How Well do Feature Visualizations Support Causal Understanding of CNN Activations?

Anonymous Author(s)
Affiliation
Address
email

Abstract

A precise understanding of why units in an artificial network respond to certain stimuli would constitute a big step towards explainable artificial intelligence. One widely used approach towards this goal is to visualize unit responses via activation maximization. These feature visualizations are purported to provide humans with precise information about the image features that *cause* a unit to be activated — an advantage over other alternatives like highly activating dataset samples. If humans indeed gain causal insight from visualizations, this should enable them to predict the effect of an intervention, such as whether occluding a certain patch of the image (say, a dog's head) changes a unit's activation. Here, we test this hypothesis by asking humans to decide which of two square occlusions causes a larger change to a unit's activation. Both a large-scale crowdsourced experiment and measurements with experts show that on average the extremely activating feature visualizations by Olah et al. [36] indeed help humans on this task ($68 \pm 4\,\%$ accuracy; baseline performance without any visualizations is $60 \pm 3\,\%$). However, they do not provide any significant advantage over other visualizations (such as e.g. dataset samples), which yield very similar performance ($65 \pm 3\,\%$ to $66 \pm 5\,\%$ accuracy). Taken together, we propose an objective psychophysical task to quantify the benefit of unit-level interpretability methods for humans, and find no evidence that a widely-used feature visualization method provides humans with better "causal understanding" of unit activations than simple alternative visualizations.

1 Introduction

It is hard to trust a black-box algorithm, and it is hard to deploy an algorithm if one does not trust its output. Many of today's best-performing machine learning models, deep convolutional neural networks (CNNs), are also among the most mysterious ones with regards to their internal information processing. CNNs typically consist of dozens of layers with hundreds or thousands of units that distributively process and aggregate information until they reach their final decision at the topmost layer. Shedding light onto the inner workings of deep convolutional neural networks has been a long-standing quest that has so far produced more questions than answers.

One of the most popular tools for explaining the behavior of individual network units is to visualize unit responses via activation maximization [14, 29, 34, 31, 35, 32, 49, 13]. The idea is to start with an image (typically white noise) and iteratively change pixel values to maximize the activation of a particular network unit via gradient ascent. The resulting synthetic images, called *feature visualizations*, often show interpretable structures, and are believed to isolate and highlight exactly those features that "cause" a unit's response [36, 46]. Several works leverage this method to study InceptionV1 [38, 37, 7, 39, 46, 8, 51, 52, 42] and other networks [5, 19, 18]; others create interactive tools [53, 40] or introduce analysis frameworks [57]. Some of the synthetic feature visualizations

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

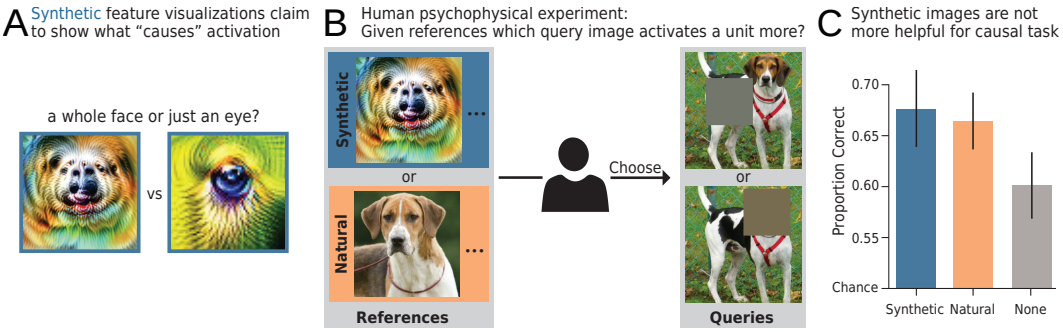


Figure 1: How useful are feature visualizations to interpret the effects of interventions? **A: "Causal" synthetic feature visualizations. B: Human experiment.** Given strongly activating reference images (e.g. synthetic or natural), a human participant chooses which out of two manipulated images activates a unit more. Note that this trial is made up — real trials are often more difficult. **C: Core results.** While participants are above chance for all visualization types, synthetic images only provide a significant advantage over *no* references and not over other alternatives such as natural references.

appear quite intuitive and precise. As shown in Fig. 1A, they might be able to distinguish whether, for example, a unit responds to just an eye or a whole dog's face.

However, other aspects cast a more critical light on feature visualization's "causality": Generating these synthetic images typically involves regularization mechanisms [32], which may influence how faithfully they visualize what "causes" a network unit's activation. Furthermore, to obtain a complete description of a mathematical function, one generally needs more information than just knowing the location of the extrema. In view of this, it is an open question how well a unit can be characterized by simply visualizing the arguments of its maxima. Finally, a crucial unknown factor is whether *humans* are able to obtain a causal understanding of CNN activations from these synthetic visualizations.

Given these points, we develop a psychophysical experiment to test whether feature visualizations by Olah et al. [36] indeed allow humans to gain a causal understanding of a unit's behavior. Our task is based on the reasoning that being able to predict the effect of an intervention is at the heart of causal understanding. Understanding the causal relation between variables implies an understanding of how changes in one variable affect another one [41]. In our proposed experiment, this means that participants could predict the effect of an intervention — in form of an image manipulation — if they knew the causal relation between image features and a unit's activations. Our experiment tests whether synthetic feature visualizations indeed provide information about such causal relations. Specifically, we ask humans which of two interventions on an image activates a CNN unit more strongly. The interventions we test are obtained by placing an occlusion patch at two different locations in an image. This experiment both probes the purported explanation method's advantage of causality as well as how informative its visualizations are for humans in a counterfactual-inspired prediction task [12].

Besides feature visualizations, other visualization methods have been used to gain an understanding of the inner workings of CNNs. In this experiment, we additionally test alternatives based on natural dataset examples and compare them with feature visualizations. This is particularly interesting because dataset examples are often assumed to provide less "causal" information about a unit's response, since they might contain misleading correlations. To continue the example above, dog eyes usually co-occur with dog faces; thus, separating the influence of one image feature from the other one using natural exemplars might be challenging.

Our data shows that:

- Synthetic feature visualizations provide humans with some helpful information about the most important patch in an image — but not much more information than no visualizations at all.
- Dataset samples as well as other combinations and types of visualizations are similarly helpful.
- How easily the most important patch is identifiable depends on the unit, the images as well as the relative activation strength attributed to the patch.

2 Related Work

Feature visualizations are a widely used method to understand the learned representations and decision-making of CNNs [29, 34, 31, 35, 32, 49, 13, 36, 33]. However, others question whether this synthetic visualization technique, first introduced by Erhan et al. [14], is too intuition-driven [24], and how representative the appealing visualizations in publications are [23]. Further, as already mentioned above, the engineering of the loss function may influence their faithfulness [32]. Another challenge is generating *diverse* feature visualizations to represent the different aspects that one single unit responds to [38, 32]. Finally, a recent human evaluation study [4] found that while these synthetic images do provide humans with helpful information in a forward simulation-inspired task, simple natural dataset examples are even more helpful.

Human evaluation studies are extensively used to quantify various aspects of interpretability. As an alternative to pure mathematical approximations [2, 58, 50, 55], researchers not only evaluate the understandability of explanation methods in psychophysical studies [6, 30, 4], but also trust in these methods [25, 56]) as well as the human cognitive load necessary for parsing explanations [1] or whether humans would follow an explained model decision [43, 11, 44]. A recent study even demonstrates that metrics of the explanation quality computed *with* human judgment are more insightful than those without [3].

Counterfactuals are a popular paradigm for both *creating* as well as *evaluating* explanation methods. Intuitively, they provide answers to the question "what should I change to achieve a different outcome?" — in the context of machine learning explanation methods, usually the smallest, realistic change to a data point is of interest. As examples, counterfactual explanation methods have been developed for vision- [20] and language-based [54] models as well as for model-agnostic scenarios [47]. To *evaluate* — rather than create — explanation methods, researchers often follow the "counterfactual simulation" task introduced by Doshi-Velez and Kim [12]: Humans are given an input, an output, and an explanation and are then asked "what must be changed to change the method's [model's] prediction to a desired output?" Based on this task, Lucic et al. [27] test their new explanation method and Hase and Bansal [22] compare different explanation methods to each other.

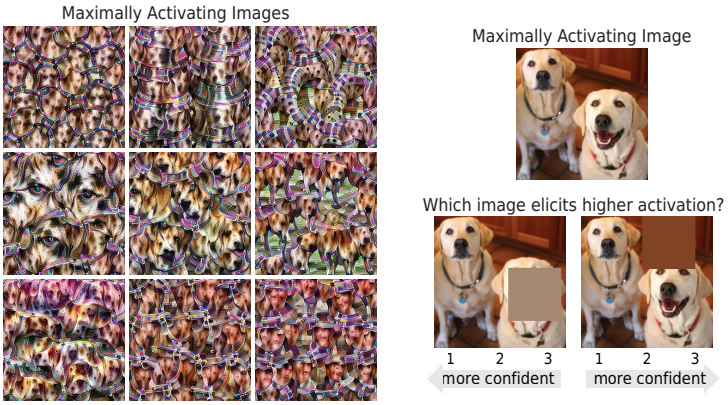


Figure 2: Schematic visualization of an example trial in our psychophysical experiment. For a certain network unit, participants are shown several maximally activating images. While the ones on the left serve as reference images, the ones on the right serve as query images: The top one is a natural maximally activating image and the bottom ones are copies of said image with square occlusions at different locations. The task is to select the image that activates the given network unit more strongly. The answer is given by clicking on the number below the corresponding image according to the participant's confidence level (1: not confident, 2: somewhat confident, 3: very confident).

In this project, we design a counterfactual-inspired task to evaluate how well feature visualizations support causal understanding of CNN activations. This is the first study to apply such a paradigm to understanding the causes of individual units' activations. In order to scale the experiments, we simplify our task by having participants choose between two intervention *options*, rather than freely determine interventions themselves.

3 Methods

We run an extensive psychophysical experiment with more than $12,000$ trials distributed over 323 crowd-sourced participants on Amazon Mechanical Turk (MTurk) and two experts (the two first authors).[1] For a more detailed description of the methods, please see Appx. Sec. A.1.

Design Principles Overall, our experimental design choices aim at (1) the *best performance possible*, meaning that we select images that make the signal as clear as possible; (2) *generality* over the network, meaning that we randomly sample units of different layers and branches (testing all units would be too costly); and (3) *easy extendability*, meaning that we chose a between-participant design (each participant sees only one reference image condition) so that other visualizations methods can be added to the comparisons in the future.

3.1 Psychophysical Task

If feature visualizations indeed support causal understanding of CNN activations, this should enable humans to predict the effect of an intervention, such as whether occluding an image region changes a unit's activation. Based on this idea, we employ a two-alternative forced choice task (chance performance: 50%) where human observers are presented with two different occlusions in an image, and asked to estimate which of them causes a smaller change to the given unit's activation (see Fig. 2 for an example trial). More specifically, participants choose the *query* image that they believe to also elicit a strong activation given a set of *reference* images. Such references could for instance consist of synthetic feature visualizations of a certain unit (purportedly "causal"), or alternative visualizations. To summarize, the task requires humans to first identify the shared aspect in the reference images and to then choose the query image in which that aspect is more visible. Since we do not make any assumptions about whether participants are familiar with machine learning, we avoid asking participants about activations of a unit in the CNN. Instead, we explain that an image would be "favored" by a machine, and the task is to select the image which is "more favored". The complete set of instructions shown to participants can be found in Appx. Fig. 8 and 9. In addition to each participant's image choice, the subjective confidence level and reaction time are also recorded.

3.2 Stimulus Generation

To generate stimuli, we use an InceptionV1 network [48] trained on ImageNet [10, 45], following Olah et al. [36]. Throughout this paper, we refer to a CNN's channel as a "unit" and imply taking the spatial average of all neurons in one channel.[2] We test units sampled from 9 layers and 2 Inception module branches (namely 3×3 and POOL). For more details on the generation procedures of the respective stimuli, see Appx. A.1.2.

We use five different types of **reference images**:

- **Synthetic references**: The synthetic images are the optimization results of the feature visualization method by Olah et al. [36] with the channel objective for 9 diverse images.
- **Natural references**: The reference images are the most strongly activating[3] dataset samples from ImageNet [10, 45].
- **Mixed references**: This is a combination of the previous two conditions: the 5 most strongly activating natural and 4 synthetic reference images are used. The motivation is that this condition combines the advantages of both worlds — namely precise information from feature visualizations and easily understandable natural images — and, thus, has the potential to give rise to higher performance in the task. Jointly looking at these two visualization types is common in practice [36].
- **Blurred references**: To increase the informativeness of natural images for this task, we modify them by blurring everything but a single patch. This patch is chosen in the same way as in the maximally activating query image (see below). Consequently, this method cues participants to the most important image feature. In a way, these images can be seen as an approximate inverse of the maximally activating query image and might improve performance on our task.

[1] Code is attached and will be made publicly available after the anonymous review period.

[2] Other papers might refer to a channel as a "feature map" [4]

[3] To reduce compute requirements, we use a random subset of the training set ($\approx 50\%$).

- **No references**: This is a control condition in which participants do not see any reference images and have to solve the task purely based on query images.

To generate **query images**, we place a square patch of 90×90 pixels of the average RGB color of the occluded pixels into a most strongly activating image chosen from ImageNet. The location of the occlusion patch is chosen such that the activation of the manipulated image is most extreme among all possible occlusion locations. Then, we select the manipulated images eliciting the largest and the smallest activation as the target and distractor query images respectively.

3.3 Structure of the Psychophysical Experiment

We test the five different reference image types as separate experimental conditions. In each condition, we collect data from a total of 50 different MTurk participants, each assigned to a single Human Intelligence Task (HIT) consisting of an instruction block, a variable number of practice blocks and a main block. The instructions extensively explain a hand-crafted example trial (see Appx. Fig. 8 and 9). The blocks of 4 practice trials each - which are randomly sampled from a pool of 10 trials - have to be repeated until reaching 100% performance; except for the none condition, as there is no obvious ground truth due to the absence of reference images. Finally, 18 main trials follow that are randomly interleaved with a total of 3 obvious catch trials. While feedback is provided during practice trials, no feedback is provided in the other trials. At the end, participants can share comments via an optional free-text field. Across all conditions, all participants see the same query images for the instruction, practice and catch trials. In contrast, the query images do differ in the main trials: In each reference image condition, we test 10 different sets of query images, responded to by 5 different MTurk participants, hence 50 HITs per condition. The order of the main and catch trials per participant is randomly arranged, and identical across conditions. Each MTurk participant takes part in only one reference image condition (i.e. reference images are a between-participants factor). For more details, see Appx. Sec. A.1.4.

3.4 Ensuring High-Quality Data in an Online Experiment

To ensure that the data we collect in our online experiment is of high quality, we take three measures: (1) We only allow MTurk participants to continue to the relevant trials in the main block once they reach perfect performance on a practice block. This means that they are likely to have understood the task when they start the main trials. (2) We integrate hidden checks in the experiment to ensure high-quality data. These *exclusion criteria* comprise a performance threshold for catch trials, a minimum image choice variability, a minimum time spent on both the instructions and the whole experiment as well as a maximum number of iterations through the practice blocks (for details, see Appx. Sec. A.1.1). We set these exclusion criteria before data collection, and only include data from participants who pass all five of them in our analysis. (3) A previous human evaluation study in a well-controlled lab environment found that natural reference images are more informative than synthetic ones when choosing which of two different images are more highly activating for a given unit [4]. We replicate this main finding on MTurk based on a subset of the originally tested units (see Appx. A.2) which indicates that the experiment's environment does not influence this task's outcome. Our decision to leverage a crowdsourcing platform is further corroborated by the result of Borowski et al. [4], that there is no significant difference between expert and lay performance.

3.5 Baselines

In order to both set MTurk participants' performance into context as well as evaluate different strategies participants could use to perform our task, we further evaluate a few baselines including expert responses and simple strategies.

- **Expert Baseline**: The two first authors answer all 18 trials in all 5 reference conditions on all 10 image sets. As they are familiar with the task design and are certainly engaged, this data serves as an upper human bound.
- **Center Baseline**: In natural images from ImageNet, important objects are likely to be closer to the center of the image. If participants were biased to assume that units respond to *objects*, a potential strategy to decide which occluding patch produces a smaller effect on the unit's activation would

therefore be to choose the image with the most eccentric occlusion. The Center Baseline model performs this strategy for all images.

- **Primary Object Baseline**: The Center Baseline is not a perfect measurement of an object-biased strategy because primary objects can appear away from the center. To account for this, the two first authors manually label all trials such that the occlusion hides as little information as possible from the most prominent object in the scene.
- **Variance Baseline**: Another assumption participants might make is that a patch in a low-contrast region, e.g. a blue sky, is unlikely to have a large effect on the unit's activation. This baseline selects the query image whose content is less affected by the introduction of the occlusion patch. To simulate this, we calculate the standard deviation over the occluded pixels and choose the one of the lower standard deviation.
- **Saliency Baseline**: As a complement to the baselines above, this baseline selects the query image whose original pixels hidden by the occlusion patch have a lower probability of being looked at by the participants. This simulates that participants select the image with a patch that occludes less prominent information and is estimated with the saliency prediction model DeepGaze IIE [26]. For more details, see Appx. Sec. A.1.3.

4 Results

The results shown in this section are based on 7350 trials from MTurk participants who passed all exclusion criteria, and experts distributed over five conditions. In all figures, *Synthetic* refers to the purportedly "causal", activation-maximizing feature visualizations, *Natural* refers to ImageNet samples, *Mixed* refers to combined presentation of synthetic and natural images, *Blur* refers to the blurred images, and *None* refers to a condition with no reference images at all. Further, error bars indicate two standard errors above and below the participant-mean over network units and image sets.

4.1 No Significant Advantage of Synthetic Feature Visualizations

If feature visualizations provide humans with useful information about the image features causing high unit activations and other visualizations do not, participants' accuracy in our task should be higher given feature visualizations than for all other visualization types or no reference images. This is only partly what we find: On average, accuracy for feature visualizations is slightly higher than when no reference images are given ($68 \pm 4\%$ vs. $60 \pm 3\%$). However, the accuracy for feature visualizations is not significantly higher than for other visualization methods (see Fig. 3A, dark bars). For the latter, MTurk participants reach between 65 ± 3 % and 66 ± 5 % depending on the visualization type. Statistically, only the condition without reference images is different from all other conditions ($p < 0.05$, Mann-Whitney U test). Taken together, these findings suggests that all visualization methods are similarly helpful for humans in our counterfactual-inspired task, and that they only seem to offer a small improvement over no visualizations at all.

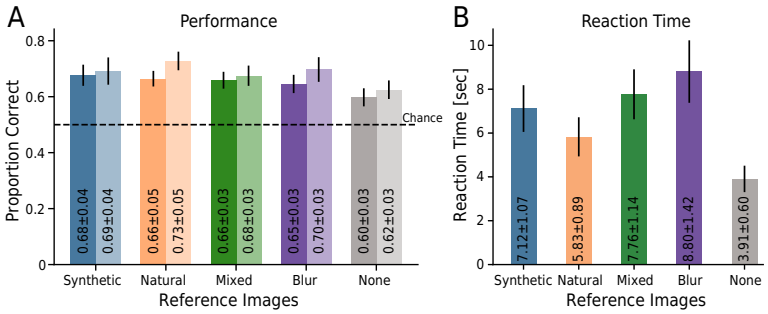


Figure 3: **A: Task accuracy.** On average, humans reach the same performance regime with any visualization method. This holds for both lay participants on MTurk (darker colors) as well as experts (brighter colors). **B: Reaction Times.** MTurk participants need several seconds to answer a trial, indicating that they carefully make their decision. For more details see Appx. Fig. 15.

4.1.1 MTurk Participants Carefully Make Their Choices

Similar performances for various conditions such as those found in Fig. 3A might suggest that participants would not give their best when doing our experiment. However, several aspects speak

against this: (1) Measurement of the two first authors, i.e. experts who designed and thus clearly understand the task, and certainly engage during the experiment, again show very similar performance (see Fig. 3A, bright bars): This estimated upper bound is just $1-7\%$ better than MTurk participant performance. (2) Reaction times per trial (see Fig. 3B) lie between ≈ 4 s and ≈ 9 s. This, as well as the fact that participants take longer for the conditions *with* references than for the *None* condition, suggest that they carefully make their decisions. (3) Relatively large performance differences ($\approx 20\%$) between no visualization and visualization methods do exist for a few feature maps (see Fig. 5). This suggests that participants understand and make use of the available information. (4) Several MTurk participants' comments in an optional free-text field indicate that they engage in the task: "[...] I did my best", "It was engaging", "interesting task". (5) More similar trial-by-trial responses between MTurk participants than expected by chance (see Fig. 4B discussed below) indicate that humans use the available information.

4.1.2 Simple Baselines Can Reach the Same Above-Chance Performance Regime

Decision-making strategies can be diverse. To set human performance into context, we evaluate several simple strategies as baselines: How high is performance if one always chooses the query image with an unoccluded center or primary object? Or such that the more varying or salient image region is unoccluded? Fig. 4A shows that these strategies have varying performances with the best ones — namely the variance and saliency baselines — reaching 61 and 63%. Since already these simple heuristics, which do not require reference visualizations, reach the same performance regime as participants, the additional advantage of visualizations (reaching just $5-7\,\%$ better performance) appears limited.

4.1.3 Decision Patterns are Most Consistent with Saliency Baseline

While accuracy is the most common metric to evaluate task performance, it does not suffice to compare two systems' decision-making processes [28, 17, 16]. Instead, a quantitative trial-by-trial error analysis is necessary to ascertain or distinguish strategies. Here, we use Cohen's kappa [9] to calculate the degree of agreement in classification while taking the expected chance agreement into account. A value of 100% corresponds to perfect agreement, while a value of 0% corresponds to as much agreement as would be expected by chance. Negative values indicate systematic disagreement.

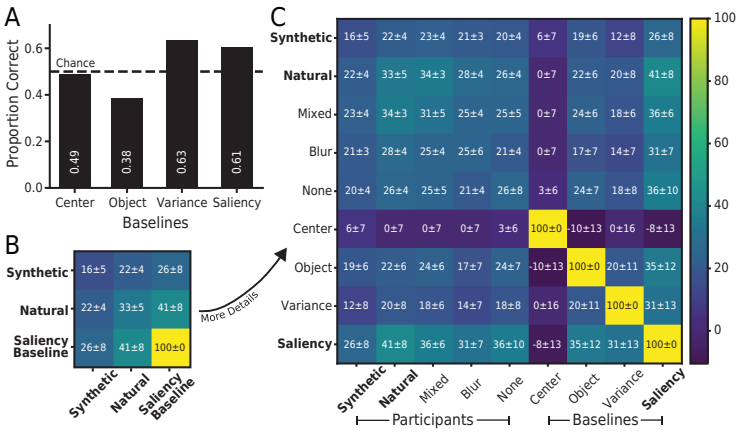


Figure 4: **A: Baseline Performances.** Simple baselines can reach above chance level. **B, C: Decision consistency.** The mean and two standard errors of the mean of Cohen's kappa (in %) averaged over participants and image sets quantifies the pairwise consistency of decision patterns. While they vary across participants, they are higher between conditions with natural references and highest between the saliency baseline and other conditions. For more details, see Appx. Fig 17.

In Fig. 4B and C, we plot consistency between MTurk participants of the same and different reference conditions as well as between MTurk participants and baselines. Since Cohen's kappa only allows for comparisons of two decision makers, we compute this statistic for all possible pairs across image sets, and report the mean and two standard errors of the mean. All values between participants as well as between participants and baselines are in an intermediate regime (up to $41\,\%$). This suggests that there is systematic agreement, but also quite some room for subjective decisions. Strikingly, highest agreement is found for comparisons with the saliency baseline, while lowest agreement is found for comparisons with the center baseline. Within participant to participant comparisons, decision strategies for conditions involving purely natural images (*Natural*, *Mixed*) are more similar to each

other as well as slightly more similar to other strategies than the *Synthetic*, *Blur* or *None* condition to other strategies. Within the *Synthetic* condition, participants are relatively inconsistent.

4.2 Performance Varies across Units, Image Sets and Activation Differences, but Not Reference Condition

Having found that feature visualizations do not offer an overall advantage over other techniques, we now ask: Is performance similar across units, query images and their activation differences?

Units and Image Sets As evident from Fig. 5, performance varies greatly by unit, but not by reference condition: While only one unit (layer 2, 3×3 branch) is clearly below chance level, many units reach around average performance and a few units stand out with high performances (e.g. layer 8, POOL). Further, all five reference conditions are relatively close to each other for all units. Fig. 6 further illustrates the different difficulty levels as well as the strong unit- and image-dependency: For the easy feature map (Fig. 6A), the feature is fairly clearly identifiable and visible in both reference and query images. In contrast, for the difficult feature map (Fig. 6B), the unit's feature selectivity is unclear not only in the reference but also in the query images.

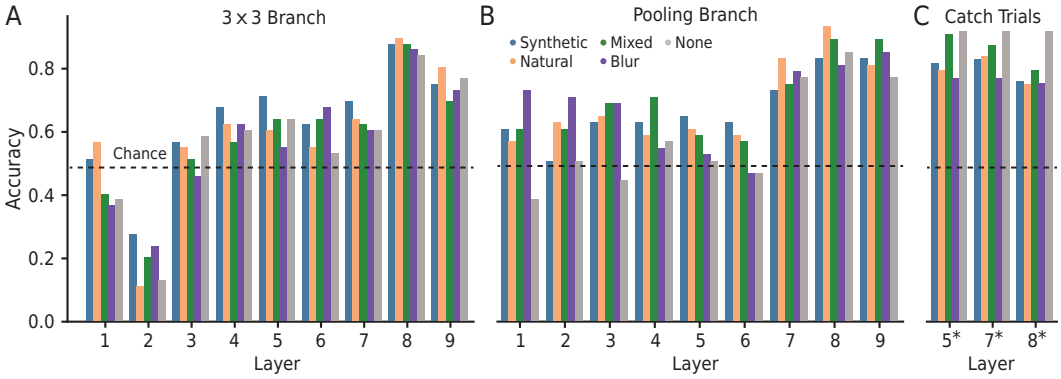


Figure 5: While for some units predicting the effect of an intervention is relatively easy, for most units performance is close to chance. **A** and **B** show the performance per unit in the main trials separated by branch (3×3 and POOL respectively) and layer. **C** shows the performance per unit in the hand-picked trials used as catch trials (hence the *), though selected from those MTurk participants who pass the exclusion criteria without the catch trial exclusion criterion. Note that each bar represents participant-averages over image sets.

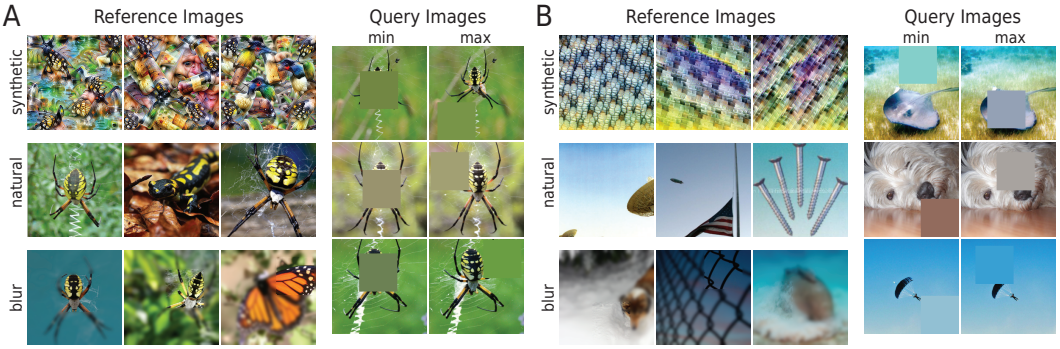
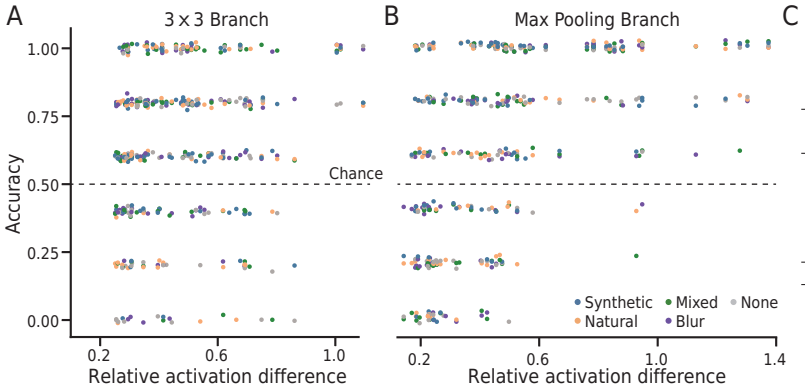


Figure 6: Example reference and query images for a unit with high (**A**) and low (**B**) performance from layer 8 and 2 of the POOL branch, respectively.

Activation Differences We hypothesize that our task might be easier if the difference in activations between the two interventions of the query images is larger. In Fig. 7A and B, we plot by-image-set performance against the relative activation differences, i.e. the difference between activations elicited by the two manipulated images normalized by the unperturbed query image's activation. The figure

shows that even though we select query images as the most strongly activating images for a feature map, the relative activation differences vary widely. Furthermore, performance indeed tends to increase with higher relative activation difference, confirming our hypothesis. This trend is stronger in the POOL than in the 3×3 branch as quantified by the Spearman's rank correlations in Fig. 7C.



	Branch	
References	3×3	Pool
Natural	0.175	0.479*
Optimized	0.103	0.493*
Mixed	0.235*	0.496*
Blur	0.156	0.601*
None	0.259*	0.518*
All	0.183*	0.515*

Figure 7: Performance tends to increase with the relative activation difference between query images. This effect is stronger for the POOL branch (**B**) than for the 3×3 branch (**A**) as quantified by Spearman's rank correlations (**C**). Stars signal significance ($p < .05$). Note that each dot in A and B represents the participant-averages, i.e. there is one dot per combination of layers, branch and image set. For an alternative visualization see Appx. Fig. 18.

5 Discussion & Conclusions

Explanation methods such as feature visualizations have been criticized as intuition-driven [24], and it is unclear whether they allow humans to gain a precise understanding of which image features "cause" high activation in a unit. Here, we propose an objective psychophysical task to quantify how well these synthetic images support causal understanding of CNN units. Through a time- and cost-intensive evaluation (based on $24,439$ trials taking more than 81 participant hours including all pilot and reported experiments), we put this widespread intuition to a quantitative test. Our data provides no evidence that humans can predict the effect of an image intervention (occlusion) particularly well when supported with feature visualizations. Instead, human performance is only moderately above a baseline condition where humans are not shown any visualization at all, and similar to that of other visualization methods such as simple dataset samples. Finally, task performance depends on the unit choice, image selections and activation differences between query images. These results add quantitative evidence against the generally-assumed usefulness of feature visualizations for understanding the causes of CNN unit activations.

Our counterfactual-inspired task is the *first* quantitative evaluation of whether feature visualizations support causal understanding of unit activations, but it is certainly not the *only* possible way to evaluate causal understanding. For example, our interventions are constrained to occlusions of a fixed size and shape, imposing an upper limit on the precision with which the occlusions can cover the part of the image that is most responsible for driving a unit's activation. Future work could explore more complex intervention techniques, extend our study to more units of InceptionV1 as well as to different networks, and investigate additional visualization methods. Thanks to its between-participant design, new conditions can be added to the data without the requirement to re-run already collected trials.

Taken together, our empirical results show that the widely used visualization method by Olah et al. [36] is more limited in its ability to convey causal understanding of CNN activations than previously assumed. This reinforces the importance of testing falsifiable hypotheses in the field of interpretable artificial intelligence [24]. Feature visualizations certainly have an important place within the fields of interpretability and explainability, and their importance is likely to grow further with increased societal applications of machine learning. It is therefore important to develop an understanding of what we can — and cannot — expect from current visualization methods. We hope that our task will serve as a challenging test case to steer further development of feature visualizations.

References

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

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

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

[4] Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021.

[5] 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.

[6] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 258–262, 2019.

[7] 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.

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

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

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848.

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

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

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

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

[15] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2950–2958. IEEE, 2019. doi: 10.1109/ICCV.2019.00304.

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

[17] Robert Geirhos, Kristof Meding, and Felix A. Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In Hugo Larochelle,

Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

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

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

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

[21] Peter Green and Catriona J MacLeod. Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498, 2016.

[22] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491.

[23] 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.

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

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

[26] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling, 2021.

[27] Ana Lucic, Hinda Haned, and Maarten de Rijke. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 90–98, 2020.

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

[29] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5188–5196. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299155.

[30] Sina Mohseni, Jeremy E Block, and Eric Ragan. Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In *26th International Conference on Intelligent User Interfaces*, pages 22–31, 2021.

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

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

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

[34] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pages 427–436. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298640.

[35] Anh Mai Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3387–3395, 2016.

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

[37] 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.

[38] 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.

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

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

[41] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

[43] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

[44] Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? *arXiv preprint arXiv:2012.13354*, 2020.

[45] 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.

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

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

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

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

[50] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

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

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

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

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

[55] Xi Ye, Rohan Nair, and Greg Durrett. Evaluating explanations for reading comprehension with realistic counterfactuals. *arXiv preprint arXiv:2104.04515*, 2021.

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

[57] Mohammad Nokhbeh Zaeem and Majid Komeili. Cause and effect: Concept-based explanation of neural networks. *arXiv preprint arXiv:2105.07033*, 2021.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In the abstract and introduction, we clearly describe our motivation and assumptions as well as the scope of our experiment and our conclusions.
   (b) Did you describe the limitations of your work? [Yes] While we already clearly describe the scope of our experiment in the introduction (e.g. testing the feature visualization method by Olah et al. [36]), we discuss caveats of our experiment in the second paragraph of Sec. 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We do not see any potential negative societal impacts of our work, but instead quite the opposite: Our study contributes to better understanding the level of informativeness of interpretability methods for CNNs. This is more broadly discussed in the third paragraph of Sec. 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We attach our code in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Sec. 3 of the main text, we provide sufficient information about the stimulus generation and experiment structure such that a reader can follow the general idea of our project. In Appx. Sec. A.1, we add more detailed information. Finally, our attached code contains all the used random seeds and other details.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] As stated in the first paragraph of Sec. 4, errors in all figures denote two standard errors of the participant-mean over network units and image sets. In Fig. 5 and Fig. 7, we decided not to plot error bars because on the one hand, these would make the plots even more cluttered, and on the other hand, these averages are taken over fewer dimensions (respectively only over subject and image sets or only over subjects) which would make the errors very large. In an effort to present our results clearly, we precisely describe what one value stands for in these figure captions. Finally, in 4A, there are no error bars as baselines give exactly one value per image.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] For computing resources, see Appx. Sec. A.1.5; for online crowdsourcing resources, see Sec. 5 with sums over all experiments and pilots, as well as Appx. Sec. A.1.6 with numbers for the counterfactual-inspired task and Appx. Sec. A.2 with numbers for the replication experiment.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Whenever we refer to other researcher's methods, we cite them, e.g. Olah et al. [36]. Further, in the one case that we use an image from a search engine, we provide the URL (see Appx. Sec. A.1.4.
   (b) Did you mention the license of the assets? [Yes] For the query image in the instruction trial, we restate the website's statement regarding the usability rights.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appx. Sec. A.1.6 as well as the instruction screenshots in Fig. 8 and Fig. 9.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appx. Sec. A.1.6.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appx. Sec. A.1.6 as well as the instruction screenshots in Fig. 8 and Fig. 9.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Our experiments did not represent any larger risk than normal computer use. For pure psychophysical experiments with non-offensive stimuli, a choice task and mouse clicks, we did not consider sending a request to our IRB.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] see Sec. A.1.6