# EXEMPLARY NATURAL IMAGES EXPLAIN CNN ACTIVATIONS BETTER THAN FEATURE VISUALIZATIONS

**Anonymous authors**
Paper under double-blind review

## 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 (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 synthetic images by a wide margin ($92\%$ accuracy). Additionally, participants are faster and more confident for natural images, whereas subjective impressions about the interpretability of feature visualization 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 conclusion, popular synthetic images from feature visualizations are significantly less informative for assessing CNN activations than natural images. We argue that future visualization methods should improve over this simple baseline.

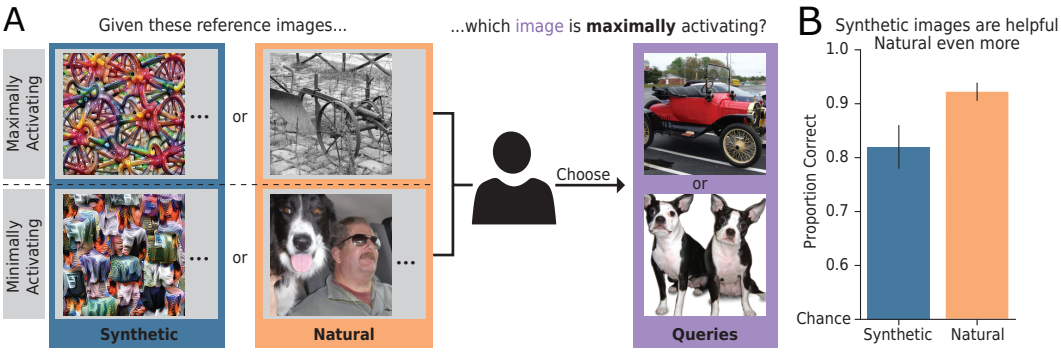


Figure 1: How useful are synthetic compared to natural images for interpreting neural network activations? **A: Human experiment.** Given extremely activating reference images (either synthetic or natural), a human participant chooses which out of two query images is also a maximally activating image. 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.

## 1 INTRODUCTION

As Deep Learning methods are being deployed across society, academia and industry, the need to understand their decisions becomes ever more pressing. 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 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). As such, a particular milestone for CNNs was understanding that features are 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 largely 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.

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. (2017). 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.
- Impressions of the interpretability of visualizations vary greatly between participants.

[1] also known as *input maximization* or *maximally exciting images*

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 the maximally activating one. For screenshots of each step in the task, cf. Appendix Fig. 7.

## 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 (e.g. (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 (e.g. 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; Ope), 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., 2019b). As overviews, we recommend Zhang & Zhu (2018); Montavon et al. (2018) and Samek et al. (2020).

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. For purely decision-level explanation methods like saliency maps (that do not aim to explain intermediate representations), studies have shown that the putative explanations can be similar for trained and randomized nets (Adebayo et al., 2018), that they can be affected by meaningless data processing steps (Kindermans et al., 2017), and that they are vulnerable to adversarial attacks (Ghorbani et al., 2019a).

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 inves-

tigate 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. Unfortunately, how differences between conditions should be interpreted is often unclear. 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)).

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: By presenting data for a forward prediction task of a CNN, we provide a quantitative estimate of the informativeness of maximally activating images. Furthermore, our data is unique as it probes 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 be a maximally activating image given some reference images (cf. Fig. 2). Apart from image choices, we record participants' confidence levels and their reaction times. Specifically, responses are given by clicking on the confidence levels belonging to either query image. 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 Appendix Sec. A.2.5).

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.1.1.

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 (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 et al., 2015). The synthetic images throughout this study are the optimization results of the feature visualization method by Olah et al. (2017). The natural stimuli are selected from the validation set of the ImageNet ILSVRC 2012 (Russakovsky et al., 2015) dataset according to their activations for the feature map of interest.

## 4 RESULTS

In this section, all figures show data from Experiment I except Fig. 5A+C (Experiment II). All figures for Experiment II, which replicate the findings of Experiment I, as well as additional figures for Experiment I, 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.

[2]Code will be released after the anonymous review period.

[3]also known as GoogLeNet

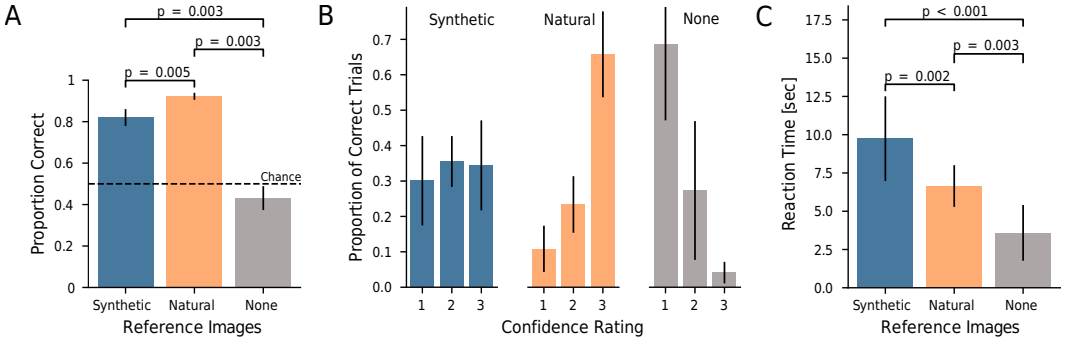


Figure 3: Participants are better, more confident and faster at judging which of two query images causes higher unit activation with natural than with synthetic reference images. **A: Performance.** Given synthetic reference images, participants are well above chance (proportion correct: $82 \pm 4\%$), but even better for natural reference images ($92 \pm 2\%$). Without 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 ($\chi^2$, $p < .001$). **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. 15); for Experiment II, see Fig. 16.). The $p$-values in A and C correspond to Wilcoxon signed-rank tests.

### 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 ($50\%$) (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 correct answers, 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 than 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 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 (cf. 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 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 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.

### 4.3 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 (cf. Fig. 5A). Contrary to our prior expectation, we find no significant differences in expert vs. lay performance (RM ANOVA, $p = .444$, for details see Appendix Sec. A.2.2). Hence, extensive experience with CNNs is not necessary to perform well in this forward simulation task. Both experts and lay participants are both better in the natural than in the synthetic condition.

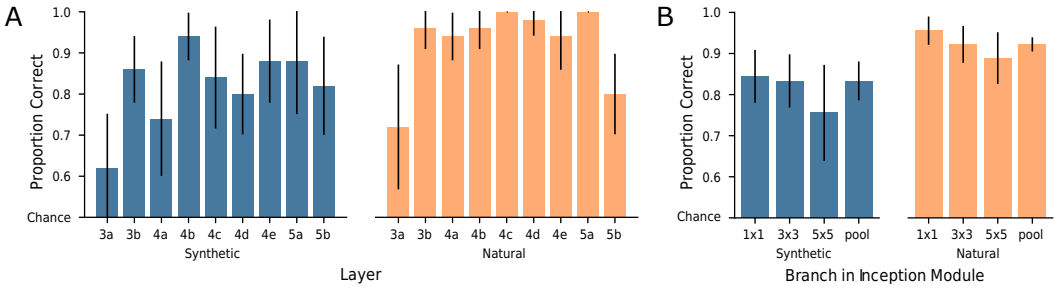


Figure 4: Performance is consistently high across (**A**) a broad range of layers and (**B**) all branches of the Inception modules. The latter differ in their kernel sizes ($1 \times 1$, $3 \times 3$, $5 \times 5$, pool). Again, natural images are (mostly) more helpful than synthetic images. Additional plots for the none condition as well as Experiment II can be found in the Appendix in Fig. 17 and Fig. 18.

### 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 interpetability method as a whole. Olah et al. (2017) 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. 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.2.3 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: Rarely, 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 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 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 ($91 \pm 3$ 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* (cf. Appendix Fig. 13) and oral impressions after the experiments. The former ask

[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 layer mixed5a and mixed5b ourselves.

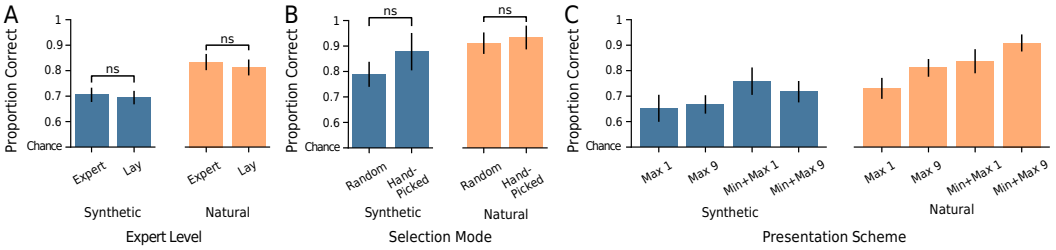


Figure 5: We found no evidence for large effects of expert level or feature map selection. However, performance does improve with additional information. **A: Expert level.** Both experts and lay participants perform equally well (RM ANOVA, $p = .44$), and consistently better on natural than on synthetic 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.

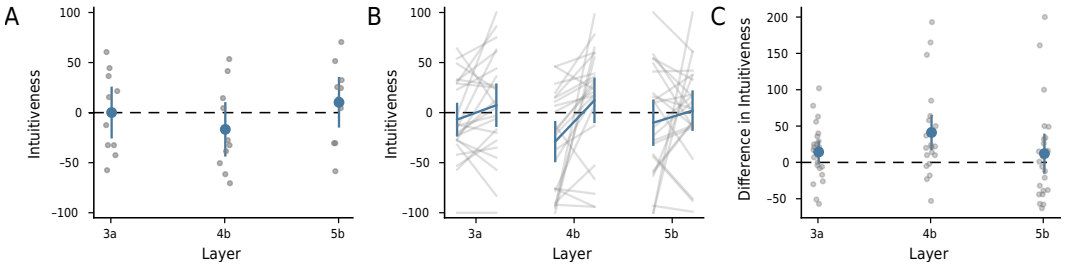


Figure 6: The subjective intuitiveness of feature visualizations varies greatly (cf. **A** for the ratings from the beginning of Experiment I and **B** for the ratings at the beginning and end of Experiment II). The means over all subjects yield a neutral result, i.e. the visualizations are neither un- 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.

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 (cf. Fig. 6B+C, Wilcoxon test, $p < .001$). 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. 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.

## 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 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 much more 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" (Olah et al., 2017), sometimes even called a "natural image prior" (Mahendran & Vedaldi, 2015). This raises the question: Are natural images maybe 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.

(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 suggested 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. 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 setup 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. 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 quantitative evaluations of feature visualizations and suggest that example natural images provide a surprisingly challenging baseline for understanding CNN activations.

## REFERENCES

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

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.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

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.

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.

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

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, 2019b.

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 `Howtovisualizeconvolutionalfeaturesin40linesofcode`.

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/`.

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.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

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.

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.

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.

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.

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.

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.