The effectiveness of feature attribution methods and its correlation with automatic evaluation scores

Giang Nguyen[*]
Auburn University
nguyengiangbkhn@gmail.com

Daeyoung Kim
KAIST
kimd@kaist.ac.kr

Anh Nguyen[*]
Auburn University
anh.ng8@gmail.com

Abstract

Explaining the decisions of an Artificial Intelligence (AI) model is increasingly critical in many real-world, high-stake applications. Hundreds of papers have either proposed new feature attribution methods, discussed or harnessed these tools in their work. However, despite humans being the target end-users, most attribution methods were only evaluated on proxy automatic-evaluation metrics [60, 78, 80]. In this paper, we conduct the first user study to measure attribution map effectiveness in assisting humans in ImageNet classification and Stanford Dogs fine-grained classification, and when an image is natural or adversarial (i.e. contains adversarial perturbations). Overall, feature attribution is surprisingly not more effective than showing humans nearest training-set examples. On a harder task of fine-grained dog categorization, presenting attribution maps to humans does not help, but instead hurts the performance of human-AI teams compared to AI alone. Importantly, we found automatic attribution-map evaluation measures to correlate poorly with the actual human-AI team performance. Our findings encourage the community to rigorously test their methods on the downstream human-in-the-loop applications and to rethink the existing evaluation metrics.

1 Introduction

Why did a computer vision system suspect that a person had breast cancer [77], or was an US capitol rioter [2], or a shoplifter [3]? The explanations for such high-stake predictions made by existing Artificial Intelligence (AI) agents can impact human lives in various aspects, from social [22], to scientific [54], and legal [3, 23, 31]. An image classifier's decisions can be explained via an *attribution map* (AM) [13], i.e. a heatmap that highlights the input pixels that are important for or against a predicted label. AMs can be useful in localizing tumors in x-rays [61], reflecting biases [41] and memory [56] of image classifiers, or teaching humans to distinguish butterflies [48].

Via AMs, part of the thought process of AIs is now transparent to humans, enabling human-AI collaboration opportunities. Can machines and humans work together to improve the accuracy of humans in image classification? This question remains largely unknown although, in other non-image domains, AMs have been found useful to humans in the downstream tasks [12, 39, 47, 66, 66].

Furthermore, it is also unknown whether any existing automatic evaluation scores may predict such effectiveness of AMs to humans, the target user of AMs. Most AMs were often only evaluated on

---

[*]Corresponding authors. Work done when GN was at KAIST.
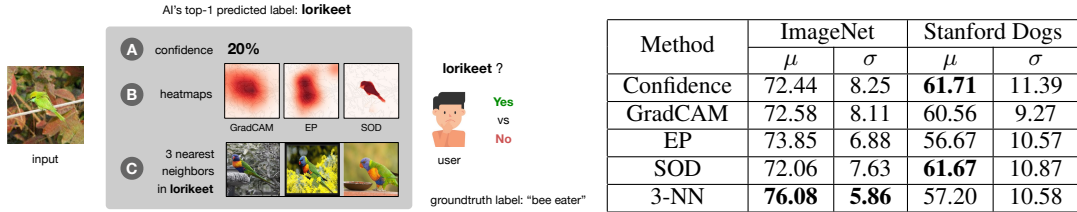
(a) Given an input image, its top-1 predicted label (here, "lorikeet") and confidence score (A), we asked the user to decide Yes or No whether the predicted label is accurate (here, the correct answer is No)—i.e. the accuracy of human-AI teams *without* visual explanations. We also compared this baseline with the treatments where *one* attribution map (B) or a set of three nearest neighbors (C) is also provided to the user (in addition to the confidence score).

Method	ImageNet		Stanford Dogs	
	μ	σ	μ	σ
Confidence	72.44	8.25	**61.71**	11.39
GradCAM	72.58	8.11	60.56	9.27
EP	73.85	6.88	56.67	10.57
SOD	72.06	7.63	**61.67**	10.87
3-NN	**76.08**	**5.86**	57.20	10.58

(b) Users of 3-NN outperform all other users including those using attribution maps on ImageNet classification. However, on fine-grained dog classification (Stanford Dogs), visual explanations (GradCAM, EP, 3-NN) surprisingly did not help but hurt the human performance. Results from testing humans on both real and adversarial images.

Figure 1: Overview of the methods compared in our user-study and its results.

proxy evaluation metrics such as pointing game [78], object localization [80], or deletion [60], which may not necessarily correlate with human-AI team accuracy on a downstream task.

In this paper, we conduct the first user-study to measure AM effectiveness in assisting humans on the 120-class Stanford Dog [38] fine-grained classification and 1000-class ImageNet [65] classification, which is the task that most attribution methods were designed for. We asked 320 lay and 11 expert users to decide whether machine decisions are correct after observing an input image, top-1 classification outputs, and explanations (Fig. 1a). This task captures two related quantities: (1) human accuracy on image categorization, here simplified to binary classification; and (2) model debugging, i.e. understanding whether AI is correct or wrong. For a fuller understanding of the effectiveness of AMs, we tested users on both real and adversarial images [75]. Our main findings include:[2]

1. AMs are, surprisingly, not more effective than nearest-neighbors (here, 3-NN) in improving human-AI team performance on both ImageNet and fine-grained dog classification (Sec. 3.2).

2. On fine-grained dog classification, a harder task for a human-AI team than 1000-class ImageNet, presenting AMs to humans interestingly does not help but instead hurts the performance of human-AI teams, compared to AI alone without humans (Fig. 2b).

3. On adversarial Stanford Dog images which are very hard for humans to label, presenting confidence scores only to humans is the most effective method compared to all other explanations tested including AMs and 3-NN examples (Sec. 3.2).

4. Despite being commonly used in the literature, some automatic evaluation metrics—here, Pointing Game [78], localization errors [80], and IoU [80]—correlate poorly with the actual human-AI team performance (Sec. 3.5).

5. According to a study on 11 AI-expert users, 3-NN is significantly more useful than Grad-CAM [67], a state-of-the-art attribution method for ImageNet classification (Sec. 3.6).

2 Methods

2.1 Image classification tasks

To evaluate human-AI team performance on image classification, for each image, we also presented to a user: (1) an AI's top-1 predicted label, (2) its confidence score, and, optionally (3) a visual explanation that is generated in attempt to explain the predicted label. The user was then asked to decide whether AI's predicted labels are correct or not (see Fig. 1a). We performed our experiments on two image datasets of ImageNet and Stanford Dogs, which are of increasing difficulty.

---

[2]Code and data are available at `http://github.com/anguyen8/effectiveness-attribution-maps`

ImageNet We tested human-AI teams on ImageNet [65]—an image classification task that most attribution methods were tested on or designed for [7, 60, 67, 80]. ImageNet is a ~1.3M-image dataset spanning across 1000 diverse categories from natural to man-made entities [65].

Stanford Dogs To test whether our findings generalize to a fine-grained classification task that is harder to human-AI teams than ImageNet, we repeated the user-study on the 120-class Stanford Dogs [38] dataset, a subset of ImageNet. This dataset is challenging due to the similarity between dog species and the large intra-class variation [38]. Compared to ImageNet, Stanford Dogs is expected to be harder to human-AI teams because lay users often have significantly less prior knowledge about fine-grained dog breeds compared to a wide variety of everyday objects in ImageNet.

2.2 Image classifiers

We took ResNet-34 [34] pretrained on ImageNet (73.31% top-1 accuracy) from torchvision [50] as the target classifier (i.e., the AI) for both ImageNet and Stanford Dogs classification tasks because the 1000-class ImageNet that the model was pretrained on includes 120 Stanford Dogs classes. The visual explanations in this paper were generated to explain this model's predicted labels. We chose ResNet-34 because ResNets were widely used in feature attribution research [7, 28, 45, 60, 67].

2.3 Visual explanation methods

To understand the causal effect of adding humans in the loop, we compared the performance of a standard AI-only system (i.e. no humans involved) and human-AI teams. In each team, besides an input image and a corresponding top-1 label, humans are also presented with the corresponding confidence score from the classifier, and a visual explanation (i.e. one heatmap or three 3-NN images in; see examples from our user-study in Sec. A13).

AI-only A common way to automating the process of accepting or rejecting a predicted label is via confidence-score thresholding [14]. That is, a top-1 predicted label is accepted if its associated confidence score is $\geq T$, a threshold. We found the optimal confidence threshold T^* that produces the highest accepting accuracy on the entire validation set by sweeping across threshold values $\in \{0.05, 0.10, ..., 0.95\}$, i.e. at a 0.05 increment. The optimal T^* values are 0.55 for ImageNet and 0.50 for Stanford Dogs (details in Sec. A6).

Confidence scores only To understand the impact of visual explanations on human decisions, we also used a baseline where users are asked to make decisions given no explanations (i.e. given only the input image, its top-1 predicted label and confidence score). To our best knowledge, this baseline has not been studied in computer vision, but has been shown useful in some non-image domains for improving human-AI team accuracy or user's trust [12, 79].

GradCAM and Extremal Perturbation (EP) We chose GradCAM [67] and Extremal Perturbation (EP) [28] as two representatives for state-of-the-art attribution methods (Fig. 1a). Representing for the class of white-box, *gradient*-based methods [16, 63, 80], GradCAM relies on gradients and the activation map at the last conv layer of ResNet-34 to compute a heatmap. GradCAM passed a weight-randomization sanity check [6] and often obtained competitive scores in proxy evaluation metrics (see Table 1 in [28] and Table 1&2 in [25]).

In contrast, EP is a representative for *perturbation*-based methods [7, 20, 29]. EP searches for a set of $a\%$ of input pixels that maximizes the target-class confidence score. We followed the authors' best hyperparameters—summing over four binary masks generated using $a \in \{2.5, 5, 10, 20\}$ and a Gaussian smoothing kernel with an std equal to 9% of the shorter side of the image (see the code [1]). We used the TorchRay package [5] to generate GradCAM and EP attribution maps.

We initially also tested vanilla Gradient [70], Integrated Gradient (IG) [74], and SHAP [46] methods in a pilot study but discarded them from our study since their AMs are too noisy to be useful to users in a pilot study.

Salient object detection (SOD) To assess the need for heatmaps to explain a *specific classifer*'s decisions, we also considered a classifier-agnostic heatmap baseline. That is, we used a pre-trained state-of-the-art salient-object detection (SOD) method called PoolNet [43], which uses a ResNet-50

backbone pre-trained on ImageNet. PoolNet was trained to output a heatmap that highlights the *salient* object in an input image (Fig. 3)— a process that does *not* take our image classifier into account. Thus, SOD serves as a strong saliency baseline for GradCAM and EP.

3-NN To further understand the pros and cons of AMs, we compared them to a representative, prototype-based explanation method (e.g. [18, 53, 55]). That is, for a given input image $\mathbf{x}$ and a predicted label $\mathbf{y}$ (e.g., "lorikeet"), we show the top-3 nearest images to $\mathbf{x}$ by retrieving them from the same ImageNet training-set class $\mathbf{y}$ (Fig. 1a). To compute the distance between two images, we used the L_2 distance in the feature space of the last conv layer of ResNet-34 classifier (i.e. layer4 per PyTorch definition [4] or conv5_x in He et al. [34]). We chose layer4 as our pilot study found it to be the most effective among the four main conv layers (i.e. layer1 to layer4) of ResNet-34.

While k-NN has a wide spectrum of applications in machine learning and computer vision [68], the effectiveness of human-AI collaboration using prototype-based explanations has been rarely studied [64]. To the best of our knowledge, we provided the first user study that evaluates the effectiveness of post-hoc, nearest-neighbor explanations (here, 3-NN) on human-AI team performance.

2.4   User-study design

2.4.1   Participants

Our user-study experiments were designed and hosted on Gorilla [11]. We recruited lay participants via Prolific [57] (at \$10.2/hr), which is known for a high-quality user base [59]. Each Prolific participant self-identified that English is their first language, which is the only demographic filter we used to select users on Prolific. Participants come from diverse geographic locations and backgrounds (see Sec. A9). Over the course of two small pilot studies and one main, large-scale study, we collected in total over 466 complete submissions (each per user) after discarding incomplete ones.

In the main study, after filtering out submissions by validation scores (described in Sec. 2.4.2), we had 161 and 159 *qualified* submissions for our ImageNet and Dogs experiments, respectively. Each of the 5 methods (i.e. except for AI-only) was experimented by at least 30 users, and each (image, explanation) pair was seen by at least two users (statistics of users and trials in Sec. A8).

2.4.2   Tasks

Our Gorilla study contains three main sets of screens: (1) Introduction—where each user is introduced to the study and relevant rules (details in Sec. A11); (2) Training; and (3) Test. Each user is randomly assigned a set of test images and *only one*[3] explanation method (e.g. GradCAM heatmaps) to work with during the entire study.

Training To familiarize with the image classification task (either ImageNet or Stanford Dogs), users were given five practice questions. After answering each question, they were shown feedback with groundtruth answers. In each training screen, we also described each component in a screen via annotations (see example screens in Sec. A11.2).

Validation and Test After training, each user was asked to answer 40 Yes/No questions in total. Before each question, a user was provided with a short WordNet [51] definition of the predicted label and three random training-set images (see an example in Fig. A5) correctly-classified into the predicted class with a confidence score ≥ 0.9. To control the quality of user responses, out of 40 trials, we used 10 trials as validation cases where we carefully chose the input images such that we expected participants who followed our instructions to answer correctly (details in Sec. A11.3). We excluded those submissions that had below 10/10 and 8/10 accuracy on the 10 validation trials from the ImageNet and Dogs experiments, respectively. For the remaining (i.e., qualified) submissions, we used the results of their 30 non-validation trials in our study.

---

[3]Showing multiple explanation types to the same user complicates the user-training procedure and also makes the task harder to users due to context switching.

2.4.3 Images

We wish to understand the effectiveness of visual explanations when AIs are correct vs. wrong, and when AIs face real vs. adversarial examples [75]. For both ImageNet and Dogs experiments, we used the ResNet-34 classifier to sample three types of images: (1) correctly-classified, real images; (2) misclassified, real images; and (3) adversarial images (i.e. also misclassified). In total, we used 3 types × 150 images = 450 images per dataset. Each image was then used to generate model predictions and explanations for comparing the 6 methods described in Sec. 2.3.

Filtering From the 50K-image ImageNet validation set, we sampled images for both ImageNet and Dogs experiments. To minimize the impact of low-quality images to users' performance, we removed all 900 grayscale images and 897 images that have either width or height smaller than 224 px, leaving 48,203 and 5,881 images available for use in our ImageNet and Dogs experiments, respectively. For Dogs, we further excluded all 71 dog images mislabeled by ResNet-34 into non-dog categories (examples in Sec. A12) because they can trivialize explanations, yielding 5,810 Dogs images available for selection for the study.

Selecting natural images To understand human-AI team performance at varying levels of difficulty for humans, we randomly divided images from the pool of filtered, real images into three sets: Easy (E), Medium (M), and Hard (H).

Hard images are those correctly labeled by the classifier with a low confidence score (i.e., $\in [0, 0.3)$) and mislabeled with high confidence (i.e., $\in [0.8, 1.0]$). Vice versa, the Easy set contains those correctly labeled with high confidence and mislabeled with low confidence. The Medium set contains both correctly and incorrectly labeled images with a confidence score $\in [0.4, 0.6]$ i.e. when the AI is unsure (see confidence-score distributions in Sec. A7). In each set (E/M/H), we sampled 50 images correctly-labeled and 50 mislabeled by the model. In sum, per dataset, there are 300 natural images divided evenly into 6 *controlled* bins (see Fig. A1 for the ratios of these bins in the original datasets).

Generating adversarial images After the filtering above, we took the remaining real images to generate adversarial examples via Foolbox [62] for the ResNet-34 classifier using the Project Gradient Descent (PGD) framework [49] with an L_∞ bound $\epsilon = 16/255$ for 40 steps, each of size of $1/30$. We chose this setup because at weaker attack settings, most adversarial images became *correctly* classified after being saved as a JPEG file [44], defeating the purpose of causing AIs to misbehave. Here, for each dataset, we randomly sampled 150 adversarial examples (e.g., the input image in Fig. 1a) that are *misclassified* at the time presented to users in the JPEG format and often contain so small artifacts that we assume to not bias human decisions. Following the natural-image sampling, we also divided the 150 adversarial images into three sets (E/M/H), each containing 50 images.

2.5 Automatic evaluation metrics for attribution maps

Many attribution methods have been published; however, most AMs were not tested on end-users but instead only assessed via proxy evaluation metrics. We aim to measure the correlation between three common metrics—Pointing Game [78], Intersection over Union (IoU) [80], and weakly-supervised localization (WSL) [80]—with the actual human-AI team performance in our user study.

All three metrics are based on the assumption that an AM for explaining a predicted label **y** should highlight an image region that overlaps with a human-annotated bounding box (BB) for that category **y**. For each of 300 real, correctly-classified images from each dataset (described in Sec. 2.4.3), we obtained its human-annotated BB from ILSVRC 2012 [65] for using in the three metrics.

Pointing game [78] is a common metrics often reported in the literature (e.g. [24, 28, 60, 63, 67, 76]). For an input image, a generated AM is considered a "correct" explanation (i.e. a *hit*) if its highest-intensity point lies inside the human-annotated BB. Otherwise, it is a *miss*. The accuracy for an explanation method is computed by averaging over all images. We used the TorchRay implementation of Pointing Game [5] and its default hyperparameters (tolerance = 15).

Intersection over Union The idea is to compute the agreement in Intersection over Union (IoU) [80] between a human-annotated BB and a binarized AM. A heatmap was binarized at the method's optimal threshold α, which was found by sweeping across values $\in \{0.05x \mid x \in \mathbb{N} \text{ and } 0 < x < 20\}$.

Weakly-supervised localization (WSL) is based on IoU scores [80]. WSL counts a heatmap correct if its binarized version has a BB that overlaps with the human-labeled BB at an IoU > 0.5.

WSL is also commonly used e.g. [7, 24, 67]. For both GradCAM and EP, we found the binarization threshold $\alpha = 0.05$ corresponds to their best WSL scores on 150 correctly-classified images (i.e., excluding mislabeled images because human-annotated BBs are for the groundtruth labels).

3 Results

3.1 3-NN is more effective than AMs on ImageNet but both hurt user accuracy on Dogs

We measure the per-user accuracy on ImageNet classification over our controlled image sets (correctly-labeled, incorrectly-labeled, and adversarial images; see Sec. 2.4.3), i.e. on both natural and adversarial images. We repeated the same study for Stanford Dogs (see Table 1b).

ImageNet We found that 3-NN is significantly better than SOD and GradCAM (Mann Whitney U-test; $p < 0.035$). The differences among feature attribution methods (GradCAM, SOD, EP) are *not* statistically significant. In terms of mean and standard deviation of per-user accuracy, users of 3-NN outperformed all other users by a large margin of ~3% (Table 1b; 76.08% and 5.86%).

Stanford Dogs We found that users of 3-NN and EP performed statistically significantly worse than the users with no explanations (Table 1b; Confidence) and SOD users (Mann Whitney U-test; $p < 0.024$). There are no significant differences among other pairs of methods. That is, interestingly, on fine-grained Stanford Dog classification where an image may be real or adversarial, visual explanations (i.e. both 3-NN and AMs) *hurt* user classification accuracy.

3.2 On natural images, how effective are attribution maps in human-AI team image classification?

We wish to understand the effectiveness of AMs (GradCAM and EP) compared to four baselines (AI-only, Confidence, SOD, and 3-NN) in image classification by human-AI teams. Here, we compare 6 methods on 2 *natural*-image sets: ImageNet and Stanford Dogs (i.e. no adversarial images).

Experiment Because a given image can be mapped to one of the 6 controlled bins as described in Sec. 2.4.3, for each of the 6 methods, we computed its human-AI team accuracy for each of the 6 bins where each bin has exactly 50 images (Sec. A1 reports per-bin accuracy scores). To estimate the overall accuracy of a method on the *original* ImageNet and Dogs dataset, we computed the weighted sum of its per-bin accuracy scores where the weights are the frequencies of images appearing in each bin in practice. For example, 70.93% of the Dogs images are correctly-classified with a high confidence score (Fig. A1; Easy Correct).

ImageNet results On ImageNet, human-AI teams where humans use heatmaps (GradCAM, EP, or SOD) and confidence scores outperformed AI-only by an absolute gain of ~6–8% in accuracy (Fig. 2a; 80.79% vs. 88.77%). That is, when teaming up, **humans and AI together achieve a better performance than AI alone**. Interestingly, only half of such improvement can be attributed to the heatmap explanations (Fig. 2a; 84.79% vs. 88.77%). That is, users when presented with the input image and the classifier's top-1 label and confidence score already obtained +4% boost over AI alone (84.79% vs. 80.79%).

Dogs results Interestingly, the trend did not carry over to fine-grained dog classification. On average, humans when presented with (1) confidence scores only or (2) confidence scores with heatmaps all *underperformed* the accuracy of AI-only (Fig. 2b; 81.14% vs. 76.45%). An explanation is that ImageNet contains ~50% of man-made entities, which contain many everyday objects that users are familiar with. Therefore, the human-AI teaming led to a substantial boost on ImageNet. In contrast, most lay users are not dog experts and therefore do not have the prior knowledge necessary to help them in dog identification, resulting in even worse performance than a trained AI (Fig. 2b; 81.14% vs. 76.45%). Interestingly, when providing users with nearest neighbors, human-AI teams with 3-NN outperformed all other methods (Fig. 2b; 82.88%) including the AI-only.

Both datasets On both ImageNet and Dog distributions, **3-NN is among the most effective**. On average over 6 controlled bins, 3-NN also outperformed all other methods by a clear margin of 2.71% (Sec. A2 & A4). Interestingly, SOD users tend to reject more often than other users (Sec. A10), inadvertently causing a high human-AI team accuracy on AI-misclassified images (Sec. A14.5).
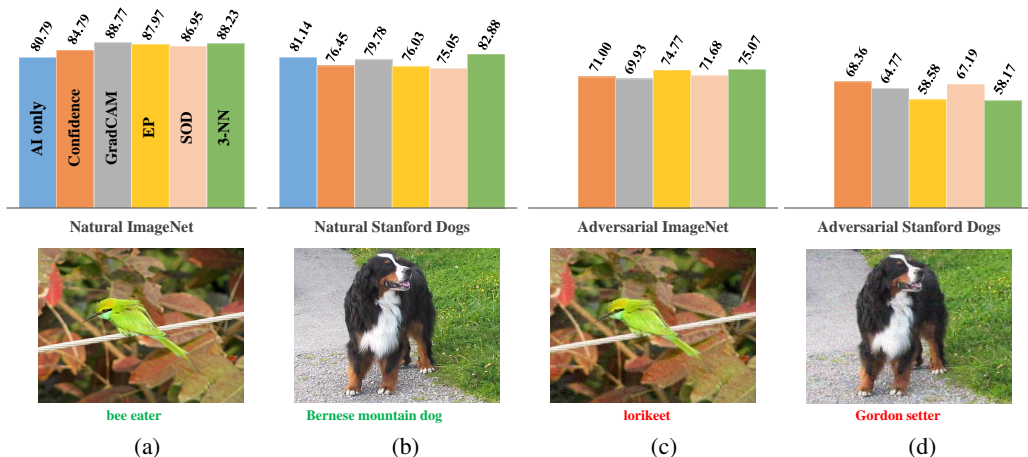
6

Figure 2: 3-NN is consistently among the most effective in improving human-AI team accuracy (%) on natural ImageNet images (a), natural Dog images (b), and adversarial ImageNet images (c). On the challenging adversarial Dogs images (d), using confidence scores only helps humans the most in detecting AI's mistakes compared to using confidence scores and one visual explanation. Below each sample dataset image is the top-1 predicted label (which was correct or wrong) from the classifier.

3.3 On adversarial images, how effective are attribution maps in human-AI team image classification?

As adversarial examples are posing a big threat to AI-only systems [8, 9, 52, 58], here, we are interested in understanding how effective explanations are in improving human-AI team performance over AI-only, which is assumed to be 0% accurate under adversarial attacks. That is, on ImageNet and Dogs, we compare the accuracy of human-AI teams on 150 adversarial examples, which all caused the classifier to misclassify. Note that a user was given a random mix of natural and adversarial test images and was not informed whether an input image is adversarial or not.

Adversarial ImageNet On both natural and adversarial ImageNet, AMs are on-par or more effective than showing confidence scores alone (Fig. 2c). Furthermore, the effect of 3-NN is a consistent +4% gain compared to Confidence (Fig. 2a & c; Confidence vs. 3-NN). Aligned with the results on natural ImageNet and Dogs, 3-NN remains the best method on adversarial ImageNet (Fig. 2c; 75.07%). See Sec. A14.4 for adversarial images that were only correctly-rejected by 3-NN users but not others.

Adversarial Dogs Interestingly, on adversarial Dogs, adding an explanation (either 3-NN or a heatmap) tend to cause users to agree with the model's incorrect decisions (Fig. 2d). That is, heatmaps often only highlight a coarse body region of a dog without pinpointing an exact feature that might be explanatory. Similarly, 3-NN often shows examples of an almost identical breed to the groundtruth (e.g. mountain dog vs. Gordon setter), which is hard for lay-users to tell apart (see qualitative examples in Sec. A14.6).

## 3.4 On ImageNet, why is 3-NN more effective than attribution maps?

Analyzing the breakdown accuracy of 3-NN in each controlled set of Easy, Medium, and Hard, we found 3-NN to be the most effective method in the Easy and Hard categories of **ImageNet** (Table A3).

Easy When AI mislabels with low confidence, 3-NN often presents contrast evidence showing that the predicted label is incorrect, i.e. the nearest examples from the predicted class is distinct from the input image (Fig. 1a; lorikeets have a distinct blue-orange pattern not found in bee eaters). The same explanation is our leading hypothesis for 3-NN's effectiveness on adversarial Easy, ImageNet images. More examples in Sec. A14.3.

Hard Hard images often contain occlusions (Figs. A9 & A12), unusual instances of common objects (Fig. A13), or could reasonably be in multiple categories (Figs. A11 & A10). When the classifier is *correct* but with low confidence, we found 3-NN to be helpful in providing extra evidence for users to

confirm AI's decisions (Fig. 3). In contrast, heatmaps often only highlight the main object regardless of whether AI is correct or not (Fig. 1a; GradCAM). More examples are in Sec. A14.1.

Medium Interestingly, 3-NN was the best method on the Easy and Hard set, but not on the Medium (Table A3). Upon a closer look, we surprisingly found ∼63% of misclassified images by 3-NN human-AI teams to have debatable groundtruth ImageNet labels (see Sec. A14.2 for failure cases of 3-NN).
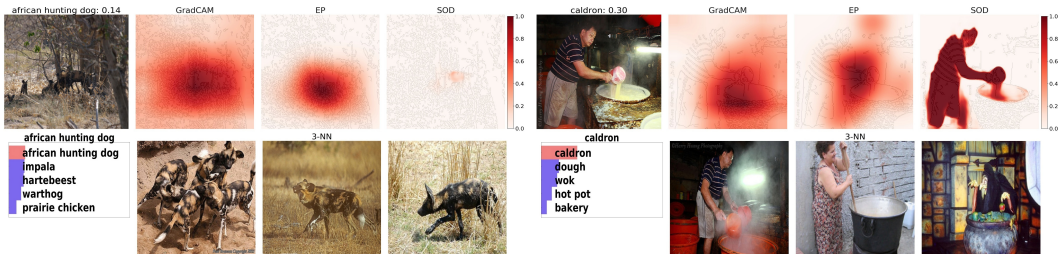
(a) "african hunting dog": dog-like mammals of Africa   (b) "caldron": a very large pot used for boiling

Figure 3: Hard images that were only correctly accepted by 3-NN users but not other users of GradCAM, EP, or SOD. Despite the animals in the input image are partially occluded, 3-NN provided closed-up examples of african hunting dogs, enabling users to correctly decide the label (a). Choosing a single label for a scene of multiple objects is challenging. However, 3-NN was able to retrieve a nearest example showing a very similar scene enabling users to accept AI's correct decisions (b). See full-res images in Figs. A9 & A10.

3.5   How do automatic evaluation scores correlate with human-AI team performance?

IoU [80], WSL [80], and Pointing Game [78] are three attribution-map evaluation metrics commonly used in the literature, e.g. in [7, 28, 29, 60, 67]. Here, we measure the correlation between these scores and the actual human-AI team accuracy to assess how high-performance on such benchmarks translate into the real performance on downstream tasks.

Experiment We took the EP and GradCAM heatmaps of the 150 real images that were correctly-classified (see Sec. 2.4.3) from each dataset and computed their IoU, WSL, and Pointing Game scores[4] (see Sec. 2.5). We computed the Pearson correlation between these scores and the human-AI team accuracy obtained in the previous human-study (Sec. 2.4).

Results While EP and GradCAM are state-of-the-art methods under Pointing Game [28] or WSL [29], we surprisingly found the accuracy of users when using these AMs to correlate poorly with the IoU, WSL, and Pointing Game scores of heatmaps. Only in the case of GradCAM heatmaps for ImageNet (Fig. 4a), the evaluation metrics showed a small positive correlation with human-AI team accuracy ($r = 0.22$ for IoU; $r = 0.15$ for WSL; and $r = 0.21$ for Pointing Game). In all other cases, i.e. GradCAM on Dogs; EP on ImageNet and Dogs (Fig. 4b–c), the correlation is negligible ($-0.12 \leq r \leq 0.07$). That is, the **pointing or localization accuracy of feature attribution methods do not necessarily reflect their effectiveness in helping users make correct decisions** in image classification.

3.6   Machine learning experts found Nearest-Neighbors more effective than GradCAM

We have found in our previous study on lay-users that AMs can be useful to human-AI teams, but not more than a simple 3-NN in most cases (Secs. 3.1–3.4). Here, we aim to evaluate whether such conclusions carry over to the case of machine learning (ML) expert-users who are familiar with feature attribution methods, nearest-neighbors, and ImageNet. Because ML experts have a deeper understanding into these algorithms than lay-users, our expert-study aims to measure the utility of these two visual explanation techniques to ML researchers and practitioners.

---

[4]Note that WSL and Pointing Game scores are binary while IoU scores are real-valued.

8

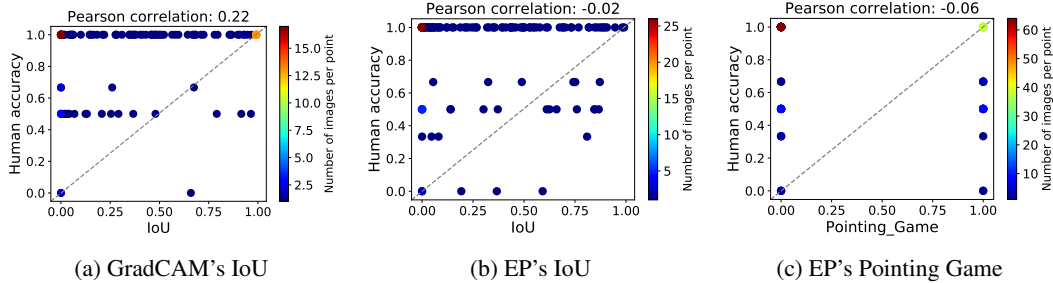|		
(a) GradCAM's IoU	(b) EP's IoU	(c) EP's Pointing Game

Figure 4: The ImageNet localization performance (here, under IoU vs. human-annotated bounding boxes) of GradCAM [67] and EP [28] attribution maps poorly correlate with the human-AI team accuracy (y-axis) when users use these heatmaps in image classification. Humans often can make correct decisions despite that heatmaps poorly localize the object (see the range 0.0–0.2 on x-axis).

As GradCAM was the best attribution method in the lay-user ImageNet study (Fig. 2), we chose GradCAM as the representative for attribution maps and compare it with 3-NN on human-AI team image classification on ImageNet.

Experiment We repeated the same lay-user study but on a set of 11 expert users who are Ph.D. students and postdoctoral researchers in ML and computer vision and from a range of academic institutions. We recruited five GradCAM users and six 3-NN users who are very familiar with feature attribution and nearest-neighbor algorithms, respectively. Similar to the lay-user study, each expert is presented with a set of randomly-chosen 30 images which include both natural and adversarial images.

Results The experts working with 3-NN performed substantially better than the GradCAM experts (Table 1; mean accuracy of 76.67% vs. 68.00%). 3-NN is consistently more effective on both natural and adversarial image subsets. Interestingly, the performance of GradCAM users also vary $3\times$ more (standard deviation of 8.69% vs. 2.98% in accuracy). Aligned with the lay-user study, here, we found 3-NN to be more effective than AMs in improving human-AI team performance where the users are domain experts familiar with the mechanics of how explanations were generated.

Table 1: 3-NN is far more effective than GradCAM in human-AI team image classification of both natural and adversarial images. See the mean (μ) and std (σ) in per-user accuracy over all images.

	Users	Avg. validation accuracy	Natural		Adversarial		μ	σ
			Accuracy	Trials	Accuracy	Trials		
GradCAM	5	9.80/10	67.31	70/104	69.57	32/46	68.00	8.69
3-NN	6	9.83/10	**78.45**	91/116	**73.44**	47/64	**76.67**	**2.98**

4 Related Work

Attribution maps improved human accuracy in non-image domains Our work is motivated by prior findings showing that AMs helped humans make more accurate decisions on non-image classification tasks including predicting book categories from text [66], movie-review sentiment analysis [12, 66], predicting hypoxemia-risk from medical tabular data [47], or detecting fake text reviews [39, 40]. In contrast, on house-price prediction, AMs showed marginal utility to users and even hurt their performance in some cases [21].

Evaluating confidence scores on humans AI confidence scores have been found to improve user's trust on AI's decisions [79] and be effective in human-AI team prediction accuracy on several NLP tasks [12]. In this work, we do not measure user trust but only the complementary human-AI team performance. To the best of our knowledge, our work is the first to perform such human-evaluation of AI confidence scores for image classification.

Evaluating explanations on humans In NLP, using attribution methods as a word-highlighter has been found to improve user performance in question answering [26]. Bansal et al. [12] found

that such human-AI team performance improvement was because the explanations tend to make users more likely to accept AI's decisions regardless of whether the AI is correct or not. We found consistent results that users with explanations tend to accept AI's predicted labels more (see Sec. A10) with the exception of SOD users who tend to reject more (possibly due to the prediction-agnosticity of SOD heatmaps).

In image classification, Chu et al. [19] found that presenting IG heatmaps to users did not significantly improve human-accuracy on predicting age from facial photos. Different from [19], our study tested multiple attribution methods (GradCAM, EP, SOD) and on the ImageNet classification task, which most attribution methods were evaluated on under proxy metrics (e.g. localization).

Shen and Huang [69] showed users GradCAM [67], EP [28], and SmoothGrad [72] heatmaps and measured user-performance in harnessing the heatmaps to identify a label that an AI incorrectly predicts. While they measured the effect of showing all three heatmaps to users, we compared each method separately (GradCAM vs. EP vs. SOD) in a larger study. Similar to [69], Alqaraawi et al. [10] and Folke et al. [27] tested the capability of attribution maps in helping users understand the decision-making process of AI. Our work differs from the above three papers [10, 27, 69] and similar studies in the text domain [15, 33] in that we did not only measure how well users understand AIs, but also the human-AI team accuracy on a standard *downstream* task of ImageNet.

Using AMs to debug models, Adebayo et al. [6] found that humans heavily rely on predicted labels rather than AMs, which is strongly aligned with our observations on the Natural ImageNet (in Sec. 3.2) where 3-NN outperforms AMs.

Adversarial attacks   Concurrently, Folke et al. [27] also perform a related study that nearest-neighbor measures the effectiveness of both AMs and examples in helping users predict when AIs misclassify *natural adversarial* images [37], which we do not study in this paper. Our work largely differs from [27] in both the scale of the study, the type of image, and the experiment setup. To our knowledge, we are the first to evaluate human users on the well-known adversarial examples [49]. Previous work also found that it is easy to minutely change the input image such that the gradient-based AMs radically change [30]. However, none of such prior attempts studied the impact of their attacks to the actual end-users of the heatmaps.

5   Discussion and Conclusion

Limitations   An inherent limitation of our study is that it is not possible to control the amount of prior knowledge that a participant has *before* entering the study. For example, a human with a strong dog expertise may perform better at the fine-grained dog classification. In that case, the utility of explanations is unknown in our study. We attempted to estimate the effect of prior knowledge to human-AI team accuracy by asking each user whether they know a class before each trial. We found prior knowledge to account for ∼1-6% in accuracy (Sec. A5). Due to COVID and the large participant pool, our study was done online; which, however, made it infeasible for us to control various physical factors (e.g. user performing other activities during the experiment) compared to a physical in-lab study. As most prior AM studies, we do consider state-of-the-art methods for training humans on explanations [71, 73] partly because each explanation method may need a different optimal training procedure.

To our knowledge, our work is the first to (1) evaluate human-AI team performance on the common ImageNet classification; (2) assess explanations on adversarial examples; (3) reveal the weak correlation between automatic evaluation metrics (Pointing Game, IoU [80], and WSL [80]) and the actual team performance. Such poor correlation encourages future interpretability research to take humans into their evaluation and to rethink the current automatic metrics [22, 42]. We also showed the first evidence in the literature that a simple 3-NN can outperform existing attribution maps, suggesting a combination of two explanation types may be useful to explore in future work. The superiority of 3-NN also suggests prototypical [17, 32] and visually-grounded explanations [35, 36] may be more effective than heatmaps.

Acknowledgement

References

[1] Torchray/attribution_benchmark.py at master · facebookresearch/torchray. `https://github.com/facebookresearch/TorchRay/blob/master/examples/attribution_benchmark.py`. (Accessed on 04/28/2021).

[2] The fbi's capitol riot investigation used surveillance technology that advocates say threatens civil liberties - the washington post. `https://www.washingtonpost.com/technology/2021/04/02/capitol-siege-arrests-technology-fbi-privacy/`. (Accessed on 04/11/2021).

[3] The new lawsuit that shows facial recognition is officially a civil rights issue | mit technology review. `https://www.technologyreview.com/2021/04/14/1022676/robert-williams-facial-recognition-lawsuit-aclu-detroit-police/`. (Accessed on 04/15/2021).

[4] vision/resnet.py at master · pytorch/vision. `https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py#L183`. (Accessed on 04/28/2021).

[5] facebookresearch/torchray: Understanding deep networks via extremal perturbations and smooth masks. `https://github.com/facebookresearch/TorchRay`. (Accessed on 04/28/2021).

[6] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

[7] Agarwal, C. and Nguyen, A. Explaining image classifiers by removing input features using generative models. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[8] Agarwal, C., Nguyen, A., and Schonfeld, D. Improving robustness to adversarial examples by encouraging discriminative features. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3801–3505. IEEE, 2019.

[9] Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4845–4854, 2019.

[10] Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 275–285, 2020.

[11] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52(1):388–407, 2020.

[12] Bansal, G., Wu, T., Zhu, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. S. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779*, 2020.

[13] Bansal, N., Agarwal, C., and Nguyen, A. Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8673–8683, 2020.

[14] Bendale, A. and Boult, T. E. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

[15] Chandrasekaran, A., Prabhu, V., Yadav, D., Chattopadhyay, P., and Parikh, D. Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*, 2018.

[16] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, 2018.

[17] Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018.

[18] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: Deep learning for interpretable image recognition. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf`.

[19] Chu, E., Roy, D., and Andreas, J. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

[20] Covert, I., Lundberg, S., and Lee, S.-I. Feature removal is a unifying principle for model explanation methods. *arXiv preprint arXiv:2011.03623*, 2020.

[21] Dinu, J., Bigham, J., and Kolter, J. Z. Challenging common interpretability assumptions in feature attribution explanations. *arXiv preprint arXiv:2012.02748*, 2020.

[22] Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[23] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., and Wood, A. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[24] Du, M., Liu, N., Song, Q., and Hu, X. Towards explanation of dnn-based prediction with guided feature inversion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1358–1367, 2018.

[25] Elliott, A., Law, S., and Russell, C. Perturbations on the perceptual ball. *arXiv preprint arXiv:1912.09405*, 2019.

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

[27] Folke, T., Li, Z., Sojitra, R. B., Yang, S. C.-H., and Shafto, P. Explainable ai for natural adversarial images. *arXiv preprint arXiv:2106.09106*, 2021.

[28] Fong, R., Patrick, M., and Vedaldi, A. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2950–2958, 2019.

[29] Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

[30] Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

[31] Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[32] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019.

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

[34] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[35] Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. Generating visual explanations. In *European Conference on Computer Vision*, pp. 3–19. Springer, 2016.

[36] Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. Generating counterfactual explanations with natural language. *arXiv preprint arXiv:1806.09809*, 2018.

[37] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

[38] Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. Novel dataset for fine-grained image categorization: Stanford dogs. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[39] Lai, V. and Tan, C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 29–38, 2019.

[40] Lai, V., Liu, H., and Tan, C. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[41] Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., and Samek, W. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2912–2920, 2016.

[42] Leavitt, M. L. and Morcos, A. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

[43] Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., and Jiang, J. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3917–3926, 2019.

[44] Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., and Wen, W. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 860–868. IEEE, 2019.

[45] Lu, Y., Guo, W., Xing, X., and Stafford Noble, W. Robust decoy-enhanced saliency maps. *arXiv e-prints*, pp. arXiv–2002, 2020.

[46] Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[47] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

[48] Mac Aodha, O., Su, S., Chen, Y., Perona, P., and Yue, Y. Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3820–3828, 2018.

[49] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[50] Marcel, S. and Rodriguez, Y. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010.

[51] Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[52] Munusamy Kabilan, V., Morris, B., and Nguyen, A. Vectordefense: Vectorization as a defense to adversarial examples. *arXiv e-prints*, pp. arXiv–1804, 2018.

[53] Nauta, M., Jutte, A., Provoost, J., and Seifert, C. This looks like that, because... explaining prototypes for interpretable image recognition. *arXiv preprint arXiv:2011.02863*, 2020.

[54] Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '15, pp. 427–436. IEEE, June 2015. doi: 10.1109/CVPR.2015.7298640.

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

[56] Nguyen, G., Chen, S., Jun, T. J., and Kim, D. Explaining how deep neural networks forget by deep visualization. In *International Conference on Pattern Recognition*, pp. 162–173. Springer, 2021.

[57] Palan, S. and Schitter, C. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[58] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.

[59] Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[60] Petsiuk, V., Das, A., and Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[61] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[62] Rauber, J., Brendel, W., and Bethge, M. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. URL http://arxiv.org/abs/1707.04131.

[63] Rebuffi, S.-A., Fong, R., Ji, X., and Vedaldi, A. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8839–8848, 2020.

[64] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021.

[65] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[66] Schmidt, P. and Biessmann, F. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.

[67] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 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.

[68] Shakhnarovich, G., Darrell, T., and Indyk, P. Nearest-neighbor methods in learning and vision. In *Neural Information Processing*, 2005.

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

[70] Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[71] Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., and Krause, A. Near-optimally teaching the crowd to classify. In *International Conference on Machine Learning*, pp. 154–162. PMLR, 2014.

[72] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[73] Su, S., Chen, Y., Mac Aodha, O., Perona, P., and Yue, Y. Interpretable machine teaching via feature feedback. 2017.

[74] Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

[75] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[76] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020.

[77] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Pysarenko, H. T. K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S. G., Heacock, L., Moy, L., Cho, K., and Geras, K. J. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019. doi: 10.1109/TMI.2019.2945514.

[78] Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[79] Zhang, Y., Liao, Q. V., and Bellamy, R. K. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020.

[80] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.