SAM: The Sensitivity of Attribution Methods to Hyperparameters

Naman Bansal*
Auburn University
bnaman50@gmail.com

Chirag Agarwal*
University of Illinois at Chicago
chiragagarwall12@gmail.com

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

Abstract

Attribution methods can provide powerful insights into the reasons for a classifier's decision. We argue that a key desideratum of an explanation method is its robustness to input hyperparameters which are often randomly set or empirically tuned. High sensitivity to arbitrary hyperparameter choices does not only impede reproducibility but also questions the correctness of an explanation and impairs the trust of end-users. In this paper, we provide a thorough empirical study on the sensitivity of existing attribution methods. We found an alarming trend that many methods are highly sensitive to changes in their common hyperparameters e.g. even changing a random seed can yield a different explanation! Interestingly, such sensitivity is not reflected in the average explanation accuracy scores over the dataset as commonly reported in the literature. In addition, explanations generated for robust classifiers (i.e. which are trained to be invariant to pixel-wise perturbations) are surprisingly more robust than those generated for regular classifiers.

1. Introduction

Why did a self-driving car decide to run into a truck [29]? Why is a patient being predicted to have breast cancer [59] or to be a future criminal [2]? The explanations for such predictions made by machine learning (ML) models can impact our lives in many ways, under scientific [53, 37], social [18] or legal [24, 19] aspects.

A popular medium for visually explaining an image classifier's decisions is an *attribution map i.e.* a heatmap that highlights the input pixels that are the evidence for and against the classification outputs [35]. Dozens of attribution methods (Fig. 1) have been proposed [44] and applied to a variety of domains including natural images [35], medical brain scans [25], text [12], videos [50], and speech [14]. Notably, attribution maps have been useful *e.g.* in localizing malignant tumors in a breast x-ray scan [41] or in revealing biases in object recognition models [30, 31]. Yet are these explanations reliable enough to convince medical doctors or judges to accept a life-critical prediction by a machine [32]?

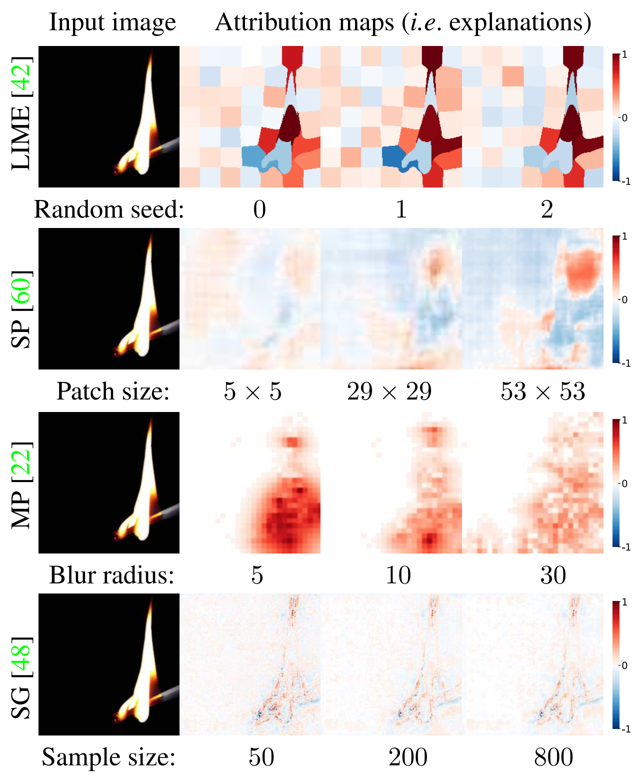


Figure 1: Attribution maps by four methods to explain the same prediction (match stick: 0.535) made by a ResNet-50 classifier to an ImageNet image. In each row, the explanations are generated by running the default settings of a method while varying only one common hyperparameter. All 12 explanations are unique and can be interpreted differently. **LIME:** an explanation changes when one re-runs the algorithm with a different random seed. **SP:** the positive evidence for the fire (top-right red blob) grows together with the patch size. **MP:** attribution maps become more scattered as the Gaussian blur radius increases. **SG:** heatmaps becomes smoother as the number of samples increases.

First, ML techniques often have a set of hyperparame-

*Equal contribution. CA performed this work during his internship at Auburn University.

ters to be tuned empirically and most attribution methods are not an exception. Second, a major cause of the current replication crisis in ML [27] is that many methods, *e.g.* in reinforcement learning, are notoriously sensitive to hyperparameters—a factor which is also often overlooked in the interpretability field. Aside from being faithful, an explanation needs to be reproducible and invariant to arbitrary hyperparameter choices. In this paper, we studied an important question: *How sensitive are attribution maps to their hyperparameters?* on 7 well-known attribution methods and found that:[1]

1. Gradient heatmaps, for *robust* image classifiers *i.e.* models trained to ignore adversarial pixel-wise noise [21], exhibit visible structures (Fig. 3) in stark contrast to the noisy, uninterpretable gradient images for regular classifiers reported in prior work [48] (Sec. 4.1).
2. The gradient images from a robust and a regular classifier are different but would appear $\sim$1.5$\times$ more similar, potentially causing misinterpretation, under several prior methods that attempted to de-noise the original explanations [49, 48] (Sec. 4.2).
3. For many attribution methods [42, 22, 60], their output heatmaps can change dramatically (Fig. 1) when a common hyperparameter changes (Sec. 4.3). This sensitivity of an individual explanation also translates into the sensitivity of its accuracy scores (Sec. 4.5).
4. Some hyperparameters cause up to 10$\times$ more variation in the accuracy of explanations than others (Sec. 4.6).
5. Explanations for robust classifiers are not only more invariant to pixel-wise image changes (Sec. 4.1) but also to hyperparameter changes (Sec. 4.3)

2. Methods and Related Work

Let $f : \mathbb{R}^{d\times d\times 3} \to [0, 1]$ be a classifier that maps a color image $\boldsymbol{x}$ of spatial size $d \times d$ onto a probability of a target class. An attribution method is a function A that takes three inputs—an image $\boldsymbol{x}$, the model f, and a set of hyperparameters $\mathcal{H}$—and outputs a matrix $\boldsymbol{a} = A(f, \boldsymbol{x}, \mathcal{H}) \in [-1, 1]^{d\times d}$. Here, the explanation $\boldsymbol{a}$ associates each input pixel x_i to a scalar $a_i \in [-1, 1]$, which indicates how much x_i contributes for or against the classification score $f(\boldsymbol{x})$.

Methods Attribution methods can be categorized into two main types: (1) exact and (2) approximate approaches. **Exact approaches** may derive an attribution map by upsampling a feature map of a convolutional network [62], or from the analytical gradients of the classification w.r.t. the input *i.e.* $\nabla_{\boldsymbol{x}} f$ [47, 8], or by combining both the gradients and the feature maps [45]. These approaches enjoy fast derivation of explanations and have no hyperparameters in principles. However, they require access to the internal network parameters—which may not be available in practice. Also, taking gradients as attributions faces several issues: (1) gradient images are often noisy [48] limiting their utility; (2) gradient saturation [51] *i.e.* when the function f flattens within the vicinity of a pixel x_i, its gradient becomes near-zero and may misrepresent the actual importance of x_i; (3) sudden changes in the gradient $\partial f / \partial x_i$ (*e.g.* from ReLUs [36]) may yield misleading interpretation of the attribution of pixel x_i [46].

Therefore, many **approximate methods** have been proposed to modify the vanilla gradients to address the aforementioned issues [48, 51, 13]. Among gradient-based methods, we chose to study the following four representatives.
Gradient [47, 8] The gradient image $\nabla_{\boldsymbol{x}} f$ quantifies how a small change of each input pixel modifies the classification and therefore commonly serves as an attribution map.
SmoothGrad (SG) [48] proposed to smooth out a gradient image by averaging out the gradients over a batch of N_{SG} noisy versions $\boldsymbol{x}_n$ of the input image $\boldsymbol{x}_0$. That is, an SG heatmap is $\frac{1}{N_{SG}} \sum_1^{N_{SG}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0 + \epsilon)$ where $\epsilon \sim \mathcal{N}(0, \sigma)$.
Gradient $\odot$ Input (GI) [46] As gradients are often noisy and thus not interpretable [48], element-wise multiplying the gradient image with the input *i.e.* $\nabla_{\boldsymbol{x}} f \odot \boldsymbol{x}$ can yield less-noisy heatmaps in practice. Here, the input image acts as a model-independent smoothing filter. GI is an approximation of a family of related LRP methods [13] as shown in [11] and is also a representative for other explicit gradient-based extensions [25, 61, 34, 46].
Integrated Gradients (IG) [51] In order to ameliorate the gradient saturation problem [51], IG intuitively replaces the gradient in GI [46] with an average of the gradients evaluated for N_{IG} images linearly sampled along a straight line between the original image $\boldsymbol{x}$ and a zero image. IG is intuitively a smooth version of GI and depends on the sample size N_{IG} while GI has no hyperparameters.

Furthermore, there exist other approximate methods that attempt to compute the attribution of an input region by replacing it with zeros [60, 42], random noise [17], or blurred versions of the original content [22]. These methods inherently depend on many empirically-chosen hyperparameters. Among the family of perturbation-based methods, we chose to study the following three famous representatives.
Sliding Patch (SP) [60] slides a square, occlusion patch of size $p \times p$ across the input image and records the prediction changes into an attribution map. This approach is applicable to any black-box classifier f and widely used [25, 4, 38, 11].
LIME [42] Instead of a square patch, LIME generates N_{LIME} masked images $\{\bar{\boldsymbol{x}}^i\}$ by masking out a random set of S non-overlapping superpixels in the input image. Intuitively, the attribution for a superpixel k is proportional to

[1] Code is available at https://github.com/anguyen8/sam

the average score $f(\bar{\boldsymbol{x}}^i)$ over a batch of N_{LIME} perturbed images where the superpixel k is not masked out.

Meaningful-Perturbation (MP) [22] finds a minimal Gaussian blur mask of radius b_R such that when applied over the input image would produce a blurred version that has a near-zero classification score. MP is the basis for many extensions [56, 40, 15, 57, 54, 9]. In this paper, we evaluate MP sensitivity to three common hyperparameters: the blur radius b_R, the number of steps N_{iter}, and the random seed (which determines the random initialization).

See Sec. S1 for a detailed description of all methods.

Explanation sensitivity First, recent work has argued that some attribution methods have a problem of being highly sensitive to small pixel-wise changes in the input image [28, 10, 23]. Our results suggest that such sensitivity to image changes also depends on the classifier f. That is, gradient-based explanations of a robust classifier stay more consistent when the input image is perturbed with pixel-wise noise (Sec. 4.1). Second, some attribution methods were found to behave akin to an edge detector *i.e.* producing similar explanations despite that f's parameters are randomized to various degrees [8]. In sum, previous work has studied the sensitivity of explanations to input image changes [28, 10, 23] and classifier changes [8]. In this paper, we present the first systematic study on the sensitivity of explanations to changes in the *hyperparameters* $\mathcal{H}$, which are often randomly or heuristically tuned [48, 58].

3. Experiment framework

Explanation evaluation metrics Currently, there is not yet a common ground-truth dataset for evaluating the accuracy of attribution methods [18]. However, researchers often approximate explanation correctness via two main techniques: (1) object localization [62]; and (2) Insertion & Deletion [38]. The **localization error** measures how accurately an attribution map localizes the main object in the input image [62]—a reasonable approximation for the ImageNet images [43], which are object-centric and paired with human-labeled segmentation masks. We did not use evaluation metrics like Pointing Game accuracy [61] and Saliency Metric [17] as they are derivatives of the localization task. The **Deletion** metric [38] measures the classification score changes as we gradually zero out the input pixels in the descending order of their attributions. The idea is if the attribution values correctly reflect the discriminative power of the input pixels, knocking out the highest-attribution pixels should quickly cause the probability to approach zero. In contrast, **Insertion** [38] tests whether inserting the highest-attribution pixels into a zero image would quickly increase the probability. We used all three above mentioned metrics[2] to quantify how much the variation of explanations translates into the sensitivity of their accuracy (Sec. 4.5).

Classifiers All of our experiments were conducted on two groups of classifiers: (a) GoogLeNet [52] & ResNet-50 [26] (hereafter, ResNet) pre-trained on the 1000-class 2012 ImageNet dataset [43]; and (b) the robust versions of them *i.e.* GoogLeNet-R & ResNet-R that were trained to also be invariant to small adversarial changes in the input image [20]. We obtained the two regular models from the PyTorch model zoo [39], the ResNet-R from [20], and we trained GoogLeNet-R by ourselves using the code released by [20]. While the two robust classifiers are more invariant to pixel-wise noise they have lower ImageNet validation-set accuracy scores (50.94% and 56.25%) than those of the original GoogLeNet & ResNet (68.86% and 75.59%).

Datasets From the 50,000 ImageNet validation-set images, we randomly sampled a set of 1735 images that all four models correctly classify. We used this set of images in all experiments throughout the paper.

Similarity metrics To quantify the sensitivity of attribution maps, we followed Adebayo et al. [8] and used three measures[3] that cover a wide range of similarity notions: Spearman rank correlation, Pearson correlation of the histogram of gradients (HOGs), and the structural similarity index (SSIM). To quantify the sensitivity of the accuracy scores of explanations, we used the standard deviation (std).

4. Experiments and Results

4.1. Gradient maps of robust classifiers are smooth and insensitive to pixel-wise image noise

Gradient saliency maps of image classifiers are (1) notoriously noisy [47, 48, 13] limiting their utility and (2) sensitive to input changes [10]. Therefore, a number of techniques have been proposed to de-noise the gradient images [46, 48, 49, 45]. However, are these smoothing techniques necessary for gradients of robust classifiers?

First, we observed, for the first time, that the vanilla gradients of robust classifiers consistently exhibit visible structures (see the outline of the goblet in Fig. 3c & e), which is surprising! They are in stark contrast to the noisy gradients of regular classifiers (Fig. 3b & d).

Second, we found that the gradient explanations of robust classifiers are significantly more invariant to a large amount of random noise added to the input image. Specifically, for each image $\boldsymbol{x}$ in the dataset, we added noise $\sim \mathcal{N}(0,\, 0.1)$ to generate a noisy version $\boldsymbol{x}_n$ (Fig. 3; bottom) and measured the similarity between the saliency maps for the pair $(\boldsymbol{x},\, \boldsymbol{x}_n)$ using all three similarity metrics described in Sec. 3. Across all images and all three quantitative metrics, the gradients of robust classifiers are substantially more invariant to noise than their regular counterparts

[2] We used the Insertion and Deletion code by the authors [38].

[3] We used the implementation by scikit-image [55].

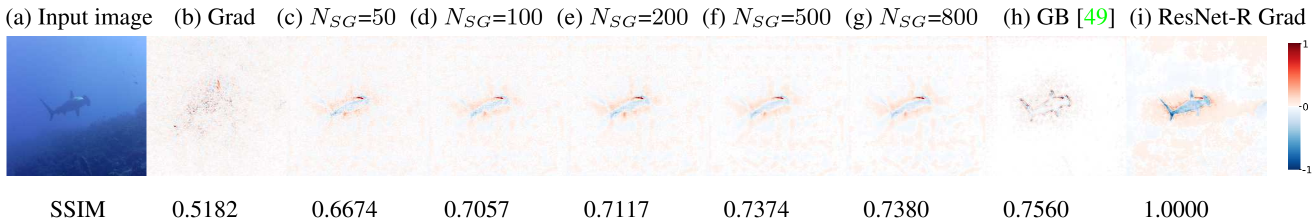


Figure 2: The SmoothGrad [48] explanations (b–g) for a prediction by ResNet are becoming increasingly similar to the explanation for a different prediction by a ResNet-R as we increase N_{SG}—a hyperparameter that governs the smoothness of SG explanations. Similarly, under GuidedBackprop (GB) [49], the explanation appears substantially closer to that of a different model (h vs. i) compared the original heatmaps (b vs. i). Below each heatmap is the SSIM similarity score between that heatmap and the ResNet-R heatmap (i). See more examples in Fig. S14.

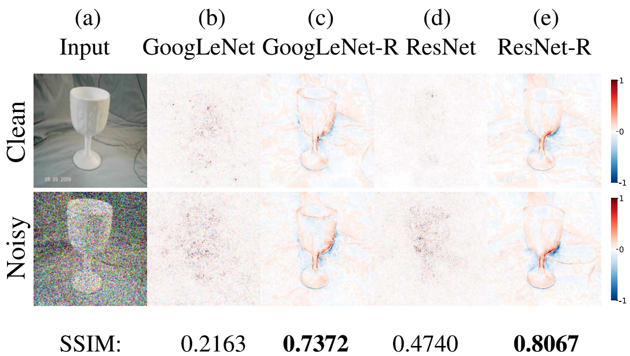


Figure 3: **Top:** The gradients of robust classifiers (c & e) reflect the structure of the goblet in an example input image (a), which is in stark contrast to the commonly reported noisy gradients of regular classifiers (b & d). **Bottom:** The gradients of robust classifiers remain similar before and after the addition of noise to the input image (c & e—higher SSIM scores). An SSIM similarity score is for the two images in each column.

(Figs. 5a & S13). For example, the average similarity of the gradient pairs from robust models is $\sim$36$\times$ higher than that of the counterparts under the Spearman rank correlation (Fig. S13; leftmost bars). This result interestingly show that the gradients of *robust* models are fairly insensitive to minor pixel-wise image changes—a concern in [28, 10, 23].

4.2. De-noising explanations may cause misinterpretation

We have shown that the vanilla gradients of robust classifiers can be fairly smooth (Sec. 4.1). That result naturally raises a follow-up question: Do the smoothing techniques [48, 45, 49] improve or mislead our interpretation of explanations? To shed light on that question, we quantify the similarity between (a) the de-noised explanations by SG [48] for a regular classifier and (b) the vanilla gradient saliency maps for a robust classifier.

Experiment For each image, we generated SG explanations for regular models by sweeping across a range of the sample size $N_{SG} \in \{0, 50, 100, 200, 500, 800\}$. Here, N_{SG} = 0 yields the vanilla gradient. We measured the similarity between each SG heatmap of a regular model and the vanilla gradient of a robust counterpart model (*e.g.* ResNet vs. ResNet-R).

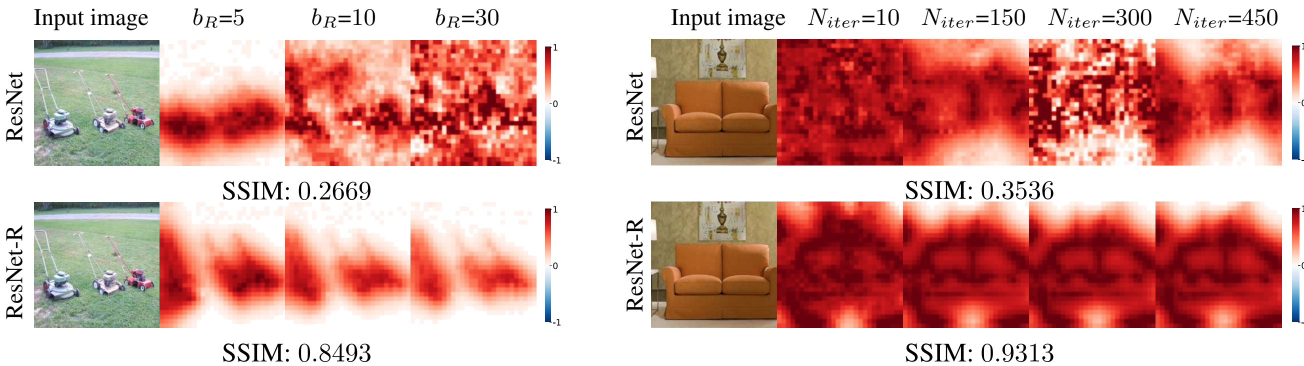
Results We observed that as the sample size N_{SG} increases, the resultant explanations of ResNet become increasingly more similar to the explanation of ResNet-R—a completely different classifier! That is, the SSIM similarity between two heatmaps increases up to $\sim$1.4$\times$ (Fig. 2; b–g) on average. This monotonic trend is also observed across three similarity metrics and two pairs of regular vs. robust models (Fig. S3).

Additionally, we generated an explanation using another popular explanation method, GuidedBackprop (GB) [49], which *modifies* the gradient by only letting the *positive* forward activations and backward gradients to flow through during backpropagation. Across the dataset, the average similarity between a pair of (ResNet GB heatmap, ResNet-R gradient heatmap) is 0.377 while the original similarity between the vanilla gradients of two models is only 0.239.

In sum, our result shows that two explanations from two completely different classifiers (ResNet vs. ResNet-R) may become substantially more similar under explanations techniques (here, SG and GB) that attempt to heuristically denoise heatmaps, potentially misleading user interpretation. We reached the same conclusion by comparing GI and its approximate version *i.e.* IG [51] (see Sec. S3).

4.3. Gradient-based attribution maps are sensitive to hyperparameters

In practice, attribution methods often have various hyperparameters that are either randomly set (*e.g.* a random seed [42]) or empirically tuned (*e.g.* the number of optimization steps [22]). It is important to understand how such choices made by the end-user vary the explanations (Fig. 1), which impedes reproducibility and can impair users' trust *e.g.* a medical doctor's trust in a model's explanation of its prediction [32, 18]. Here, we quantify the sensitivity of attribution maps generated by two representative meth-



(a) Sensitivity to changes in the blur radius b_R

(b) Sensitivity to changes in the number of iterations N_{iter}

Figure 4: MP attribution maps generated for a regular model (ResNet) are highly sensitive to changes (*i.e.* low SSIM scores) in the Gaussian blur radius b_R (a) and in the number of iterations N_{iter} (b). In contrast, the same MP explanations for a robust model (ResNet-R) are substantially more stable (see Fig. S8 for quantitative results). Two reference images in this figure are the top-2 that cause the largest differences between the SSIM scores of ResNet vs. ResNet-R heatmaps. MP being more unstable with ResNet compared to that with ResNet-R can be seen quantitatively in the loss plot (Fig. S9) and qualitatively in the evolution of the MP heatmaps (Fig. S10). See Fig. S16 for more examples of the blur sensitivity experiments.

ods (SG and MP) as a common hyperparameter changes. In all experiments, we compare the average pair-wise similarity between a *reference* heatmap—the explanation generated using the default settings provided by the authors—and those generated by changing one hyperparameter.

4.3.1 SmoothGrad is sensitive to sample sizes

SG was created to combat the issue that gradient images for image classifiers are often too noisy to be human-interpretable—an issue reported in many previous papers [48, 49, 13, 47] and also shown in Sec. 4.1. While SG does qualitatively sharpen the explanations [48] (see Fig. 2b vs. c), the method also introduces two hyperparameters (1) the sample size N_{SG} and (2) the Gaussian std σ that were empirically tuned [48]. Here, we test the sensitivity of SG explanations when varying these two hyperparameters.

Experiment To test the sensitivity to sample sizes, we measure the average pair-wise similarity between a reference heatmap at $N_{SG} = 50$ (Fig. S12b; ii)—*i.e.* the default value in [48]—and each of the four heatmaps generated by sweeping across $N_{SG} \in \{100, 200, 500, 800\}$ (Fig. S12b; iii—vi) on the same input image. σ is constant at 0.15.

Results We found that the SG explanations for robust models exhibited near-maximum consistency (Fig. S12a; all scores are near 1.0). In contrast, the robustness of SG when running on regular models is consistently lower under all three metrics (Fig. S12a; light vs. dark red or light vs. dark green). SG heatmaps for robust classifiers appear sharper and less noisy compared to those of regular models (Fig. S12b; top vs. bottom). Furthermore, while SG heatmaps may appear qualitatively stable (Fig. S12b; ii–vi), the actual pixel-wise variations are not. For example, the L_1 pixel-wise difference between the ResNet heatmaps at the two extreme settings (*i.e.* $N_{SG} = 50$ vs. 800) is over $5\times$ larger than the difference between the respective ResNet-R explanations (Fig. S12b; vii).

In sum, we showed that it is non-trivial how to tune a hyperparameter, here N_{SG}, to yield an accurate explanation because the heatmaps vary differently for different classifiers. Similarly, we further found SG heatmaps to be highly sensitive to changes in the amount of noise *i.e.* Gaussian std σ (Sec. S4.1) added to the input image.

4.3.2 Meaningful-Perturbation is sensitive to the number of iterations, the Gaussian blur radius, and the random seed

MP [22] is a representative of a family of methods that attempt to learn an explanation via iterative optimization [56, 40, 15, 57, 54, 9]. However, in practice, optimization problems are often non-convex and thus the stopping criteria for iterative solvers are heuristically set. For instance, it can be controlled by a pre-defined number of iterations N_{iter}. Also, MP learns to blur the input image to minimize the classification scores and thus depends on the Gaussian blur radius b_R. Here, we test MP sensitivity to three common hyperparameters: N_{iter}, b_R, and the random seed which governs random initializations.

Experiment In order to test the sensitivity to the number of iterations, we measure the average similarity between a reference heatmap at $N_{iter} = 300$ which is the default setting in [22] and each of the three heatmaps generated by sweeping across $N_{iter} \in \{10, 150, 450\}$ (Fig. 4b) on the same input image. To measure the sensitivity to the blur radius settings, we repeated a similar comparison to the above for a reference heatmap at $b_R = 10$ and other heatmaps by sweeping across $b_R \in \{5, 30\}$ (Fig. 4a). For other hyperparameters, we used all default settings as in [22].

Results We found that MP explanations are sensitive to changes in the blur radius but interestingly in opposite ways for two different types of classifiers. That is, as we increase b_R, the heatmaps for ResNet tend to be more noisy and sparse; however, those for ResNet-R become gradually more localized and smoother (Fig. 4a; top vs. bottom). See Fig. S16 for more examples.

Across the number of iterations, MP explanations for regular classifiers vary dramatically. In contrast, the heatmaps for robust models are 1.4× more consistent under SSIM similarity metrics (Figs. 4b & S10). The MP optimization runs for robust models converged substantially faster within only ∼10 steps (compared to the default $N_{iter} = 300$ [22]) which can be seen in both the loss plot (Fig. S9) and the sequence of heatmaps (Fig. S10). This inconsistent behavior of MP suggests that when comparing MP explanations between these two classifiers, an end-user may draw an entirely different conclusion depending on when optimization stops (which is heuristically chosen).

Sensitivity to the random seed Our previous experiments followed exactly the setup in [22] where the authors used a blur circular mask that suppresses the target probability by 99% as the initial heatmap. This initialization, however, strongly biases the optimization towards a certain type of explanation. To avoid that, in practice, MP users randomly initialize the explanation before optimization [16]. By running experiments similar to the previous ones, we found that MP is also sensitive to the random seed, which controls the random initializations. That is, on average across 3 similarity metrics, heatmaps for robust classifiers are 1.22× more consistent than those for regular classifiers (see Sec. S4.3 for more details and Fig. S7 for results).

In sum, consistent with SG results (Sec. 4.3.1), robust classifiers yield more stable explanations than regular models for the three aforementioned hyperparameters of MP (Fig. S8). That is, not only the gradients of robust classifiers are more interpretable but also more invariant to pixel-wise image changes, yielding more robust explanations (Fig. 4b).

4.4. Non-gradient attribution maps are sensitive to hyperparameters

4.4.1 Sliding-Patch is sensitive to the patch size

Sec. 4.3 shows that *gradient*-based explanation methods are sensitive to hyperparameters and their sensitivity depends on the robustness of the gradients with respect to the input changes (Sec. 4.3.2). Here, we test whether methods that are *not gradient-based* would have similar shortcomings. We chose SP [60] which slides a square patch of size $p \times p$ across the input image and records the classification probability changes into the corresponding cells in the attribution map. While SP has been widely used [60, 4, 11], it remains unknown how to choose the patch size.

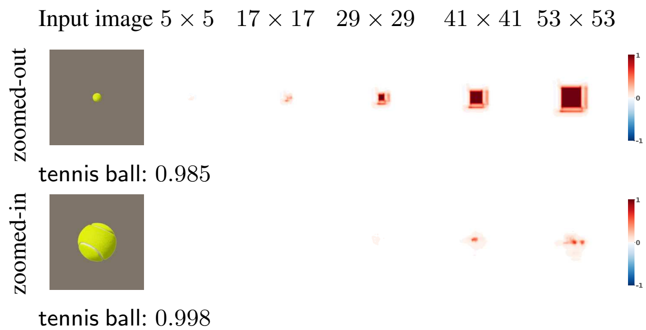
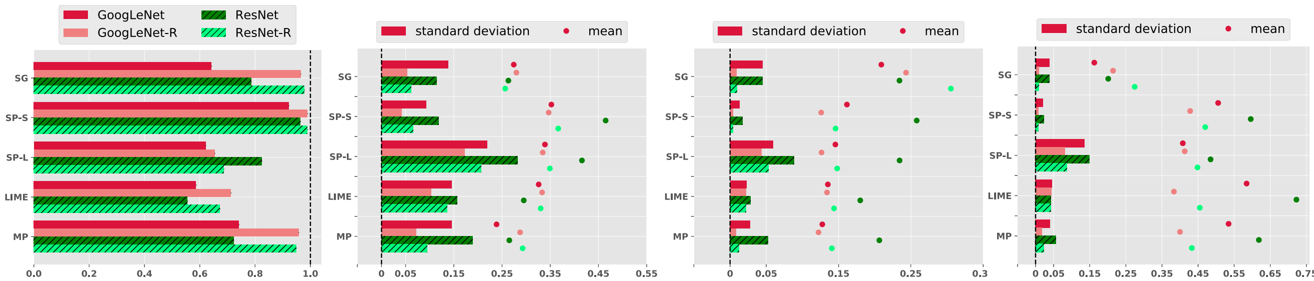


Figure 6: SP explanations are sensitive to patch sizes. **Zoomed-out:** SP attribution region (red squares) for a tennis ball of size 19×19 (rendered on a 224×224 zero image) grows as the patch size increases. **Zoomed-in:** SP outputs blank heatmaps at patch sizes of 5×5, 17×17, and 29×29, which are much smaller than the size of the tennis ball (here, 84 × 84) in a zoomed-in version of the top image.

To understand **the relation between SP patch size and the size of the object** in an input image, we generated two images, each containing a tennis ball of size 19 × 19 or 84×84 on a zero background of size 224×224 (Fig. 6). We ran SP on these two images sweeping across 5 patch sizes of $p \times p$ where $p \in \{5, 17, 29, 41, 53\}$. We observed that the heatmaps tend to be blank when the patch size is much smaller than the object size (Fig. 6; zoomed-in) because the occlusion patch is too small to substantially change the classification score. In contrast, if the patch size is much larger than the object size (Fig. 6; zoomed-out), the attribution areas tend to be exaggerated *i.e.* even larger than the object size (Fig. 6; the size of the red square increases from left to right). Therefore, SP explanations are subject to errors as the size of the object in the image is unknown.

Sensitivity to large changes To quantify the sensitivity of SP explanations to the patch size, here, we measure the average similarity between a reference SP attribution map at $p = 29$ and each of the four attribution maps generated by sweeping across $p \in \{5, 17, 41, 53\}$ on the same input image. This set of patch sizes covers a large range of settings (hence, denoted by SP-L) used in the literature [11, 60, 4]. We kept the stride constant at 3. We observed that across all classifiers, SP is highly sensitive to changes within the SP-L set. In contrast to the case of gradient-based methods, SP explanations for robust classifiers are not significantly more consistent than those for regular models (Fig. S11). Compared to other methods, SP sensitivity to patch sizes is higher than the sensitivity of SG and MP (Fig. 5a; SP-L bars are the shortest on average). See Fig. S15 for more examples on sensitivity to large changes in patch size.

Sensitivity to small changes We further repeated the previous experiment but comparing the similarity of SP explanations at $p = 53$ with those generated at $p \in \{52, 54\}$ i.e. a



(a) SSIM (b) Localization error (c) Deletion (d) Insertion

Figure 5: Average sensitivity of an individual attribution map measured in the pixel space (a) and three accuracy metric spaces: the Localization error (b), Deletion (c) and Insertion (d) scores (Sec. 4.5). The results were produced by varying the random seed of LIME and MP (bottom two rows), the patch size in SP (SP-S and SP-L), and the sample size of SG (top row). SP-S and SP-L are two variants of the SP experiments (Sec. 4.4.1). For Localization performance of SP-S (b), even a change of $\pm$1px in patch size results in a std of $\sim$10% for GoogLeNet (dark red) and ResNet (dark green). Compared to regular models, robust models (here, GoogLeNet- and ResNet-R) cause the attribution maps to be more consistent pixel-wise under hyperparameter changes—*i.e*. higher SSIM scores (a)—and also more consistent in the three accuracy metrics—*i.e*. lower standard deviations (b–d). See Table S1 for the exact numbers.

small range (hence, denoted by SP-S). We observed that SP explanations are not 100% consistent even when the patch dimension changes within only $\pm$1px (Fig. 5a; SSIM scores for SP-S are < 1.0).

4.4.2 LIME is sensitive to random seeds and sample sizes

LIME [42] is a black-box explanation method. Instead of masking out a single square patch (as in SP), which can yield the "square artifact" (Fig. 6; zoomed-out), LIME masks out a finite set of random *superpixels*.

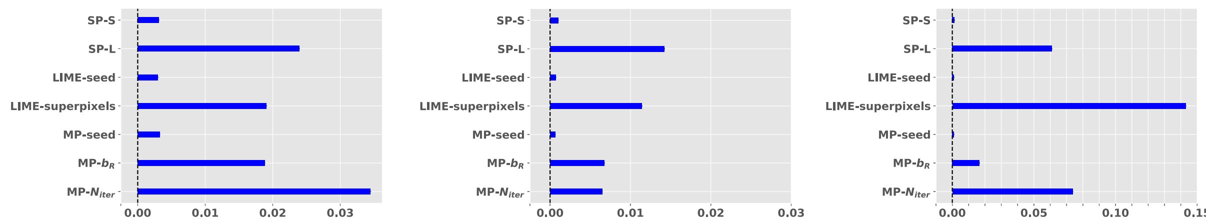
Our experiments show that LIME is highly sensitive to its two common hyperparameters. First, LIME attribution maps interestingly often change as the random seed (which controls the random sampling of superpixel masks) changes! Second, LIME is also sensitive to the changes in the number of perturbation samples. See Sec. S4.2 for more details. Aligned with the results with SP (Sec. 4.4.1), here, we did not find robust classifiers to yield more stable LIME heatmaps than regular classifiers consistently under all three similarity metrics. An explanation is that GoogLeNet-R and ResNet-R are robust to pixel-wise changes but not patch-wise or superpixel-wise changes (as done by SP and LIME) in the input image. See Fig. S17 for a list of the most sensitive cases across all the LIME sensitivity experiments.

4.5. How do the accuracy scores of an explanation vary when a hyperparameter changes?

In Sec. 4.3 and Sec. 4.4, we have shown that many attribution methods are highly sensitive to changes in their common hyperparameters. For example, under SSIM, the average explanation consistency is often far from the maximum (Fig. 5a; GoogLeNet and ResNet scores are far below 1.0). However, there is still a need to quantify how the variation in pixel-wise heatmaps translates into the variation in accuracy scores. That is, two heatmaps that are different pixel-wise may have the same accuracy score. Therefore, it is important for users to understand: *How much does the correctness of an explanation varies, on average, when a given hyperparameter changes?* To answer that, here, we quantify the variance of three explanation accuracy scores (*i.e*. the Localization error, Insertion, and Deletion scores described in Sec. 3) upon varying the *most common* hyperparameters of the considered attribution methods: (1) the sample size in SG (Sec. 4.3.1); (2) the patch size in SP (Sec. 4.4.1; both sweeping across a small range *i.e*. SP-S and a large range *i.e*. SP-L); (3) the random seed in LIME (Sec. 4.4.2); and (4) the random seed in MP (Sec. 4.3.2).

Experiment For each hyperparameter, we swept across N values to generate the corresponding N explanations for each input image. Using an accuracy metric, we evaluated each set of N attribution maps per image to produce N accuracy scores. From the N scores, we then obtained a mean and a std, for each image. From the per-image means and standard deviations, we then calculated the global mean and average std across the dataset (Fig. 5). We repeated the same procedure for each accuracy metric and each classifier.

Results First, we found that changing the tested hyperparameters (*i.e*. which are the most common) does not only change the explanations (Fig. 5a; average SSIM scores are under 1.0) but also their three downstream accuracy scores (Fig. 5b–d; the average std bars are above 0). However, explanation accuracy varies differently between the metrics. That is, compared to the mean scores (Fig. 5; circles), the score variation (in std) are higher for object local-



(a) Localization error (b) Deletion (c) Insertion

Figure 7: Comparisons of the variation in three accuracy scores of attribution methods when changing different hyperparameters. Here, the horizontal bars show standard deviations (std) for the Localization error (a), Deletion (b) and Insertion (c) scores obtained by marginalizing over all images and classifiers (see Sec. 4.6). Changing the number of superpixels (in LIME) and the number of iterations (in MP) causes the largest sensitivity to the accuracy of the two methods, respectively.

ization (Fig. 5b) and lower for deletion and insertion metrics (Fig. 5c–d). Notably, the localization scores are highly sensitive—the average stds of regular and robust models are $0.51\times$ and $0.31\times$ of their respective mean accuracy scores.

Second, varying the patch size of SP by only 1px caused a small variation in the explanation (Fig. 5a; mean SSIM scores are ≈ 1 for SP-S) but a large variation in object localization performance (Fig. 5b; for SP-S, the stds are $\sim$10% of the mean statistics).

Third, across all four tested hyperparameters and three accuracy metrics, the correctness of explanations for robust models is on average $2.4\times$ less variable than that for regular models. In sum, we found that explanations for robust classifiers are not only more consistent but also more similarly accurate upon varying the common hyperparameters (compared to the darker bars *i.e.* regular classifiers, lighter bars are longer in Fig. 5a and shorter in Fig. 5b–d).

4.6. Which hyperparameter when changed causes a higher variation in explanation accuracy?

In Sec. 4.5, we show that the accuracy of an *individual* explanation, on average, can vary substantially as we change a hyperparameter. Here, we ask a different important question: *Which hyperparameter when varied causes a higher variation in explanation accuracy?* That is, we attempt to compare hyperparameters by computing the marginal effects of changing each hyperparameter to the variation in accuracy scores (when marginalizing over all images and four classifiers).

Experiment As a common practice in the literature, for each classifier, we computed an accuracy score for each generated explanation and took a mean accuracy score over the entire dataset. Repeating the computation for N values of each hyperparameter (*e.g.* N random seeds of LIME), we obtained N mean accuracy scores from which we computed an std s. For each hyperparameter, we averaged over $\{s\}_4$ *i.e.* four such stds, each computed for a classifier, yielding one global std, which is used for comparing hyperparameters. Here, we compare the global stds for different hyperparameters within and between methods (see Fig. 7): (1) the patch size in SP (Sec. 4.4.1; SP-S and SP-L); (2) the random seed and the number of superpixels in LIME (Sec. 4.4.2); (3) the random seed, the blur radius, and the number of iterations of MP (Sec. 4.3.2).

SP results Within SP, we found that varying the patch size across a larger range yields a higher variation in accuracy scores (Fig. 7a; SP-L vs. SP-S).

LIME results Our results enable quantitatively comparing the effects of changing different hyperparameters. In LIME, varying the number of superpixels causes far more sensitivity in the correctness of explanations compared to varying the LIME random seed (Fig. 7; row 3 vs. 4). Specifically, the std of Insertion scores when changing the number of superpixels was $130.5\times$ higher as compared to the std when changing the random seed (Fig. 7c).

MP results In MP, changing the number of optimization iterations causes the largest sensitivity in explanation accuracy (among the three MP hyperparameters). Precisely, the std of Insertion scores, when changing the blur radius b_R and the number of iterations N_{iter}, was $16.6\times$ and $74\times$ higher than that when changing the random seed (Fig. 7c; bottom three rows).

Across methods Changing the random seed in LIME vs. in MP (two different methods) interestingly causes a similar variation in all three accuracy metrics (Fig. 7; row 3 vs. 5).

5. Discussion and Conclusion

We present the first thorough study on the sensitivity of attribution methods to changes in their input hyperparameters. Our findings show that the attribution maps for many gradient-based and perturbation-based interpretability methods can change radically upon changing a hyperparameter, causing their accuracy scores to vary as well. We propose to evaluate the sensitivity to hyperparameters as an evaluation metric for attribution methods. It is important to carefully evaluate the pros and cons of interpretability methods with no hyperparameters and those that have.

References

[1] ankurtaly/integrated-gradients: Attributing predictions made by the inception network using the integrated gradients method. https://github.com/ankurtaly/Integrated-Gradients. (Accessed on 11/08/2019). 12

[2] Machine bias — propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Accessed on 06/10/2019). 1

[3] marcotcr/lime: Lime: Explaining the predictions of any machine learning classifier. https://github.com/marcotcr/lime. (Accessed on 11/08/2019). 13

[4] Network visualization based on occlusion sensitivity » deep learning - matlab & simulink. https://blogs.mathworks.com/deep-learning/2017/12/15/network-visualization-based-on-occlusion-sensitivity/. (Accessed on 11/15/2019). 2, 6, 13

[5] Pair-code/saliency: Tensorflow implementation for smoothgrad, grad-cam, guided backprop, integrated gradients and other saliency techniques. https://github.com/pair-code/saliency. (Accessed on 11/08/2019). 13

[6] ruthcfong/perturb_explanations: Code for fong and vedaldi 2017, "interpretable explanations of black boxes by meaningful perturbation". https://github.com/ruthcfong/perturb_explanations. (Accessed on 11/08/2019). 14

[7] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 13

[8] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 2, 3, 12

[9] Chirag Agarwal, Dan Schonfeld, and Anh Nguyen. Removing input features via a generative model to explain their attributions to classifier's decisions. *arXiv preprint arXiv:1910.04256*, 2019. 3, 5

[10] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018. 3, 4

[11] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. 2, 6

[12] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168. Association for Computational Linguistics, 2017. 1

[13] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 2, 3, 5

[14] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 2018. 1

[15] Marco Carletti, Marco Godi, Maedeh Aghaei, and Marco Cristani. Understanding deep architectures by visual summaries. *arXiv preprint arXiv:1801.09103*, 2018. 3, 5

[16] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019. 6

[17] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017. 2, 3

[18] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 1, 3, 4

[19] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017. 1

[20] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations, 2019. 3, 14

[21] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Learning perceptually-aligned representations via adversarial robustness. *arXiv preprint arXiv:1906.00945*, 2019. 2

[22] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 1, 2, 3, 4, 5, 6, 17, 18

[23] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019. 3, 4

[24] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. 1

[25] Luisa M graf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017. 1, 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[27] Matthew Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. 2

[28] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019. 3, 4

[29] F Lambert. Understanding the fatal tesla accident on autopilot and the nhtsa probe. *Electrek, July*, 2016. 1

[30] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Muller, and Wojciech Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2912–2920, 2016. 1

[31] Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Muller, and Wojciech Samek. Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1629–1638, 2017. 1

[32] Zachary C Lipton. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037*, 2017. 1, 4

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 14

[34] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. 2

[35] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. 1

[36] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 2

[37] Anh Nguyen, Jason Yosinski, and Jeff Clune. 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, pages 427–436. IEEE, June 2015. 1

[38] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2, 3

[39] PyTorch. torchvision.models — pytorch master documentation. https://pytorch.org/docs/stable/torchvision/models.html, 2019. (Accessed on 09/21/2019). 3, 14

[40] Zhongang Qi, Saeed Khorram, and Fuxin Li. Visualizing deep networks by optimizing with integrated gradients. *arXiv preprint arXiv:1905.00954*, 2019. 3, 5

[41] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016. 1, 2, 4, 7

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[44] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 5–22. Springer, 2019. 1

[45] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. 2, 3, 4

[46] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016. 2, 3, 15

[47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2, 3, 5

[48] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 1, 2, 3, 4, 5, 15, 16

[49] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2, 3, 4, 5, 14, 24

[50] Vignesh Srinivasan, Sebastian Lapuschkin, Cornelius Hellge, Klaus-Robert Müller, and Wojciech Samek. Interpretable human action recognition in compressed domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 1

[51] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017. 2, 4, 14, 15

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[54] Hristina Uzunova, Jan Ehrhardt, Timo Kepp, and Heinz Handels. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094911. International Society for Optics and Photonics, 2019. 3, 5

[55] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014. 3

[56] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019. 3, 5

[57] Yulong Wang, Xiaolin Hu, and Hang Su. Learning attributions grounded in existing facts for robust visual explanation. *XAI 2018*, page 178. 3, 5

[58] Yi Wei, Ming-Ching Chang, Yiming Ying, Ser Nam Lim, and Siwei Lyu. Explain black-box image classifications using superpixel-based interpretation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1640–1645. IEEE, 2018. 3

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

[60] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1, 2, 6

[61] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 2, 3

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2, 3