforward DNNs on same-different tasks (**B**, left red dots). This led researchers to hypothesize that feedback mechanisms would be necessary. However, as our deeper, pure feedforward ResNet-50 model achieves high performance on same-different tasks (**B**, right red dots), this result can be interpreted as if feedback mechanisms are *not* strictly necessary (point iii).

finite-time recurrent neural networks can be rolled out into pure feedforward neural networks (Liao and Poggio, 2016; van Bergen and Kriegeskorte, 2020), this point remains debatable. Finally, regarding different learning efficiencies, we claim that they are a tricky argument to support the hypothesis of feedback connections being necessary. After all, it is possible that humans would find same-different tasks more difficult to learn, too. However, due to their lifelong exposure to visual input and learning, this is impossible to assess.

Discussion

In this second case study, our experiments unveil that modern, pure feedforward DNNs can correctly categorize *both* task types of a popular abstract visual reasoning test. This result contrasts previous works and claims: Other researchers believe feedback mechanisms would be *necessary* for a DNN to solve the task type of distinguishing whether shapes are identical or not (Kim et al., 2018b; Serre, 2019; Schofield et al., 2018). Our findings indicate that feedforward connections are *sufficient* for this (point iii).

Zooming out, our results add to the bigger picture of where DNNs stand with respect to abstract visual reasoning. Moreover, they illustrate what mechanisms are necessary and sufficient for such tasks. Multiple works in the literature (e.g. Barrett et al., 2018; Yan and Zhou, 2017; Zhang et al., 2019; Villalobos et al., 2020) measure DNN performance on custom data sets targeted at e.g. symmetry recognition (Yan and Zhou, 2017). Often, DNNs are discovered to not be able to succeed at all tasks. Yet, specific architectures designed to support reasoning are demonstrated to help

overcome these limitations (Barrett et al., 2018; Zhang et al., 2019; Villalobos et al., 2020). Compared to humans, for whom abstract reasoning is an essential element of intelligence, these abilities and mechanisms are different.

Related work

The body of literature specifically investigating DNNs on SVRT's same-different tasks is growing quickly: On the one hand, a lot of focus is put on exploring and developing new architectures based on different mechanisms. On the other hand, better understanding and augmenting the data set is becoming a new subject of study. This is possible because previous data set versions are coming within reach.

To start out, an overview of the investigated architectures and their performances is given. Similar to the experiments of Stabinger et al. (2016); Kim et al. (2018b) and ours, a couple of works replicate the finding that shallow feedforward DNNs cannot perform well on same-different tasks (Messina et al., 2019; Messina et al., 2021b), but that a ResNet-inspired DNN *can* (Messina et al., 2019). Further, a few works test the Relation Network (Santoro et al., 2017) or a DNN inspired by it (Kim et al., 2018b; Puebla and Bowers, 2021; Messina et al., 2021a). This is an architecture specifically designed for relational reasoning. While Kim et al. (2018b) and Puebla and Bowers (2021) demonstrate that this model performs similarly to pure feedforward DNNs, Messina et al. (2021a) show that it is outperformed by their newly designed DNN. They call this latter successful model "Recurrent Vision Transformer" (Messina et al., 2021a). It is inspired by different transformer models and contains recurrent connections, an attention mechanism as well as a convolutional feature extractor. On same-different tasks, its performance is high, often almost perfect. In total, this is the third study on the SVRT data set from this group. In it, Messina et al. (2021a) also highlight that a pure Vision Transformer cannot learn same-different tasks. Earlier work of theirs lays out that evidence for the importance of residual, recurrent and more generally skip connections varies, and that network depth would not play an important role (Messina et al., 2019; Messina et al., 2021b). Focusing on attention, the group around Thomas Serre finds that DNNs *with* attention mechanisms, be they spatial or feature-based, permit higher performance in 22 out of the 23 SVRT tasks (Vaishnav et al., 2021).

Zooming out from the debate of what specific architectural mechanisms enable DNNs to perform same-different tasks, researchers inspect the data set itself: As a first observation, Stabinger et al. (2016) suspect that the generation procedure of the SVRT images allows their DNN to pick up other than the intended cues ("shortcut-learning" (Geirhos et al., 2020a)). In later work, Stabinger et al. (2021) argue that the data set should be made more difficult. Concretely, they suggest the shape generation procedure should be adjusted to prevent trivial clues giving the correct answer away. While the early study of Kim et al. (2018b) already introduces a parameterized version of the SVRT data set with varying task complexity, Puebla and Bowers (2021) go even further: In their new generalization test set, stimuli differ on the pixel-level. Here, for example, line widths may vary — a modification that we also investigated in our closed contour comparison case study.

Taking the research on architectural mechanisms and data sets together, it seems that modern, pure feedforward DNNs can solve same-different tasks, as long as they are tested on i.i.d. data (Messina et al., 2019). When moving to the out of distribution (o.o.d.) regime (Puebla and Bowers, 2021), not only DNNs like a ResNet-50 but also specifically augmented architectures like a Relation Network fail. Overall, a number of studies (Messina et al., 2019; Messina et al., 2021a,b; Vaishnav et al., 2021) suggests that feedback mechanisms and especially attention can help with same-different tasks. This view is endorsed in both a couple of review articles (Lindsay and Serre, 2021; Stabinger et al., 2021) as well as in studies on other abstract visual reasoning tests, such as in e.g. Barrett et al., 2018 and Villalobos et al., 2020.

To summarize, progress is being made toward DNNs performing well on abstract visual reasoning tasks. Based on the current data, categorizing feedback mechanisms as necessary or sufficient depends on the experimental conditions and therefore remains an open question.

Future directions

To further improve our understanding of DNN behavior on abstract visual reasoning, different directions can be explored.

As a first step, the portfolio of DNN capabilities on new SVRT data set versions can be augmented. To this end, the most promising model by Messina et al. (2021a), the Recurrent Vision Transformer, can be evaluated on the most challenging o.o.d. data set by Puebla and Bowers (2021). This experiment would reveal whether one of the most modern kind of DNN models may have an edge in o.o.d. regimes, which are currently unreachable for less recent DNN versions.

Other interesting directions are investigating what data set augmentations during training as well as how meta-learning can improve generalization. Finally, developing a deeper understanding of the decision-making strategies can reveal insights into the aspects that give certain mechanisms an advantage. In the bigger picture, this can add to the understanding of which mechanism is necessary and which one is sufficient.

Summary

Within the goal of understanding machine visual perceptions, the experiments of this case study suggest that modern, pure feedforward DNNs *can* solve the abstract visual reasoning task of categorizing images into containing same or different shapes.

Taking a step back, this comparison study and its discussion illustrate that training DNNs is a complex endeavor and involves many choices including but not limited to architecture, network depth and width, regularization schemes and the optimizer. Moreover, generalizing results beyond the tested setup is tricky and it is essential to differentiate between necessary and sufficient mechanisms (point iii).

2.1.3.3 Comparison case study III "Recognition gap"

Motivation

This third comparison case study investigates the minimally necessary visual information required for object recognition. Expressed as small image crops, the question is how the human and machine visual systems behave on them. In the bigger picture,

this performance as well as the nature of these minimal features grant insights into how DNNs as well as humans achieve object recognition.

A previous study claims that humans and machine algorithms would utilize different features and processes for object recognition (Ullman et al., 2016). In their experiment, Ullman et al. (2016) successively cropped or reduced the resolution of images until humans could not recognize them anymore. Interestingly, they discovered that human recognition performance drops sharply if a smallest recognizable crop is reduced any further. The authors termed this sudden transition the "recognition gap". Specifically, it is evaluated as the difference between the fraction of people who correctly identify the minimal recognizable crop (e.g. 0.9) and the fraction of people who correctly identify the next smaller, i.e. maximal *un*recognizable, crop (e.g. 0.2). In the given example, the recognition gap would evaluate to $0.9 - 0.2 = 0.7$. For humans, the authors found a *large* recognition gap: 0.71 ± 0.05 (see Figure 6B). In contrast, for a number of machine algorithms, they found a *small* recognition gap: 0.14 ± 0.24. Notably, these latter models were evaluated on *human*-selected stimuli. Ullman et al. (2016) interpreted their results as "that the human visual system uses features and processes that are not used by current models and that are critical for recognition".

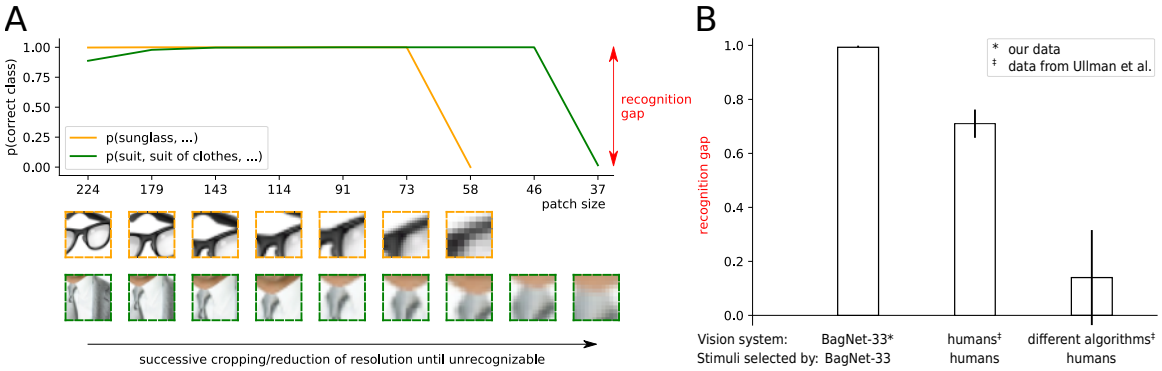


Figure 6: **Comparison case study III "Recognition gap"**: With the goal of identifying the minimal information necessary to recognize an object, Ullman et al. (2016) discovered that humans' ability to recognize small image crops drops precipitously. They call this phenomenon the "recognition gap". It is measured as the fraction of humans that correctly classify the recognizable crop minus the fraction of humans that correctly classify the *un*recognizable crop. While Ullman et al. (2016) found this recognition gap to be large for humans (**B**, middle bar), they identified it to be small for machine visual systems evaluated on the same human-selected stimuli (**B**, right bar). In our experiment (**A**), we implement a search algorithm with BagNet-33 that closely mimics the human procedure of Ullman et al. (2016) (point ii). It reveals this machine algorithm's recognition gap to be *large* (**B**, left bar). As a consequence, we conclude that both the human and a machine visual system can recognize small image crops.

Experiments and results

In our experiment, we align experimental conditions (point ii) and find that object recognition between humans and a DNN is more similar than previously thought. In particular, our goal is to investigate if at least part of the differences shown by Ullman et al. (2016) might be explainable by different experimental conditions. Therefore, we implement the search procedure for the smallest recognizable and the largest unrecognizable image crop such that it mimics the original human psychophysical experiment (see Figure 6A). As shown in Figure 6B, we find a *large* recognition gap of 0.99 ± 0.01 for our machine algorithm, a BagNet-33 (Brendel and Bethge, 2018). Crucially, we evaluate it on image crops *it* selected. Compared to the human recognition gap evaluated on human-selected crops by Ullman et al. (2016), this gap is similar.

Discussion

In this third case study, our experiment unveils that a modern DNN *can* recognize small image crops and that this capability drops sharply when reducing a crop just a little further. Compared to a previous study (Ullman et al., 2016), this finding is similar to the one for humans and contrasts the one for machines: Ullman et al. (2016) claimed that humans would exhibit a large recognition gap, but not machine algorithms. The crucial change for our new result is aligning experimental conditions (point ii): We evaluate our DNN on crops *it* selected — just like Ullman et al. (2016) evaluated humans on crops *they* selected. These two recognition gaps turn out to be similar in size. Tying this comparison case study back to our checklist, it illustrates that aligning experimental conditions can influence the results (point ii).

Zooming out, our findings add to the bigger picture of what kind of information the human and machine visual system can successfully process. Specifically, we conclude that *both* the human *and* the machine visual systems can recognize small image crops. Also, a sudden drop in recognizability occurs in *both* systems. This means that reducing the amount of information just a little too much has a large effect on both human and machine recognition capabilities. This interpretation is different to the statement from Ullman et al. (2016) (see the quote above) and clarifies that our human perspective can bias our interpretation (point v).

Related work

In the literature, different parallels to this third case study can be drawn. Here, I first highlight previous works that also pursue the goal of simplifying input images while maintaining correct classification. Then, I outline how the work of Ullman et al. (2016) is directly being extended.

The idea of identifying the minimally necessary information for successful object recognition has been studied in both biological and artificial systems for quite a long time. At the core, information from the original image is removed and the remaining crop is tested for recognizability. In 1995, Biederman (1995) reviews such ideas for the behavioral level in humans. Even earlier, namely in 1993, Tanaka (1993) summarizes and presents a similar procedure for the cellular level in primates. Its goal

is to understand receptive fields in inferior temporal cortex (IT). Regarding machine visual systems, Zhou et al. (2015) perform a study fairly recently, in 2015: Specifically, they investigate *scene-classifying* DNNs. To arrive at the minimal necessary image area for successful recognition, the researchers iteratively remove unimportant image segments. They uncover that in e.g. a bedroom-scene, the bed itself remains visible while other furniture or decoration does not. As a consequence, Zhou et al. (2015) infer that scene-classifying DNNs learn recognizing objects as a by-product.

The body of literature immediately extending the work by Ullman et al. (2016) is growing. For example, in 2019, Srivastava et al. (2019) again investigate minimal image crops for humans and machines. Similar to us, they implement a machine-based search algorithm. However, their procedure differs greatly from the original human psychophysical experiment and only a subset of their identified crops corresponds to the definition introduced by Ullman et al. (2016). On a subset of this subset, Srivastava et al. (2019) measure a moderate recognition gap of 56 % for their DNN. Despite the differences in their search procedure, these results again highlight that humans *and* DNNs are susceptible to small image changes. Going further, Srivastava et al. (2019) additionally test humans on their machine-selected stimuli. Here, they detect a small recognition gap of only 14.5 %. This suggests that the content of small crops differs between the human and the machine visual system.

In a series of works, Shimon Ullman and his colleagues themselves add to their findings from 2016. For example, they apply their idea of identifying the minimally necessary information to videos (Ben-Yosef et al., 2020, 2021). As for static images, they find a large drop in recognizability for humans, but only a small recognition gap for DNNs. However, just like in their previous work, they evaluate machine algorithms on *human*-selected data. This means that they do not align experimental conditions (point ii). As a consequence, it is an open question if their results can be replicated when letting machine algorithms determine their own minimal videos. Regarding the work on *static* minimal recognizable images, different directions are pursued: For example, Benoni et al. (2020) examine the time trajectory of the recognition process, and Holzinger et al. (2019) investigate the brain activations of category-selective areas in an fMRI study. Furthermore, Ben-Yosef et al. (2018, 2017) create a model that automatically provides a "full interpretation" of the minimal recognizable images. Finally, a review-like article summarizes image interpretation above and below the object level (Ben-Yosef and Ullman, 2018).

To summarize, the body of work around the minimal information necessary for successful recognition is growing. However, as experimental designs differ a lot, the robustness of the identified similarities and differences between humans and machines remains an open question.

Future directions

In pursuance of deepening our understanding of how humans and DNNs process small pieces of information, different follow-up directions can be examined.

Specific to our implementation of the machine-based search for minimal crops, there is potential to align experimental conditions even more. As such, programming an exhaustive search instead of always following the best-performing crop would mir-

ror the psychophysics from Ullman et al. (2016) more closely. Whether such additional effort would change the result is questionable. The undeniable advantage of such a more complicated procedure would be identifying *several* minimal recognizable image crops for each original image. Going further, the diversity of these crops can be analyzed and compared to the human ones. Similar to Srivastava et al. (2019), this may reveal additional similarities and differences between humans and machines.

Zooming out, different experiments can be performed to deepen our understanding of what the phenomenon "recognition gap" really means. Within the visual domain, more modern DNNs such as transformers can be tested. This would reinforce or diversify our understanding of how machine algorithms process small image crops. In a different direction, exploring different data modalities may turn out insightful. For example, is a sharp drop in recognizability present for text or speech data? And, more specifically, do both humans *and* machine algorithms exhibit a recognition gap for these types of input? Either way, this may reveal interesting observations about different data modalities and the processing of small pieces of information.

Summary
Within the overall goal of improving our understanding of visual perception, the experiment of this case study indicates that a DNN - similar to humans - can correctly recognize small image crops. Further, it underlines that both systems behave similarly on crops at the edge of recognizability.

In the bigger picture, this third case study illustrates that unequal testing procedures can confound conclusions. Therefore, aligning experimental conditions between systems as much as possible is of utmost importance (point ii).

2.1.4 Discussion

Comparison studies are notoriously difficult. Designing, conducting and interpreting experiments for two different systems can be challenging. Nonetheless, such investigations can reveal great insights into their similarities and differences. In our checklist, we presented good practices that can help prevent pitfalls of comparison studies. In three comparison case studies, we illustrated the application of its points and thereby deepened our understanding of visual processing in DNNs.

Going further, the presented case studies not only reflect our checklist suggestions but also other good practices. As mentioned in Section 1.5.1, recommendations include for example (1) short presentation times for forward processing comparisons (Tang et al., 2018; Thorpe et al., 1996; Serre et al., 2007b; DiCarlo et al., 2012), (2) creating challenging stimuli (Wichmann et al., 2017), and (3) limiting human bias (Buckner, 2019). In our closed contour detection case study, (1) was realized with a stimulus presentation time of 100 msec, and (2) and (3) were addressed by variations of the closed and open contour data set. In our recognition gap study, (3) was accomplished by designing a new experiment for a machine algorithm and thereby aligning experimental conditions fairly to both humans and DNNs (point ii of our checklist). Regarding the related work on the SVRT data set of the second case study, the stimulus variations of Kim et al. (2018b) and Puebla and Bowers (2021) can further be

seen as realizations of (2), i.e. making similarities and differences visible via creating more challenging experiments.

Overall, many good practices exist. Some have even become so much second nature that mentioning them explicitly is forgotten. While our checklist complements the mentioned existing good practices, concurrent work has been developing even more. For a discussion of a few of those, the reader is referred to Section 3.1.2.

To summarize, comparison studies are powerful, yet they require care. When thoroughly designed, conducted and interpreted, they can reveal great insights into the two investigated systems. As such, this publication deepened our understanding of DNN's visual information processing with respect to closed contour detection, same-different classification and behavior on minimally necessary information. Using the language from above (Section 2.1.1), comparing the fundamentally different systems of humans and machines — such as apples and oranges — turned out fruitful.

2.2 Exemplary natural images explain CNN activations better than state-of-the-art feature visualization

Judy Borowski*, Roland Simon Zimmermann*, Judith Schepers, Robert Geirhos, Thomas S.A. Wallis‡, Matthias Bethge‡, Wieland Brendel‡. ICLR, 2021.

This and the next paper belong together and they shift gears from better understanding machine visual perception on the basis of comparison studies to the explainability method feature visualization. This latter tool is specifically designed to grant insights into the inner workings of a DNN. Within the overall story line, this means that the next two papers reflect a finer level of granularity than the previous publication. In essence, the goal of the feature visualizations is to make the *features* that a deep convolutional neural network learns understandable.

Similar to the previous paper, we here also collect human behavioral data — however, the use is different. In the previous paper, we compared human behavioral performances directly to DNN performances. In the next steps, we then drew conclusions about their abilities, limits, features and mechanisms. In these works, we interpret human accuracy as an indicator of the explainability method's helpfulness. In short: we *evaluate* feature visualizations. More precisely, human participants perform specific tasks with the explainability tool as clues. The logic is that high performance indicates that the explainability method indeed helps humans in the task, whereas chance performance indicates that it does not.

An evaluation study of an explainability method is different in nature from a comparison study. Despite differences, and as alluded to in the previous paragraph, there are parallels between the two. As such, most of our points from the checklist for comparison studies are reflected in the evaluation experiments. The only point that does not apply is point ii of aligning experimental conditions between humans and machines. As a matter of fact, this point is irrelevant for evaluating an explainability method with humans. For a detailed discussion of the checklist points with respect to our evaluations of an explainability method, please see Section 3.1.1. Without doubt, comparison studies and explainability tools are generally two complementary approaches to deepen our understanding of DNNs.

In this and the following publications, the focus is on the popular feature visualization method by Olah et al. (2017). It counteracts high frequency artifacts by performing gradient steps in Fourier space as well as applying jittering, rotating, scaling, padding and cropping transformations to the image before updating it. In order to visualize different facets of a feature, Olah et al. (2017) use a diversity term. Please note that I refer to only this one method of feature visualizations throughout this thesis when using the term "feature visualization" — unless I explicitly remark otherwise.

2.2.1 Motivation

The artificial images of the explainability method "feature visualization" show what maximally activates a certain unit in a DNN (see Section 1.5.2). As such, they are intended as a lens into the *features* that a DNN learns.

Despite their promising insights, there is a great controversy around them and opinions in the research community diverge: On the one hand, feature visualizations are extensively used (Olah et al., 2020b,a; Cammarata et al., 2020; Cadena et al., 2018; Mahendran and Vedaldi, 2015; Nguyen et al., 2015; Mordvintsev et al., 2015; Nguyen et al., 2016a, 2017; Tsipras et al., 2019; Engstrom et al., 2019a; Olah et al., 2017; Nguyen et al., 2019) and even combined with other techniques (Olah et al., 2018; Carter et al., 2019; Addepalli et al., 2020; Hohman et al., 2019b) to better understand the learned representations and decision-making processes of DNNs. In fact, many scientists believe that "features can be rigorously studied and understood" (Olah et al., 2020b) and that features are meaningful (Zhou et al., 2014; Bau et al., 2017, 2020).

On the other hand, though, several aspects around feature visualizations raise doubts: As an example, pure vanilla methods produce noisy and "nonsensical high-frequency" images (Olah et al., 2017), which is why introducing regularization mechanisms has become the norm (Mahendran and Vedaldi, 2015; Offert and Bell, 2020; Nguyen et al., 2017, 2015; Mordvintsev et al., 2015). While this makes the artificial visualizations more human-understandable, their faithfulness gets impaired and it is a difficult question how to choose the best regularization mechanism. Other open challenges concern the following aspects: What is the appropriate "unit" that represents a single feature in a network? A single neuron, several neurons or a whole channel? And how much can we learn from a single feature, given that Morcos et al. (2018) suggested that units of an easily understandable feature play a less important role compared to units that respond to several inputs? Or how can we guarantee to generate diverse enough feature visualizations to account for e.g. "polysemantic" neurons, i.e. neurons that fire for different, unrelated inputs? Can we know that *maximizing* the activation is the right choice or should we also look for ways to visualize features that elicit e.g. only 70 % of the maximal activation? So far, the development of explainability methods often relies on intuition, a criticism raised by Leavitt and Morcos (2020). They further lament that falsifiable hypotheses are missing. Last but not least, as e.g. Kriegeskorte (2015) writes, it is unclear how representative the appealing, most likely "hand-picked" (Olah et al., 2017) images in articles are of the entirety of a whole network.

In this project, our idea is to further advance the discussion around feature visualizations by quantitatively evaluating the informativeness of one such method. To the best of our knowledge, this study is the first to evaluate feature visualizations with humans and to test intermediate representations[7].

[7]Post-publication, I realized that Bau et al. (2017) evaluated an interpretability method on all five convolutional layers of AlexNet.

2.2.2 Experiments and results

To measure how helpful feature visualizations are for humans, we conduct two well-controlled human psychophysical experiments in our lab. Specifically, we test the method developed by Olah et al. (2017) and ask participants to simulate a DNN's behavior. The task (see Figure 7A) is to choose one out of two natural query images (two-alternative forced choice (2AFC) paradigm (Fechner, 1860)) that the participants expect to elicit a strong unit activation given synthetic reference images. To set participant performance into context besides the chance level of 50 %, we additionally test baseline conditions such as natural data set samples.

Our main result is that even though synthetic images do provide humans with helpful information about feature map activations ($82 \pm 4\,\%$), natural images are even more helpful ($92 \pm 2\,\%$, see Figure 7). Further, we discover that this superiority of natural images mostly holds across various conditions. These span different network parts (layers and inception module branches), participant expertise levels, hand- and randomly-picked feature visualizations as well as different presentation schemes of reference images. As to the comparisons between expert and lay participants as well as between hand- and randomly-picked feature visualizations, we do not find the corresponding differences to be statistically significant. As to the presentation schemes, our experiments unveil that providing several reference images as well as presenting both minimally *and* maximally activating images improve human performance. Finally, our data demonstrates that subjective impressions of feature visualizations' interpretability vary greatly between participants.

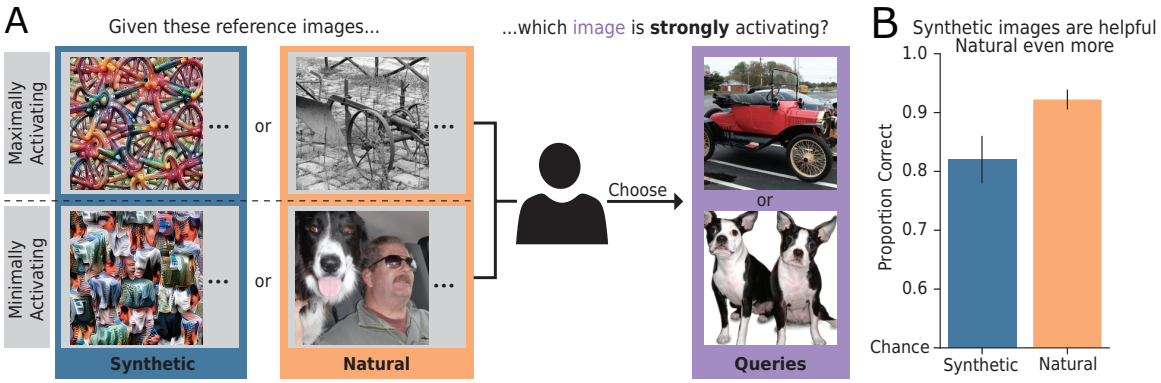


Figure 7: **Informativeness of feature visualizations in feedforward prediction task**: In our psychophysical experiments, humans are presented either synthetic feature visualizations or extremely activating natural data set samples. Based on these references, their task is to decide which of two natural query images elicits a stronger activation in the given DNN unit (**A**). Our main result is that while humans achieve more than chance performance with synthetic images, natural images are even more helpful (**B**).

2.2.3 Discussion

Feature visualizations via activation maximization are a popular explainability method to better understand the behavior of DNNs. At the same time, they are criticized for being intuition-driven (Leavitt and Morcos, 2020). Here, we quantitatively evaluate the helpfulness of these synthetic visualizations with humans for the first time: How useful are synthetic feature visualizations for humans in predicting which image elicits stronger DNN activation?

Our experiments reveal that even though feature visualizations by Olah et al. (2017) do provide useful information, natural images are even more helpful. On the most basic level, the above chance performance for both conditions indicates that humans *can* simulate the preference of a DNN unit with the help of extremely activating images. In other words, we can gain an understanding about the inner workings of a machine algorithm. Going further, the fact that natural data set samples represent the more helpful source of information implies that this understanding is deeper with natural and not synthetic feature visualizations.

Beyond our main result, our findings shed light on a number of other aspects. These concern the expert level, the representativeness of hand-picked visualizations, the presentation scheme as well as subjective impressions of the helpfulness of feature visualizations.

To the best of our knowledge, our study is the first one to compare expert and lay people in an evaluation study of an explainability method. In existing work, publications focus either on experts (Hase and Bansal, 2020; Kumarakulasinghe et al., 2020) or on laypeople (Schmidt and Biessmann, 2019; Alufaisan et al., 2021). Our data shows no significant difference between the two groups. Consequently, we suggest that future experiments may not have to rely on expensive expert participants. Instead, leveraging lay participant pools may be sufficient.

Further, we test whether hand-picking particularly appealing feature visualizations allows participants to achieve higher performance. The motivation for this is to evaluate whether the visualizations shown in publications represent the general level of understandability conveyed via the explainability method. Our data indicates no significant difference between hand-picked and randomly-picked feature visualizations. Therefore, the selection manner seems to be a minor aspect. Nonetheless, I note that the gap is fairly large (see publication, Figure 5B) and particularly larger than for the natural condition. Consequently, investigating this hypothesis with more data can bring more certainty.

Regarding the best presentation scheme of reference images, our experiments indicate that providing several reference images as well as presenting both minimally *and* maximally activating images improves human performance. This finding is in line with previous publications. For example, Offert (2017) and Kim et al. (2016) respectively suggest that more than one example and particularly negative examples help humans develop an understanding of data. Finally, how to best choose data set samples is an active area of research (Crabbé et al., 2021).

With respect to how intuitive participants perceive synthetic feature visualizations, we discover that opinions vary greatly. Some participants find certain synthetic images straight-forward, others find them difficult to interpret. This result is similar to the one of Hase and Bansal (2020) and suggests that each explainability method has to be tested individually. Overall, these intuitiveness impressions add to the picture of other often evaluated aspects. As such, researchers often measure e.g. trust, satisfaction or confidence (Alqaraawi et al., 2020; Schmidt and Biessmann, 2019; Kumarakulasinghe et al., 2020). After all, a machine algorithm should only be deployed if people trust it and its explanations.

Future directions
To extend the evaluation of feature visualizations, future work can explore different directions. For example, the specific statement from our concrete setup can be extended to a more general one. To this end, the very first and very last layers of the InceptionV1 network, combinations of units, the neuron objective, as well as different networks and other feature visualization methods can be investigated. Moreover, the amount of information provided to participants can be varied. For instance, the signal from the query images can be decreased by sampling them from less extreme activations. Alternatively, additional information can be provided by sharing which layer a unit is taken from. With the described undertakings, a more general statement about the helpfulness of the investigated explainability method would come into reach.

Summary
In the broad view of understanding DNNs, the experiments of this study demonstrate that human participants can comprehend the inner workings of a DNN well enough such that they can simulate its feedforward behavior. Specifically, they develop this understanding with extremely activating images. As such, we evaluate synthetic images from the popular explainability method feature visualizations as well as natural data set samples. Most of our evidence hints toward natural images being more helpful. Overall, this study illustrates that humans can extract helpful information from the two visualization types of natural and synthetic images and gain some understanding of visual perception in machines.

2.3 How well do feature visualizations support causal understanding of CNN activations?

Roland Simon Zimmermann*, Judy Borowski*, Robert Geirhos, Matthias Bethge‡, Thomas S.A. Wallis‡, Wieland Brendel‡. NeurIPS, 2021.

In this third publication, the explainability method feature visualization is again subject to investigation. Such tools are specifically designed to convey insights into DNNs in human-understandable terms. Evidently, they are a great means for the big goal of improving our understanding of machine visual perception.

Compared to the second paper, this one has many parallels with it: Here, we evaluate the same explainability technique and we do so with the same approach; by leveraging human responses from psychophysical experiments. What is different, though, is the kind of task we test and the kind of helpfulness we deduce from it: In the previous paper, humans performed a so-called feedforward prediction task. Thus, we inferred the *general informativeness* of the tested explainability method. In this paper, we aim to explicitly examine a *specific quality* of the tool. It is referred to as "causal understanding". To quantitatively evaluate this aspect, we design a new task in a counterfactual-inspired setup. Taken together, this third paper extends and complements the second one.

2.3.1 Motivation

The explainability method "feature visualization" is intended to grant insights into the features that a deep convolutional neural network learns. This is achieved via its synthetic images. They are the result of an optimization procedure whose objective is to create an input that elicits maximal activation for a certain DNN unit. In essence, these "favorite" inputs are how feature visualizations impart a unit's semantic meaning.

Many researchers think feature visualizations establish a "causal link" (Schubert et al., 2021; Olah et al., 2017). In fact, they consider this the method's core motivation (Olah, 2021a). Hereby, they mean that a synthetic image reveals what feature "causes" a unit to fire. This relation is inherent to the generation procedure: A pixel only changes because it will then elicit higher unit activation. Put differently, feature visualizations are believed to isolate and highlight exactly those features that "cause" a strong unit response. A popular example of this is the two commonly used images in Figure 8A. Here, one feature visualization displays a whole dog's face. And the other one displays just an eye. With such pure features, an observer is able to distinguish whether a unit responds to the whole object or just a part of it. To summarize, this reasoning is why feature visualizations are believed to support causal understanding.

The purported advantage of feature visualization becomes even clearer when contrasting its synthetic images to natural ones: The latter often contain more features than the one(s) in question (Olah et al., 2017). Sticking to the example above, finding a picture of only a dog's eye is rare. Instead, a dog's eye is almost always captured along with its whole head, if not (parts of the) whole body. Such correlations may

mislead an observer and convey an incorrect impression of a DNN's activations. Researchers like Olah et al. (2017) consider the detachedness of feature visualizations from the potentially confusing natural image manifold the method's advantage.

Despite the arguments for feature visualizations supporting causal understanding, a few other aspects cast a more critical light on this matter: For example, regularization mechanisms, intended to make the synthetic images look more human-interpretable, influence their faithfulness. As such, a pixel's value does not purely emerge anymore only because it "causes" high unit activation, but partly also because it then contributes to a more human-understandable appearance. Further, a complete understanding of a function is unlikely to be gained just based on the argument of said function's maxima. In other words, it is unclear how much about the entire role of a DNN unit strongly activating images can reveal. And finally, it is an open question whether *humans* actually gain causal understanding via feature visualizations.

Given these points, we want to test the widely spread intuition around feature visualization's causality in this study. In particular, our goal is to quantitatively measure how well feature visualizations support causal understanding of DNN activations.

2.3.2 Experiments and results

To test causal understanding of feature visualizations, we design a human psychophysical experiment. Our assumption is that if synthetic visualizations indeed grant causal insight, then they should allow humans to predict the effect of an intervention better. For our crowd-sourced experiment (see Figure 8B), we translate this as follows: To start, we provide reference images such as feature visualizations. Their role is to grant potential causal insight into what features elicit high unit activation. Next, and as the intervention whose effect our participants predict, we present image manipulations that change a unit's activation. Such an image manipulation is realized as a square occlusion superimposed on a query image. Specifically, we place it such that it either maximizes or minimizes the activation of the unperturbed image. Ultimately, our human participants are asked to choose which of two partly occluded images activates a given unit more strongly (2AFC paradigm (Fechner, 1860)). In essence, this requires two steps: At first and on the basis of the reference images, participants identify the important feature that elicit high activation. Then and with respect to the manipulated query images, they select the one where as much as possible of that important feature is visible. Overall, and just like in our previous publication, we compare different reference conditions in this experiment. For example, we test how well natural, the combination of synthetic and natural, or no reference images support causal understanding. This helps us set the performance of feature visualizations into context besides the chance level of 50 % from the 2AFC-task.

The main result from our experiment is that feature visualizations do provide humans with helpful information about the most important patch in an image — but not much more than other or no visualizations at all. As Figure 8C depicts, we specifically find that performance for synthetic images is at 67 ± 4 %. Even though this

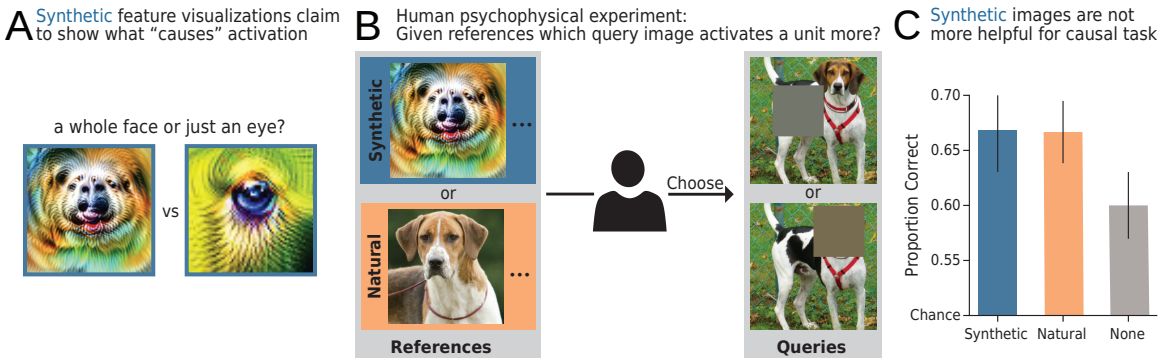


Figure 8: **Causal understanding of feature visualizations in counterfactual-inspired task**: Synthetic feature visualizations are believed to isolate and highlight exactly those features that "cause" a strong unit response (**A**). In our psychophysical experiments, humans are given reference images such as synthetic feature visualizations and decide which of two partly occluded images elicits stronger activation in the given network unit (**B**). Our main result is that while humans achieve more than chance performance with synthetic images, natural images are similarly helpful and the gap to no reference images ("None") is small (**C**).

result is clearly above chance, comparing it to other conditions makes it appear less powerful: In the none condition, where no reference images are given and participants make their choices purely based on query images, performance is already at $60 \pm 3\,\%$. This suggests that feature visualizations provide a small advantage only. The same observation holds true when looking at other conditions: Strongly activating natural data set samples, as an example, also reveal similar performance levels as feature visualizations.

Since performances between reference conditions are so similar, we thoroughly investigate a number of aspects to verify their plausibility. For example, we compare accuracies of the first two authors of this paper to the crowd-sourced data. As these measurements yield similar scores, we are confident to claim that our lay participants make their choices carefully. Further, simple baselines set our human data into context: For example, a straight-forward strategy is to always choose the query image with an unoccluded primary object. Results from such baselines reveal that they can reach above chance performance regimes between the no reference and visualizations conditions. Thus, our human data seems plausible. From a different perspective, though still in light of these simple decision-strategies, the advantage of feature visualizations is again not large. Finally, extensive analyses show that how easily the most important image patch is identifiable depends on various factors. For instance, the DNN unit, image choice, and activation difference between the manipulated query images, but not the reference condition show a dependence with performance.

2.3.3 Discussion

In this project, we put the intuition of feature visualizations supporting causal understanding to a quantitative test. As a matter of fact, this is the first time this widely

spread assumption is being evaluated. Specifically, we ask human participants in an online crowd-sourced experiment about the causal relation between manipulated images and a unit's activation. Our data show that feature visualizations do not support causal understanding particularly well: While they do provide useful information as indicated by above chance level accuracy, performance is only marginally higher than the one for a baseline where no reference images are provided and similar to the one where other visualizations such as natural data set samples are displayed. This suggests that the anticipated advantage of feature visualizations is not as large as expected. More precisely, the benefit of the method's detachedness from the natural image manifold and potentially misleading, spurious correlations is not clear. From a different perspective, another interpretation is that spurious correlations in natural images seem to not impair human performance as much as previously anticipated.

While our experiments reveal no particular advantage for feature visualizations, there is also no asset detectable for natural reference images or even the joint display of synthetic and natural images. The former aspect contrasts results from the previous publication: Here, natural data set samples repeatedly and almost consistently yield the highest task performance. In this second publication on feature visualizations, we do not find evidence in this direction on the performance level. However, a trial-by-trial analysis does reveal that humans are indeed most consistent with each other in the *natural* condition (see Figure 4 in publication) — and they happen to be least consistent in the synthetic condition. Regarding the mixed condition, the combination of synthetic and natural images is intended to leverage the best from two worlds: Pure features from synthetic images should give hints at what "causes" the unit to fire, and familiar contexts from natural images should help locate them. Since performance for this condition is similarly high as for other conditions, no benefit is detected. Taken together, this second experiment reveals no favorite visualization method as particularly helpful in causal understanding.

Future directions

Similar to the previous publication, there are many possible future directions.

To start off, and again similar to the other evaluation study of feature visualizations, the generality of our statement beyond our specific setup can be explored. As such, all the mentioned aspects from the previous publication can be examined. For example, more units of InceptionV1, combinations of units, the neuron objective, varying levels of information, as well as other networks and feature visualization methods can be tested. As of now, it is an open question whether, e.g. the presentation of feature visualizations for a single unit, a whole channel, or a combination of them would elicit different human performance levels. Similarly, another interesting direction is investigating what effect units that fire for multiple features have on how easily humans understand visualizations, if any.

Specific to the task presented in this second publication on feature visualization, variations of our intervention can be tested. So far, our square occlusions limit the kinds of features that can be sensibly covered. Chris Olah relates to exactly this shortcoming when claiming that our experiment would not test causality: "Both data set examples and feature visualizations give examples of *where* a feature fired. The

challenge is *what about that portion* caused it to fire, especially given that things like eyes and head are extremely correlated (in this case, part-whole relationship)?" (Olah, 2021b). To adequately address this, occlusions must become adjustable in size and shape. For example, a dog's eyes or its head can be covered by variably sized, round-ish patches. As another example, low-level line orientations can be occluded by multiple narrow, elongated areas. Even though realizing such adjustable occlusions at scale is not straight-forward with state-of-the-art algorithms, they would allow testing causal understanding more precisely.

Changing the psychophysical task even further to letting humans *themselves* describe variably sized regions could open up insights beyond causality. For instance, participants *themselves* can shade image region(s) that they believe to correspond to the feature(s) in question given references[8]. In order to evaluate human agreement and infer how humans interpret DNN features, these shaded regions then have to be compared to each other. To this end, measures like the intersection over union may be useful (see Bau et al. (2017)). An alternative approach can be transferring the task to a natural language setup: Based on references, participants describe *in words* what they interpret as the relevant feature(s)[9]. Again, and just like in the previous option, these natural language descriptions have to be analyzed for similarities and differences in a following step. Altogether, the suggested approaches permit identifying and illustrating a DNN's features more precisely. Notably, they do not test the quality of *causal understanding* but rather whether humans gain the same *feature understanding*. Going even further, experiments could investigate where humans' uncertainties lie and what kind of additional information they would need to improve their feature understanding.

Regarding the data from both our publications, the patterns of how our participants come to their decisions can be analyzed. Specifically, the predictive power of certain image aspects can be tested. For example, how much of the human decisions can be explained by a predictor that models, say color? Other interesting candidates would be for example orientation, object class, and spatial frequency. Going further, linear mixed models can reveal plausible combinations of predictors. Overall, such results could reveal what image aspects play an important role in how humans make their decisions. From a different perspective, this can grant insight into what the most helpful explanation would look like for humans.

Summary of feature visualization evaluations

Within the goal of deepening our understanding of visual perception in machines, the experiments of this study reveal that human participants can comprehend the causal relation between input images and DNN activations to a certain extent, namely such that they can moderately well predict the effect of an intervention. Similar to the

[8]Ultimately, the described approach of identifying certain image regions is similar to searching for the minimally necessary information in an image discussed by Biederman (1995); Tanaka (1993); Zhou et al. (2015) and Ullman et al. (2016). I presented all of the latter in Section 2.1.3.3. The main difference between the approaches is that the proposed method here aims at *specific features* and *intermediate network units* whereas the latter works aim at *general classification*.

[9]Caplette and Turk-Browne (2022) already did this in a recent study, though for the purpose of understanding mental representations and for one late layer only.

previous publication, they develop this understanding with extremely activating images. As such, we compare the popular feature visualization method to natural data set samples and other reference image variations. Our data illustrates that all visualization methods provide similarly helpful information. Therefore, the purported advantage of feature visualizations supporting causal understanding particularly well is not endorsed. Nonetheless, this study illustrates that humans can extract helpful information from different visualization variations and gain a limited causal understanding of information processing in DNNs.

To summarize the last two publications, we put the feature visualization method by Olah et al. (2017) to quantitative tests in order to evaluate their informativeness for humans. In fact, these are the *first* human evaluations of this explainability method. So far, only one other human study was done with feature visualizations (see Section 1.5.2). With our evaluation experiments, insights are gained regarding how well we understand the internal information processing of machine algorithms for object classification. Specifically, we test variants of two of the tasks that Doshi-Velez and Kim (2017) suggested, namely a "counterfactual"-inspired and a feed-"forward prediction" paradigm. These different evaluations allow us to draw conclusions on the general informativeness as well as a specific quality of it: causal understanding. We are convinced that such psychophysical tests are a great way to transfer intuitions via falsifiable hypotheses (Leavitt and Morcos, 2020) into quantitative results. As such, they reveal realistic estimates regarding what information explainability methods can and cannot convey. Finally, we hope that our objective psychophysical tasks serve and inspire further development of challenging evaluations and that they help steer future advancement of feature visualizations.

3 Discussion

The papers summarized in this thesis aim to deepen our understanding of visual perception in machine algorithms. In all of them, we utilize human psychophysics — however, in two different ways: In the first publication (see Section 2.1 and Funke et al. (2021)), we conduct comparison studies between DNNs and humans. By *directly* assessing similarities and difference in behavioral performances, we draw conclusions on the investigated DNNs' abilities, limits, and mechanisms. In the latter two publications (see Section 2.2 and Borowski et al. (2021) as well as Section 2.3 and Zimmermann et al. (2021)), we perform human evaluations of an explainability method. This means we infer how well humans understand the inner workings of a DNN based on feature visualizations by testing how well they can simulate the machine algorithm.

Here, I discuss the results from our publications in a broader perspective. Following the story line of understanding from a coarse to a fine-grained level, I first discuss the comparison and then the explanation works. Specifically, I illustrate the relevance of our checklist by depicting its applicability not only in other comparison studies of the literature but also beyond comparison studies, namely in our evaluation experiments. Then, I outline other good practices for comparison studies that are suggested in concurrent work. Shifting toward feature visualizations, I first delineate potential reasons for why feature visualizations are only moderately helpful, and why natural data set samples are surprisingly informative in our tasks. Next, I emphasize that we only test one feature visualization technique in our publications and present an additional experiment with another technique. As many of our findings point toward the helpfulness of naturalness, I discuss how important this kind of appearance may be. And finally, I relate our findings to other explainability methods' evaluations. In a last and third section of the Discussion, I zoom out from the presented approaches toward understanding visual perception in machines with human psychophysics and briefly outline alternatives as well as what important insights they grant.

3.1 Comparisons between DNN and human behavior

Comparison studies are notoriously difficult. Designing, conducting and interpreting experiments for two different systems can disguise pitfalls. Nonetheless, such studies between DNNs and humans have revealed great insights. In our publication (see Section 2.1 and Funke et al. (2021)), we present a checklist to support adequate comparisons. Besides suggesting general good practices regarding designing, conducting and interpreting comparison studies, we illustrate the application of these points in case studies.

3.1.1 Where else is our checklist reflected?

The relevance of our checklist extends beyond the three presented case studies. In fact, several studies from the literature reflect our suggested points. What is more, our checklist can prove useful beyond comparison studies. For example, many of the

suggested checklist points can be identified in our feature visualization evaluation studies.

Reflections of our checklist in other comparison studies of the literature

Zooming out from our own case studies, I here highlight three sets of publications that reflect our checklist. Specifically, they demonstrate how (aligning) experimental design choices can influence results (point ii) and that resisting human bias is important (point v).

The publications of Zhou and Firestone (2019) and Dujmović et al. (2020) realize certain experimental factors differently and therefore unveil diverging evidence of how humans perceive adversarial examples. Specifically, Zhou and Firestone (2019) suggest that human perception of adversarial examples would be more similar to machine perception than previously assumed. In contrast, Dujmović et al. (2020) indicate the opposite, namely "much weaker and more variable" agreement than reported by Zhou and Firestone (2019). Key differences lie in e.g. how the stimuli are generated and selected or which of the 1000 ImageNet label options are provided to human participants. As an example and with respect to this latter aspect, Zhou and Firestone (2019) demonstrate that when label options are chosen randomly, agreement between human and machine perception of adversarial images is high. In contrast, Dujmović et al. (2020) report that when label options are adopted to represent realistic, "competitive response alternatives," agreement is close to chance. This example illustrates that the design choice of e.g. label options influences the resulting human susceptibility to DNN adversarials. In connection to our checklist, these varying findings can be directly linked to both point ii and v: Adequately aligning experimental conditions as well as resisting human bias are crucial to be able to draw robust conclusions.

Similar to the previous pair of publications, the experiments of Geirhos et al. (2018a) and Tartaglini et al. (2022) differ and consequently reveal opposing findings regarding the texture vs. shape bias in DNNs: Geirhos et al. (2018a) reveal that DNNs are biased toward texture. In their experiments, the researchers introduce cue-conflict images, i.e. images where the shape and texture information point toward different classes. They discover that a DNN classifies an image with the shape of e.g. a cat but the texture of an elephant as an elephant. This is in contrast to grown-up humans who are known to exhibit a shape bias (Smith et al., 2002; Diesendruck and Bloom, 2003; Gershkoff-Stowe and Smith, 2004; Biederman, 1995; Colunga and Smith, 2005). In fact, Geirhos et al. (2018a) observe that human participants classify the elephanty cat as a cat. Contrary to the findings of Geirhos et al. (2018a), Tartaglini et al. (2022) unveil evidence that DNNs — similar to humans — are biased toward shape. The different approach of their experiments is rooted in practices of developmental psychology. For example, Tartaglini et al. (2022) adapt the cue-conflict stimuli from Geirhos et al. (2018a) such that the texture only covers the inside of a shape, but not the background. Also, they evaluate *relative* similarities between texture- and shape-consistent stimuli, not absolute, single output decisions like Geirhos et al. (2018a). Overall, the procedures of both studies have their justifications. The diverging evi-

dence, though, illustrates how differently aligning experimental conditions (point ii) can have a large effect on the results.

In a third set of publications, distinct choices in a psychophysical task paradigm lead to different labels for the widely-used data set ImageNet. More specifically, Tsipras et al. (2020) and Beyer et al. (2020) re-evaluate the original labels generated by Deng et al. (2009); Russakovsky et al. (2015), and analyze whether training models on ImageNet still corresponds to progress on the real-world task of object recognition. In the original data collection pipeline, human annotators only *confirm* that an image contains one or more instances of a presented object Deng et al. (2009); Russakovsky et al. (2015). The logic is simple: If this is indeed the case, the image receives that object's class label. However, shortcomings of this paradigm with its leading question are that annotators do not know about alternative class candidates or the granularity of classes. Also, multiple objects in one image create ambiguity: For example, how should a participant know whether "paddle" or "canoe" is more appropriate, if an image contains both? Finally, many classes such as the more than 120 dog breeds can be challenging for laymen. Altogether, this task paradigm can be seen as expressing human bias (point v) and not providing humans suitable label options (point ii). To counteract these aspects, Tsipras et al. (2020) and Beyer et al. (2020) concurrently design a new task: Here, participants select *all* present objects in an image. Potential candidates are of course not all 1000 ImageNet classes, but a subset generated by top predictions of DNNs. With these new labels, the two groups reveal that DNNs are still making progress on object recognition, but also partly overfit to ImageNet. Overall, this set of publication illustrates that *repeated* direct comparisons of human and machine data can deepen our understanding of visual perception in machines.

Taken together, the varying results of our three case studies as well as of these three sets of publications demonstrate how challenging comparison studies are. We hope that applying our checklist helps designing, conducting and interpreting experiments in a sound, robust and reliable manner.

Reflections of our checklist beyond comparison studies: in our feature visualization studies

Even though we presented our checklist for *comparison* studies, it can also be useful for other types of studies. To exemplify this, I here outline how it relates to our human evaluations of feature visualizations. As already alluded to in Section 2.2, all points but point ii are reflected in our evaluation studies — sometimes even in multiple ways.

To start, point i suggests to isolate functional or implementational properties. In our evaluation studies, we aim to clearly bring out the different qualities of informativeness by designing two different tasks: The feedforward prediction task assesses how helpful this explainability method is to anticipate an input's activation, whereas the counterfactual-inspired task estimates causal understanding. Ultimately, this is closely related to the already quoted statement of Doshi-Velez and Kim (2017): "The claim of the research should match the type of the evaluation."

Point ii is about aligning experimental conditions between the two systems subject to comparison. In a human evaluation study of an explainability method, this point does not apply: By definition, *humans* are tested on *machine*-selected stimuli and *machine*-based explanations. Judging this crisscross of humans and machines with the mindset of a comparison study, experimental conditions would clearly count as misaligned. However, in this scenario, the very assessment of human responses on the basis of machine ground-truth allows us to infer how well humans can simulate machine behavior. It is exactly what provides the estimate of the explainability method's usefulness. This means that the setup of testing one system on the other is absolutely intended.

In point iii, the proposal is to differentiate between necessary and sufficient mechanisms. If we consider feature visualizations a "mechanism" that conveys an understanding of DNNs, our evaluation studies implicitly unveil that this method is only sufficient: By testing various reference conditions such as natural data set samples, we demonstrate that humans can extract useful information from such other "mechanisms" as well. In fact, even the purely machine-based baselines in our second publication illustrate that straight-forward decision-making strategies suffice for above chance performance in our task. In the bigger picture, most evaluation studies of explainability methods do not assess them as candidates for necessary explanations but just sufficient ones.

Next, point iv concerns testing generalization scenarios of the investigated mechanisms. In our evaluation studies, we make an effort regarding this by measuring many different units across the investigated network. For example, in our first publication, we evaluate all layers (except for the first and last one) of InceptionV1 and all branches in the Inception Modules. As discussed in the "Future directions" sections, there is still room to explore other generalization scenarios in order to make more general statements about feature visualizations.

Finally, point v centers around resisting human bias. In our evaluation studies, reflections of this can be seen in several aspects. Here, I describe two: To begin, the whole endeavor of *evaluating* feature visualizations is an action of verifying human intuition. As our results show, feature visualizations turn out less helpful in our psychophysical tasks than previously anticipated. Therefore, our studies proved useful to objectify human impressions. On another level, providing more context of performance levels with other baselines can also be seen as a way to resist overhasty conclusions. At first sight, the above chance performance levels of feature visualizations can be interpreted as a big success. However, additional data points from alternative references allow us to more objectively draw conclusions. Altogether, our human bias cannot be removed. As exemplified, though, various efforts can help counteract it.

In summary, our checklist for comparing human and machine visual perception can be applied to not only comparison studies but also human evaluations of an explainability method. This means it is useful for designing, conducting and interpreting various studies. Further, the reflections illustrate that even though comparison studies and explainability methods are different in nature, parallels do exist. Above all

and without any doubt, they are both great approaches to further strengthen our understanding of machine visual perception.

3.1.2 What are other good practices in comparison studies?

Besides our publication, concurrent work develops myriad suggestions regarding how to best perform comparison studies.

As already presented in the Introduction (see Section 1.5.1), good practices regarding how to compare human and machine visual perception have been developing for quite some time. For feedforward DNNs, choosing a short presentation time of stimuli in human psychophysics has been established to adequately mirror the machine processing (Tang et al., 2018; Thorpe et al., 1996; Serre et al., 2007b; DiCarlo et al., 2012). Further, designing challenging experiments (Wichmann et al., 2017) has become common in order to discover where the ever improving DNNs fall short. And finally, awareness around limiting our human bias in comparison studies has increased (Buckner, 2019). While our case studies reflected the five points proposed in our checklist, they also put the three, just mentioned suggestions into practice.

Going further, the perspective article by Firestone (2020) enriches the endeavor of comparison studies between humans and DNNs with ideas from other comparative research fields. Specifically, he discusses that it can make sense to (1) "limit machines like humans", to (2) "limit humans like machines" and to (3) perform "species-specific task alignment". The underlying motivation comes from cognitive science: Here, it is common insight that what a system *knows* and what a system *does* may not always correspond. In other words, *competence* and *performance* may differ.

A straight-forward example for the discrepancy between competence and performance and how to work around it via species-specific task alignment or limiting humans like machines can be found in human-machine comparisons of object recognition: For machine algorithms, this problem is commonly translated into a many-class classification task. As such, DNNs can trivially select a label from e.g. 1000 options. In contrast, such a plethora of options would be inappropriate for humans. Our working memory is not able to cope with so many items. In other words, we are limited in our *performance*. Nonetheless, it is out of doubt that we are capable of comprehending and correctly attributing all those classes. This means we do have the *competence*. To adapt the machine task setting for humans, the workaround is selecting a reasonable subset of labels. As illustrated in two of the publication sets presented in Section 3.1.1, these choices are not trivial in practice, and they can influence a study's results. Nonetheless, the idea of species-specific task alignment does theoretically permit a fair comparison between human and DNNs.

Comparing the perspective article by Firestone (2020) and our work, the two publications have different strengths and weaknesses. For example, Firestone (2020) provides plenty of background and examples from cognitive science and even developmental psychology — whereas we only briefly touch on the broad history of comparison studies. Further, Firestone (2020) explicitly formulates *three* fairly specific recommendations — out of which the third advice on species-specific task alignment can be seen as summarizing the former two, namely limiting machines (humans) like

humans (machines). In contrast, our checklist addresses a rather general level. The most similar point compared to the ones by Firestone (2020) is point ii of aligning experimental conditions. Nonetheless, certain design choices in our case studies do address Firestone's specific recommendations. For instance, we limit humans like machines by presenting the closed contour stimuli for 100 msec only. This presentation time is believed to mirror the pure feedforward pass in a DNN (Tang et al., 2018; Thorpe et al., 1996; Serre et al., 2007b; DiCarlo et al., 2012). Finally, a difference is that Firestone (2020) mentions many different studies, whereas we thoroughly investigate three case studies and contribute new experiments. Altogether, the perspective article by Firestone (2020) and our work complement each other well.

Yet another suggestion for comparing DNNs and biological vision is motivated by comparative biology and implies to "focus on differences, not similarities" (Lonnqvist et al., 2021). Specifically, the authors advocate considering DNNs and human vision as distinct "species" (Lonnqvist et al., 2021). Then, investigating differences between their information processing can reveal what is really needed for a task. Connecting this to our own checklist, point iii regarding differentiating between necessary and sufficient mechanisms is most similar. Here, we also allude to the fact that there is often more than one way to approach a task, and that its pure existence does not yet translate to its certain deployment. In the word of Lonnqvist et al. (2020), such a mechanism would not be "crucial". As an example, the authors point out that machine large-scale object recognition does not require "attention, segmentation or recurrence", even though these functions play an important role in the human visual system. All in all, focusing on differences instead of similarities can be a fruitful way forward, and particularly to *generally* deepen our understanding of vision, not only vision in machines (Lonnqvist et al., 2020).

A third suggestion for a good practice regarding comparing DNNs and biological vision is to use language carefully, and more generally to avoid anthropomorphization (Mitchell, 2021b; Shevlin and Halina, 2019). Even though using human-associated terms like "understand" or "winning" may be easier to convey high-level ideas around algorithms, this can not only mislead the general public in science communication but also unconsciously bias experts (Mitchell, 2021b). Specifically, Shevlin and Halina (2019) points out that "rich psychological concepts" such as "awareness, perception, agency and theory of mind" "require greater caution when employed to describe the capabilities of machine intelligence". As illustrated in the first and third publication sets of Section 3.1.1, language as well as more generally our human reference point can influence experimental design choices. With respect to our own checklist, point v of resisting human bias, which was also already mentioned by Buckner (2019), is most closely connected. More generally, the phenomenon of anthropomorphization has been well known in the literature from comparative psychology (e.g. Romanes, 1883; Haun et al., 2011; Koehler, 1943; Köhler, 1925; Boesch, 2007; Tomasello and Call, 2008). In all, thoughtful and accurate language choice not only support realistic expectation management but also adequate design, and interpretation of experiments.

Besides the developments with respect to rather high-level recommendations from e.g. us, Firestone (2020), Lonnqvist et al. (2021) or Mitchell (2021b) and Shevlin and Halina (2019), comparison *metrics* as well as analysis techniques and experimental

setups are being further developed in the literature. The necessity for these advancements is growing for two reasons: One, the machine visual systems are becoming more and more similar to the human one. And two, high performance regimes make it difficult to identify differences. As often mentioned in this thesis, these two aspects are exactly the case for DNNs and humans on specific tasks like object recognition (He et al., 2015; Kheradpisheh et al., 2016). In such a situation, the similarities and differences between two systems can be obfuscated when only evaluating aggregate measures like class-average performance. To address this situation, more fine-grained measures than accuracy (Ma and Peters, 2020; Geirhos et al., 2020b; Chollet, 2019; Lonnqvist et al., 2021) as well as more advanced analysis techniques and experimental designs are needed.

As a matter of fact, more sophisticated ideas are explicitly described in publications and are starting to be deployed. For example, Ma and Peters (2020) suggest analyzing *error* patterns using a confusion matrix. With this approach, Wichmann et al. (2017) indeed identify DNNs' biases toward a few classes when stimulus manipulation is high. Ma and Peters (2020) further recommend investigating different aggregation statistics of a confusion matrix. A popular choice of this is the trial-by-trial metric called Cohen's kappa (Cohen, 1960). It takes into account both the observed error overlap as well as the error overlap expected by chance. As an example, Geirhos et al. (2020b) use it to compare humans and DNNs on object recognition, finding that they are not much more similar than expected by chance, while *different* DNNs, on the other hand, are "remarkably consistent". Tuli et al. (2021) further extend this error analyses by varying the levels of aggregation granularity: Applying the Jensen-Shannon distance in different ways to the confusion matrix, they evaluate *which* classes are misclassified and *what* classes are misclassified as *what*.[10] Moreover, Ma and Peters (2020) suggest enriching the human-machine comparison by not only measuring the distance between single accuracy-values but between the *distributions* reflecting by-trial variability. For humans, this variability is natural, and machine answers can be generated by using Bayesian neural networks. Other porpositions by Ma and Peters (2020) include measuring the receiver operating characteristic, reaction times, or the learning trajectory (Ratcliff, 1990). This latter aspect of measuring "skill-acquisition efficiency" is what Chollet (2019) advocates for as well — in fact in the form of a general Artificial Intelligence benchmark. Overall, there are lots of metrics to analyze similarities and differences on a deeper level.

Going beyond measures, Ma and Peters (2020) propose advanced techniques to compare humans and machines. For example, fitting cognitive process models to both the DNN and human data could be one approach (Wang et al., 2016). In a second step, these models' parameter estimates would then be compared against each other (Wang et al., 2016). As another idea, Ma and Peters (2020) describe the Turing test: Here, humans guess the generation origin of stimuli, which are either produced by humans or generative algorithms. Similarly, many other experimental setups can

[10] In our third publication, we also use the measure of Cohen's kappa to compare decision patterns between humans in different reference conditions as well as between humans and baseline decision strategies.

be designed and there are (almost) no limitations. To summarize, the portfolio of analysis metrics and experimental setups is large.

While the big advantage of the mentioned methods is that they can reveal deeper insights, they also come with disadvantage. For example, they are both more complex to understand and explain as well as more expensive to compute. Whether they are worth the effort depends on many factors. In principle, though, certain experimental designs might still be sufficiently analyzed by simple metrics. As an example, class-average accuracy was an appropriate measure in case study II (see Section 2.1.3) to show that our pure feedforward DNN can perform well on a challenging abstract visual reasoning task. One the whole, each method has its own justification and it is a case-by-base decision which one is most suitable.

Altogether, a multitude of not only good high-level practices but also specific approaches and metrics exist for comparison studies. As illustrated, it can be helpful to take inspiration from adjacent fields and apply the lessons learned to benefit from this head start. The checklist presented in our first publication and the suggestions listed here complement each other well. What recommendations to follow and what metrics to use is ultimately always a case-by-case decision. Altogether, the various mentioned aspects can help with comparing fundamentally different systems to gain more nuanced and robust insights.

This concludes the broader discussion of the first publication, where I illustrated applications of our checklist and outlined other good practices for comparison studies. In the bigger picture of this thesis, our behavioral experiments with humans and DNNs added to a better understanding of machine vision on a rather coarse level. This means that our case studies revealed insights on a functional and algorithmic level. Next, the focus remains on the latter level and publications two and three are put into a broader perspective.

3.2 Explainability method: feature visualization

With the goal of understanding machine visual processing at a more fine-grained level, we investigated the explainability method feature visualization. Specifically, our human evaluations reveal that feature visualizations are helpful in our two psychophysical tasks. However, participants do not achieve perfect performance and hence there is still room for improvement. Moreover, other visualization options such as natural data set images can also convey helpful information — depending on the task paradigm, similarly much as feature visualizations or even more than them.

3.2.1 Why are feature visualizations by Olah et al. (2017) only moderately informative?

In the bigger picture, a first interesting question is why the explainability method of feature visualizations by Olah et al. (2017) is only moderately helpful. As already discussed, various factors from the tool — such as the challenge of choosing the appropriate unit — or the human evaluation setup — such as difficulty of providing

the appropriate amount of information — may contribute to suboptimal results (see Sections 2.2.1, 2.2.3 and 2.3.3). In the following, I discuss other potential, and more high-level reasons.

First off, we humans are not familiar with feature visualizations. This means we may have more difficulties recognizing understandable structure and extracting useful information from these synthetic images. Even though we steered against this by letting participants perform practice trials and granting them unlimited or plenty of time in the experiments in the two publications respectively, the artificial appearance may have impaired performance.

Another reason for imperfect performance in our task is that some DNN features do not represent human-understandable concepts. In other words, we humans are simply unlikely to understand them. Given the different numbers of units in a DNN that represent the input throughout its processing stages (order of millions in InceptionV1) and the number of dimensions that humans mentally represent natural objects with (49 according to Hebart et al. (2020)), it is not surprising that not all machine representations are intuitive to us. In fact, these non-recognizable features represent a challenge irrespective of the reference condition. As a consequence, answering trials for such units is reduced to a guessing game.

Yet another aspect is the poor representative power and diverse visualizations of feature visualizations for certain low-level features. As such, the latter seem surprisingly poorly identifiable in the synthetic reference condition. Figure 11 of the Supplementary Material in Borowski et al. (2021) exemplifies that diverse synthetic reference visualizations for low-level units do not allow to infer a common feature: Not only do they differ in color, but also in global appearance as well as local texture. In addition, and even more confusingly, *weakly* and *strongly* activating inputs sometimes look similar. This anecdotal evidence suggests that low-level units can be difficult to understand with synthetic images. Further research is of course needed to further test this hypothesis's representativeness.

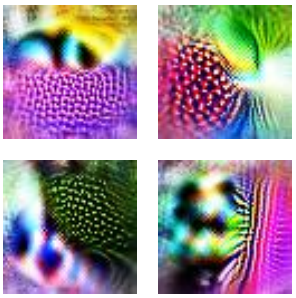


Figure 9: **High-low frequency features**: Long hours of investigations reveal that they detect object boundaries (Schubert et al., 2021). Image credit: OpenAI Microscope (OpenAI, 2020).

Finally, some feature visualizations require enormous efforts to understand their meaning. Evidently, such synthetic images are unrealistic aids in our experiments. One example of them is the family of high-low frequency detectors (Olah et al., 2020b; Schubert et al., 2021) (see Figure 9). At first glance, they do not seem intuitive. They probe for high frequency in one part of the receptive field and low frequency in the other one. What can this kind of feature correspond to? After investing a lot of time — the group's overall efforts are in the "thousands of hours" (Olah et al., 2020b) —, Schubert et al. (2021) suggest that they are boundary detectors of objects. This means that they fire strongly for "a highly-textured, in-focus foreground object against a blurry background" (Schubert et al., 2021). In a natural image, such a feature might detect, for example, a "microphone's lattice-work" against a blurry face (Schubert et al., 2021). This insight is a big achievement

for better understanding DNNs and — from an optimistic perspective — exemplifies that feature visualizations can even "teach" humans new DNN features. However, this boundary detector represents just one example out of an unknown number of unclear DNN features. As such, it is unknown how many more such theoretically identifiable units exist. Taken together, ad-hoc non-intuitive features represent challenges and do not reveal straight-forward insights into DNNs to humans.

Overall, the unfamiliarity, (putative) detachedness from human concepts as well as the limited representative power explain why feature visualizations — as well as partly other reference visualizations — turn out only moderately helpful in our experiments.

3.2.2 Why are natural data set samples surprisingly helpful?

A related question to why the tested feature visualizations provide only moderately helpful information in our tasks is why natural data set samples are so useful. Originally, we had intended them as a baseline. However, our data show that they can be even more helpful than (see Section 4.3) or similarly helpful as the explainability method' explanations (see Section 2.3).

To start, humans are simply more familiar with natural images, and hence they can extract helpful information from this source. Even though co-occurrences of (parts of) objects can potentially mislead an observer, they are *natural*. Therefore, the total composition of objects may actually help our participants in better orientation. In fact, they seem to interpret natural reference images most similarly across all conditions, as indicated by their highest by-trial decision consistency (see Figure 4 in Zimmermann et al. (2021)). Furthermore, an unusual parallel regarding merely correlated image features is the decision-making of DNNs: The latter are known to take advantage of such features, even though they are not necessarily causally related with the target class (e.g. feature "fingers" for class "band aid") (Singla and Feizi, 2021; Ilyas et al., 2019). Albeit this behavior is unwanted and considered as learning unintended shortcuts (Geirhos et al., 2020a), artificial models do make use of them. Altogether, expressing an explanation in a familiar way — such as with natural images — may represent an advantage.

Another aspect in favor of natural images — and as mentioned earlier, not for synthetic ones — is their representative power of certain low-level features. As such, the latter seem surprisingly easily identifiable in the natural reference condition. Figure 11 of the Supplementary Material in Borowski et al. (2021) exemplifies that the unrelated objects in natural reference images presumably encourage observers to shift attention away from the object level. Instead, the common low-level information stands out. As already mentioned, this anecdotal evidence suggests that low-level units may be more easily understood via natural than synthetic images. However, further research is needed to test the generality of this hypothesis.

In summary, the higher familiarity and the surprisingly powerful representation of low-level features represent potential clarifications why natural data set samples are so helpful for our participants.

3.2.3 How useful are other feature visualization methods?

Moving away from why different reference images grant different degrees of information, another open question is how helpful *other* tools of the same explainability family are. More specifically, in our two publications, we test exactly one feature visualization method, namely the popular tool from Olah et al. (2017). Of course, one implementation is not representative of the whole family. Therefore, an open question is whether humans can extract useful information from other feature visualizations and how they compare to the ones from Olah et al. (2017).

Additional study with feature visualizations from Nguyen et al. (2017) in feedforward prediction task

To extend our main studies beyond testing the feature visualization method by Olah et al. (2017), we conduct a small experiment with another technique[11]: Feature visualizations by Nguyen et al. (2017) are created with a generative adversarial network (GAN) and they look remarkably "photo-realistic" (see Figure 10A for example images, as well as Appx. Sec. 4.3). Here, we evaluate them with five expert participants. Specifically, the task is the same as in our first feature visualization evaluation publication Borowski et al. (2021): In a feedforward prediction task, the question is which of two query images elicits higher activation given extremely activation references. As before, we compare synthetic images to natural data set samples (see Appx. Sec. 4.3 for screenshots of the tasks).

The data from this new experiment suggests that feature visualizations are more helpful than natural images: $87 \pm 3\,\%$ vs. $79 \pm 2\,\%$ (see Figure 10). However, a Wilcoxon signed-rank test indicates that the difference in reference conditions is just not statistically significant ($p = 0.058$, JASP (JASP Team, 2021, version 0.16)).

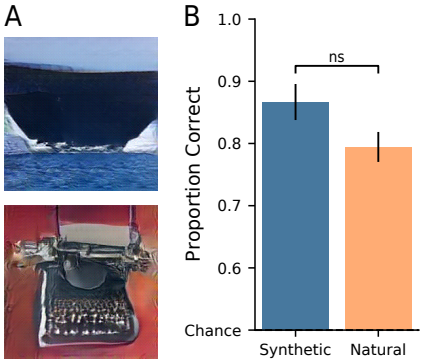


Figure 10: **Informativeness of feature visualizations by Nguyen et al. (2017)**: In a small study, we discover that synthetic images (example images displayed in **A**) are more helpful than natural ones (**B**).

This new result partly contrasts the findings from our earlier experiments. The aspect that remains consistent is the above chance performance for both conditions. It reinforces that, in general, extremely activating images are a helpful source for our feedforward prediction task. The difference, though, is in the ordering of helpful visualization options: In our first feature visualization evaluation publication Borowski et al. (2021), we repeatedly and almost consistently discovered that *natural* images provide more information than synthetic images. On the contrary, the data in this additional experiment suggests that *synthetic* images are more helpful than natural images.

[11] This small study was again joint work with the same authors and similar contributions as in the publication (Borowski et al., 2021). Until the publication of this thesis, it remained unpublished.

The partly diverging evidence regarding the most helpful visualization option may be due to several differences between the experiments. For example, the selection of natural reference and query images is made by different networks. The reason is that the two feature visualization methods build on different networks in their original implementations. As we want to guarantee comparability *within* the experiments presented in the publication and the additional one here, we adjust the network choice for each of them: In the publication's experiments with the method of Olah et al. (2017), we use the classification network InceptionV1 (Szegedy et al., 2015). In contrast, in this small additional study with the method of Nguyen et al. (2017), we use the classification network AlexNet (Krizhevsky et al., 2012). Notably, the top-5 accuracies of these two DNNs differ on ImageNet: They are $\approx 93\,\%$ and $\approx 80\,\%$[12], respectively. As a consequence, it is possible that the natural images in this additional study simply *cannot* have been as informative due to this method choice in feature visualization. Further, there is no layer- or unit-correspondence between the networks. This makes the comparison between the networks and methods difficult. Moreover, we test varying numbers of units in the experiments: 45 and 80 in the first publication, and 72 in this additional experiment. Yet another difference is that we use the channel-objective with the method of Olah et al. (2017), and the neuron-objective with the method of Nguyen et al. (2015). Finally, the sheer numbers as well as the expertise level of participants varies: In the first two experiments, 10 expert and 23 expert/lay participants take part, whereas in this third experiment, only 5 expert participants take part.

In summary, the additional study with synthetic images from Nguyen et al. (2017) adds to the picture of how helpful the explainability method feature visualization is for humans. As such, this new experiment supports the stance that feature visualizations indeed provide useful information. Further it diversifies the results of what visualization option is best.

From a larger perspective, the partly diverging findings in our various experiments suggest that each experiment only has limited representativeness for the whole family of feature visualization methods. In other words, generalization beyond the specific tested tool and task is not possible. The latter is corroborated by yet another ordering of the visualization options in the counterfactual-inspired task compared to the ones discussed above: Here, all tested visualizations are equally helpful. In conclusion, as generalizing is not possible, each specific scenario has to be evaluated individually.

3.2.4 How important is "naturalness" in explanations?

While the helpfulness of feature visualizations seems to depend on the specific method, another interpretation of our partly diverging findings in the publications and the additional study is possible: Maybe the best visualization options suggest that "naturalness" of an image plays an important role? Below, I discuss arguments around this aspect.

[12] in the Caffe-implementation

To start, a few elements suggest that "naturalness" of an image matters. For one, *natural* images provide more helpful information in the experiments of our first evaluation publication. This result is largely consistent across various conditions. Second, the more useful reference type in our small study is the *photo-realistic* feature visualization. The outstanding natural appearance of these synthetic images can be interpreted as evidence for our hypothesis. Finally, the common practice of augmenting the vanilla generation procedure of feature visualizations with regularization mechanisms points toward the importance of references images' naturalness. As such, these instruments are sometimes even called "natural image prior" (Mahendran and Vedaldi, 2015). Taken together, these three elements can be interpreted as that the naturalness of extremely activating reference images plays a crucial role for humans to predict a DNN's feedforward behavior.

On the other hand, an argument against the proposed advantage of natural-looking images may be raised on the ground of our experimental setup. As such, the query images for which the activation has to be predicted are also natural — and not synthetic. Therefore, it can be argued that high performance for natural(-looking) reference images may be expected. Although this is indeed plausible, we nevertheless consider our choice of natural query images reasonable: Ultimately, explainability methods are designed to explain behavior on real-world data. As that corresponds to natural images, natural images are exactly what we should test.

On a speculative note, it can be conjectured that a photo-realistic feature visualization method like the one from Nguyen et al. (2017) may combine the best aspects of two worlds. On the one hand, people are familiar with natural-looking images (see Section 3.2.2). As a consequence, they are good at extracting useful information from them. On the other hand, feature visualizations isolate the pure feature(s) that cause(s) strong network activation (see Section 2.3). This means that merely correlating features are not displayed and therefore cannot confuse an observer. Altogether, the purity of feature visualizations and the rendering in a photo-realistic way may constitute a powerful explanation and therefore explain why the method from Nguyen et al. (2017) was so informative for humans.

Taken together, our data provides evidence that naturalness of explanation images is important. Whether purely natural images or feature visualizations presented in a natural look grant humans better insights into DNNs under various experimental conditions remains to be investigated in future research.

3.2.5 How do our findings relate to other explainability methods' evaluations?

While we find feature visualizations to be helpful, albeit in a limited manner, an interesting question is how these findings relate to other explainability method's evaluations. Unfortunately, several aspects complicate this comparison: For example, evaluation results depend on various factors such as participant pool, data type, and experimental setup (including e.g. physical environment, psychophysical task, and instructions). What is more, the lack of common quantitative measures makes

contrasting studies challenging. Below, I consequently concentrate on qualitative comparisons.

In terms of the general, moderate informativeness of feature visualizations, the literature contains findings in all directions: Some evaluations of explainability methods also reveal a positive effect of an explanation (e.g. Kumarakulasinghe et al., 2020), others report inconclusive results (e.g. Alufaisan et al., 2021; Chu et al., 2020), and yet others even find negative effects on human performance (e.g. Shen and Huang, 2020). In a new meta-analysis of evaluation papers, Schemmer et al. (2022) observe a general "statistically positive impact of XAI on user's performance". Even though the latter reflects our finding, future work will have to show whether this remains the overall trend.

Regarding our strong results for the natural reference images, other publications also describe similar findings. In fact, several researchers also observe that data set samples provide (more) helpful information to humans than the tested explainability tool (e.g. Nguyen et al., 2021; Jeyakumar et al., 2020; Hase and Bansal, 2020). For example, Nguyen et al. (2021) measure attribution (or saliency) maps, i.e. visualizations of the contributions of each pixel toward the DNN's classification. Regarding a feedforward prediction task based on ImageNet, they discover that these explanations are not more helpful than nearest training examples. Further, regarding a more difficult and more fine-grained dog classification task, they reveal that attribution maps even *deter* human performance. Combining this finding with our own results, there is quite some evidence that data set samples can already grant good insights into visual processing of DNNs.

In summary, not only our main finding of moderate usefulness but also the one about natural data set samples being helpful is indeed reflected in the literature.

This concludes the broader discussion of the second and third publication, where I debated potential reasons for our findings and set them in a wider context. In the bigger picture of this thesis, these works deepened our understanding of machine vision on a fine-grained, or algorithmic level. Altogether, the presented three publications cover a range of granularity levels from coarse to fine-grained understanding (see Figure 3). Zooming out from behavioral human-DNN comparisons and explainability methods, other approaches exist and they can further extend our knowledge around visual perception in machines.

3.3 Other approaches toward understanding machine visual perception

In general, there are myriad approaches to understand vision in machines. This section complements the two presented approaches of behavioral comparison studies and explainability methods with *comparisons between DNNs and neural data* and *investigations of DNN performance in isolation*.

Figure 11 shows an overview of the four mentioned approaches and how they relate to each other. Specifically, I span the space of approaches by two orthogonal

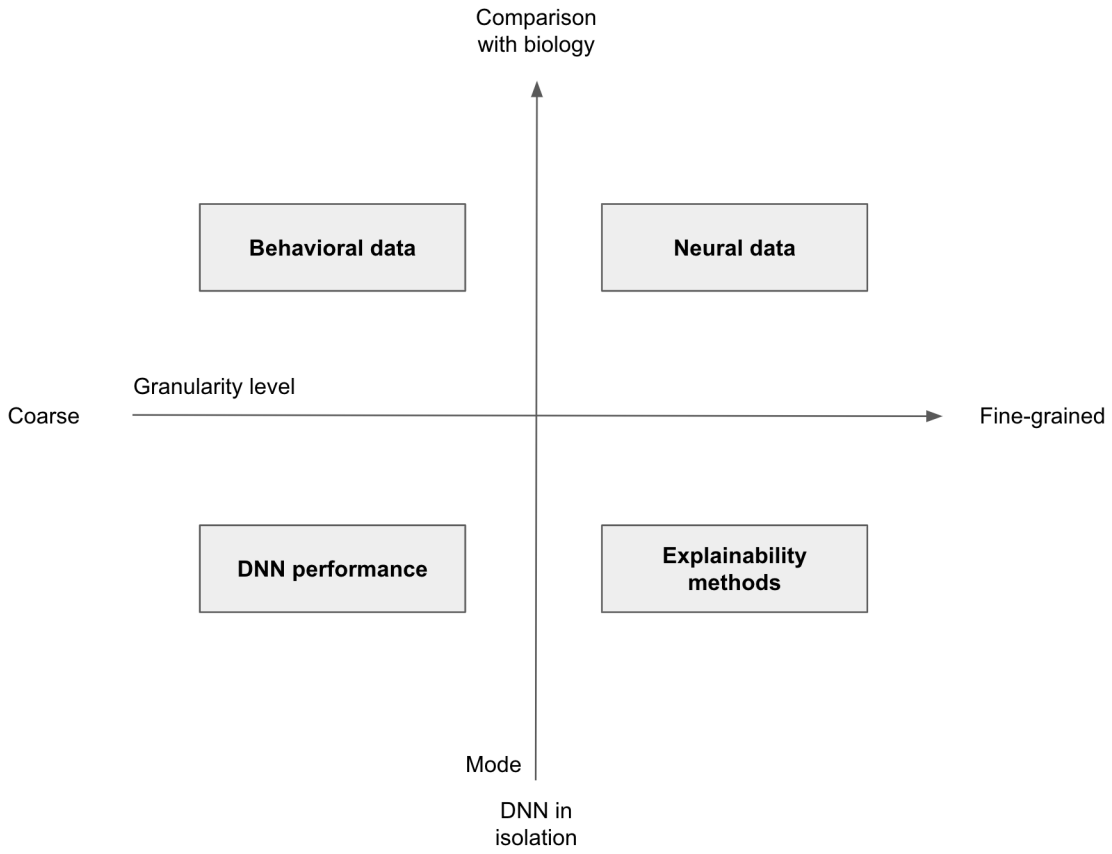


Figure 11: **Approaches toward understanding visual perception in DNNs**: Zooming out from the two approaches presented in our publications, there are other ways to deepen our understanding of DNNs. Specifically, this figure extends the axis of granularity (here: horizontal, see also Figure 3) by the axis of mode (here: vertical). The latter differentiates between approaches of comparing DNNs with biology and studying them in isolation. While our publications move to the top left and bottom right quadrants, the top right and bottom left quadrant appear as new approaches in the picture: DNNs can also be understood via *comparisons to neural data* and via studying their *performance in isolation*. Importantly, approaches positioned in the quadrants may also reach beyond their quadrant, e.g. an explainability method may address a more coarse level of understanding.

axes, namely the *granularity level* and the approach's *mode*. The former dimension varies from coarse fine-grained, and the latter dimension varies from investigations in isolation to comparisons with biology. Compared to Figure 3, the mode-axis is newly added and allows for a more accurate distinction of approaches. As a consequence of extending the space by this second dimension, the approaches presented in our three publications are moved to the upper left and lower right quadrant. Importantly, the human aspect of our explainability work does *not* correspond to the upper part of the mode-axis. The reason is that the explainability method itself is generated *in isolation* and our *human evaluations* are a means to infer the explainability method's informativeness (see Section 2.2). In contrast, the upper part of the mode axis refers to *direct* comparisons between the two systems.

Often, the four approaches are not clearly mappable to the dimension grid and can extend beyond one quadrant. What is more, studies not seldomly combine different approaches to obtain a more holistic understanding of a single phenomenon. Accordingly, the sections below outline a few examples for the four approaches as well as their combination, and also highlight new insights gained by them regarding visual perception in DNNs.

Tangent: Marr's levels of descriptions for computational vision
As a tangent, there is a close connection between the axis "granularity level" of Figure 11 and Marr's famous levels of descriptions for computational vision. In fact, the latter would be a suitable alternative name for the figure's vertical dimension.

Splitting up the "complex [visual] information processing task" into three levels of descriptions and explanations was a big advancement at the end of the 1970's (Poggio, 1981): Specifically, David Marr suggested to treat (1) the *computational* level, (2) the *algorithmic* — or sometimes representational — level, and (3) the *implementational* level. More specifically, the first level defines the problem and what the system does to solve it, the second level determines the processes of how the problem is solved, and the third level concerns the realization on the physical hardware. Importantly, the researcher claimed that explanations would only be complete if they cover this whole range (Poggio, 1981). Today, 40 years later, these three levels are still relevant and researchers like (Kay, 2018) advocate that all of them should be pursued[13].

Relating Marr's levels to the segments shown in Figure 11, the computational level largely covers the left half, and the algorithmic and implementational levels cover the right half. As explained throughout this thesis, the presented publications relate to Marr's first two levels. However, I referred to them as the "functional" and "algorithmic" levels in Sections 1.5.1 and 1.5.2. Please also note that the "implementational properties" we refer to in point i of the checklist are not meant in Marr's sense, but rather refer to the algorithmic level. For examples of machine

[13]For a balanced description, it is important to mention that opinions regarding the degree to which DNNs are appropriate models for all three levels and particularly the third one diverge. For a range of approaches as well as findings on similarities and differences, see e.g. Ma and Peters (2020); Kriegeskorte (2015); Yamins and DiCarlo (2016); Cichy and Kaiser (2019); Lonnqvist et al. (2021); Markram et al. (2011); Albrecht et al. (2002); Horwitz and Hass (2012); Lennie and Movshon (2005); Crick (1989); Lindsay (2021); Whittington and Bogacz (2019); Bartunov et al. (2018); Sacramento et al. (2018); Richards and Lillicrap (2019); Chollet (2021).

systems describing the implementational level (i.e. in Marr's sense), the reader is referred to e.g. literature on spiking neural networks and neuromorphic engineering (e.g. Cao et al., 2015; Pfeiffer and Pfeil, 2018; Liao et al., 2021), or examples in Sections 3.3.3 and 3.3.5.

3.3.1 Comparisons between DNNs and biological behavioral data

As already explained at various locations throughout this thesis, many different human behavioral aspects are investigated in DNNs in order to deepen our understanding of them. Typically, such studies reveal insights on a rather coarse, i.e. Marr's computational or our the functional level. Consequently, these approaches are positioned in the upper left quadrant (see Figure 11). Nonetheless, certain experiment designs can also reveal more detailed observations. These ones, in contrast, would be more adequately mapped to a more central area in the upper half corresponding to Marr's algorithmic level (as examples see case study II and III in Section 2.1.3).

Besides the numerous examples mentioned throughout this thesis, I here provide a few more. To start, a prominent finding regarding human-machine comparisons concerns stimulus augmentations with noise: Geirhos et al. (2018b) and Wichmann et al. (2017) found that DNN performance drops more sharply than that of humans on increasingly manipulated stimuli. Another interesting finding relates to similarity and typicality judgments of images. Here, Lake et al. (2015) found that DNNs *can* predict how typical humans find an image for a category, though evidence varied regarding whether DNNs can account for human similarity judgments between images (Jozwik et al., 2017; Rosenfeld et al., 2018). Finally, an example of a benchmark for comparisons between DNNs and humans is the "MIT/Tuebingen Saliency Benchmark" (Kümmerer et al.). It evaluates how well models can explain "what drives human eye movements" (Kümmerer et al.). Strictly speaking, it is debatable whether eye movements should be considered behavior or more fine-grained biological data. Therefore, this benchmark is another example that could be mapped further to the right in Figure 11.

3.3.2 Comparisons between DNNs and neural data

Another approach to deepen our understanding of visual perception in DNNs is comparisons between DNNs and neural activations. Because these types of studies usually unveil findings on a more fine-grained, i.e. algorithmic level, they are placed in the upper right quadrant in Figure 11.

In fact, comparisons on the neural level uncovered surprising similarities between biological and artificial systems and added to the excitement around DNNs. In numerous studies, this modern type of algorithm was shown to account well for neural activation in visual cortex, and in fact better than other models (Cadieu et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Kubilius et al., 2018; Cadena et al., 2019a). What is more, these studies revealed astounding hierarchical correspondence between DNN layers and areas of the ventral stream (Yamins et al., 2014; Güçlü and van Gerven, 2015; Seeliger et al., 2018; Eickenberg et al., 2017). As

a specific example, in one such early study, Yamins et al. (2014) recorded extracellular activity in macaque monkeys while they were viewing complex object images. The authors then found that DNNs of better classification performance also predict V4 and IT activity better, and more specifically, that the penultimate (last) layer best predicts activity in V4 (IT). These kinds of findings were made with data from different measuring modalities (fMRI (Han et al., 2019; Cichy et al., 2016; Güçlü and van Gerven, 2015; Eickenberg et al., 2017; Khaligh-Razavi and Kriegeskorte, 2014), magnetoencephalography (Cichy et al., 2016; Tacchetti et al., 2017; Seeliger et al., 2018), and electrophysiological cell recordings (Yamins et al., 2014; Cadena et al., 2019a,b; de Vries et al., 2020; Kuzovkin et al., 2018; Cadieu et al., 2014; Tripp, 2017; Khaligh-Razavi and Kriegeskorte, 2014)), in both primates (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Cadena et al., 2019a) and rodents (Cadena et al., 2019b,a; de Vries et al., 2020), and for static image (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Cadena et al., 2019b) as well as for temporally dynamic video data (Tacchetti et al., 2017; Eickenberg et al., 2017). Finally, an example of a benchmark for comparisons between DNNs and neural data is the "brain hierarchy" score, which quantifies *hierarchical* correspondence between DNN and human brain activity measured via fMRI (Nonaka et al., 2021).

3.3.3 DNN performance in isolation

Yet another way to increase our understanding of DNNs is to focus on scrutinizing their behavior under many different circumstances. Given most of these studies expose their abilities and limitations on a behavioral, i.e. coarse level, this approach is mapped to the lower left quadrant in Figure 11. However, as with the behavioral comparisons (see Section 3.3.1), some experiments can also show insights on a more fine-grained level or have connections to other approaches. Therefore, such studies can also be mapped to a more central area in the lower half.

Regarding investigations based on just DNNs, there are innumerous publications. To start, benchmarks have played a crucial role in accelerating advances in DNNs' visual perception. For example, object classification improved on the "ImageNet Large Scale Visual Recognition Challenge" (Deng et al., 2009; Russakovsky et al., 2015) from 74.2 % to 96.42 % within only four years (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016). Despite these big successes, there are still limitations for DNNs. As such, DNNs have been reported to fail in both the independent and identically distributed (i.i.d.) as well as the out of distribution (o.o.d.) cases. For instance, rare objects (Buolamwini and Gebru, 2018) as well as objects in unusual contexts (de Vries et al., 2019; Beery et al., 2018), poses (Alcorn et al., 2019), viewpoints (Alcorn et al., 2019) or geometrical formations (Engstrom et al., 2019b; Webster et al., 2018) represent challenges. These findings can mirror biases and fit the bigger picture of DNN's risks to overfit (Srivastava et al., 2014; Zhang et al., 2017; Morcos et al., 2018) and learn shortcuts (Geirhos et al., 2020a). As a way forward, more and more data sets with challenging conditions (e.g. corruptions) are being developed (e.g. Hendrycks and Dietterich, 2019; Wang et al., 2019; Geirhos et al., 2018a) and evaluating models on them is becoming common

practice (e.g. Michaelis et al., 2019). Finally, other interesting directions to obtain a deeper understanding of DNNs within investigations of pure DNN behavior are testing different training scenarios, including different objective functions, training schemes and architectures (e.g. Ulyanov et al., 2018; Gaier and Ha, 2019; Geirhos et al., 2021). Not seldomly, inspiration for such new hypotheses is taken from biology. For instance, Choksi et al. (2021) explore recurrent connections, and Pogodin et al. (2021) test biologically plausible training procedures. As before, these approaches can be seen as not exclusively belonging to the lower left quadrant but as also having a connection to the upper (very) right area due to the experiment's neural (implementational) nature.

3.3.4 Explainability methods

As already explained a few times in this thesis, explainability methods are a great way to deepen our understanding of machine visual perception. Depending on their nature, they can grant insights in various ways. While there are a few typical dimensions to differentiate these methods (see Section 1.5.2), the one that corresponds best to the axis of *granularity levels* is local vs. global methods. As most tools rather aim at a more fine-grained level, they are mapped to the lower right quadrant in Figure 11. Nonetheless, their insights can also extend beyond fine-grained levels and concern rather coarse levels.

Aside from feature visualizations, there are many other explainability methods, from which I here describe a few local and global ones. Prominent examples of local methods, which produce explanations for only one individual prediction of a data point, include SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016) and saliency maps (Simonyan et al., 2014; Zeiler and Fergus, 2014; Zhou et al., 2016; Sundararajan et al., 2017; Smilkov et al., 2017; Samek et al., 2021; Baehrens et al., 2010; Smilkov et al., 2017; Shrikumar et al., 2017; Ancona et al., 2018; Springenberg et al., 2015; Bach et al., 2015; Lewis et al., 2021; Shitole et al., 2021). More specifically, SHAP is a game-theory-based approach that estimates the contribution of each feature toward the model's prediction. LIME stands for "local interpretable model-agnostic explanations" and finds a surrogate model to explain the relative importance of different features. And saliency maps, which are also called "heatmaps", "sensitivity" or "pixel attribution maps", visualize the contribution of each pixel toward the DNN's classification. Regarding global methods, whose explanations stand for a whole model, a popular example is partial dependence plots (Friedman, 2001). They show the expected prediction when all but e.g. one feature are marginalized out. Finally, local explainability methods can be combined to grant more general insight into models.

3.3.5 Investigations beyond one quadrant

To deepen our understanding of visual perception in DNNs, investigations can integrate various aspects, such that a mapping to more than one quadrant is sensible. As an example, the benchmark "BrainScore" (Schrimpf et al., 2018) combines several neural predictivity scores as well as behavioral measures. Hence, it covers the whole

upper half of Figure 11 corresponding to comparisons between DNNs and biology. Another example is the study by Langlois et al. (2021). Here, the authors compare the importances of image regions derived with an explainability method (lower right quadrant) and from human psychophysical tasks (upper left quadrant). Finally, a prominent phenomenon studied from various angles is adversarial examples. As such, publications purely focusing on DNNs (lower left quadrant) revealed that adversarial examples not only transfer between model architectures and training sets (Goodfellow et al., 2015; Szegedy et al., 2013; Papernot et al., 2016; Charles et al., 2019; Ilyas et al., 2019; Liu et al., 2016) but also exist in 3D (Athalye et al., 2018). What is more, several potential explanations for the phenomenon of adversarials were explored. Two of them, which also fall into the lower left quadrant, identify certain image features as being responsible for misleading DNNs (Wang et al., 2020b) or granting humans robustness (Harrington and Deza, 2022). In contrast, other studies (Kim et al., 2020; Dapello et al., 2020) augment DNNs with biologically inspired features or building blocks, which grants the algorithms better robustness. Hence, Kim et al. (2020) hypothesize that "humans do not even see most adversarial perturbations." Relating these latter two studies to Figure 11, they not only relate to the lower left quadrant, but also to the upper (very) right area due to the studies' neural (implementational) motivation. Again corresponding to the upper right quadrant, Han et al. (2019) find that responses of human fMRI and DNNs to adversarial examples do not show strong correlations. Finally, regarding comparisons between DNNs and humans (upper left quadrant), researchers investigated how susceptible humans are to adversarials (Zhou and Firestone, 2019; Dujmović et al., 2020) and constructed adversarials to deceive humans (Elsayed et al., 2018). Taken together, these examples illustrate that leveraging ideas from more than one quadrant can be advantageous to understand machine vision more holistically.

In summary, there are numerous approaches toward understanding visual perception in machines. Certainly, each of them comes with its own strengths and weaknesses, and no single one is per se better than the other ones. Instead, which approach and experimental setup to choose always depends on both the objective as well as resource constraints.

3.4 Summary

All in all, I not only broadly discussed the two approaches from the presented publications but also zoomed out to outline other approaches. As explained, comparison studies and explainability methods do not only grant application opportunities and advantages but also come with challenges and limits. In the next chapter, I therefore propose future directions.

4 Outlook: Future directions

Over the last decade, deep convolutional neural networks dominated the field of visual perception. To a limited degree, we understand these algorithms. Specifically, they provide solutions for not only image classification (Krizhevsky et al., 2012) but also object detection (Ren et al., 2015), segmentation (Girshick et al., 2014) or facial recognition (Schroff et al., 2015). In the future, we hope that we can both build more powerful models suitable for more than one specific task as well as bring more light into our understanding of these rather black boxes.

4.1 Comparisons between DNN and human behavior

With respect to comparisons between DNNs and humans on a behavioral level, different future directions can be pursued.

To start off, research can be continued regarding *how* comparison studies are conducted. For example, the toolbox of experimental setups and metrics can be further widened. What is more, common good practices can be re-evaluated as well as further expanded. On the whole, better methods and procedures will permit greater insights.

Regarding the *subject* of future comparison studies, there are countless options. Besides further gestalt phenomena or single tasks, various more abstract abilities can be investigated[14]. As a matter of fact, the latter are where today's DNNs fall short by a large margin. For example, they do not possess a genuine understanding of e.g. physics or psychology (Lake et al., 2017), or a "common sense" (Zhu et al., 2020). Further, they fall short regarding meta-cognition, i.e. the ability to "notice when a task is hard or when they are likely to fail" (Wichmann et al., 2017). Also, current DNNs typically master just one task, whereas we humans are capable of combining our knowledge and abilities not only to solve different types of tasks but also to process input from different sensory modalities (Lake et al., 2017). Lastly, DNNs typically require lots of labeled data points and are not flexible in their learning strategy. From an optimistic point of view, fields like continual learning (Delange et al., 2021), meta-learning (Vanschoren, 2018), multi-task learning (Ruder, 2017), multi-modal learning (Wang, 2021) as well as semi-, self- and unsupervised learning (Schmarje et al., 2021) are contributing to progress with respect to the previously mentioned fields. Without any doubt, there is a lot of room for improvement and it is still a long journey until machines will be able to mirror more human abilities. Systematically and regularly evaluating how well they compare on these various phenomena will be an important pillar on this path.

Given the fast developments in Deep Learning, future comparisons should always be performed on *state-of-the-art* models. Great examples for this are evaluations of self-supervised (e.g. He et al., 2020) as well as vision transformer models (e.g. Dosovitskiy et al., 2020). These algorithms are the results of remarkable advancements

[14]Certainly, not all human functions should be transferred to machines, as is discussed for e.g. illusions by Lonnqvist et al. (2021). In fact, it is an entire discussion by itself whether machine systems should become completely human-like or to what extent these similarities are desirable.

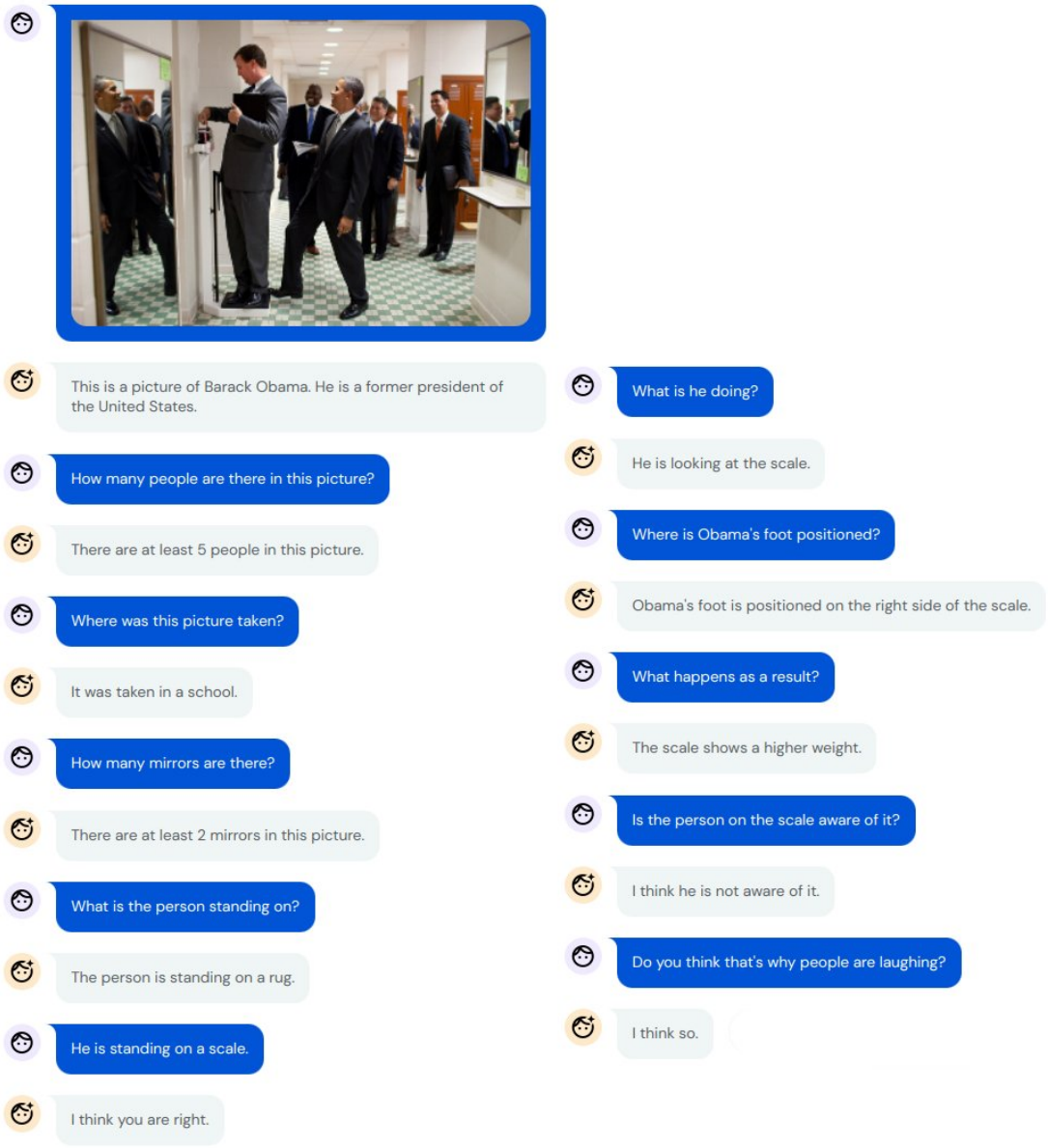


Figure 12: **An understanding machine?** The multi-modal algorithm Flamingo (Alayrac et al., 2022) seems to comprehend various details of the complex scene in the photograph, and converses about them in a (heavily engineered) dialogue (Ring, 2022b). 10 years ago, when Karpathy (2012) outlined numerous challenges in Computer Vision illustrated by this very picture, such advancements seemed out of reach. In the future, continuously and systematically evaluating how similar machine systems become to humans will continue to be important. (The figure is taken from (Ring, 2022a).)

based on the training styles and network architecture. In fact, the former reveal higher similarities with human perception (Geirhos et al., 2021; Storrs et al., 2021), human brain representations (Konkle and Alvarez, 2020) and neural data (Zhuang et al., 2021), and the latter exhibit a higher shape bias (Naseer et al., 2021; Tuli et al., 2021; Geirhos et al., 2018b). Finally, a third great example comes from the multimodal model Flamingo (Alayrac et al., 2022). As shown in Figure 12, this DNN can demonstrate significant visual and language understanding, as well as maybe even some common sense (Ring, 2022a). Indubitably, this dialogue is heavily engineered (Ring, 2022b) and the model does fail badly in numerous other situations (e.g. Alayrac, 2022). Nonetheless, it is a great example of where comparisons of machine and human behavior are essential. Altogether, future work should always focus on state-of-the-art models and reveal these best models' new abilities and remaining limitations.

In summary, it is of utmost importance to continue comparisons between humans and DNNs. Specifically, method improvements, evaluations of more human phenomena and novel models will contribute to advancing our understanding of DNNs. The hope is that we will then in turn be able to build even more powerful models.

4.2 Explainability methods

Besides comparison studies, explainability methods are another way to deepen our understanding of machine systems. However — and as exemplified by the presented second and third publication —, they do not always meet the expectations regarding their informativeness, increased transparency and human interpretability. Below, I outline implications of the investigated method's limited helpfulness as future directions for both feature visualizations as well as the bigger field of XAI.

4.2.1 Feature visualizations

An exciting future direction is to further develop a new feature visualization method. Ideally, these visualizations would be both easily interpretable by humans and faithfully highlight the relevant feature of the network. As alluded to above (see Section 3.2.4), improving the method by Nguyen et al. (2016a) may be a promising candidate: Photo-realistic feature visualizations combine the advantages of familiar appearances and focus on the structure in question. Specifically, updating the generation procedure of Nguyen et al. (2017) with a more modern GAN and classification network may be a low-hanging fruit.

Zooming out from the specific feature visualization method, future work can explore ways in which non-intuitive features can augment human feature understanding. Previous work on high-low frequency detectors (Schubert et al., 2021; Olah et al., 2020b) demonstrates that this is possible (see Figure 9 for an example). However, as this achievement was very labor-intensive, creating faster, systematic approaches of translating features to human-understandable terms would open up new opportunities. Questions like whether the appealing, "hand-picked" (Olah et al., 2017) feature visualizations in publications represent the generality of DNN units (Kriegeskorte,

2015) would then become obsolete. Realistically speaking, though, it is still a long way to understand the meaning of all DNN features.

Zooming out even further from machine visual systems and connecting this work to biological networks, an interesting direction is investigating the transferability of the presented psychophysical paradigms to *biological* networks. It is out of question that feature visualizations generated for mice or macaque monkeys elicit strong activations (Walker et al., 2019; Bashivan et al., 2019; Ponce et al., 2019). Does this transferability also hold for our psychophysical paradigms? I.e. are humans able to predict a biological neuron's activations based on references such as feature visualizations? If so, this would not only strengthen the parallels between artificial and biological visual systems but also increase the generality of our statements on the method of feature visualization.

Altogether, there are many exciting future directions for the method feature visualization.

4.2.2 Field of XAI

Widening the perspective to the field of XAI, our evaluation studies of the explainability method feature visualization and especially its limited helpfulness mirror more general challenges. Several of these are discussed below and represent ideas for directions of future work.

Evaluate the methods

To start off, explainability methods need to be evaluated. Just relying on intuition and judging what looks appropriate, intuitive and meaningful is not enough (Leavitt and Morcos, 2020; Poursabzi-Sangdeh et al., 2021). As such, the "seductive allure of visualization" (Leavitt and Morcos, 2020) can be misleading. Not only did our two publications show this for feature visualizations, but similar findings were made for other methods like saliency maps and brain images for cognitive sciences (McCabe and Castel, 2008). Specifically, and as explained in our third project, feature visualizations are believed to grant causal insight — though our data did not confirm this. Similarly, saliency maps often look convincing and reasonable. Nonetheless, various evaluations revealed that these explanations are misleading (Adebayo et al., 2018; Nie et al., 2018; Ghorbani et al., 2019; Sundararajan et al., 2017; Zhou et al., 2021; Hooker et al., 2019; Lin et al., 2020) and not as helpful as expected for humans (Jin et al., 2022; Nguyen et al., 2021; Fel et al., 2021; Alqaraawi et al., 2020; Chu et al., 2020; Shen and Huang, 2020). Finally, McCabe and Castel (2008) demonstrate that brain images where colorful blobs represent brain activity that is associated with cognitive processes increase scientific credibility compared to mere bar graphs or topographical maps.

Overall, the number of evaluations of explainability methods in the literature is growing, however, there is still a long way to go. As such, the fact that the evaluation overview article by Nauta et al. (2022) reviews more than 300 conference papers from the last 7 years illustrates that this topic is gaining importance. In fact, some explainability methods like attribution methods have been evaluated fairly extensively

(Fel and Vigouroux, 2020; Feng and Boyd-Graber, 2019; Schmidt and Biessmann, 2019; Adebayo et al., 2018; Lin et al., 2020; Tjoa and Guan, 2020; Zhou et al., 2021; Arras et al., 2021; Hooker et al., 2019; Nie et al., 2018; Sundararajan et al., 2017; Ghorbani et al., 2019; Prasad et al., 2021; Alqaraawi et al., 2020; Shitole et al., 2021; Bansal et al., 2021; Lai and Tan, 2019; Lai et al., 2020; Dinu et al., 2020; Folke et al., 2021; Shen and Huang, 2020; Chu et al., 2020; Chandrasekaran et al., 2018; Nguyen et al., 2021; Zhang et al., 2020). However, that only "1 in 3 papers evaluate exclusively with anecdotal evidence, and 1 in 5 papers evaluate with users" (Nauta et al., 2022) reflects that measuring the informativeness of explainability methods is by far no common practice yet. As such, feature visualizations had also not been evaluated before our works.

The core challenge of evaluating explainability methods is that there is no ground truth. As we do not fully understand how DNNs process information, we simply cannot generate the "perfect" explanation. And in turn, this means that it is tricky to decide when an explanation grants useful insights into DNNs. The lack of a ground truth moreover entails that it is unclear in what directions existing explainability methods should be further developed. Lord Kelvin described such a situation with the words: "If you cannot measure it, you cannot improve it."

To counteract the core challenge of evaluating explainability methods, several suggestions have been made and practices developed. For example, multiple explainability methods can be compared against each other and reveal which one grants the most useful insight (e.g. Hase and Bansal, 2020). Often, such comparisons also include simple baselines (e.g. Nguyen et al., 2021). Another, though less frequent, approach is to generate a custom data set. Here, ground truth is typically straight-forward (e.g. Zhou et al., 2021).

Besides workarounds for the lack of ground truth, many other suggestions regarding how to evaluate explainability methods have been put forward. For example, Lipton (2018) advocates evidence-based approaches, and Adadi and Berrada (2018) call for "developing formalized rigorous evaluation metrics and methods". This need for quantification is echoed by many other researchers, e.g. Amparore et al. (2021); Nguyen and Martínez (2020) and Leavitt and Morcos (2020). The authors of the latter paper further emphasize that "falsifiable hypotheses" have to be tested and that "merely proving the existence of something rarely tells us much about whether that phenomenon is relevant to the network." Following a different approach, Nauta et al. (2022) captured the many requirements of a good explanation in twelve properties concerning the content, presentation and user of explainability methods. And finally, on a different level, Doshi-Velez and Kim (2017) emphasize that the evaluation of an explainability method should "only [depend] on the quality of the explanation, [...] regardless of the correctness of the associated prediction".

The nature of evaluation studies can be categorized into either human psychophysical experiments or mathematical experiments and theoretical analysis. In this section, the focus is put on the latter, as the former was already a topic in our publications (see Sections 2.2 and 2.3) as well as Section 1.5.2. Furthermore, it is discussed again in Section 4.2.2.

The advantage of mathematical evaluations is their scalability and reproducibility. Typically, aspects such as consistency, robustness, and sensitivity of an explainability method are measured (e.g. Adebayo et al., 2018). Regarding theoretical analyses, their clarity is the biggest asset: A hypothesis is either supported or rejected.

The existence of human and mathematical evaluation studies raises the question whether these two types qualitatively yield the same findings. A few studies indeed conducted both types of measurements, i.e. mathematical and human evaluations, and measured their consistency. The result is discouraging: Mathematical and human evaluations are poorly correlated (Biessmann and Refiano, 2019; Nguyen et al., 2021; Fel et al., 2021) or even anti-correlated (Fel et al., 2021). In the bigger picture, this suggests that future work should focus on human evaluations — ultimately also because these are the agents affected in the real-world scenarios.

Taken together, the field of XAI has an enormous need for evaluation studies. Luckily, more and more researchers have been responding to this and are finding workarounds for the challenges around this endeavor. The biggest problem, i.e. the lack of ground truth, is closely related to the following topic.

Define explainability

Until today, there is no generally agreed upon definition for *explainability*. This also holds true for the term *interpretability*, that is often used interchangeably. While not uncommon for a new and rapidly developing field, the danger is that different researchers fill this gap differently. And as a consequence, unspoken expectations can arise or misunderstandings can emerge.

On the bright side, several scientists made suggestions how to define explainability. For example, Doshi-Velez and Kim (2017) see it as "the ability to explain or present in understandable terms to a human". Similarly, Biran and Cotton (2017) state that "systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation." While Adadi and Berrada (2018); Miller (2019) and Diakopoulos et al. (2021) largely agree with the above, Pion-Tonachini et al. (2021) make an explicit distinction between the two terms. They claim that explainability "concerns how well the internal mechanics of a specific AI model can be explained in human terms, i.e. *how* a model works", whereas interpretability concerns "how well properties of the world, e.g. cause and effect, can be observed and discovered and understood, when using that AI model as a lens on the natural or artificial system that generated that data, i.e. *why* an AI model works, in terms of properties of the world". Finally, Murdoch et al. (2019) define a framework of three desiderata for interpretability: Besides predictive accuracy, i.e. a measure of how good the model is, they suggest reporting descriptive accuracy, i.e. a value of how faithful an explainability method is, as well as relevancy, i.e. a judgment of how relevant the information is to the target audience.

Similar to the lack of a definition in words, there is also no mathematical term for explainability. As such, a general interpretability-term that could simply be added to the optimization function does not exist (Lipton, 2018). Nonetheless, approaches like including a lasso-sparsity term go in the direction of "developing richer loss functions"

(Lipton, 2018) and have indeed been shown to grant more interpretability (Kim, 2020).

In the bigger picture, the lack of an interpretability-term is an example of misaligned objectives. Specifically, Lipton (2018) states that "While the machine learning objective might be to reduce error, the real-world purpose is to provide useful information." Clearly, capturing this latter goal mathematically is (currently) out of reach. Instead, our workaround is designing simple and small proxy tasks — such as object recognition. In turn, (Doshi-Velez and Kim, 2017) argue that an "incompleteness in the problem formalization" of these substitute problems is the reason for the need of interpretability.

Widening the perspective even more, the history of other fields may permit a more optimistic outlook. For example, formalizing the criteria in fairness (Hardt et al., 2016) and privacy (Toubiana et al., 2010; Dwork et al., 2012) "allowed for a blossoming of rigorous research in these fields" (Doshi-Velez and Kim, 2017). Seeing the big difference that steps toward clear definitions can make is encouraging for the future of XAI.

In summary, the current lack of a definition for explainability or interpretability represents a challenge for the young field. Evidently, there is also no mathematical term. The next section depicts a minimal workaround for these difficulties.

Communicate clearly

The kind of explainability that a tool grants — or is believed to grant — should be *clearly communicated*. Lipton (2018) laments that "few authors articulate precisely what interpretability means or precisely how their proposed solution is useful." Instead, he suggests that researchers should "fix a specific definition" of what they mean by their model being interpretable. In a similar direction, Doshi-Velez and Kim (2017) propose a taxonomy to describe dimensions of interpretability such as what the basic unit of an explanation is and how different units of explanations are related. What is more, they advocate researchers can "do each other a service" by explaining aspects in their papers like how the problem formulation, that a specific explainability method aims to solve, is incomplete. Finally — and as already mentioned —, Doshi-Velez and Kim (2017) emphasize that "the claim of the research should match the type of the evaluation" (Doshi-Velez and Kim, 2017).

Regarding the explainability method feature visualizations, we gained a clearer impression of the kind of advantages that Chris Olah attributes to them in a video-call with him[15] as well as what he wishes we/I had communicated more clearly in personal communication. For example, Chris Olah considers feature visualizations only helpful for experts, not for laypeople. Further, they are valuable when deciding between competing theories of what a unit fires for. In the bigger picture, these examples are narrower than what we took away from the blog posts around this popular implementation by Olah et al. (2017). Finally, Chris Olah notes that modern works would use feature visualizations as "part of a set of tools for understanding neurons"

[15]The video call took place on October 15th, 2021, following interaction on twitter (Olah, 2021b). The two first authors as well as the two last authors of the third publication participated.

as well as that "many neurons are difficult to understand due to polysemanticity" (personal communication).

While clear communication of the envisioned — and potentially evaluated — advantages of an explainability tool is indispensable, the next section proposes another way to advance the field of XAI.

Develop better methods

To deepen our understanding of DNNs, better explainability methods have to be designed. On the one hand, many evaluation studies show that humans do not yet gain a satisfying level of understanding (Shen and Huang, 2020; Nguyen et al., 2021; Jeyakumar et al., 2020; Chandrasekaran et al., 2017; Hase and Bansal, 2020; Alufaisan et al., 2021; Dieber and Kirrane, 2020; Dinu et al., 2020). And on the other hand, studies reveal that humans can tend to place too much trust in algorithms and their explanations (Kaur et al., 2020; Kim et al., 2021b; Parliament, 2021; Poursabzi-Sangdeh et al., 2021) and that "there is no effect of explanations on users' performance compared to sole AI predictions" (Schemmer et al., 2022).

It is out of question that a myriad of different explainability methods already exists and that more and more techniques are being developed and improved. As such, each method yields a different "notion[s] of transparency" (Weller, 2019), and is therefore useful in a "different setting[s]" (Weller, 2019) as well as for a different target group. For example, counterfactual explanations (Wachter et al., 2017; Goyal et al., 2019; Wu et al., 2021; Sharma et al., 2019; Ustun et al., 2019; Mahajan et al., 2020; Karimi et al., 2020) answer the question what change in the input would alter the model's prediction and can be intuitive for laypeople. In contrast, feature visualizations (Olah et al., 2017; Erhan et al., 2009; Nguyen et al., 2017) reveal insights into DNN features and are more suitable for experts according to e.g. Chris Olah (see above) or Huang et al. (2020). And yet another explanation approach called "TCAV", which stands for "testing with concept activation vectors", is to take human concepts as the starting point, let them define the aspect they are interested in with pictures (e.g. stripes) and test a network's sensitivity to it (Kim et al., 2018a). Clearly, the diversity of methods and their use cases illustrate that "there is no universally appropriate approach" (Weller, 2019). Nonetheless, evaluation studies rarely reveal that explanations do exactly what they are expected to do. Therefore, further tuning explainability methods and developing new ones will remain an elementary pillar of XAI.

When developing an explainability method we should keep in mind that its informativeness is limited both by the model it depicts as well as by our human perception. The former assumes that the explainability tool illustrates the model in question *faithfully*. If that is the case, explanations reveal aspects of the algorithm's decision-making, regardless of whether those decisions are correct or not. While this is a limit of an AI system imposed by the performance of a model, another limit can be our human perception. As such, some aspects of DNNs are simply not recognizable to us. For example, adversarial examples, i.e. certain minuscule pixel changes in an input image, can have a large effects on DNN predictions. However, we humans do

not detect them. Consequently, when an explainability method faithfully depicts such aspects, they may not be meaningful to humans (Ilyas et al., 2019). Taken together, this trade-off between faithfulness and meaningfulness to humans, as well as limits of models can complicate the endeavor of developing effective explainability tools.

Finally, it is of utmost importance to critically evaluate whether the deployment of an explainability method is a good step in the first place. For low-stake decisions, it is agreed that explainability is not necessary (Bunt et al., 2012; Doshi-Velez and Kim, 2017). In contrast, for high-stake decisions, opinions are more diverse. For example, the AI Act requires transparency for technologies that interact with humans (Parliament, 2021). Also, the AI-healthcare start-up Pacmed emphasizes that explainability adds value to their product (Cina, 2021) and Jin et al. (2022) even consider "[b]eing able to explain the prediction [of an AI model] to clinical end-users [...] a necessity". Nonetheless, assessing when to invest in interpretability efforts and when to focus on improving the system itself can be tricky (Lipton, 2018; Weller, 2019). After all, the latter could potentially save more lives by e.g. better autonomous cars or cancer detection algorithms (Weller, 2019). Already Albert Einstein described such a situation as the "perfection of means and confusion of goals" (Einstein, 1941). In the more recent XAI literature, the trade-off between interpretability and accuracy is often discussed (Breiman, 2001). Zooming out further, transparency can open up even more harms when considering IP rights and efficiency in economy as well as privacy aspects (Weller, 2019). In the most extreme case against transparency, some voices advocate against explaining black-box decisions. For example, Cynthia Rudin defends the idea to only construct inherently interpretable models for high-stake decisions (Rudin, 2019). Further, and with respect to AI systems in healthcare, Ghassemi et al. (2021) favor "rigorous internal and external validation of AI models" instead of explainability. Altogether, these different opinions illustrate that deciding whether to deploy an explainability method is not trivial and ultimately depends on the specific case.

To summarize, there is still room for improvement of explainability tools, and development choices depend on many factors. The following and final section specifies aspects to consider beyond the field of XAI when creating explainability tools.

Collaborate with other fields

Many voices call for more interdisciplinary collaborations in XAI as the streams of research are currently fairly isolated (Abdul et al., 2018; Miller et al., 2017). As a result of the latter, the development of explainability methods and intelligible systems does not leverage its full potential. For example, explainability methods may seem rather designed for *researchers* than for end-users (Miller et al., 2017; Bhatt et al., 2020). Miller et al. (2017) believes that this is because *programmers* and not "interaction designers" make design decisions. As a way forward, researchers working together with both practitioners (Kaur et al., 2020) and designers (Zhu et al., 2018) could turn out fruitful. Similarly, more collaboration between XAI and Human-Computer-Interaction (HCI) (Abdul et al., 2018) as well as social sciences (Miller, 2019) such as experimental psychology (Taylor and Taylor, 2021) could be beneficial.

Concretely, evaluations of explainability methods can benefit from the "strong ethos" in the HCI-community that a system should "deliver on its intended task" (Doshi-Velez and Kim, 2017; Antunes et al., 2008; Lazar et al., 2017). As a matter of fact, many aspects beyond the usefulness of explainability tools are being evaluated. For example, how much do humans trust explainability methods and artificial decision making systems (e.g. Lim et al., 2009; Yin et al., 2019; Kaur et al., 2020; Kim et al., 2021b)? Or what is an appropriate cognitive load for parsing explanations (e.g. Abdul et al., 2020; Lage et al., 2019)? And how likely are humans to follow a machine's decision (e.g. Diprose et al., 2020; Poursabzi-Sangdeh et al., 2021)? Such findings grant useful insight and could even be a start for design principles of new explainability methods or their presentation (Lage et al., 2019)[16].

More generally, i.e. beyond just evaluations, Miller (2019) suggests that XAI should build on existing work in philosophy, cognitive science and social psychology. As such, he summarizes four main findings from mainly "everyday", human explanations that "explanatory agents" could be improved with (Miller, 2019): (1) "Explanations are contrastive", i.e. humans are usually not interested in why a certain event happened, but rather why that event happened instead of another one. (2) "Explanations are selected", i.e. no comprehensive list of causes is expected, but one or two reasons usually suffice. (3) Explanations are a social conversation, i.e. in most cases, they are interactive and relative to the questioner's beliefs. (4) Probabilities are not as important as causal links. Clearly, not every point is "feasible for all applications" (Miller, 2019), and some researchers have also found opposing results (Kulesza et al., 2013). Nonetheless, the above points may be useful inspiration for AI researchers.

Other insights from psychology may also serve as useful background information for the development of future explainability methods. For example, the subjective flavors mentioned earlier are reflected in a definition for explanations given in the field of psychology: Here, explanations are considered the "currency in which we exchanged beliefs" (Lombrozo, 2006). Further, humans are known to come up with very good explanations, but they are often wrong (Nisbett and Wilson, 1977). This means that our explanations are not based on any "true introspection", but "on a priori, implicit causal theories" (Nisbett and Wilson, 1977). Taking a step back, it is funny to realize that we accept our explanations even though we do not comprehend the biochemical processes of the human brain (Lipton, 2018). This contrasts the criticism that post-hoc explanations sometimes receive (Lipton, 2018; Rudin, 2019). Finally, (Weller, 2019) describes that humans "are not good at estimating how transparent we are ourselves when communicating with others". As such, the "illusion of transparency" depicts that we believe others would be able to discern our internal state better than they actually can (Gilovich et al., 1998).

The so-called "Copy machine study" (Langer et al., 1978; Weller, 2019) from the field of psychology highlights the potentially dangerous effect of a meaningless expla-

[16]Despite these different evaluation angles, critical voices raise the concern that "subjective views [...] and asking people what they prefer" (e.g. Jeyakumar et al., 2020) is no reasonable evaluation of the "correctness [of explainability methods]" Lipton (2018). In fact, Lipton (2018) claims that "the literature has dodged the issue of correctness [of explainability methods]." For a discussion on the latter, see Section 4.2.2.

nation. The setting is that participants request not to wait in line to make copies at a busy copy machine, but to be allowed to jump the line. Their questions differ in that they provide no explanation ("May I use the machine?"), an empty explanation ("May I use the machine because I have to make copies?") or a real explanation ("May I use the machine because I'm in a rush?"). The success rates of respectively 60 %, 93 %, and 94 % reveal that humans can be influenced by a meaningless explanation. In the bigger picture, this illustrates how complex the topic of explanations is.

Zooming out even further, views from e.g. philosophers and spiritual leaders complement the picture of explanations. For example, the Dalai Lama is reported to have said "A lack of transparency results in distrust and a deep sense of insecurity" (Weller, 2019). While this is generally accepted, the previous study shows that care has to be taken regarding whether that transparency is real. Going further than the Copy machine study, work in psychology (e.g. Levine and Schweitzer, 2015) shows that even incorrect explanations can increase trust: *Prosocial lies* have "benevolent motives or socially useful effects" (Dietz, 2018). The utilitarian perspective supports this view: Lying may be justified under certain circumstances (e.g. conflicting obligations, rights, interests) if an overall betterment is achieved (Dietz, 2018). In contrast, Kant's view is that lying is unacceptable under any circumstances (Kant, 1797). Altogether, explanations are complicated and often a case-to-case decision.

Relating these various views back to XAI, they mirror a few questions and trade-offs already discussed in other places of this thesis. For example, the Copy machine study can be seen as raising the question whether a reasonable amount of trust is placed in explanations for machine systems and whether acting in accordance with the provided explanations is reasonable behavior. As research in XAI has shown, this is not always the case (e.g. Kaur et al., 2020; Kim et al., 2021b; Schemmer et al., 2022). However, on a brighter note, the EU's AI Act Proposal (Parliament, 2021) already includes a section to counteract this pitfall: Humans shall be enabled to "remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system ('automation bias')". As another example, the aspect of pure transparency and well-intended augmentations of explanations can be related to the discussion around faithfulness to the machine system vs. meaningfulness to humans within XAI (see Section 4.2.2). Overall, these parallels are encouraging and suggest that the field of XAI is moving in a promising direction.

Narrowing the focus again on Machine Learning, XAI is closely related to fields such as fairness, privacy, reliability, robustness and causality (Doshi-Velez and Kim, 2017). In fact, explainable methods can be utilized to investigate and "confirm" "desiderata" related to these topics (Doshi-Velez and Kim, 2017). As such, insights from explainability methods may help evaluate fairness, i.e. whether an algorithm is biased or even discriminates against certain groups (e.g. Mothilal et al., 2020). With respect to privacy, where the general goal is to "protect[s] sensitive information in the data" (Doshi-Velez and Kim, 2017), there are two sides in relation to XAI: On the one hand, explainability and privacy can go hand in hand (e.g. Nori et al., 2021), and on the other hand, solutions must be found to avoid the risk of interpretable models revealing "characteristics of individual data points" (Harder et al., 2020). While the

list of examples can be extended, the general gist is that an active dialogue between the different communities is essential.

In summary, both insights from fields such as Human-Computer-Interaction, psychology and philosophy as well as desiderata from sub-fields specific to Machine Learning such as fairness or privacy are useful for XAI. Fortunately, parallels between findings in the previously mentioned research areas and XAI already exist. In the future, even stronger, interdisciplinary collaboration and active exchanges are likely to continue being beneficial to steer XAI in promising directions.

As illustrated, there are plenty of directions for future work in XAI and the method feature visualization. Given the increasing applications of DNNs as well as the growing legal demands for explanations, this young field is likely to become even more important. Presumably, establishing not only a definition but also common practices regarding evaluations and communication will unlock, again, even faster progress. Without any doubt, XAI will keep playing a major role in facilitating to deepen our understanding of visual perception in machines.

4.3 Summary

Understanding visual perception in machines is important. This is not only the case because of the increasingly many areas of DNN applications that directly affect human lives but also because of the growing legal demands. As illustrated in this thesis, our approaches leveraging human psychophysics, namely comparisons and evaluations of an explainability method, represent a step toward a deeper understanding of DNNs. In the future, further studies that both follow various approaches as well as foster interdisciplinary connections are necessary to expand insights at various levels. Ultimately, this better understanding will facilitate and guide developing even more powerful, more robust, and fairer machine systems of visual perception.

Acknowledgments

The work described in this thesis is the result of many collaborative efforts. I consider myself fortunate to have had the opportunity to be part of an inspiring, diverse and fun environment. Here, I want to thank these people.

First and foremost, I would like to express sincere gratitude to my three supervisors. Their support was invaluable, allowed me to develop as a researcher, and taught me the immense lessons of how many different perspectives a single topic can entail.

To Matthias Bethge, I am grateful for opening the door to his interdisciplinary and vivid lab full of bright people. His bird's eye perspective influenced our projects and put them into perspective for the community. Further, his enthusiasm not only for science but various topics beyond is inspiring. As such, his role in how the Machine Learning community in Tübingen is developing is impressive. Having had the opportunity to dip my toes into endeavors beyond pure research is something I am thankful for.

To Wieland Brendel, I am thankful for regular supervision. Knowing to have steady opportunities to discuss project progress, coding challenges, communications with other researchers, rebuttals and reviews was a great pillar. Besides his passion for academic research, his open-mindedness for optimizing work arrangements as well as his activities in outreach are inspiring and impressive. I value the freedom and support from him to pursue outreach activities myself.

Third, I appreciate the supervision and mentoring of Tom Wallis. Even though he officially took on other roles outside the University Tübingen, he regularly provided guidance, scientific advice and could always cheer a situation up with a good story. Without his expertise in psychophysics and statistical analyses, our experiments would by far not have reached their rigor. His factual approach and attitude of reporting clearly was a grounding counterbalance to the unprecedented speed and news-eagerness of the Deep Learning field.

Next, I would like to thank my co-authors. Without them, these projects would not have been possible. Scientific discussions and iteration after iteration of drafts taught me again how the a whole is so much greater than the sum of its parts.

Furthermore, I would like to express my gratitude to Heike König, Melanie Ertle-Palm, and Tina Gauger for support with administration.

Also, I would like to thank the lab members for many opportunities for discussions — be they of coding, infrastructure, scientific or geo-political nature — as well as a social atmosphere.

Further, I would like to thank my Thesis Advisory Committee members Isabel Valera and Felix Wichmann. Checking in on a yearly basis and receiving feedback from different perspectives was valuable.

Moreover, I am grateful for the IMPRS-IS community, a great network of smart and inspiring young scientists.

What is more, I would like to thank Marieke Mur and Niko Kriegeskorte for giving me the opportunity to dip my toes into research. The functional read-out project and experience in Cambridge is what inspired me to pursue a PhD. Similarly, a shout-out

goes to Clemens Grewe for continuously discussing career options as well as other topics with me since 2014.

Finally, I would like to thank all the people who provided feedback on drafts of this thesis, here listed in alphabetical order: Claudio, Lukas, Matthias B., Matthias K., Sabine, Simon, Steffi, Tom, and Wieland. Also, I appreciate Christina's and Dr. Sarah Müller's support on the final steps for the submission.

Lastly, I am immensely thankful to my family and friends for their endless support during this journey. They made Cyber Valley my home again. Not only their patience and encouragement, but also their critical opinions have shaped me. Thank you.

References

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–18, 2018.

Ashraf M. Abdul, Christian von der Weth, Mohan S. Kankanhalli, and Brian Y. Lim. COGAM: measuring and moderating cognitive load in machine learning model explanations. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM, 2020. doi: 10.1145/3313831.3376615.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Sravanti Addepalli, Dipesh Tamboli, R Venkatesh Babu, and Biplab Banerjee. Saliency-driven class impressions for feature visualization of deep neural networks. *arXiv preprint arXiv:2007.15861*, 2020.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

JB Alayrac. "it is great to see the excitement about flamingo! as shown in different examples during the last few days, interacting with flamingo has been quite fun, unique and sometimes mind blowing. however, flamingo has clear limitations as detailed in this flamingo! 1/11", 2022. URL `https://twitter.com/jalayrac/status/1524025887271829504?t=3kTFo3VQWnVsSbfeO4PEEw&s=19`. tweeted on 2022-05-10.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Duane G Albrecht, Wilson S Geisler, Robert A Frazor, and Alison M Crane. Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. *Journal of neurophysiology*, 88(2):888–913, 2002.

Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019.

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks:

a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.

Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6618–6626, 2021.

Elvio Amparore, Alan Perotti, and Paolo Bajardi. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Computer Science*, 7:e479, 2021.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks, 2018.

Pedro Antunes, Valeria Herskovic, Sergio F Ochoa, and Jose A Pino. Structuring dimensions for collaborative systems evaluation. *ACM computing surveys (CSUR)*, 44(2):1–28, 2008.

Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 2021.

arXiv. arxiv submission rate statistics, 2020. URL `https://arxiv.org/help/stats/2018_by_area#cs_yearly`. accessed on 2022-02-04.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

Nicholas Baker, Gennady Erlikhman, Philip J Kellman, and Hongjing Lu. Deep convolutional networks do not perceive illusory contours. In *CogSci*, 2018a.

Nicholas Baker, Hongjing Lu, Gennady Erlikhman, and Philip J Kellman. Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613, 2018b.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 511–520, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/barrett18a.html`.

Sergey Bartunov, Adam Santoro, Blake A Richards, Luke Marris, Geoffrey E Hinton, and Timothy Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *arXiv preprint arXiv:1807.04587*, 2018.

Pouya Bashivan, Kohitij Kar, and James J DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439), 2019.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

Guy Ben-Yosef and Shimon Ullman. Image interpretation above and below the object level. *Interface focus*, 8(4):20180020, 2018.

Guy Ben-Yosef, Liav Assif, and Shimon Ullman. Structured learning and detailed interpretation of minimal object images. *arXiv preprint arXiv:1711.11151*, 2017.

Guy Ben-Yosef, Liav Assif, and Shimon Ullman. Full interpretation of minimal images. *Cognition*, 171:65–84, 2018.

Guy Ben-Yosef, Gabriel Kreiman, and Shimon Ullman. Minimal videos: Trade-off between spatial and temporal information in human and machine vision. *Cognition*, 201:104263, 2020.

Guy Ben-Yosef, Gabriel Kreiman, and Shimon Ullman. What can human minimal videos tell us about dynamic recognition models? *arXiv preprint arXiv:2104.09447*, 2021.

Ari S Benjamin, Cheng Qiu, Ling-Qi Zhang, Konrad P Kording, and Alan A Stocker. Shared visual illusions between humans and artificial neural networks. In *Proceedings of the Annual Conference of Cognitive Computational Neuroscience. Available online at: https://ccneuro. org/2019/proceedings/0000585. pdf*, 2019.

Hanna Benoni, Daniel Harari, and Shimon Ullman. What takes the brain so long: Object recognition at the level of minimal images develops for up to seconds of presentation time. *arXiv preprint arXiv:2006.05249*, 2020.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3375624. URL https://doi.org/10.1145/3351095.3375624.

Irving Biederman. *Visual object recognition*, volume 2. MIT press Cambridge, MA, USA, 1995.

Felix Biessmann and Dionysius Irza Refiano. A psychophysics approach for quantitative comparison of interpretable computer vision models. *arXiv preprint arXiv:1912.05011*, 2019.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

Christophe Boesch. What makes us human (homo sapiens)? the challenge of cognitive cross-species comparison. *Journal of Comparative Psychology*, 121(3):227, 2007.

Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QO9-y8also-.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2018.

Cameron Buckner. The comparative psychology of artificial intelligences. May 2019. Added some missing references Corrected a misattribution of the animal-AI Olympics to Cambridge. Leverhulme CFI is a multi-institutional organization and the competition is being held more at Imperial College London.

Andrea Bunt, Matthew Lount, and Catherine Lauzon. Are explanations always important? a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pages 169–178, 2012.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Santiago A Cadena, Marissa A Weis, Leon A Gatys, Matthias Bethge, and Alexander S Ecker. Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–232, 2018.

Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019a.

Santiago A Cadena, Fabian H Sinz, Taliah Muhammad, Emmanouil Froudarakis, Erick Cobos, Edgar Y Walker, Jake Reimer, Matthias Bethge, Andreas Tolias, and Alexander S Ecker. How well do deep neural networks trained on object recognition characterize the mouse visual system? 2019b.

Charles Cadieu, Minjoon Kouh, Anitha Pasupathy, Charles E Connor, Maximilian Riesenhuber, and Tomaso Poggio. A model of v4 shape selectivity and invariance. *Journal of neurophysiology*, 98(3):1733–1750, 2007.

Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963, 2014.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 6(1):e00024–006, 2021.

Yongqiang Cao, Yang Chen, and Deepak Khosla. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113(1):54–66, 2015.

Laurent Caplette and Nicholas B. Turk-Browne. Computational reconstruction of mental representations using human behavior. *PsyArXiv preprint*, 2022. URL doi:10.31234/osf.io/7fdvw.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*, 2017.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*, 2018.

Zachary Charles, Harrison Rosenberg, and Dimitris Papailiopoulos. A geometric perspective on the transferability of adversarial directions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1960–1968. PMLR, 2019.

Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, B. ADOR, Andrea Alamia, and Rufin VanRullen. Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=v4vjMuXF-B.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):1–13, 2016.

Giovanni Cina. Personal communication with pacmed https://pacmed.ai/, 2021.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Eliana Colunga and Linda B Smith. A connectionist account of the object-substance distinction. In *Psychological Review*. Citeseer, 2005.

Jonathan Crabbé, Zhaozhi Qian, Fergus Imrie, and Mihaela van der Schaar. Explaining latent representations with a corpus of examples. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203): 129–132, 1989.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

Saskia EJ de Vries, Jerome A Lecoq, Michael A Buice, Peter A Groblewski, Gabriel K Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23(1):138–151, 2020.

Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. Workshop series on fairness, accountability and transparency in machine learning. principles for accountable algorithms and a social impact statement for algorithms, 2021. URL `https://www.fatml.org/resources/principles-for-accountable-algorithms`. accessed on 2021-09-21.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020.

Gil Diesendruck and Paul Bloom. How specific is the shape bias? *Child development*, 74(1):168–178, 2003.

Simone Dietz. White and prosocial lies. In *The Oxford handbook of lying*. 2018.

Jonathan Dinu, Jeffrey Bigham, and J Zico Kolter. Challenging common interpretability assumptions in feature attribution explanations. *arXiv preprint arXiv:2012.02748*, 2020.

William K Diprose, Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, and Reece Robinson. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, 27(4):592–600, 2020.

Adrien Doerig, Alban Bornet, Oh-Hyeon Choung, and Micahel H Herzog. Crowding reveals fundamental differences in local vs. global processing in humans and machines. *Vision research*, 167:39–45, 2020.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Marin Dujmović, Gaurav Malhotra, and Jeffrey S Bowers. What do adversarial images tell us about human vision? *Elife*, 9:e55978, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

Albert Einstein. The common language of science. *Out of my later years*, pages 111–113, 1941.

James Elder and Steven Zucker. The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research*, 33(7):981–991, 1993.

Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *arXiv preprint arXiv:1802.08195*, 2018.

Daniel C Elton. Common pitfalls when explaining ai and why mechanistic explanation is a hard problem. In *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 401–408. Springer, 2022.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019a.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019b.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

Gustav Theodor Fechner. *Elemente der Psychophysik*. Breitkopf & Härtel, 1860.

Reuben Feinman and Brenden M Lake. Learning inductive biases with simple neural networks. *arXiv preprint arXiv:1802.02745*, 2018.

Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. *arXiv preprint arXiv:2009.04521*, 2020.

Thomas FEL, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=hA-PHQGOjqQ`.

Thomas Fel, Julien Colin, Remi Cadene, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *arXiv preprint arXiv:2112.04417*, 2021.

Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 229–239, 2019.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.

Tomas Folke, ZhaoBin Li, Ravi B Sojitra, Scott Cheng-Hsin Yang, and Patrick Shafto. Explainable ai for natural adversarial images. *arXiv preprint arXiv:2106.09106*, 2021.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8730–8738, 2018.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021.

Adam Gaier and David Ha. Weight agnostic neural networks. *arXiv preprint arXiv:1906.04358*, 2019.

Alberto Gallace and Charles Spence. The cognitive and neural correlates of tactile memory. *Psychological bulletin*, 135(3):380, 2009.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32 (11):1231–1237, 2013.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018a.

Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. *arXiv preprint arXiv:1808.08750*, 2018b.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020a.

Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: Quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *arXiv preprint arXiv:2006.16736*, 2020b.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *arXiv preprint arXiv:2106.07411*, 2021.

Richard J Gerrig, Philip G Zimbardo, Andrew J Campbell, Steven R Cumming, and Fiona J Wilkes. *Psychology and life*. Pearson Higher Education AU, 2015.

Lisa Gershkoff-Stowe and Linda B Smith. Shape and the first hundred nouns. *Child development*, 75(4):1098–1114, 2004.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

Thomas Gilovich, Kenneth Savitsky, and Victoria Husted Medvec. The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology*, 75(2):332, 1998.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

EB Goldstein. Sensation and perception 8th edn (belmont, ca: Thomson wadsworth), 2010.

Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, and Marcelo Bertalmío. Convolutional neural networks can be deceived by visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12317, 2019.

Nicolas Gonthier, Yann Gousseau, and Saïd Ladjal. An analysis of the transfer learning of convolutional neural networks for artistic images. *arXiv preprint arXiv:2011.02727*, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 2019.

David M Green. Psychoacoustics and detection theory. *The Journal of the Acoustical Society of America*, 32(10):1189–1203, 1960.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386, 2020.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Chihye Han, Wonjun Yoon, Gihyun Kwon, Seungkyu Nam, and Daeshik Kim. Representation of white-and black-box adversarial examples in deep neural networks and humans: A functional magnetic resonance imaging study. *arXiv preprint arXiv:1905.02422*, 2019.

Frederik Harder, Matthias Bauer, and Mijung Park. Interpretable and differentially private predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4083–4090, Apr. 2020. doi: 10.1609/aaai.v34i04.5827. URL `https://ojs.aaai.org/index.php/AAAI/article/view/5827`.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Anne Harrington and Arturo Deza. Finding biological plausibility for adversarially robust features via metameric tasks. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=yeP_zx9vqNm`.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.

Daniel BM Haun, Fiona M Jordan, Giorgio Vallortigara, and Nicky S Clayton. Origins of spatial, temporal, and numerical cognition: Insights from comparative psychology. *Space, Time and Number in the Brain*, pages 191–206, 2011.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19000–19015. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/db5f9f42a7157abe65bb145000b5871a-Paper.pdf`.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6): 82–97, 2012a.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012b.

Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26(1): 1096–1106, 2019a.

Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26(1): 1096–1106, 2019b.

Yael Holzinger, Shimon Ullman, Daniel Harari, Marlene Behrmann, and Galia Avidan. Minimal recognizable configurations elicit category-selective responses in higher order visual cortex. *Journal of cognitive neuroscience*, 31(9):1354–1367, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32:9737–9748, 2019.

Gregory D Horwitz and Charles A Hass. Nonlinear analysis of macaque v1 color tuning reveals cardinal directions for cortical color processing. *Nature neuroscience*, 15(6):913–919, 2012.

Audrey Huang, Jeffrey Li, and Naveen Shankar. Interpretability, 2020. URL `https://blog.ml.cmu.edu/2020/08/31/6-interpretability/`. accessed on 2022-02-10.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160 (1):106–154, 1962.

Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020.

JASP Team. JASP (Version 0.16), 2021. URL `https://jasp-stats.org/`.

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33, 2020.

Li-Feng Jiang-Xie, Luping Yin, Shengli Zhao, Vincent Prevosto, Bao-Xia Han, Kafui Dzirasa, and Fan Wang. A common neuroendocrine substrate for diverse general anesthetics and sleep. *Neuron*, 102(5):1053–1065, 2019.

Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Evaluating explainable ai on a multimodal medical imaging task: Can existing algorithms fulfill clinical requirements? 2022.

Kamila M Jozwik, Nikolaus Kriegeskorte, Katherine R Storrs, and Marieke Mur. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, 8:1726, 2017.

Farnaz Khun Jush, Markus Biele, Peter Michael Dueppenbecker, Oliver Schmidt, and Andreas Maier. Dnn-based speed-of-sound reconstruction for automated breast ultrasound. In *2020 IEEE International Ultrasonics Symposium (IUS)*, pages 1–7. IEEE, 2020.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, and Sarah Mack. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

Immanuel Kant. Über ein vermeintes recht aus menschenliebe zu lügen, 1797.

Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions, 2020.

Andrej Karpathy. The state of computer vision and ai: we are really, really far away., 2012. URL `http://karpathy.github.io/2012/10/22/state-of-computer-vision/`. accessed on 2022-02-01.

D Katz. Chapter 1: Studies on surface touch, section 18-28. *The world of touch Translation by LE Krueger, Hillsdale, NJ: Lawrence Erlbaum Associates (Original work published 1925)*, 1989.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376219. URL `https://doi.org/10.1145/3313831.3376219`.

Kendrick N Kay. Principles for models of neural information processing. *NeuroImage*, 180:101–109, 2018.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Salman Khan, Alexander Wong, and Bryan P. Tripp. Task-driven learning of contour integration responses in a v1 model. In *NeurIPS 2020 Workshop SVRHM*, 2020. URL `https://openreview.net/forum?id=q2NlN7NpIm8`.

Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6(1):1–24, 2016.

Been Kim. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018a.

Been Kim, Emily Reif, Martin Wattenberg, and Samy Bengio. Do neural networks show gestalt phenomena? an exploration of the law of closure. *arXiv preprint arXiv:1903.01069*, 2(8), 2019.

Been Kim, Emily Reif, Martin Wattenberg, Samy Bengio, and Michael C Mozer. Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain & Behavior*, pages 1–13, 2021a.

Edward Kim, Jocelyn Rego, Yijing Watkins, and Garrett T Kenyon. Modeling biological immunity to adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4666–4675, 2020.

Eunji Kim. Interpretable and accurate convolutional neural networks for human activity recognition. *IEEE Transactions on Industrial Informatics*, 16(11):7190–7198, 2020.

Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-so-clevr: learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4): 20180011, 2018b.

Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. Hive: Evaluating the human interpretability of visual explanations, 2021b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

O Koehler. Counting experiments on a common raven and comparative experiments on humans. *Zeitschrift für Tierpsychologie*, 5(3):575–712, 1943.

Kurt Koffka. *Principles of Gestalt psychology*. Harcourt Brace, New York, 1935.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

Wolfgang Köhler. The mentality of apes. *New York: Kegan Paul, Trench, Trubner & Co*, 1925.

Talia Konkle and George A Alvarez. Instance-level contrastive learning yields human brain-like representation without category-supervision. *bioRxiv*, 2020.

Ilona Kovacs and Bela Julesz. A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences*, 90(16):7495–7497, 1993.

Nikolaus Kriegeskorte. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1(1):417–446, 2015. doi: 10.1146/annurev-vision-082114-035447. URL `https://doi.org/10.1146/annurev-vision-082114-035447`. PMID: 28532370.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, page 408385, 2018.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.

Nesaretnam Barr Kumarakulasinghe, Tobias Blomberg, Jintai Liu, Alexandra Saraiva Leao, and Panagiotis Papapetrou. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 7–12. IEEE, 2020.

Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017.

Ilya Kuzovkin, Raul Vicente, Mathilde Petton, Jean-Philippe Lachaux, Monica Baciu, Philippe Kahane, Sylvain Rheims, Juan R Vidal, and Jaan Aru. Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Communications biology*, 1(1):1–12, 2018.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67, 2019.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Brenden M Lake, Wojciech Zaremba, Rob Fergus, and Todd M Gureckis. Deep neural networks predict category typicality ratings for images. In *CogSci*, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Himabindu Lakkaraju, Julius Adebayo, and Sameer Singh. Explaining ml predictions: State-of-the-art, challenges, and opportunities, 2020. URL `https://www.youtube.com/watch?v=EbpU4p_0hes`. accessed on 2021-10-11.

Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.

Thomas A Langlois, Haicheng Charles Zhao, Erin Grant, Ishita Dasgupta, Thomas L Griffiths, and Nori Jacoby. Passive attention in artificial neural networks predicts human visual selectivity. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Peter Lennie and J Anthony Movshon. Coding of color and form in the geniculostriate visual pathway (invited review). *JOSA A*, 22(10):2013–2033, 2005.

Dennis M Levi, Cong Yu, Shu-Guang Kuai, and Elizabeth Rislove. Global contour processing in amblyopia. *Vision Research*, 47(4):512–524, 2007.

Emma E Levine and Maurice E Schweitzer. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106, 2015.

Noah Lewis, Robyn Miller, Harshvardhan Gazula, Md Mahfuzur Rahman, Armin Iraji, Vince. D. Calhoun, and Sergey Plis. Can recurrent models know more than we do? In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 243–247, 2021. doi: 10.1109/ICHI52183.2021.00046.

Fuyou Liao, Feichi Zhou, and Yang Chai. Neuromorphic vision sensors: Principle, progress and perspectives. *Journal of Semiconductors*, 42(1):013105, 2021.

Qianli Liao and Tomaso Poggio. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*, 2016.

Timothy P Lillicrap and Konrad P Kording. What does it mean to understand a neural network? *arXiv preprint arXiv:1907.06374*, 2019.

Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 2119–2128, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi: 10.1145/1518701.1519023.

Yi-Shan Lin, Wen-Chuan Lee, and Z Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. *arXiv preprint arXiv:2009.10639*, 2020.

Akis Linardos, Matthias Kummerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021.

Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

Grace W Lindsay and Thomas Serre. Deep learning networks and visual perception. In *Oxford Research Encyclopedia of Psychology*. 2021.

Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Gunter Loffler, Hugh R Wilson, and Frances Wilkinson. Local and global contributions to shape discrimination. *Vision Research*, 43(5):519–530, 2003.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Ben Lonnqvist, Alasdair DF Clarke, and Ramakrishna Chakravarthi. Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, 126: 262–274, 2020.

Ben Lonnqvist, Alban Bornet, Adrien Doerig, and Michael H Herzog. A comparative biology approach to dnn modeling of vision: A focus on differences, not similarities. *Journal of Vision*, 21(10):17–17, 2021.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Wei Ji Ma and Benjamin Peters. A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181*, 2020.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers, 2020.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

Henry Markram, Karlheinz Meier, Thomas Lippert, Sten Grillner, Richard Frackowiak, Stanislas Dehaene, Alois Knoll, Haim Sompolinsky, Kris Verstreken, Javier DeFelipe, et al. Introducing the human brain project. *Procedia Computer Science*, 7:39–42, 2011.

Birgit Mathes and Manfred Fahle. Closure facilitates contour integration. *Vision research*, 47(6):818–827, 2007.

Rainer Mausfeld. No psychology in–no psychology out. *Psychologische Rundschau*, 54(3):185–91, 2003.

David P McCabe and Alan D Castel. Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition*, 107(1):343–352, 2008.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

David A Mély, Drew Linsley, and Thomas Serre. Complementary surrounds explain diverse contextual phenomena across visual modalities. *Psychological review*, 125 (5):769, 2018.

N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro. Testing deep neural networks on the same-different task. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2019. doi: 10.1109/CBMI.2019.8877412.

Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Recurrent vision transformer for solving visual reasoning problems. *arXiv preprint arXiv:2111.14576*, 2021a.

Nicola Messina, Giuseppe Amato, Fabio Carrara, Claudio Gennaro, and Fabrizio Falchi. Solving the same-different task with convolutional neural networks. *Pattern Recognition Letters*, 143:75–80, 2021b.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.

Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pages 1–66, 2021.

Marvin Minsky and Seymour Papert. Perceptrons. 1969.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Melanie Mitchell. Abstraction and analogy-making in artificial intelligence. *arXiv preprint arXiv:2102.10717*, 2021a.

Melanie Mitchell. Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021b.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. URL `https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`. accessed on 2022-01-24.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=o2mbl-Hmfgd`.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, pages 3387–3395, 2016a.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016b.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3510–3520. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.374.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 55–76. Springer, 2019.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=OKPS9YdZ8Va.

Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International Conference on Machine Learning*, pages 3809–3818. PMLR, 2018.

Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.

Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013, 2021. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2021.103013. URL https://www.sciencedirect.com/science/article/pii/S2589004221009810.

Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. Accuracy, interpretability, and differential privacy via explainable boosting. In *International Conference on Machine Learning*, pages 8227–8237. PMLR, 2021.

Fabian Offert. "i know it when i see it". visualization and intuitive interpretability. *arXiv preprint arXiv:1711.08042*, 2017.

Fabian Offert and Peter Bell. Perceptual bias and technical metapictures: critical machine vision as a humanities challenge. *AI & SOCIETY*, pages 1–12, 2020.

Chris Olah. "it's great to see critical analysis of feature visualization! but i'm not sure this really gets at the core motivation of feature visualization. we know dataset examples are easier to read. feature visualization is helpful because it gets at causality.", 2021a. URL https://twitter.com/ch402/status/1321140564964765696. tweeted on 2020-10-27.

Chris Olah. "unfortunately, i don't think this gets at causality. to use the example of the dog head vs eye that i gave earlier, a square occluding a dog head would also occlude an eye.", 2021b. URL https://twitter.com/ch402/status/1447623297064071170. tweeted on 2021-10-11.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 2020a. doi: 10.23915/distill.00024.002. https://distill.pub/2020/circuits/early-vision.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020b. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

Chris Olah, Nick Cammarata, Chelsea Voss, Ludwig Schubert, and Gabriel Goh. Naturally occurring equivariance in neural networks. *Distill*, 5(12):e00024–004, 2020c.

OpenAI. Openai microscope. `https://microscope.openai.com/models`, 2020. (Accessed on 25/02/2022).

Zhaoyang Pang, Callum Biggs O'May, Bhavin Choksi, and Rufin VanRullen. Predictive coding feedback results in perceived illusory contours in a recurrent neural network, 2021.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Seymour A Papert. The summer vision project. 1966.

European Parliament. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206, 2021. accessed on 2021-12-17.

Michael Petrov, Chelsea Voss, Ludwig Schubert, Nick Cammarata, Gabriel Goh, and Chris Olah. Weight banding. *Distill*, 6(4):e00024–009, 2021.

Michael Pfeiffer and Thomas Pfeil. Deep learning with spiking neurons: opportunities and challenges. *Frontiers in neuroscience*, 12:774, 2018.

Luca Pion-Tonachini, Kristofer Bouchard, Hector Garcia Martin, Sean Peisert, W Bradley Holtz, Anil Aswani, Dipankar Dwivedi, Haruko Wainwright, Ghanshyam Pilania, Benjamin Nachman, et al. Learning from learning machines: a new generation of ai technology to meet the needs of science. *arXiv preprint arXiv:2111.13786*, 2021.

T. Poggio. Marr's computational approach to vision. *Trends in Neurosciences*, 4:258–262, 1981. ISSN 0166-2236. doi: https://doi.org/10.1016/0166-2236(81)90081-3. URL `https://www.sciencedirect.com/science/article/pii/0166223681900813`.

Roman Pogodin, Yash Mehta, Timothy P Lillicrap, and Peter E. Latham. Towards biologically plausible convolutional networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=ibD-yZEVBUX`.

Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior?, 2021.

Guillermo Puebla and Jeffrey Bowers. Can deep convolutional neural networks support relational reasoning in the same-different task? *bioRxiv*, 2021.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

Xuanchi Ren, Tao Yang, Li Erran Li, Alexandre Alahi, and Qifeng Chen. Safety-aware motion prediction with unseen vehicles for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15731–15740, October 2021.

research.com. Top computer science conferences for machine learning, data mining & artificial intelligence, 2022. URL `https://research.com/conference-rankings/computer-science/machine-learning`. accessed on 2022-02-04.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Blake A Richards and Timothy P Lillicrap. Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology*, 54:28–36, 2019.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Maximilian Riesenhuber and Tomaso Poggio. Computational models of object recognition in cortex: A review. 2000.

Roman Ring. "10 yrs ago @karpathy wrote a blog post on the outlook of ai: https://karpathy.github.io/2012/10/22/state-of-computer-vision/ in which he describes how difficult it would be for an ai to understand a given photo, concluding "we are very, very far and this depresses me." today, our flamingo steps up to the challenge.", 2022a. URL `https://twitter.com/Inoryy/status/1522621712382234624`. tweeted on 2022-05-06.

Roman Ring. "to be clear, of course we're not quite there yet. i had to lead the dialogue, correct rug <> scale, and explicitly ask about the joke. andrej's "challenge" greatly influenced my outlook on ai and it was really exciting for me to see flamingo handle it at all. i'll take "cute" :)", 2022b. URL `https://twitter.com/Inoryy/status/1522636310267240450?t=QsufgAOqpHSGchJ9Xgs1Qg&s=19`. tweeted on 2022-05-06.

Dario L Ringach and Robert Shapley. Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision research*, 36(19):3037–3050, 1996.

Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2940–2949. JMLR. org, 2017.

George John Romanes. *Animal intelligence*. D. Appleton, 1883.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Amir Rosenfeld, Markus D Solbach, and John K Tsotsos. Totally looks like-how humans compare, compared to machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1961–1964, 2018.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge.

International Journal of Computer Vision (IJCV), 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *arXiv preprint arXiv:1810.11393*, 2018.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4967–4976. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/e6acf4b0f69f6f6e60e9a815938aa1ff-Paper.pdf`.

Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis on the utility of explainable artificial intelligence in human-ai decision-making. *arXiv preprint arXiv:2205.05126*, 2022.

Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, 9:82146–82168, 2021.

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.

Andrew J Schofield, Iain D Gilchrist, Marina Bloj, Ales Leonardis, and Nicola Bellotto. Understanding images in biological and computer vision, 2018.

scholar.google.com. Google scholar: Top publications. categories - engineering & computer science - subcategories, 2022. URL `https://scholar.google.de/citations?view_op=top_venues&hl=en&vq=eng`. accessed on 2022-02-04.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL `https://www.biorxiv.org/content/10.1101/407007v2`.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Ludwig Schubert, Chelsea Voss, Nick Cammarata, Gabriel Goh, and Chris Olah. High-low frequency detectors. *Distill*, 6(1):e00024–005, 2021.

Marco Seeland and Patrick Mäder. Multi-view classification with convolutional neural networks. *Plos one*, 16(1):e0245230, 2021.

Katja Seeliger, Matthias Fritsche, Umut Güçlü, Sanne Schoenmakers, J-M Schoffelen, Sander E Bosch, and MAJ Van Gerven. Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, 180:253–266, 2018.

Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399–426, 2019.

Thomas Serre, Gabriel Kreiman, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, and Tomaso Poggio. A quantitative theory of immediate visual recognition. *Progress in brain research*, 165:33–56, 2007a.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15): 6424–6429, 2007b.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172, 2020.

Henry Shevlin and Marta Halina. Apply rich psychological terms in ai with care. *Nature Machine Intelligence*, 1(4):165–167, 2019.

Vivswan Shitole, Li Fuxin, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. One explanation is not enough: Structured attention graphs for image classification. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Edward H Shortliffe and Bruce G Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4):351–379, 1975.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences, 2017.

Stefan Sietzen, Mathias Lechner, Judy Borowski, Ramin Hasani, and Manuela Waldner. Interactive analysis of cnn robustness. *Computer Graphics Forum (Proceedings of Pacific Graphics 2021)*, 40(7), 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

Sahil Singla and Soheil Feizi. Causal imagenet: How to discover spurious features in deep learning?schroff2015facenet. *arXiv preprint arXiv:2110.04301*, 2021.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Linda B Smith, Susan S Jones, Barbara Landau, Lisa Gershkoff-Stowe, and Larissa Samuelson. Object name learning provides on-the-job training for attention. *Psychological science*, 13(1):13–19, 2002.

Pete Souza. President barack obama jokingly puts his toe on the scale as trip director marvin nicholson, unaware to the president's action, weighs himself as the presidential entourage passed through the volleyball locker room at the university of texas in austin, texas, aug. 9, 2010, 2010. URL `https://www.flickr.com/photos/obamawhitehouse/4921383047/`. accessed on 2022-02-02.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Sanjana Srivastava, Guy Ben-Yosef, and Xavier Boix. Minimal images in deep neural networks: Fragile object recognition in natural images. *arXiv preprint arXiv:1902.03227*, 2019.

Sebastian Stabinger, Antonio Rodríguez-Sánchez, and Justus Piater. 25 years of cnns: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks*, pages 380–387, Cham, 2016. Springer, Springer International Publishing.

Sebastian Stabinger, David Peer, Justus Piater, and Antonio Rodríguez-Sánchez. Evaluating the progress of deep learning for visual relational concepts. *Journal of Vision*, 21(11):8–8, 2021.

Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, pages 1–16, 2021.

Eric D Sun and Ron Dekel. Imagenet-trained deep neural networks exhibit illusion-like response to the scintillating grid. *Journal of Vision*, 21(11):15–15, 2021.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

William R Swartout. Xplain: A system for creating and explaining expert consulting programs. *Artificial intelligence*, 21(3):285–325, 1983.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1–9. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298594.

Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

Andrea Tacchetti, Leyla Isik, and Tomaso Poggio. Invariant recognition drives neural representations of action sequences. *PLoS computational biology*, 13(12):e1005859, 2017.

Keiji Tanaka. Neuronal mechanisms of object recognition. *Science*, 262(5134):685–688, 1993.

Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018.

Jie Tang. Conference rank: Computer science, 2022. URL `https://www.aminer.org/ranks/conf`. accessed on 2022-02-04.

Michael J Tarr. News on views: pandemonium revisited. *nature neuroscience*, 2(11): 932–935, 1999.

Alexa R Tartaglini, Wai Keen Vong, and Brenden M Lake. A developmentally-inspired examination of shape versus texture bias in machines. *arXiv preprint arXiv:2202.08340*, 2022.

J Eric T Taylor and Graham W Taylor. Artificial cognition: How experimental psychology can help generate explainable artificial intelligence. *Psychonomic Bulletin & Review*, 28(2):454–475, 2021.

Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *nature*, 381(6582):520–522, 1996.

Erico Tjoa and Cuntai Guan. Quantifying explainability of saliency methods in deep neural networks. *arXiv preprint arXiv:2009.02899*, 2020.

Michael Tomasello and Josep Call. Assessing the validity of ape-human comparisons: A reply to boesch (2007). 2008.

Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. Adnostic: Privacy preserving targeted advertising. In *Proceedings Network and Distributed System Symposium*, 2010.

Bryan P Tripp. Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3551–3560. IEEE, 2017.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

Tal Tversky, Wilson S Geisler, and Jeffrey S Perry. Contour grouping: Closure effects are explained by good continuation and proximity. *Vision Research*, 44(24):2769–2777, 2004.

Mike Tyka. Class visualization with bilateral filters, February 2016. URL `https://mtyka.github.io/deepdream/2016/02/05/bilateral-class-vis.html`. (Accessed on 09/26/2020).

Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10):2744–2749, 2016.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

Mohit Vaishnav, Remi Cadene, Andrea Alamia, Drew Linsley, Rufin VanRullen, and Thomas Serre. Understanding the computational demands underlying visual reasoning. *arXiv preprint arXiv:2108.03603*, 2021.

Ruben S van Bergen and Nikolaus Kriegeskorte. Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65: 176–193, 2020.

Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

Kimberly M Villalobos, Vilim Stih, Amineh Ahmadinejad, Jamell Dozier, Andrew Francl, Frederico Azevedo, Tomotake Sasaki, and Xavier Boix. Do deep neural networks for segmentation understand insideness? 04/2020 2020.

Anna Volokitin, Gemma Roig, and Tomaso Poggio. Do deep neural networks suffer from crowding? *arXiv preprint arXiv:1706.08616*, 2017.

Chelsea Voss, Nick Cammarata, Gabriel Goh, Michael Petrov, Ludwig Schubert, Ben Egan, Swee Kiat Lim, and Chris Olah. Visualizing weights. *Distill*, 6(2):e00024–007, 2021a.

Chelsea Voss, Gabriel Goh, Nick Cammarata, Michael Petrov, Ludwig Schubert, and Chris Olah. Branch specialization. *Distill*, 6(4):e00024–008, 2021b.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

Han-Ying Wang, Kohgaku Eguchi, Takayuki Yamashita, and Tomoyuki Takahashi. Frequency-dependent block of excitatory neurotransmission by isoflurane via dual presynaptic mechanisms. *Journal of Neuroscience*, 40(21):4103–4115, 2020a.

Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *arXiv preprint arXiv:1905.13549*, 2019.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020b.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Yang Wang. Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–25, 2021.

Emily J Ward. Exploring perceptual illusions in deep neural networks. *bioRxiv*, page 687905, 2019.

Eiji Watanabe, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka. Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, 9:345, 2018.

Brandon Richard Webster, Samuel E Anthony, and Walter J Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2280–2286, 2018.

Donglai Wei, Bolei Zhou, Antonio Torrabla, and William Freeman. Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379*, 2015.

Daniel S Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6):70–79, 2019.

Adrian Weller. Transparency: Motivations and challenges, 2019.

Max Wertheimer. Laws of organization in perceptual forms. *A source book of Gestalt Psychology*, 1, 1923.

James CR Whittington and Rafal Bogacz. Theories of error back-propagation in the brain. *Trends in cognitive sciences*, 23(3):235–250, 2019.

Felix A Wichmann, David HJ Janssen, Robert Geirhos, Guillermo Aguilar, Heiko H Schütt, Marianne Maertens, and Matthias Bethge. Methods and measurements to compare men against machines. *Electronic Imaging*, 2017(14):36–45, 2017.

Eric Wong, Shibani Santurkar, and Aleksander Mądry. Leveraging sparse linear layers for debuggable deep networks, 2021.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*, 2021.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Zhennan Yan and Xiang Sean Zhou. How intelligent are convolutional neural networks? *arXiv preprint arXiv:1709.06126*, 2017.

Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. Understanding the effect of accuracy on trust in machine learning models. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 279. ACM, 2019. doi: 10.1145/3290605.3300509.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017.

Renqiao Zhang, Jiajun Wu, Chengkai Zhang, William T Freeman, and Joshua B Tenenbaum. A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *arXiv preprint arXiv:1605.01138*, 2016.

Xi Zhang, Xiaolin Wu, and Jun Du. Challenge of spatial cognition for deep learning. *arXiv preprint arXiv:1908.04396*, 2019.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *arXiv preprint arXiv:2104.14403*, 2021.

Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature communications*, 10(1):1–9, 2019.

Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2018.

Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118 (3), 2021.

Roland S Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas SA Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of cnn activations? *arXiv preprint arXiv:2106.12447*, 2021.

Audun M. Øygard. Visualizing googlenet classes, Jun 2016. URL `https://www.auduno.com/2015/07/29/visualizing-googlenet-classes/`. (Accessed on 09/26/2020).

Appendix A: Additional results on Section 2.2 "Exemplary natural images explain CNN activations better than state-of-the-art feature visualization"

The following four figures display example trials from the additional experiment with the feature visualizations from Nguyen et al. (2017) and CaffeNet.

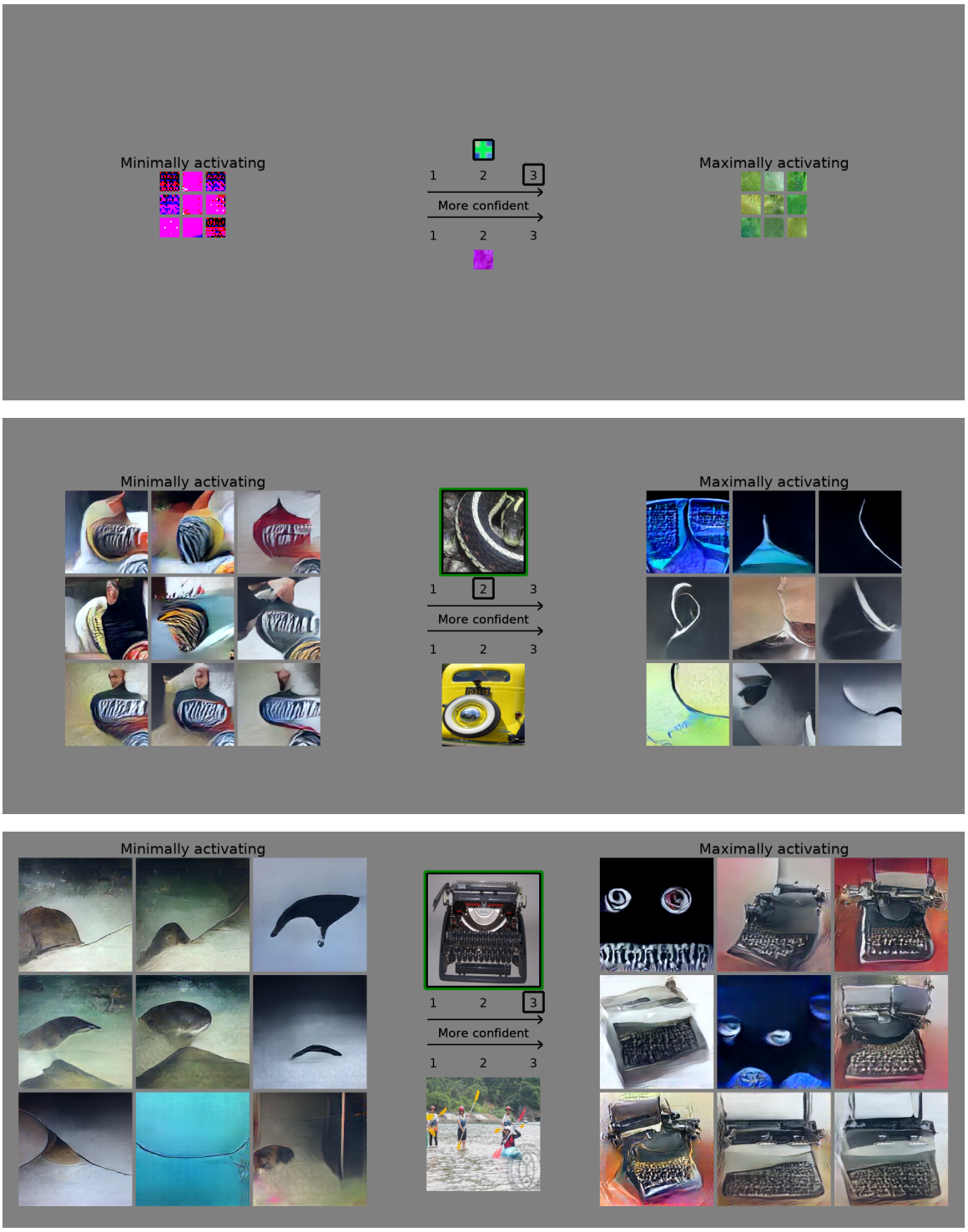


Figure 13: Example trials of the synthetic condition that seem fairly easy.

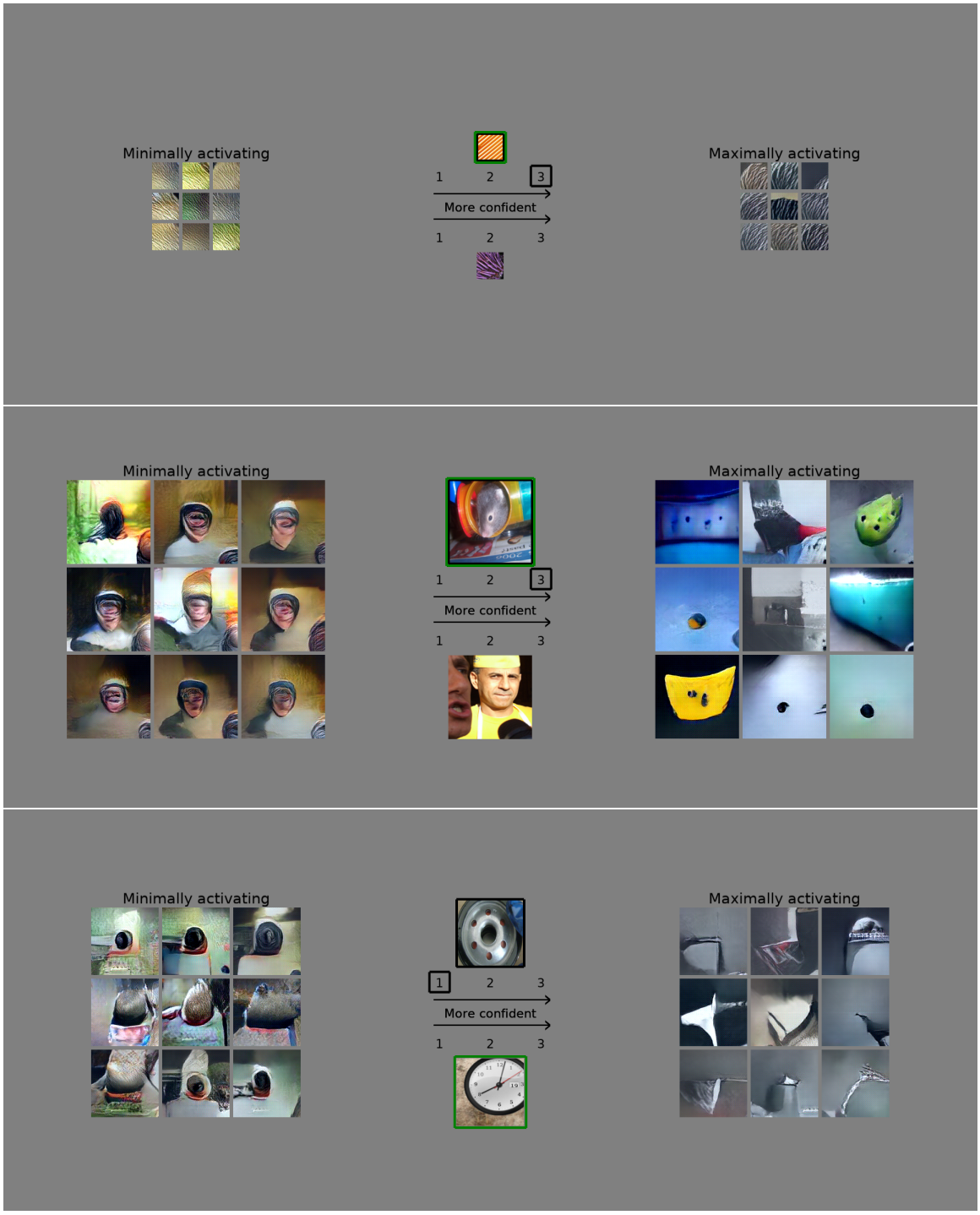


Figure 14: Example trials of the synthetic condition that seem more difficult.

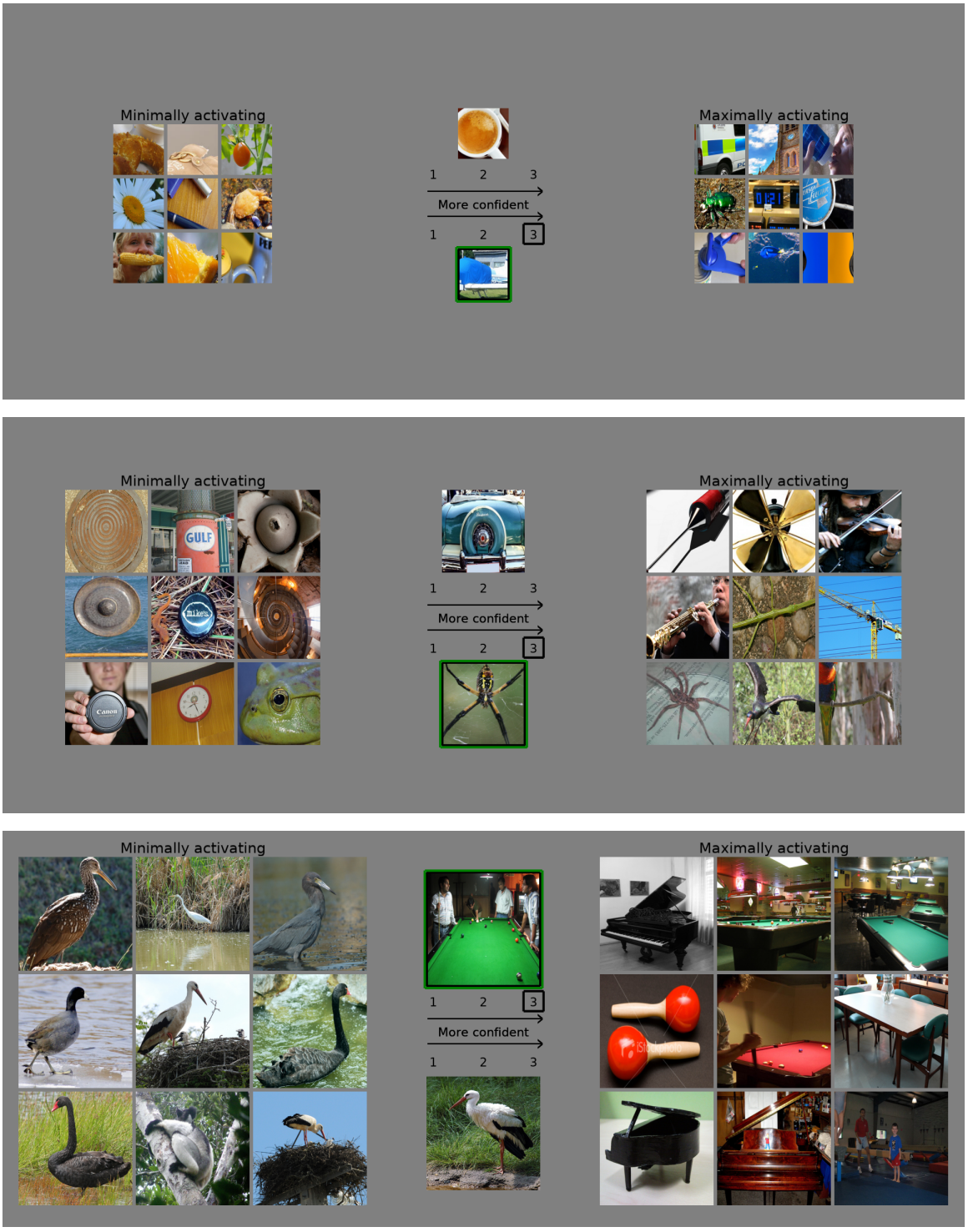


Figure 15: Example trials of the natural condition that seem fairly easy.

Figure 16: Example trials of the natural condition that seem more difficult.

Appendix B: Publications

* indicates joint first authorship, and ‡ indicates joint senior authorship.

The contribution statements are copied verbatim from the original publications.

Publication 1: Five points to check when comparing visual perception in humans and machines

Christina Maria Funke*, Judy Borowski*, Karolina Stosio, Wieland Brendel‡, Thomas S.A. Wallis‡, Matthias Bethge‡. Journal of Vision, 2021.

Contributions:

"The closed contour case study was designed by CMF, JB, TSAW and MB and later with WB. The code for the stimuli generation was developed by CMF. The neural networks were trained by CMF and JB. The psychophysical experiments were performed and analysed by CMF, TSAW and JB. The SVRT case study was conducted by CMF under supervision of TSAW, WB and MB. KS designed and implemented the recognition gap case study under the supervision of WB and MB, JB extended and refined it under the supervision of WB and MB. The initial idea to unite the three projects was conceived by WB, MB, TSAW and CMF, and further developed including JB. The first draft was jointly written by JB and CMF with input from TSAW and WB. All authors contributed to the final version and provided critical revisions."

Earlier versions of this work were presented at the following venues:

- as a poster at the Vision Sciences Society Conference (2018) under the title "Comparing the ability of humans and DNNs to recognise closed contours in cluttered images",
- as a poster at Conference on Cognitive Computational Neuroscience (2019) under the title "The Notorious Difficulty of Comparing Human and Machine Perception", and
- as a poster, which won the best paper award, at the NeurIPS Workshop *Shared Visual Representations in Human and Machine Intelligence* (2019) under the title "The Notorious Difficulty of Comparing Human and Machine Perception".

What is more, this work was featured in the following online articles:

- Challenges of Comparing Human and Machine Perception on The Gradient,
- Same or Different? The Question Flummoxes Neural Networks. on Quantamagazine,
- Why It's Notoriously Difficult to Compare AI and Human Perception on The New Stack,
- Computer vision: Why it's hard to compare AI and human perception on TechTalks, and
- AI vs. Human: A Comparison of Human Perception with Artificial Intelligence (AI) on ThinkML.

Five points to check when comparing visual perception in humans and machines

Christina M. Funke[*] University of Tübingen, Tübingen, Germany

Judy Borowski[*] University of Tübingen, Tübingen, Germany

Karolina Stosio University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen and Berlin, Germany
Volkswagen Group Machine Learning Research Lab, Munich, Germany

Wieland Brendel[†] University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen and Berlin, Germany
Werner Reichardt Centre for Integrative Neuroscience, Tübingen, Germany

Thomas S. A. Wallis[†] University of Tübingen, Tübingen, Germany
Present address: Amazon.com, Tübingen

Matthias Bethge[†] University of Tübingen, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen and Berlin, Germany
Werner Reichardt Centre for Integrative Neuroscience, Tübingen, Germany

With the rise of machines to human-level performance in complex recognition tasks, a growing amount of work is directed toward comparing information processing in humans and machines. These studies are an exciting chance to learn about one system by studying the other. Here, we propose ideas on how to design, conduct, and interpret experiments such that they adequately support the investigation of mechanisms when comparing human and machine perception. We demonstrate and apply these ideas through three case studies. The first case study shows how human bias can affect the interpretation of results and that several analytic tools can help to overcome this human reference point. In the second case study, we highlight the difference between necessary and sufficient mechanisms in visual reasoning tasks. Thereby, we show that contrary to previous suggestions, feedback mechanisms might not be necessary for the tasks in question. The third case study highlights the importance of aligning experimental conditions. We find that a previously observed difference in object recognition does not hold when adapting the experiment to make conditions more equitable between humans and machines. In presenting a checklist for comparative studies of visual reasoning in humans and machines, we hope to highlight how to overcome potential pitfalls in design and inference.

Introduction

Until recently, only biological systems could abstract the visual information in our world and transform it into a representation that supports understanding and action. Researchers have been studying how to implement such transformations in artificial systems since at least the 1950s. One advantage of artificial systems for understanding these computations is that many analyses can be performed that would not be possible in biological systems. For example, key

Citation: Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S. A., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3):16, 1–23, https://doi.org/10.1167/jov.21.3.16.

https://doi.org/10.1167/jov.21.3.16 Received April 21, 2020; published March 16, 2021 ISSN 1534-7362

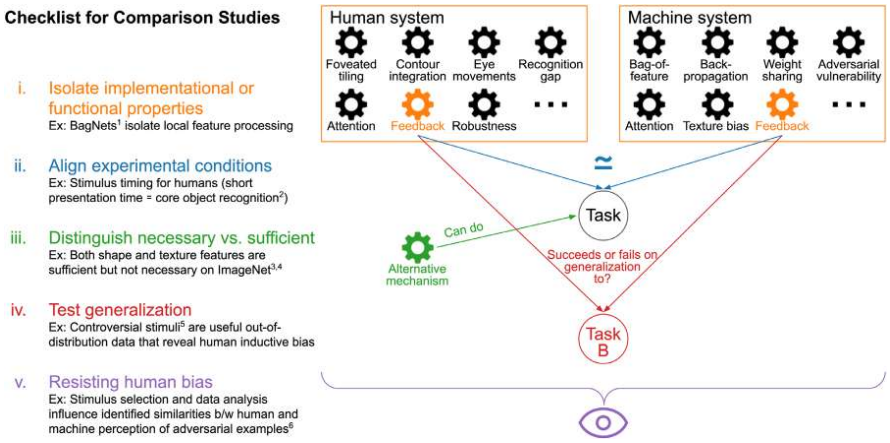


Figure 1. i: The human system and a candidate machine system differ in a range of properties. Isolating a specific mechanism (for example, feedback) can be challenging. ii: When designing an experiment, equivalent settings are important. iii: Even if a specific mechanism was important for a task, it would not be clear if this mechanism is necessary, as there could be other mechanisms (that might or might not be part of the human or machine system) that can allow a system to perform well. iv: Furthermore, the identified mechanisms might depend on the specific experimental setting and not generalize to, for example, another task. v: Overall, our human bias influences how we conduct and interpret our experiments. [1]Brendel and Bethge (2019); [2]DiCarlo et al. (2012); [3]Geirhos, Rubisch, et al. (2018); [4]Kubilius et al. (2016); [5]Golan et al. (2019); [6]Dujmović et al. (2020).

components of visual processing, such as the role of feedback connections, can be investigated, and methods such as ablation studies gain new precision.

Traditional models of visual processing sought to explicitly replicate the hypothesized computations performed in biological visual systems. One famous example is the hierarchical HMAX-model (Fukushima, 1980; Riesenhuber & Poggio, 1999). It instantiates mechanisms hypothesized to occur in primate visual systems, such as template matching and max operations, whose goal is to achieve invariance to position, scale, and translation. Crucially, though, these models never got close to human performance in real-world tasks.

With the success of learned approaches in the past decade, and particularly that of convolutional deep neural networks (DNNs), we now have much more powerful models. In fact, these models are able to perform a range of constrained image understanding tasks with human-like performance (Krizhevsky et al., 2012; Eigen & Fergus, 2015; Long et al., 2015).

While matching machine performance with that of the human visual system is a crucial step, the inner workings of the two systems can still be very different. We hence need to move beyond comparing accuracies to understand how the systems' mechanisms differ (Geirhos et al., 2020; Chollet, 2019; Ma & Peters, 2020; Firestone, 2020).

The range of frequently considered mechanisms is broad. They not only concern the architectural level (such as feedback vs. feed-forward connections, lateral connections, foveated architectures or eye movements, …), but also involve different learning schemes (back-propagation vs. spike-timing-dependent plasticity/Hebbian learning, …) as well as the nature of the representations themselves (such as reliance on texture rather than shape, global vs. local processing, …). For an overview of comparison studies, please see Appendix A.

Checklist for psychophysical comparison studies

We present a checklist on how to design, conduct, and interpret experiments of comparison studies that investigate relevant mechanisms for visual perception. The diagram in Figure 1 illustrates the core ideas that we elaborate on below.

i. **Isolating implementational or functional properties.** Naturally, the systems that are being compared often differ in more than just one aspect, and hence pinpointing one single reason for an observed difference can be challenging. One approach is to design an artificial network constrained such that the mechanism of interest will show its effect as clearly as possible. An example of such an attempt

is Brendel and Bethge (2019), which constrained models to process purely local information by reducing their receptive field sizes. Unfortunately, in many cases, it is almost impossible to exclude potential side effects from other experimental factors such as architecture or training procedure. Therefore, making explicit if, how, and where results depend on other experimental factors is important.

ii. **Aligning experimental conditions for both systems.** In comparative studies (whether humans and machines, or different organisms in nature), it can be exceedingly challenging to make experimental conditions equivalent. When comparing the two systems, any differences should be made as explicit as possible and taken into account in the design and analysis of the study. For example, the human brain profits from lifelong experience, whereas a machine algorithm is usually limited to learning from specific stimuli of a particular task and setting. Another example is the stimulus timing used in psychophysical experiments, for which there is no direct equivalent in stateless algorithms. Comparisons of human and machine accuracies must therefore be considered with the temporal presentation characteristics of the experiment. These characteristics could be chosen based on, for example, a definition of the behavior of interest as that occurring within a certain time after stimulus onset (as for, e.g., "core object recognition"; DiCarlo et al., 2012). Firestone (2020) highlights that as aligning systems perfectly may not be possible due to different "hardware" constraints such as memory capacity, unequal performance of two systems might still arise despite similar competencies.

iii. **Differentiating between necessary and sufficient mechanisms.** It is possible that multiple mechanisms allow good task performance — for example, DNNs can use either shape or texture features to reach high performance on ImageNet (Geirhos, Rubisch, et al., 2018; Kubilius et al., 2016). Thus, observing good performance for one mechanism does not imply that this mechanism is strictly necessary or that it is employed by the human visual system. As another example, Watanabe et al. (2018) investigated whether the rotating snakes illusion (Kitaoka & Ashida, 2003; Conway et al., 2005) could be replicated in artificial neural networks. While they found that this was indeed the case, we argue that the mechanisms must be different from the ones used by humans, as the illusion requires small eye movements or blinks (Hisakata & Murakami, 2008; Kuriki et al., 2008), while the artificial model does not emulate such biological processes.

iv. **Testing generalization of mechanisms.** Having identified an important mechanism, one needs to make explicit for which particular conditions (class of tasks, data sets, …) the conclusion is intended to hold. A mechanism that is important for one setup may or may not be important for another one. In other words, whether a mechanism works under generalized settings has to be explicitly tested. An example of outstanding generalization for humans is their visual *robustness* against various variations in the input. In DNNs, a mechanism to improve robustness is to "stylize" (Gatys et al., 2016) training data. First presented as raising performance on parametrically distorted images (Geirhos, Rubisch, et al., 2018), this mechanism was later shown to also improve performance on images suffering from common corruptions (Michaelis et al., 2019) but would be unlikely to help with adversarial robustness. From a different perspective, the work of Golan et al. (2019) on controversial stimuli is an example where using stimuli outside of the training distribution can be insightful. Controversial stimuli are synthetic images that are designed to trigger distinct responses for two machine models. In their experimental setup, the use of these out-of-distribution data allows the authors to reveal whether the inductive bias of humans is similar to one of the candidate models.

v. **Resisting human bias.** Human bias can affect not only the design but also the conclusions we draw from comparison experiments. In other words, our human reference point can influence, for example, how we interpret the behavior of other systems, be they biological or artificial. An example is the well-known Braitenberg vehicles (Braitenberg, 1986), which are defined by very simple rules. To a human observer, however, the vehicles' behavior appears as arising from complex internal states such as fear, aggression, or love. This phenomenon of anthropomorphizing is well known in the field of comparative psychology (Romanes, 1883; Köhler, 1925; Koehler, 1943; Haun et al., 2010; Boesch, 2007; Tomasello & Call, 2008). Buckner (2019) specifically warns of human-centered interpretations and recommends to apply the lessons learned in comparative psychology to comparing DNNs and humans. In addition, our human reference point can influence how we design an experiment. As an example, Dujmović et al. (2020) illustrate that the selection of stimuli and labels can have a big effect on finding similarities or differences between humans and machines to adversarial examples.

In the remainder of this article, we provide concrete examples of the aspects discussed above using three case studies[1]:

(1) **Closed contour detection**: The first case study illustrates how tricky overcoming our human bias

can be and that shedding light on an alternative decision-making mechanism may require multiple additional experiments.

(2) **Synthetic Visual Reasoning Test**: The second case study highlights the challenge of isolating mechanisms and of differentiating between necessary and sufficient mechanisms. Thereby, we discuss how human and machine model learning differ and how changes in the model architecture can affect the performance.

(3) **Recognition gap**: The third case study illustrates the importance of aligning experimental conditions.

Case study 1: Closed contour detection

Closed contours play a special role in human visual perception. According to the Gestalt principles of prägnanz and good continuation, humans can group distinct visual elements together so that they appear as a "form" or "whole." As such, closed contours are thought to be prioritized by the human visual system and to be important in perceptual organization (Koffka, 2013; Elder & Zucker, 1993; Kovacs & Julesz, 1993; Tversky et al., 2004; Ringach & Shapley, 1996). Specifically, to tell if a line closes up to form a closed contour, humans are believed to implement a process called "contour integration" that relies at least partially on global information (Levi et al., 2007; Loffler et al., 2003; Mathes & Fahle, 2007). Even many flanking, open contours would hardly influence humans' robust closed contour detection abilities.

Our experiments

We hypothesize that, in contrast to humans, closed contour detection is difficult for DNNs. The reason is that this task would presumably require long-range contour integration, but DNNs are believed to process mainly local information (Geirhos, Rubisch, et al., 2018; Brendel & Bethge, 2019). Here, we test how well humans and neural networks can separate closed from open contours. To this end, we create a custom data set, test humans and DNNs on it, and investigate the decision-making process of the DNNs.

DNNs and humans reach high performance

We created a data set with two classes of images: The first class contained a closed contour; the second one did not. In order to make sure that the statistical properties of the two classes were similar, we included a main contour for both classes. While this contour line closed up for the first class, it remained open for the second class. This main contour consisted of 3–9 straight-line segments. In order to make the task more difficult, we added several flankers with either one or two line segments that each had a length of at least 32 pixels (Figure 2A). The size of the images was 256 × 256 pixels. All lines were black and the background was uniformly gray. Details on the stimulus generation can be found in Appendix B.

Humans identified the closed contour stimulus very reliably in a two-interval forced-choice task. Their performance was 88.39% ($SEM = 2.96\%$) on stimuli whose generation procedure was identical to the training set. For stimuli with white instead of black lines, human participants reached a performance of 90.52% ($SEM = 1.58\%$). The psychophysical experiment is described in Appendix B.

We fine-tuned a ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) on the closed contour data set. Similar to humans, it performed very well and reached an accuracy of 99.95% (see Figure 2A [i.i.d. to training]).

We found that both humans and our DNN reach high accuracy on the closed contour detection task. From a human-centered perspective, it is enticing to infer that the model had learned the concept of open and closed contours and possibly that it performs a similar contour integration-like process as humans. However, this would have been overhasty. To better understand the degree of similarity, we investigated how our model performs on variations of the data sets that were not used during the training procedure.

Generalization tests reveal differences

Humans are expected to have no difficulties if the number of flankers, the color, or the shape of lines would differ. We here test our model's robustness on such variants of the data set. If our model used similar decision-making processes as humans, it should be able to generalize well without any further training on the new images. This procedure is another perspective to shed light on whether our model really understood the concept of closedness or just picked up some statistical cues in the training data set.

We tested our model on 15 variants of the data set (out of distribution test sets) without fine-tuning on these variations. As shown in Figure 2A, B, our trained model generalized well to many but not all modified stimulus sets.

On the following variations, our model achieved high accuracy: Curvy contours (1, 3) were easily distinguishable for our model, as long as the diameter remained below 100 pixels. Also, adding a dashed, closed flanker (2) did not lower its performance. The classification ability of the model remained similarly high for the no-flankers (4) and the asymmetric

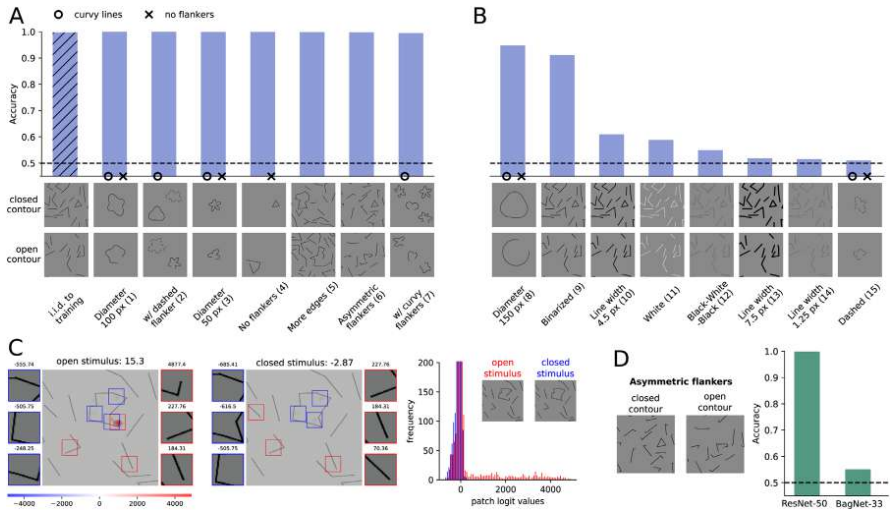


Figure 2. (A) Our ResNet-50-model generalized well to many data sets without further retraining, suggesting it would be able to distinguish closed and open contours. (B) However, the poor performance on many other data sets showed that our model did *not* learn the concept of closedness. (C) The heatmaps of our BagNet-33-based model show which parts of the image provided evidence for closedness (blue, negative values) or openness (red, positive values). The patches on the sides show the most extremely, nonoverlapping patches and their logit values. The logit distribution shows that most patches had logit values close to zero (y-axis truncated) and that many more patches in the open stimulus contributed positive logit values. (D) Our BagNet- and ResNet-models showed different performances on generalization sets, such as the asymmetric flankers. This indicates that the local decision-making process of the substitute model BagNet is not used by the original model ResNet. Figure best viewed electronically.

flankers condition (6). When testing our model on main contours that consisted of more edges than the ones presented during training (5), the performance was also hardly impaired. It remained high as well when multiple curvy open contours were added as flankers (7).

The following variations were more difficult for our model: If the size of the contour got too large, a moderate drop in accuracy was found (8). For binarized images, our model's performance was also reduced (9). And finally, (almost) chance performance was observed when varying the line width (14, 10, 13), changing the line color (11, 12), or using dashed curvy lines (15).

While humans would perform well on all variants of the closed contour data set, the failure of our model on some generalization tests suggests that it solves the task differently from humans. On the other hand, it is equally difficult to prove that the model does not understand the concept. As described by Firestone (2020), models can "perform differently despite similar underlying competences." In either way, we argue that it is important to openly consider alternative mechanisms to the human approach of global contour integration.

Our closed contour detection task is partly solvable with local features

In order to investigate an alternative mechanism to global contour integration, we here design an experiment to understand how well a decision-making process based on purely local features can work. For this purpose, we trained and tested BagNet-33 (Brendel & Bethge, 2019), a model that has access to local features only. It is a variation of ResNet-50 (He et al., 2016), where most 3×3 kernels are replaced by 1×1 kernels and therefore the receptive field size at the top-most convolutional layer is restricted to 33×33 pixels.

We found that our restricted model still reached close to 90% performance. In other words, contour integration was not necessary to perform well on the task.

To understand which local features the model relied on mostly, we analyzed the contribution of each patch to the final classification decision. To this end, we used the log-likelihood values for each 33×33 pixels patch from BagNet-33 and visualized them as a *heatmap*. Such a straightforward interpretation of the

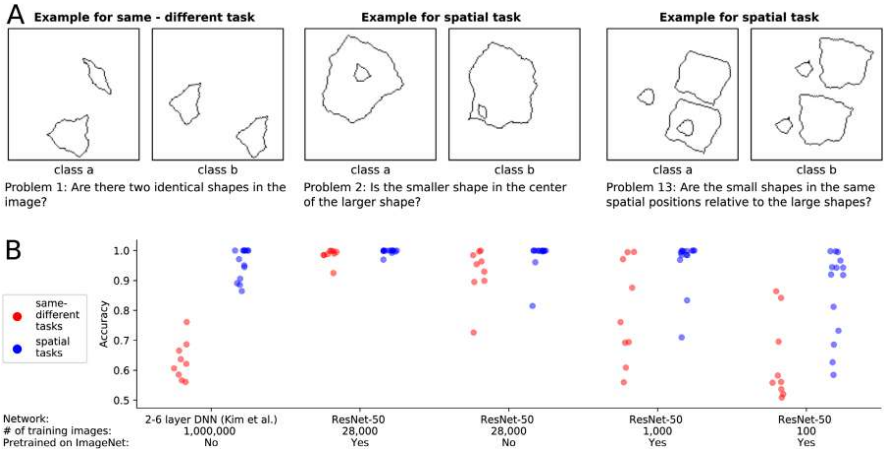


Figure 3. (A) For three of the 23 SVRT problems, two example images representing the two opposing classes are shown. In each problem, the task was to find the rule that separated the images and to sort them accordingly. (B) Kim et al. (2018) trained a DNN on each of the problems. They found that same-different tasks (red points), in contrast to spatial tasks (blue points), could not be solved with their models. Our ResNet-50-based models reached high accuracies for all problems when using 28,000 training examples and weights from pretraining on ImageNet.

contributions of single image patches is not possible with standard DNNs like ResNet (He et al., 2016) due to their large receptive field sizes in top layers.

The heatmaps of BagNet-33 (see Figure 2C) revealed which local patches played an important role in the decision-making process: An open contour was often detected by the presence of an endpoint at a short edge. Since all flankers in the training set had edges larger than 33 pixels, the presence of this feature was an indicator of an open contour. In turn, the absence of this feature was an indicator of a closed contour.

Whether the ResNet-50-based model used the same local feature as the substitute model was unclear. To answer this question, we tested BagNet on the previously mentioned generalization tests. We found that the data sets on which it showed high performance were sometimes different from the ones of ResNet (see Figure 7 in the Appendix B). A striking example was the failure of BagNet on the ”asymmetric flankers” condition (see Figure 2D). For these images, the flankers often consisted of shorter line segments and thus obscured the local feature we assumed BagNet to use. In contrast, ResNet performed well on this variation. This suggests that the decision-making strategy of ResNet did not heavily depend on the local feature found with the substitute BagNet model.

In summary, the generalization tests, the high performance of BagNet as well as the existence of a distinctive local feature provide evidence that our human-biased assumption was misleading. We saw that other mechanisms for closed contour detection besides global contour integration do exist (see Introduction, *“Differentiating between necessary and sufficient mechanisms”*). As humans, we can easily miss the many statistical subtleties by which a task can be solved. In this respect, BagNets proved to be a useful tool to test a purportedly “global” visual task for the presence of local artifacts. Overall, various experiments and analyses can be beneficial to understand mechanisms and to overcome our human reference point.

Case study 2: Synthetic Visual Reasoning Test

In order to compare human and machine performance at learning abstract relationships between shapes, Fleuret et al. (2011) created the Synthetic Visual Reasoning Test (SVRT) consisting of 23 problems (see Figure 3A). They showed that humans need only few examples to understand the underlying concepts. Stabinger et al. (2016) as well as Kim et al. (2018) assessed the performance of deep convolutional neural networks on these problems. Both studies found a dichotomy between two task categories: While high accuracy was reached on spatial problems, the performance on same-different problems was poor. In order to compare the two types of tasks more systematically, Kim et al. (2018) developed a parameterized version of the SVRT data set called

PSVRT. Using this data set, they found that for same-different problems, an increase in the complexity of the data set could quickly strain their models. In addition, they showed that an attentive version of the model did not exhibit the same deficits. From these results, the authors concluded that feedback mechanisms as present in the human visual system such as attention, working memory, or perceptual grouping are probably important components for abstract visual reasoning. More generally, these studies have been perceived and cited with the broader claim of feed-forward DNNs not being able to learn same-different relationships between visual objects (Serre, 2019; Schofield et al., 2018) – at least not "efficiently" (Firestone, 2020).

We argue that the results of Kim et al. (2018) cannot be taken as evidence for the importance of feedback components for abstract visual reasoning:

(1) While their experiments showed that same-different tasks are harder to *learn* for their models, this might also be true for the human visual system. Normally sighted humans have experienced lifelong visual input; only looking at human performance with this extensive learning experience cannot reveal differences in learning difficulty.
(2) Even if there is a difference in learning complexity, this difference is not necessarily due to differences in the inference mechanism (e.g., feed-forward vs. feedback)—the large variety of other differences between biological and artificial vision systems could be critical causal factors as well.
(3) In the same line, small modifications in the learning algorithm or architecture can significantly change learning complexity. For example, changing the network depth or width can greatly improve learning performance (Tan & Le, 2019).
(4) Just because an attentive version of the model can learn both types of tasks does not prove that feedback mechanisms are necessary for these tasks (see Introduction, *"Differentiating between necessary and sufficient mechanisms"*).

Determining the necessity of feedback mechanisms is especially difficult because feedback mechanisms are not clearly distinct from purely feed-forward mechanisms. In fact, any finite-time recurrent network can be unrolled into a feed-forward network (Liao & Poggio, 2016; van Bergen & Kriegeskorte, 2020).

For these reasons, we argue that the importance of feedback mechanisms for abstract visual reasoning remains unclear.

In the following paragraph we present our own experiments on the SVRT data set and show that standard feed-forward DNNs can indeed perform well on same-different tasks. This confirms that feedback mechanisms are not strictly necessary for same-different tasks, although they helped in the specific experimental setting of Kim et al. (2018). Furthermore, this experiment highlights that changes of the network architecture and training procedure can have large effects on the performance of artificial systems.

Our experiments

The findings of Kim et al. (2018) were based on rather small neural networks, which consisted of up to six layers. However, typical network architectures used for object recognition consist of more layers and have larger receptive fields. For this reason, we tested a representative of such networks, namely, ResNet-50. The experimental setup can be found in Appendix C.

We found that our feed-forward model can in fact perform well on the same-different tasks of SVRT (see Figure 3B; see also concurrent work of Messina et al., 2019). This result was not due to an increase in the number of training samples. In fact, we used fewer images (28,000 images) than Kim et al. (2018) (1 million images) and Messina et al. (2019) (400,000 images). Of course, the results were obtained on the SVRT data set and might not hold for other visual reasoning data sets (see Introduction, *"Testing generalization of mechanisms"*).

In the very low-data regime (1,000 samples), we found a difference between the two types of tasks. In particular, the overall performance on same-different tasks was lower than on spatial reasoning tasks. As for the previously mentioned studies, this cannot be taken as evidence for systematic differences between feed-forward neural networks and the human visual system. In contrast to the neural networks used in this experiment, the human visual system is naturally pretrained on large amounts of visual reasoning tasks, thus making the low-data regime an unfair testing scenario from which it is almost impossible to draw solid conclusions about differences in the internal information processing. In other words, it might very well be that the human visual system trained from scratch on the two types of tasks would exhibit a similar difference in sample efficiency as a ResNet-50. Furthermore, the performance of a network in the low-data regime is heavily influenced by many factors other than architecture, including regularization schemes or the optimizer, making it even more difficult to reach conclusions about systematic differences in the network structure between humans and machines.

Case study 3: Recognition gap

Ullman et al. (2016) investigated the minimally necessary visual information required for object recognition. To this end, they successively cropped or reduced the resolution of a natural image until more than 50% of all human participants failed to

identify the object. The study revealed that recognition performance drops sharply if the minimal recognizable image crops are reduced any further. They referred to this drop in performance as the "recognition gap." The gap is computed by subtracting the proportion of people who correctly classify the largest unrecognizable crop (e.g., 0.2) from that of the people who correctly classify the smallest recognizable crop (e.g., 0.9). In this example, the recognition gap would evaluate to $0.9 - 0.2 = 0.7$. On the same human-selected image crops, Ullman et al. (2016) found that the recognition gap is much smaller for machine vision algorithms (0.14 ± 0.24) than for humans (0.71 ± 0.05). The researchers concluded that machine vision algorithms would not be able to "explain [humans'] sensitivity to precise feature configurations" and "that the human visual system uses features and processes that are not used by current models and that are critical for recognition." In a follow-up study, Srivastava et al. (2019) identified "fragile recognition images" (FRIs) with an exhaustive machine-based procedure whose results include a subset of patches that adhere to the definition of minimal recognizable configurations (MIRCs) by Ullman et al. (2016). On these machine-selected FRIs, a DNN experienced a moderately high recognition gap, whereas humans experienced a low one. Because of the differences between the selection procedures used in Ullman et al. (2016) and Srivastava et al. (2019), the question remained open whether machines would show a high recognition gap on machine-selected minimal images, if the selection procedure was similar to the one used in Ullman et al. (2016).

Our experiment

Our goal was to investigate if the differences in recognition gaps identified by Ullman et al. (2016) would at least in part be explainable by differences in the experimental procedures for humans and machines. Crucially, we wanted to assess machine performance on *machine*-selected, and not *human*-selected, image crops. We therefore implemented the psychophysics experiment in a machine setting to search the smallest recognizable images (or MIRCs) and the largest unrecognizable images (sub-MIRCs). In the final step, we evaluated our machine model's recognition gap using the *machine*-selected MIRCs and sub-MIRCs.

Methods

Our machine-based search algorithm used the deep convolutional neural network BagNet-33 (Brendel & Bethge, 2019), which allows us to straightforwardly analyze images as small as 33×33 pixels. In the first step, the classification accuracy was evaluated for the whole image. If it was above 0.5, the image was successively cropped and reduced in resolution. In each step, the best-performing crop was taken as the new parent. When the classification probability of all children fell below 0.5, the parent was identified as the MIRC, and all its children were considered sub-MIRCs. In order to evaluate the recognition gap, we calculate the difference in accuracy between the MIRC and the *best-performing* sub-MIRC. This definition is more conservative than the one from Ullman et al. (2016), who evaluated the difference in accuracy between the MIRC and the *worst-performing* sub-MIRC. For more details on the search procedure, please see Appendix D.

Results

We evaluated the recognition gap on two data sets: the original images from Ullman et al. (2016) and a subset of the ImageNet validation images (Deng et al., 2009). As shown in Figure 4A, our model has an average recognition gap of 0.99 ± 0.01 on the machine-selected crops of the data set from Ullman et al. (2016). On the machine-selected crops of the ImageNet validation subset, a large recognition gap occurs as well. Our values are similar to the recognition gap in humans and differ from the machines' recognition gap (0.14 ± 0.24) between human-selected MIRCs and sub-MIRCs as identified by Ullman et al. (2016).

Discussion

Our findings contrast claims made by Ullman et al. (2016). The latter study concluded that machine algorithms are not as sensitive as humans to precise feature configurations and that they are missing features and processes that are "critical for recognition." First, our study shows that a machine algorithm *is* sensitive to small image crops. It is only the precise minimal features that differ between humans and machines. Second, by the word "critical," Ullman et al. (2016) imply that object recognition would not be possible without these human features and processes. Applying the same reasoning to Srivastava et al. (2019), the low human performance on machine-selected patches should suggest that humans would miss "features and processes critical for recognition." This would be an obviously overreaching conclusion. Furthermore, the success of modern artificial object recognition speaks against the conclusion that the purported processes are "critical" for recognition, at least within this discretely defined recognition task. Finally, what we can conclude from the experiments of Ullman et al. (2016) and from our own is that both the human and a machine visual system can recognize small image crops and that there is a sudden drop in recognizability when reducing the amount of information.

In summary, these results highlight the importance of testing humans and machines in as similar settings

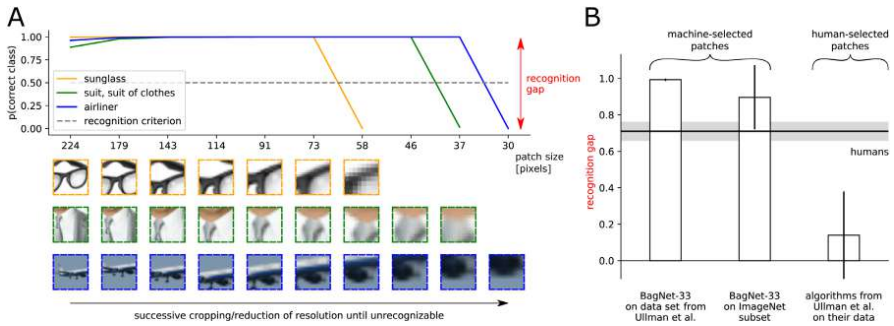


Figure 4. (A) BagNet-33's probability of correct class for decreasing crops: The sharp drop when the image becomes too small or the resolution too low is called the "recognition gap" (Ullman et al., 2016). It was computed by subtracting the model's predicted probability of the correct class for the sub-MIRC from the model's predicted probability of the correct class for the MIRC. As an example, the glasses stimulus was evaluated as $0.9999 - 0.0002 = 0.9997$. The crop size on the x-axis corresponds to the size of the original image in pixels. Steps of reduced resolution are not displayed such that the three sample stimuli can be displayed coherently. (B) Recognition gaps for machine algorithms (vertical bars) and humans (gray horizontal bar). A recognition gap is identifiable for the DNN BagNet-33 when testing machine-selected stimuli of the original images from Ullman et al. (2016) and a subset of the ImageNet validation images (Deng et al., 2009). Error bars denote standard deviation.

as possible, and of avoiding a human bias in the experiment design. All conditions, instructions, and procedures should be as close as possible between humans and machines in order to ensure that observed differences are due to inherently different decision strategies rather than differences in the testing procedure.

Conclusion

Comparing human and machine visual perception can be challenging. In this work, we presented a checklist on how to perform such comparison studies in a meaningful and robust way. For one, isolating a single mechanism requires us to minimize or exclude the effect of other differences between biological and artificial and to align experimental conditions for both systems. We further have to differentiate between necessary and sufficient mechanisms and to circumscribe in which tasks they are actually deployed. Finally, an overarching challenge in comparison studies between humans and machines is our strong internal human interpretation bias.

Using three case studies, we illustrated the application of the checklist. The first case study on closed contour detection showed that human bias can impede the objective interpretation of results and that investigating which mechanisms could or could not be at work may require several analytic tools. The second case study highlighted the difficulty of drawing robust conclusions about mechanisms from experiments. While previous studies suggested that feedback mechanisms might be important for visual reasoning tasks, our experiments showed that they are not necessarily required. The third case study clarified that aligning experimental conditions for both systems is essential. When adapting the experimental settings, we found that, unlike the differences reported in a previous study, DNNs and humans indeed show similar behavior on an object recognition task.

Our checklist complements other recent proposals about how to compare visual inference strategies between humans and machines (Buckner, 2019; Chollet, 2019; Ma & Peters, 2020; Geirhos et al., 2020) and helps to create more nuanced and robust insights into both systems.

Acknowledgments

The authors thank Alexander S. Ecker, Felix A. Wichmann, Matthias Kümmerer, Dylan Paiton, and Drew Linsley for helpful discussions. We thank Thomas Serre, Junkyung Kim, Matthew Ricci, Justus Piater, Sebastian Stabinger, Antonio Rodríguez-Sánchez, Shimon Ullman, Liav, Assif, and Daniel Harari for discussions and feedback on an earlier version of this manuscript. Additionally, we thank Nikolas Kriegeskorte for his detailed and constructive feedback, which helped us make our manuscript stronger. Furthermore, we thank Wiebke Ringels for helping with data collection for the psychophysical experiment.

We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for

supporting CMF and JB. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the competence center for machine learning (FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellence Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the Deutsche Forschungsgemeinschaft (DFG; Projektnummer 276693517 SFB 1233).

The closed contour case study was designed by CMF, JB, TSAW, and MB and later with WB. The code for the stimuli generation was developed by CMF. The neural networks were trained by CMF and JB. The psychophysical experiments were performed and analyzed by CMF, TSAW, and JB. The SVRT case study was conducted by CMF under supervision of TSAW, WB, and MB. KS designed and implemented the recognition gap case study under the supervision of WB and MB; JB extended and refined it under the supervision of WB and MB. The initial idea to unite the three projects was conceived by WB, MB, TSAW, and CMF, and further developed including JB. The first draft was jointly written by JB and CMF with input from TSAW and WB. All authors contributed to the final version and provided critical revisions.

Elements of this work were presented at the Conference on Cognitive Computational Neuroscience 2019 and the Shared Visual Representations in Human and Machine Intelligence Workshop at the Conference on Neural Information Processing Systems 2019.

The icon image is modified from the image by Gerd Leonhard, available under https://www.flickr.com/photos/gleonhard/33661762360 on December 17, 2020. The original license is CC BY-SA 2.0, and therefore so is the one for the icon image.

Commercial relationships: Matthias Bethge: Amazon scholar Jan 2019 – Jan 2021, Layer7AI, DeepArt.io, Upload AI; Wieland Brendel: Layer7AI.
Corresponding authors: Christina M. Funke; Judy Borowski.
Email: christina.funke@bethgelab.org; judy.borowski@bethgelab.org.
Address: Maria-von-Linden-Strasse 6, 72076, Tübingen, Germany.

*CMF and JB are both first authors on this work.
†WB, TSAW and MB are joint senior authors.

Footnote

[1]The code is available at https://github.com/bethgelab/notorious_difficulty_of_comparing_human_and_machine_perception.

References

Barrett, D. G., Hill, F., Santoro, A., Morcos, A. S., & Lillicrap, T. (2018). Measuring abstract reasoning in neural networks. In J. Dy, & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning, 80*, 511–520. PMLR, http://proceedings.mlr.press/v80/barrett18a.html.

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current Opinion in Neurobiology, 55*, 55–64.

Boesch, C. (2007). What makes us human (homo sapiens)? The challenge of cognitive cross-species comparison. *Journal of Comparative Psychology, 121*(3), 227.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Braitenberg, V. (1986). *Vehicles: Experiments in synthetic psychology*. MIT Press, https://mitpress.mit.edu/books/vehicles.

Brendel, W., & Bethge, M. (2019). Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint, arXiv:1904.00760.

Buckner, C. (2019). The comparative psychology of artificial intelligences, http://philsci-archive.pitt.edu/16128/.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., . . . Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Computational Biology, 15*(4), e1006897.

Chollet, F. (2019). The measure of intelligence. arXiv preprint, arXiv:1911.01547.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317, https://doi.org/10.1016/j.tics.2019.01.009.

Conway, B. R., Kitaoka, A., Yazdanbakhsh, A., Pack, C. C., & Livingstone, M. S. (2005). Neural basis for a powerful static motion illusion. *Journal of Neuroscience, 25*(23), 5651–5656.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2019). Crowding reveals fundamental

differences in local vs. global processing in humans and machines. bioRxiv, 744268.

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife, 9*, e55978. Retrieved from https://doi.org/10.7554/eLife.55978.

Eberhardt, S., Cader, J. G., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems, 29*, 1100–1108. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2016/file/42e77b63637ab381e8be5f8318cc28a2-Paper.pdf.

Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision* (pp. 2650–2658), doi:10.1109/ICCV.2015.304.

Elder, J., & Zucker, S. (1993). The effect of contour closure on the rapid discrimination of two-dimensional shapes. *Vision Research, 33*(7), 981–991.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., … Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31*, 3910–3920. Curran Associates, Inc.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*. Retrieved from https://www.pnas.org/content/early/2020/10/13/1905334117.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences, 108*(43), 17621–17625.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414–2423), doi:10.1109/CVPR.2016.265.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. arXiv preprint, arXiv:2006.16736.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint, arXiv:1811.12231.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, 31*, 7538–7550. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2019). Controversial stimuli: Pitting neural networks against each other as models of human recognition. arXiv preprint, arXiv:1911.09288.

Gomez-Villa, A., Martin, A., Vazquez-Corral, J., & Bertalmio, M. (2019). Convolutional neural networks can be deceived by visual illusions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12301–12309), doi:10.1109/CVPR.2019.01259.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning, 70*, 1321–1330.

Han, C., Yoon, W., Kwon, G., Nam, S., & Kim, D. (2019). Representation of white-and black-box adversarial examples in deep neural networks and humans: A functional magnetic resonance imaging study. arXiv preprint, arXiv:1905.02422.

Haun, D. B. M., Jordan, F. M., Vallortigara, G., & Clayton, N. S. (2010). Origins of spatial, temporal, and numerical cognition: Insights from comparative psychology. *Trends in Cognitive Sciences, 14*(12), 552–560, https://doi.org/10.1016/j.tics.2010.09.006, http://www.sciencedirect.com/science/article/pii/S1364661310002135.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778), doi:10.1109/CVPR.2016.90.

Hisakata, R., & Murakami, I. (2008). The effects of eccentricity and retinal illuminance on the illusory motion seen in a stationary luminance gradient. *Vision Research, 48*(19), 1940–1948.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models

may explain it cortical representation. *PLoS Computational Biology, 10*(11), e1003915.

Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do neural networks show gestalt phenomena? An exploration of the law of closure. arXiv preprint, arXiv:1903.01069.

Kim, J., Ricci, M., & Serre, T. (2018). Not-so-clevr: Learning same–different relations strains feedforward neural networks. *Interface Focus, 8*(4), 20180011.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint, arXiv:1412.6980.

Kitaoka, A., & Ashida, H. (2003). Phenomenal characteristics of the peripheral drift illusion. *Vision, 15*(4), 261–262.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C., (2007). What's new in psychtoolbox-3. *Perception, 36*(14), 1.

Koehler, O. (1943). Counting experiments on a common raven and comparative experiments on humans. *Zeitschrift für Tierpsychologie, 5*(3), 575–712.

Koffka, K. (2013). *Principles of Gestalt psychology*. New York: Routledge.

Köhler, W. (1925). *The mentality of apes*. New York, NY: Kegan Paul, Trench, Trubner & Co.

Kovacs, I., & Julesz, B. (1993). A closed curve is much more than an incomplete one: Effect of closure in figure-ground segmentation. *Proceedings of the National Academy of Sciences, 90*(16), 7495–7497.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems, 25*, 1097–1105. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology, 12*(4), e1004896.

Kuriki, I., Ashida, H., Murakami, I., & Kitaoka, A. (2008). Functional brain imaging of the rotating snakes illusion by fMRI. *Journal of Vision, 8*(10), 16, https://doi.org/10.1167/8.6.64.

Levi, D. M., Yu, C., Kuai, S.-G., & Rislove, E. (2007). Global contour processing in amblyopia. *Vision Research, 47*(4), 512–524.

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv preprint, arXiv:1604.03640.

Loffler, G., Wilson, H. R., & Wilkinson, F. (2003). Local and global contributions to shape discrimination. *Vision Research, 43*(5), 519–530.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431–3440), doi:10.1109/CVPR.2015.7298965.

Luo, L., Xiong, Y., Liu, Y., & Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint, arXiv:1902.09843.

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. arXiv preprint, arXiv:2005.02181.

Majaj, N. J., & Pelli, D. G. (2018). Deep learning—using machine learning to study biological vision. *Journal of Vision, 18*(13), 2, https://doi.org/10.1167/18.13.2.

Mathes, B., & Fahle, M. (2007). Closure facilitates contour integration. *Vision Research, 47*(6), 818–827, https://doi.org/10.1167/18.13.2.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint, arXiv:1902.01007.

Messina, N., Amato, G., Carrara, F., Falchi, F., & Gennaro, C. (2019). Testing deep neural networks on the same-different task. In *International Conference on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6). IEEE, doi:10.1109/CBMI.2019.8877412.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, M., ... Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint, arXiv:1907.07484.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38*(11), 39–41.

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. arXiv preprint, arXiv:1907.07355.

Pelli, D. G., & Vision, S. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. arXiv preprint, arXiv:1608.02164.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Ringach, D. L., & Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision Research, 36*(19), 3037–3050.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *Proceedings of the 34th International Conference on Machine Learning, 70*, 2940–2949.

Romanes, G. J. (1883). *Animal intelligence*. D. Appleton, https://books.google.de/books?hl=en&lr=&id=Vx8aAAAAYAAJ&oi=fnd&pg=PA1&dq=animal+intelligence+1883&ots=IUOqpa2YRA&sig=JIVJfeIN7HlireTKzBd2tdv8IzM&redir_esc=y#v=onepage&q=animal%20intelligence%201883&f=false.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., . . . Lillicrap, T. (2017). A simple neural network module for relational reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, . . . R. Garnett (Eds.), *Advances in Neural Information Processing Systems, 30*, 4967–4976. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2017/file/e6acf4b0f69f6f6e60e9a815938aa1ff-Paper.pdf.

Schofield, A. J., Gilchrist, I. D., Bloj, M., Leonardis, A., & Bellotto, N. (2018). Understanding images in biological and computer vision. *Interface Focus, 8*, https://doi.org/10.1098/rsfs.2018.0027

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . Schmidt, K. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? BioRxiv, 407007.

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science, 5*, 399–426.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology, 8*, 1551.

Srivastava, S., Ben-Yosef, G., & Boix, X. (2019). Minimal images in deep neural networks: Fragile object recognition in natural images. arXiv preprint, arXiv:1902.03227.

Stabinger, S., Rodríguez-Sánchez, A., & Piater, J. (2016). 25 years of CNNs: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks*. (pp. 380–387). Cham: Springer.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., . . . Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint, arXiv:1312.6199.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint, arXiv:1905.11946.

Tomasello, M., & Call, J. (2008). Assessing the validity of ape-human comparisons: A reply to boesch (2007). *Journal of Comparative Psychology, 122*(4), 449–452. American Psychological Association.

Tversky, T., Geisler, W. S., & Perry, J. S. (2004). Contour grouping: Closure effects are explained by good continuation and proximity. *Vision Research, 44*(24), 2769–2777.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences, 113*(10), 2744–2749.

van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: The role of recurrence in visual inference. arXiv preprint, arXiv:2003.12128.

Villalobos, K. M., Stih, V., Ahmadinejad, A., Dozier, J., Francl, A., Azevedo, F., Sasaki, T., . . . Boix, X. (2020). Do deep neural networks for segmentation understand insideness? https://cbmm.mit.edu/publications/do-neural-networks-segmentation-understand-insideness.

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do deep neural networks suffer from crowding? In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, . . . R. Garnett (Eds.), *Advances in Neural Information Processing Systems, 30*, 5628–5638. Curran Associates, Inc, https://proceedings.neurips.cc/paper/2017/file/c61f571dbd2fb949d3fe5ae1608dd48b-Paper.pdf.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology, 9*, 345.

Wu, X., Zhang, X., & Du, J. (2019). Challenge of spatial cognition for deep learning. arXiv preprint, arXiv:1908.04396.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624.

Yan, Z., & Zhou, X. S. (2017). How intelligent are convolutional neural networks? arXiv preprint, arXiv:1709.06126.

Zhang, R., Wu, J., Zhang, C., Freeman, W. T., & Tenenbaum, J. B. (2016). A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. arXiv preprint, arXiv:1605.01138.

Zhang, X., Watkins, Y., & Kenyon, G. T. (2018). Can deep learning learn the principle of closed

contour detection? In G. Bebis, R. Boyle, B. Parvin, D. Koracin, M. Turek, S. Ramalingam, K. Xu, S. Lin, B. Alsallakh, J. Yang, E. Cuervo, . . . J. Ventura (Eds.), *International Symposium on Visual Computing* (pp. 455–460). Cham: Springer, https://doi.org/10.1007/978-3-030-03801-4_40.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*(1), 1334.

Appendix A: Literature overview of comparison studies

A growing body of work discusses comparisons of humans and machines on a higher level. Majaj and Pelli (2018) provide a broad overview how machine learning can help vision scientists to study biological vision, while Barrett et al. (2019) review methods on how to analyze representations of biological and artificial networks. From the perspective of cognitive science, Cichy and Kaiser (2019) stress that deep learning models *can* serve as scientific models that not only provide both helpful predictions and explanations but that can also be used for exploration. Furthermore, from the perspective of psychology and philosophy, Buckner (2019) emphasizes often-neglected caveats when comparing humans and DNNs such as human-centered interpretations and calls for discussions regarding how to properly align machine and human performance. Chollet (2019) proposes a general artificial intelligence benchmark and suggests to rather evaluate intelligence as "skill-acquisition efficiency" than to focus on skills at specific tasks.

In the following, we give a brief overview of studies that compare human and machine perception. In order to test if DNNs have similar cognitive abilities as humans, a number of studies test DNNs on abstract (visual) reasoning tasks (Barrett et al., 2018; Yan & Zhou, 2017; Wu et al., 2019; Santoro et al., 2017; Villalobos et al., 2020). Other comparison studies focus on whether human visual phenomena such as illusions (Gomez-Villa et al., 2019; Watanabe et al., 2018; Kim et al., 2019) or crowding (Volokitin et al., 2017; Doerig et al., 2019) can be reproduced in computational models. In the attempt to probe intuition in machine models, DNNs are compared to intuitive physics engines, that is, probabilistic models that simulate physical events (Zhang et al., 2016).

Other works investigate whether DNNs are sensible models of human perceptual processing. To this end, their prediction or internal representations are compared to those of biological systems, for example, to human and/or monkey behavioral representations (Peterson et al., 2016; Schrimpf et al., 2018; Yamins et al., 2014; Eberhardt et al., 2016; Golan et al., 2019), human fMRI representations (Han et al., 2019; Khaligh-Razavi & Kriegeskorte, 2014) or monkey cell recordings (Schrimpf et al., 2018; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Cadena et al., 2019).

A great number of studies focus on manipulating tasks and/or models. Researchers often use generalization tests on data dissimilar to the training set (Zhang et al., 2018; Wu et al., 2019) to test whether machines understood the underlying concepts. In other studies, the degradation of object classification accuracy is measured with respect to image degradations (Geirhos et al., 2018) or with respect to the type of features that play an important role for human or machine decision-making (Geirhos, Rubisch, et al., 2018; Brendel & Bethge, 2019; Kubilius et al., 2016; Ullman et al., 2016; Ritter et al., 2017). A lot of effort is being put into investigating whether humans are vulnerable to small, adversarial perturbations in images (Elsayed et al., 2018; Zhou & Firestone, 2019; Han et al., 2019; Dujmović et al., 2020), as DNNs are shown to be (Szegedy et al., 2013). Similarly, in the field of natural language processing, a trend is to manipulate the data set itself by, for example, negating statements to test whether a trained model gains an understanding of natural language or whether it only picks up on statistical regularities (Niven & Kao, 2019; McCoy et al., 2019).

Further work takes inspiration from biology or uses human knowledge explicitly in order to improve DNNs. Spoerer et al. (2017) found that recurrent connections, which are abundant in biological systems, allow for higher object recognition performance, especially in challenging situations such as in the presence of occlusions—in contrast to pure feed-forward networks. Furthermore, several researchers suggest (Zhang et al., 2018; Kim et al., 2018) or show (Wu et al., 2019; Barrett et al., 2018; Santoro et al., 2017) that designing networks' architecture or features with human knowledge is key for machine algorithms to successfully solve abstract (reasoning) tasks.

Appendix B: Closed contour detection

Data set

Each image in the training set contained a main contour, multiple flankers, and a background image. The main contour and flankers were drawn into an image of size $1,028 \times 1,028$ pixels. The main contour and flankers could be straight or curvy lines, for which the generation processes are respectively described in the next two subsections. The lines had a default thickness of 10 pixels. We then resized the image to 256×256 pixels using anti-aliasing to transform the black and white pixels into smoother lines that had

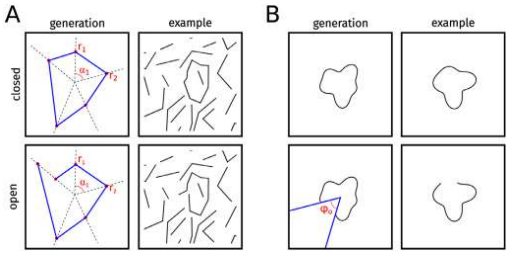


Figure 5. Closed contour data set. (A) Left: The main contour was generated by connecting points from a random sampling process of angles and radii. Right: Resulting line-drawing with flankers. (B) Left: Generation process of curvy contours. Right: Resulting line-drawing.

gray pixels at the borders. Thus, the lines in the resized image had a thickness of 2.5 pixels. In the following, all specifications of sizes refer to the resized image (i.e., a line described of final length 10 pixels extended over 40 pixels when drawn into the $1,028 \times 1,028$-pixel image). For the psychophysical experiments (see Appendix B, Psychophysical experiment), we added a white margin of 16 pixels on each side of the image to avoid illusory contours at the borders of the image.

Varying contrast of background. An image from the ImageNet data set was added as background to the line drawing. We converted the image into LAB color space and linearly rescaled the pixel intensities of the image to produce a normalized contrast value between 0 (gray image with the RGB values [118, 118, 118]) and 1 (original image) (see Figure 8A). When adding the image to the line drawing, we replaced all pixels of the line drawing by the values of the background image for which the background image had a higher grayscale value than the line drawing. For the experiments in the main body, the contrast of the background image was always 0. The other contrast levels were used only for the additional experiment described in Appendix B, Additional experiment: Increasing the task difficulty by adding a background image.

Generation of image pairs. We aimed to reduce the statistical properties that could be exploited to solve the task without judging the closedness of the contour. Therefore, we generated image pairs consisting of an "open" and a "closed" version of the same image. The two versions were designed to be almost identical and had the same flankers. They differed only in the main contour, which was either open or closed. Examples of such image pairs are shown in Figure 5. During training, either the closed or the open image of a pair was used. However, for the validation and testing, both versions were used. This allowed us to compare the predictions and heatmaps for images that differed only slightly but belonged to different classes.

Line-drawing with polygons as main contour

The data set used for training as well as some of the generalization sets consisted of straight lines. The main contour consisted of $n \in \{3, 4, 5, 6, 7, 8, 9\}$ line segments that formed either an open or a closed contour. The generation process of the main contour is depicted on the left side of Figure 5A. To get a contour with n edges, we generated n points, which were defined by a randomly sampled angle α_n and a randomly sampled radius r_n (between 0 and 128 pixels). By connecting the resulting points, we obtained the closed contour. We used the python PIL library (PIL 5.4.1, python3) to draw the lines that connect the endpoints. For the corresponding open contour, we sampled two radii for one of the angles such that they had a distance of 20 to 50 pixels from each other. When connecting the points, a gap was created between the points that share the same angle. This generation procedure could allow for very short lines with edges being very close to each other. To avoid this, we excluded all shapes with corner points closer to 10 pixels from nonadjacent lines.

The position of the main contour was random, but we ensured that the contour did not extend over the border of the image.

Besides the main contour, several flankers consisting of either one or two line segments were added to each stimulus. The exact number of flankers was uniformly sampled from the range [10,25]. The length of each line segment varied between 32 and 64 pixels. For the flankers consisting of two line segments, both lines had the same length, and the angle between the line segments was at least 45°. We added the flankers successively to the image and thereby ensured a minimal distance of 10 pixels between the line centers. To ensure that the corresponding image pairs would have the same flankers, the distances to both the closed and open version of the main contour were accounted for when re-sampling flankers. If a flanker did not fulfill this criterion, a new flanker was sampled of the same size and the same number of line segments, but it was placed somewhere else. If a flanker extended over the border of the image, the flanker was cropped.

Line-drawing with curvy lines as main contour

For some of the generalization sets, the contours consisted of curvy instead of straight lines. These were generated by modulating a circle of a given radius r_c with a radial frequency function that was defined by two sinusoidal functions. The radius of the contour was thus given by

$$r(\phi) = A_1 \sin(f_1(\phi + \theta_1)) + A_2 \sin(f_2(\phi + \theta_2)) + r_c, \quad (1)$$

with the frequencies f_1 and f_2 (integers between 1 and 6), amplitudes A_1 and A_2 (random values between 15

and 45), and phases θ_1 and θ_2 (between 0 and 2π). Unless stated otherwise, the diameter (diameter $= 2 \times r_c$) was a random value between 50 and 100 pixels, and the contour was positioned in the center of the image. The open contours were obtained by removing a circular segment of size $\phi_o = \frac{\pi}{3}$ at a random phase (see Figure 5B).

For two of the generalization data sets, we used dashed contours that were obtained by masking out 20 equally distributed circular segments each of size $\phi_d = \frac{\pi}{20}$.

Details on generalization data sets

We constructed 15 variants of the data set to test generalization performance. Nine variants consisted of contours with straight lines. Six of these featured varying line styles like changes in line width (10, 13, 14) and/or line color (11, 12). For one variant (5), we increased the number of edges in the main contour. Another variant (4) had no flankers, and yet another variant (6) featured asymmetric flankers. For variant 9, the lines were binarized (only black or gray pixels instead of different gray tones).

In another six variants, the contours as well as the flankers were curved, meaning that we modulated a circle with a radial frequency function. The first four variants did not contain any flankers and the main contour had a fixed size of 50 pixels (3), 100 pixels (1), and 150 pixels (8). For another variant (15), the contour was a dashed line. Finally, we tested the effect of different flankers by adding one additional closed, yet dashed contour (2) or one to four open contours (7).

Below, we provide more details on some of these data sets:

Black-white-black lines (12). For all contours, black lines enclosed a white one in the middle. Each of these three lines had a thickness of 1.5 pixels, which resulted in a total thickness of 4.5 pixels.

Asymmetric flankers (6). The two-line flankers consisted of one long and one short line instead of two equally long lines.

W/ dashed flanker (2). This data set with curvy contours contained an additional dashed, yet closed contour as a flanker. It was produced like the main contour in the dashed main contour set. To avoid overlap of the contours, the main contour and the flanker could only appear at four determined positions in the image, namely, the corners.

W/ multiple flankers (7). In addition to the curvy main contour, between one and four open curvy contours were added as flankers. The flankers were generated by the same process as the main contour. The circles that were modulated had a diameter of 50 pixels and could appear at either one of the four corners of the image or in the center.

Psychophysical experiment

To estimate how well humans would be able to distinguish closed and open stimuli, we performed a psychophysical experiment in which observers reported which of two sequentially presented images contained a closed contour (two-interval forced choice [2-IFC] task).

Stimuli

The images of the closed contour data set were used as stimuli for the psychophysical experiments. Specifically, we used the images from the test sets that were used to evaluate the performance of the models. For our psychophysical experiments, we used two different conditions: The images contained either black (i.i.d. to the training set) or white contour lines. The latter was one one of the generalization test sets.

Apparatus

Stimuli were displayed on a VIEWPixx 3D LCD (VPIXX Technologies; spatial resolution $1,920 \times 1,080$ pixels, temporal resolution 120 Hz, operating with the scanning backlight turned off). Outside the stimulus image, the monitor was set to mean gray. Observers viewed the display from 60 cm (maintained via a chinrest) in a darkened chamber. At this distance, pixels subtended approximately 0.024° on average (41 pixels per degree of visual angle). The monitor was linearized (maximum luminance 260 cd/m^2 using a Konica-Minolta LS-100 photometer. Stimulus presentation and data collection were controlled via a desktop computer (Intel Core i5-4460 CPU, AMD Radeon R9 380 GPU) running Ubuntu Linux (16.04 LTS), using the Psychtoolbox Library (Pelli & Vision, 1997; Kleiner et al., 2007; Brainard & Vision, 1997, version 3.0.12) and the iShow library (http://dx.doi.org/10.5281/zenodo.34217) under MATLAB (The Mathworks, Inc., R2015b).

Participants

In total, 19 naïve observers (4 male, 15 female, age: 25.05 years, $SD = 3.52$) participated in the experiment. Observers were paid 10€ per hour for participation. Before the experiment, all subjects had given written informed consent for participating. All subjects had normal or corrected-to-normal vision. All procedures conformed to Standard 8 of the American Psychological 405 Association's "Ethical Principles of Psychologists and Code of Conduct" (2010).

Procedure

On each trial, one closed and one open contour stimulus were presented to the observer (see Figure 6A).

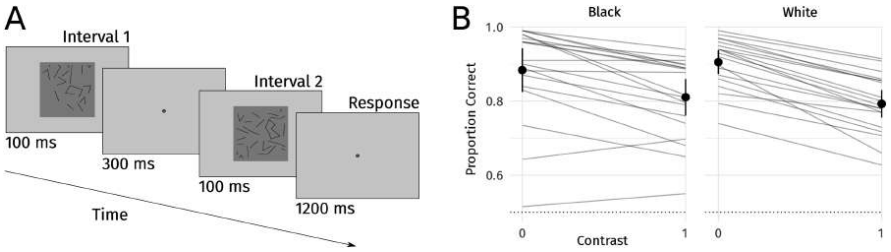


Figure 6. (A) In a 2-IFC task, human observers had to tell which of two images contained a closed contour. (B) Accuracy of the 20 naïve observers for the different conditions.

The images used for each trial were randomly picked, but we ensured that the open and closed images shown in the same trial were not the ones that were almost identical to each other (see Appendix B, Generation of image pairs). Thus, the number of edges of the main contour could differ between the two images shown in the same trial. Each image was shown for 100 ms, separated by a 300-ms interstimulus interval (blank gray screen). We instructed the observer to look at the fixation spot in the center of the screen. The observer was asked to identify whether the image containing a closed contour appeared first or second. The observer had 1,200 ms to respond and was given feedback after each trial. The intertrial interval was 1,000 ms. Each block consisted of 100 trials and observers performed five blocks. Trials with different line colors and varying background images (contrasts including 0, 0.4, and 1) were blocked. Here, we only report the results for black and white lines of contrast 0. Upon the first time that a block with a new line color was shown, observers performed a practice session with 48 trials of the corresponding line color.

Training of ResNet-50 model

We fine-tuned a ResNet-50 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) on the closed contour task. We replaced the last fully connected, 1,000-way classification layer by layer with only one output neuron to perform binary classification with a decision threshold of 0. The weights of all layers were fine-tuned using the optimizer Adam (Kingma & Ba, 2014) with a batch size of 64. All images were preprocessed to have the same mean and standard deviation and were randomly mirrored horizontally and vertically for data augmentation. The model was trained on 14,000 images for 10 epochs with a learning rate of 0.0003. We used a validation set of 5,600 images.

Generalization tests. To determine the generalization performance, we evaluated the model on the test sets without any further training. Each of the test sets contained 5,600 images. Poor accuracy could simply result from a suboptimal decision criterion rather than because the network would not be able to tell the stimuli apart. To account for the distribution shift between the original training images and the generalization tasks, we optimized the decision threshold (a single scalar) for each data set. To find the optimal threshold for each data set, we subdivided the interval, in which 95% of all logits lie, into 100 sub points and picked the threshold that would lead to the highest performance.

Training of BagNet-33 model

To test an alternative decision-making mechanism to global contour integration, we trained and tested a BagNet-33 (Brendel & Bethge, 2019) on the closed contour task. Like the ResNet-50 model, it was pretrained on ImageNet (Deng et al., 2009) and we replaced the last fully connected, 1,000-way classification layer by layer with only one output neuron. We fine-tuned the weights using the optimizer AdaBound (Luo et al., 2019) with an initial and final learning rate of 0.0001 and 0.1, respectively. The training images were generated on-the-fly, which meant that new images were produced for each epoch. In total, the fine-tuning lasted 100 epochs, and we picked the weights from the epoch with the highest performance.

Generalization tests. The generalization tests were conducted equivalently to the ones with ResNet-50. The results are shown in Figure 7.

Additional experiment: Increasing the task difficulty by adding a background image

We performed an additional experiment, where we tested if the model would become more robust and thus generalized better if we trained on a more difficult task. This was achieved by adding an image to the

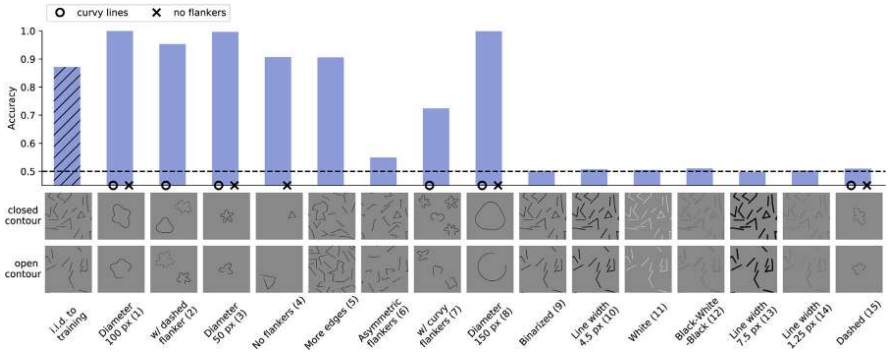


Figure 7. Generalization performances of BagNet-33.

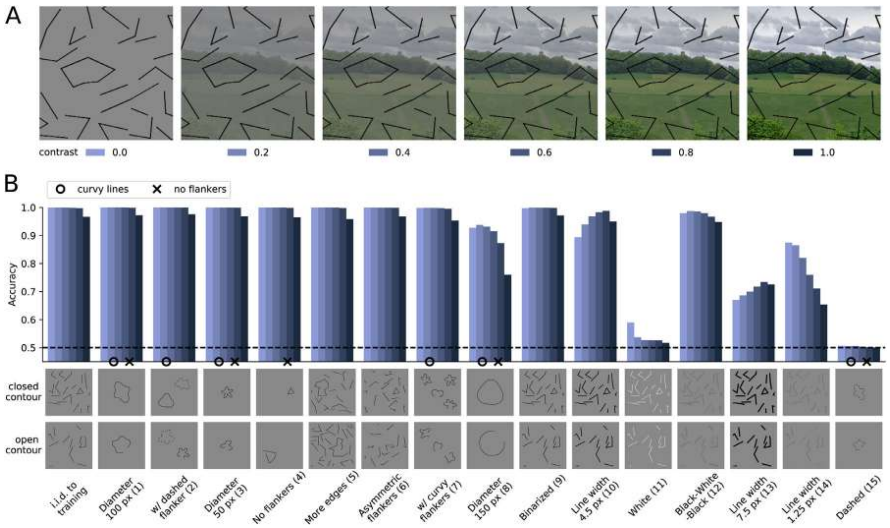


Figure 8. (A) An image of varying contrast was added as background. (B) Generalization performances of our models trained on random contrast levels and tested on single contrast levels.

background, such that the model had to learn how to separate the lines from the task-irrelevant background.

In our experiment, we fine-tuned our ResNet-50-based model on images with a background image of a uniformly sampled contrast. For each data set, we evaluated the model separately on six discrete contrast levels {0, 0.2, 0.4, 0.6, 0.8, 1} (see Figure 8A). We found that the generalization performance varied for some data sets compared to the experiment in the main body (see Figure 8B).

Appendix C: SVRT

Methods

Data set. We used the original C-code provided by Fleuret et al. (2011) to generate the images of the SVRT data set. The images had a size of 128 × 128 pixels. For each problem, we used up to 28,000 images for training,

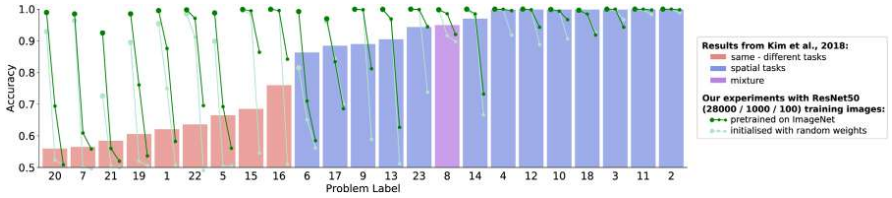


Figure 9. Accuracy of the models for the individual problems. Problem 8 is a mixture of same-different task and spatial task. In Figure 3, this problem was assigned to the spatial tasks. Bars replotted from Kim et al. (2018).

5,600 images for validation, and 11,200 images for testing.

Experimental procedures. For each of the SVRT problems, we fine-tuned a ResNet-50 that was pretrained on ImageNet (Deng et al., 2009) (as described in Appendix B, Training of ResNet-50 model). The same preprocessing, data augmentation, optimizer, and batch size as for the closed contour task were used.

For the different experiments, we varied the number of training images. We used subsets containing 28,000, 1,000, or 100 images. The number of epochs depended on the size of the training set: The model was fine-tuned for respectively 10, 280, or 2800 epochs. For each training set size and SVRT problem, we used the best learning rate after a hyper-parameter search on the validation set, where we tested the learning rates $[6 \times 10^{-5}, 1 \times 10^{-4}, 3 \times 10^{-4}]$.

As a control experiment, we also initialized the model with random weights, and we again performed a hyper-parameter search over the learning rates $[3 \times 10^{-4}, 6 \times 10^{-4}, 1 \times 10^{-3}]$.

Results

In Figure 9, we show the results for the individual problems. When using 28,000 training images, we reached above 90% accuracy for all SVRT problems, including the ones that required same-different judgments (see also Figure 3B). When using less training images, the performance on the test set was reduced. In particular, we found that the performance on same-different tasks dropped more rapidly than on spatial reasoning tasks. If the ResNet-50 was trained from scratch (i.e., weights were randomly initialized instead of loaded from pretraining on ImageNet), the performance dropped only slightly on all but one spatial reasoning task. Larger drops were found on same-different tasks.

Appendix D: Recognition gap

Details on methods

Data set. We used two data sets for this experiment. One consisted of 10 natural, color images whose grayscale versions were also used in the original study by Ullman et al. (2016). We discarded one image from the original data set as it does not correspond to any ImageNet class. For our ground truth class selection, please see Table 1. The second data set consisted of 1,000 images from the ImageNet (Deng et al., 2009) validation set. All images were preprocessed like in standard training of ResNet (i.e., resizing to 256 × 256 pixels, cropping centrally to 224 × 224 pixels and normalizing).

Model. In order to evaluate the recognition gap, the model had to be able to handle small input images. Standard networks like ResNet (He et al., 2016) are not equipped to handle small images. In contrast, BagNet-33 (Brendel & Bethge, 2019) allows us to straightforwardly analyze images as small as 33 × 33 pixels and hence was our model of choice for this experiment. It is a variation of ResNet-50 (He et al., 2016), where most 3 × 3 kernels are replaced by 1 × 1 kernels such that the receptive field size at the top-most convolutional layer is restricted to 33 × 33 pixels.

Machine-based search procedure for minimal recognizable images. Similar to Ullman et al. (2016), we defined minimal recognizable images or configurations (MIRCs) as those patches of an image for which an observer—by which we mean an ensemble of humans or one or several machine algorithms—reaches $\geq 50\%$ accuracy, but any additional 20% cropping of the corners or 20% reduction in resolution would lead to an accuracy $< 50\%$. MIRCs are thus inherently observer-dependent. The original study only searched for MIRCs in humans. We implemented the following procedure to find MIRCs in our DNN: We passed each preprocessed image through BagNet-33 and selected the most predictive crop according to its

Image	WordNet Hierarchy ID	WordNet Hierarchy description	Neuron number in ResNet-50 (indexing starts at 0)
fly	n02190166	fly	308
ship	n02687172	aircraft carrier, carrier, flattop, attack aircraft carrier	403
	n03095699	container ship, containership, container vessel	510
	n03344393	fireboat	554
	n03662601	lifeboat	625
	n03673027	liner, ocean liner	628
eagle	n01608432	kite	21
	n01614925	bald eagle, American eagle, Haliaeetus leucocephalus	22
glasses	n04355933	sunglass	836
	n04356056	sunglasses, dark glasses, shades	837
bike	n02835271	bicycle-built-for-two, tandem bicycle, tandem	444
	n03599486	jinrikisha, ricksha, rickshaw	612
	n03785016	moped	665
	n03792782	mountain bike, all-terrain bike, off-roader	671
	n04482393	tricycle, trike, velocipede	870
suit	n04350905	suit, suit of clothes	834
	n04591157	windsor tie	906
plane	n02690373	airliner	404
horse	n02389026	sorrel	339
	n03538406	horse cart, horse-cart	603
car	n02701002	ambulance	407
	n02814533	beach wagon, station wagon, wagon estate car, beach waggon, station waggon, waggon	436
	n02930766	cab, hack, taxi, taxicab	468
	n03100240	convertible	511
	n03594945	jeep, landrover	609
	n03670208	limousine, limo	627
	n03769881	minibus	654
	n03770679	minivan	656
	n04037443	racer, race car, racing car	751
	n04285008	sports car, sport car	817

Table 1. Selection of ImageNet classes for stimuli of Ullman et al. (2016).

probability. See Appendix D, Selecting best crop when probabilities saturate on how to handle cases where the probability saturates at 100% and Appendix D, Analysis of different class selections and different number of descendants for different treatments of ground truth class selections. If this probability of the full-size image for the ground-truth class was $\geq$ 50%, we again searched for the 80% subpatch with the highest probability. We repeated the search procedure until the class probability for all subpatches fell below 50%. If the 80% subpatches would be smaller than 33×33 pixels, which is BagNet-33's smallest natural patch size, the crop was increased to 33×33 pixels using bilinear sampling. We evaluated the recognition gap as the difference in accuracy between the MIRC and the *best-performing* sub-MIRC. This definition was more conservative than the one from Ullman et al. (2016), who considered the maximum difference between a MIRC and its sub-MIRCs, that is, the difference between the MIRC and the *worst-performing* sub-MIRC. Please note that one difference between our machine procedure and the psychophysics experiment by Ullman et al. (2016) remained: The former was greedy, whereas the latter corresponded to an exhaustive search under certain assumptions.

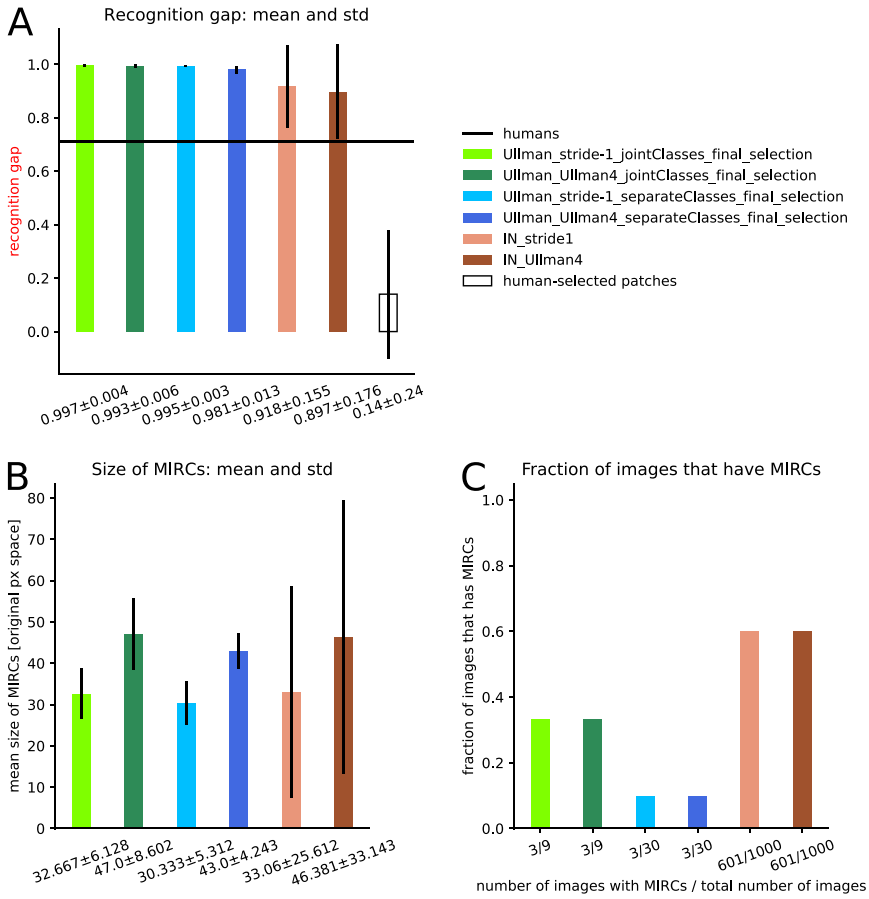


Figure 10. (A) Recognition gaps. The legend holds for all subplots. (B) Size of MIRCs. (C) Fraction of images that have MIRCS.

Analysis of different class selections and different number of descendants

Treating the 10 stimuli from Ullman et al. (2016) in our machine algorithm setting required two design choices: We needed to both pick suitable ground truth classes from ImageNet for each stimulus as well as choose if and how to combine them. The former is subjective, and using relationships from WordNet Hierarchy (Miller, 1995) (as Ullman et al. [2016] did in their psychophysics experiment) only provides limited guidance. We picked classes to our best judgment (for our final ground truth class choices, please see Table 1). Regarding the aspect of handling several ground truth classes, we extended our experiments: We tested whether considering all classes as one ("joint classes," i.e., summing the probabilities) or separately ("separate classes," i.e., rerunning the stimuli for each ground truth class) would have an effect on the recognition gap. As another check, we investigated whether the number of descendant options would alter the recognition gap: Instead of only considering the four corner crops as in the psychophysics experiment by Ullman et al. (2016) ("Ullman4"), we looked at every crop shifted by 1 pixel as a potential new parent ("stride-1"). The results reported in the main body correspond to joint classes and corner crops. Finally, besides analyzing the recognition gap, we also analyzed the sizes of MIRCs and the fractions of images that possess MIRCs for the mentioned conditions.

Figure 10A shows that all options result in similar values for the recognition gap. The trend of smaller

MIRC sizes for stride-1 compared to four corner crops shows that the search algorithm can find even smaller MIRCs when all crops are possible descendants (see. Figure 10B). The final analysis of how many images possess MIRCs (see Figure 10C) shows that recognition gaps only exist for fractions of the tested images: In the case of the stimuli from Ullman et al. (2016), three out of nine images, and in the case of ImageNet, about 60% of the images have MIRCs. This means that the recognition performance of the initial full-size configurations was $\geq$ 50% for those fractions only. Please note that we did not evaluate the recognition gap over images that did not meet this criterion. In contrast, Ullman et al. (2016) average only across MIRCs that have a recognition rate above 65% and sub-MIRCs that have a recognition rate below 20% (personal communication, 2019). The reason why our model could only reliably classify three out of the nine stimuli from Ullman et al. (2016) can partly be traced back to the oversimplification of single-class attribution in ImageNet as well as to the overconfidence of deep learning classification algorithms (Guo et al., 2017): They often attribute a lot of evidence to one class, and the remaining ones only share very little evidence.

Selecting best crop when probabilities saturate

We observed that several crops had very high probabilities and therefore used the "logit" measure $logit(p)$, where p is the probability. It is defined as the following: $logit(p) = \log(\frac{p}{1-p})$. Note that this measure is different from what the deep learning community usually refers to as "logits," which are the values before the softmax layer. In the following, we denote the latter values as $\mathbf{z}$. The logit $logit(p)$ is monotonic w.r.t. to the probability p, meaning that the higher the probability p, the higher the logit $logit(p)$. However, while p saturates at 100%, $logit(p)$ is unbounded. Therefore, it yields a more sensitive discrimination measure between image patches j that all have $p(\mathbf{z}^j) = 1$, where the superscript j denotes different patches.

In the following, we will provide a short derivation for the logit $logit(p)$. Consider a single patch with the correct class c. We start with the probability p_c of class c, which can be obtained by plugging the logits z_i into the softmax formula, where i corresponds to the classes [0, …, 1,000].

$$p_c(\mathbf{z}) = \frac{exp(z_c)}{exp(z_c) + \sum_{i \neq c} exp(z_i)} \quad (2)$$

Since we are interested in the probability of the correct class, it holds that $p_c(\mathbf{z}) \neq 0$. Thus, in the regime of interest, we can invert both sides of the equation. After simplifying, we get

$$\frac{1}{p_c(\mathbf{z})} - 1 = \frac{\sum_{i \neq c} exp(z_i)}{exp(z_c)}. \quad (3)$$

When taking the negative logarithm on both sides, we obtain

$$\Leftrightarrow -log\left(\frac{1}{p_c(\mathbf{z})} - 1\right) = -log\left(\frac{\sum_{i \neq c} exp(z_i)}{exp(z_c)}\right) \quad (4)$$

$$\Leftrightarrow - log\left(\frac{1 - p_c(\mathbf{z})}{p_c(\mathbf{z})}\right) = -log\left(\sum_{i \neq c} exp(z_i)\right) - (-log(exp(z_c))) \quad (5)$$

$$\Leftrightarrow log\left(\frac{p_c(\mathbf{z})}{1 - p_c(\mathbf{z})}\right) = z_c - log\left(\sum_{i \neq c} exp(z_i)\right) \quad (6)$$

The left-hand side of the equation is exactly the definition of the logit $logit(p)$. Intuitively, it measures in log-space how much the network's belief in the correct class outweighs the belief in all other classes taken together. The following reassembling operations illustrate this:

$$\begin{aligned} logit(p_c) &= log\left(\frac{p_c(\mathbf{z})}{1 - p_c(\mathbf{z})}\right) \\ &= \underbrace{log(p_c(\mathbf{z}))}_{\text{log probability of correct class}} \\ &\quad - \underbrace{log(1 - p_c(\mathbf{z}))}_{\text{log probability of all incorrect classes}} \end{aligned} \quad (7)$$

The above formulations regarding one correct class hold when adjusting the experimental design to accept several classes k as correct predictions. In brief, the logit $logit(p_C(z))$, where C stands for several classes, then states

$$\begin{aligned} logit(p_C(\mathbf{z})) &= -log\left(\frac{1}{p_{c_1}(\mathbf{z}) + p_{c_2}(\mathbf{z}) + ... + p_{c_k}(\mathbf{z})} - 1\right) \\ &= -log\left(\frac{1}{\sum_k p_k(\mathbf{z})} - 1\right) \\ &= \underbrace{log\left(\sum_k p_k(\mathbf{z})\right)}_{\text{log probability of all correct classes}} \end{aligned}$$

$$- \underbrace{log\left(1 - \sum_k p_k(\mathbf{z})\right)}_{\text{log probability of all incorrect classes}}$$

$$= log\left(\sum_k exp(z_k)\right) - log\left(\sum_{i \neq k} exp(z_i)\right) \quad (8)$$

Selection of ImageNet classes for stimuli of Ullman et al. (2016)

Note that our selection of classes is different from the one used by Ullman et al. (2016). We went through all classes for each image and selected the ones that we considered sensible. The 10th image of the eye does not have a sensible ImageNet class; hence, only nine stimuli from Ullman et al. (2016) are listed in Table 1.

Publication 2: Exemplary natural images explain CNN activations better than state-of-the-art feature visualization

Judy Borowski*, Roland Simon Zimmermann*, Judith Schepers, Robert Geirhos, Thomas S.A. Wallis‡, Matthias Bethge‡, Wieland Brendel‡. ICLR, 2021.

Contributions:

"The initiative of investigating human predictability of CNN activations came from WB. JB, WB, MB and TSAW jointly combined it with the idea of investigating human interpretability of feature visualizations. JB led the project. JB, RSZ and JS jointly designed and implemented the experiments (with advice and feedback from TSAW, RG, MB and WB). The data analysis was performed by JB and RSZ (with advice and feedback from RG, TSAW, MB and WB). JB designed, and JB and JS implemented the pilot study. JB conducted the experiments (with help from JS). RSZ performed the statistical significance tests (with advice from TSAW and feedback from JB and RG). MB helped shape the bigger picture and initiated intuitiveness trials. WB provided day-to-day supervision. JB, RSZ and RG wrote the initial version of the manuscript. All authors contributed to the final version of the manuscript."

An earlier version of this work was presented as a poster at the NeurIPS Workshop *Shared Visual Representations in Human and Machine Intelligence* (2020) under the title "Natural Images Are More Informative for Interpreting CNN Activations than Synthetic Feature Visualizations."

EXEMPLARY NATURAL IMAGES EXPLAIN CNN ACTIVATIONS BETTER THAN STATE-OF-THE-ART FEATURE VISUALIZATION

Judy Borowski*, **Roland S. Zimmermann***, **Judith Schepers**, **Robert Geirhos**, **Thomas S. A. Wallis**†‡, **Matthias Bethge**‡, **Wieland Brendel**‡
University of Tübingen, Germany

ABSTRACT

Feature visualizations such as synthetic maximally activating images are a widely used explanation method to better understand the information processing of convolutional neural networks (CNNs). At the same time, there are concerns that these visualizations might not accurately represent CNNs' inner workings. Here, we measure how much extremely activating images help humans to predict CNN activations. Using a well-controlled psychophysical paradigm, we compare the informativeness of synthetic images by Olah et al. (2017) with a simple baseline visualization, namely exemplary natural images that also strongly activate a specific feature map. Given either synthetic or natural reference images, human participants choose which of two query images leads to strong positive activation. The experiment is designed to maximize participants' performance, and is the first to probe *intermediate* instead of final layer representations. We find that synthetic images indeed provide helpful information about feature map activations ($82 \pm 4\%$ accuracy; chance would be 50%). However, natural images — originally intended to be a baseline — outperform these synthetic images by a wide margin ($92 \pm 2\%$). Additionally, participants are faster and more confident for natural images, whereas subjective impressions about the interpretability of the feature visualizations by Olah et al. (2017) are mixed. The higher informativeness of natural images holds across most layers, for both expert and lay participants as well as for hand- and randomly-picked feature visualizations. Even if only a single reference image is given, synthetic images provide less information than natural images ($65 \pm 5\%$ vs. $73 \pm 4\%$). In summary, synthetic images from a popular feature visualization method are significantly less informative for assessing CNN activations than natural images. We argue that visualization methods should improve over this simple baseline.

1 INTRODUCTION

As Deep Learning methods are being deployed across society, academia and industry, the need to understand their decisions becomes ever more pressing. Under certain conditions, a "right to explanation" is even required by law in the European Union (GDPR, 2016; Goodman & Flaxman, 2017). Fortunately, the field of *interpretability* or *explainable artificial intelligence* (XAI) is also growing: Not only are discussions on goals and definitions of interpretability advancing (Doshi-Velez & Kim, 2017; Lipton, 2018; Gilpin et al., 2018; Murdoch et al., 2019; Miller, 2019; Samek et al., 2020) but the number of explanation methods is rising, their maturity is evolving (Zeiler & Fergus, 2014; Ribeiro et al., 2016; Selvaraju et al., 2017; Kim et al., 2018) and they are tested and

*Joint first and corresponding authors: firstname.lastname@uni-tuebingen.de
†Current affiliation: Institute of Psychology and Center for Cognitive Science, Technische Universität Darmstadt
‡Joint senior authors

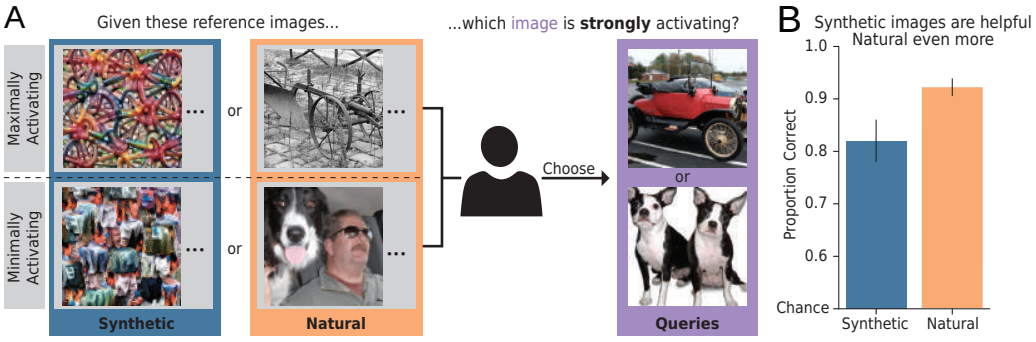


Figure 1: How useful are synthetic compared to natural images for interpreting neural network activations? **A: Human experiment.** Given extremely activating reference images (either synthetic or natural), a human participant chooses which out of two query images is also a strongly activating image. Synthetic images were generated via feature visualization (Olah et al., 2017). **B: Core result.** Participants are well above chance for synthetic images — but even better when seeing *natural* reference images.

used in real-world scenarios like medicine (Cai et al., 2019; Kröll et al., 2020) and meteorology (Ebert-Uphoff & Hilburn, 2020).

We here focus on the popular post-hoc explanation method (or interpretability method) of *feature visualizations via activation maximization*[1]. First introduced by Erhan et al. (2009) and subsequently improved by many others (Mahendran & Vedaldi, 2015; Nguyen et al., 2015; Mordvintsev et al., 2015; Nguyen et al., 2016a; 2017), these synthetic, maximally activating images seek to visualize features that a specific network unit, feature map or a combination thereof is selective for. However, feature visualizations are surrounded by a great controversy: How accurately do they represent a CNN's inner workings—or in short, how useful are they? This is the guiding question of our study.

On the one hand, many researchers are convinced that feature visualizations are interpretable (Graetz, 2019) and that "features can be rigorously studied and understood" (Olah et al., 2020b). Also other applications from Computer Vision and Natural Language Processing support the view that features are meaningful (Mikolov et al., 2013; Karpathy et al., 2015; Radford et al., 2017; Zhou et al., 2014; Bau et al., 2017; 2020) and might be formed in a hierarchical fashion (LeCun et al., 2015; Güçlü & van Gerven, 2015; Goodfellow et al., 2016). Over the past few years, extensive investigations to better understand CNNs are based on feature visualizations (Olah et al., 2020b;a; Cammarata et al., 2020; Cadena et al., 2018), and the technique is being combined with other explanation methods (Olah et al., 2018; Carter et al., 2019; Addepalli et al., 2020; Hohman et al., 2019).

On the other hand, feature visualizations can be equal parts art and engineering as they are science: vanilla methods look noisy, thus human-defined regularization mechanisms are introduced. But do the resulting beautiful visualizations accurately show what a CNN is selective for? How representative are the seemingly well-interpretable, "hand-picked" (Olah et al., 2017) synthetic images in publications for the entirety of all units in a network, a concern raised by e.g. Kriegeskorte (2015)? What if the features that a CNN is truly sensitive to are imperceptible instead, as might be suggested by the existence of adversarial examples (Szegedy et al., 2013; Ilyas et al., 2019)? Morcos et al. (2018) even suggest that units of easily understandable features play a less important role in a network. Another criticism of synthetic maximally activating images is that they only visualize extreme features, while potentially leaving other features undetected that only elicit e.g. 70% of the maximal activation. Also, polysemantic units (Olah et al., 2020b), i.e. units that are highly activated by different semantic concepts, as well as the importance of combinations of units (Olah et al., 2017; 2018; Fong & Vedaldi, 2018) already hint at the complexity of how concepts are encoded in CNNs.

One way to advance this debate is to measure the utility of feature visualizations in terms of their helpfulness for *humans*. In this study, we therefore design well-controlled psychophysical experiments that aim to quantify the informativeness of the popular visualization method by Olah et al. (2017). Specifically, participants choose which of two natural images would elicit a higher activa-

[1]Also known as *input maximization* or *maximally exciting images (MEIs)*.

tion in a CNN given a set of reference images that visualize the network selectivities. We use natural query images because real-world applications of XAI require understanding model decisions to natural inputs. To the best of our knowledge, our study is the first to probe how well humans can predict *intermediate* CNN activations. Our data shows that:

- Synthetic images provide humans with helpful information about feature map activations.
- Exemplary natural images are even more helpful.
- The superiority of natural images mostly holds across the network and various conditions.
- Subjective impressions of the interpretability of the synthetic visualizations vary greatly between participants.

2 Related Work

Significant progress has been made in recent years towards understanding CNNs for image data. Here, we mention a few selected methods as examples of the plethora of approaches for understanding CNN decision-making: *Saliency maps* show the importance of each pixel to the classification decision (Springenberg et al., 2014; Bach et al., 2015; Smilkov et al., 2017; Zintgraf et al., 2017), *concept activation vectors* show a model's sensitivity to human-defined concepts (Kim et al., 2018), and other methods - amongst feature visualizations - focus on explaining individual units (Bau et al., 2020). Some tools integrate interactive, software-like aspects (Hohman et al., 2019; Wang et al., 2020; Carter et al., 2019; Collaris & van Wijk, 2020; OpenAI, 2020), combine more than one explanation method (Shi et al., 2020; Addepalli et al., 2020) or make progress towards automated explanation methods (Lapuschkin et al., 2019; Ghorbani et al., 2019). As overviews, we recommend Gilpin et al. (2018); Zhang & Zhu (2018); Montavon et al. (2018) and Carvalho et al. (2019).

Despite their great insights, challenges for explanation methods remain. Oftentimes, these techniques are criticized as being over-engineered; regarding feature visualizations, this concerns the loss function and techniques to make the synthetic images look interpretable (Nguyen et al., 2017). Another critique is that interpretability research is not sufficiently tested against falsifiable hypotheses and rather relies too much on intuition (Leavitt & Morcos, 2020).

In order to further advance XAI, scientists advocate different directions. Besides the focus on developing additional methods, some researchers (e.g. Olah et al. (2020b)) promote the "natural science" approach, i.e. studying a neural network extensively and making empirical claims until falsification. Yet another direction is to quantitatively evaluate explanation methods. So far, only decision-level explanation methods have been studied in this regard. Quantitative evaluations can either be realized with humans directly or with mathematically-grounded models as an approximation for human perception. Many of the latter approaches show great insights (e.g. Hooker et al. (2019); Nguyen & Martínez (2020); Fel & Vigouroux (2020); Lin et al. (2020); Tritscher et al. (2020); Tjoa & Guan (2020)). However, a recent study demonstrates that metrics of the explanation quality computed without human judgment are inconclusive and do not correspond to the *human* rankings (Biessmann & Refiano, 2019). Additionally, Miller (2019) emphasizes that XAI should build on existing research in philosophy, cognitive science and social psychology.

The body of literature on human evaluations of explanation methods is growing: Various combinations of data types (tabular, text, static images), task set-ups and participant pools (experts vs. laypeople, on-site vs. crowd-sourcing) are being explored. However, these studies all aim to investigate final model decisions and do not probe intermediate activations like our experiments do. For a detailed table of related studies, see Appendix Sec. A.3. A commonly employed task paradigm is the "forward simulation / prediction" task, first introduced by Doshi-Velez & Kim (2017): Participants guess the model's computation based on an input and an explanation. As there is no absolute metric for the goodness of explanation methods (yet), comparisons are always performed within studies, typically against baselines. The same holds for additional data collected for confidence or trust ratings. According to the current literature, studies reporting positive effects of explanations (e.g. Kumarakulasinghe et al. (2020)) slightly outweigh those reporting inconclusive (e.g. Alufaisan et al. (2020); Chu et al. (2020)) or even negative effects (e.g. Shen & Huang (2020)).

Figure 2: Example trial in psychophysical experiments. A participant is shown minimally and maximally activating reference images for a certain feature map on the sides and is asked to select the image from the center that also strongly activates that feature map. The answer is given by clicking on the number according to the participant's confidence level (1: not confident, 2: somewhat confident, 3: very confident). After each trial, the participant receives feedback which image was indeed the maximally activating one. For screenshots of each step in the task, see Appendix Fig. 7.

To our knowledge, no study has yet evaluated the popular explanation method of feature visualizations and how it improves human understanding of intermediate network activations. This study therefore closes an important gap: By presenting data for a forward prediction task of a CNN, we provide a quantitative estimate of the informativeness of maximally activating images generated with the method of Olah et al. (2017). Furthermore, our experiments are unique as they probe for the first time how well humans can predict *intermediate* model activations.

3 METHODS

We perform two human psychophysical studies[2] with different foci (Experiment I ($N = 10$) and Experiment II ($N = 23$)). In both studies, the task is to choose the one image out of two natural query images (two-alternative forced choice paradigm) that the participant considers to also elicit a strong activation given some reference images (see Fig. 2). Apart from the image choice, we record the participant's confidence level and reaction time. Specifically, responses are given by clicking on the confidence levels belonging to either query image. In order to gain insights into how intuitive participants find feature visualizations, their subjective judgments are collected in a separate task and a dynamic conversation after the experiment (for details, see Appendix Sec. A.1.1 and Appendix Sec. A.2.6).

All design choices are made with two main goals: (1) allowing participants to achieve the *best performance possible* to approximate an upper bound on the helpfulness of the explanation method, and (2) gaining a *general* impression of the helpfulness of the examined method. As an example, we choose the natural query images from among those of lowest and highest activations ($\rightarrow$ best possible performance) and test many different feature maps across the network ($\rightarrow$ generality). For more details on the human experiment besides the ones below, see Appendix Sec. A.1.

In Experiment I, we focus on comparing the performance of synthetic images to two baseline conditions: natural reference images and no reference images. In Experiment II, we compare lay vs. expert participants as well as different presentation schemes of reference images. Expert participants qualify by being familiar or having practical experience with feature visualization techniques or at least CNNs. Regarding presentation schemes, we vary whether only maximally or both maximally and minimally activating images are shown; as well as how many example images of each of these are presented (1 or 9).

Following the existing work on feature visualization (Olah et al., 2017; 2018; 2020b;a), we use an Inception V1 network[3] (Szegedy et al., 2015) trained on ImageNet (Deng et al., 2009; Russakovsky

[2]Code and data is available at `https://bethgelab.github.io/testing_visualizations/`

[3]also known as GoogLeNet

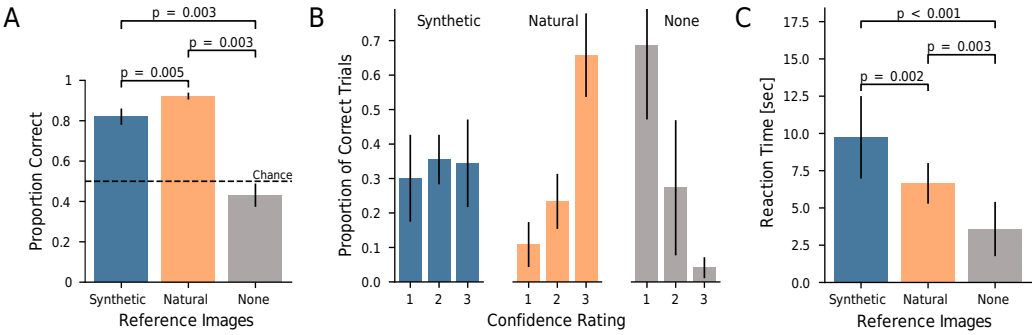


Figure 3: Participants are better, more confident and faster at judging which of two query images causes higher feature map activation with natural than with synthetic reference images. **A: Performance.** Given synthetic reference images, participants are well above chance (proportion correct: $82 \pm 4\%$), but even better for natural reference images ($92 \pm 2\%$). Without reference images (baseline comparison "None"), participants are close to chance. **B: Confidence.** Participants are much more confident (higher rating = more confident) for natural than for synthetic images on correctly answered trials (χ^2, $p < .001$). **C: Reaction time.** For correctly answered trials, participants are on average faster when presented with natural than with synthetic reference images. We show additional plots on confidence and reaction time for incorrectly answered trials and all trials in the Appendix (Fig. 16); for Experiment II, see Fig. 17.). The p-values in A and C correspond to Wilcoxon signed-rank tests.

et al., 2015). The synthetic images throughout this study are the optimization results of the feature visualization method by Olah et al. (2017) with the spatial average of a whole feature map ("channel objective"). The natural stimuli are selected from the validation set of the ImageNet ILSVRC 2012 dataset (Russakovsky et al., 2015) according to their activations for the feature map of interest. Specifically, the images of the most extreme activations are sampled, while ensuring that each lay or expert participant sees different query and reference images. A more detailed description of the specific sampling process for natural stimuli and the generation process of synthetic stimuli is given in Sec. A.1.2.

4 Results

In this section, all figures show data from Experiment I except for Fig. 5A+C, which show data from Experiment II. All figures for Experiment II, which replicate the findings of Experiment I, as well as additional figures for Experiment I (such as a by-feature-map analysis), can be found in the Appendix Sec. A.2. Note that (unless explicitly noted otherwise), error bars denote two standard errors of the mean of the participant average metric.

4.1 Participants are better, more confident and faster with natural images

Synthetic images can be helpful: Given synthetic reference images generated via feature visualization (Olah et al., 2017), participants are able to predict whether a certain network feature map prefers one over the other query image with an accuracy of $82 \pm 4\%$, which is well above chance level (50%) (see Fig. 3A). However, performance is even higher in what we intended to be the baseline condition: natural reference images ($92 \pm 2\%$). Additionally, for correct answers, participants much more frequently report being highly certain on natural relative to synthetic trials (see Fig. 3B), and their average reaction time is approximately 3.7 seconds faster when seeing natural than synthetic reference images (see Fig. 3C). Taken together, these findings indicate that in our setup, participants are not just better overall, but also more confident and substantially faster on natural images.

4.2 Natural images are more helpful across a broad range of layers

Next, we take a more fine-grained look at performance across different layers and branches of the Inception modules (see Fig. 4). Generally, feature map visualizations from lower layers show low-level features such as striped patterns, color or texture, whereas feature map visualizations from

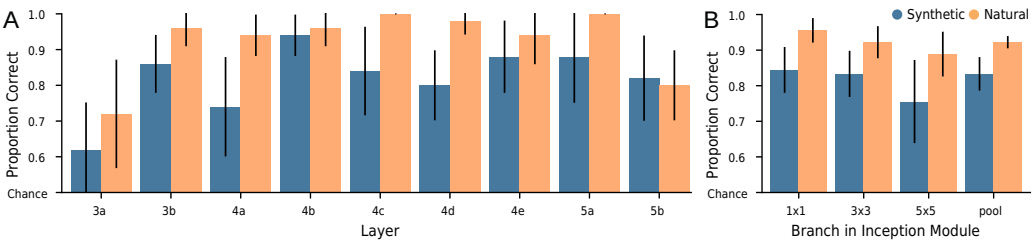


Figure 4: Performance is high across (**A**) a broad range of layers and (**B**) all branches of the Inception modules. The latter differ in their kernel sizes (1×1, 3×3, 5×5, pool). Again, natural images are (mostly) more helpful than synthetic images. Additional plots for the none condition as well as Experiment II can be found in the Appendix in respectively Fig. 18 and Fig. 19.

higher layers tend to show more high-level concepts like (parts of) objects (LeCun et al., 2015; Güçlü & van Gerven, 2015; Goodfellow et al., 2016). We find performance to be reasonably high across most layers and branches: participants are able to match both low-level and high-level patterns (despite not being explicitly instructed what layer a feature map belonged to). Again, natural images are mostly more helpful than synthetic images.

4.3 For expert and lay participants alike: natural images are more helpful

Explanation methods seek to explain aspects of algorithmic decision-making. Importantly, an explanation should not just be amenable to experts but to anyone affected by an algorithm's decision. We here test whether the explanation method of feature visualization is equally applicable to expert and lay participants (see Fig. 5A). Contrary to our prior expectation, we find no significant differences in expert vs. lay performance (RM ANOVA, $p = .44$, for details see Appendix Sec. A.2.2). Hence, extensive experience with CNNs is not necessary to perform well in this forward simulation task. In line with the previous main finding, both experts and lay participants are both better in the natural than in the synthetic condition.

4.4 Even for hand-picked feature visualizations, performance is higher on natural images

Often, explanation methods are presented using carefully selected network units, raising the question whether author-chosen units are representative for the interpretability method as a whole. Olah et al. (2017) identify a number of particularly interpretable feature maps in Inception V1 in their appendix overview. When presenting either these hand-picked visualizations[4] or randomly selected ones, performance for hand-picked feature maps improves slightly (Fig. 5B); however this performance difference is small and not significant for both natural (Wilcoxon test, $p = .59$) and synthetic (Wilcoxon test, $p = .18$) reference images (see Appendix Sec. A.2.4 for further analysis). Consistent with the findings reported above, performance is higher for natural than for synthetic reference images *even on carefully selected hand-picked feature maps*.

4.5 Additional information boosts performance, especially for natural images

Publications on feature visualizations vary in terms of how optimized images are presented: Often, a single maximally activating image is shown (e.g. Erhan et al. (2009); Carter et al. (2019); Olah et al. (2018)); sometimes a few images are shown simultaneously (e.g. Yosinski et al. (2015); Nguyen et al. (2016b)), and on occasion both maximally and minimally activating images are shown in unison (Olah et al. (2017)). Naturally, the question arises as to what influence (if any) these choices have, and whether there is an optimal way of presenting extremely activating images. For this reason, we systematically compare approaches along two dimensions: the number of reference images (1 vs. 9) and the availability of minimally activating images (only Max vs. Min+Max). The results can

[4]All our hand-picked feature maps are taken from the pooling branch of the Inception module. As the appendix overview in Olah et al. (2017) does not contain one feature map for each of these, *we* select interpretable feature maps for the missing layers mixed5a and mixed5b ourselves.

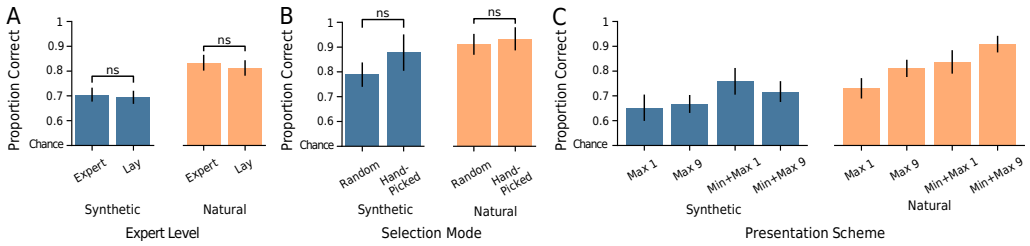


Figure 5: We found no evidence for large effects of expert level or feature map selection. However, performance does improve with additional information. **A: Expert level.** Both experts and lay participants perform equally well (RM ANOVA, $p = .44$), and consistently better on natural than on synthetic images. **B: Selection mode.** There is no significant performance difference between hand-picked feature maps selected for interpretability and randomly selected ones (Wilcoxon test, $p = .18$ for synthetic and $p = .59$ for natural reference images). **C: Presentation scheme.** Presenting both maximally and minimally activating images simultaneously (Min+Max) and presenting nine instead of one single reference image tend to improve performance, especially for natural reference images. "ns" highlights non-significant differences.

be found in Fig. 5C. When just a single maximally activating image is presented (condition Max 1), natural images already outperform synthetic images ($73 \pm 4\%$ vs. $64 \pm 5\%$). With additional information along either dimension, performance improves both for natural as well as for synthetic images. The stronger boost in performance, however, is observed for natural reference images. In fact, performance is higher for natural than for synthetic reference images in all four conditions. In the Min+Max 9 condition, a replication of the result from Experiment I shown in Fig. 3A, natural images now outperform synthetic images by an even larger margin (91 ± 3 vs. $72 \pm 4\%$).

4.6 Subjectively, interpretability of feature visualizations varies greatly

While our data suggests that feature visualizations are indeed helpful for humans to predict CNN activations, we want to emphasize again that our design choices aim at an upper bound on their informativeness. Another important aspect of evaluating an explanation method is the subjective impression. Besides recording confidence ratings and reaction times, we collect judgments on *intuitiveness trials* (see Appendix Fig. 14) and oral impressions after the experiments. The former ask for ratings of how intuitive feature visualizations appear for natural images. As Fig. 6A+B show, participants perceive the intuitiveness of synthetic feature visualizations for strongly activating natural dataset images very differently. Further, the comparison of intuitiveness judgments before and after the main experiments reveals only a small significant average improvement for one out of three feature maps (see Fig. 6B+C, Wilcoxon test, $p < .001$ for mixed4b). The interactive conversations paint a similar picture: Some synthetic feature visualizations are perceived as intuitive while others do not correspond to understandable concepts. Nonetheless, four participants report that their first "gut feeling" for interpreting these reference images (as one participant phrased it) is more reliable. Further, a few participants point out that the synthetic visualizations are exhausting to understand. Finally, three participants additionally emphasize that the minimally activating reference images played an important role in their decision-making.

In a by-feature-map analysis (see Appendix A.2.7 for details and images, as well as Supplementary Material 1 for more images), we compare differences and commonalities for feature maps of different performance levels. According to our observations, easy feature maps seem to contain clear object parts or shapes. In contrast, difficult feature maps seem to have diverse reference images, features that do not correspond to human concepts, or contain conflicting information as to which commonalities between query and reference images matter more. Bluntly speaking, we are also often surprised that participants identified the correct image — the reasons for this are unclear to us.

5 Discussion & Conclusion

Feature visualizations such as synthetic maximally activating images are a widely used explanation method, but it is unclear whether they indeed help humans to understand CNNs. Using well-controlled psychophysical experiments with both expert and lay participants, we here conduct the

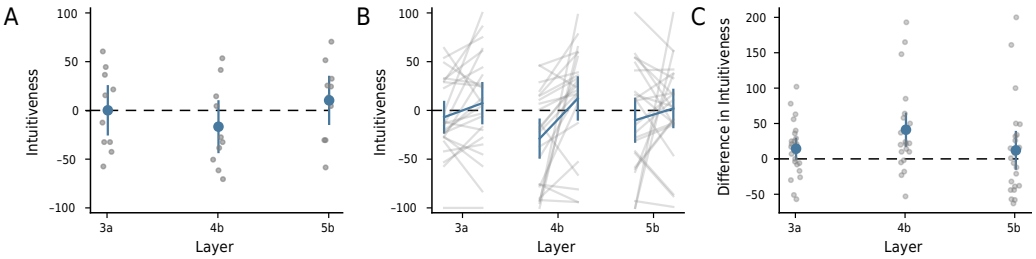


Figure 6: The subjective intuitiveness of feature visualizations varies greatly (see **A** for the ratings from the beginning of Experiment I and **B** for the ratings at the beginning and end of Experiment II). The means over all participants yield a neutral result, i.e. the visualizations are neither un- nor intuitive, and the improvement of subjective intuitiveness before and after the experiment is only significant for one feature map (mixed4b). **C:** On average, participants found feature visualizations slightly more intuitive after doing the experiment as the differences larger than zero show. In all three subfigures, gray dots and lines show data per participant.

very first investigation of intermediate synthetic feature visualizations by Olah et al. (2017): Can participants predict which of two query images leads to a strong activation in a feature map, given extremely activating visualizations? Specifically, we shed light on the following questions:

(1.) *How informative are synthetic feature visualizations — and how do they compare to a natural image baseline?* We find above-chance performance given synthetic feature visualizations, but to our own surprise, synthetic feature visualizations are systematically *less* informative than the simple baseline of strongly activating natural images. Interestingly, many synthetic feature visualizations contain regularization mechanisms to introduce more "natural structure" (Olah et al., 2017), sometimes even called a "natural image prior" (Mahendran & Vedaldi, 2015; Offert & Bell, 2020). This raises the question: Are natural images maybe all you need? One might posit that extremely activating natural (reference) images would have an unfair advantage because we also test on extremely activating natural (query) images. However, our task design ultimately reflects that XAI is mainly concerned with explaining how units behave on *natural* inputs. Furthermore, the fact that feature visualization are not bound to the natural image manifold is often claimed as an advantage because it supposedly allows them to capture more precisely which features a unit is sensitive to (Olah et al., 2017). Our results, though, demonstrate that this is not the case if we want to understand the behavior of units on natural inputs.

(2.) *Do you need to be a CNN expert in order to understand feature visualizations?* To the best of our knowledge, our study is the first to compare the performances of expert and lay people when evaluating explanation methods. Previously, publications either focused on only expert groups (Hase & Bansal, 2020; Kumarakulasinghe et al., 2020) or only laypeople (Schmidt & Biessmann, 2019; Alufaisan et al., 2020). Our experiment shows no significant difference between expert and lay participants in our task — both perform similarly well, and even better on natural images: a replication of our main finding. While a few caveats remain when moving an experiment from the well-controlled lab to a crowdsourcing platform (Haghiri et al., 2019), this suggests that future studies may not have to rely on selected expert participants, but may leverage larger lay participant pools.

(3.) *Are hand-picked synthetic feature visualizations representative?* An open question was whether the visualizations shown in publications represent the general interpretability of feature visualizations (a concern voiced by e.g. Kriegeskorte, 2015), even though they are hand-picked (Olah et al., 2017). Our finding that there is no large difference in performance between hand- and randomly-picked feature visualizations suggests that this aspect is minor.

(4.) *What is the best way of presenting images?* Existing work suggests that more than one example (Offert, 2017) and particularly negative examples (Kim et al., 2016) enhance human understanding of data distributions. Our systematic exploration of presentation schemes provides evidence that increasing the number of reference images as well as presenting both minimally *and* maximally activating reference images (as opposed to only maximally activating ones) improve human performance. This finding might be of interest to future studies aiming at peak performance or for developing software for understanding CNNs.

(5.) *How do humans subjectively perceive feature visualizations?* Apart from the high informativeness of explanations, another relevant question is how much trust humans have in them. In our experiment, we find that subjective impressions of how reasonable synthetic feature visualizations are for explaining responses to natural images vary greatly. This finding is in line with Hase & Bansal (2020) who evaluated explanation methods on text and tabular data.

Caveats. Despite our best intentions, a few caveats remain: The forward simulation paradigm is only one specific way to measure the informativeness of explanation methods, but does not allow us to make judgments about their helpfulness in other applications such as comparing different CNNs. Further, we emphasize that all experimental design choices were made with the goal to measure the best possible performance. As a consequence, our finding that synthetic reference images help humans predict a network's strongly activating image may not necessarily be representative of a less optimal experimental set-up with e.g. query images corresponding to less extreme feature map activations. Knobs to further de- or increase participant performance remain (e.g. hyper-parameter choices could be tuned to layers). Finally, while we explored one particular method in depth (Olah et al., 2017); it remains an open question whether the results can be replicated for other feature visualizations methods.

Future directions. We see many promising future directions. For one, the current study uses query images from extreme opposite ends of a feature map's activation spectrum. For a more fine-grained measure of informativeness, we will study query images that elicit more similar activations. Additionally, future participants could be provided with even *more* information—such as, for example, where a feature map is located in the network. Furthermore, it has been suggested that the combination of synthetic and natural reference images might provide synergistic information to participants (Olah et al., 2017), which could again be studied in our experimental paradigm. Finally, further studies could explore single neuron-centered feature visualizations, combinations of units as well as different network architectures.

Taken together, our results highlight the need for thorough human quantitative evaluations of feature visualizations and suggest that example natural images provide a surprisingly challenging baseline for understanding CNN activations.

AUTHOR CONTRIBUTIONS

The initiative of investigating human predictability of CNN activations came from WB. JB, WB, MB and TSAW jointly combined it with the idea of investigating human interpretability of feature visualizations. JB led the project. JB, RSZ and JS jointly designed and implemented the experiments (with advice and feedback from TSAW, RG, MB and WB). The data analysis was performed by JB and RSZ (with advice and feedback from RG, TSAW, MB and WB). JB designed, and JB and JS implemented the pilot study. JB conducted the experiments (with help from JS). RSZ performed the statistical significance tests (with advice from TSAW and feedback from JB and RG). MB helped shape the bigger picture and initiated intuitiveness trials. WB provided day-to-day supervision. JB, RSZ and RG wrote the initial version of the manuscript. All authors contributed to the final version of the manuscript.

ACKNOWLEDGMENTS

We thank Felix A. Wichmann and Isabel Valera for helpful discussions. We further thank Alexander Böttcher and Stefan Sietzen for support as well as helfpul discussions on technical details. Additionally, we thank Chris Olah for clarifications via `slack.distill.pub`. Moreover, we thank Leon Sixt for valuable feedback on the introduction and related work. From our lab, we thank Matthias Kümmerer, Matthias Tangemann, Evgenia Rusak and Ori Press for helping in piloting our experiments, as well as feedback from Evgenia Rusak, Claudio Michaelis, Dylan Paiton and Matthias Kümmerer. And finally, we thank all our participants for taking part in our experiments.

We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting JB, RZ and RG. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the Competence Center for Machine Learning (TUE.AI, FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the Cluster of Excellence Machine Learning: New Perspectives for Sciences (EXC2064/1), and the German Research Foundation (DFG; SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP3, project number 276693517).

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Sravanti Addepalli, Dipesh Tamboli, R Venkatesh Babu, and Biplab Banerjee. Saliency-driven class impressions for feature visualization of deep neural networks. *arXiv preprint arXiv:2007.15861*, 2020.

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 275–285, 2020.

Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. Does explainable artificial intelligence improve human decision-making? *arXiv preprint arXiv:2006.11194*, 2020.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6541–6549, 2017.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 2020.

Felix Biessmann and Dionysius Irza Refiano. A psychophysics approach for quantitative comparison of interpretable computer vision models. *arXiv preprint arXiv:1912.05011*, 2019.

Santiago A Cadena, Marissa A Weis, Leon A Gatys, Matthias Bethge, and Alexander S Ecker. Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 217–232, 2018.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 2019. doi: 10.23915/distill.00015. https://distill.pub/2019/activation-atlas.

Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*, 2017.

Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

Dennis Collaris and Jarke J van Wijk. Explainexplore: Visual exploration of machine learning explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 26–35. IEEE, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Jürgen Dieber and Sabrina Kirrane. Why model why? assessing the strengths and limitations of lime. *arXiv preprint arXiv:2012.00093*, 2020.

Jonathan Dinu, Jeffrey Bigham, and J Zico Kolter. Challenging common interpretability assumptions in feature attribution explanations. *arXiv preprint arXiv:2012.02748*, 2020.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Imme Ebert-Uphoff and Kyle Hilburn. Evaluation, tuning and interpretation of neural networks for working with images in meteorological applications. *Bulletin of the American Meteorological Society*, pp. 1–49, 2020.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Thomas Fel and David Vigouroux. Representativity and consistency measures for deep neural network explanations. *arXiv preprint arXiv:2009.04521*, 2020.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8730–8738, 2018.

GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pp. 9277–9286, 2019.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE, 2018.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Fabio M. Graetz. How to visualize convolutional features in 40 lines of code, Jan 2019. URL `https://towardsdatascience.com/how-to-visualize-convolutional-features-in-40-lines-of-code-70b7d87b0030`.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Siavash Haghiri, Patricia Rubisch, Robert Geirhos, Felix Wichmann, and Ulrike von Luxburg. Comparison-based framework for psychophysics: Lab versus crowdsourcing. *arXiv preprint arXiv:1905.07234*, 2019.

Peter Hase and Mohit Bansal. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*, 2020.

Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE transactions on visualization and computer graphics*, 26(1):1096–1106, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 9737–9748, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

JASP Team. JASP (Version 0.13.1), 2020. URL `https://jasp-stats.org/`.

Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33, 2020.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in neural information processing systems*, pp. 2280–2288, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

Jean-Philippe Kröll, Simon B Eickhoff, Felix Hoffstaedter, and Kaustubh R Patil. Evolving complex yet interpretable representations: application to alzheimer's diagnosis and prognosis. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8. IEEE, 2020.

Nesaretnam Barr Kumarakulasinghe, Tobias Blomberg, Jintai Liu, Alexandra Saraiva Leao, and Panagiotis Papapetrou. Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 7–12. IEEE, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Yi-Shan Lin, Wen-Chuan Lee, and Z Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. *arXiv preprint arXiv:2009.10639*, 2020.

Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. Model agnostic multilevel explanations. *Advances in Neural Information Processing Systems*, 33, 2020.

An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*, 2020.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, pp. 3387–3395, 2016a.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016b.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477, 2017.

Fabian Offert. " i know it when i see it". visualization and intuitive interpretability. *arXiv preprint arXiv:1711.08042*, 2017.

Fabian Offert and Peter Bell. Perceptual bias and technical metapictures: critical machine vision as a humanities challenge. *AI & SOCIETY*, pp. 1–12, 2020.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 2020a. doi: 10.23915/distill.00024.002. https://distill.pub/2020/circuits/early-vision.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020b. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

OpenAI. OpenAI Microscope. `https://microscope.openai.com/models`, 2020. (Accessed on 09/12/2020).

Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51(1):195–203, 2019.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*, 2020.

Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Hua Shen and Ting-Hao'Kenneth' Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. *arXiv preprint arXiv:2008.11721*, 2020.

Rui Shi, Tianxing Li, and Yasushi Yamaguchi. Group visualization of class-discriminative features. *Neural Networks*, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Erico Tjoa and Cuntai Guan. Quantifying explainability of saliency methods in deep neural networks. *arXiv preprint arXiv:2009.02899*, 2020.

Julian Tritscher, Markus Ring, Daniel Schlr, Lena Hettinger, and Andreas Hotho. Evaluation of post-hoc xai approaches through synthetic tabular data. In *International Symposium on Methodologies for Intelligent Systems*, pp. 422–430. Springer, 2020.

Zijie J Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Chau. Cnn explainer: Learning convolutional neural networks with interactive visualization. *arXiv preprint arXiv:2004.15004*, 2020.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

A APPENDIX

A.1 DETAILS ON METHODS

A.1.1 HUMAN EXPERIMENTS

In our two human psychophysical studies, we ask humans to predict a feature map's strongly activating image ("forward simulation task", Doshi-Velez & Kim 2017). Answers to the two-alternative forced choice paradigm are recorded together with the participants' confidence level (1: not confident, 2: somewhat confident, 3: very confident, see Fig. 7). Time per trial is unlimited and we record reaction time. After each trial, feedback is given (see Fig. 7). A progress bar at the bottom of the screen indicates how many trials of a block are already completed. As reference images, either synthetic, natural or no reference images are given. The synthetic images are the feature visualizations from the method of Olah et al. (2017). Trials of different reference images are arranged in blocks. Synthetic and natural reference images are alternated, and, in the case of Experiment I, framed by trials without reference images (see Fig. 8A, B). The order of the reference image types is counter-balanced across subjects.

The main trials in the experiments are complemented by practice, catch and intuitiveness trials. To avoid learning effects, we use different feature maps for each trial type per participant. Specifically, *practice trials* give participants the opportunity to familiarize themselves with the task. In order to monitor the attention of participants, *catch trials* appear randomly throughout blocks of main trials. Here, the query images are a copy of one of the reference images, i.e., there is an obvious correct answer (see Fig. 15). This control mechanism allows us to decide whether trial blocks should be excluded from the analysis due to e.g. fatigue. To obtain the participant's subjective impression of the helpfulness of maximally activating images, the experiments are preceded (and also succeeded in the case of Experiment II) by three *intuitiveness trials* (see Fig. 14). Here, participants judge in a slightly different task design how intuitive they consider the synthetic stimuli for the natural stimuli. For more details on the intuitiveness trials, see below.

At the end of the experiment, all expert participants in Experiment I and all lay (but not expert) participants in Experiment II are asked about their strategy and whether it changed over time. The information gained through the first group allows us to understand the variety of cues used and paves the way to identify interesting directions for follow-up experiments. The information gained through the second group allowed comparisons to experts' impressions reported in Experiment I.

Experiment I The first experiment focuses on comparing performance of synthetic images to two baselines: natural reference images and no reference images (see Fig. 8A). Screenshots of trials are shown in Fig. 12. In total, 45 feature maps are tested: 36 of these are uniformly sampled from the feature maps of each of the four branches for each of the nine Inception modules. The other nine feature maps are uniformly hand-picked for interpretability from the Inception modules' pooling branch based on the appendix overview selection provided by Olah et al. (2017) or based on our own choices. In the spirit of a *general* statement about the explainability method, different participants see different natural reference and query images, and each participant sees different natural query images for the same feature maps in different reference conditions. To check the consistency of participants' responses, we repeat six randomly chosen main trials for each of the three tested reference image types at the end of the experiment.

Experiment II The second experiment (see Fig. 8B) is about testing expert vs. lay participants as well as comparing different presentation schemes[5] (Max 1, Min+Max 1, Max 9 and Min+Max 9, see Fig. 8E). Screenshots of trials are shown in Fig. 13. In total, 80 feature maps are tested: They are uniformly sampled from every second layer with an Inception module of the network (hence a total of 5 instead of 9 layers), and from all four branches of the Inception modules. Given the focus on four different presentation schemes in this experiment, we repeat the sampling method four times without overlap. In terms of reference image types, only synthetic and natural images are tested. Like in Experiment I, different participants see different natural reference and query images.

[5] In pilot experiments, we learned that participants preferred 9 over 4 reference images, hence this "default" choice in Experiment I.

(a) Screen at the beginning of a trial. The question is which of the two natural images at the center of the screen also strongly activates the CNN feature map given the reference images on the sides.

(b) Screen including a participant's answer visualized by black boxes around the image and the confidence level. A participant indicates which natural image at the center would also be a strongly activating image by clicking on the number corresponding to his/her confidence level (1: not confident, 2: somewhat confident, 3: confident). The time until a participant selects an answer is recorded ("reaction time").

(c) Screen including a participant's answer (black boxes) and feedback on which image is indeed also a strongly activating image (green box).

Figure 7: Forward Simulation Task. The progress bar at the bottom of the screen indicates the progress within one block of trials.

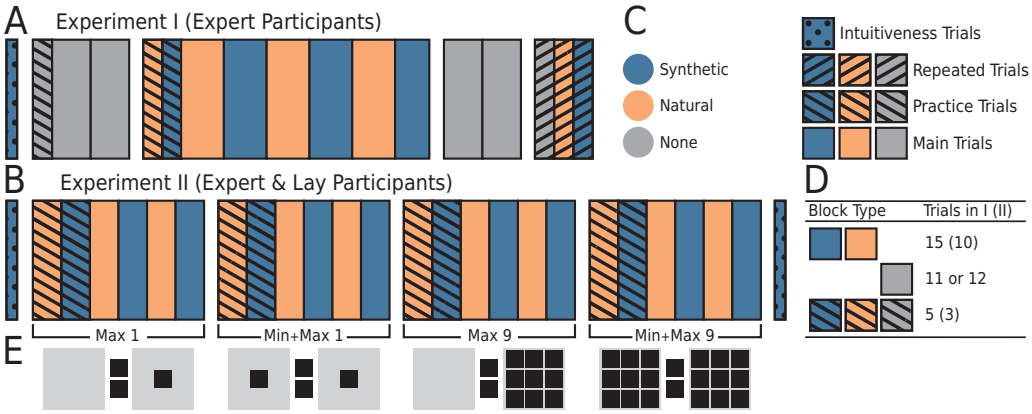


Figure 8: Detailed structure of the two experiments with different foci. **A: Experiment I.** Here, the focus is on comparing performance of synthetic and natural reference images to the most simple baseline: no reference images ("None"). To counter-balance conditions, the order of natural and synthetic blocks is alternated across participants. For each of the three reference image types (synthetic, natural and none), 45 relevant trials are used plus additional catch, practice and repeated trials. **B: Experiment II.** Here, the focus is on testing expert and lay participants as well as comparing different presentation schemes (Max 1, Min+Max 1, Max 9 and Min+Max 9, see **E** for illustrations). Both the order of natural and synthetic blocks as well as the four presentation conditions are counter-balanced across participants. To maintain a reasonable experiment length for each participant, only 20 relevant trials are used per reference image type and presentation scheme, plus additional catch and practice trials. **C:** Legend. **D:** Number of trials per block type (i.e. reference image type and main vs. practice trial) and experiment. Catch trials are not shown in the figure; there was a total of 3 (2) catch trials per each synthetic and natural main block in Experiment I (II). **E:** Illustration of presentation schemes. In Experiment II, all four schemes are tested, in Experiment I only Min+Max 9 is tested.

However, expert and lay participants see the same images. For details on the counter-balancing of all conditions, please refer to Tab. 1.

Intuitiveness Trials In order to obtain the participants' subjective impression of the helpfulness of maximally activating images, we add trials at the beginning of the experiments, and also at the end of Experiment II. The task set-up is slightly different (see Fig. 14): Only maximally activating (i.e. no minimally activating) images are shown. We ask participants to rate how intuitive they find the explanation of the entirety of the synthetic images for the entirety of the natural images. Again, all images presented in one trial are specific to one feature map. By moving a slider to the right (left), participants judge the explanation method as intuitive (not intuitive). The ratings are recorded on a continuous scale from -100 (not intuitive) to $+100$ (intuitive). All participants see the same three trials in a randomized order. The trials are again taken from the hand-picked (i.e. interpretable) feature maps of the appendix overview in Olah et al. (2017). In theory, this again allows for the highest intuitiveness ratings possible. The specific feature maps are from a low, intermediate and high layer: feature map 43 of mixed3a, feature map 504 of mixed4b and feature map 17 of mixed 5b.

Participants Our two experiments are within-subject studies, meaning that every participant answers trials for all conditions. This design choice allows us to test fewer participants. In Experiment I, 10 expert participants take part (7 male, 3 female, age: 27.2 years, SD = 1.75). In Experiment II, 23 participants take part (of which 10 are experts; 14 male, 9 female, age: 28.1 years, SD = 6.76). Expert participants qualify by being familiar or having worked with convolutional neural networks and most of them even with feature visualization techniques. All participants are naive with respect to the aim of the study. Expert (lay) participants are paid 15€ (10 €), per hour for participation. Before the experiment, all participants give written informed consent for participating. All participants have normal or corrected to normal vision. All procedures conform to Standard

8 of the American Psychological 405 Association's "Ethical Principles of Psychologists and Code of Conduct" (2016). Before the experiment, the first author explains the task to each participant and ensures complete understanding. For lay participants, the explanation is simplified: Maximally (minimally) activating images are called "favorite images" ("non-favorite images") of a "computer program" and the question is explained as which of the two query images would also be a "favorite" image to the computer program.

Apparatus Stimuli are displayed on a VIEWPixx 3D LCD (VPIXX Technologies; spatial resolution 1920×1080 px, temporal resolution $120\,\mathrm{Hz}$). Outside the stimulus image, the monitor is set to mean gray. Participants view the display from $60\,\mathrm{cm}$ (maintained via a chinrest) in a darkened chamber. At this distance, pixels subtend approximately $0.024°$ degrees on average ($41\,\mathrm{ps}$ per degree of visual angle). Stimulus presentation and data collection is controlled via a desktop computer (Intel Core i5-4460 CPU, AMD Radeon R9 380 GPU) running Ubuntu Linux (16.04 LTS), using PsychoPy (Peirce et al., 2019, version 3.0) under Python 3.6.

A.1.2 Stimuli Selection

Model Following the existing work on feature visualization by Olah et al. (2017; 2018; 2020b;a), we use an Inception V1 network[6] (Szegedy et al., 2015) trained on ImageNet (Deng et al., 2009; Russakovsky et al., 2015). Note that the Inception V1 network used in previously mentioned work slightly deviates from the original network architecture: The 3×3 branch of Inception module mixed4a only holds 204 instead of 208 feature maps. To stay as close as possible to the aforementioned work, we also use their implementation and trained weights of the network[7]. We investigate feature visualizations for all branches (i.e. kernel sizes) of the Inception modules and sample from layers mixed3a to mixed5b before the ReLU non-linearity.

Synthethic Images from Feature Visualization The synthetic images throughout this study are the optimization results of the feature visualization method from Olah et al. (2017). We use the channel objective to find synthetic stimuli that maximally (minimally) activate the spatial mean of a given feature map of the network. We perform the optimization using lucid 0.3.8 and TensorFlow 1.15.0 (Abadi et al., 2015) and use the hyperparameter as specified in Olah et al. (2017). For the experimental conditions with more than one minimally/maximally activating reference image, we add a diversity regulariztion across the samples. In hindsight, we realized that we generated 10 synthetic images in Experiment I, even though we only needed and used 9 per feature map.

Selection of Natural Images The natural stimuli are selected from the validation set of the ImageNet ILSVRC 2012 (Russakovsky et al., 2015) dataset. To choose the maximally (minimally) activating natural stimuli for a given feature map, we perform three steps, which are illustrated in Fig. 9 and explained in the following: First, we calculate the activation of said feature map for all pre-processed images (resizing to 256×256 pixels, cropping centrally to 224×224 pixels and normalizing) and take the spatial average to get a scalar representing the excitability of the given feature map caused by the image. Second, we order the images according to the collected activation values and select the $(N_{stimuli}+1) \cdot N_{batches}$ maximally (respectively minimally) activating images. Here, $N_{stimuli}$ corresponds to the number of reference images used (either 1 or 9, see Fig. 8, **E**), the $+1$ comes from the query image, and $N_{batches} = 20$ determines the maximum number of participants we can test with our setup. Third, we distribute the selected images into $N_{stimuli}+1$ blocks. Within each block, we randomly shuffle the order of the images. Lastly, we create $N_{batches}$ batches of data by selecting one image from each of the blocks for every batch.[8]

[6] This network is considered very interpretable (Olah et al., 2018), yet other work also finds deeper networks more interpretable (Bau et al., 2017). More recent work, again, suggests that "analogous features [...] form across models [...]," i.e. that interpretable feature visualizations appear "universally" for different CNNs (Olah et al., 2020b; OpenAI, 2020).

[7] `github.com/tensorflow/lucid/tree/v0.3.8/lucid`

[8] After having performed Experiment I and II, we realized a minor bug in our code: Instead of moving every 20th image into the same batch for one participant, we moved every 10th image into the same batch for one participant. This means that we only use a total of 110 different images, instead of 200. The minimal query image is still always selected from the 20 least activating images; the maximal query image is selected from the 91st to 110th maximally activating images - and we do not use the 111th to 200th maximally activating images.

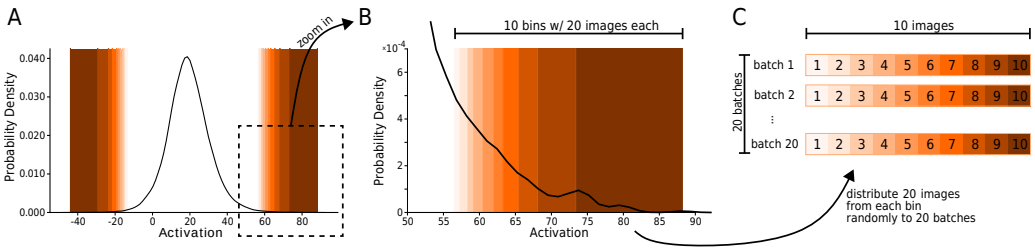


Figure 9: Sampling of natural images. **A:** Distribution of activations. For an example channel (mixed3a, kernel size 1×1, feature map 25), the smoothed distribution of activations for all $50,000$ ImageNet validation images is plotted. The natural stimuli for the experiment are taken from the tails of the distribution (shaded background). **B:** Zoomed-in tail of activations distribution. In the presentation schemes with 9 images, 10 bins with 20 images each are created (10 because of 9 reference plus 1 query image). **C:** In order to obtain 20 batches with 10 images each, the 20 images from one bin are randomly distributed to the 20 batches. This guarantees that each batch contains a fair selection of extremely activating images. The query images are *always* sampled from the most extreme bins in order to give the best signal possible. In the case of the presentation schemes with 1 reference image, the number of bins in B is reduced to 2 and the number of images per batch in C is also reduced to 2.

Subject	**Order of presentation schemes (0-3) and batch-blocks (A-D)**	**Batches**		**Order of synthetic and natural**
		Practice	**Main**	
1 2 3 4	0 (A) 1 (B) 2 (C) 3 (D) 0 (B) 2 (D) 1 (C) 3 (A) 3 (B) 1 (D) 2 (A) 0 (C) 3 (C) 2 (B) 1 (A) 0 (D)		natural: 1 synthetic: 2	natural - synthetic
5 6 7 8	see subject 1-4	0	natural: 3 synthetic: 4	synthetic - natural
9 10 11 12	see subject 1-4		natural: 5 synthetic: 6	natural - synthetic
13	see subject 1-4		natural: 7 synthetic: 8	synthetic - natural

Table 1: **Counter-balancing of conditions in Experiment II.** In total, 13 naive and 10 lay participants are tested. Each "batch block" contains 20 feature maps (sampled from five layers and all Inception module branches). Batches indicate which batch number the natural query (and reference images) are taken from.

The reasons for creating several batches of extremely activating natural images are two-fold: (1) We want to get a *general* impression of the interpretability method and would like to reduce the dependence on single images, and (2) in Experiment I, a participant has to see different query images in the three different reference conditions. A downside of this design choice is an increase in variability. The precise allocation was done as follows: In Experiment I, the natural query images of the none condition were always allocated the batch with $batch\_nr = subject\_id$, the query and reference images of the natural condition were allocated the batch with $batch\_nr = subject\_id+1$, and the natural query images of the synthetic condition were allocated the batch with $batch\_nr = subject\_id+2$. The allocation scheme in Experiment II can be found in Table 1.

Selection of Feature Maps The selection of feature maps used in Experiment I is shown in Table 2; the selection of feature maps used in Experiment II is shown in Table 3.

Layer	Branch	Feature Map	Layer	Branch	Feature Map
mixed3a	1×1	25	mixed4d	1×1	95
	3×3	189		3×3	342
	5×5	197		5×5	451
	Pool	227		Pool	483
	Pool*	230		Pool*	516
mixed3b	1×1	64	mixed4e	1×1	231
	3×3	178		3×3	524
	5×5	390		5×5	656
	Pool	430		Pool	816
	Pool*	462		Pool*	809
mixed4a	1×1	68	mixed5a	1×1	229
	3×3	257		3×3	278
	5×5	427		5×5	636
	Pool	486		Pool	743
	Pool*	501		Pool*	720
mixed4b	1×1	45	mixed5b	1×1	119
	3×3	339		3×3	684
	5×5	438		5×5	844
	Pool	491		Pool	1007
	Pool*	465		Pool*	946
mixed4c	1×1	94			
	3×3	247			
	5×5	432			
	Pool	496			
	Pool*	449			

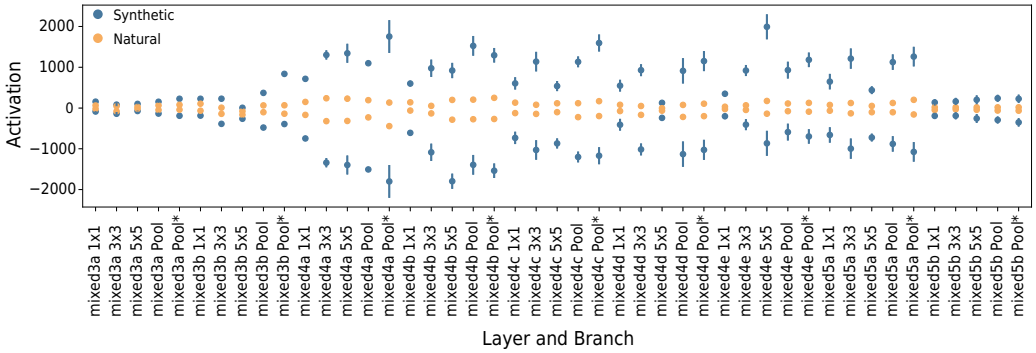
Table 2: Feature maps analyzed in Experiment I. For each of the 9 layers with an Inception module, one randomly chosen feature map per branch (1×1, 3×3, 5×5 and pool) and one additional hand-picked feature map (highlighted with *) are used.

A.1.3 Different Activation Magnitudes

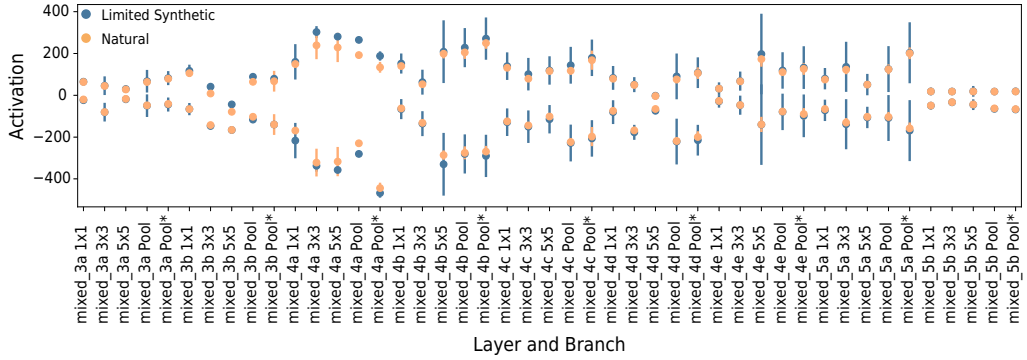
We note that the elicited activations of synthetic images are almost always about one magnitude larger than the activations of natural images (see Fig. 10a). This constitutes an inherent difference in the synthetic and natural reference image condition. A simple approach to make the two conditions more comparable is to limit the optimization process such that the resulting feature visualizations elicit activations similar to that of natural images. This can be achieved by halting the optimization process once the activations approximately match. By following that procedure one finds limited synthetic images which are indistinguishable from natural images in terms of their activations (see Fig. 10b). Importantly though, these images are visually not more similar to natural images, have a much lower color contrast than normal feature visualizations, and above all hardly resemble meaningful features (see Fig. 11).

A.1.4 Data Analysis

Significance Tests All significance tests are performed with JASP (JASP Team, 2020, version 0.13.1). For the analysis of the distribution of confidence ratings (see Fig. 3B), we use contingency tables with χ^2-tests. For testing pairwise effects in accuracy, confidence, reaction time and intuitiveness data, we report Wilcoxon signed-rank tests with uncorrected p-values (Bonferroni-corrected critical alpha values with family-wise alpha level of 0.05 reported in all figures where relevant). These non-parametric tests are preferred for these data because they do not make distributional assumptions like normally-distributed errors, as in e.g. paired t-tests. For testing marginal effects (main effects of one factor marginalizing over another) we report results from repeated measures ANOVA (RM ANOVA), which does assume normality.



(a) Activations of natural and synthetic images.



(b) Activations of natural and limited synthetic images.

Figure 10: Mean activations and standard deviations (not two standard errors of the mean!) of the minimally (below 0) and maximally (above 0) activating synthetic and natural images used in Experiment I. Note that there are 10 (i.e. accidentally not 9) synthetic images and $20 \cdot 10 = 200$ natural images (because of 20 batches) in Experiment I for both minimally and maximally activating images. Please also note that the standard deviations for the selected natural images are invisible because they are so small. Limited synthetic images refer to feature visualizations which are the result of stopping the optimization process early with the goal of matching the activation level of natural stimuli.

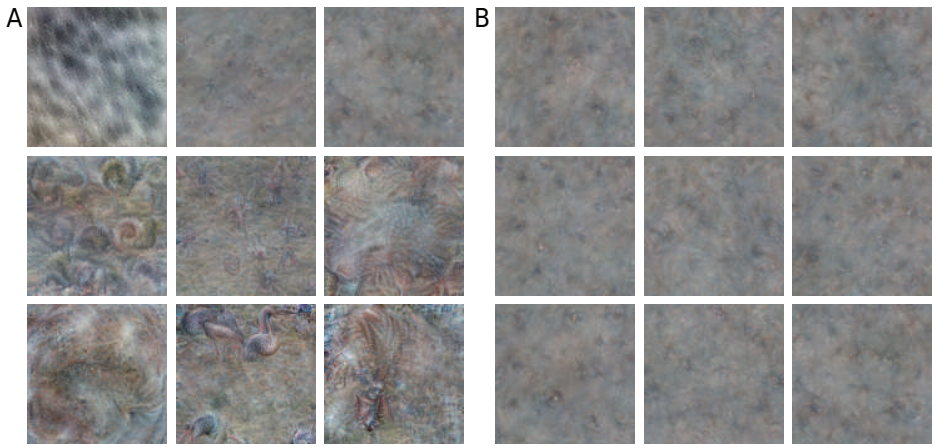


Figure 11: Limited feature visualizations, which are the result of stopping the optimization process early with the goal of matching the activation level of the chosen extreme natural stimuli. **A**: Feature visualizations for mixed_4a pool* feature map of Experiment I. **B**: Feature visualizations for all nine pool* feature maps of Experiment I.

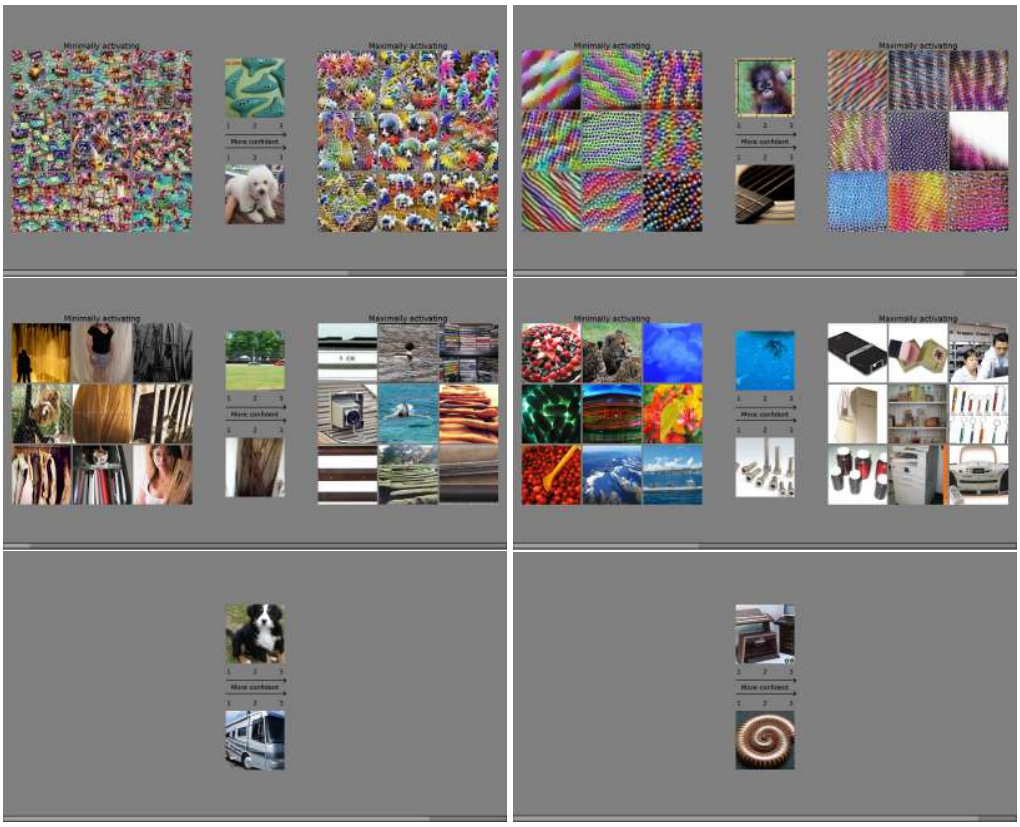


Figure 12: Experiment I: Example trials of the three reference images conditions: synthetic reference images (first row), natural reference images (second row) or no reference images (third row). The query images in the center are always natural images.

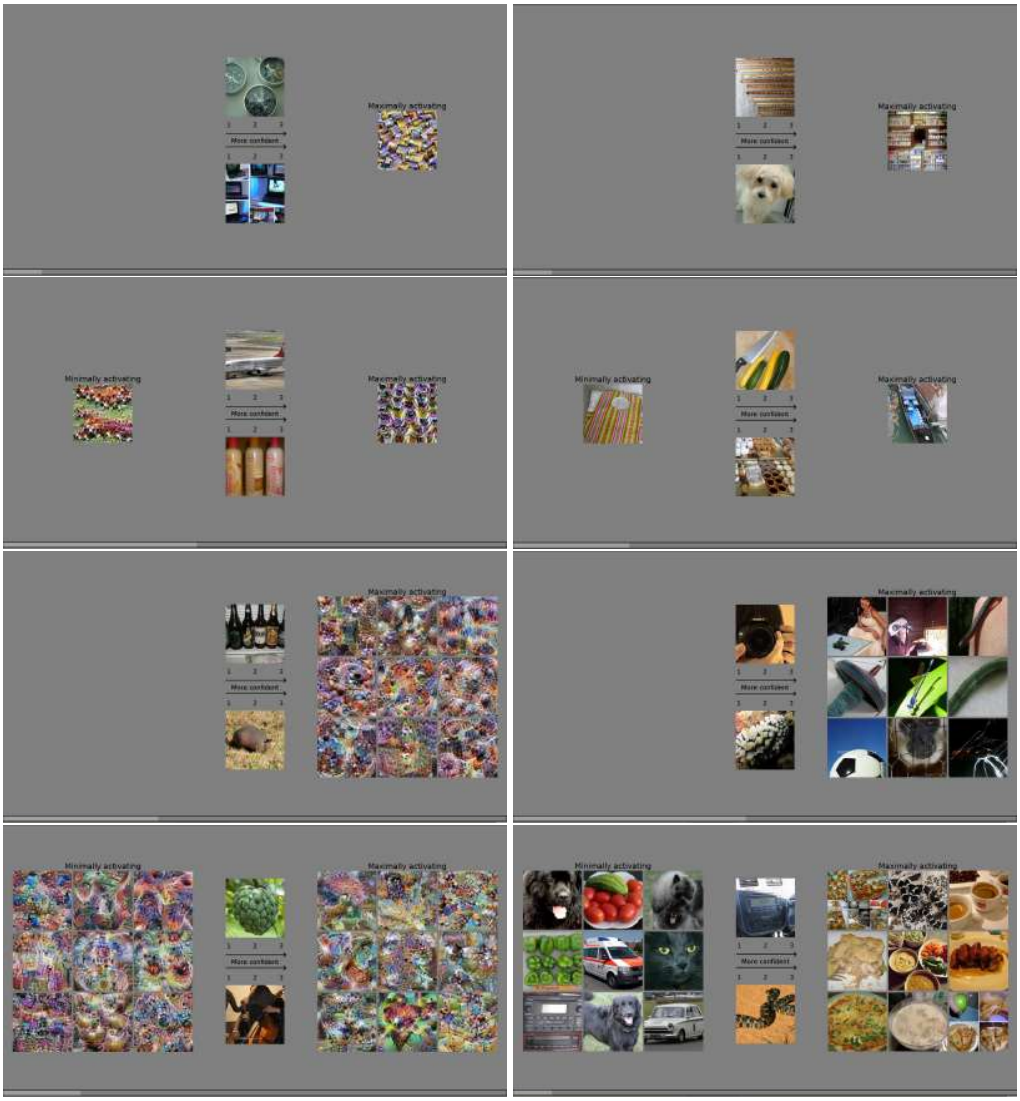


Figure 13: Experiment II: Example trials of the four presentation schemes: Max 1, Min+max 1, Max 9, Min+Max 9. The left column contains synthetic reference images, the right column contains natural reference images.

Figure 14: Trials for intuitiveness judgment. The tested feature maps are from layer mixed3a (channel 43), mixed4b (channel 504) and mixed 5b (channel 17). They are the same in Experiment I and in Experiment II.

Layer	Branch	Feature Map for Batch Block (A-D)			
		A	B	C	D
mixed3a	1×1	25	14	12	53
	3×3	189	97	171	106
	5×5	197	203	212	204
	Pool	227	238	232	247
mixed4a	1×1	68	33	45	17
	3×3	257	355	321	200
	5×5	427	425	429	423
	Pool	486	497	478	506
mixed4c	1×1	94	53	59	95
	3×3	247	237	357	209
	5×5	432	402	400	416
	Pool	496	498	473	497
mixed4e	1×1	231	83	6	89
	3×3	524	323	401	373
	5×5	656	624	642	620
	Pool	816	755	724	783
mixed5b	1×1	119	14	266	300
	3×3	684	592	657	481
	5×5	844	829	839	875
	Pool	1007	913	927	903

Table 3: Feature maps analyzed in Experiment II. Four sets of feature maps (batch blocks A to D) are sampled: For every second layer with an Inception module (5 layers in total), one feature map is randomly selected per branch of the Inception module (1×1, 3×3, 5×5 and pool). For the practice, catch and intuitiveness trials additional randomly chosen feature maps are used.

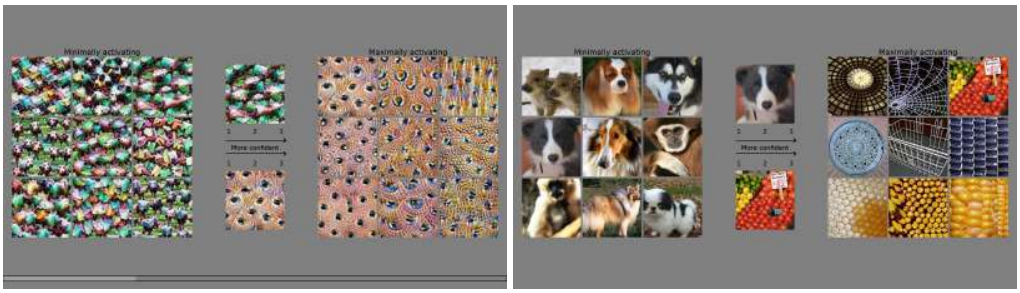


Figure 15: Catch trials. An image from the reference images is copied as a query image, which makes the answer obvious. The purpose of these trials is to integrate a mechanism into the experiment which allows us to check post-hoc whether a participant was still paying attention.

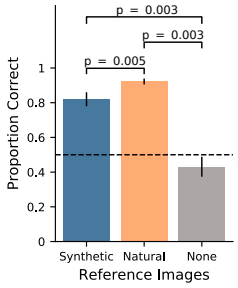
A.2 Details on results

A.2.1 Complementing figures for main results

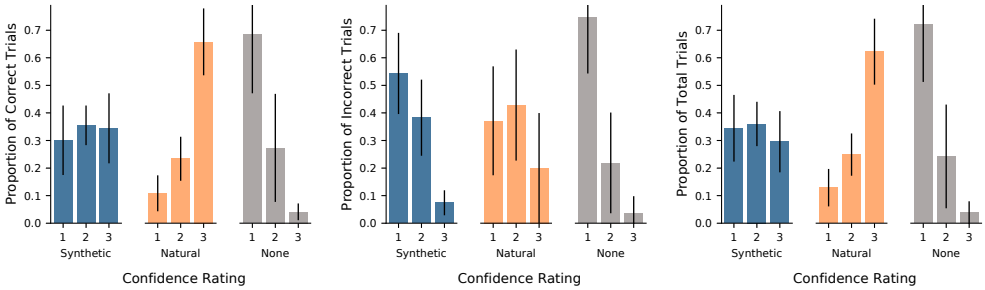
Figures 16 - 21 complement the results and figures presented in Section 4. Here, all experimental conditions are shown.

A.2.2 Details on performance of expert and lay participants

As reported in the main body of the paper, a mixed-effects ANOVA revealed no significant main effect of expert level ($F(1, 21) = 0.6$, $p = 0.44$, between-subjects effect). Further, there is no significant interaction with the reference image type ($F(1, 21) = 0.4, p = 0.53$), and both expert and lay participants show a significant main effect of the reference image type ($F(1, 21) = 230.2, p < 0.001$).



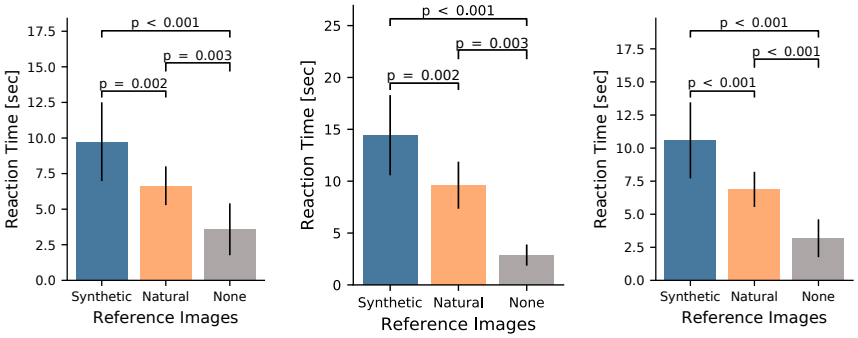
(a) Performance.



(b) Confidence ratings on correctly answered trials.

(c) Confidence ratings on incorrectly answered trials.

(d) Confidence ratings on all trials.

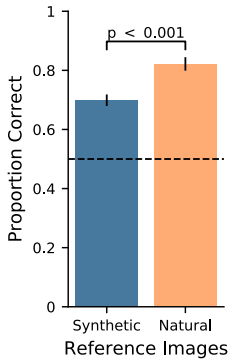


(e) Reaction time on correctly answered trials.

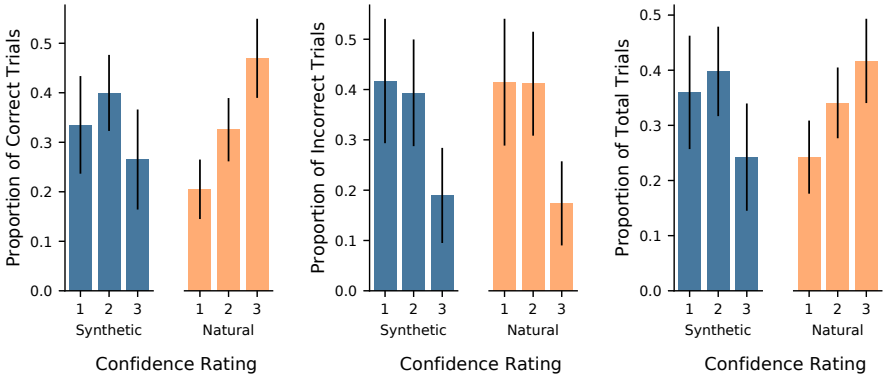
(f) Reaction time on incorrectly answered trials.

(g) Reaction time on all trials.

Figure 16: Task performance (a), distribution of confidence ratings (b-d) and reaction times (e-g) of Experiment I. The p-values are calculated with Wilcoxon sign-rank tests. Note that unlike in the main paper, these figures consistently include the "None" condition. For explanations, see Sec. 4.1.



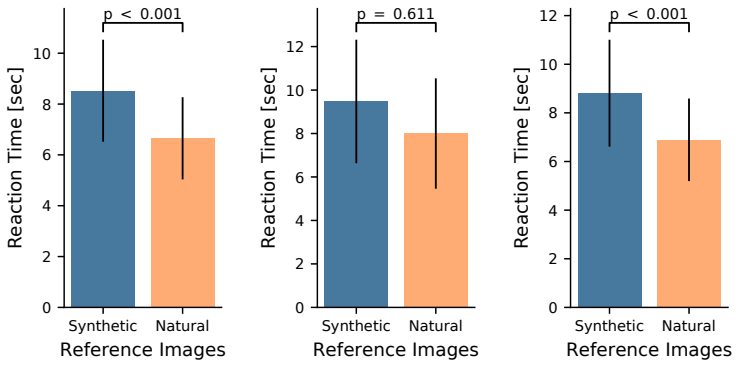
(a) Performance.



(b) Confidence ratings on correctly answered trials.

(c) Confidence ratings on incorrectly answered trials.

(d) Confidence ratings on all trials.

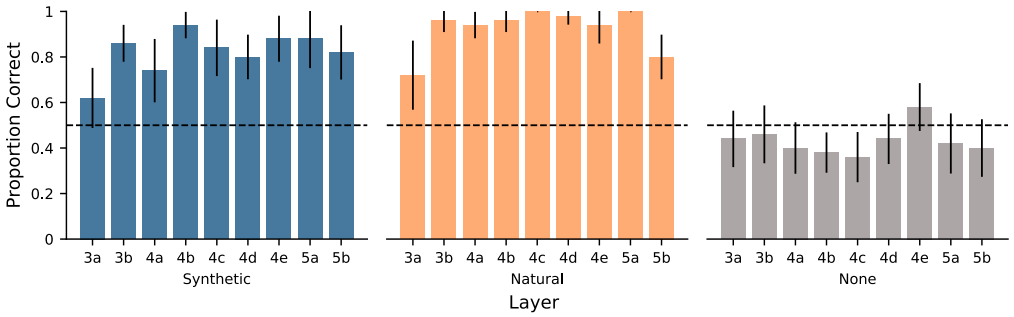


(e) Reaction time on correctly answered trials.

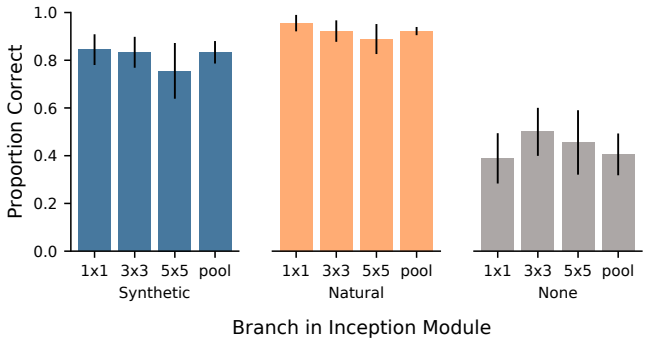
(f) Reaction time on incorrectly answered trials.

(g) Reaction time on all trials.

Figure 17: Task performance (a), distribution of confidence ratings (b-d) and reaction times (e-g) of Experiment II, averaged over expert level and presentation schemes. The p-values are calculated with Wilcoxon sign-rank tests. The results replicate our findings of Experiment I. For explanations on the latter, see Sec. 4.1.

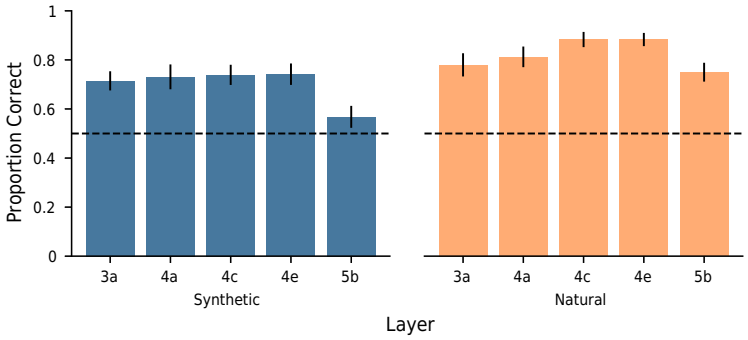


(a) Performance across layers.

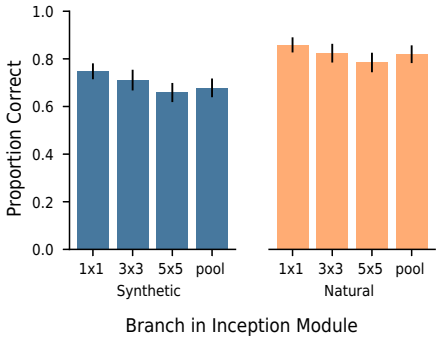


(b) Performance across branches.

Figure 18: High performance across (a) layers and (b) branches of the Inception modules in Experiment I. Note that unlike in the main paper these figures consistently include the "None" condition. For explanations, see Sec. 4.2.



(a) Performance across layers.



(b) Performance across branches in Inception module.

Figure 19: High performance across (a) layers and (b) branches of the Inception modules in Experiment II. Note that only every second layer is tested here (unlike in Experiment I). The results replicate our findings of Experiment I. For explanations, see Sec. 4.2

A.2.3 DETAILS ON PERFORMANCE OF EXPERTS SPLIT BY DIFFERENT LEVELS OF EXPERTISE

Even though Experiment II does not show a significant performance difference for lay and expert participants, it is an open question whether the level of expertise or the background of experts matters. For the data from experts, we hence further divide participants into subgroups according to their expertise (see Fig. 20a-f) and background level (see Fig. 20g-h). Expertise level 1 means that participants are familiar with CNNs, but not feature visualizations; expertise level 2 means that participants have heard of or read about feature visualizations; and expertise level 3 means that participants have used feature visualizations themselves. We note that we also accepted feature visualizations methods other than the one by Olah et al. (2017), e.g. DeepDream (Mordvintsev et al., 2015) for level 2 and 3. Regarding background, we distinguished computational neuroscientists from researchers working on computer vision and / or machine learning. We note that some subgroups only hold one participant and hence may not be representative.

Our data shows varying trends for the three expert levels (see Fig. 20a-f): For synthetic images, performance decreases with increasing expertise in Experiment I, but increases for Experiment II. For natural images, performance first increases for participants of expertise level 2, and then slightly decreases for participants with expertise level 3 - a trend that holds for both Experiment I and II. In the none condition of Experiment I, performance is highest for the participant of expertise level 1, but decreases for participants of expertise level 2, and again slightly increases for expertise level 3.

Regarding expert's different backgrounds, our hypothesis is that many of the computational neuroscientists are very familiar with maximally exciting images for monkeys or rodents, and hence might perform better than pure computer vision / machine learning experts. Fig. 20g-h suggest that this is not the case: The bars for all three reference image types are very similar.

Not finding clear trends in our data between different expertise levels or experts is not surprising as there is even no significant difference between participants whose professional backgrounds are much further apart: lay people vs. people familiar with CNNs.

A.2.4 DETAILS ON PERFORMANCE OF HAND- AND RANDOMLY-PICKED FEATURE MAPS

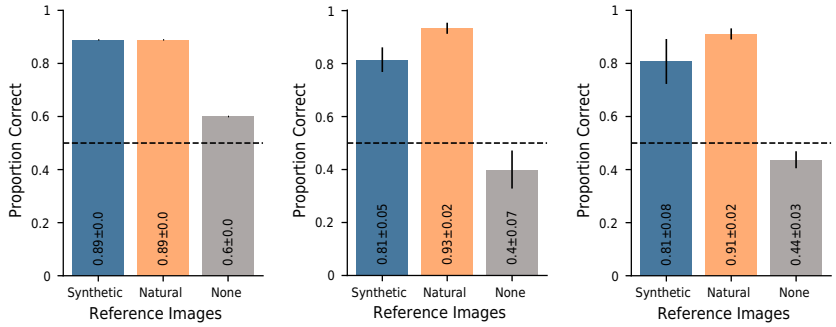
As described in the main body of the paper, pairwise Wilcoxon sign-rank tests reveal no significant differences between hand-picked and randomly-selected feature maps within each reference image type ($Z(9) = 27.5$, $p = 0.59$ for natural reference images and $Z(9) = 41$ $p = 0.18$ for synthetic references). However, marginalizing over reference image type using a repeated measures ANOVA reveals a significant main effect of the feature map selection mode: $F(1, 9) = 6.14, p = 0.035$. Therefore, while there may be a small effect of hand-picking feature maps, our data indicates that this effect, if present, is small.

A.2.5 REPEATED TRIALS

To check the consistency of participants' responses, we repeat six main trials for each of the three tested reference image types at the end of the experiment. Specifically, the six trials correspond to the three highest and three lowest absolute confidence ratings. Results are shown in Fig. 21. We observe consistency to be high for both the synthetic and natural reference image types, and moderate for no reference images (see Fig. 21A). In absolute terms, the largest increase in performance occurs for the none condition; for natural reference images there was also a small increase; for synthetic reference images, there was a slight decrease (see Fig. 21B and C). In the question session after the experiments, many participants reported remembering the repeated trials from the first time.

A.2.6 QUALITATIVE FINDINGS

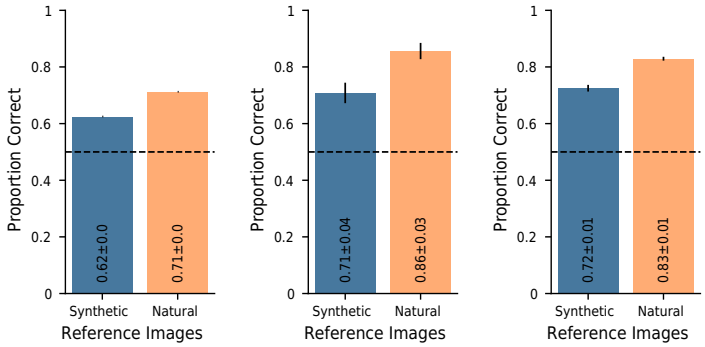
In a qualitative interview conducted after completion of the experiment, participants reported to use a large variety of strategies. Colors, edges, repeated patterns, orientations, small local structures and (small) objects were commonly mentioned. Most but not all participants reported to have adapted their decision strategy throughout the experiment. Especially lay participants from Experiment II emphasized that the trial-by-trial feedback was helpful and that it helped to learn new strategies. As already described in the main text, participants reported that the task difficulty varied greatly; while some trials were simple, others were challenging. A few participants highlighted that the comparison between minimally and maximally activating images was a crucial clue and allowed employing the



(a) Expertise level 1: one participant.

(b) Expertise level 2: six participants.

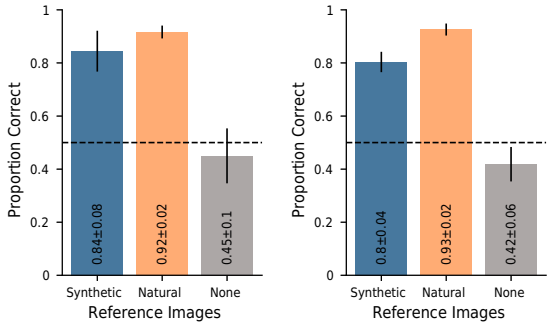
(c) Expertise level 3: three participants.



(d) Expertise level 1: one participant.

(e) Expertise level 2: six participants.

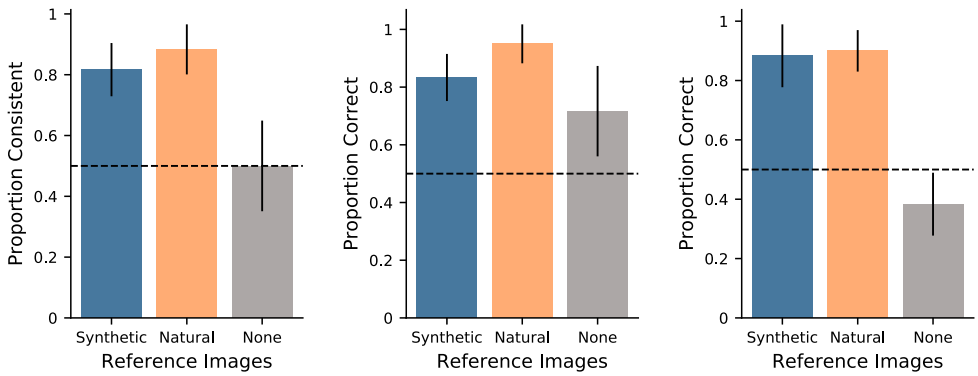
(f) Expertise level 3: three participants.



(g) Computational Neuroscience background: six participants.

(h) Computer Vision / Machine Learning background: four participants.

Figure 20: Performance of experts split by different levels of expertise: The first (second) row shows the data of Experiment I (II) split up by different levels of familiarity with CNNs and feature visualizations. The third row shows the data of Experiment I split up by different backgrounds.



(a) Proportion of trials that were answered the same upon repetition.

(b) Performance for repeated trials upon repetition.

(c) Performance for repeated trials when first shown.

Figure 21: Repeated trials in Experiment I.

exclusion criterion: If the minimally activating query image was easily identifiable, the choice of the maximally activating query image was trivial. This aspect motivated us to conduct an additional experiment where the presentation scheme was varied (Experiment II).

A.2.7 By-Feature-Map Analysis

For Experiment I, we look at each feature map separately and analyze which feature maps participants find easy and which they find difficult. Further, we investigate commonalities and differences between feature maps. We note that the data for this analysis relies on only 10 responses for each feature map and hence may be noisy.

In Fig. 22, we show the number of correct answers split up by reference image type. The patterns look similar to the trend in Fig. 4: Across most layers, there is no clearly identifiable trend that feature maps of a certain network depth would be easier or more difficult; only the lowest (3a) and the highest layer (5b) seem slightly more difficult for both the synthetic and the natural reference images.

Easy Feature Maps When feature maps are easy (synthetic: $10/10$, natural: $10/10$ correct responses), their features seem to correspond to clear object parts (e.g. dogs vs. humans, food vs. cats), or shapes (e.g. round vs. edgy (see Supplementary Material Fig. 2- 5)). In Fig. 23, we show the query as well as natural and synthetic reference images for one such easy feature map for one participant. For the images shown to two more participants, see Supplementary Material Fig. 1. Other relatively easy feature maps (where eight to ten participants choose the correct query image for both reference image types) additionally contained other low level cues such as color or texture (see Supplementary Material Fig. 4-5).

Difficult Feature Maps The most difficult feature maps for synthetic and natural reference images are displayed in Fig. 24. Only four participants predicted the correct query image. Interestingly, the other reference image type was much more easily predictable for both feature maps: Nine out of ten participants correctly simulated the network's decision. Our impression is that the reason for these feature maps being so difficult in one reference condition is the diversity in the images. In the case of synthetic reference images, we also consider identifying a concept difficult and consequently are unsure what to compare.

From studying several feature maps, our impression is that one or more of the following aspects make feature maps difficult to interpret:

- Reference images are diverse (see Fig. 24a for synthetic reference images and d for natural reference images)

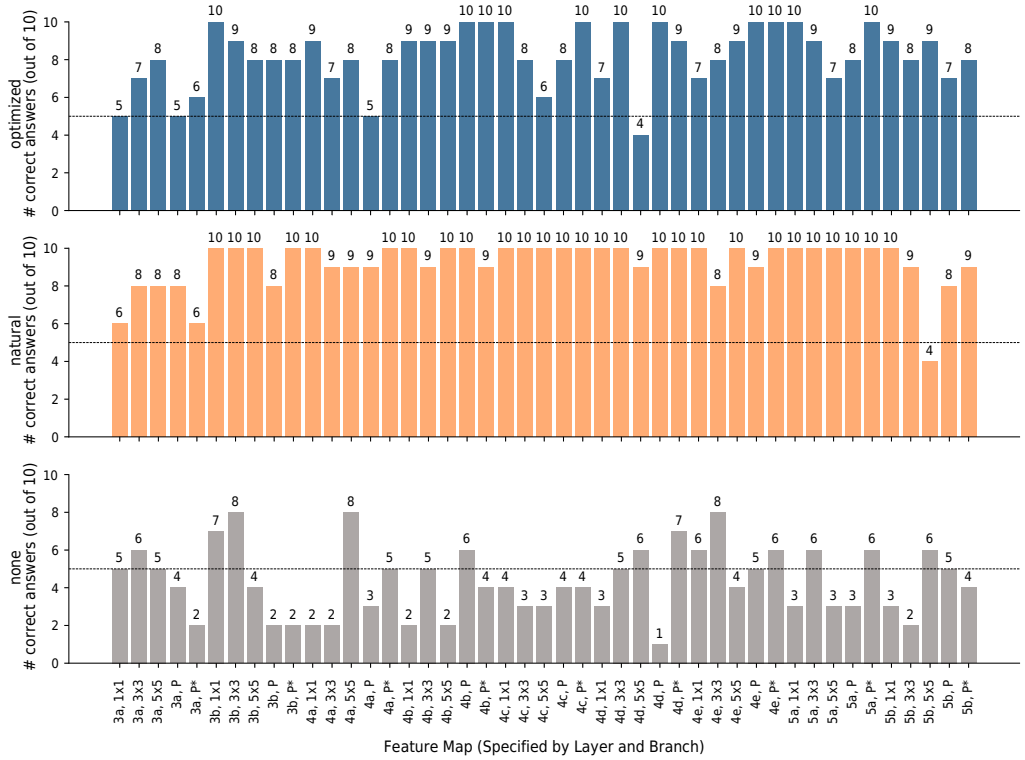


Figure 22: Data for Experiment I split up by feature maps: For each reference image type, the number of correct answers (out of ten) is shown. There is no clear trend that certain feature maps would be easier or more difficult.

Figure 23: An easy feature map (here: 5a, pool*) from Experiment I where all participants answered correctly for both synthetic and natural reference images. The shown stimuli were shown to participant 1, for stimuli shown to participant 2 and 3, see Supplementary Material Fig 1.

- The common feature(s) seem to not correspond to common human concepts (see Fig. 24a and c)
- Conflicting information, i.e. commonalities can be found between one query image and both the minimal and maximal reference images (see Fig. 25a: eyes and extremity-like structure in synthetic min reference images vs. eyes and earth-colors in synthetic max reference images - both could be considered similar to the max query image of a frog)
- Very small object parts such as eyes or round, earth-colored shapes seem to be the decisive features (see Fig. 25a and b)
- Low level cues such as the orientation of lines appear random in the synthetic reference images[9] (see Fig. 26a)

Finally, when we speak bluntly, we are often surprised that participants identified the correct image — the reasons for this are unclear to us (see for example Supplementary Material Fig. 6-7).

A.2.8 High quality data as shown by high performance on catch trials

We integrate a mechanism to probe the quality of our data: In *catch trials*, the correct answer is trivial and hence incorrect answers might suggest the exclusion of specific trial blocks (for details, see Sec. A.1.1). Fortunately, very few trials are missed: In Experiment I, only two (out of ten) participants miss one trial each (i.e. a total of 2 out of 180 catch trials were missed); in Experiment II, five participants miss one trial and four participants miss two trials (i.e. a total of 13 out of 736 catch

[9]We expected lower layers to be easier than higher layers for synthetic reference images, but our data showed that this was not the case (see Fig. 22. We can imagine that the diversity term as well as the non-custom hyper-parameters contribute to these sub-optimal images.

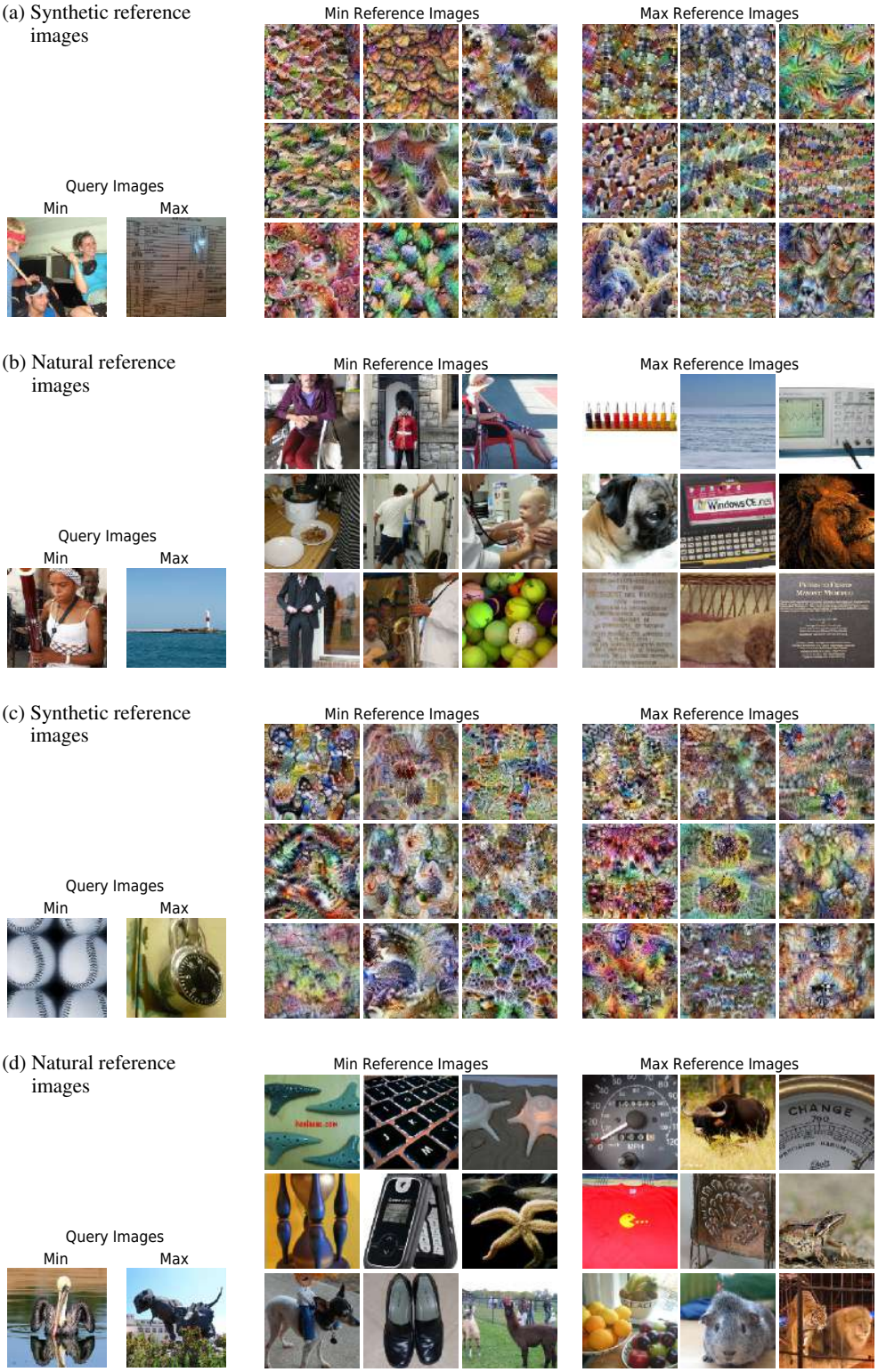


Figure 24: Two difficult feature maps (4d, 5x5 in a and b; 5b, 5x5 in c and d) from Experiment I where only four participants answered correctly for synthetic (a and b) and natural (c and d) reference images. The displayed stimuli were shown to participant 1, for stimuli shown to participant 2 (3), see Supplementary Material Fig. 8 (9).

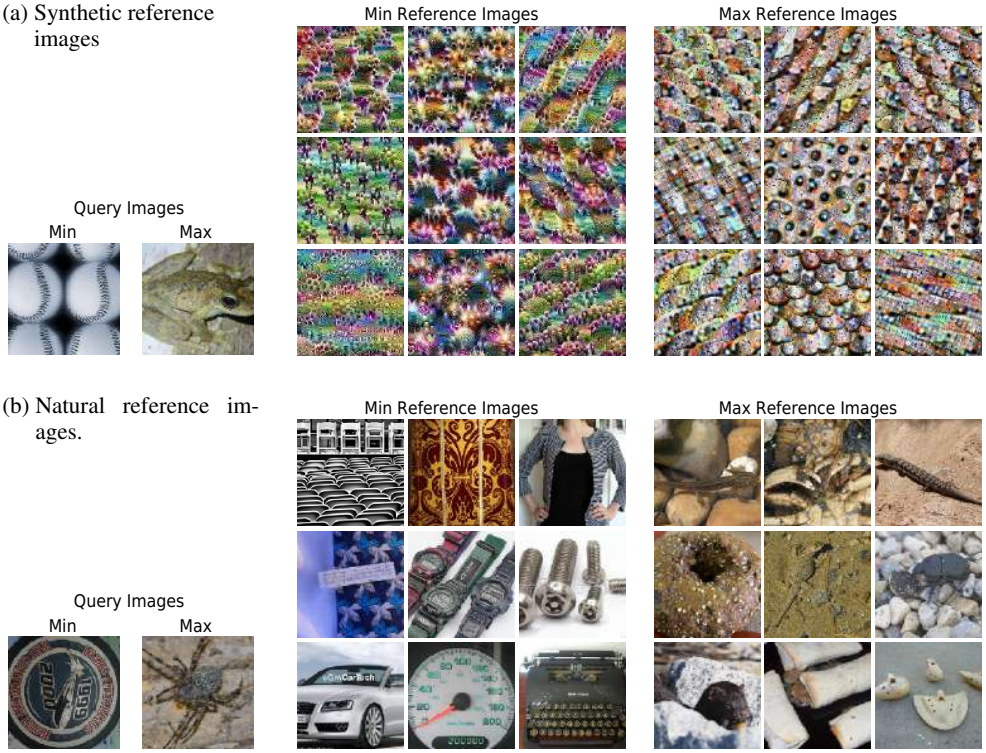


Figure 25: A feature map (here: 4a, Pool) from Experiment I where the feature is small (eyes) and a participant might perceive conflicting information (eyes and extremity-like structure in min reference images vs. eyes and earth-colors in max reference images). In this specific example, eight (nine) out of ten participants gave the correct answer for this feature map given synthetic (natural) reference images. The displayed stimuli were shown to participant 1, for stimuli shown to participant 2 and 3, see Supplementary Material Fig. 10.

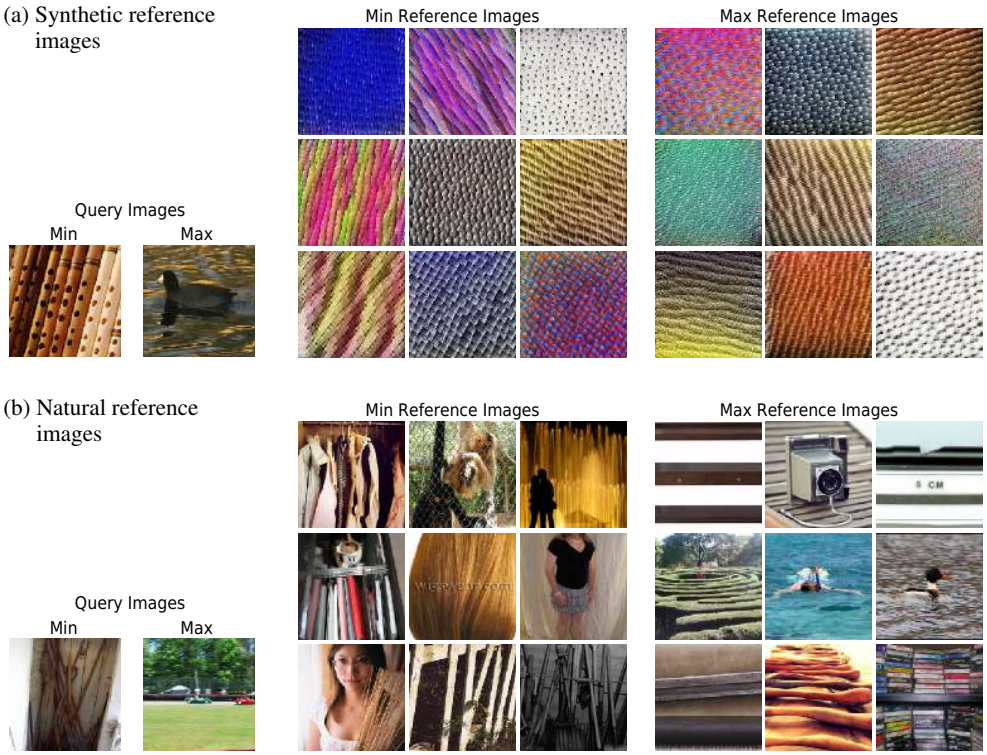


Figure 26: A feature map from a low layer (here: 3a, 3x3) from Experiment I where the feature seems to be a low level cue (horizontal vs. vertical striped) that is surprisingly clear in the natural, but surprisingly unclear in the synthetic reference images. In this specific example, seven (eight) out of ten subjects gave the correct answer for this feature map given synthetic (natural) reference images. The displayed stimuli were shown to participant 1, for stimuli shown to participant 2 and 3, see Supplementary Material Fig. 11.

trials were missed). As this indicates that our data is of high quality, we do not perform the analysis with excluded trials as we expect to find the same results.

[8] Baseline condition.

[9] Metrics of explanation quality computed without human judgment are inconclusive and do not correspond to human rankings.

[10] Task has an additional "I don't know"-option for confidence rating.

[11] Comparison is only performed between methods but no absolute measure of interpretability for a method is obtained.

A.3 DETAILS ON RELATED WORK

Paper	Analyzes Intermediate Features?	Explanation Methods Analyzed	Results: Explanation helpful?	Results: Confidence/Trust
Ours	**yes**	• Feature Visualization • natural images[8] • no explanation[8]	yes	• high variance in confidence ratings • natural images are more helpful
Biessmann & Refiano (2019)	no	• LRP • Guided Backprop • simple gradient[8]	yes	• highest confidence for guided backprop[9]
Chu et al. (2020)	no	• prediction + gradients • prediction[8] • no information[8]	no	• faulty explanations do not decrease trust
Shen & Huan (2020)	no	• Extremal Perturb • GradCAM • SmoothGrad • no explanation[8]	no	• -
Jeyakumar et al. (2020)	no	• LIME • Anchor • SHAP • Saliency Maps • Grad-CAM++ • Ex-Matchina	unclear[11]	• -
Alqaraawi et al. (2020)	no	• LRP • classification scores • no explanation[8]	yes	• confidence similar across conditions
Chandra-sekaran et al. (2017)	no	• prediction confidence • attention maps • Grad-CAM • no explanation[8]	no	• -
Schmidt & Biessmann (2019)	no	• LIME • custom method • random/no explanation[8]	yes	• humans trust own judgement regardless explanations, except in one condition
Hase & Bansal (2020)	no	• LIME • Prototype • Anchor • Decision Boundary • combination of all 4	partly	• high variance in helpfulness • helpfulness cannot predict user performance
Kumaraku-lasinghe et al. (2020)	no	• LIME	yes	• fairly high trust and reliance
Ribeiro et al. (2018)	no	• LIME • Anchor • no explanation[8]	yes	• high confidence for Anchor • low for LIME & no explanation
Alufaisan et al. (2020)	no	• prediction + Anchor • prediction[8] • no information[8]	partly	• explanations do not increase confidence
Ramamurthy et al. (2020)	no	• MAME • SP-LIME • Two Step	• unclear[11]	• users can adjust MAME which increased trust
Dieber & Kirrane (2020)	no	• LIME	partly	• -
Dinu et al. (2020)	no	• SHAP • ridge • lasso • random explanation[8]	partly	• no statement on confidence ratings

Paper	Experimental Setup			
	Dataset	**Task**	**Participants**	**Collected Data**
Ours	• natural images (ImageNet)	• CNN activation classification	• experts • laypeople	• decision • confidence • reaction time • post-hoc evaluation
Biessmann & Refiano (2019)	• face images (Cohn-Kanade)	• 2-way classification[10]	• laypeople	• decision • confidence • reaction time
Chu et al. (2020)	• face images (APPA-REAL)	• age regression	• laypeople	• decision • trust • reaction time • post-hoc evaluation
Shen & Huan (2020)	• natural images (ImageNet)	• model error identification	• laypeople	• decision
Jeyakumar et al. (2020)	• natural images (CIFAR-10) • text (Sentiment140) • audio (Speech Commands) • sensory data (MIT-BIH Arrhythmia)	• preference for one out of two explanation methods	• laypeople	• decision
Alqaraawi et al. (2020)	• natural images (Pascal VOC)	• classification	• technical background (neither lay nor expert)	• decision • confidence • free answer on features
Chandra-sekaran et al. (2017)	• VQA (visualqa.org)	• model error identification • regression	• laypeople	• decision
Schmidt & Biessmann (2019)	• book categories • Movie reviews (IMDb)	• 9-/2-way classification	• laypeople	• decision • reaction time • trust
Hase & Bansal (2020)	• movie reviews (Movie Review) • tabular (Adult)	• 2-way classification	• experts	• decision • helpfulness rating • explanation helpfulness
Kumaraku-lasinghe et al. (2020)	• tabular (Patient data)	• 2-way classification	• experts	• decision • feature ranking • satisfaction • questionnaire
Ribeiro et al. (2018)	• tabular (Adult, rcdv)	• 2-way classification[10] • VQA	• experts	• decision • reaction time • confidence
Alufaisan et al. (2020)	• tabular (COMPAS, Census Income)	• 2-way classification	• laypeople	• decision • confidence • reaction time
Ramamurthy et al. (2020)	• tabular (HELOC, pump failure)	• 2-way classification	• experts • laypeople	• decision
Dieber & Kirrane (2020)	• tabular (Rain in Australia)	• interview	• laypeople • experts	• how interpretable LIME output is
Dinu et al. (2020)	• tabular (Airbnb price listings)	• interview	• laypeople	• decision: which model would perform better in practice • confidence

Table 4: Overview of publications that evaluate explanation methods in human experiments. Note that the table already starts on the previous page and that the footnotes are displayed on page 39.

SUPPLEMENTARY MATERIAL FOR *Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization*

1 EXTENDED: BY-FEATURE-MAP ANALYSIS

On the following pages, we provide more images of trials that participant two and three saw during Experiment I.

Figure 1: An easy feature map (here: 5a, pool*) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. Our impression is that the decisive feature (dog vs. bird) is well understandable. The shown stimuli were shown to participants two (a and b) and three (c and d); for stimuli shown to participant one, see Appendix Fig. 23.

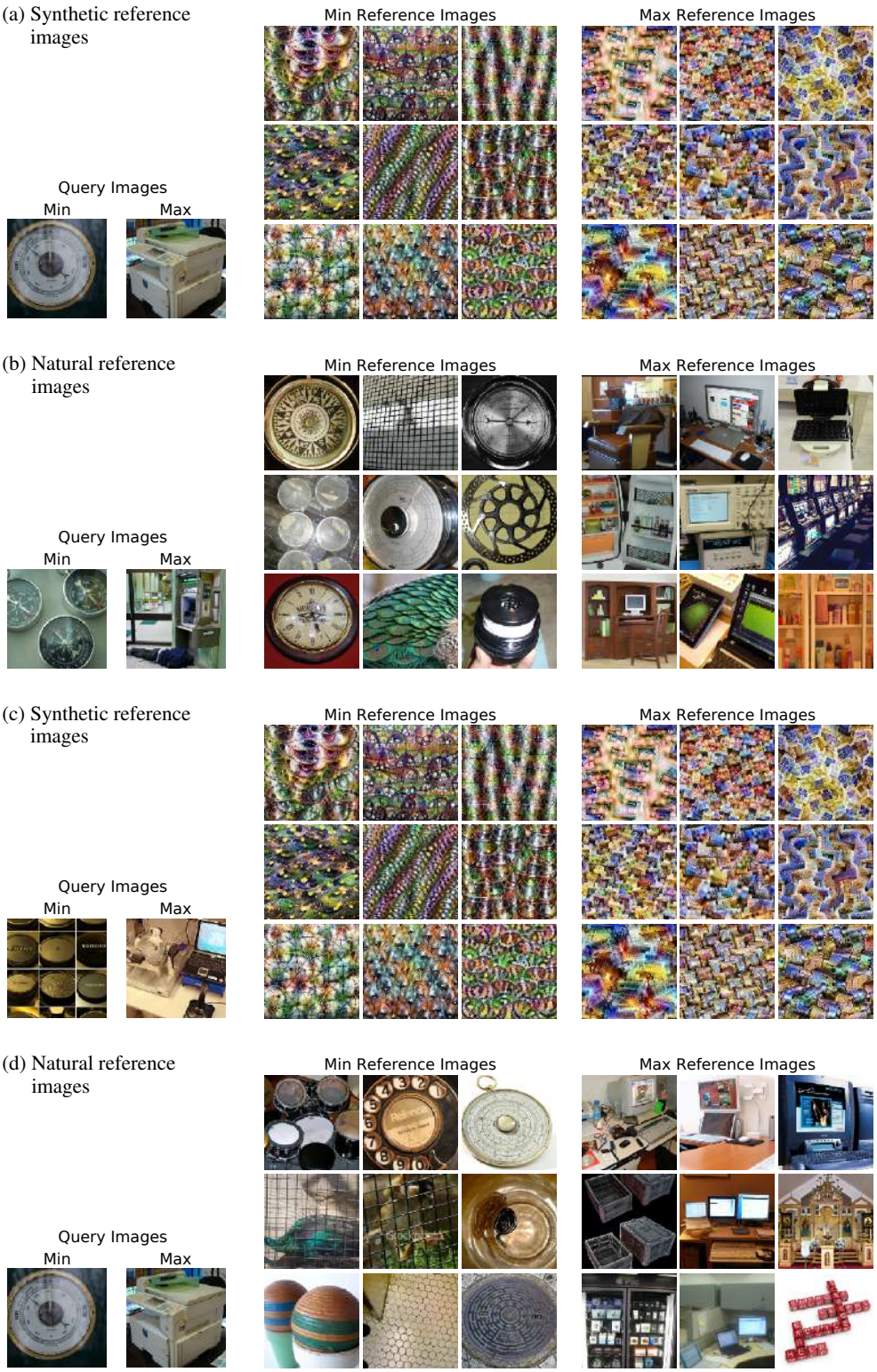


Figure 2: Another easy feature map (here: 4c, 1x1) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. The decisive feature seems to be round (minimal) vs. edgy (max). The shown stimuli were shown to participants one (a and b) and two (c and d).

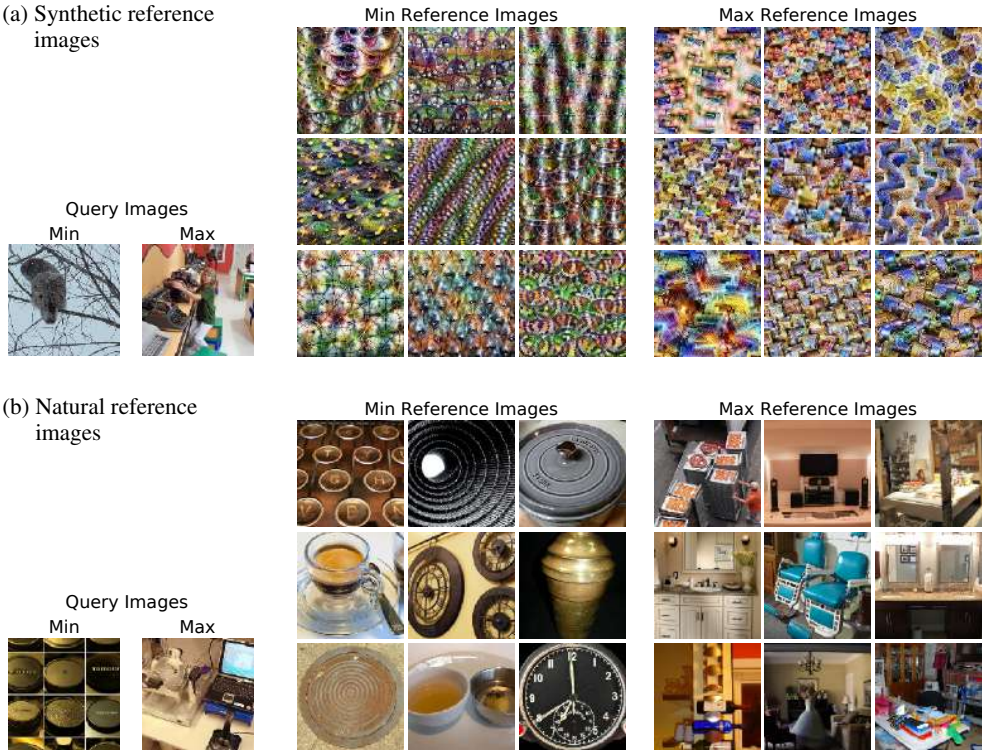


Figure 3: Another easy feature map (here: 4c, 1x1) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. The decisive feature seems to be round (minimal) vs. edgy (max). The shown stimuli were shown to participant three.

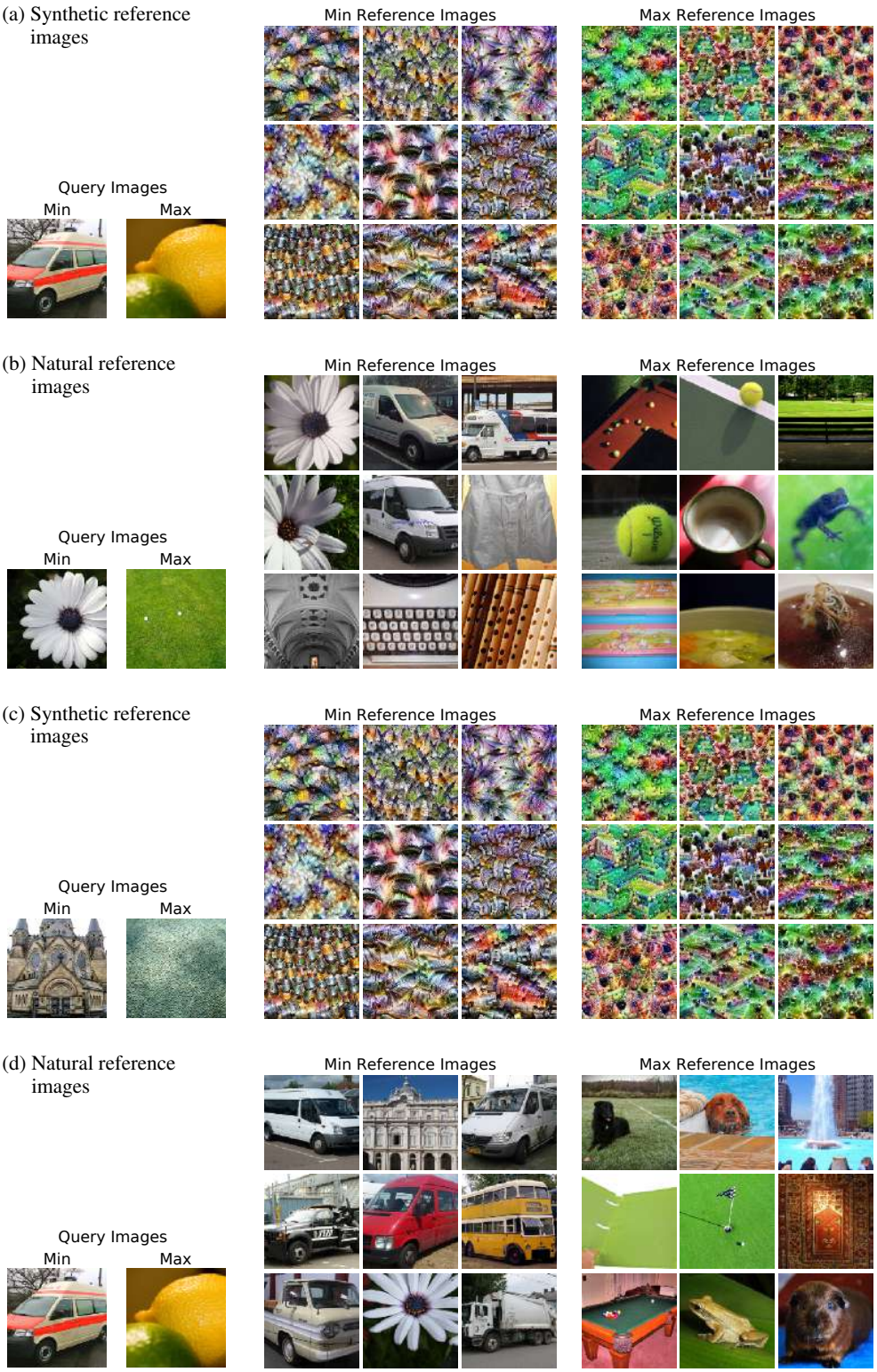


Figure 4: Another easy feature map (here: 4d, 3x3) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. The decisive maximal feature seems to have to do with green color and round shapes. The shown stimuli were shown to participants one (a and b) and two (c and d).

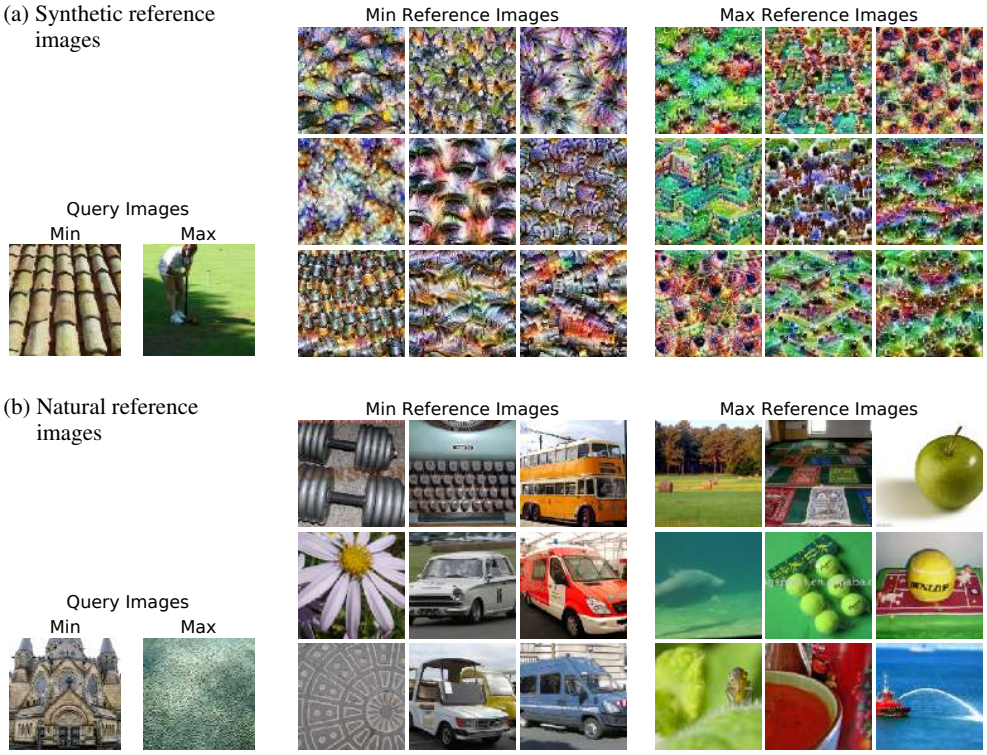


Figure 5: Another easy feature map (here: 4d, 3x3) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. The decisive maximal feature seems to have to do with green color and round shapes. The shown stimuli were shown to participant three.

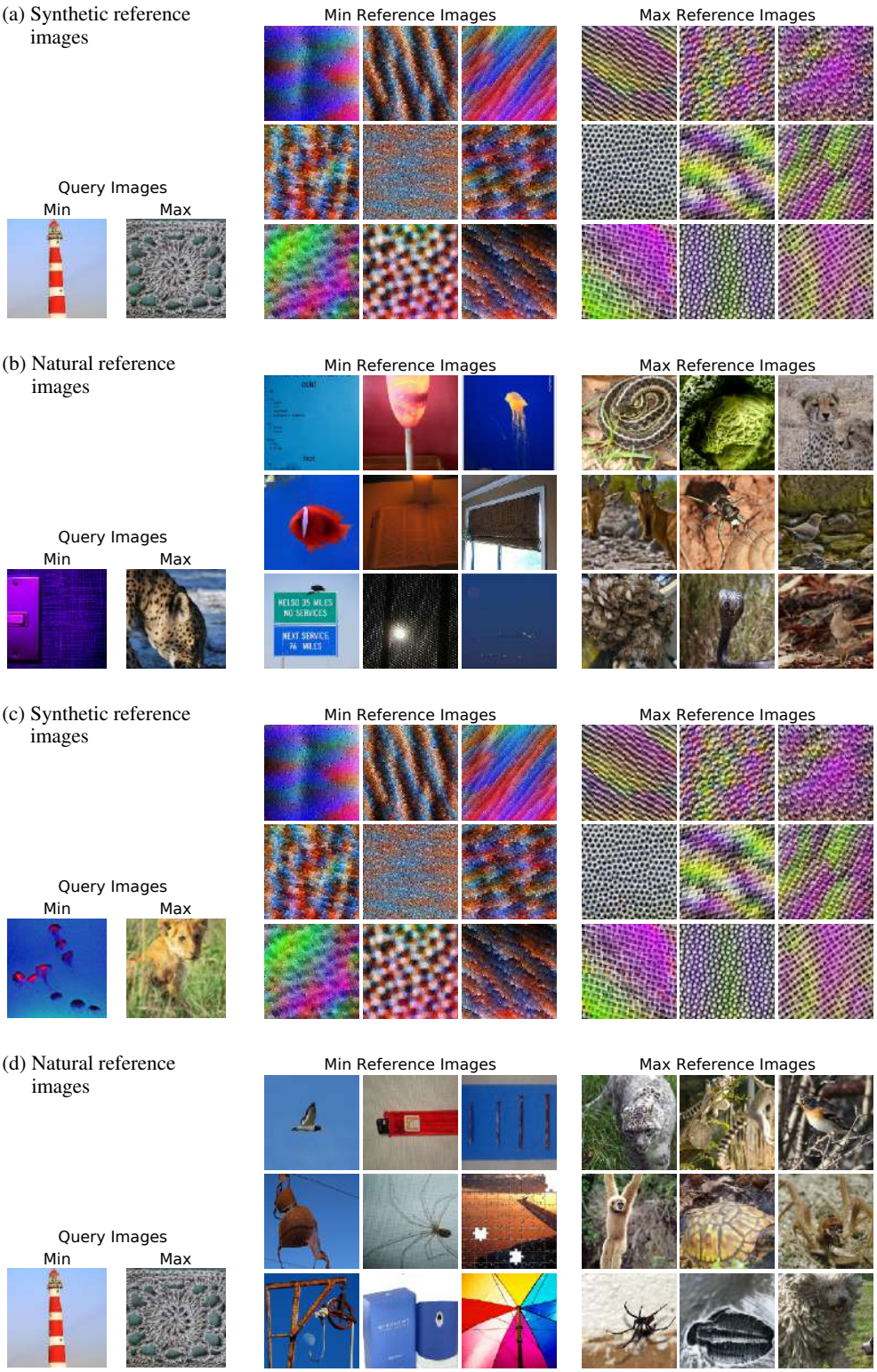


Figure 6: Another easy feature map (here: 3b, 1x1) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. We are surprised that all the trials for synthetic reference images were answered correctly; to us, the decisive feature is not easily identifiable. The shown stimuli were shown to participants one (a and b) and two (c and d).

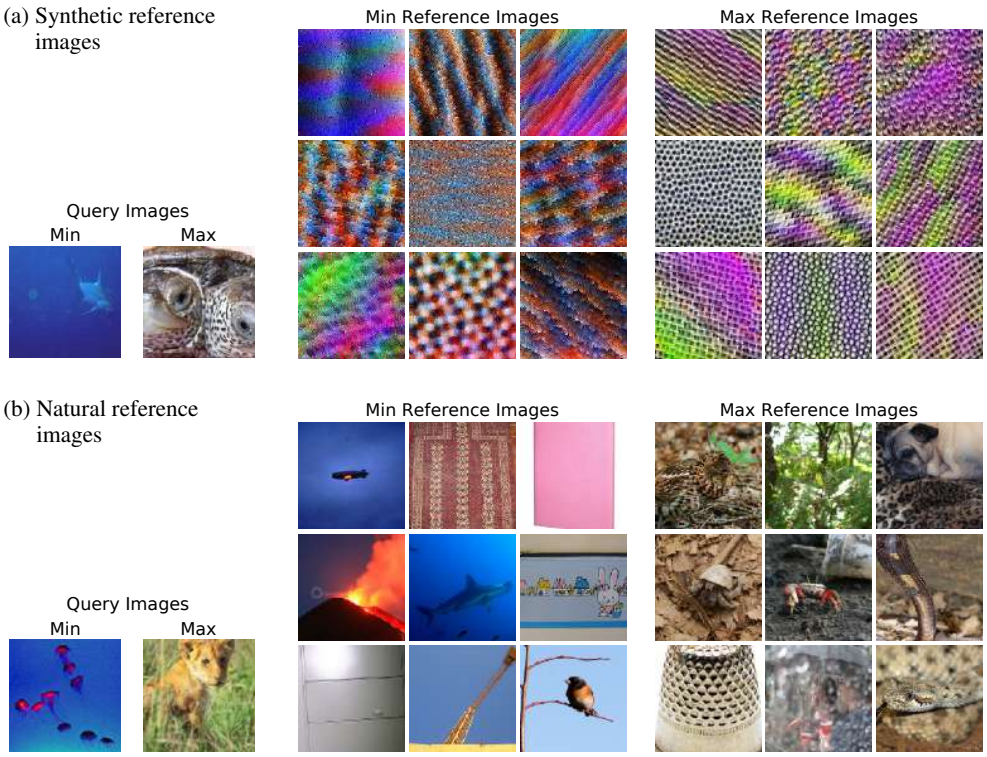


Figure 7: Another easy feature map (here: 3b, 1x1) from Experiment I where all subjects answered correctly for both synthetic and natural reference images. We are surprised that all the trials for synthetic reference images were answered correctly; to us, the decisive feature is not easily identifiable. The shown stimuli were shown to participant three.

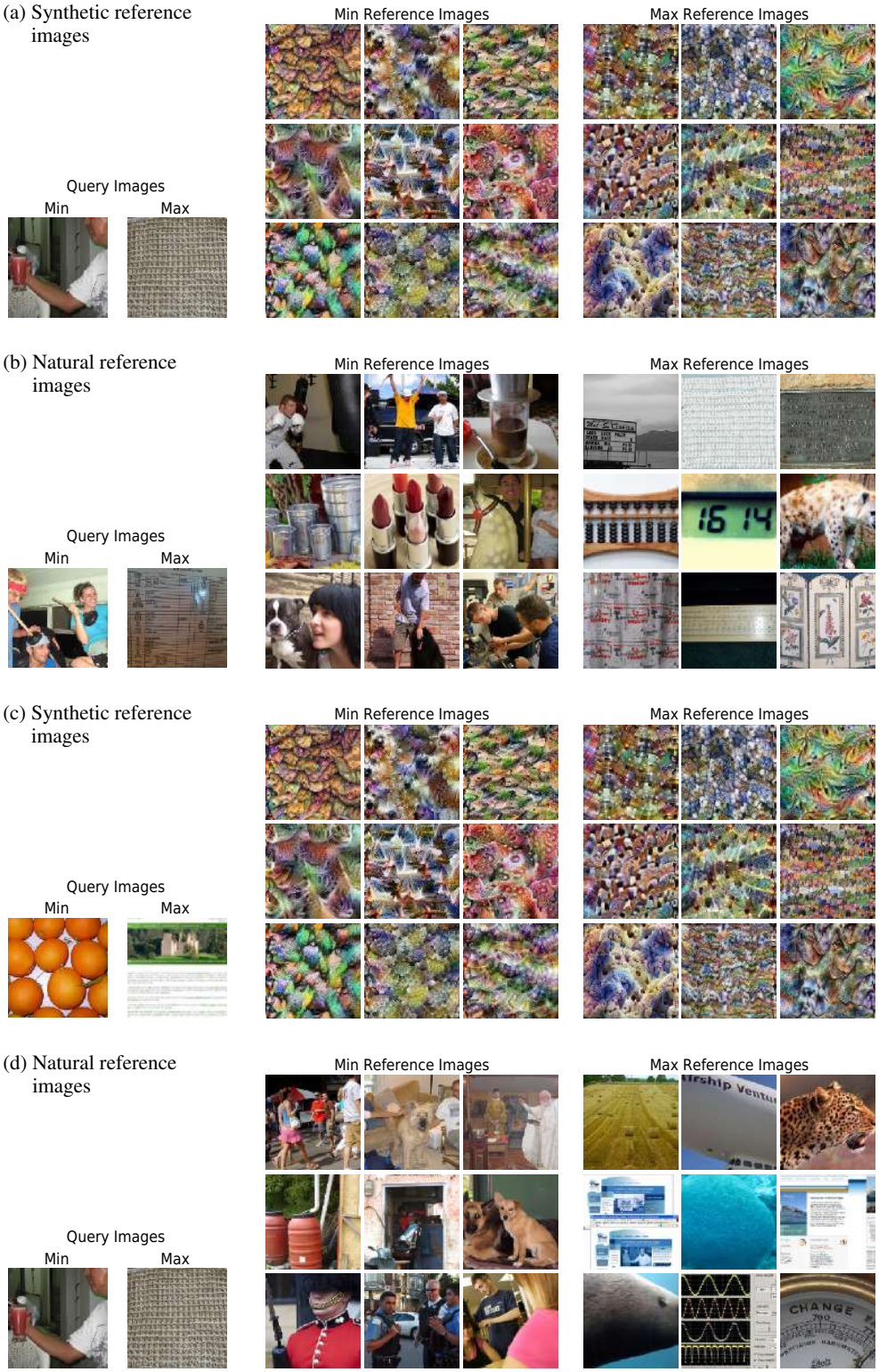


Figure 8: Example of a difficult feature map (4d, 5x5) from Experiment I where only four subjects answered correctly for synthetic reference images. The displayed stimuli were shown to participant two (a, b) and three (c, d), for stimuli shown to participant one, see Appendix Fig. 24.

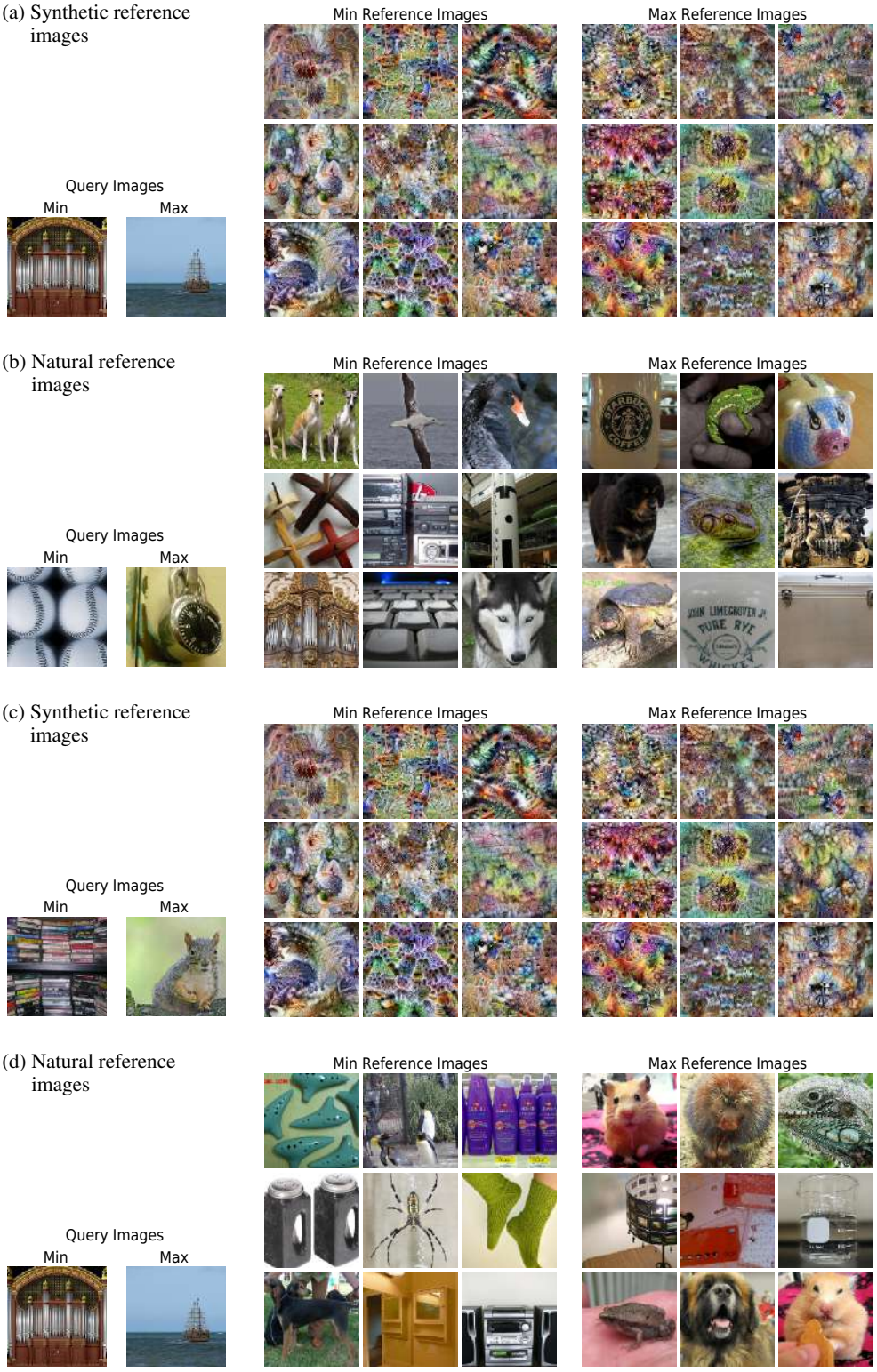


Figure 9: Example of a difficult feature map (5b, 5x5) from Experiment I where only four subjects answered correctly for natural (e-h) reference images. The displayed stimuli were shown to participant two (a, b) and three (c, d), for stimuli shown to participant one, see Appendix Fig. 24.

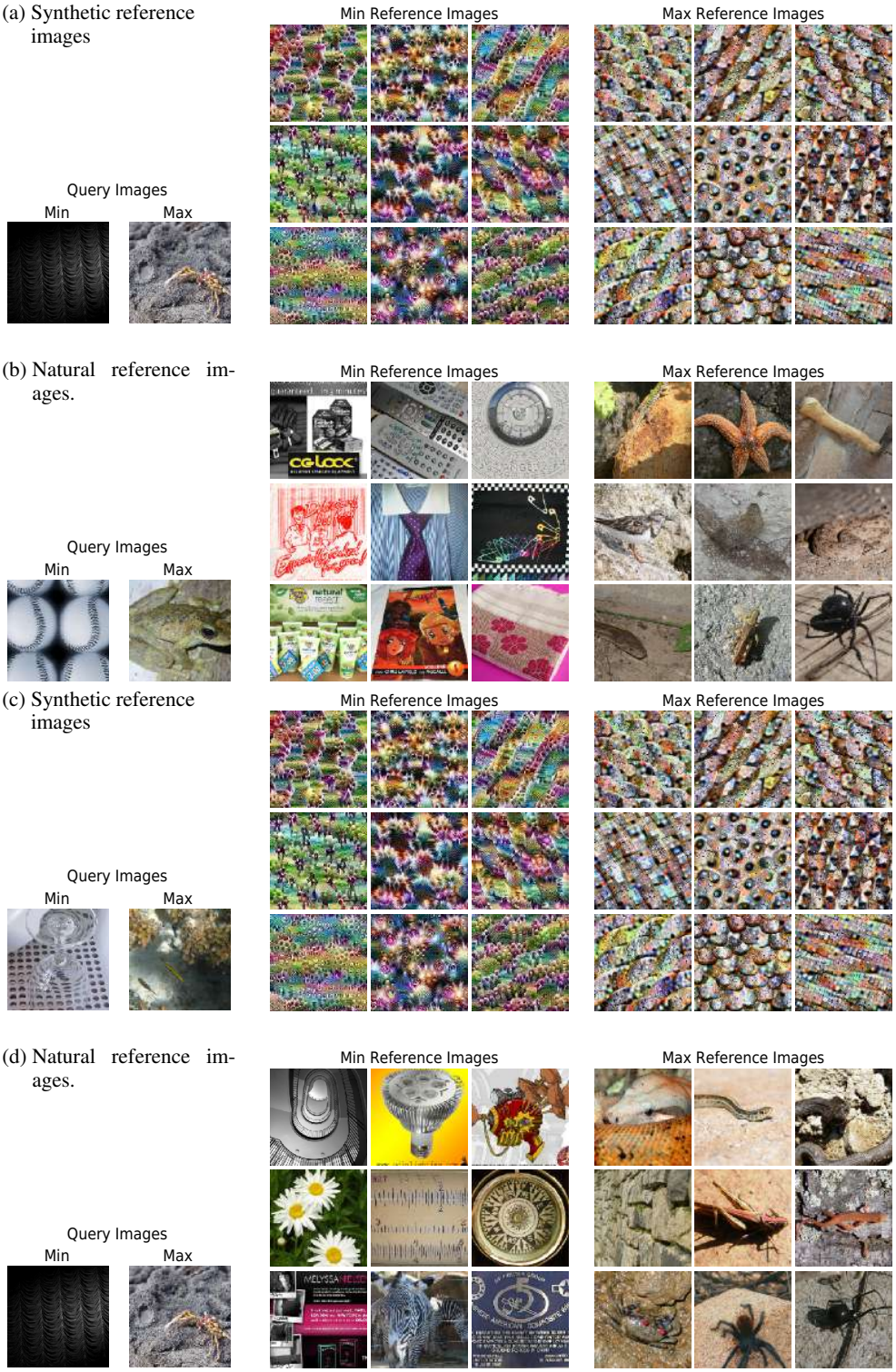


Figure 10: A feature map (here: 4a, Pool) from Experiment I where the feature is small (eyes) and a participant might perceive conflicting information (eyes and extremity-like structure in min reference images vs. eyes and earth-colors in max reference images). In this specific example, eight (nine) out of ten subjects gave the correct answer for this feature map given synthetic (natural) reference images. The displayed stimuli were shown to participants two (a and b) and three (c and d), for stimuli shown to participant one, see Appendix Fig. 25.

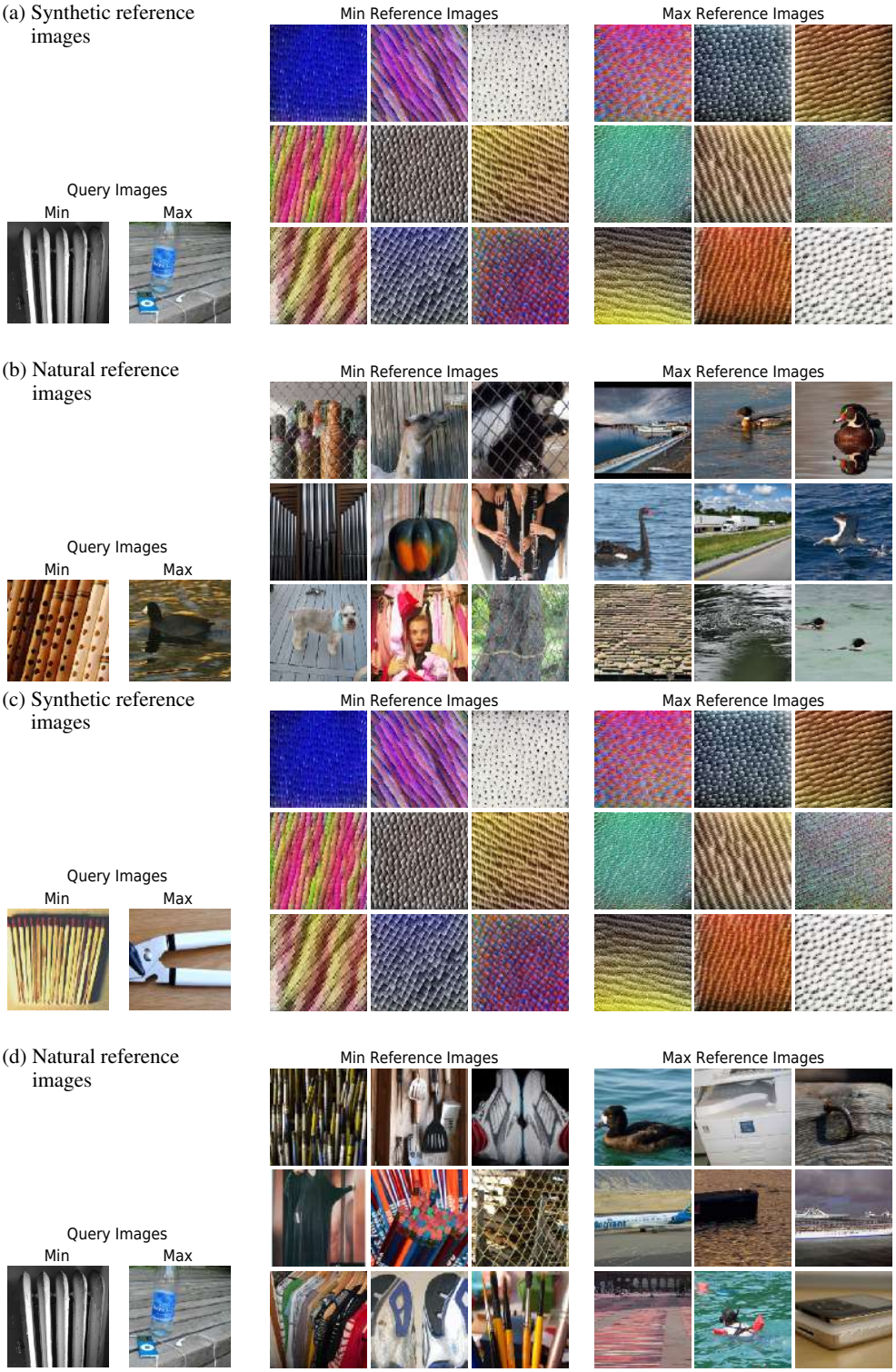


Figure 11: A feature map from a low layer (here: 3a, 3x3) from Experiment I where the feature seems to be a low level cue (horizontal vs. vertical striped) that is surprisingly clear in the natural, but surprisingly unclear in the synthetic reference images. In this specific example, seven (eight) out of ten subjects gave the correct answer for this feature map given synthetic (natural) reference images. The displayed stimuli were shown to participants two (a and b) and three (c and d), for stimuli shown to participant one, see Appendix Fig. 26.

Publication 3: How well do feature visualizations support causal understanding of CNN activations?

Roland Simon Zimmermann*, Judy Borowski*, Robert Geirhos, Matthias Bethge‡, Thomas S.A. Wallis‡, Wieland Brendel‡. NeurIPS, 2021.

Contributions:

"The idea to test how well feature visualizations support causal understanding of CNN activations was born out of several reviewer and audience comments on our previous paper (Borowski et al., 2021). The first idea of how to test this in a psychophysical experiment came from TSAW. JB led the project. JB, RSZ, WB and TSAW jointly improved the experimental set-up with input from MB and RG. RSZ led and JB helped with the implementation and execution of the experiment; JB led and RSZ contributed to the generation of stimuli. RSZ and JB both coded the baselines, and TSAW guided the replication experiment with statistical power simulations. The data analysis was performed by RSZ and JB with advice and feedback from RG, TSAW, WB and MB. TSAW and WB provided day-to-day supervision. While JB and RSZ created the first draft of the manuscript, RG and TSAW heavily edited the manuscript and all authors contributed to the final version."

An earlier version of this work was presented as a poster at the ICML Workshop *Theoretic Foundation, Criticism, and Application Trend of Explainable AI* (2021) under the same title.

How Well do Feature Visualizations Support Causal Understanding of CNN Activations?

Roland S. Zimmermann[*1] **Judy Borowski**[*1]

Robert Geirhos[1] **Matthias Bethge**[†1] **Thomas S. A. Wallis**[†2] **Wieland Brendel**[†1]

[1] Tübingen AI Center, University of Tübingen, Germany.
[2] Institute of Psychology and Centre for Cognitive Science, Technical University of Darmstadt, Germany.
[*] Shared first authorship, determined by coin flip. firstname.lastname@uni-tuebingen.de
[†] Joint supervision.

Abstract

A precise understanding of why units in an artificial network respond to certain stimuli would constitute a big step towards explainable artificial intelligence. One widely used approach towards this goal is to visualize unit responses via activation maximization. These synthetic feature visualizations are purported to provide humans with precise information about the image features that *cause* a unit to be activated — an advantage over other alternatives like strongly activating natural dataset samples. If humans indeed gain causal insight from visualizations, this should enable them to predict the effect of an intervention, such as how occluding a certain patch of the image (say, a dog's head) changes a unit's activation. Here, we test this hypothesis by asking humans to decide which of two square occlusions causes a larger change to a unit's activation. Both a large-scale crowdsourced experiment and measurements with experts show that on average the extremely activating feature visualizations by Olah et al. [40] indeed help humans on this task ($68 \pm 4\,\%$ accuracy; baseline performance without any visualizations is $60 \pm 3\,\%$). However, they do not provide any substantial advantage over other visualizations (such as e.g. dataset samples), which yield similar performance ($66 \pm 3\,\%$ to $67 \pm 3\,\%$ accuracy). Taken together, we propose an objective psychophysical task to quantify the benefit of unit-level interpretability methods for humans, and find no evidence that a widely-used feature visualization method provides humans with better "causal understanding" of unit activations than simple alternative visualizations.

1 Introduction

It is hard to trust a black-box algorithm, and it is hard to deploy an algorithm if one does not trust its output. Many of today's best-performing machine learning models, deep convolutional neural networks (CNNs), are also among the most mysterious ones with regards to their internal information processing. CNNs typically consist of dozens of layers with hundreds or thousands of units that distributively process and aggregate information until they reach their final decision at the topmost layer. Shedding light onto the inner workings of deep convolutional neural networks has been a long-standing quest that has so far produced more questions than answers.

One of the most popular tools for explaining the behavior of individual network units is to visualize unit responses via activation maximization [16, 33, 38, 35, 39, 36, 54, 15]. The idea is to start with an image (typically random noise) and iteratively change pixel values to maximize the activation

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

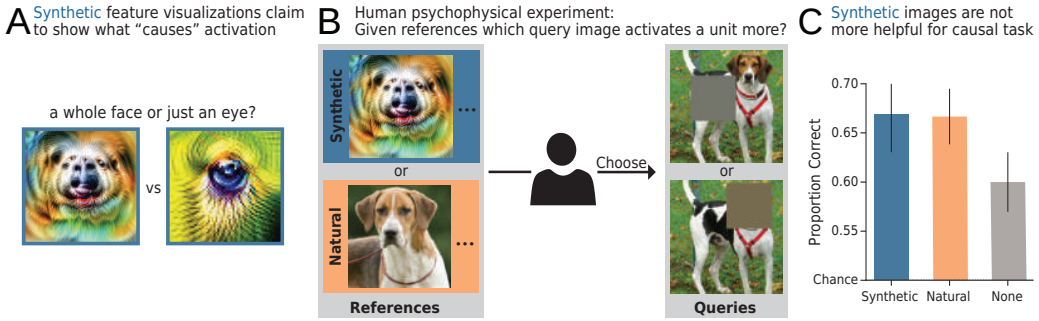


Figure 1: How useful are feature visualizations to interpret the effects of interventions? **A: "Causal" synthetic feature visualizations. B: Human experiment.** Given strongly activating reference images (e.g. synthetic or natural), a human participant chooses which out of two manipulated images activates a unit more. Note that this trial is made up — real trials are often more difficult. **C: Core result.** While participants are above chance for all visualization types, synthetic images only provide a substantial advantage over *no* references and not over other alternatives such as natural references.

of a particular network unit via gradient ascent. The resulting synthetic images, called *feature visualizations*, often show interpretable structures, and are believed to isolate and highlight exactly those features that "cause" a unit's response [40, 50]. Some of the synthetic feature visualizations appear quite intuitive and precise. As shown in Fig. 1A, they might facilitate distinguishing whether, for example, a unit responds to just an eye or a whole dog's face.

However, other aspects cast a more critical light on feature visualization's "causality": Generating these synthetic images typically involves regularization mechanisms [36, 33, 38, 35], which may influence how faithfully they visualize what "causes" a network unit's activation. Furthermore, to obtain a complete description of a mathematical function, one generally needs more information than just knowing its extrema. In view of this, it is an open question how well a unit can be characterized by simply visualizing the arguments of its maxima. Finally, a crucial unknown factor is whether *humans* are able to obtain a causal understanding of CNN activations from these synthetic visualizations.

Given these points, we develop a psychophysical experiment to test whether feature visualizations by Olah et al. [40] indeed allow humans to gain a causal understanding of a unit's behavior. Our task is based on the reasoning that being able to predict the effect of an intervention is at the heart of causal understanding. Understanding the causal relation between variables implies an understanding of how changes in one variable affect another one [45]. In our proposed experiment, this means that participants can predict the effect of an intervention — in form of an image manipulation — if they know the causal relation between image features and a unit's activations. Our experiment tests whether synthetic feature visualizations indeed provide information about such causal relations. Specifically, we ask humans which of two manipulated images activates a CNN unit more strongly. The interventions we test are obtained by placing an occlusion patch at two different locations in an image. Taken together, this experiment probes the purported explanation method's advantage of causality in a counterfactual-inspired prediction set-up [14].

Besides feature visualizations, other visualization methods have been used to gain an understanding of the inner workings of CNNs. In this experiment, we additionally test alternatives based on natural dataset examples and compare them with feature visualizations. This is particularly interesting because dataset examples are often assumed to provide less "causal" information about a unit's response as they might contain misleading correlations [40]. To continue the example above, dog eyes usually co-occur with dog faces; thus, separating the influence of one image feature from the other one using natural exemplars might be challenging.

Our data shows that:

- Synthetic feature visualizations provide humans with some helpful information about the most important patch in an image — but not much more information than no visualizations at all.
- Dataset samples as well as other combinations and types of visualizations are similarly helpful.
- How easily the most important patch is identifiable depends on the unit, the images as well as the relative activation strength attributed to the patch.

2 Related Work

Feature visualizations are a widely used method to understand the learned representations and decision-making mechanisms of CNNs [33, 38, 35, 39, 36, 54, 15, 40, 37]. As such, several works leverage this method to study InceptionV1 [42, 41, 8, 43, 50, 9, 58, 59, 46] and other networks [6, 21, 20]; others create interactive tools [61, 44, 52] or introduce analysis frameworks [65]. In contrast, some researchers question whether this synthetic visualization technique, first introduced by Erhan et al. [16], is too intuition-driven [27], and how representative the appealing visualizations in publications are [26]. Further, as already mentioned above, the engineering of the loss function may influence their faithfulness [36, 33, 38, 35]. Another challenge is generating *diverse* feature visualizations to represent the different aspects that one single unit may respond to [42, 36]. Finally, our recent human evaluation study [5] found that while these synthetic images do provide humans with helpful information in a forward simulation-inspired task, simple natural dataset examples are even more helpful.

Human evaluation studies are extensively used to quantify various aspects of interpretability. As an alternative to pure mathematical approximations [2, 66, 57, 63], researchers not only evaluate the understandability of explanation methods in psychophysical studies [7, 34, 5], but also trust in these methods [28, 64]) as well as the human cognitive load necessary for parsing explanations [1] or whether humans would follow an explained model decision [47, 13, 48]. A recent study even demonstrates that metrics of the explanation quality computed *with* human judgment are more insightful than those without [4].

Counterfactuals are a popular paradigm for both *creating* as well as *evaluating* explanation methods. Intuitively, they provide answers to the question "what should I change to achieve a different outcome?" — in the context of machine learning explanation methods, usually the smallest, realistic change to a data point is of interest. As examples, counterfactual explanation methods have been developed for vision- [22] and language-based [62] models as well as for model-agnostic scenarios [51]. Further, they are set into context of the EU General Data Protection Regulation [60]. Ustun et al. [56] investigate feasible and least-cost counterfactuals, while Mahajan et al. [32] and Karimi et al. [25] take feature interactions into account. To *evaluate* — rather than create — explanation methods, researchers often follow the "counterfactual simulation" task introduced by Doshi-Velez and Kim [14]: Humans are given an input, an output, and an explanation and are then asked "what must be changed to change the method's [model's] prediction to a desired output?" Doshi-Velez and Kim [14]. Based on this task, Lucic et al. [30] test their new explanation method and Hase and Bansal [24] compare different explanation methods to each other.

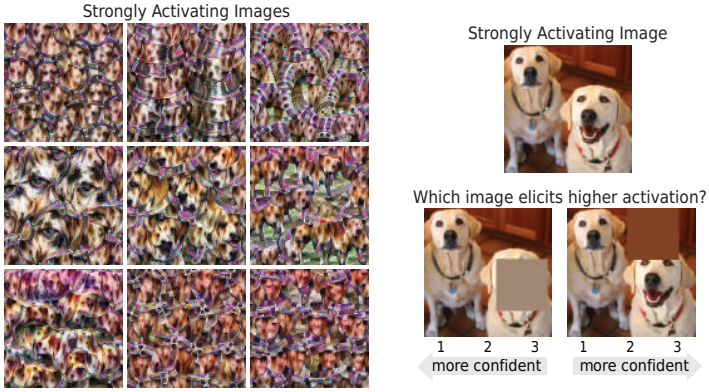


Figure 2: **Schematic visualization of an example trial** in our psychophysical experiment. For a certain network unit, participants are shown several maximally activating images. While the ones on the left serve as reference images, the ones on the right serve as query images: The top one is a natural maximally activating image and the bottom ones are copies of said image with square occlusions at different locations. The task is to select the image that activates the given network unit more strongly. Participants answer by clicking on the number below the corresponding image according to their confidence level (1: not confident, 2: somewhat confident, 3: very confident). Correct answer: right image.

In this project, we design a counterfactual-inspired task to evaluate how well feature visualizations support causal understanding of CNN activations. This is the first study to apply such a paradigm to understanding the causes of individual units' activations. In order to scale the experiments, we

simplify our task by having participants choose between two intervention *options*, rather than having them freely determine interventions themselves.

3 Methods

We run an extensive psychophysical experiment with more than $12,000$ trials distributed over 323 crowdsourced participants on Amazon Mechanical Turk (MTurk) and two experts (the two first authors).[1] For more details than provided below, please see Appx. Sec. A.1.

Design Principles Overall, our experimental design choices aim at (1) the *best performance possible*, meaning that we select images that make the signal as clear as possible; (2) *generality* over the network, meaning that we randomly sample units of different layers and branches (testing all units would be too costly); and (3) *easy extendability*, meaning that we choose a between-participant design (each participant sees only one reference image condition) so that other visualizations methods can be added to the comparisons in the future.

3.1 Psychophysical Task

If feature visualizations indeed support causal understanding of CNN activations, this should enable humans to predict the effect of an intervention, such as how occluding an image region changes a unit's activation. Based on this idea, we employ a two-alternative forced choice task (chance performance: 50%) where human observers are presented with two different occlusions in an image, and asked to estimate which of them causes a smaller change to the given unit's activation (see Fig. 2 for an example trial). More specifically, participants choose the *query* image that they believe to also elicit a strong activation given a set of 9 *reference* images. Such references could for instance consist of synthetic feature visualizations of a certain unit (purportedly "causal"), or alternative visualizations. To summarize, the task requires humans to first identify the shared aspect in the reference images and to then choose the query image in which that aspect is more visible. Since we do not make any assumptions about whether participants are familiar with machine learning, we avoid asking participants about activations of a unit in the CNN. Instead, we explain that an image would be "favored" by a machine, and the task is to select the image which is "more favored". The complete set of instructions shown to participants can be found in Appx. Fig. 9 and 10. In addition to each participant's image choice, the subjective confidence level and reaction time are also recorded.

3.2 Stimulus Generation

To generate stimuli, we follow Olah et al. [40] and use an InceptionV1 network [53] trained on ImageNet [12, 49]. Throughout this paper, we refer to a CNN's channel as a "unit" and imply taking the spatial average of all neurons in one channel.[2] We test units sampled from 9 layers and 2 Inception module branches (namely 3×3 and POOL). For more details on the generation procedures of the respective stimuli, see Appx. A.1.2.

We use five different types of **reference images**:

- **Synthetic references**: The synthetic images are the optimization results of the feature visualization method by Olah et al. [40] with the channel objective for 9 diverse images.
- **Natural references**: The reference images are the most strongly activating[3] dataset samples from ImageNet [12, 49].
- **Mixed references**: This is a combination of the previous two conditions: the 5 most strongly activating natural and 4 synthetic reference images are used. The motivation is that this condition combines the advantages of both worlds — namely precise information from feature visualizations and easily understandable natural images — and, thus, has the potential to give rise to higher performance in the task. Jointly looking at these two visualization types is common in practice [40].

[1] Code and data are available at github.com/brendel-group/causal-understanding-via-visualizations.

[2] Other papers might refer to a channel as a "feature map", e.g. [5].

[3] To reduce compute requirements, we use a random subset of the training set ($\approx 50\%$).

- **Blurred references**: To increase the informativeness of natural images for this task, we modify them by blurring everything but a single patch. This patch is chosen in the same way as in the maximally activating query image (see below). Consequently, this method cues participants to the most important image feature. In a way, these images can be seen as an approximate inverse of the maximally activating query image and might improve performance on our task.
- **No references**: This is a control condition in which participants do not see any reference images and have to solve the task purely based on query images.

To generate **query images**, we place a square patch of 90×90 pixels of the average RGB color of the occluded pixels into a most strongly activating image chosen from ImageNet. The location of the occlusion patch is chosen such that the activation of the manipulated image is either minimal or maximal among all possible occlusion locations. These images then yield the distractor and target query images respectively.

3.3 Structure of the Psychophysical Experiment

We test the five different reference image types as separate experimental conditions. In each condition, we collect data from a total of 50 different MTurk participants, each assigned to a single Human Intelligence Task (HIT) consisting of an instruction block, a variable number of practice blocks and a main block. The instructions extensively explain a hand-crafted example trial (see Appx. Fig. 9 and 10). The blocks of 4 practice trials each - which are randomly sampled from a pool of 10 trials - have to be repeated until reaching 100% performance; except in the none condition, as there is no obvious ground truth due to the absence of reference images. Finally, 18 main trials follow that are randomly interleaved with a total of 3 obvious catch trials. While feedback is provided during practice trials, no feedback is provided in the other trials. At the end, participants can share comments via an optional free-text field. Across all conditions, all participants see the same query images for the instruction, practice and catch trials. In contrast, the query images differ across participants in the main trials: In each reference image condition, we test 10 different sets of query images, each responded to by 5 different MTurk participants, hence 50 HITs per condition. The order of the main and catch trials per participant is randomly arranged, and identical across conditions. Each MTurk participant takes part in only one reference image condition (i.e. reference images are a between-participants factor). For more details, see Appx. Sec. A.1.4.

3.4 Ensuring High-Quality Data in an Online Experiment

To ensure that the data we collect in our online experiment is of high quality, we take two measures: (1) We integrate hidden checks which were set before data collection. Only if a participant passes all five of them do we include his/her data in our analysis. First, these *exclusion criteria* comprise a performance threshold on the practice trials as well as a maximum number of blocks a participant may attempt. Further, they include a performance threshold for catch trials, a minimum image choice variability as well as a minimum time spent on both the instructions and the whole experiment. For more details, see Appx. Sec. A.1.1. (2) Our previous human evaluation study in a well-controlled lab environment found that natural reference images are more informative than synthetic feature visualizations when choosing which of two different images is more highly activating for a given unit [5]. We replicate this main finding on MTurk based on a subset of the originally tested units (see Appx. A.3) which indicates that the experiment's environment does not influence this task's outcome. Our decision to leverage a crowdsourcing platform is further corroborated by our result in Borowski et al. [5], that there is no significant difference between expert and lay performance.

3.5 Baselines

In order to both set MTurk participants' performance into context as well as evaluate different strategies participants could use to perform our task, we further evaluate a few baselines.

- **Expert Baseline**: The two first authors answer all 18 trials in all 5 reference conditions on all 10 image sets. As they are familiar with the task design and are certainly engaged, this data serves as an upper human bound.
- **Center Baseline**: In natural images from ImageNet, important objects are likely to be closer to the center of the image. If participants were biased to assume that units respond to *objects*, a potential

strategy to decide which occluding patch produces a smaller effect on the unit's activation would therefore be to choose the image with the most eccentric occlusion. The Center Baseline model performs this strategy for all images.

- **Primary Object Baseline**: The Center Baseline is not a perfect measurement of an object-biased strategy because primary objects can appear away from the center. To account for this, the two first authors and the last author manually label all trials, choosing the image for which the occlusion hides as little information as possible from the most prominent object in the scene. In approximately one third of the trials ($58/180$), the authors' confidence ratings are very low (reflecting e.g. the absence of a primary object); in these cases we repeatedly replace the decisions by random binomial choices. Thus, in the results, we report the estimated expected values, but cannot perform a by-trial analysis. For more details, see Appx. Sec. A.1.3.
- **Variance Baseline**: Another assumption participants might make is that a patch in a low-contrast region, e.g. a blue sky, is unlikely to have a large effect on the unit's activation. This baseline selects the query image whose content is less affected by the introduction of the occlusion patch. To simulate this, we calculate the standard deviation over the occluded pixels and choose the one of the lower standard deviation.
- **Saliency Baseline**: As a complement to the baselines above, this baseline selects the query image whose original pixels hidden by the occlusion patch have a lower probability of being looked at by the participants. This simulates that participants select the image with a patch that occludes less prominent information and is estimated with the saliency prediction model DeepGaze IIE [29]. For more details, see Appx. Sec. A.1.3.

4 Results

The results shown in this section are based on 7350 [4] trials from MTurk participants, who passed all exclusion criteria, and experts distributed over five conditions. In all figures, *Synthetic* refers to the purportedly "causal", activation-maximizing feature visualizations, *Natural* to ImageNet samples, *Mixed* to the combined presentation of synthetic and natural images, *Blur* to the blurred images, and *None* to the condition with no reference images at all. Further, error bars indicate two standard errors above and below the participant-mean over network units and image sets, unless stated otherwise.

4.1 No Significant Advantage of Synthetic Feature Visualizations

If feature visualizations provide humans with useful information about the image features causing high unit activations and other visualizations do not, participants' accuracy in our task should be higher given feature visualizations than for all other visualization types or no reference images. This is only partly what we find: On average, accuracy for feature visualizations is slightly higher than when no reference images are given ($67 \pm 4\%$ vs. $60 \pm 3\%$). However, the accuracy for feature visualizations is not significantly higher than for other visualization methods (see Fig. 3A, dark bars). For the latter, MTurk participants reach between $66 \pm 3\,\%$ and $67 \pm 5\,\%$ depending on the visualization type. Statistically, only the condition without reference images is

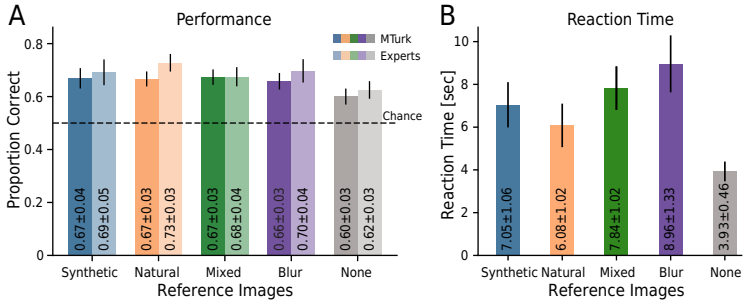


Figure 3: **A: Task accuracy.** On average, humans reach the same performance regime with any visualization method. This holds for both lay participants on MTurk (darker colors) as well as experts (brighter colors). **B: Reaction times.** MTurk participants need several seconds to answer a trial, indicating that they carefully make their decision. For more details see Appx. Fig. 13.

[4] (18 main + 3 catch trials)×50 MTurk participants ×5 conditions + (18 main + 3 catch trials)×20 expert measurements ×5 conditions.

different from all other conditions ($p < 0.05$, Mann-Whitney U test). Taken together, these findings suggest that all visualization methods are similarly helpful for humans in our counterfactual-inspired task, and that they only seem to offer a small improvement over no visualizations at all.

4.1.1 MTurk Participants Carefully Make Their Choices

Similar performances for various conditions such as those found in Fig. 3A might suggest that participants would not give their best when doing our experiment. However, several aspects speak against this: (1) Measurement of the two first authors, i.e. experts who designed and thus clearly understand the task, and certainly engage during the experiment, again show very similar performance (see Fig. 3A, bright bars): This estimated upper bound is just $1 - 6\%$ better than MTurk participant performance. (2) With our strict exclusion criteria, we check for doubtful participant behavior and only include data from participants who pass all five criteria. (3) Reaction times per trial (see Fig. 3B) lie between $\approx 4\,\mathrm{s}$ and $\approx 9\,\mathrm{s}$. This, as well as the fact that participants take longer for the conditions *with* references than for the *None* condition, suggest that they carefully make their decisions. (4) Several MTurk participants' comments in an optional free-text field indicate that they engage in the task: "[...] I did my best", "It was engaging", "interesting task". (5) Trial-by-trial responses between MTurk participants are more similar than expected by chance (see Fig. 4B discussed below), which suggests that humans use the available information.

4.1.2 Simple Baselines Can Reach the Same Above-Chance Performance Regime

Decision-making strategies can be diverse. To set human performance into context, we evaluate several simple strategies as baselines: How high is performance if one always chooses the query image with an unoccluded center (Center Baseline) or primary object (Object Baseline)? Or such that the more varying or salient image region is unoccluded (Variance and Saliency Baseline)? Fig. 4A shows that these strategies have varying performances with the best ones — namely the Object and Variance baselines — reaching $63 \pm 1\,\%$ and $63\,\%$, respectively. Since already these simple heuristics, which do not require reference visualizations, can reach the same performance regime as participants, the additional advantage of visualizations (reaching just up to $4\,\%$ better performance) appears limited.

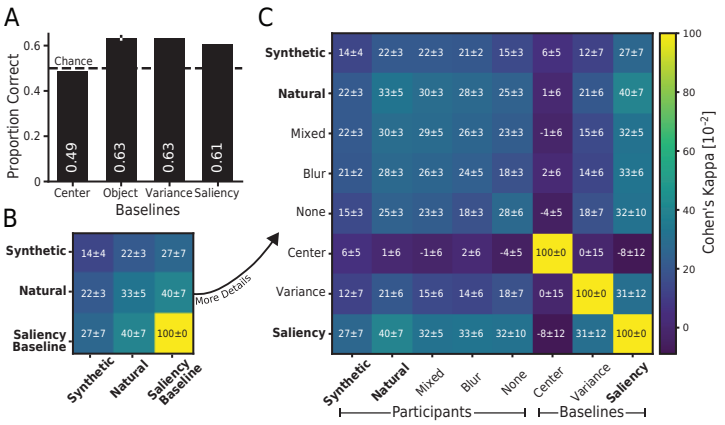


Figure 4: **A: Baseline performances.** Simple baselines can reach above chance level.[5] **B, C: Decision consistency.** The mean and two standard errors of the mean of Cohen's kappa averaged over participants and image sets quantifies the pairwise consistency of decision patterns.[6] While they vary across participants, they are higher between conditions with natural references and highest between the Saliency Baseline and other conditions. For more details, see Appx. Fig 15.

4.2 By-trial Decisions Show Systematic but Fairly Low Agreement

While accuracy is the most common metric to evaluate task performance, it does not suffice to compare two systems' decision-making processes [31, 19, 18]. Instead, a quantitative trial-by-trial error analysis is necessary to ascertain or distinguish strategies. Here, we use Cohen's kappa [10] to

[5] Only the Object Baseline has an error bar because in trials with, e.g. no clear primary object, we replace decisions by random binomial choices. The reported values are the estimated expectation value and standard deviation.

[6] There is no data for the Object Baseline because about one third of the trials do not have a clear answer from the three author responses. For more details, see Appx. A.1.3.

calculate the degree of agreement in classification while taking the expected chance agreement into account. A value of 1 corresponds to perfect agreement, while a value of 0 corresponds to as much agreement as would be expected by chance. Negative values indicate systematic disagreement.

In Fig. 4B and C, we plot consistency between MTurk participants of the same and different reference conditions as well as between MTurk participants and baselines. Since Cohen's kappa only allows for comparisons of two decision makers, we compute this statistic for all possible pairs across image sets, and report the mean over participants and image sets and two standard errors of the mean. All values between participants as well as between participants and baselines are in an intermediate regime (up to 0.40). This suggests that there is systematic agreement, but also quite some room for subjective decisions. Among participant-baseline comparisons, highest agreement is found for the saliency baseline[7] , while lowest agreement is found for the Center Baseline. Within participant to participant comparisons, decision strategies for conditions involving unmodified natural images (*Natural*, *Mixed*) are more similar to each other as well as slightly more similar to other strategies than the *Synthetic*, *Blur* or *None* condition to other strategies. Within the *Synthetic* condition, participants are relatively inconsistent. We hypothesize that due to the fact that humans are more familiar with natural images, they use more consistent information from these types of reference images and, thus, their decisions are more similar.

4.3 Performance Varies across Units, Image Sets and Activation Differences, but Less So for Reference Conditions

Having found that feature visualizations do not offer an overall advantage over other techniques, we now ask: Is performance similar across units, query images and their activation differences?

Units and Image Sets As evident from Fig. 5, performance varies by unit, but usually not much by reference condition: While only one unit (layer 2, POOL) is clearly below chance level, many units reach around average performance and a few units stand out with high performances (e.g. layer 8, POOL). Further, the five reference conditions are relatively close to each other for most units. Finally, on the image set level, we observe fairly high variance - probably partly due to the limited number of participants per image set (see Appx. Fig. 14).

Fig. 6 further illustrates the different difficulty levels as well as the strong unit- and image-dependency: For the shown easy unit (Fig. 6A), the (presumably yellow-black) feature is fairly clearly identifiable and visible in the diverse reference and query images. In contrast, for the shown difficult unit (Fig. 6B), the unit's feature selectivity is unclear not only in the reference but also in the query images.

Activation Differences We hypothesize that our task might be easier if the difference in activations between the two interventions of the query images is larger. In Fig. 7A and B, we plot by-image-set performance against the relative activation differences, i.e. the difference between activations elicited by the two manipulated images normalized by the unperturbed query image's activation. The figure shows that even though we select query images as the most strongly activating images for a unit, the relative activation differences vary widely. Furthermore, human performance indeed tends to increase with higher relative activation difference, confirming our hypothesis. This trend is stronger in the POOL than in the 3×3 branch as quantified by the Spearman's rank correlations in Fig. 7C.

5 Discussion & Conclusions

Explanation methods such as feature visualizations have been criticized as intuition-driven [27], and it is unclear whether they allow humans to gain a precise understanding of which image features "cause" high activation in a unit. Here, we propose an objective psychophysical task to quantify how well these synthetic images support causal understanding of CNN units. Through a time- and cost-intensive evaluation (based on $24,439$ trials taking more than 81 participant hours including all pilot and reported experiments), we put this widespread intuition to a quantitative test. Our data provides no evidence that humans can predict the effect of an image intervention (occlusion) particularly well when supported with feature visualizations. Instead, human performance is only moderately above a

[7] From a different perspective, this result can be seen as a confirmation that the CNN learned to look at the "important" part of the image for downstream classification.

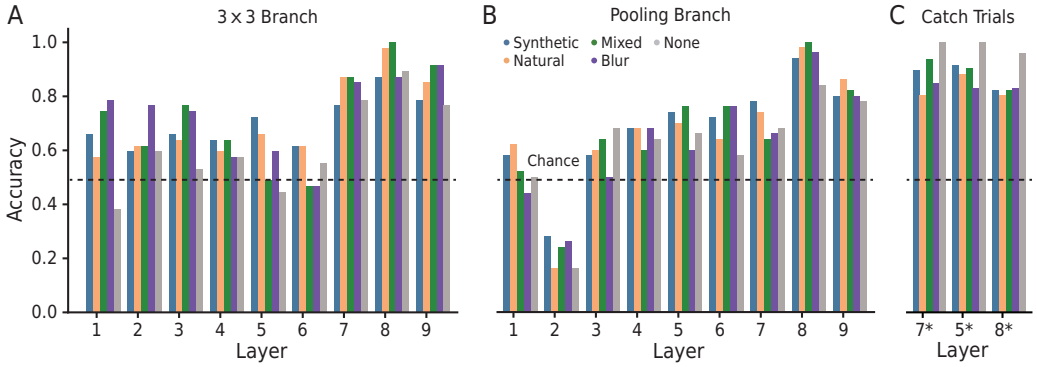


Figure 5: While for some units predicting the effect of an intervention is relatively easy, for most units performance is close to or just above chance. **A** and **B** show the **performance per unit** in the main trials separated by branch (3×3 and POOL respectively) and layer. **C** shows the performance per unit in the hand-picked trials used as catch trials (hence the *), though selected from those MTurk participants who pass the exclusion criteria without the catch trial exclusion criterion. Note that each bar represents averages over participants and image sets.

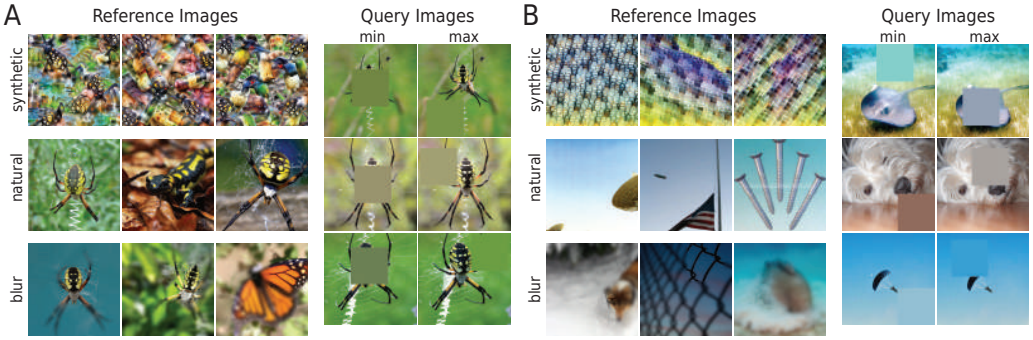
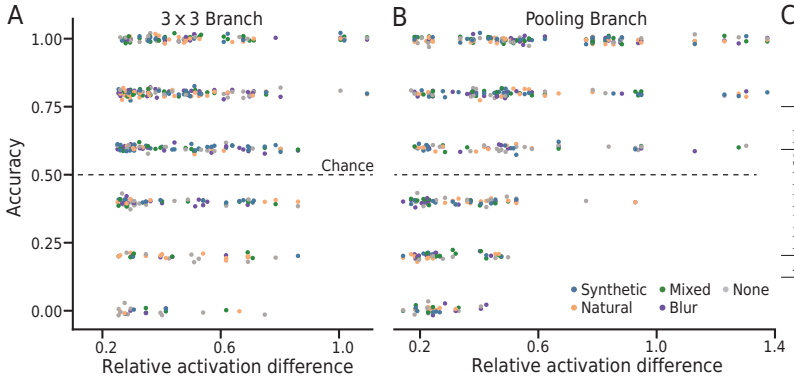


Figure 6: **Example reference and query images** for a unit with high (**A**) and low (**B**) performance from layer 8 and 2 of the POOL branch, respectively.



	Branch	
References	3×3	Pool
Synthetic	0.047	0.487*
Natural	0.162	0.447*
Mixed	0.080	0.488*
Blur	0.159	0.597*
None	0.242*	0.432*
All	0.135*	0.491*

Figure 7: Performance tends to increase with the relative activation difference between query images. This effect is stronger for the POOL branch (**B**) than for the 3×3 branch (**A**) as quantified by Spearman's rank correlations (**C**). Stars signal significance ($p < .05$). Note that each dot in A and B represents the participant-averages, i.e. there is one dot per combination of layers, branch and image set. For an alternative visualization see Appx. Fig. 16.

baseline condition where humans are not shown any visualization at all, and similar to that of other visualization methods such as simple dataset samples. Further, by-trial decisions show systematic but fairly low agreement between participants. Finally, task performance depends on the unit choice, image selections and activation differences between query images. These results add quantitative evidence against the generally-assumed usefulness of feature visualizations for understanding the causes of CNN unit activations.

Our counterfactual-inspired task is the *first* quantitative evaluation of whether feature visualizations support causal understanding of unit activations, but it is certainly not the *only* possible way to evaluate causal understanding. For example, our interventions are constrained to occlusions of a fixed size and shape, imposing an upper limit on the precision with which the occlusions can cover the part of the image that is most responsible for driving a unit's activation. Future work could explore more complex intervention techniques, extend our study to more units of InceptionV1 as well as to different networks, and investigate additional visualization methods. Thanks to the between-participant design, new conditions can be added to the data without the requirement to re-run already collected trials.

Taken together, the empirical results of our quantitative evaluation method indicate that the widely used visualization method by Olah et al. [40] does not provide causal understanding of CNN activations beyond what can be obtained from much simpler baselines. This finding is contrary to wide-spread community intuition and reinforces the importance of testing falsifiable hypotheses in the field of interpretable artificial intelligence [27]. With increasing societal applications of machine learning, the importance of feature visualizations and interpretable machine learning methods is likely to continue to increase. Therefore, it is important to develop an understanding of what we can — and cannot — expect from explainability methods. We think that human benchmarks, like the one presented in this study, help to expose a precise notion of interpretability that is quantitatively measurable and comparable to competing methods or baselines. The paradigm we developed in this work can be easily adapted to account for other notions of causality and, more generally, interpretability as well. For the future, we hope that our task will serve as a challenging test case to steer further development of feature visualizations.

Author Contributions

The idea to test how well feature visualizations support causal understanding of CNN activations was born out of several reviewer and audience comments on our previous paper [5]. The first idea of how to test this in a psychophysical experiment came from TSAW. JB led the project. JB, RSZ, WB and TSAW jointly improved the experimental set-up with input from MB and RG. RSZ led and JB helped with the implementation and execution of the experiment; JB led and RSZ contributed to the generation of stimuli. RSZ and JB both coded the baselines, and TSAW guided the replication experiment with statistical power simulations. The data analysis was performed by RSZ and JB with advice and feedback from RG, TSAW, WB and MB. TSAW and WB provided day-to-day supervision. While JB and RSZ created the first draft of the manuscript, RG and TSAW heavily edited the manuscript and all authors contributed to the final version.

Acknowledgments

We thank Felix A. Wichmann and Isabel Valera for a helpful discussion. We further thank Ludwig Schubert for information on technical details via `slack.distill.pub`. In addition, we thank our colleagues for helpful discussions, and especially Matthias Kümmerer, Dylan Paiton, Wolfram Barfuss, and Matthias Tangemann for valuable feedback on our task, and/or technical support. Moreover, we thank our various reviewers and other researchers for comments on our previous paper inspiring us to investigate causal understanding of visualization methods. And finally, we thank all our participants for taking part in our experiments.

Funding

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting JB, RSZ and RG. This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Competence Center for Machine Learning (TUE.AI, FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ 01GQ1002), the Cluster of Excellence Machine Learning: New Perspectives for Sciences (EXC2064/1), and the German Research Foundation (DFG, SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP3, project number 276693517). MB and WB acknowledge funding from the MICrONS program of the Intelligence Advanced Research Projects

Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. WB acknowledges financial support via the Emmy Noether Research Group on The Role of Strong Response Consistency for Robust and Explainable Machine Vision funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1.

References

[1] Ashraf M. Abdul, Christian von der Weth, Mohan S. Kankanhalli, and Brian Y. Lim. COGAM: measuring and moderating cognitive load in machine learning model explanations. In Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik, editors, *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14. ACM, 2020. doi: 10.1145/3313831.3376615.

[2] Elvio Amparore, Alan Perotti, and Paolo Bajardi. To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Computer Science*, 7:e479, 2021.

[3] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[4] Felix Biessmann and Dionysius Irza Refiano. A psychophysics approach for quantitative comparison of interpretable computer vision models. *arXiv preprint arXiv:1912.05011*, 2019.

[5] Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021.

[6] Santiago A Cadena, Marissa A Weis, Leon A Gatys, Matthias Bethge, and Alexander S Ecker. Diverse feature visualizations reveal invariances in early layers of deep neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–232, 2018.

[7] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 258–262, 2019.

[8] Nick Cammarata, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov, and Chris Olah. Curve detectors. *Distill*, 2020. doi: 10.23915/distill.00024.003. https://distill.pub/2020/circuits/curve-detectors.

[9] Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 6(1):e00024–006, 2021.

[10] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[11] Joshua R de Leeuw and Benjamin A Motz. Psychophysics in a web browser? comparing response times collected with javascript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, 48(1):1–12, 2016.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848.

[13] William K Diprose, Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, and Reece Robinson. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association*, 27(4):592–600, 2020.

[14] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[15] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

[16] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[17] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2950–2958. IEEE, 2019. doi: 10.1109/ICCV.2019.00304.

[18] Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021.

[19] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[20] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3): e30, 2021.

[21] Nicolas Gonthier, Yann Gousseau, and Saïd Ladjal. An analysis of the transfer learning of convolutional neural networks for artistic images. *arXiv preprint arXiv:2011.02727*, 2020.

[22] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 2019.

[23] Peter Green and Catriona J MacLeod. Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498, 2016.

[24] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491.

[25] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions, 2020.

[26] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[27] Matthew L Leavitt and Ari Morcos. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

[28] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 2119–2128, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi: 10.1145/1518701.1519023.

[29] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling, 2021.

[30] Ana Lucic, Hinda Haned, and Maarten de Rijke. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 90–98, 2020.

[31] Wei Ji Ma and Benjamin Peters. A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181*, 2020.

[32] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers, 2020.

[33] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5188–5196. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299155.

[34] Sina Mohseni, Jeremy E Block, and Eric Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th International Conference on Intelligent User Interfaces*, pages 22–31, 2021.

[35] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015.

[36] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3510–3520. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.374.

[37] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 55–76. Springer, 2019.

[38] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427–436. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298640.

[39] Anh Mai Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3387–3395, 2016.

[40] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11): e7, 2017.

[41] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 2020. doi: 10.23915/distill.00024.002. https://distill.pub/2020/circuits/early-vision.

[42] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001. https://distill.pub/2020/circuits/zoom-in.

[43] Chris Olah, Nick Cammarata, Chelsea Voss, Ludwig Schubert, and Gabriel Goh. Naturally occurring equivariance in neural networks. *Distill*, 5(12):e00024–004, 2020.

[44] OpenAI. OpenAI Microscope. `https://microscope.openai.com/models`, 2020. (Accessed on 09/12/2020).

[45] Judea Pearl. *Causality*. Cambridge university press, 2009.

[46] Michael Petrov, Chelsea Voss, Ludwig Schubert, Nick Cammarata, Gabriel Goh, and Chris Olah. Weight banding. *Distill*, 6(4):e00024–009, 2021.

[47] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

[48] Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? *arXiv preprint arXiv:2012.13354*, 2020.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[50] Ludwig Schubert, Chelsea Voss, Nick Cammarata, Gabriel Goh, and Chris Olah. High-low frequency detectors. *Distill*, 6(1):e00024–005, 2021.

[51] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[52] Stefan Sietzen, Mathias Lechner, Judy Borowski, Ramin Hasani, and Manuela Waldner. Interactive analysis of cnn robustness. *Computer Graphics Forum (Proceedings of Pacific Graphics 2021)*, 40(7), 2021.

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA,*

USA, June 7-12, 2015, pages 1–9. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298594.

[54] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[55] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR, 2020.

[56] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[57] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

[58] Chelsea Voss, Nick Cammarata, Gabriel Goh, Michael Petrov, Ludwig Schubert, Ben Egan, Swee Kiat Lim, and Chris Olah. Visualizing weights. *Distill*, 6(2):e00024–007, 2021.

[59] Chelsea Voss, Gabriel Goh, Nick Cammarata, Michael Petrov, Ludwig Schubert, and Chris Olah. Branch specialization. *Distill*, 6(4):e00024–008, 2021.

[60] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018.

[61] Eric Wong, Shibani Santurkar, and Aleksander Mądry. Leveraging sparse linear layers for debuggable deep networks, 2021.

[62] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*, 2021.

[63] Xi Ye, Rohan Nair, and Greg Durrett. Evaluating explanations for reading comprehension with realistic counterfactuals. *arXiv preprint arXiv:2104.04515*, 2021.

[64] Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. Understanding the effect of accuracy on trust in machine learning models. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 279. ACM, 2019. doi: 10.1145/3290605.3300509.

[65] Mohammad Nokhbeh Zaeem and Majid Komeili. Cause and effect: Concept-based explanation of neural networks. *arXiv preprint arXiv:2105.07033*, 2021.

[66] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *arXiv preprint arXiv:2104.14403*, 2021.

A Appendix

A.1 Details on Methods of Counterfactual-Inspired Experiment

We closely follow our previous work [5] and hence often refer to specific sections of it in this Appendix.

A.1.1 Data Collection

Exclusion Criteria In order to acquire data of high quality from MTurk, we integrate five exclusion criteria. If one or more of these criteria is not met, we post the same HIT again:

- Maximal number of attempts to reach 100% performance in practice trials: 5
- Performance threshold for catch trials: two out of three trials have to be correctly answered
- Answer variability: at least one trial must be chosen from the less frequently selected side (to discard participants who only responded with "left" or "right")
- Time to read the instructions: at least $20\,\mathrm{s}$ ($15\,\mathrm{s}$ in the none condition)
- Time for the whole experiment: at least $90\,\mathrm{s}$ and at most $900\,\mathrm{s}$ (at least $40\,\mathrm{s}$, and at most $900\,\mathrm{s}$ in the none condition)

Minimize Biases To minimize a bias to either query image, the location of the truly maximally activating query image is randomized and participants have to center their mouse cursor by pressing a centered button "Continue" after each trial.

Expert Measurements The two first authors complete all 10 image sets in multiple conditions: At first, they label the query images for the Primary Object Baseline. Then they answer the none, synthetic or natural (counterbalanced between the two authors), mixed, and blur condition. Clicking through the trials several times means that they see identical images repeatedly.

A.1.2 Stimulus Generation

Model In line with previous work (e.g. Borowski et al. [5], Olah et al. [40]), we use an Inception V1 network [53] trained on ImageNet [12, 49]. For more details, see Sec. A.1.2 "Stimuli Selection - Model" in Borowski et al. [5].

Natural Images as Query and Reference Images The natural reference and query images are selected from a random subset of $599,552$ training images of the ImageNet ILSVRC 2012 dataset [49]. For each unit, we select those images that elicit a maximal activation. More specifically, we choose the very most activating images as the query images and the next most activating images as reference images and ensure no overlap between query and references images as well as between image sets. As we follow our work published in Borowski et al. [5], please see A.1.2 for more details on the sampling procedure. In total, we generate 20 different image sets per unit. In the presented data, we only use half of these sets.

Query Images For the query images, we use the 20 maximally activating images for a given unit. To produce the manipulated query images, a square patch of 90×90 pixels is placed on the unperturbed query image. The side length of a patch corresponds to 40% of a preprocessed image's side length. The position of the occlusion patch is chosen such that the manipulated image's activation for a given unit is minimal (maximal) among all possible manipulated images' activations. This maximizes the signal in the query images and means that patches of the two query images can overlap.

In a control experiment, we test whether the partial occlusions of the natural ImageNet images cause the manipulated images to lie outside the natural image distribution. If this was the case, the query images would fail to be representative of the network's activity for natural images. Here, we test how similar the response to the unperturbed and partially occluded images is. Specifically, we count how often there is an overlap of the top-5 predictions. If network activations were drastically different for the occluded than for the unperturbed images, we should find low agreement. However, we do find an agreement for $97.8,\%$ of all tested images. Therefore, the square occlusions only have a marginal effect on the network's overall activity/predictions.

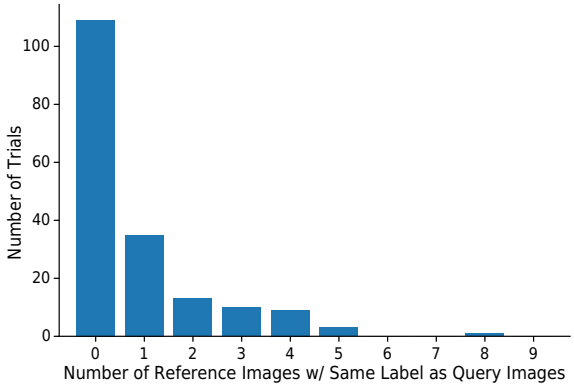


Figure 8: Distribution of the number of natural reference images that have the same label as the query image over the main trials used in the counterfactual-inspired experiment.

Reference Images: Natural Images In a control experiment, we test how often the label of the reference images coincide with the query image's label. If there was a high correspondence of these ImageNet labels, this could suggest that our experiment would rather reveal insights on how well humans would be able to *classify* images according to *labels* rather than to answer a counterfactual-inspired task based on the unit activations. Fig. 8 shows that the overlap of labels between query and reference images is low.

Reference Images: Blurred Images The blurred reference images are created by blurring all but one patch with a Gaussian kernel of size $(21, 21)$. This parameter choice allows participants to still get a general impression of an image, but not recognize details. Further, it is in line with work by Fong et al. [17]. The image choices are identical to the natural condition. Further — and just like for the query images — the position of the unblurred patch is chosen such that the manipulated image's activation for a given unit is maximal among all possible manipulated images' activations. Finally, the size of the unblurred patch is identical to the occlusion patch size: 40% of a preprocessed image's side length.

Reference Images: Synthetic Images from Feature Visualization Depending on the condition, we adjust the number of feature visualizations we generate: For the purely synthetic condition, we generate 9 visualizations, for the mixed condition, we generate 4 visualizations. As we follow our work published in Borowski et al. [5], please see A.1.2 for further details.

A.1.3 Baselines

Primary Object Baseline The Primary Object Baseline simulates that the more strongly activating manipulated image would be the one where the occlusion hides as little as possible from the most prominent object of the query image. To this end, the first two authors and the last author label all images. When doing so, they use a slightly modified logic: They select the image whose most prominent object is *most* occluded. If they cannot clearly identify a primary object in the image, they flag these trials, which are then treated differently in the analysis. For the analysis, the image choice is inverted again to counteract the inverted task that the authors responded to.

The performance reported in Fig. 4 is calculated by averaging over the three individual performances. Each individual performance itself is in turn estimated as the expectation value over random sampling for query images with no clear primary object. This analysis is in line with how the performance of MTurk participants is analyzed. An alternative option would be to take the majority vote of the three answers. When randomly sampling the choice for query images with no clear primary object, taking the majority votes and evaluating the expected accuracy, the performance would evaluate to 0.70 ± 0.02. Notably, 58 of all 180 trials are affected by the sampling as two or more authors responded with a confidence of 1 in 36 trials, and one author responded with a confidence of 1 while the other two gave opposing answers in 22 trials. This represents a fairly large fraction and reflects that many images on ImageNet have more than one prominent object [55, 3]. Consequently, there may not be a ground-truth for each trial in the Primary Object Baseline.

Saliency Baseline The Saliency Baseline simulates that participants select the image with a patch occluding the less prominent image region. To this end, we pass the unoccluded query image through the saliency prediction model DeepGaze IIE [29] which yields a probability density over the entire image. Next, we integrate said density over each of the two square patches. We then select the image with a lower value indicating that less important information is hidden by the occlusion patch.

A.1.4 Trials

Main trials For both the 3×3 and the POOL branch of each of the 9 layers with an Inception module, one randomly chosen unit is tested (see Table 1). These are the same units as in Experiment I of Borowski et al. [5].

Table 1: Units used as main trials in the 3×3 as well as the POOL branch in the counterfactual-inspired experiment. The numbers in brackets after each layer's name correspond to the numbering used in all our plots.

	Unit	
Layer	3×3	POOL
mixed3a (1)	189	227
mixed3b (2)	178	430
mixed4a (3)	257	486
mixed4b (4)	339	491
mixed4c (5)	247	496
mixed4d (6)	342	483
mixed4e (7)	524	816
mixed5a (8)	278	743
mixed5b (9)	684	1007

Instruction, Practice and Catch Trials The instruction, practice and catch trials are hand-picked by the two first authors. As a pool of units, the appendix overview of Olah et al. [40] as well as the "interpretable" POOL units used in Experiment I and all units used in Experiment II of Borowski et al. [5] are used. After generating all 20 reference and query image sets for these units, the authors select those units and image sets that they consider easiest (see Table 2).

Instruction Trial To explain the task as intuitively as possible, we construct an easy, artificial instruction trial (see Fig. 9 and 10): At first, we select a unit with easily understandable feature visualizations: The synthetic images of unit 720 of the POOL branch in layer 8 show relatively clear bird-like structures. From a popular image search engine, we then select an image[8] which not only clearly shows a bird but also other objects, namely a dog and water. To construct the minimally and maximally activating query images, we place the occlusion patches manually on the bird and dog.

Practice Trials In each attempt to pass the practice block, the trials are randomly sampled from a pool of 10 trials (see Table 2). Please note that unlike in any other trial type, participants receive feedback in the practice block: After each trial, they learn whether their chosen image truly is the query image of higher activation.

Catch Trials While all conditions with reference images use hand-picked easy trials (see Table 2), the none condition cannot rely on straight-forward clues from references. Therefore, we exchange the minimal query image with a minimal query image of a different, otherwise unused unit. This ensures that the catch trials in the none condition are also obvious.

A.1.5 Infrastructure

The online experiment is hosted on an Ubuntu 18.04 server running on an Intel(R) Xeon(R) Gold 5220 CPU. The experiment is implemented in JavaScript using jspsych 6.3.1 [11] and flask via

[8] `https://pixnio.com/fauna-animals/dogs/dog-water-bird-swan-lake-waterfowl-animal-swimming` released into public domain under CC0 license by Bicanski.

Table 2: Hand-picked unit choices for instruction, catch and practice trials in the counterfactual-inspired experiment.

Trial Type	Layer	Branch	Unit	Difficulty Level
instruction	mixed5a	pool	720	very easy
catch	mixed4e	pool	783	very easy
	mixed4c	pool	484	very easy
	mixed5a	3×3	557	very easy
practice	mixed4e	1×1	6	very easy
	mixed4a	pool	505	very easy
	mixed4e	pool	809	very easy
	mixed4c	pool	449	easy
	mixed4b	pool	465	easy
	mixed4c	1×1	59	easy
	mixed4e	1×1	83	easy
	mixed3a	1×1	43	easy
	mixed3b	pool	472	easy
	mixed4b	1×1	5	easy

Python 3.6. The generation of the stimuli shown in the experiment was completed in approximately 35 hours on a single GeForce GTX 1080 GPU. The calculation of all baselines required 8 additional GPU hours.

A.1.6 Amazon Mechanical Turk

MTurk participants To increase the chance that all MTurk participants understand the English instructions at the beginning of the experiment, we restrict access to workers from the following English-speaking countries: USA, Canada, Great Britain, Australia, New Zealand and Ireland.

Financial Compensation Based on an estimated duration and pilot experiments as well as a targeted hourly rate of US$ 15, we calculate the pay to be US$ 0.70 for the none condition and US$ 1.95 for all other conditions. MTurk participants whose data we include need a mean time of 209.64 ± 79.53 s and 396.87 ± 145.78 s for the whole experiment for the none condition and for all other conditions, respectively, which results in an hourly compensation of ≈ 12.02 US$/hour and 17.69 US$/hour, respectively. All MTurk participants who fully complete a HIT are paid, regardless of whether their responses meet the exclusion criteria. A total of US$ 1989.06 is spent on all pilot and real replication and counterfactual-inspired experiments.

Rights to Data We do not gather personal identifiable data from any MTurk participant. According to the MTurk Participation Agreement 3a [9], workers agree to vest all ownership and intellectual property rights to the requester (i.e., the authors of this study). Besides informing MTurk participants in the HIT preview about the academic and image classification nature of the experiment, we restate that "By completing this HIT, you consent to your anonymized data being shared with us for a scientific study." Further, we provide an email address, which some MTurk participants used to share feedback.

[9] `https://www.mturk.com/participation-agreement`, accessed on May 22nd, 2021

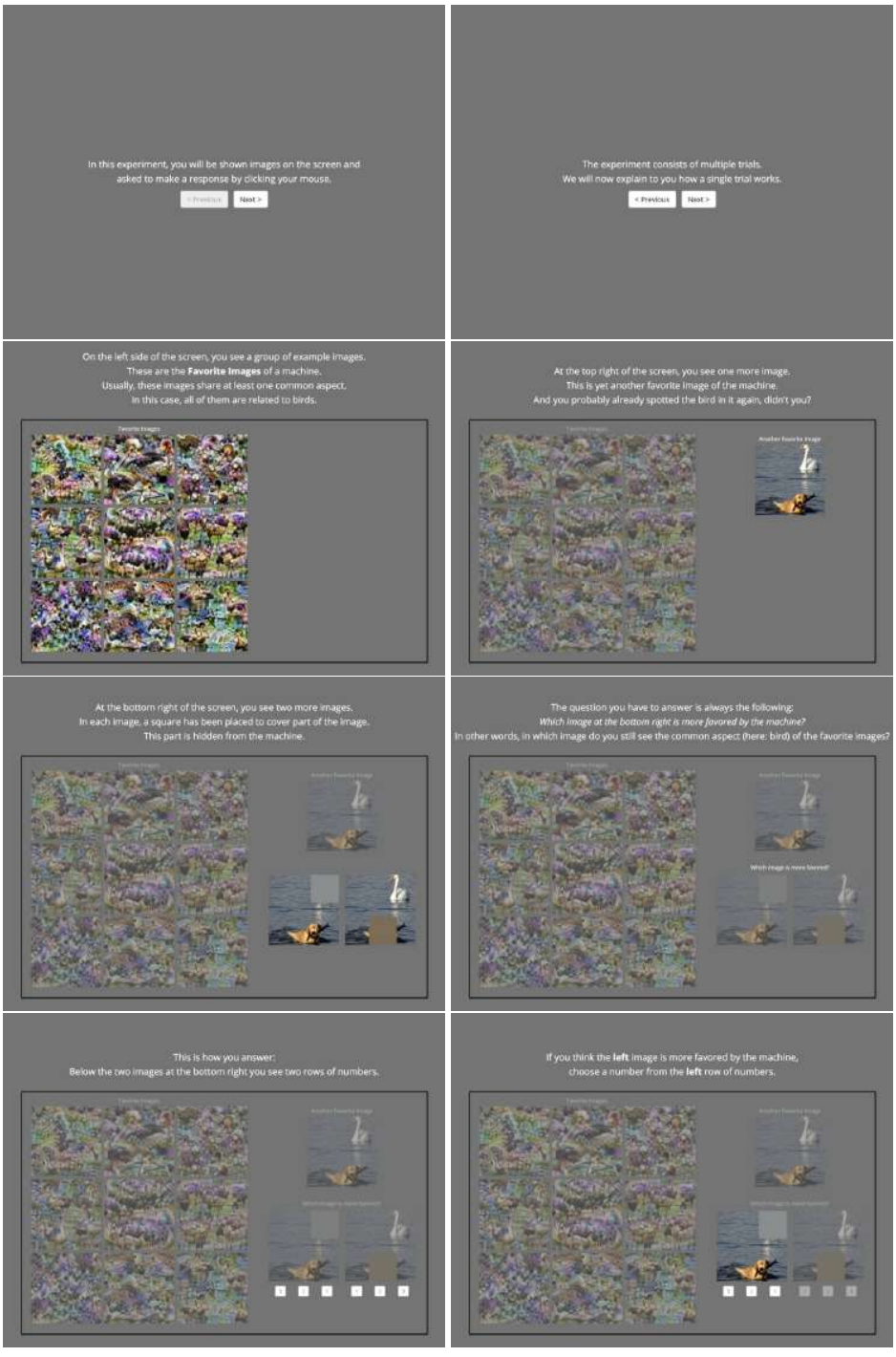


Figure 9: First eight instructions at the beginning of the counterfactual-inspired experiment.

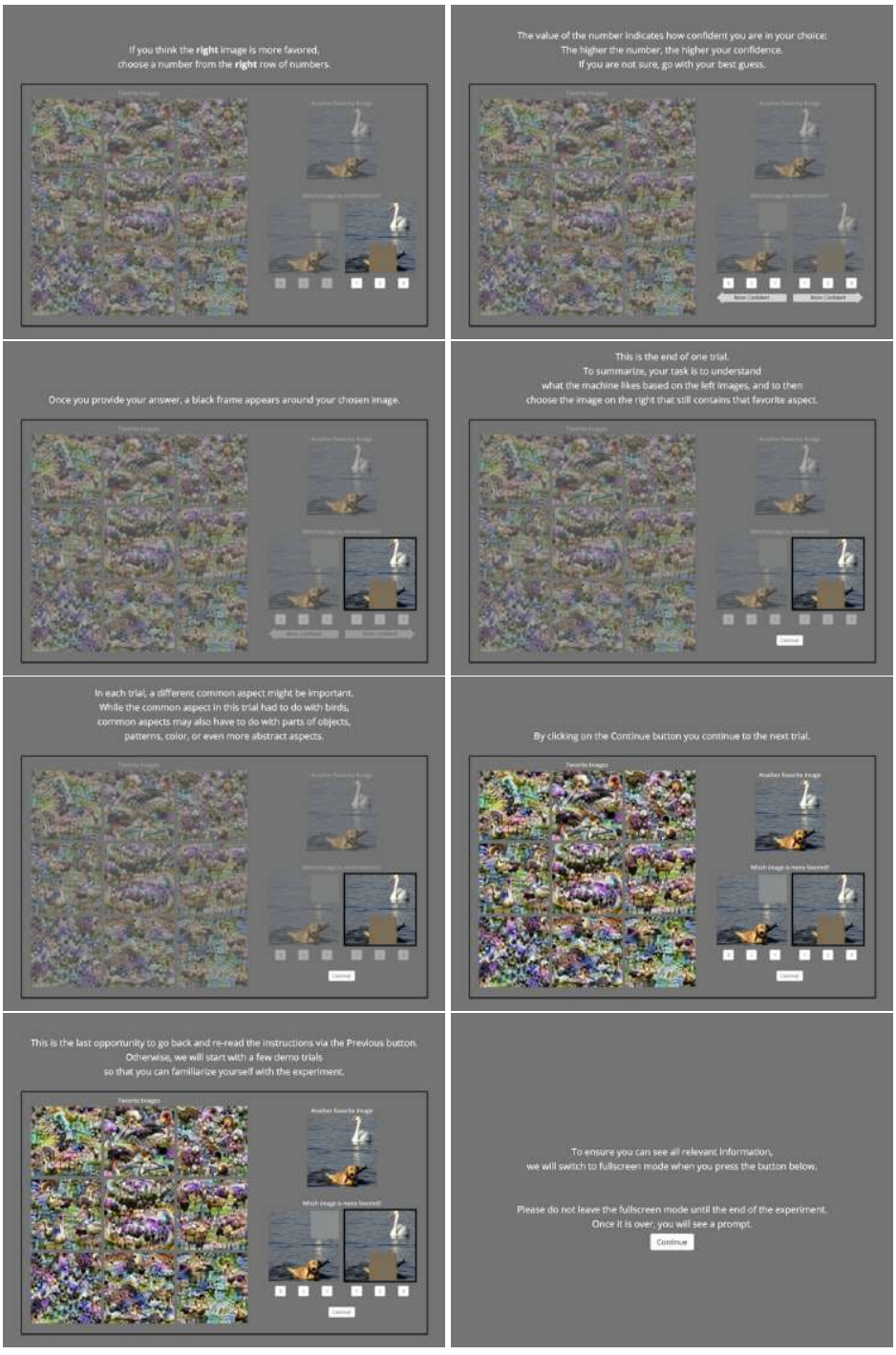


Figure 10: Second eight instructions at the beginning of the counterfactual-inspired experiment.

A.2 Details on Results of Counterfactual-Inspired Experiment

A.2.1 Different Query images

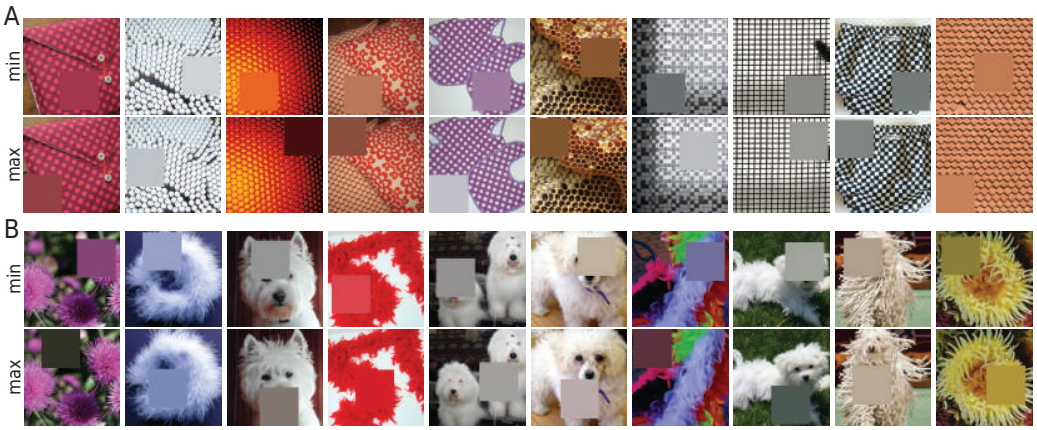
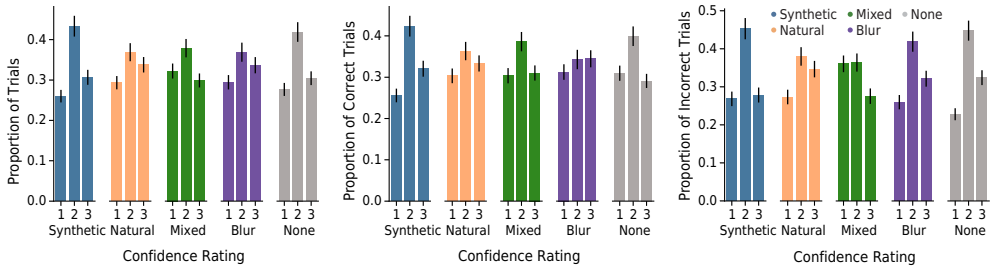


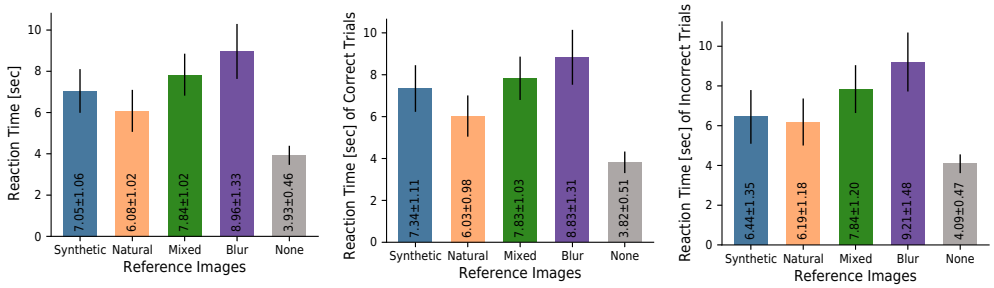
Figure 11: For each unit, we test 10 different image sets in the counterfactual-inspired experiment. The diversity of query images for layer 3 of the 3×3 branch (**A**), and layer 7 of the POOL branch (**B**) gives an intuitive explanation for varying performances.

A.2.2 Confidence Ratings and Reaction Times



(a) (b) (c)

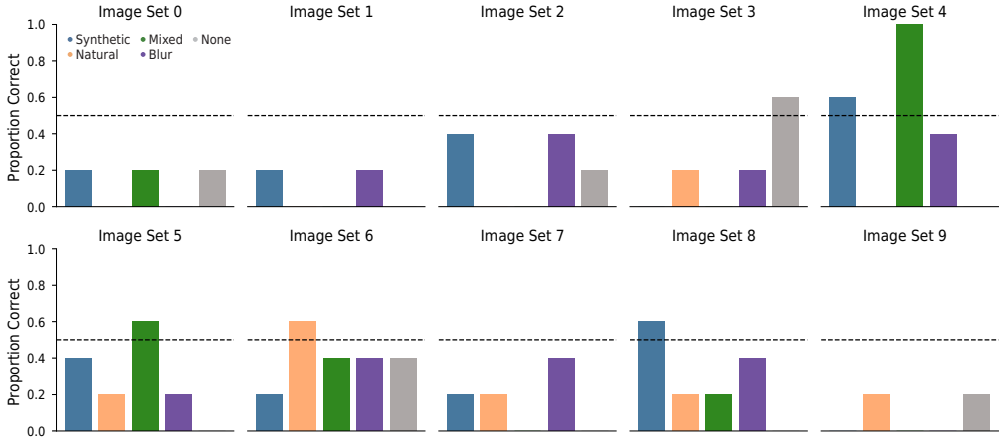
Figure 12: Confidence ratings of MTurk participants in the different reference conditions for (a) all, (b) only correct or (c) only incorrect trials of the counterfactual-inspired experiment.



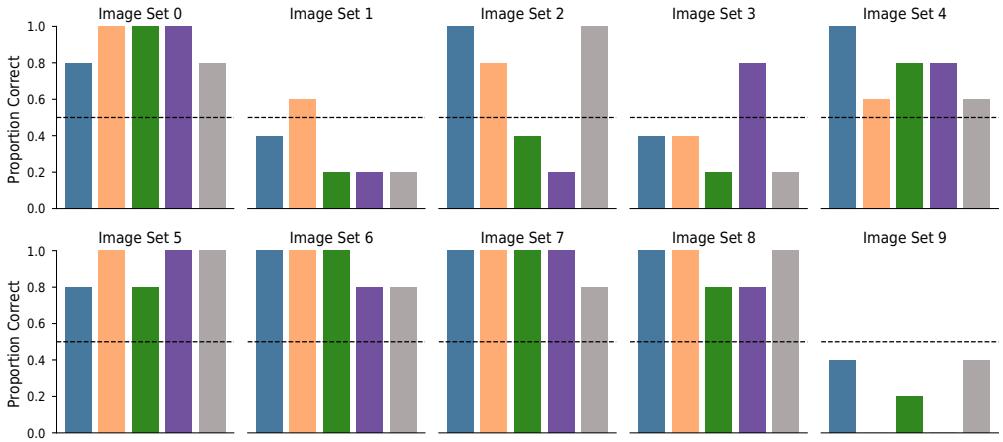
(a) (b) (c)

Figure 13: Reaction times of MTurk participants in the different reference conditions for (a) all, (b) only correct or (c) only incorrect trials of the counterfactual-inspired experiment.

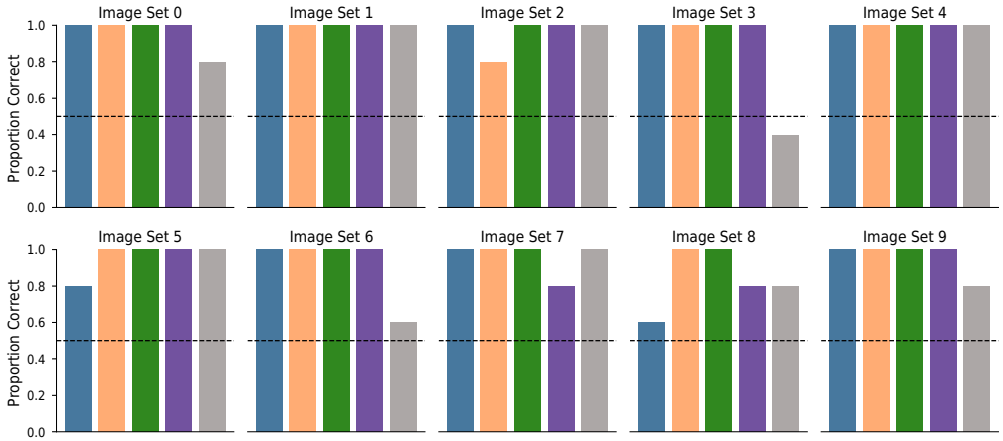
A.2.3 Performance per Image Set



(a) Difficult unit.



(b) Intermediate unit.



(c) Easy unit.

Figure 14: Performance in the counterfactual-inspired experiment split up by image sets and conditions for a difficult (layer 3, POOL), intermediate (layer 7, POOL) and easy unit (layer 8, POOL). Each bar shows the average over 5 MTurk participants.

A.2.4 Strategy Comparisons

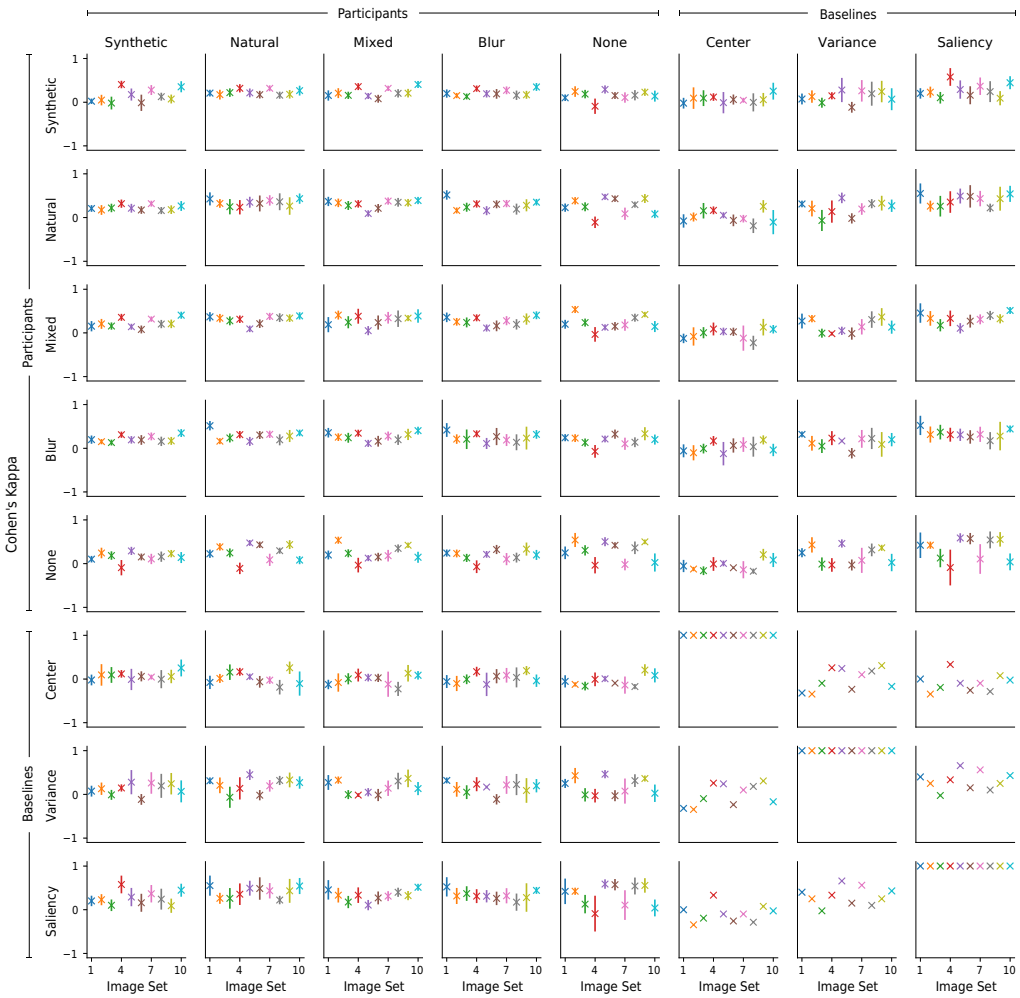
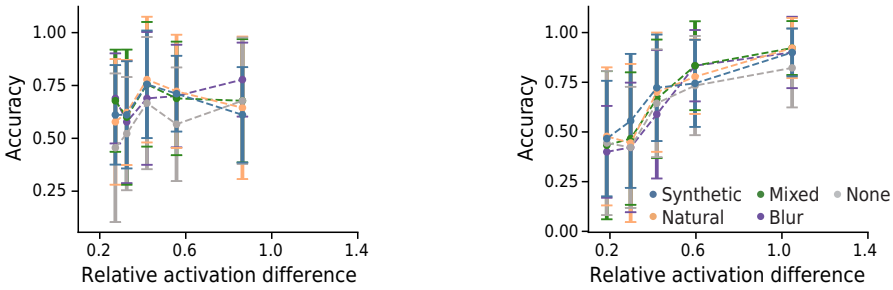


Figure 15: Cohen's kappa per image set in the counterfactual-inspired experiment (averages over participant-participant-, participant-baseline- or baseline-baseline-pairs). Error bars denote two standard errors of the mean.

A.2.5 Relative Activation Differences

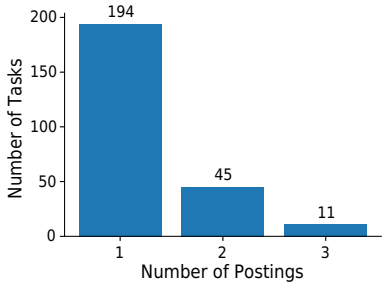


(a) 3×3 branch.

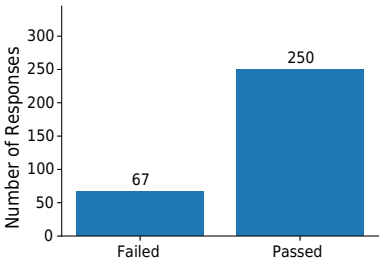
(b) POOL branch.

Figure 16: Accuracy in the counterfactual-inspired experiment as a function of the relative activation difference between the two query images for the (a) 3×3 branch and the (b) POOL branch. Here, the data points shown in Fig. 7 are summarized in 5 bins of equal counts; the plot shows the mean and standard deviation for each of the bins.

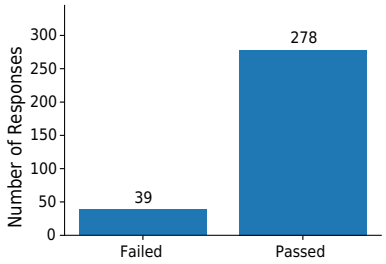
A.2.6 Exclusion Criteria



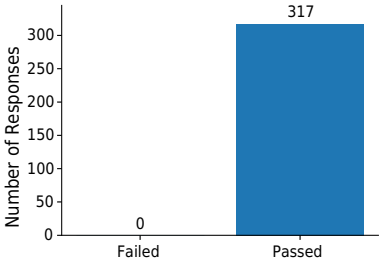
(a) Number of times a HIT is posted.



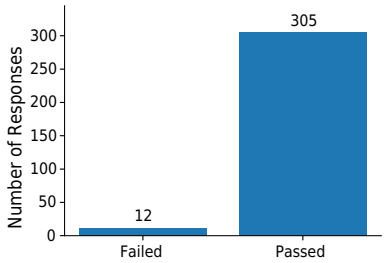
(b) All exclusion criteria.



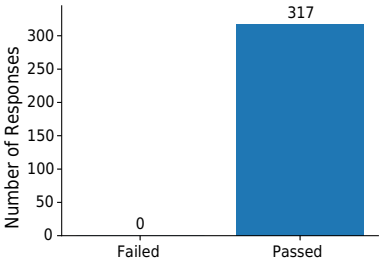
(c) Exclusion criterion: catch trials.



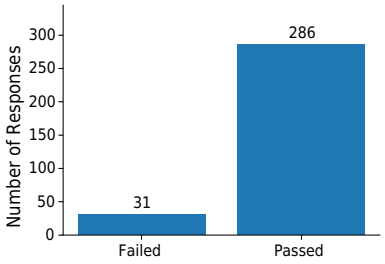
(d) Exclusion criterion: row variability.



(e) Exclusion criterion: instruction time.

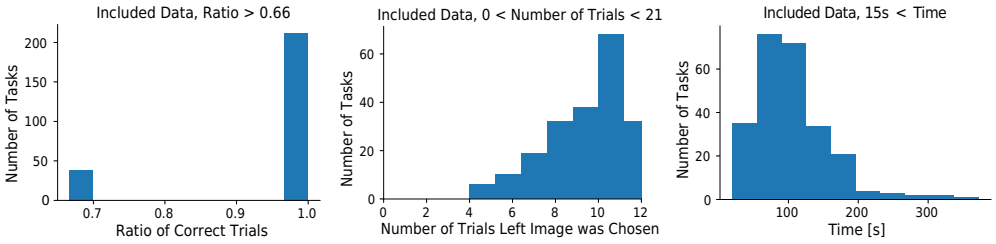


(f) Exclusion criterion: total response time.

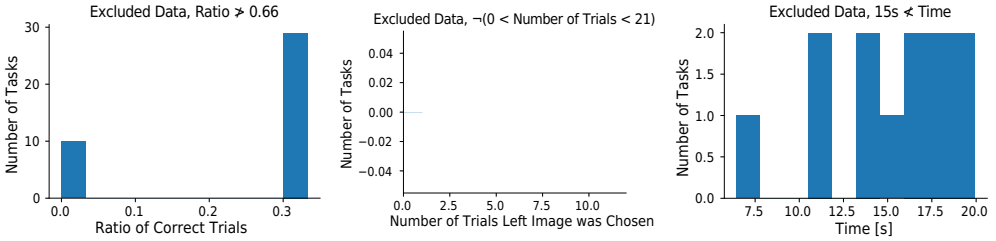


(g) Exclusion criterion: practice block.

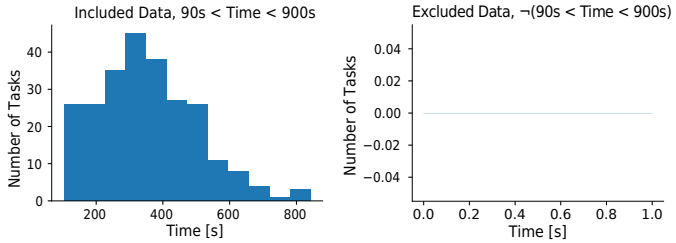
Figure 17: (a) Number of times a HIT is posted. To limit the financial risk, we limit the maximal number of times that a HIT can be posted at 3. (b-g) Distributions of MTurk participants that passed/failed the exclusion criteria in the counterfactual-inspired experiment on MTurk. Note that the sum of the counts of responses for the individual exclusion criteria in c-f is higher than the summary in b because a participant may have failed more than one exclusion criterion.



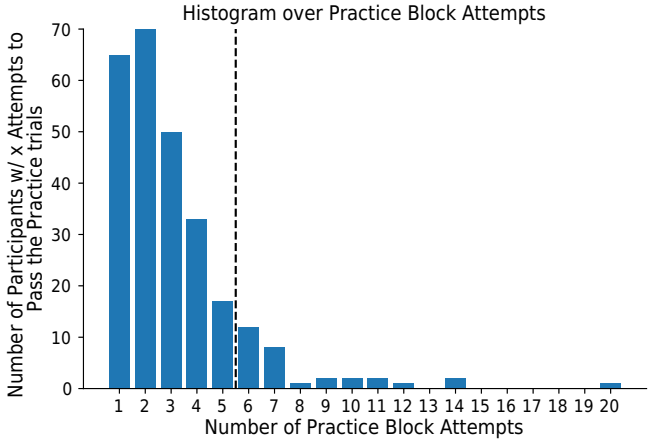
(a) Catch trials from included data. (b) Row variability from included data. (c) Instruction time from included data.



(d) Catch trials from excluded data. (e) Row variability from excluded data. (f) Instruction time from excluded data.



(g) Total response time from included data. (h) Total response time from excluded data.



(i) Practice Block Attempts: We include data from people who needed 5 or fewer attempts.

Figure 18: Distributions of the individual values controlled by the exclusion criteria in the counterfactual-inspired experiment on MTurk. For the first four criteria, a - c and g (d - f and h) show the data for the included (excluded) data. The final criterion in i shows a joint distribution.

A.3 Replication of the Main Result of Borowski et al. [5]

To check whether collecting data on a crowdsourcing platform yields sensible data in our case, we first test whether we can replicate the main finding of our previous human psychophysical experiment on feature visualizations [5]. In the latter, we found in a well-controlled lab environment that natural reference images are more informative than synthetic ones when choosing which of two different images are more highly activating for a given unit. Below, we report how we alter the experimental set-up to turn the lab experiment into an online experiment on MTurk and what results we find.

A.3.1 Experimental Set-up

While keeping as many aspects as possible consistent with our original study [5], we make a few changes: (1) We run an online crowdsourced experiment on MTurk, instead of in a lab. (2) Instead of testing the 45 units used in the original Experiment I, we only test one single branch of each Inception module, namely the 3×3 kernel size. This is a reasonable decision given that the main finding of the superiority of natural over synthetic images was present in all branches and that there was no significant difference per condition between different branches. (3) We exchange the within-participant design for a between-participant design, i.e. one MTurk participant does one condition only, namely either the natural or synthetic reference condition. This version is more suitable for short online experiments. (4) Instead of testing 10 participants in the lab, we test 130 MTurk participants per condition, i.e. 260 in total. This number of participants is estimated with an a priori power analysis using the SIMR package [23] to allow us to detect an effect half as large as the one observed in Borowski et al. [5] 80% of the time. Assumptions about variance, average performance, and effect size are chosen to be conservative relative to the original study because we expect MTurk participants' responses to be noisier.

One HIT on MTurk consists of 1 extensively explained instruction trial, 2 practice trials, and then 9 main trials that are randomly interleaved with a total of 3 catch trials. Each trial type is sampled from a disjoint pool of units: All participants see the same unit for the instruction trial; the catch trials are sampled from the same pool as in the original experiment, and the practice trials are the units that were used as interpretability judgment trials in [5], namely mixed3a, kernel size 1×1, unit 43; mixed4b, POOL, unit 504; mixed5b, 1×1, unit 17. A total of 13 participants see the same main trials that one lab participant saw. The order of the main and catch trials per participants is randomly arranged.

Exclusion Criteria If a participant's response does not meet one or more of the following criteria, which were determined before data collection, we discard it and post the same HIT again:

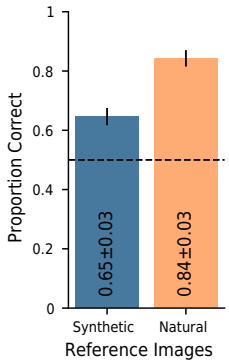
- Performance threshold for catch trials: two out of three trials have to be correctly answered
- Answer variability: at least one trial must be chosen from the less frequently selected side (to discard participants who only responded with "up" or "down")
- Time to read the instructions: at least 15 s
- Time for the whole experiment: at least 90 s and at most 600 s

MTurk compensation Based on an estimated and pilot experiment duration as well as an hourly rate of US$ 15, we calculate the pay to be US$ 1.25. We pay all MTurk participants who fully complete the experiment regardless of whether they succeed or fail in the exclusion criteria. The experiment without pilot experiments costs US$ 447. MTurk participants whose data we include need a mean time of 220.70 ± 71.58 s for the whole experiment, which results in an hourly compensation of ≈ 20.39 US$/hour.

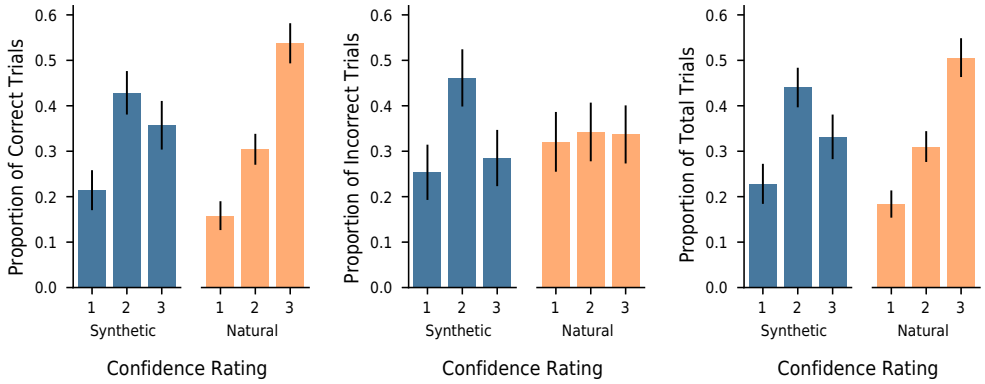
A.3.2 Results

MTurk participants achieve a higher performance when given natural than synthetic reference images: $84 \pm 3\,\%$ vs. $65 \pm 3\,\%$ (see Fig. 19a). Qualitatively, this result is the same as in the original Experiment I, see Figure 16 in Borowski et al. [5]. More precisely, the data shows a 1.35 (2.1) times larger odds (accuracy) difference for the replication. Compared to the lab data, MTurk participants seem more confident on the synthetic condition (see Fig. 19b-d), are faster in the synthetic condition (see Fig. 19e-g), and are about as fast in the natural condition (see Fig. 19e-g).

Fig. 20 shows that most participants passed the exclusion criteria. For more details on the number of postings per HIT and for more details on the MTurk participants' performance on the exclusion criteria, see 21.



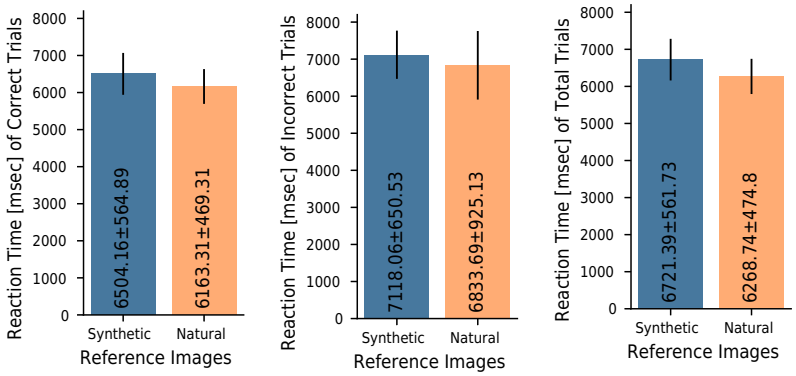
(a) Performance.



(b) Confidence ratings on correctly answered trials.

(c) Confidence ratings on incorrectly answered trials.

(d) Confidence ratings on all trials.

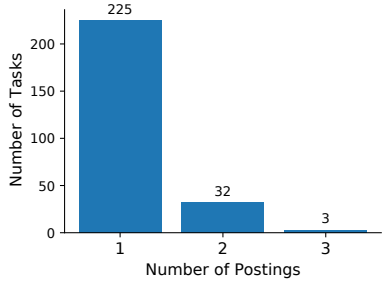


(e) Reaction time on correctly answered trials.

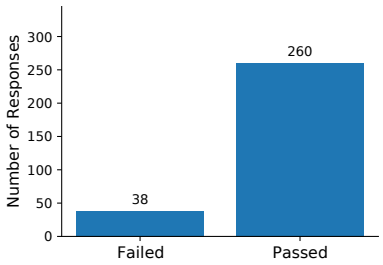
(f) Reaction time on incorrectly answered trials.

(g) Reaction time on all trials.

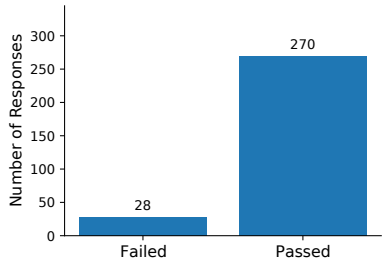
Figure 19: Results of the replication experiment of Borowski et al. [5] on MTurk for kernel size 3×3: task performance (a), distribution of confidence ratings (b-d) and reaction times (e-g).



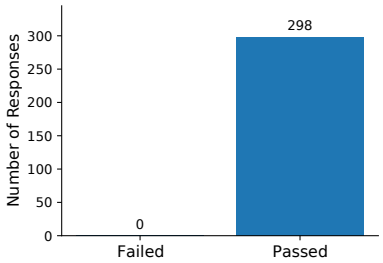
(a) Number of times a HIT is posted.



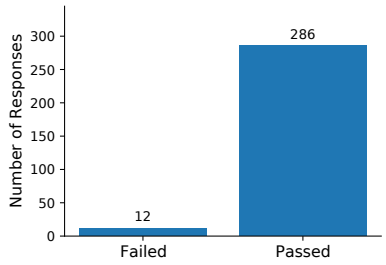
(b) All exclusion criteria.



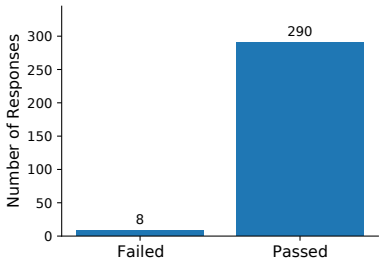
(c) Exclusion criterion: catch trials.



(d) Exclusion criterion: row variability.

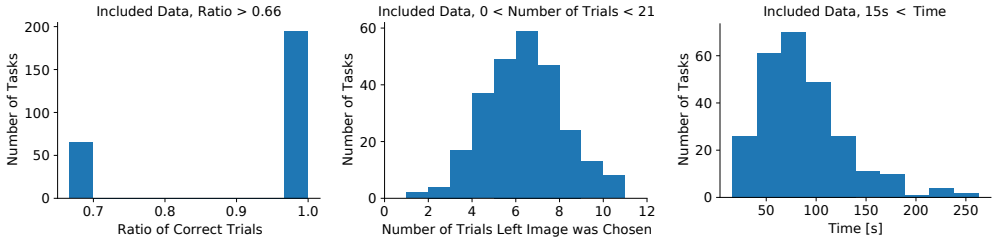


(e) Exclusion criterion: instruction time.

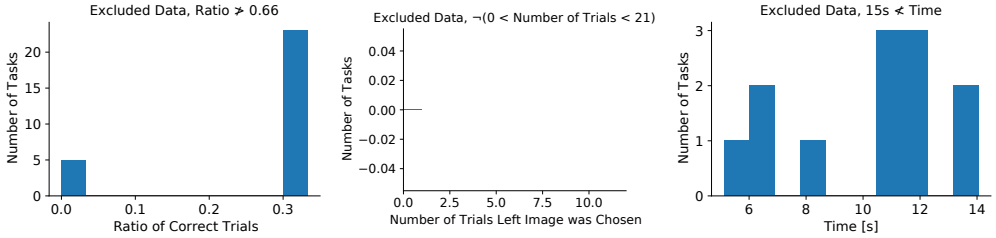


(f) Exclusion criterion: total response time.

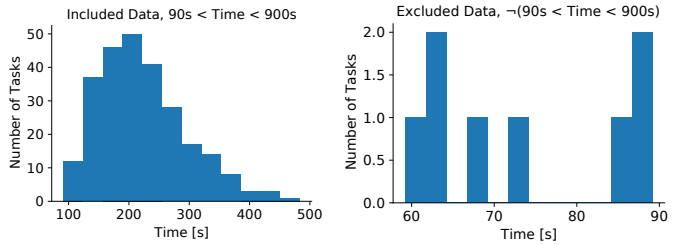
Figure 20: (a) Number of times a HIT is posted. (b-f) Distributions of MTurk participants that passed/failed the exclusion criteria in the replication experiment on MTurk. Note that the sum of the counts of responses for the individual exclusion criteria in c-f is higher than the summary in b because a participant may have failed more than one exclusion criterion.



(a) Catch trials from included data. (b) Row variability from included data. (c) Instruction time from included data.



(d) Catch trials from excluded data. (e) Row variability from excluded data. (f) Instruction time from excluded data.



(g) Total response time from included data. (h) Total response time from excluded data.

Figure 21: Distributions of the individual values controlled by the exclusion criteria in the replication experiment on MTurk. Figures a - c and g (d - f and h) show the data for the included (excluded) data.

Publication 4: Interactive Analysis of CNN Robustness

Stefan Sietzen, Mathias Lechner, Judy Borowski, Ramin Hasani, Manuela Waldner. Pacific Graphics 2021.

This paper is not included in the presented thesis. I contributed by participating in the expert user study and editing the manuscript.

Pacific Graphics 2021
E. Eisemann, K. Singh, and F.-L Zhang
(Guest Editors)

Volume 40 (2021), Number 7

Interactive Analysis of CNN Robustness

Stefan Sietzen[1], Mathias Lechner[2], Judy Borowski[3], Ramin Hasani[1,4], and Manuela Waldner[1]

[1]TU Wien, Vienna, Austria
[2] Institute of Science and Technology Austria (IST Austria), Klosterneuburg, Austria
[3] University of Tübingen, Germany
[4] Massachusetts Institute of Technology (MIT), Cambridge, USA

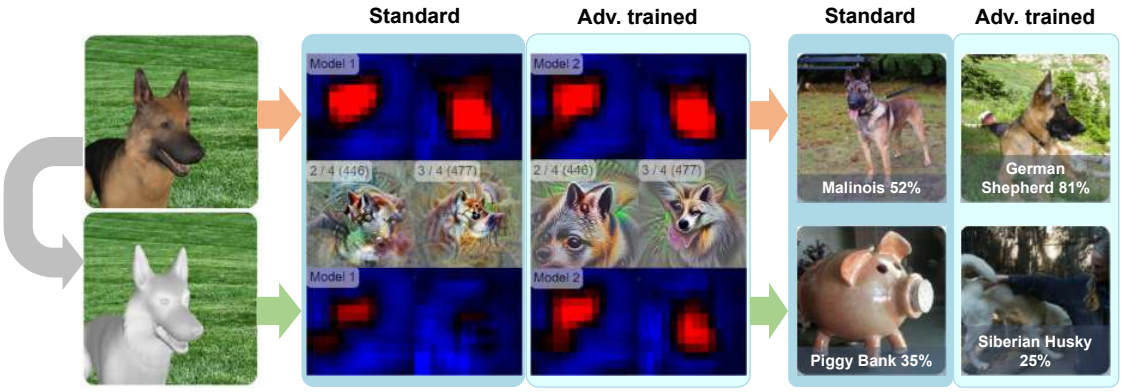


Figure 1: *Rapid interactive comparison of two models' network responses through activations of two selected "dog-relevant" neurons (middle) and predicted target classes (right) to input scene perturbations of a dog (left): Removing the texture leaves the 3D model with a smooth surface and causes a standard model to identify a piggy bank instead of a dog, while an adversarially trained model identifies a white dog breed (Siberian Husky). Activations of dog-relevant neurons in the standard model decrease due to missing texture information.*

Abstract

While convolutional neural networks (CNNs) have found wide adoption as state-of-the-art models for image-related tasks, their predictions are often highly sensitive to small input perturbations, which the human vision is robust against. This paper presents Perturber, a web-based application that allows users to instantaneously explore how CNN activations and predictions evolve when a 3D input scene is interactively perturbed. Perturber offers a large variety of scene modifications, such as camera controls, lighting and shading effects, background modifications, object morphing, as well as adversarial attacks, to facilitate the discovery of potential vulnerabilities. Fine-tuned model versions can be directly compared for qualitative evaluation of their robustness. Case studies with machine learning experts have shown that Perturber helps users to quickly generate hypotheses about model vulnerabilities and to qualitatively compare model behavior. Using quantitative analyses, we could generalize users' insights to other CNN architectures and input images, yielding new insights about the vulnerability of adversarially trained models.

CCS Concepts

• ***Human-centered computing*** → ***Visualization systems and tools;*** • ***Computing methodologies*** → ***Machine learning;***

1. Introduction

Convolutional neural networks (CNNs) achieve impressive results in a variety of applications, such as image classification [KSH17] or object detection [RHGS17]. The performance of a CNN trained on a given training data set is typically assessed in terms of prediction accuracy on a held-out validation dataset. If the statistical distributions of the training and validation set differ, the high performance can drop precipitously. In that case, the model is not robust

against the variability within the validation data. CNN robustness has been shown to be affected by image perturbations such as cropping [ZSLG16], blur [DK16], high-frequency noise [YLS*19], or texture variations [GRM*19]. Prominent examples of noise perturbation are adversarial attacks: Single pixels of an input image are changed slightly such that a CNN misclassifies it [SZS*14]. To humans, these changes are typically imperceptible, and they would still assign the correct labels.

Robustness is a highly safety-critical aspect of CNNs in various applications, such as self-driving cars [LHA*20]. For example, researchers have demonstrated how targeted yet unsuspicious changes of traffic signs can cause CNNs to consistently miss stop signs [EEF*18]. Understanding which factors influence the robustness of CNNs and, consequently, designing and evaluating more robust models are therefore research topics of central importance in the machine learning community [GSS14, SZS*14, MMS*19, LHG*21].

To create a basis for tackling robustness, researchers aim to gain a better general understanding of which image features CNNs are most sensitive to and how this sensitivity differs from human visual perception [BHL*21]. For example, Geirhos et al. [GRM*19] investigated the influence of shape versus texture on humans and CNNs in image classification and found a strong bias for texture in CNNs. This texture bias was also confirmed by Zhang and Zhu [ZZ19] by evaluating the effect of image saturation or random shuffling of image patches. Through systematic analysis of image perturbation in the Fourier domain, Yin et al. [YLS*19] could show that CNNs are highly sensitive to high-frequency perturbations.

A common way to improve robustness is to train models on more variable input images, using data augmentation methods [PW17], image stylization [GRM*19], or adversarial examples [MMS*19]. To evaluate the performance of these improved models, they are then compared to standard CNN models. To perform these analyses, researchers automatically perturb images offline, log the predictions for each perturbed input image, and compare the results between models. According to our collaborating domain experts, such experiments take several hours to set up and include time-consuming parameter searches. For some experiments, it is not even possible to determine network performance fully automatically as the ground truth of the input images also becomes unclear for humans because of the performed image perturbations. Clearly, these factors severely limit machine learning experts to quickly explore factors of image perturbation that may impact CNN performance.

We hypothesize that *being able to interactively manipulate a synthetic input scene with a large and diverse set of visual perturbation parameters and observing the changing activations and predictions instantly will allow researchers to quickly generate hypotheses and build a stronger intuition for potential vulnerabilities of CNNs*. In this work, we therefore introduce Perturber – a novel real-time experimentation interface for exploratory analysis of model robustness (see Figure 1). Our work provides the following contributions:

1. Interactive input perturbations of 3D scenes in combination with feature visualizations, activation maps, and model predictions as a novel approach to interactively explore model robustness.
2. A direct comparison interface for qualitative visual validation of more robust, fine-tuned model variants.
3. Implementation of a publicly accessible web-based VA tool[†] that supports a large variety of perturbation methods of 3D input scenes and visualizes the model responses in real-time.
4. Observations from case studies with machine learning experts demonstrating that live inspection of input perturbations allows experts to visually explore known vulnerabilities, to compare model behaviors, and to generate new hypotheses concerning model robustness.
5. Quantitative verification of selected user observations from the case study shedding new light on the robustness of adversarially trained models.

2. Related Work

In recent years, a wide spectrum of deep learning visualization methods has emerged. For a comprehensive overview of visual analytics (VA) for deep learning, we refer the reader to a survey by Hohman et al. [HKPC19]. For example, graph structure visualizations, such as the *TensorFlow Graph Visualizer* [WSW*18], help users to get a better understanding of their models' structure, which comprise numerous layers and connections. Others, such as *DeepEyes* [PHVG*18] and *DeepTracker* [LCJ*19], track detailed metrics throughout the training process to facilitate the identification of model problems or anomalous iterations. *ExplAIner* [SSSEA20] goes beyond monitoring and also integrates different steering mechanisms to help users understand and optimize their models. A case study using a VA system to assess a model's performance to detect and classify traffic lights has shown that interactive VA systems can successfully guide experts to improve their training data [GZL*20].

While these examples all focus on the inspection of a single model, others support model comparisons. For example, using *REMAP* [CPCS20], users can rapidly create model architectures through ablations (i.e., removing single layers of an existing model) and variations (i.e., creating new models through layer replacements) and compare the created models by their structure and performance. Ma et al. [MFH*20] designed multiple coordinated views to help experts analyze network model behaviors after transfer learning. *CNNComparator* [ZHP*17] uses multiple coordinated views to compare the architecture and the prediction of a selected input image between two CNNs. Model comparison is also a crucial aspect of our work. However, the focus of the present work lies on fluid modification of input stimuli and instantaneous analysis and comparison of the network responses. Thus, we let users *generate and perturb* input images from 3D scenes in a "playground-like" manner rather than letting the user select input images with a ground-truth class label.

Interactive "playgrounds" require relatively little underlying deep learning knowledge and can be used for educational purposes. For more informed users, they are valuable for building an intuition and validating knowledge from literature [KTC*19]. Notable examples are *TensorFlow Playground* [SCS*17], which supports

[†] http://perturber.stefansietzen.at/

the interactive modification and training of DNNs in the browser, and *GANLab* [KTC*19], which is designed in a similar style and supports experimentation with Generative Adversarial Networks. *CNN Explainer* [WTS*20] provides a visual explanation of the inner workings of a CNN by showing connections between layers and activation maps, allowing the user to choose the input from a fixed collection of images. More similarly to our work, Harley [Har15] provides an online tool where users can draw digits onto a canvas. Then, the responses of all neurons in a simple MNIST-trained network are visualized in real-time. *Adversarial Playground* [NQ17] allows users to interactively generate adversarial attacks and instantly observe the predictions of a simple MNIST-trained network. To support interactive probing of model responses based on input modifications, *Prospector* [KPN16], the *what-if-tool* [WPB*20], and *NLIZE* [LLL*18] allow users to interrogate the model by varying the input in the domains of tabular data and natural language processing, respectively. These works inspired us to build a system that lets users *interactively explore* model robustness through input perturbations. In contrast to prior work, Perturber operates on complex CNN models, such as Inception-V1 trained on ImageNet, and can be used to discover and explain vulnerabilities to complex input scenes, like animals or man-made objects in different environments.

To explain a CNN model, there are powerful methods to reveal the role of a network's hidden units by visualizing their learned features. *Feature visualization* is an activation maximization technique, which was improved by combining a variety of regularization techniques [YCN*15, OMS17]. Feature visualizations have been used to identify and characterize causal connections of neurons in CNNs [CCG*20], and to comprehensively document the role of individual neurons in large CNNs [Ope]. Within VA tools, feature visualizations have been used to compare learned features before and after transfer learning [MFH*20] or to visualize a graph of the most relevant neurons and their connections for a selected target class [HPRC20]. Similarly, *Bluff* [DPW*20] shows a graph containing the most relevant neurons explaining precomputed adversarial attacks, where neurons are represented by their feature visualizations.

Other powerful interpretability methods are *saliency maps* (or *attribution maps*), which show the saliency of the input image's regions with respect to the selected target class or network component [SVZ13]. Saliency maps and other gradient-based methods like *LRP* [LBM*16], *Integrated Gradients* [STY17], or *Grad-CAM* [SCD*17], however, require a computationally expensive backpropagation pass. A very simple solution to reveal relevant image regions for a model's prediction is to directly visualize the forward-propagation activations of selected feature maps in intermediate layers. For example, the *DeepVis Toolbox* [YCN*15] shows live visualizations of CNN activations from a webcam feed. The goal is to get a general intuition what features a CNN has learned. *AEVis* [LLS*18] shows activations of neurons to a pre-defined set of input images along a "datapath visualization". This visualization allows users to trace the effects of adversarial attacks through the hidden layers of a network. Datapaths are formed by critical neurons and their connections that are responsible for the predictions. Like the *DeepVis Toolbox* [YCN*15], Perturber visualizes activations and predictions based on live input. The major difference is that Perturber generates input images from an interactive 3D scene and provides a rich palette of input perturbation methods to gradually explore potential vulnerabilities of a model. In addition, it facilitates direct comparison between a standard model and a more robust variation thereof to explore the benefits and limitations of model variations.

3. Perturber Interface

The high-level goal of Perturber is to interactively explore potential sources of vulnerabilities for CNNs to facilitate the design of more robust models. Through constant exchange between visualization researchers and machine learning experts investigating model interpretability and robustness, we identified four central requirements of a VA application to support exploratory analysis and qualitative validation of model robustness:

R1: Rich **online input perturbation** of a representative scene is essential to quickly generate input images for which model responses can be investigated. The system should provide a large variety of possible perturbation parameters and allow a flexible combination of these perturbations.

R2: To understand what makes a model robust, it is not sufficient to understand *how* a model responds to the perturbed input but also *why* the model's responses change. To this end, looking not only at the input and output, but particularly at the numerous **hidden layers** is inevitable when trying to understand what makes a model react unpredictably for humans.

R3: Interactivity can help users to build intuitions through dynamic experiments [KC19]. We thus aim to make the model responses to input perturbations **instantly** visible to support a fluid feedback loop in a playground-like environment.

R4: Robustness can be achieved by training models on more versatile input images, using data augmentation methods, such as affine image transformations, image stylization, or adversarial training. A direct visual **comparison** between the standard CNN model and its more robust fine-tuned version are necessary for a first qualitative validation whether a fine-tuned model is generally more robust to input perturbations or only selectively more robust to specific perturbations it has been trained on.

Perturber supports these four core requirements through a highly interactive web-based playground consisting of the following components: the 3D input scene (Section 3.1, Figure 2 A), the perturbation control (Section 3.2, Figure 2 B-C), the prediction view (Section 3.3, Figure 2 F), and the (comparative) neuron activation view (Section 3.4, Figure 2 E-D). These views allow for a qualitative inspection of the effects of input scene perturbations on one or two selected models (Section 3.5).

Conceptually, Perturber can handle any CNN architecture. The current version supports models based on the Inception-V1 (also called GoogLeNet) [SLJ*15] architecture. Our standard model was trained on ImageNet [RDS*15]. We use a pre-trained model with weights accessible through the *Lucid*[‡] feature visualization library [OMS17].

[‡] https://github.com/tensorflow/lucid

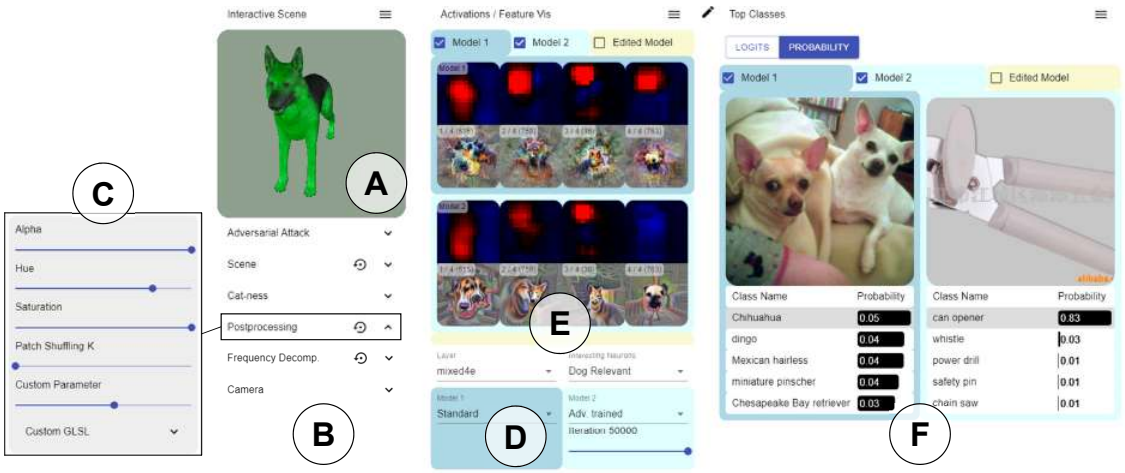


Figure 2: *Perturber interface: users manipulate the 3D scene (A) through a large variety of input perturbation methods that are grouped into functional categories (B). Per category, input perturbations can be seamlessly controlled through sliders (see inset (C)). Users can select two models to compare, as well as a set of neurons from dedicated layers (D). For the selected neurons and models, respectively, neuron responses are visualized live (E). Top-5 predictions for both models are shown as logits or probabilities along with an image example (F).*

3.1. Input Scene

The input image is generated from an interactive 3D scene that can be manipulated in numerous ways (Section 3.2). In computer vision projects, rendered images are often used instead of – or in addition to – photographs for training CNNs [SQLG15]. Quantitative experiments have shown that, for high-level computer vision algorithms, the gap between synthetic and real images is fairly small [GWCV16]. CNN responses were shown to be consistent between simple rendered 3D models and natural images [AR15]. Interactive exploration of CNN responses based on synthetic scenes therefore seems to be sufficiently representative of real-world applications. The major advantage of a synthetic 3D scene is that perturbation factors can be easily disentangled. That means, input modifications can be flexibly applied independently from each other to assess their isolated effects as well as their interactions.

Out of the 1000 ImageNet classes, 90 of them are dog breeds. Therefore, there are numerous neurons in our Inception-V1 model that specialize in dog-related patterns, such as snouts, head-orientations, fur, eyes, ears, etc. As a consequence, we chose a dog as our main foreground 3D object to represent this special significance and to feed an input image that many hand-identified neuron circuits [CCG*20] respond to strongly. As a second 3D model, we provide a vehicle, which has very different characteristics from a dog and is a commonly analyzed object in the machine learning community (e.g., [GWCV16, AR15]). In addition, users can upload custom 3D models. This way, users can verify observations they have made also with other models.

3.2. Perturbation Control

The core principle of Perturber is that the user can flexibly vary the perturbation factors they want to explore (R1) and observe their effects instantly (R3). Below the input scene view, Perturber provides sliders for seamless control of the perturbations (Figure 2 C). Perturber provides more than 20 scene perturbation factors, as well as all possible combinations between these factors. We group these perturbation methods into the following categories:

Geometry perturbations, such as rotation, translation, or cropping an image, are classic perturbations a CNN might be sensitive to [ZSLG16, ACW18, ETT*19]. Therefore, affine image transformations are also a common data augmentation method for more robust training [PW17]. We support geometric perturbations through simple camera controls, which allow the user to arbitrarily orbit around the 3D model, as well as to freely zoom and pan the scene to create image cropping effects. In addition, users can rotate the camera around the z-axis to simulate image rotation.

Scene perturbations that may seem irrelevant for human observers may easily confuse a standard CNN model. For example, changing the background [XEIM20] can have a tremendous effect on the prediction. We, therefore, support various scene perturbation operations, such as changing the background image, as well as modifying the lighting and texture parameters of the main object (see Figure 3). Through combinations of these parameters, specialized perturbations, such as silhouette images, can be generated (see Figure 3 bottom right).

Shape perturbations let users morph the shape of the main scene object into another one. Perturber currently supports morphing between dog and cat, as well as between a firetruck and a race car. By morphing the shape of an object independently from its texture, the texture-shape cue conflict [GRM*19] can be investigated.

Color perturbations act on the rasterized 2d image. We support individual post-processing effects, such as alpha blending to black, hue shifting, and (de-)saturation. In addition to these parameters,

Figure 3: *Scene perturbations – first row: original scene, no lighting influence, full background blur, different background image; second row: no texture influence, full texture blur, lowest background saturation, combination of background saturation, lighting influence, texture influence, and background blur. All perturbation parameters can be controlled seamlessly.*

users are provided with a text field where they can write their own GLSL code, taking a single parameter which is controlled by the slider. The code snippet defaults to code for contrast adjustment. Using these post-processing effects, users can assess the models' surprisingly high robustness against low contrast [DK16] and low image saturation [SLL20].

Frequency perturbations selectively modify different image frequencies. Perturber provides three-parameter frequency decomposition that splits the image into low and high-frequency bands, which is achieved by Laplacian decomposition. Through selective frequency suppression, users can investigate phenomena such as the one described by Yin et al. [YLS*19], where the authors show that adversarially trained networks are robust against high-frequency perturbations but very sensitive to low-frequency perturbations.

Spatial perturbations are post-processing effects systematically changing the image's pixel order. Perturber supports patch shuffling, which was used by Zhang and Zhu [ZZ19] to reveal a model's sensitivity to global structure, which gets highly disturbed for human observers by patch shuffling. The image is divided into a grid of $k \times k$ cells, which are then randomly re-ordered.

Adversarial perturbations, finally, are a core feature to understanding model robustness. Users can perform projected gradient descent (PGD) adversarial attacks [MMS*19] on the current scene image. To do that, they choose a model to generate the attack from, a target class or the option to suppress the original prediction, as well as the attack ε and L_p norm (L_2 or L_∞). With each button press, the user performs one PGD step with a step length of $\varepsilon/8$. This perturbation is costly and cannot be performed instantaneously. The first step typically takes around 10 seconds on a powerful consumer PC as the gradient function needs to be computed for the current input image first. To let users inspect the effect of an adversarial attack on the model fluidly like the other perturbation parameters, we overlay the current input image with the perturbation vector once it has been computed. This way, users can interactively fade the attack alpha using a slider and observe the system's response instantaneously. They can also fade the original image to inspect the perturbation vector itself.

3.3. Prediction View

The prediction view shows the top-5 classification results for each model (Figure 2 F) either as probability or as logits. This allows the user to observe the classification changes resulting from input perturbations in real-time (R3). Each model's top result is represented by a class image example. Perturber shows the first image of the respective class in the ImageNet validation dataset.

3.4. Neuron Activation View

The neuron activation view is the central interface for the analysis of how input perturbations affect the models' hidden layers (R2). Perturber represents neurons through feature visualizations. Feature visualizations aim at providing a lens into networks to visualize what patterns certain network units respond to [OMS17]. These patterns might be, for example, edges in a particular direction in earlier layers (cf., Figure 4), or specific objects, like dog heads (cf., Figure 1) or car windows, in later layers.

Activation maps show which regions of the input image cause the respective neuron to be activated. After experimenting with an implementation of the gradient-based visualization method *Grad-CAM* [RSG16], we decided to focus on the computationally far less expensive activation maps showing the neuron activations for a forward propagation pass. This way, neuron activations can be observed instantly (R3). Input regions that highly activate the neuron are shown in red, while blue regions cause a negative activation (see Figure 4). As the user manipulates the input scene, the activation maps update instantaneously so that the user can observe in real-time to which image features a neuron responds in one of the models. For example, in Figure 4, neurons 412 and 418 of layer mixed4a respond to oriented patterns. Rotating the textured race car causes these two neurons to strongly change their activations.

Inception-V1 contains thousands of neurons. Clearly, it is impossible to show feature visualizations and interactive activation

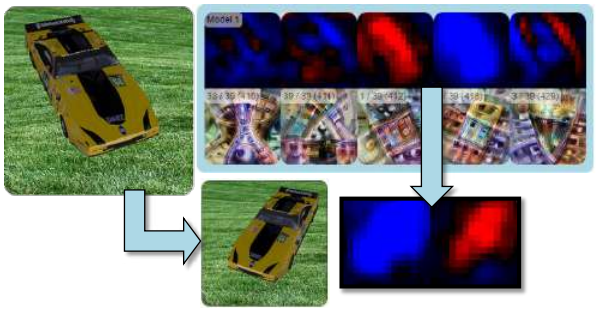


Figure 4: *Feature visualizations for neurons associated with complex shapes and curvatures in layer mixed4a for the standard model, as well as their activation maps for the input on the left. Note how rotating the input model causes an activation change for oriented shape detectors (insets on the bottom).*

maps for all neurons concurrently. Instead, we provide pre-selected neuron groups and categories for each layer, which were characterized into semantically meaningful neuron families in the course of the *Circuits* project [CCG*20]. In addition to these previously presented neuron groups, we identified further sets of dog-, cat-, race car-, and firetruck-related neurons in later layers using *Summit* [HPRC20]. These pre-selected neuron groups consisting of one to up to around 70 neurons can be selected from drop-down menus (Figure 2 D).

3.5. Model Comparison

After exploring potential sources of vulnerability, model designers commonly fine-tune existing models using augmented training data covering the identified vulnerability. To leverage Perturber for direct comparison of robustness (R4), we employ a transfer learning approach, which is initialized with the weights from the standard model. This is necessary to guarantee that feature correspondence of individual network units is kept intact. Only this way, we can assume that units across models respond similarly to identical input images.

To illustrate what individual network units respond to, Perturber relies on feature visualizations. Feature visualizations are independent of the input image, but they differ between models. Indeed, Table 1 shows noticeable differences between a single neuron's feature visualizations of three different model variations.

To showcase model comparison, Perturber already includes two robust model variations of Inception-V1: The first model was adversarially trained with projected gradient descent (PGD), which was described by Madry et al. [MMS*19]. Adversarial training increases model robustness by incorporating strong adversarial attacks into the training procedure. The second model was trained by Stylized-ImageNet, which is a variation of ImageNet, where images were transformed into different painting styles using a style transfer algorithm [GRM*19]. The purpose of Stylized-ImageNet was to induce a shape bias, while the standard ImageNet-trained models were found to be biased unproportionally towards texture [GRM*19]. Users can choose two models for comparison from the interface (Figure 2 D). In addition, they can choose multiple checkpoints along the incremental fine-tuning process to analyze the development of the models during training. Corresponding neuron activation views are then juxtaposed in the Perturber interface for direct comparison (Figure 2 E).

4. Web-Based Implementation

Perturber runs purely on the client-side in the user's browser and without any server-side computations at runtime. We precomputed all data that does not depend on interactively changeable elements to make the UI as efficient as possible. For transfer learning, we initialized the weights with those of the standard model (i.e., Inception-V1 trained on ImageNet), which were obtained through the *Lucid* library [OMS17]. For adversarial fine-tuning, we used the open-source implementation by Tsipras et al. [TSE*19]. No layers were frozen for transfer learning. For both fine-tuned models, we used a learning rate of 0.003, a batch size of 128, and we trained until the models reached a top-5 training error of around 0.5, which took approximately 50K iterations for the adversarial fine-tuning and around 90K iterations for the Stylized-ImageNet fine-tuning.

Table 1: *Comparison between naive pixel (top) and Fourier basis (bottom) parametrized feature visualizations of neuron 222 in layer mixed4a for three model variations.*

	standard	**adversarial**	**stylized**
naive pixel param.			
Fourier basis param.			

To generate feature visualization, we employed the implementation provided by the *Lucid* [OMS17] library. For each model, we generated feature visualizations for 5808 neurons from the most relevant layers (i.e., the first three convolutional layers and the concatenation layers at the end of each mixed-block). During fine-tuning, we obtained feature visualizations for 17 checkpoints of adversarial fine-tuning and seven checkpoints of Stylized-ImageNet fine-tuning. For each neuron, we computed feature visualizations using two parametrization methods – naive pixel or Fourier basis [OMS17] (see Table 1). The second performs gradient ascent in Fourier space and leads to a more equal distribution of frequencies, resulting in more naturally looking feature visualizations. We use *transformation robustness* [OMS17] in addition to both methods. Without Fourier basis parametrization, the differences between the models are more visually distinct (Table 1 top row). The computation of all ∼300K generated feature visualizations required around one month on a machine with two NVIDIA GTX 1070 GPUs. [§].

Manipulation of the 3D scene, model inference based on the rendered image, and computation of activation maps are performed live in the web browser. The client relies on GPU acceleration for both, 3D scene rendering and CNN inference. We use the WebGL-based libraries *Three.js* and *TensorFlow.js* [STA*19] for these tasks, respectively. The front-end GUI is based on *React.js*. Input perturbations based on post-processing effects are implemented as multiple sequential render passes with custom GLSL shaders. For model inference, we use *Tensorflow.js* [STA*19], which enables fast GPU-accelerated CNN inference. *Tensorflow.js* is also used for computing adversarial attacks.

A major requirement of Perturber is that the effects of input perturbations can be observed instantaneously (R3). To assess requirement R3, we measured the client's performance while constantly orbiting the camera around the object. Figure 5 shows the recorded frame rates for two client notebooks: a MacBook Pro 13" 2018 with Intel Iris Plus Graphics 655 (MBP) and an AORUS 15G Gaming Notebook with an NVIDIA GeForce GTX 2080 Super GPU

§ All generated feature visualizations can be downloaded from https://github.com/stefsietz/perturber/

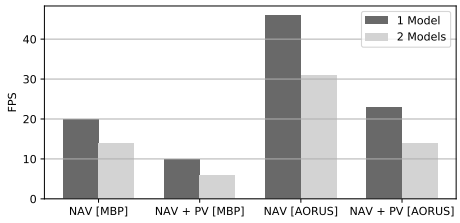


Figure 5: *Performance benchmarks for four different output configurations, measured on two machines (MBP or AORUS): neuron activation view (NAV) only, or neuron activation view in combination with prediction view (NAV+PV) visualized for one or two visualized models concurrently.*

(AORUS). It is clearly visible that the GPU of the client machine has a strong influence on the frame rate. The application performance also depends on the enabled visualizations, with the prediction view (Section 3.3) being more computationally expensive than showing four live activation maps (Section 3.4). When using less powerful client machines, users can selectively switch off views.

5. Exploration Scenario: Texture-Shape Conflict

To demonstrate the capabilities of Perturber, we selected one important aspect of robustness: the influence of shape and texture on CNN classification results. It has been shown that CNNs are biased towards texture, which affects robustness [GRM*19]. In this scenario, we first aim to investigate whether this effect could have been discovered using Perturber. Secondly, we aim to qualitatively validate whether a more robust model variation, which was fine-tuned using Stylized-ImageNet [GRM*19], can improve upon this bias.

For our first analysis, we explore the texture-shape cue conflict through shape and texture perturbations. Figure 6 illustrates that shape perturbations alone do not cause the standard model to predict a cat breed. However, when morphing the texture, the model predictions switch to cat breeds – even when the object shape remains unchanged. This is a first indicator that the model is indeed much more sensitive to texture than to shape perturbations.

To further investigate the texture sensitivity of the standard model, we replicated the patch shuffling experiment by Zhang and Zhu [ZZ19]. Using Perturber, we can inspect single neurons' activations during this experiment. We illustrate our observations on a hand-picked neuron in Figure 7, which is strongly activated by dog faces looking to the left. Note how this neuron is activated by strong texture contrasts, especially around the mouth of the dog. Image regions containing ears do not activate this particular neuron. For this scene, the standard model still predicts a dog breed up to $k = 7$ randomly shuffled image patches. This illustrates that the decomposed shape has indeed very little influence on the model prediction.

In the next step, we analyze if the standard model is indeed more sensitive to texture than the model fine-tuned on Stylized-ImageNet [GRM*19] by gradually removing the texture of the

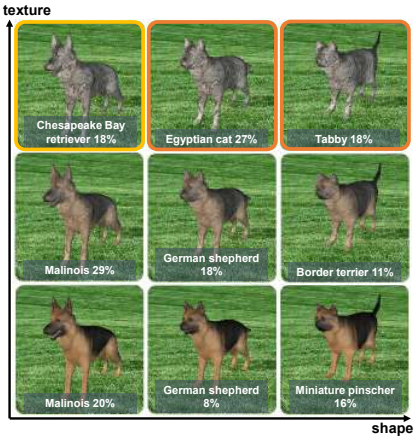


Figure 6: *Object morphing between dog (bottom left) and cat (top right) along the texture dimension (y axis) and shape dimension (x axis): Top-1 standard model predictions with their probabilities are shown as text labels. Input images leading to cat breeds within the top-5 or top-1 predicted classes are indicated by a yellow and orange frame, respectively (top row).*

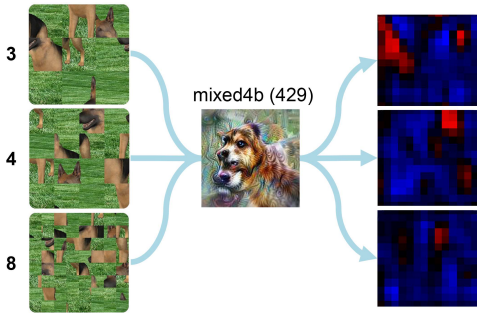


Figure 7: *Randomly shuffling the dog scene into 3 (top), 4 (middle), and 8 (bottom) patches: A dog-related neuron in mixed4b of the standard model is activated by patches containing parts of the dog's face and textured body parts.*

main object. The predictions in Figure 8 confirm that the Stylized-ImageNet trained model consistently predicts a dog breed, even in the absence of a texture. The standard model, on the other hand, seems to rely much more on the texture. The model trained with Stylized-ImageNet is also more sensitive to patch shuffling, which is another indication that it relies less on texture information than the standard model (see Section A in the Supplemental Document for image examples).

Finally, we compare shape sensitivity between the two models. To this end, we combine various scene perturbations to generate a silhouette image of the dog. We then gradually change the pose of the dog. Figure 9 shows the predictions of the standard model and the model fine-tuned with Stylized-ImageNet. Clearly, the predictions of the standard model are unstable, especially when the

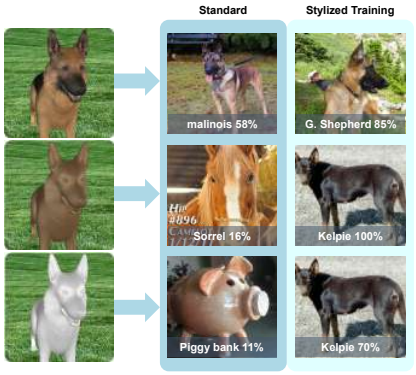


Figure 8: *Inspecting the shape vs. texture conflict through scene perturbations for the standard (left) and the Stylized-ImageNet trained model (right): while the standard model gets confused by missing textures, the Stylized-ImageNet trained model predicts a dog breed even if the 3D model is untextured.*

dog is directly facing the camera. The model trained with Stylized-ImageNet, however, reliably predicts a white wolf with high probability. The activations of relevant neurons indicate that the Stylized-ImageNet-based model has learned a more stable representation of a frontal dog head, which also get activated by a silhouette image.

These examples illustrate that Perturber provides flexible ways to explore and better understand potential threats to robustness, such as the texture-shape conflict of CNNs. Section A in the supplemental document contains more exploration scenarios.

6. Case Studies

While the previous scenario investigated a known vulnerability, we further explored whether Perturber can help to discover unknown threats to robustness through case studies with machine learning experts. We conducted case studies with five machine learning researchers (four PhD students, one post-doctoral researcher; one female, four males). Except for one user, sessions were conducted individually through video conferencing and lasted approximately one hour each. One user preferred to perform the case study offline and sent a textual text report instead.

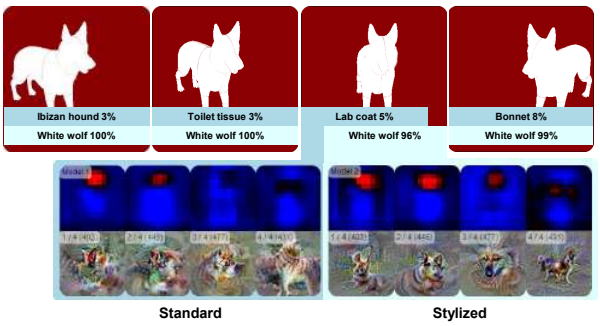


Figure 9: *Rotating a dog silhouette: the corresponding predictions of the standard model and the model trained on Stylized-ImageNet, as well as activations of dog-related neurons for the front-facing dog image in mixed4d of the standard model (left) and the robust model (right).*

The researchers cover various topics of expertise in the field of AI interpretability and the design of robust machine learning models. The concrete research areas are listed in Section B of the supplemental document. Two users were involved in the co-design process of Perturber and are co-authors of the paper. Three users were unfamiliar with the system before the evaluation. One of these users entered the co-design iteration process after the case study and is also a co-author. Paper co-authorship is a common collaboration role in design studies [SMM12].

During the video conference, the participant used the online tool while sharing his/her screen with the first and last author. Every session was recorded while conversations and observations were transcribed on-the-fly or in retrospective through automatic speech-to-text.

Every video conferencing session started with a short introduction by the participant describing his/her research focus. Afterwards, we gave a short demonstration of Perturber's features. We then asked the participant to shortly comment on his/her first impression and his/her expected insights from the analysis. Then, the participant freely played around with Perturber while thinking aloud. In particular, we asked the user to always state his/her intent before performing an action and whether he/she would have any particular hypotheses about the response and behavior of the network based on the chosen input. If a user could not find the respective functionality of the tool, the first and/or last author provided oral assistance. At the end of the study, the user was encouraged to summarize his/her impressions, the potential benefits of the tool, and to provide suggestions for improvements.

6.1. Observations and Feedback

Users praised the fact that Perturber works "live" and therefore allows for ad-hoc **exploratory analyses**. They liked that Perturber allows to play around with simple examples to quickly find patterns and form hypotheses. Generally, the main focus of the exploratory analyses was trying to identify input perturbations where a model would respond unexpectedly. One user described this process as trying to answer the following questions: *"How can I break a model? What do I need to do so that the resulting prediction is wrong?"* For example, three users were surprised to see how vulnerable the adversarially trained model seems to be to some geometric changes, such as zooming or rotating. Two users also discovered a high sensitivity of adversarially trained models to background modifications, as illustrated in Figure 10. The live approach was praised in particular for cases without a clear ground truth, such as object morphing (Figure 6). Such scenarios would not be possible to assess quantitatively without human subject studies. Being able to quickly generate hypotheses that could then be formally tested in a more controlled setting was considered very useful.

We also observed indications that Perturber is practically helpful

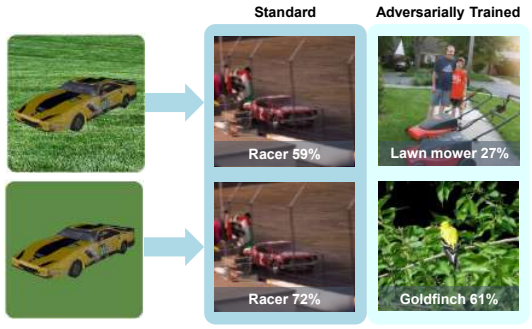


Figure 10: *Blurring the background of the race car has little influence on the standard model's predictions (left). The adversarially trained model (right) is very sensitive to the chosen background. The probability for "racer" is 11% with a grass background and 7% with a uniform green background.*

for **visual confirmation**. For example, using the model predictions and activations of selected neurons, one user verified that adversarial attacks only affect the standard model. He also observed how animal-related neurons of the standard model got activated during an attack with the target class "badger" (Figure 11). Not all assumptions were actually confirmed. For example, one user expected that the model trained on Stylized-ImageNet would be noticeably more vulnerable to image blur than the standard model. Unexpectedly, the model's responses were not more sensitive to blur than the standard model's for the chosen input scene (see Section A in the supplemental document).

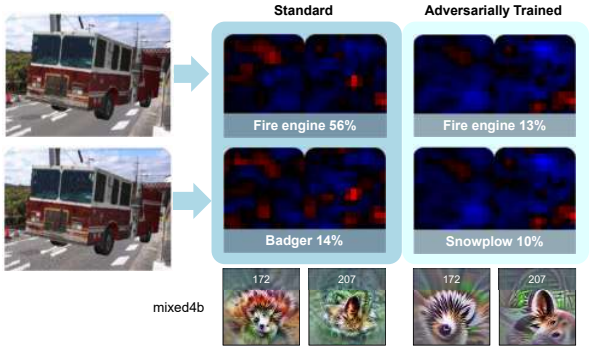


Figure 11: *Adversarial attack with target class "badger" on the fire truck scene (top) and the effect on two cat-related neurons in layer mixed4b for the standard model (left) and adversarially trained model (right): As the attack strength is increased (bottom), activations of some standard model neurons increase while the activations of the adversarially trained model are hardly affected. The prediction probabilities for the fire engine decrease to 0.1% for the standard model and to 6% for the adversarially trained model after the attack.*

Finally, a user pointed out that playing around with Perturber would **let non-experts get an intuition** how easily CNNs are fooled. For example, one user demonstrated how the standard model sometimes rapidly changes its predictions during a simple translation of the dog. Another expert demonstrated that a rotation of the dog along the z-axis in combination with an unusual background (street) was sufficient to disturb the adversarially trained model. Accordingly, he stated that Perturber could be informative for "people fearing that AI will take over the world".

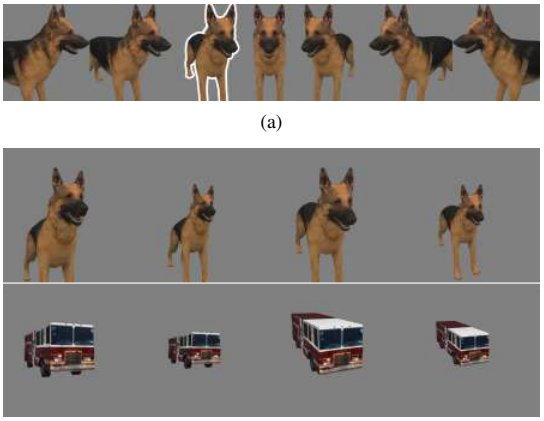
A comprehensive list of observations reported by the individual users can be found in Section B of the supplemental document. Overall, the most frequently performed perturbations leading to the most interesting observations in our case studies were 1) geometric transformations, such as object rotation, zooming, and translation, 2) modification of the background, 3) combinations of (1) and (2), as well as 4) object morphing. The prediction view was perceived as giving instant, easily interpretable, and useful feedback. It was thus the primary view to observe model behavior. Feature visualizations were considered useful to characterize the difference between the models. Participants described feature visualizations of the adversarially trained model as more "intuitive" or "cartoonish" compared to the corresponding standard model's neurons. One participant found feature visualizations sometimes hard to interpret. One recommendation therefore was to additionally show strongly activating natural image examples from a dataset.

6.2. Quantitative Evaluation of User Observations

We performed quantitative measurements to see whether what users observed visually can be generalized beyond the given input scene, synthesized input images, and the Inception-V1 network architecture. To test the generalizability of the users' observations, we performed quantitative measurements using different models than the ones used in the online tool. Specifically, we used the pre-trained ResNet50 from the *torchvision* library of *PyTorch* [PGM*19] as the standard model. For comparison, we used weights of an adversarially trained version of the same model from the *robustness* library [EIS*19] (ResNet50 ImageNet L_2-norm ε 3/255).

First, we investigated the adversarially trained model's sensitivity to background changes, which was reported by two users in the case study. For verification, we performed the *Background Challenge* by Xiao et al. [XEIM20], where a model is tested with (natural image) adversarial backgrounds. The adversarially trained model can only correctly predict 12.3% under background variations. The standard model achieves 22.3% accuracy. Using this test dataset, random guessing would yield 11.1% accuracy [XEIM20]. This shows that the robustified network is considerably more susceptible to adversarial backgrounds than the standard model.

Second, we tested if the adversarially trained model is indeed more sensitive to geometric scene transformations than the standard model (reported by three users). In particular, we looked into camera rotation. We generated a synthetic dataset, where we rendered the four 3D models provided in Perturber from seven yaw angles, ranging from -70° to 70° (Figure 12a), two pitch angles, and two distances of the camera to the object, as shown in Figure 12b. In total, we generated 28 views for each of the four 3D models.



(a)

(b)

Figure 12: *Seven yaw angles variations (a) tested for four prototype views (b) per model (two of the four models are shown here). The prototype view in (a) is highlighted.*

To compare how much the predictions fluctuate, we chose a prototype view with a yaw angle of -23.3° for each of the four pitch / distance combinations and 3D model (Figure 12b). For each of the 16 resulting prototypes, the logits of the top-10 classes served as ground truth vector $\mathbf{l}^{\star}_{10}$. For the 16 prototype views, we then calculated a fluctuation score f_p:

$$f_p = \sum_y \frac{\|\mathbf{l}^{\star}_{10} - \mathbf{l}^{y}_{10}\|_2}{s(\mathbf{l}^{\star}_{n})}, \tag{1}$$

where $\mathbf{l}^{\star}_{10}$ are the top-10 predictions of the prototype view, $\mathbf{l}^{y}_{10}$ is the logit vector of these 10 classes for the view associated with yaw y, and $s(\mathbf{l}^{\star}_{n})$ is the standard deviation of the logits of all n classes in the prototype view. In other words, the fluctuation score measures how strongly the logits of the top-10 prototype classes diverge in the rotated input images.

Figure 13 shows the average fluctuation score of all 28 prototype views. The fluctuation scores are considerably higher for the adversarially trained model for three of the four 3D models. This verifies that the adversarially trained model can be indeed more vulnerable to rotations of the main object.

7. Discussion & Conclusions

We showed that interactive perturbations in combination with live activations can be an effective method to explore potential vulnerabilities of CNNs and to perform qualitative evaluations of more robust model variations. In an exploration scenario, we could replicate the known texture-shape conflict [GRM*19] through multiple perturbation examples. Machine learning experts participating in our case study observed a variety of known but also unexpected network behaviors. They could successfully replicate known CNN properties, such as models' varying sensitivity to adversarial attacks or patch shuffling, using our synthetic input scenes. In addition, our quantitative post-study experiments of selected user observations could replicate their observations using a different network architecture and natural images. Through these experiments, we demonstrated vulnerabilities of adversarially trained models to background modifications and yaw rotations of the main object that, to the best of our knowledge, have not been discussed yet in the machine learning community.

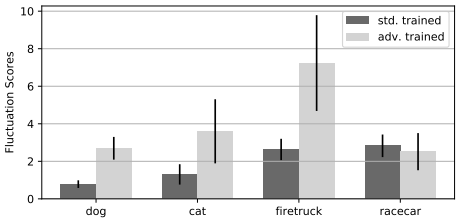


Figure 13: *Average yaw fluctuation scores for the standard model and the adversarially trained model across the four 3D models. Error bars show the standard deviation.*

The majority of insights reported by our users were based on observations how the model predictions changed when perturbing the image. This implies that the principle of live input perturbations could also be useful for pure black-box models. To get a better intuition about which image features are influential for the model's final decision, our exploration scenarios indicate that neuron activations are also essential. But due to the large number of logical neuron units distributed among multiple layers, it can be time-consuming to find a set of neurons that eventually is affected strongly by the performed input perturbation. In the future, we thus plan to investigate alternative methods to select the displayed neurons in the neuron activation view. Also visual guidance to support the discovery of potentially harmful perturbation factors would be helpful. However, traditional guidance mechanisms may require costly computation, which will hamper interactivity (R3). Effective guidance mechanisms can therefore be considered interesting future work. Another interesting line of future work would be the support for encoder-decoder architectures, used prominently for semantic segmentation and image translation tasks among others. This could be facilitated by replacing the prediction view with a continuously updating display of the generated image.

Acknowledgments

We thank Robert Geirhos and Roland Zimmermann for their participation in the case study and valuable feedback, Chris Olah and Nick Cammarata for valuable discussions in the early phase of the project, as well as the Distill Slack workspace as a platform for discussions. M.L. is supported in part by the Austrian Science Fund (FWF) under grant Z211-N23 (Wittgenstein Award). J.B. is supported by the German Federal Ministry of Education and Research (BMBF) through the Competence Center for Machine Learning (TUE.AI, FKZ 01IS18039A) and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). R.H. is partially supported by Boeing and Horizon-2020 ECSEL (grant 783163, iDev40).

References

[ACW18] ATHALYE A., CARLINI N., WAGNER D.: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *arXiv:1802.00420 [cs]* (July 2018). arXiv: 1802.00420. URL: http://arxiv.org/abs/1802.00420. 4

[AR15] AUBRY M., RUSSELL B. C.: Understanding deep features with computer-generated imagery. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 2875–2883. doi:10.1109/ICCV.2015.329. 4

[BHL*21] BABAIEE Z., HASANI R., LECHNER M., RUS D., GROSU R.: On-off center-surround receptive fields for accurate and robust image classification. In *International Conference on Machine Learning* (2021), PMLR, pp. 478–489. 2

[CCG*20] CAMMARATA N., CARTER S., GOH G., OLAH C., PETROV M., SCHUBERT L.: Thread: Circuits. *Distill 5*, 3 (Mar. 2020). URL: https://distill.pub/2020/circuits, doi:10.23915/distill.00024. 3, 4, 6

[CPCS20] CASHMAN D., PERER A., CHANG R., STROBELT H.: Ablate, Variate, and Contemplate: Visual Analytics for Discovering Neural Architectures. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 863–873. doi:10.1109/TVCG.2019.2934261. 2

[DK16] DODGE S., KARAM L.: Understanding how image quality affects deep neural networks. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)* (2016), pp. 1–6. doi:10.1109/QoMEX.2016.7498955. 2, 5

[DPW*20] DAS N., PARK H., WANG Z. J., HOHMAN F., FIRSTMAN R., ROGERS E., CHAU D. H.: Bluff: Interactively Deciphering Adversarial Attacks on Deep Neural Networks. *arXiv:2009.02608 [cs]* (Sept. 2020). arXiv: 2009.02608. URL: http://arxiv.org/abs/2009.02608. 3

[EEF*18] EYKHOLT K., EVTIMOV I., FERNANDES E., LI B., RAHMATI A., XIAO C., PRAKASH A., KOHNO T., SONG D.: Robust Physical-World Attacks on Deep Learning Visual Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2018), pp. 1625–1634. ISSN: 2575-7075. doi:10.1109/CVPR.2018.00175. 2

[EIS*19] ENGSTROM L., ILYAS A., SALMAN H., SANTURKAR S., TSIPRAS D.: Robustness (Python Library), 2019. URL: https://github.com/MadryLab/robustness. 9

[ETT*19] ENGSTROM L., TRAN B., TSIPRAS D., SCHMIDT L., MADRY A.: Exploring the landscape of spatial robustness. In *International Conference on Machine Learning* (2019), PMLR, pp. 1802–1811. URL: https://arxiv.org/abs/1712.02779. 4

[GRM*19] GEIRHOS R., RUBISCH P., MICHAELIS C., BETHGE M., WICHMANN F. A., BRENDEL W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)* (May 2019). URL: https://openreview.net/forum?id=Bygh9j09KX. 2, 4, 6, 7, 10

[GSS14] GOODFELLOW I. J., SHLENS J., SZEGEDY C.: Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014). URL: https://arxiv.org/abs/1412.6572. 2

[GWCV16] GAIDON A., WANG Q., CABON Y., VIG E.: Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4340–4349. doi:10.1109/CVPR.2016.470. 4

[GZL*20] GOU L., ZOU L., LI N., HOFMANN M., SHEKAR A. K., WENDT A., REN L.: Vatld: a visual analytics system to assess, understand and improve traffic light detection. *IEEE transactions on visualization and computer graphics* (2020). doi:10.1109/TVCG.2020.3030350. 2

[Har15] HARLEY A. W.: An Interactive Node-Link Visualization of Convolutional Neural Networks. In *Advances in Visual Computing*, vol. 9474. Springer International Publishing, Cham, 2015, pp. 867–877. doi:10.1007/978-3-319-27857-5_77. 3

[HKPC19] HOHMAN F., KAHNG M., PIENTA R., CHAU D. H.: Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics 25*, 8 (Aug. 2019), 2674–2693. doi:10.1109/TVCG.2018.2843369. 2

[HPRC20] HOHMAN F., PARK H., ROBINSON C., CHAU D. H. P.: Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 1096–1106. doi:10.1109/TVCG.2019.2934659. 3, 6

[KC19] KAHNG M., CHAU D. H.: How does visualization help people learn deep learning? evaluation of GAN Lab. In *Workshop on EValuation of Interactive VisuAl Machine Learning systems* (2019). 3

[KPN16] KRAUSE J., PERER A., NG K.: Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. pp. 5686–5697. doi:10.1145/2858036.2858529. 3

[KSH17] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: ImageNet classification with deep convolutional neural networks. *Communications of the ACM 60*, 6 (May 2017), 84–90. URL: https://dl.acm.org/doi/10.1145/3065386, doi:10.1145/3065386. 1

[KTC*19] KAHNG M., THORAT N., CHAU D. H., VIÉGAS F. B., WATTENBERG M.: GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 310–320. doi:10.1109/TVCG.2018.2864500. 2, 3

[LBM*16] LAPUSCHKIN S., BINDER A., MONTAVON G., MÜLLER K.-R., SAMEK W.: The LRP Toolbox for Artificial Neural Networks. *Journal of Machine Learning Research 17*, 114 (2016), 1–5. 3

[LCJ*19] LIU D., CUI W., JIN K., GUO Y., QU H.: DeepTracker: Visualizing the Training Process of Convolutional Neural Networks. *ACM Transactions on Intelligent Systems and Technology 10*, 1 (Jan. 2019), 1–25. URL: https://dl.acm.org/doi/10.1145/3200489, doi:10.1145/3200489. 2

[LHA*20] LECHNER M., HASANI R., AMINI A., HENZINGER T. A., RUS D., GROSU R.: Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence 2*, 10 (2020), 642–652. doi:10.1038/s42256-020-00237-3. 2

[LHG*21] LECHNER M., HASANI R., GROSU R., RUS D., HENZINGER T. A.: Adversarial training is not ready for robot learning. *arXiv preprint arXiv:2103.08187* (2021). 2

[LLL*18] LIU S., LI Z., LI T., SRIKUMAR V., PASCUCCI V., BREMER P.-T.: Nlize: A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 651–660. doi:10.1109/TVCG.2018.2865230. 3

[LLS*18] LIU M., LIU S., SU H., CAO K., ZHU J.: Analyzing the Noise Robustness of Deep Neural Networks. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Oct. 2018), pp. 60–71. doi:10.1109/VAST.2018.8802509. 3

[MFH*20] MA Y., FAN A., HE J., NELAKURTHI A. R., MACIEJEWSKI R.: A Visual Analytics Framework for Explaining and Diagnosing Transfer Learning Processes. *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. doi:10.1109/TVCG.2020.3028888. 2, 3

[MMS*19] MADRY A., MAKELOV A., SCHMIDT L., TSIPRAS D., VLADU A.: Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:1706.06083 [cs, stat]* (Sept. 2019). arXiv: 1706.06083. URL: http://arxiv.org/abs/1706.06083. 2, 5, 6

[NQ17] NORTON A. P., QI Y.: Adversarial-Playground: A visualization suite showing how adversarial examples fool deep learning. In *2017 IEEE Symposium on Visualization for Cyber Security (VizSec)* (2017), IEEE, pp. 1–4. doi:10.1109/VIZSEC.2017.8062202. 3

[OMS17] OLAH C., MORDVINTSEV A., SCHUBERT L.: Feature Visualization. *Distill 2*, 11 (Nov. 2017), e7. URL: https://distill.pub/2017/feature-visualization, doi:10.23915/distill.00007. 3, 5, 6

[Ope] OpenAI Microscope. https://microscope.openai.com/models. (Accessed on 10/12/2020). 3

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J., CHINTALA S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. 9

[PHVG*18] PEZZOTTI N., HOLLT T., VAN GEMERT J., LELIEVELDT B. P., EISEMANN E., VILANOVA A.: DeepEyes: Progressive Visual Analytics for Designing Deep Neural Networks. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 98–108. doi:10.1109/TVCG.2017.2744358. 2

[PW17] PEREZ L., WANG J.: The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017). URL: https://arxiv.org/abs/1712.04621. 2, 4

[RDS*15] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M., ET AL.: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*, 3 (2015), 211–252. doi:10.1007/s11263-015-0816-y. 3

[RHGS17] REN S., HE K., GIRSHICK R., SUN J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 6 (June 2017), 1137–1149. doi:10.1109/TPAMI.2016.2577031. 1

[RSG16] RIBEIRO M., SINGH S., GUESTRIN C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (June 2016), Association for Computational Linguistics, pp. 97–101. doi:10.18653/v1/N16-3020. 5

[SCD*17] SELVARAJU R. R., COGSWELL M., DAS A., VEDANTAM R., PARIKH D., BATRA D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct. 2017), pp. 618–626. ISSN: 2380-7504. doi:10.1109/ICCV.2017.74. 3

[SCS*17] SMILKOV D., CARTER S., SCULLEY D., VIÉGAS F. B., WATTENBERG M.: Direct-Manipulation Visualization of Deep Networks. *arXiv:1708.03788 [cs, stat]* (Aug. 2017). arXiv: 1708.03788. URL: http://arxiv.org/abs/1708.03788. 2

[SLJ*15] SZEGEDY C., LIU W., JIA Y., SERMANET P., REED S., ANGUELOV D., ERHAN D., VANHOUCKE V., RABINOVICH A.: Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1–9. doi:10.1109/CVPR.2015.7298594. 3

[SLL20] STURMFELS P., LUNDBERG S., LEE S.-I.: Visualizing the impact of feature attribution baselines. *Distill 5*, 1 (2020), e22. URL: https://distill.pub/2020/attribution-baselines/. 5

[SMM12] SEDLMAIR M., MEYER M., MUNZNER T.: Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics 18*, 12 (2012), 2431–2440. doi:10.1109/TVCG.2012.213. 8

[SQLG15] SU H., QI C. R., LI Y., GUIBAS L. J.: Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 2686–2694. doi:10.1109/ICCV.2015.308. 4

[SSSEA20] SPINNER T., SCHLEGEL U., SCHÄFER H., EL-ASSADY M.: explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 1064–1074. doi:10.1109/TVCG.2019.2934629. 2

[STA*19] SMILKOV D., THORAT N., ASSOGBA Y., YUAN A., KREEGER N., YU P., ZHANG K., CAI S., NIELSEN E., SOERGEL D., BILESCHI S., TERRY M., NICHOLSON C., GUPTA S. N., SIRAJUDDIN S., SCULLEY D., MONGA R., CORRADO G., VIÉGAS F. B., WATTENBERG M.: TensorFlow.js: Machine Learning for the Web and Beyond. *arXiv:1901.05350 [cs]* (Feb. 2019). arXiv: 1901.05350. URL: http://arxiv.org/abs/1901.05350. 6

[STY17] SUNDARARAJAN M., TALY A., YAN Q.: Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]* (June 2017). arXiv: 1703.01365. URL: http://arxiv.org/abs/1703.01365. 3

[SVZ13] SIMONYAN K., VEDALDI A., ZISSERMAN A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013). URL: https://arxiv.org/abs/1312.6034. 3

[SZS*14] SZEGEDY C., ZAREMBA W., SUTSKEVER I., BRUNA J., ERHAN D., GOODFELLOW I., FERGUS R.: Intriguing properties of neural networks. *arXiv:1312.6199 [cs]* (Feb. 2014). arXiv: 1312.6199. URL: http://arxiv.org/abs/1312.6199. 2

[TSE*19] TSIPRAS D., SANTURKAR S., ENGSTROM L., TURNER A., MADRY A.: Robustness May Be at Odds with Accuracy. *arXiv:1805.12152 [cs, stat]* (Sept. 2019). arXiv: 1805.12152. URL: http://arxiv.org/abs/1805.12152. 6

[WPB*20] WEXLER J., PUSHKARNA M., BOLUKBASI T., WATTENBERG M., VIÉGAS F., WILSON J.: The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (Jan. 2020), 56–65. doi:10.1109/TVCG.2019.2934619. 3

[WSW*18] WONGSUPHASAWAT K., SMILKOV D., WEXLER J., WILSON J., MANÉ D., FRITZ D., KRISHNAN D., VIÉGAS F. B., WATTENBERG M.: Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 1–12. doi:10.1109/TVCG.2017.2744878. 2

[WTS*20] WANG Z. J., TURKO R., SHAIKH O., PARK H., DAS N., HOHMAN F., KAHNG M., CHAU D. H.: CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* (2020). doi:10.1109/TVCG.2020.3030418. 3

[XEIM20] XIAO K., ENGSTROM L., ILYAS A., MADRY A.: Noise or signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994* (2020). URL: https://arxiv.org/abs/2006.09994. 4, 9

[YCN*15] YOSINSKI J., CLUNE J., NGUYEN A., FUCHS T., LIPSON H.: Understanding neural networks through deep visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)* (2015). URL: https://arxiv.org/abs/1506.06579. 3

[YLS*19] YIN D., LOPES R. G., SHLENS J., CUBUK E. D., GILMER J.: A Fourier Perspective on Model Robustness in Computer Vision. *arXiv:1906.08988 [cs, stat]* (Oct. 2019). arXiv: 1906.08988. URL: http://arxiv.org/abs/1906.08988. 2, 5

[ZHP*17] ZENG H., HALEEM H., PLANTAZ X., CAO N., QU H.: Cncomparator: Comparative analytics of convolutional neural networks. *arXiv preprint arXiv:1710.05285* (2017). URL: https://arxiv.org/abs/1710.05285. 2

[ZSLG16] ZHENG S., SONG Y., LEUNG T., GOODFELLOW I.: Improving the robustness of deep neural networks via stability training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 4480–4488. doi:10.1109/CVPR.2016.485. 2, 4

[ZZ19] ZHANG T., ZHU Z.: Interpreting Adversarially Trained Convolutional Neural Networks. *arXiv:1905.09797 [cs, stat]* (May 2019). arXiv: 1905.09797. URL: http://arxiv.org/abs/1905.09797. 2, 5, 7

Interactive Analysis of CNN Robustness
— Supplemental Document —

Stefan Sietzen*, Mathias Lechner, Judy Borowski,
Ramin Hasani, and Manuela Waldner

A Extended and Additional Exploration Scenarios

In addition to the exploration scenario shown in the main paper, we show here results of the following exploratory analyses:

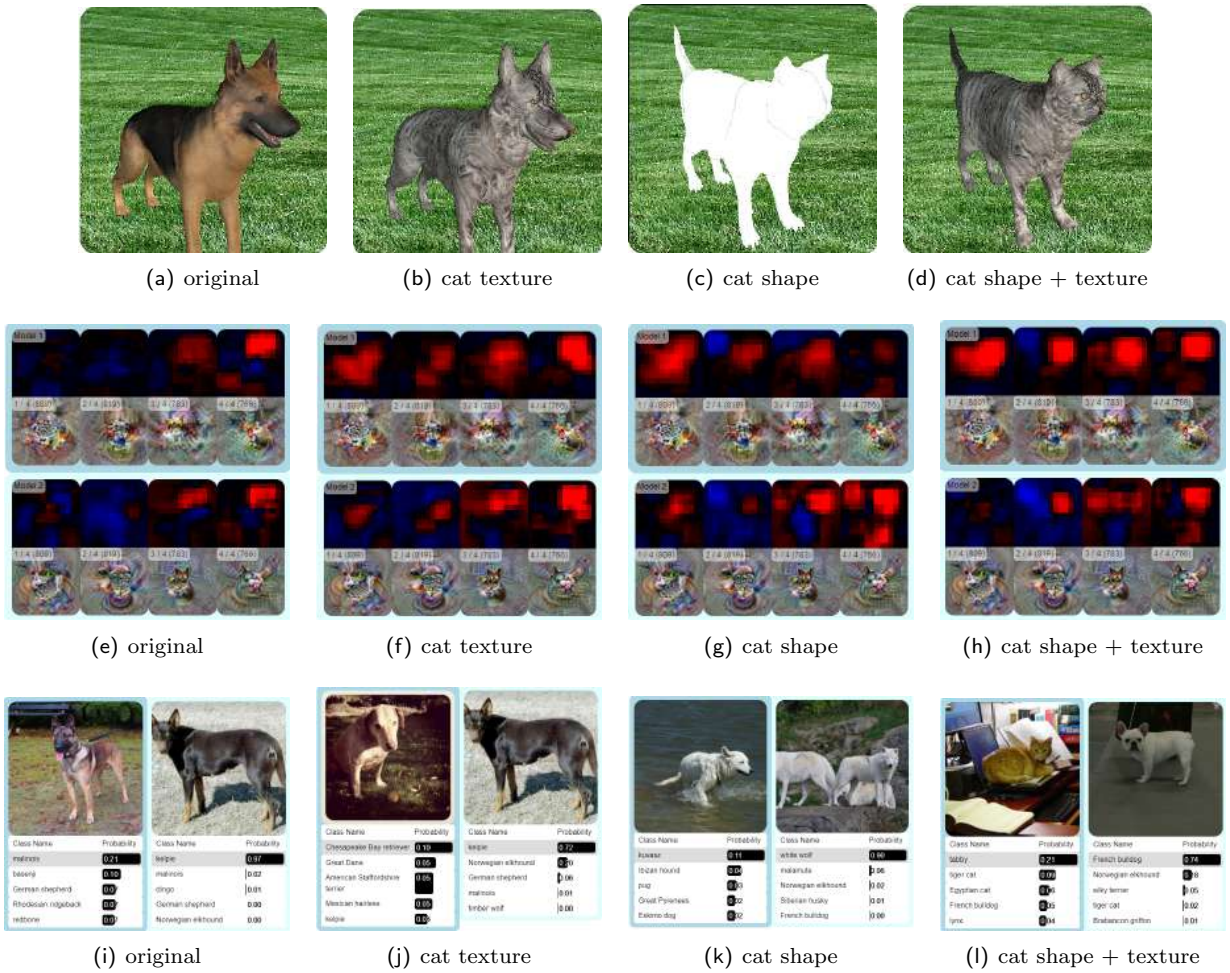
- texture influence (Section A.1 and Section A.2),
- shape sensitivity (Section A.3),
- low frequency information (Section A.4),
- high frequency information (Section A.5),
- adversarial attacks (Section A.6),
- fading to black (Section A.7),
- geometric transformations (Section A.8),
- as well as geometric transformations in combination with background modifications (Section A.9).
- development of activations & feature visualizations during fine-tuning (Section A.10).

For the respective scenarios, we compare the standard model (Inception-V1 trained on ImageNet) with the Stylized-ImageNet trained model (Inception-V1 fine-tuned with Stylized-ImageNet [1]) in Sections A.1, A.2, A.3, A.4, A.5 and with the adversarially trained model (Inception-V1 adversarially fine-tuned [2, 5]) in Section A.6, A.7, A.8, A.9. Finally, we show the development of feature visualizations during fine-tuning for both models in Section A.10.

*stefan.sietzen@gmx.at

A.1 Texture vs. Shape

Figure A.1 shows a comparison between the standard trained model and the model trained by Stylized-ImageNet. The middle row shows how cat-related neurons get activated by morphing the texture and shape, respectively. Please note how the standard model (top row in Figure A.1 e-h) gets strongly activated by the cat texture, while the respective neurons of the Stylized-ImageNet trained model (bottom row in Figure A.1 e-h) seem to get more activated by shape changes. Also note that the Stylized-ImageNet trained model never predicts a cat.



(a) original (b) cat texture (c) cat shape (d) cat shape + texture

(e) original (f) cat texture (g) cat shape (h) cat shape + texture

(i) original (j) cat texture (k) cat shape (l) cat shape + texture

Figure A.1: Object morphing to assess the texture vs. shape conflict: input scene (a-d), activations of four cat-related neurons in mixed4e (e-h) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (i-l) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.2 Patch Shuffling

Figure A.2 shows the effect of randomly shuffling image patches on the standard and the Stylized-ImageNet trained model. The dog is still predicted correctly by both models when shuffling 2×2 image patches (Figure A.2 g). The Stylized-ImageNet trained model is more sensitive to patch shuffling than the standard model (Figure A.2 h-i). Note how the activations of neuron 429 in layer mixed4b (third column in Figure A.2 d-f) follow certain regions in the dog face.

(a) $k = 2$ (b) $k = 5$ (c) $k = 8$

(d) $k = 2$ (e) $k = 5$ (f) $k = 8$

(g) $k = 2$ (h) $k = 5$ (i) $k = 8$

Figure A.2: Patch shuffling: increasing the number of randomly shuffled image patches k (a-c), activations of dog-relevant neurons in mixed4b (d-f) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (g-i) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.3 Silhouette

Figure A.3 investigates the models' shape sensitivity by analyzing the dog's silhouette against a red background in different poses. The activations of dog-related neurons in mixed4d (Figure A.3 e-h) show that the standard model seems to be more sensitive to pose changes. Indeed, the predictions (Figure A.3 i-l) of the standard model fluctuate with the pose modifications, while the Stylized-ImageNet trained model consistently predicts "white wolf" with a very high probability.

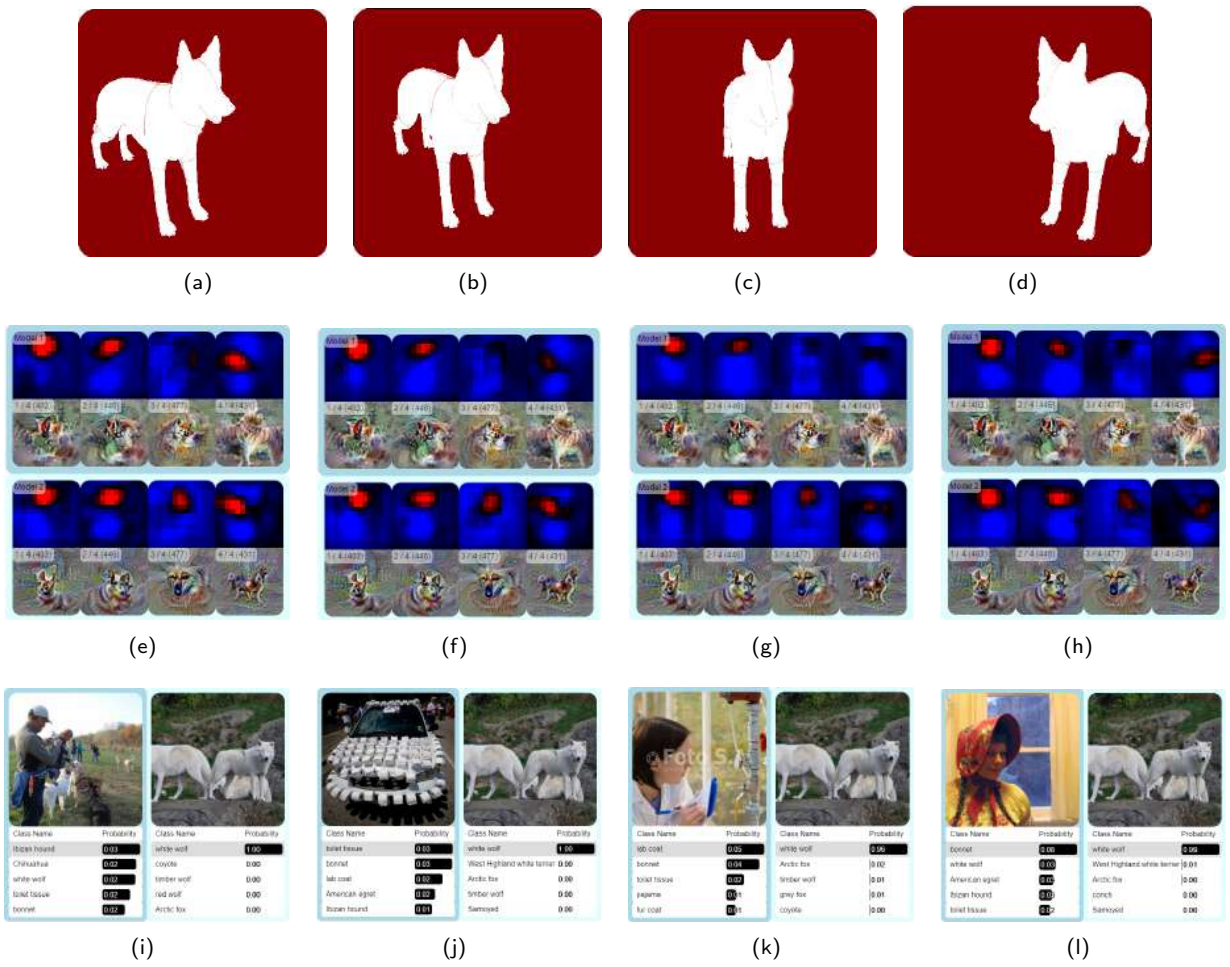


Figure A.3: Analyzing shape influence: different silhouette poses as input (a-d), activations of four dog-related neurons in mixed4d (e-h) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (i-l) by the standard model (left) and the Stylized-ImageNet trained model (right).

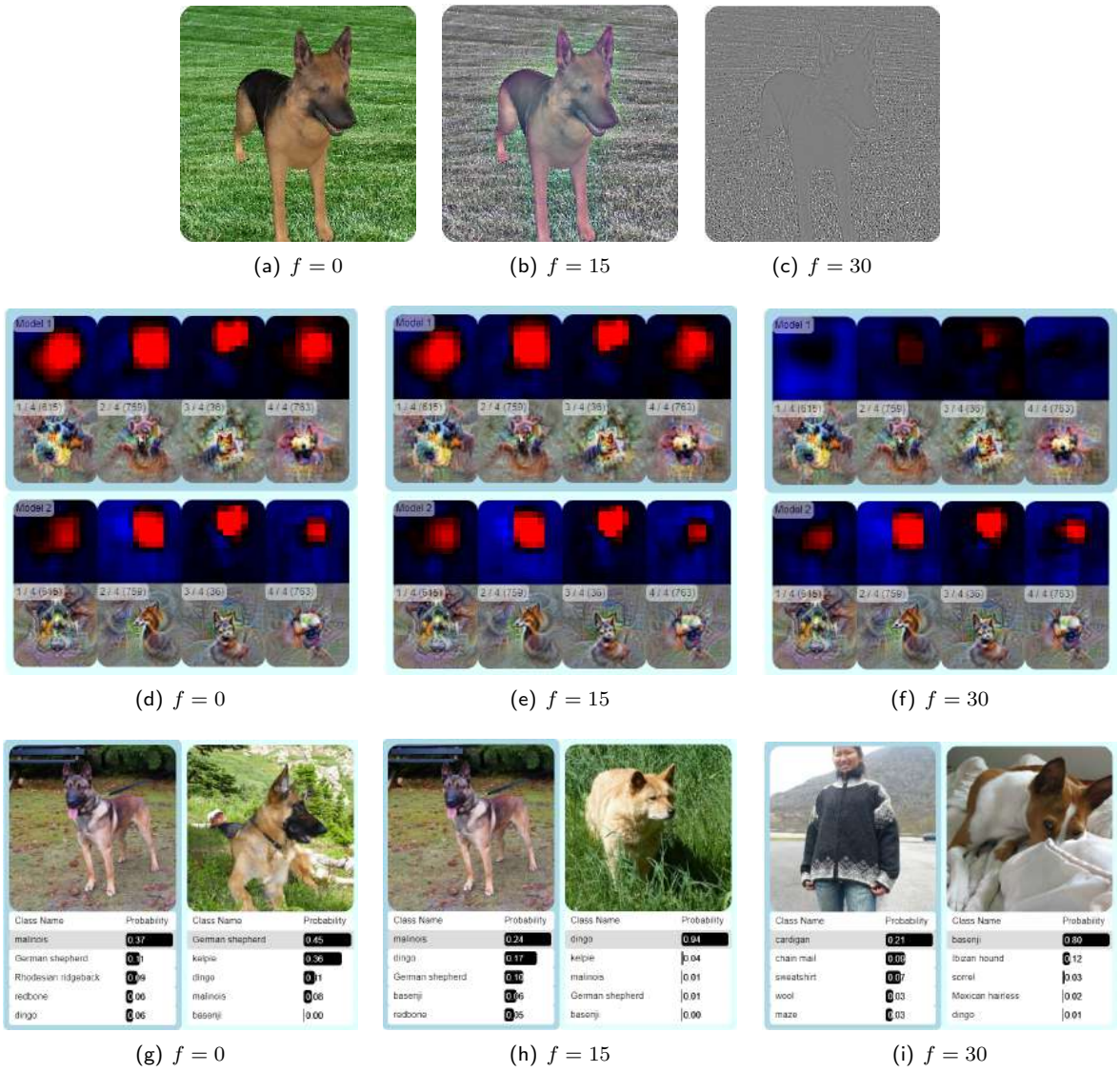
A.4 Blur

In the case study, a suspicion of one user was that the model trained on Stylized-ImageNet would be more sensitive to high-frequency information. Blurring the image would therefore disturb this model more heavily than the standard model. Figure A.4 illustrates that this is not necessarily the case: Activations of dog-related neurons gradually degrade by applying blur for both models (Figure A.4 d-f). At a high blur level, both models have very uncertain predictions (Figure A.4 i).

Figure A.4: Blurring the image: input image with gradual blur (a-c), activations of oriented dog heads in mixed4a (d-f) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (g-i) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.5 High-Pass Filtering

In contrast to low-pass filtering shown in the previous section, here we show the effects of high-pass filtering on the standard trained model and the Stylized-ImageNet trained model (Figure A.5). As the frequency threshold is increased to a high level, the activations of dog-related neurons considerably decrease for the standard model (Figure A.5 f, top row), while the same neurons are still highyly activated for the Stylized-ImageNet trained model (Figure A.5 f, bottom row), and it also still predicts a canine (Figure A.5 i, right column).



(a) $f = 0$ (b) $f = 15$ (c) $f = 30$

(d) $f = 0$ (e) $f = 15$ (f) $f = 30$

(g) $f = 0$ (h) $f = 15$ (i) $f = 30$

Figure A.5: Applying a high-pass filter on the input image: input image with increasing frequency threshold (a-c), activations of oriented dog-related neurons in mixed4e (d-f) by the standard model (top) and the Stylized-ImageNet trained model (bottom), and the predictions (g-i) by the standard model (left) and the Stylized-ImageNet trained model (right).

A.6 Adversarial Attack

Figure A.6 shows an adversarial attack (target class "Egyptian cat") for a standard and an adversarially trained model. Figure A.6 c-d, bottom row shows how the activations of the adversarially trained model remain unaffected, while the cat-related neurons get strongly activated by the attack for the standard model (Figure A.6 c-d, top row). Consequently, the standard model's prediction switches to the attack's target class (Figure A.6 f, left column), while the adversarially trained model still predicts a German shepherd with very high confidence.

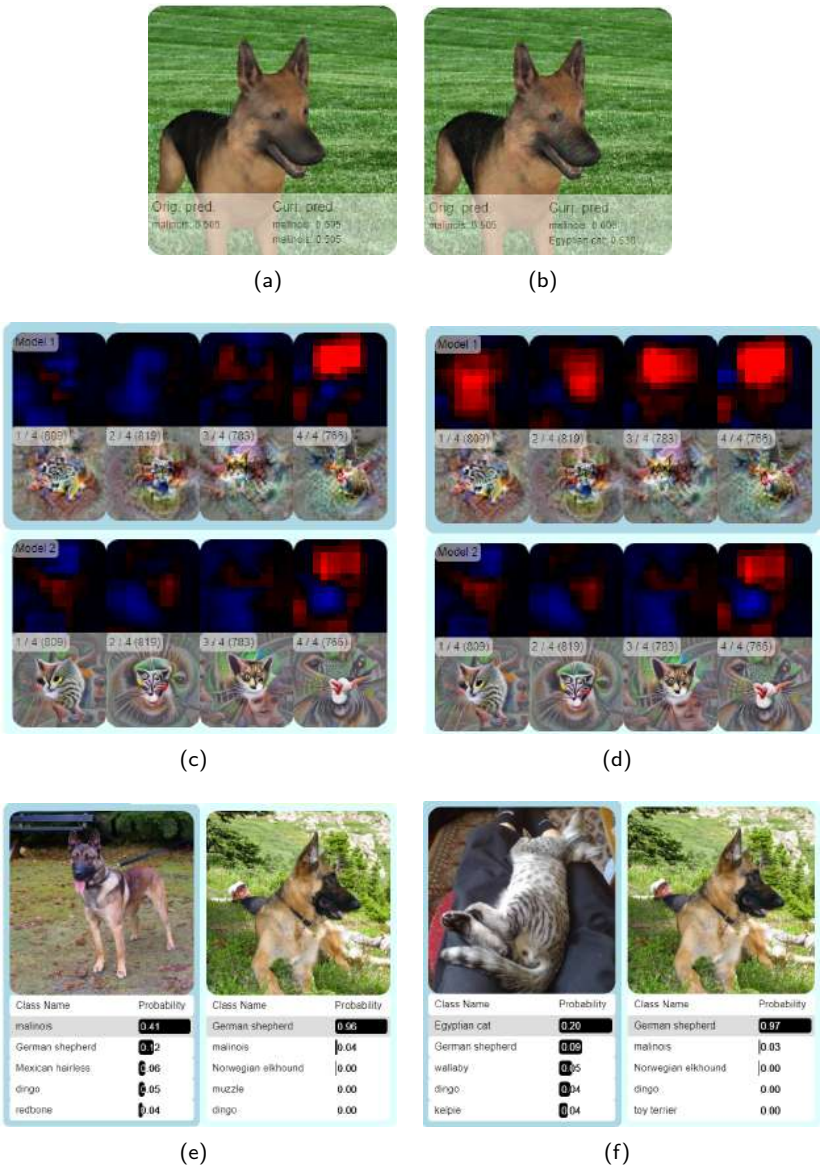
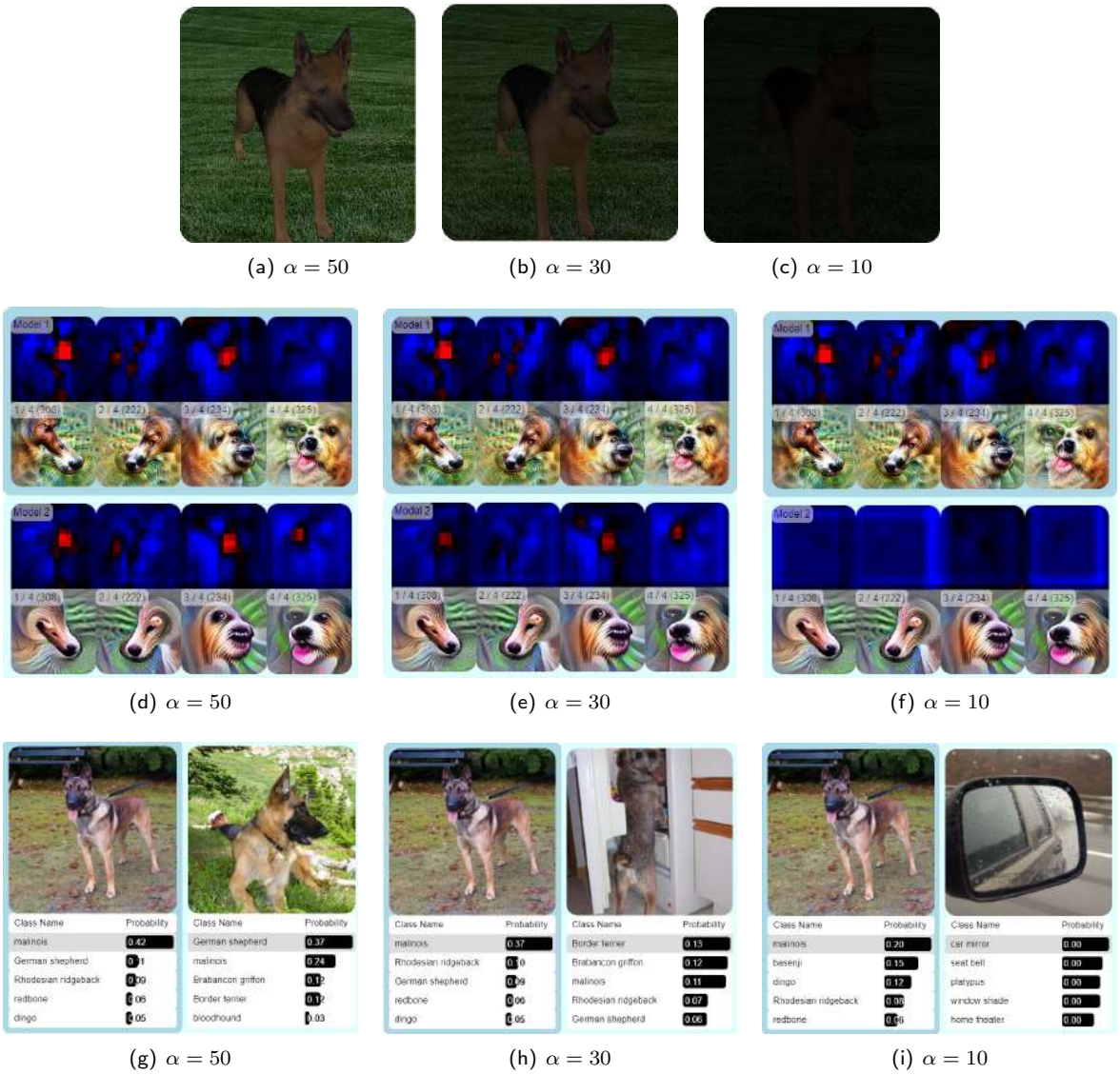


Figure A.6: Adversarial attack with target class "Egyptian cat": input image before (a) and after (b) a successful attack, activations of cat-related neurons (c,d) by the standard model (top) and the adversarially trained model (bottom), and the predictions (e,f) by the standard model (left) and the adversarially trained model (right).

A.7 Alpha

Figure A.7 illustrates the *saturation* effect [3] by gradually blending the input image to black. Note how the activations of the standard model (Figure A.7 d-f, top row) remain high and the predictions (Figure A.7 g-i, left column) remain correct even though the α is already very low on the last image so that humans can no longer perceive any content. The adversarially trained model, however, is more sensitive to the reduced image contrast.



(a) $\alpha = 50$ (b) $\alpha = 30$ (c) $\alpha = 10$

(d) $\alpha = 50$ (e) $\alpha = 30$ (f) $\alpha = 10$

(g) $\alpha = 50$ (h) $\alpha = 30$ (i) $\alpha = 10$

Figure A.7: Reducing the overall alpha and blending into black: input image with different alpha values (a-c), activations of oriented dog heads in mixed4a (d-f) by the standard model (top) and the adversarially trained model (bottom), and the predictions (g-i) by the standard model (left) and the adversarially trained model (right).

A.8 Rotation

Figure A.8 compares two models' sensitivities to viewpoint changes. From a side view (Figure A.8 a-b), both models fairly reliably predict a race car (Figure A.8 i-j). However, when looking at the car from a front-top view (Figure A.8 c), predictions are getting unstable for both models (Figure A.8 k) – in particular for the adversarially trained model, which does not predict any car-like object as top-5 target (Figure A.8 k, right). Looking at the activations of neurons that are important for the prediction of race cars in mixed4b (Figure A.8 e-h), it seems that wheels play a very important role. As the car is rotated and wheels disappear, the activations of these neurons decrease considerably (Figure A.8 g-h).

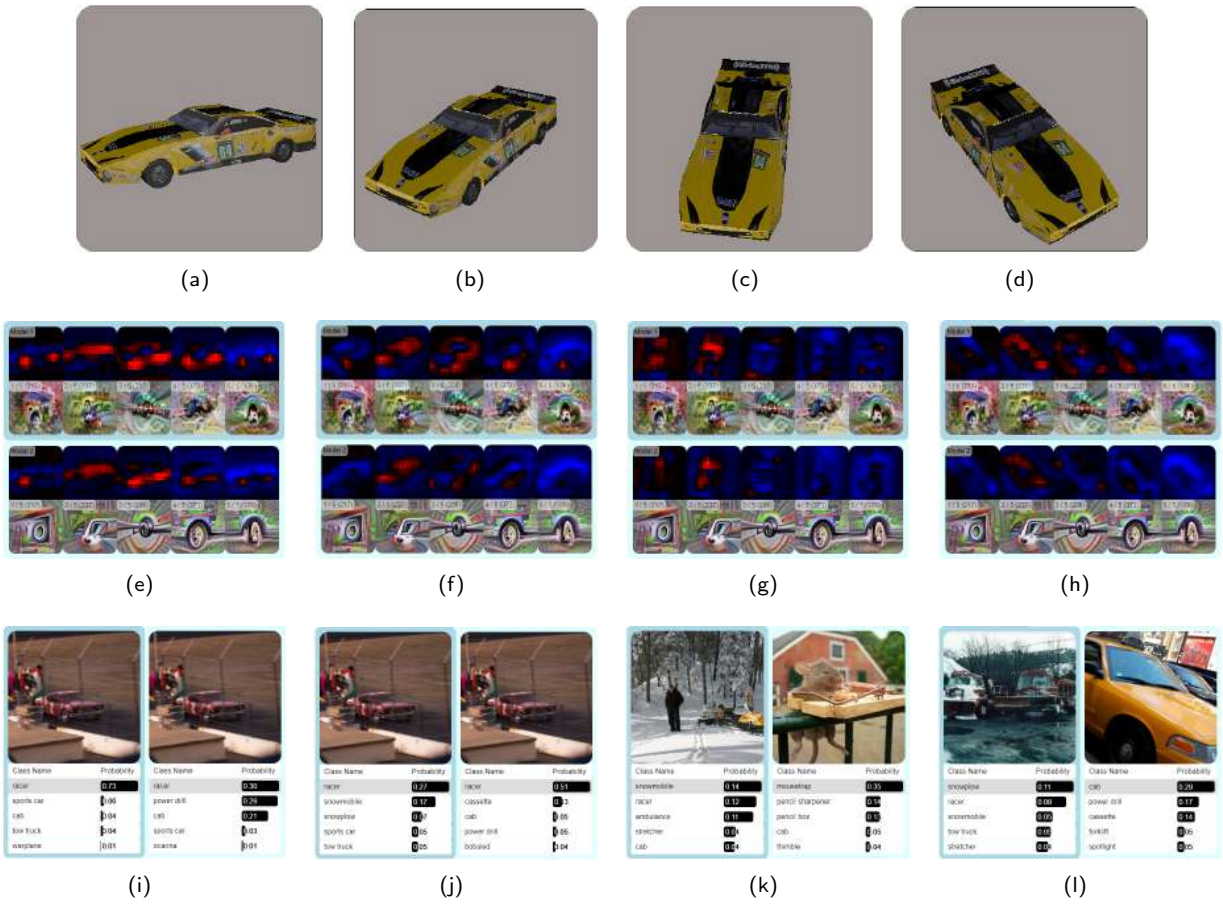
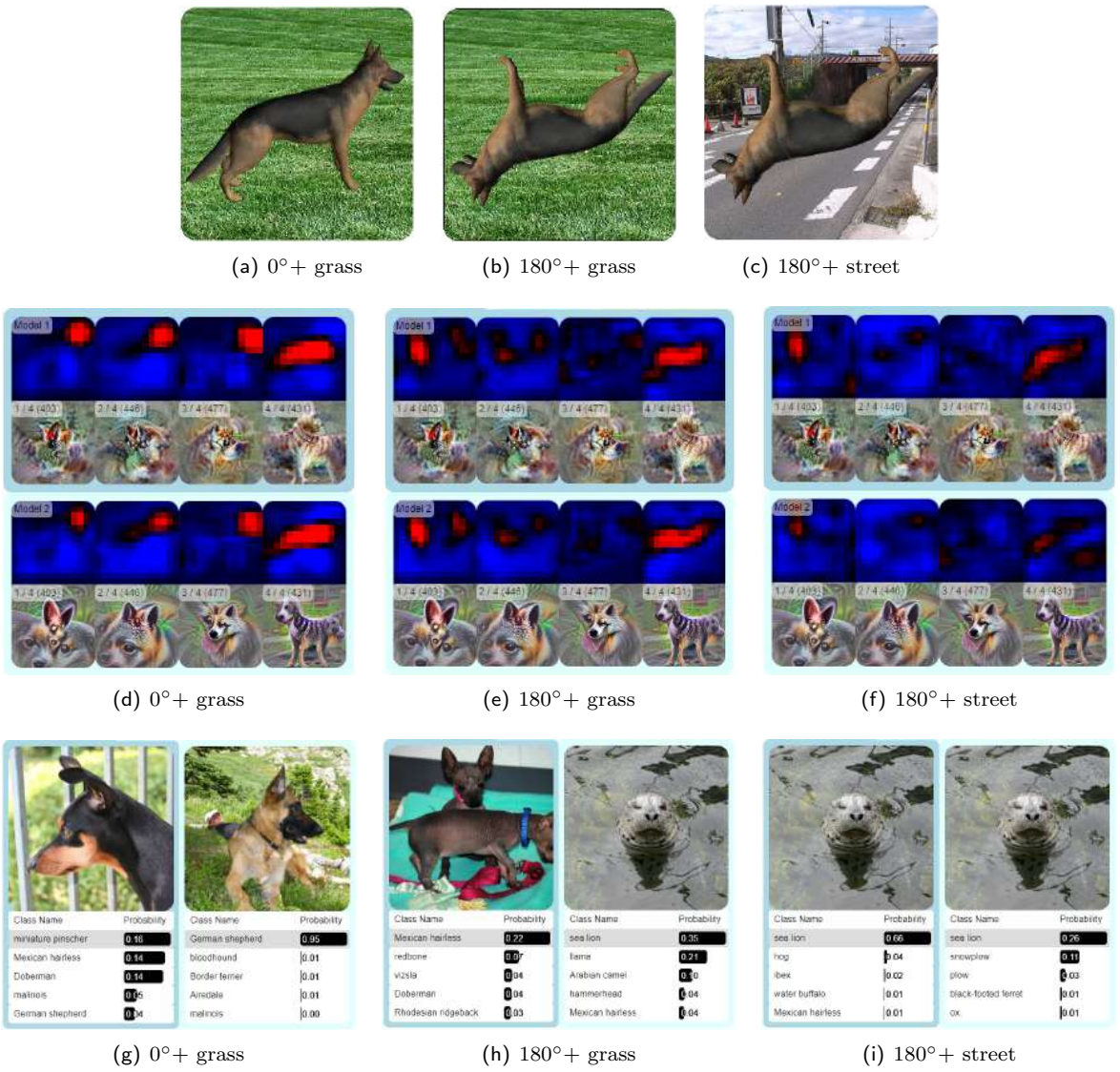


Figure A.8: Orbiting around the race car: input scene from different camera angles (a-d), activations of four race car-related neurons in mixed4b (e-h) by the standard model (top) and the adversarially trained model (bottom), and the predictions (i-l) by the standard model (left) and the adversarially trained model (right).

A.9 Roll + Background

To assess the sensitivity to the context, we analyze the influence of the main object's rotation *and* the background image in Figure A.9. While the standard model still predicts dog breeds after a 180°rotation (Figure A.9 h, left column), the adversarially trained model has a tendency to predict sea animals (Figure A.9 h, right column). After changing the background, none of the models predicts a dog (Figure A.9 i). The adversarially trained model also has artifacts in the top-5 in case of a street background (Figure A.9 i, right column). Also, note how the dog-related activations decrease once the object is rotated (Figure A.9 e) and the background is swapped (Figure A.9 f) for both models.



(a) 0°+ grass (b) 180°+ grass (c) 180°+ street

(d) 0°+ grass (e) 180°+ grass (f) 180°+ street

(g) 0°+ grass (h) 180°+ grass (i) 180°+ street

Figure A.9: Rotating the main object and changing the background: input image with different rotations and / or background images (a-c), activations of dog-related neurons in mixed4d (d-f) by the standard model (top) and the adversarially trained model (bottom), and the predictions (g-i) by the standard model (left) and the adversarially trained model (right).

A.10 Model Fine-Tuning

To better understand what happens during fine-tuning, users can compare models at various intermediate checkpoints of the fine-tuning process. This is similar to the transfer learning visualizations by Szabo et al. [4]. However, they investigated transfer learning with different datasets, while we fine-tuned the models with variants of the same dataset.

Perturber provides 17 selected checkpoints during adversarial fine-tuning and seven selected checkpoints during Stylized-ImageNet fine-tuning. We chose to include more checkpoints for the adversarial fine-tuning because it appears more dynamic compared to the Stylized-ImageNet fine-tuning (as can be seen in Figure A.10), and generating the required data is computationally expensive.

Figure A.10 shows activations and feature visualizations of neuron 222 in layer mixed4a at various fine-tuning checkpoints. During Stylized-ImageNet fine-tuning, the activations and feature visualizations stay relatively consistent. During intermediate steps of adversarial fine-tuning, however, the positive response vanishes before reappearing after iteration 10K. The corresponding feature visualizations also reflect this phenomenon by becoming less similar to a dog head intermediately before assuming the appearance of a smoother version of a dog head than before fine-tuning.

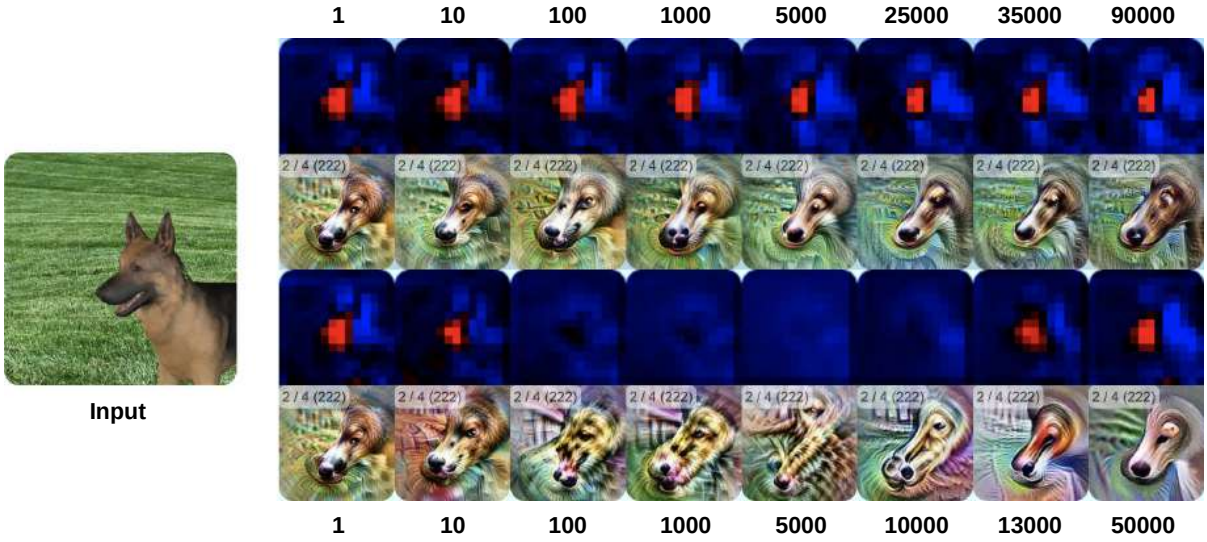


Figure A.10: Activations and feature visualizations of neuron 222 in layer mixed4a at selected fine-tuning checkpoints. Numbers above and below show the fine-tuning iteration.

B Case Study Observations & Feedback

Here, we list all observations and feedback recorded during the case studies. Table B.1 shows the research focus of the individual study participants.

Table B.2 lists all reported observations. Some of these observations are visual confirmations of known facts, some observations are highly speculative, some are just descriptions of what the users saw.

In Table B.3, finally, we list all suggestions for future improvements mentioned by the users.

Table B.1: Research focus of the case study participants.

User	Research Focus
P1	Understanding vision in humans and machines, with a special focus on Deep Learning interpretability and feature visualizations.
P2	On the interface between psychophysics and deep learning, in particular understanding how object recognition differs between humans and machines.
P3	Detection and interpretation of failure cases of computer vision models.
P4	Learning more robust, safe, and verifiable machine learning models.
P5	Designing interpretable deep learning models.

Table B.2: Observations reported by the participants in our case studies, including references to corresponding exemplary scenarios.

User	Observation	Reference
Geometric Perturbations		
P3	The adversarially trained model is robust to translation when the dog is viewed from the side, while the standard model fluctuates.	
P3	Zooming into the race car makes the Stylized-ImageNet trained model predict a school bus, which is incorrect.	
P4	Rotation affects the class output of the adversarially trained model more than that of the standard model.	Section A.8
P5	The adversarially trained model tends to misclassify the scene more often upon viewpoint changes than the standard model.	Section A.8
P5	The adversarially trained model seems to be less sensitive to object distance (zoom).	
Scene Perturbations		
P3	Background blur makes the adversarially trained model less consistent. The adversarially trained model seems to use the background more than the standard model.	Section 6.2 (main paper)
P5	Background significantly alters the decisions made by the adversarially trained model. This is less apparent for the standard model.	Section 6.2 (main paper)
Object Morphing		
P1	The cat is predicted surprisingly "late".	Section A.1
P2	Predictions first change to another dog class before they switch to a cat.	Section A.1
P3	Activations for dog-related neurons do not necessarily peak at "pure" dog images.	
Frequency Decomposition		
P1	The strong influence of frequency decomposition operations on the class predictions is surprising.	
P2	The Stylized-ImageNet trained model is more robust under blur than expected.	Section A.4
Patch Shuffling		
P2	The target class is soon difficult to predict for a human, but it is still correctly predicted by the model	Section A.2
Adversarial Attacks		
P2	Adversarial attacks only affect the standard model.	Section A.6
P4	Adversarial attacks change the activations of the early layers very little. The activations seem to change more on the later layers.	
P4	An adversarial attack on a car scene towards "badger" leads to fur neurons getting activated.	
Complex Perturbations		
P2	A small rotation in combination with an unusual background is sufficient to disturb the adversarially trained model.	Section A.9
P3	The untextured dog head can be quite certainly predicted by the adversarially trained model, but leads to a hammerhead prediction upon close-up for the standard model. Texture makes predictions more certain, but blurred texture (coloring) also helps.	
P3	Rotation has a strong influence in combination with low texture influence and zooming.	
Feature Visualizations & Activation Maps		
P1, P3	Feature visualizations of the adversarially trained model look more "intuitive" / "cartoonish".	
P2	Eye detectors react to surprisingly many regions in the street background image.	
P4	The lack of differences of activation maps between models is suprising. The only major differences were observable during adversarial attacks.	

Table B.3: Suggestions for improvement provided by the participants of the study.

User	Suggestion for Improvement
Scene Perturbations	
P1	Allow users to upload custom background images.
P2	Support background rotation.
P3	Object texture could be more detailed.
Adversarial Attacks	
P4	Change the scene behind the adversarial attack instead of adversarial attack being tied to a fixed image.
Feature Visualizations	
P1	Show dataset examples (i.e., strongly activating examples from the training data) instead of / in addition to feature visualizations.
P3	Support different feature visualizations.
General Suggestions	
P1	Provide more guidance through the interface, otherwise it can be overwhelming.
P1, P3	Provide more 3D models.
P3	Show dataset examples with similar activations as the current input scene.
P5	Perform a grid search to systematically generate input images and store the results for quantitative evaluation.

References

[1] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *International Conference on Learning Representations (ICLR)* (2019).

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *arXiv:1706.06083 [cs, stat]* (Sept. 2019). arXiv: 1706.06083.

[3] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. "Visualizing the impact of feature attribution baselines". In: *Distill* 5.1 (2020), e22.

[4] Róbert Szabó, Dániel Katona, Márton Csillag, Adrián Csiszárik, and Dániel Varga. "Visualizing Transfer Learning". In: *arXiv:2007.07628 [cs]* (July 2020). arXiv: 2007.07628.

[5] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness May Be at Odds with Accuracy". en. In: *arXiv:1805.12152 [cs, stat]* (Sept. 2019). arXiv: 1805.12152.