Natural images are more informative for interpreting CNN activations than synthetic feature visualizations

Anonymous Author(s)
Affiliation
Address
email

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 in predicting CNN activations. Using a well-controlled psychophysical paradigm, we compare the informativeness of synthetic images by Olah et al. [39] with a simple baseline visualization, namely 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 synthetic images by a wide margin ($92\% \pm 2\%$ accuracy). The superiority of natural images holds across the investigated network and various conditions. Therefore, we argue that future 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 and the interest in explainable Artificial Intelligence (XAI) is growing. A "right to explanation" is even required by law in the European Union [14, 16]. We here focus on the popular post-hoc explanation method (or interpretability method) of *feature visualizations via activation maximization*, also known as *input maximization* or *maximally exciting images*. First introduced by Erhan et al. [12] and subsequently improved by many others [29, 34, 32, 35, 37], 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 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 [17] and that "features can be rigorously studied and understood" [42]. Over the past few years, extensive investigations to better understand CNNs are based on feature visualizations [42, 41, 8, 7], and the technique is being combined with other explanation methods [40, 9, 3, 20].

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

Submitted to 2nd Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM).

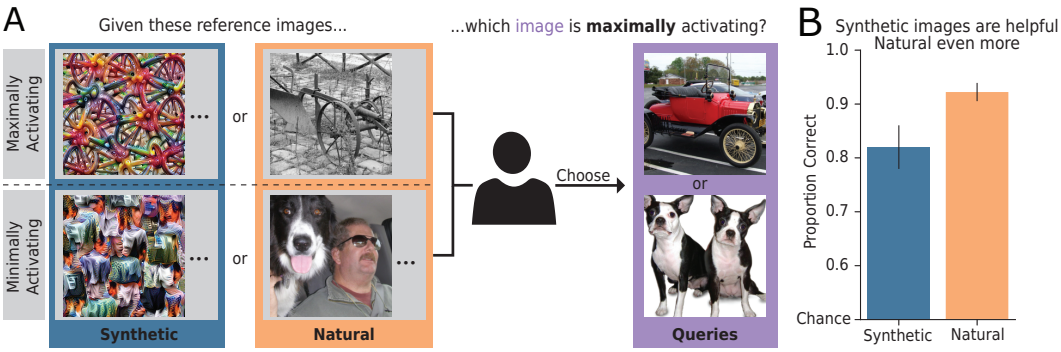


Figure 1: How useful are synthetic compared to natural images for interpreting neural network activations? **A: Human experiment.** Given extremely activating reference images (either synthetic or natural), a human participant chooses which out of two query images is also a strongly activating image. Synthetic images are generated via feature visualization [39]. **B: Core result.** Participants are well above chance for synthetic images—but even better when seeing *natural* reference images.

are the well-interpretable, hand-picked [39] synthetic images in publications for the entirety of all units in a network, a concern raised by e.g. Kriegeskorte [25]? And what if the features that a CNN is truly sensitive to are imperceptible instead, as might be suggested by the existence of adversarial examples [47, 22]? Morcos et al. [31] even suggest that units of 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.

One way to advance this debate is to objectively 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. [39]. Specifically, participants choose which of two natural images would elicit a higher activation 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 holds across the network and various conditions.

2 Related Work

Thanks to explanation methods, significant progress has been made in recent years towards understanding CNNs for image data. Nevertheless, challenges remain and concern for example over-engineering. As such, the loss function and techniques to make the synthetic images look more interpretable are often discussed for feature visualizations [37].

In order to further advance XAI, scientists advocate different research directions. Besides the focus on developing additional methods, some researchers (e.g. Olah et al. [42]) 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 [21, 33, 13, 28, 50, 49]. 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 [6]. Additionally, Miller [30] 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, 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 Supplementary Material. A commonly employed task paradigm is the "forward simulation / prediction" task, first introduced by Doshi-Velez and Kim [11]: 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, comparisons are performed within studies, often against baselines. According to the current literature, studies reporting positive effects of explanations [26] slightly outweigh those reporting inconclusive [4] or even negative effects [46].

To our knowledge, no study has yet evaluated the popular explanation method of feature visualizations and how it improves our understanding of intermediate network activations. This study therefore closes an important gap.

3 Methods

We perform two human psychophysical studies 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 be a strongly activating image given a set of reference images (cf. Fig. 2). Apart from image choices, we record participants' confidence levels and their reaction times. In order to gain insights on how intuitive participants find feature visualizations, their subjective judgments are collected in a separate task and a dynamic conversation (for details, see Appendix Sec. A.1.1 and Sec. A.2.6).

Figure 2: Example trial in psychophysical experiments. A participant sees minimally and maximally activating reference images for a certain feature map on the sides and has 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. After each trial, the participant receives feedback which image was the maximally activating one.

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. To this end, 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 with 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 [39, 40, 42, 41], we use an Inception V1 network [48] trained on ImageNet [10]. The synthetic images are the optimization results of the feature visualization method by Olah et al. [39]. The natural stimuli are selected from the validation set of the ImageNet ILSVRC 2012 [44] dataset according to their activations for the feature maps of interest.

4 Results

In this section, Fig. 3 (4) shows data from Experiment I (II). Additional figures as well as additional results (hand- vs. randomly picked, subjective impressions), can be found in the Appendix Sec. A.2. Error bars always 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 [39], 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%) (cf. Fig. 3A). However, performance is even higher in what we intended to be the baseline condition: natural reference images ($92 \pm 2\%$). Additionally, for correctly answered trials, participants much more frequently report being highly certain on natural relative to synthetic trials (cf. Fig. 3B), and their average reaction time is approximately 3.7 seconds faster when seeing natural instead of synthetic reference images (cf. Fig. 3C). Taken together, these findings indicate that in our setup, participants are not just better overall, but also more confident and substantially faster when provided with natural images.

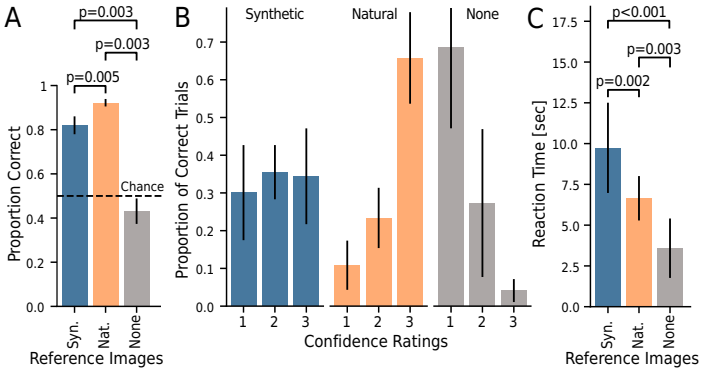


Figure 3: Participants are better, more confident and faster at judging which of two query images causes higher unit activation with natural than with synthetic reference images. **A: Performance.** Given synthetic reference images, participants are well above chance (proportion correct: $82 \pm 4\%$), but even better for natural reference images ($92 \pm 2\%$). Without references ("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. **C: Reaction time.** For correctly answered trials, participants are on average faster when presented with natural than with synthetic reference images. The p-values correspond to Wilcoxon signed-rank tests.

4.2 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 (cf. Fig. 4A). 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. Interestingly, both experts and lay participants are better in the natural than in the synthetic condition.

4.3 Additional information boosts performance, especially for natural images

Publications on feature visualizations vary in terms of how optimized images are presented [12, 9, 40, 51, 36, 39]. Naturally, the question arises as to what influence (if any) these presentation choices have. We here systematically compare presentations schemes along two dimensions (cf. Fig. 4B): the number of reference images (1 vs. 9) and the availability of minimally activating images (only Max vs. Min+Max). 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 strongest boost in performance, however, is observed for natural reference images. In the Min+Max 9 condition, a replication of the result from Experiment I shown in Fig. 3A, natural images outperform synthetic images by an even larger margin (91 ± 3 vs. $72 \pm 4\%$).

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 psychophysical experiments with both expert and lay participants, we conduct the very first investigation of *intermediate* (rather than decision-level) synthetic feature visualizations: 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 compared 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 natural strongly activating images. Interestingly, many synthetic feature visualizations contain regularization mechanisms to introduce more "natural structure" [39], sometimes even called a "natural image prior" [29]. This raises the question: Are natural images all you need? One might posit that highly-activating natural (reference) images simply appear more similar to other highly-activating natural (query) images. While that might be true, feature visualizations are meant to explain feature map activations for natural images, and this is ultimately what real-world applications of XAI are concerned with. To put the above-chance performance with synthetic feature visualizations into perspective, we point out that subjective impressions of the interpretability of visualizations vary greatly between participants (cf. Appendix Sec. A.2.5).

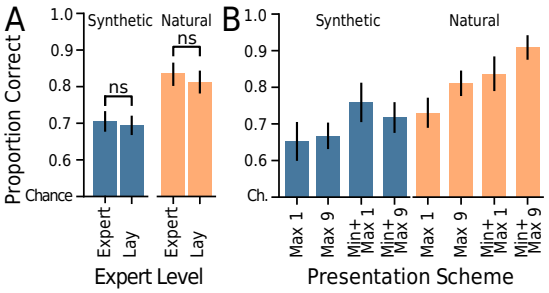


Figure 4: There is no evidence for large effects of expert level; however, performance does improve with additional information. **A: Expert level.** Both experts and lay participants perform equally well (RM ANOVA, $p = .44$), and consistently better on natural than on synthetic images. **B: Presentation scheme.** Presenting both maximally and minimally activating images simultaneously (Min+Max) and showing nine instead of one reference image tends to improve performance, especially for natural reference images. "ns" highlights non-significant differences.

(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 [19, 26] or only laypeople [45, 4]. 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. Consequently, future studies may not have to rely on selected expert participants, but may leverage larger lay participant pools.

(3.) *What is the best way of presenting images?* Existing work suggested that more than one example [38] and particularly negative examples [24] 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.

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. 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. Finally, while we explored one particular method in depth [39]; it remains an open question whether the results can be replicated for other feature visualizations methods.

Future directions. Besides using query images with more similar activations, future participants could be provided with more information, e.g. where a feature map is located in the network. Additionally, further studies could explore single unit visualizations, combinations of units and different networks.

Taken together, our results highlight the need for thorough quantitative evaluations of feature visualizations and suggest that natural images provide a surprisingly challenging baseline for understanding CNN activations.

References

[1] OpenAI Microscope. https://microscope.openai.com/models. (Accessed on 09/12/2020).

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[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] 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.

[10] 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.

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

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

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

[14] 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.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

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

[17] Fabio M. Graetz. How to visualize convolutional features in 40 lines of code, Jan 2019. URL `Howtovisualizeconvolutionalfeaturesin40linesofcode`.

[18] 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.

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

[20] 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.

[21] 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*, pages 9737–9748, 2019.

[22] 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*, pages 125–136, 2019.

[23] JASP Team. JASP (Version 0.13.1), 2020. URL `https://jasp-stats.org/`.

[24] 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*, pages 2280–2288, 2016.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015.

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

[34] 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.

[35] 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, 2016.

[36] 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*, 2016.

[37] 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*, pages 4467–4477, 2017.

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

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

[40] 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.

[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] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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*, pages 1–9, 2015.

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

[50] 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*, pages 422–430. Springer, 2020.

[51] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

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 [11]). Answers to the two-alternative forced choice paradigm (cf. Supplementary Material for screenshots) are recorded together with the participants' confidence level (1: not confident, 2: somewhat confident, 3: very confident). Time per trial is unlimited and we record reaction time. After each trial, feedback is given. As reference images, either synthetic, natural or no reference images are shown. 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 (cf. Fig. 5A, 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*. Practice trials give participants the opportunity to familiarize themselves with the task. In order to monitor the attention of participants, *catch trials* appear 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 (cf. Supplementary Material). 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* (cf. Supplementary Material). 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, expert participants in Experiment I and 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 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.

Finally, to avoid learning effects, we use different feature maps for explanation, practice, and catch trials than for main trials.

Experiment I The first experiment focuses on comparing performance of synthetic images to two baselines: natural reference images and no reference images (cf. Fig. 5A). In total, 45 feature maps are tested: 36 of these are sampled from all nine Inception modules and all branches within these modules. The other nine feature maps are hand-picked for interpretability from the modules' pooling branch based on the appendix overview selection provided by Olah et al. [39] or based on our own choices. In the spirit of a *general* statement about the explanation method, different participants see different natural 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 is about comparing different presentation schemes[1] (Max 1, Min+Max 1, Max 9 and Min+Max 9, cf. Fig. 5E) as well as testing expert vs. lay participants (cf. Fig. 5B). In total, 80 feature maps are tested: They are 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. However, expert and lay participants see the same images. For details on the counter-balancing of all conditions, please refer to the Supplementary Material.

[1] In pilot experiments, we learned that participants preferred 9 over 4 reference images, hence this "default" choice in Experiment I.

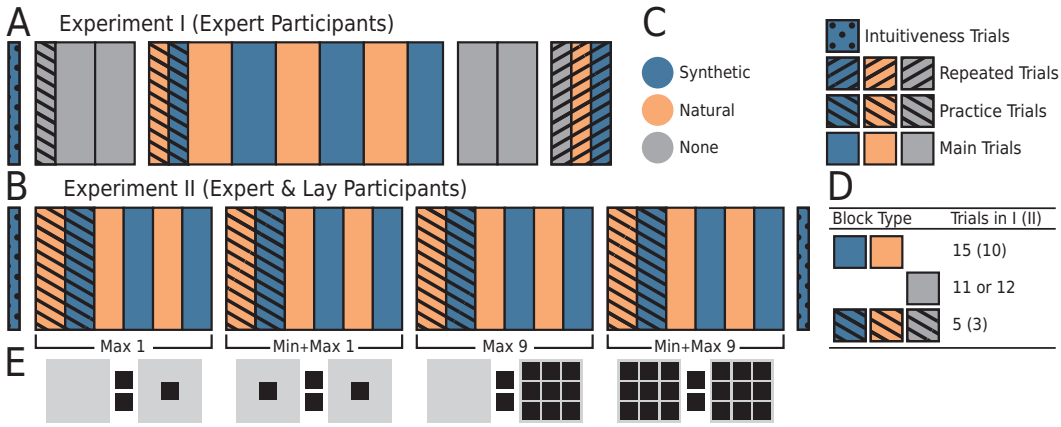


Figure 5: Detailed structure of the two experiments with different foci. **A: Experiment I.** Here, the focus is on comparing performance of synthetic and natural reference images to the most simple baseline: no reference images ("None"). To counter-balance conditions, the order of natural and synthetic 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 comparing different presentation schemes (Max 1, Min+Max 1, Max 9 and Min+Max 9, see **E** for illustrations) as well as testing expert and lay participants. Both the order of natural and synthetic blocks as well as the four presentation conditions are counter-balanced. 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. **E:** Illustration of presentation schemes. In Experiment II, all four schemes are tested, in Experiment I only Min+Max 9 is tested.

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 (cf. Supplementary Material): 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. [39]. In theory, this again allows for the highest intuitiveness ratings possible. The specific units are from a low, intermediate and high layer: feature map 43 in mixed3a, feature map 504 in mixed4b and feature map 17 in mixed 5b.

Participants Our two experiments are within-subject studies, meaning that every participant answers trials for all conditions. 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 familiarity with convolutional neural networks. All subjects 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 subjects give written informed consent for participating. All subjects 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 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 [43, version 3.0] under Python 3.6.

A.1.2 Stimuli Selection

Model Following the existing work on feature visualization [39, 40, 42, 41], we use an Inception V1 network[2] (also known as "GoogLeNet") [48] trained on ImageNet [10, 44]. Note that the Inception V1 network used in the mentioned previous 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[3]. We investigate feature visulizations for all branches (i.e. kernel sizes) of the Inception modules and sample from layers mixed3a to mixed5b before the ReLU non-linearity.

Synthetic Images from Feature Visualization The synthetic images throughout this study are the optimization results of the Feature Visualization method by Olah et al. [39]. We use the channel objective to find synthetic stimuli that maximally (minimally) activate a given feature map of the network. We perform the optimization using lucid 0.3.8 and TensorFlow 1.15.0 [2] and use the hyperparameter as specified in Olah et al. [39]. For the experimental conditions with more than one minimally (maximally) activating reference image, we add a diversity regularization 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 [44] dataset. To choose the maximally (minimally) activating natural stimuli for a given feature map, we perform three steps (for a visual illustration, see Supplementary Material): 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 crop. Second, we order the stimuli 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. 5, **E**) and $N_{batches} = 20$ determines the maximum number of participants we can test with our setup. Third, we distribute the selected stimuli into $N_{stimuli} + 1$ blocks. Lastly, we create $N_{batches}$ batches of data by randomly choosing one image from each of the blocks for every batch.

The reason for creating several batches of extremely activating natural images is 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 is done as follows: In Experiment I, the none condition is always allocated the batch with $batch_nr = subject_id$, the natural condition with $batch_nr = subject_id + 1$, and the optimized condition with $batch_nr = subject_id + 2$. The allocation scheme in Experiment II can be found in the Supplementary Material.

Selection of Feature Maps The selection of feature maps used in Experiment I is shown in the Supplementary Material.

A.1.3 Data Analysis: Significance Tests

All significance tests were performed with JASP [23, version 0.13.1]. For the analysis of the distribution of confidence ratings (cf. Fig. 3B) we used contingency tables with χ^2-tests. For testing pairwise effects in accuracy, confidence, reaction time as well as intuitiveness data, we report

[2] This network is considered very interpretable [40], yet other work also finds deeper networks more interpretable [5]. More recent work, again, suggests that "analogous features [...] form across models [...]," i.e. that interpretable feature visualizations appear "universally" for different CNNs [42, 1].

[3] `github.com/tensorflow/lucid/tree/v0.3.8/lucid`

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.2 Details on results

A.2.1 Complementing figures for main results

Fig. 6 complements the results and figures of Sec. 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 ($F(1, 21) = 0.6$, $p = .44$, 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.2.3 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 interpetability method as a whole. Olah et al. [39] identify a number of particularly "interpretable" network 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. 8); however this performance difference is small and not significant for both natural (Wilcoxon test, $Z(9) = 27.5$, $p = 0.59$) and synthetic (Wilcoxon test, $Y(9) = 41$, $p = 0.18$) reference images. 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. Nonetheless, consistent with the findings reported in the main paper, performance is higher for natural than for synthetic reference images *even on carefully selected hand-picked feature maps*.

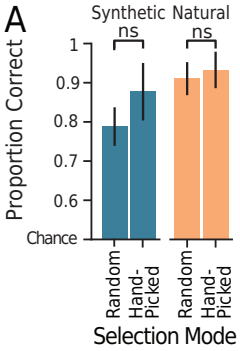
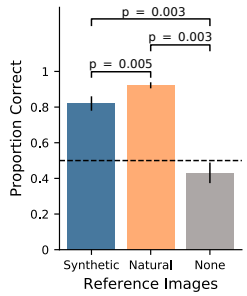


Figure 8: There is no significant performance difference between hand-picked feature maps selected for "interpretability" and randomly selected ones (Wilcoxon test, $p = 0.18$ for synthetic and $p = 0.59$ for natural reference images).

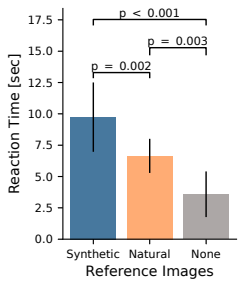
A.2.4 Natural images are more helpful across a broad range of layers

In our experiments, we also take a more fine-grained look at performance across different layers and branches of the Inception modules (cf. Fig. 7). Generally, feature map visualizations from lower layers show low-level features such as striped patterns, color or texture, whereas feature map visualizations from higher layers tend to show more high-level concepts like (parts of) objects [27, 18, 15]. We find performance to be reasonably high across most layers and branches: participants were 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.

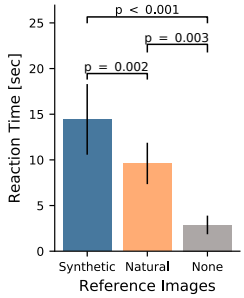
[4] All our hand-picked feature maps are taken from the pooling branch of the Inception module. As the appendix overview in Olah et al. [39] does not contain one feature map for each of these, *we* select interpretable feature maps for the missing layer mixed5a and mixed5b ourselves.



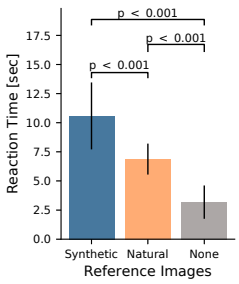
(a) Task performance.



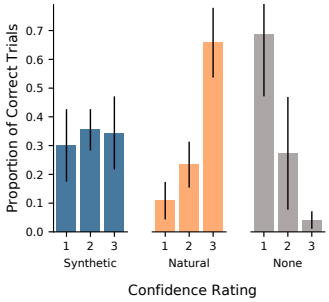
(b) Reaction time on correct trials.



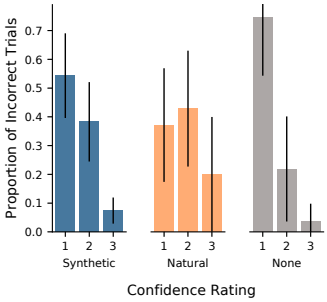
(c) Reaction time on incorrect trials.



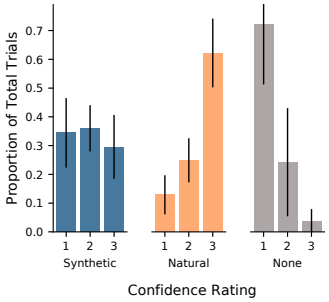
(d) Reaction time on all trials.



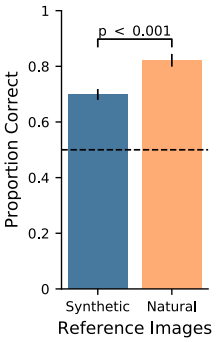
(e) Confidence ratings on correctly answered trials.



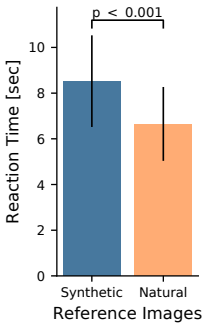
(f) Confidence ratings on incorrectly answered trials.



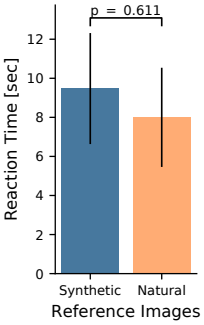
(g) Confidence ratings on all trials.



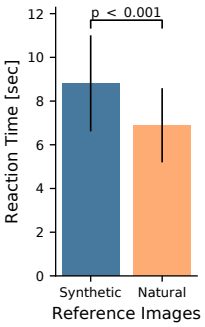
(h) Performance.



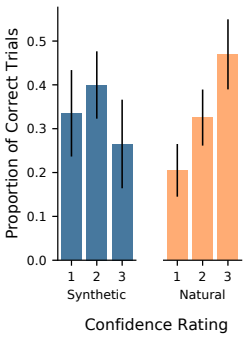
(i) Reaction time on correctly answered trials.



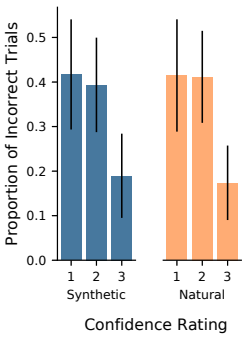
(j) Reaction time on incorrectly answered trials.



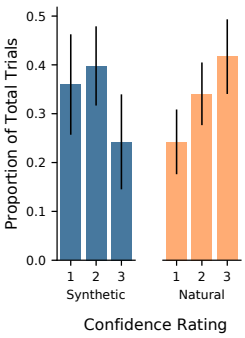
(k) Reaction time on all trials.



(l) Confidence ratings on correctly answered trials.

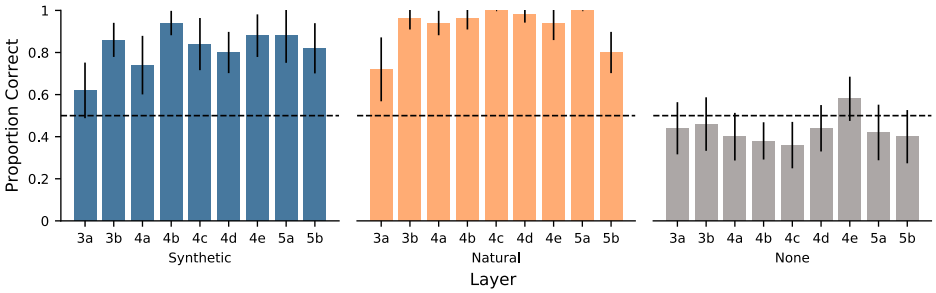


(m) Confidence ratings on incorrectly answered trials.

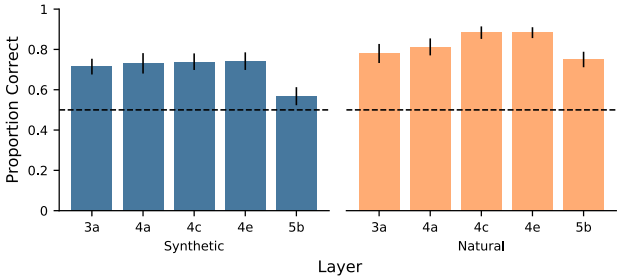


(n) Confidence ratings on all trials.

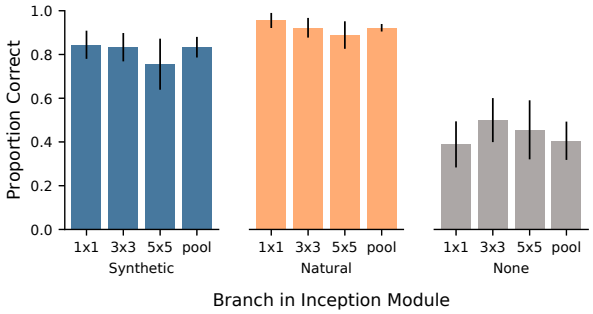
Figure 6: Results for Experiment I (II) are shown in the first (last) two rows: task performance (a, h), reaction times (b-d, i-k) and distribution of the confidence ratings (e-g, l-n). The p-values are calculated with Wilcoxon sign-rank tests.



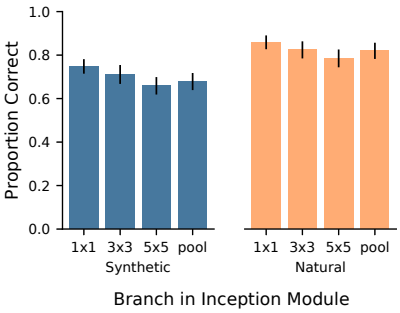
(a) Performance across layers in Experiment I.



(b) Performance across layers in Experiment II.



(c) Performance across branches in Experiment I.



(d) Performance across branches in Experiment II.

Figure 7: High performance across (a, b) layers and (c, d) branches of Inception module.

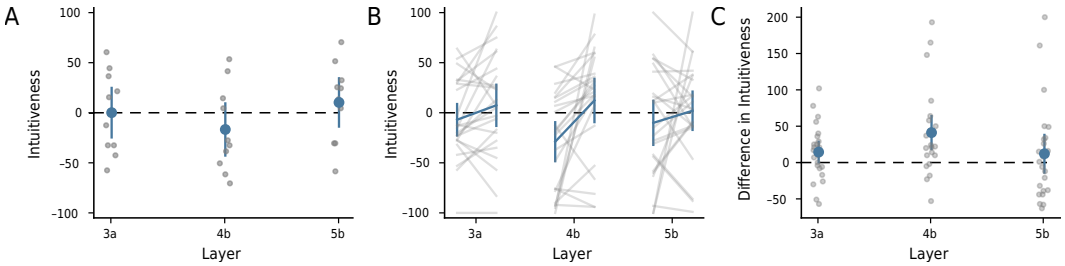


Figure 9: The subjective intuitiveness of feature visualizations varies greatly (cf. **A** for the ratings from the beginning of Experiment I and **B** for the ratings at the beginning and end of Experiment II). The means over all subjects yield a neutral result, i.e. the visualizations are neither un- 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.

A.2.5 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 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* (cf. Supplementary Material), i.e. ratings of how intuitive feature visualizations appear for natural images. As Fig. 9A+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 (cf. Fig. 9B+C, Wilcoxon test, $p < .001$).

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 alluded to in the Discussion of the main paper, 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 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).

Furthermore, the interactive conversations painted a similar picture as the intuitiveness trials (cf. Sec. A.2.5): 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. A few participants point out that the synthetic visualizations are exhausting to understand. Three participants additionally emphasize that the minimally activating reference images played an important role in their decision-making.

A.2.7 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; in Experiment II, five participants miss one trial and four participants miss two trials. 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.