Rethinking Robustness of Model Attributions

Sandesh Kamath[1], **Sankalp Mittal**[1], **Amit Deshpande**[2], **Vineeth N Balasubramanian**[1]

[1]Indian Institute of Technology, Hyderabad
[2]Microsoft Research, Bengaluru
sandesh.kamath@gmail.com

Abstract

For machine learning models to be reliable and trustworthy, their decisions must be interpretable. As these models find increasing use in safety-critical applications, it is important that not just the model predictions but also their explanations (as feature attributions) be robust to small human-imperceptible input perturbations. Recent works have shown that many attribution methods are fragile and have proposed improvements in either these methods or the model training. We observe two main causes for fragile attributions: first, the existing metrics of robustness (e.g., top-k intersection) overpenalize even reasonable local shifts in attribution, thereby making random perturbations to appear as a strong attack, and second, the attribution can be concentrated in a small region even when there are multiple important parts in an image. To rectify this, we propose simple ways to strengthen existing metrics and attribution methods that incorporate locality of pixels in robustness metrics and diversity of pixel locations in attributions. Towards the role of model training in attributional robustness, we empirically observe that adversarially trained models have more robust attributions on smaller datasets, however, this advantage disappears in larger datasets. Code is made available at https://github.com/ksandeshk/LENS.

1 Introduction

The explosive increase in the use of deep neural network (DNN)-based models for applications across domains has resulted in a very strong need to find ways to interpret the decisions made by these models (Gade et al. 2020; Tang et al. 2021; Yap et al. 2021; Oviedo et al. 2022; Oh and Jeong 2020). Interpretability is an important aspect of responsible and trustworthy AI, and model explanation methods (also known as attribution methods) are an important aspect of the community's efforts towards explaining and debugging real-world AI/ML systems. Attribution methods (Zeiler et al. 2010; Simonyan, Vedaldi, and Zisserman 2014; Bach et al. 2015; Selvaraju et al. 2017; Chattopadhyay et al. 2018; Sundararajan, Taly, and Yan 2017; Shrikumar et al. 2016; Smilkov et al. 2017; Lundberg and Lee 2017) attempt to explain the decisions made by DNN models through input-output attributions or saliency maps. (Lipton 2018; Samek et al. 2019; Fan et al. 2021; Zhang et al. 2020) present detailed surveys on these methods. Recently, the growing numbers of attribution methods has led to a concerted focus on studying the robustness of attributions to input perturbations to handle potential security hazards (Chen et al. 2019; Sarkar, Sarkar, and Balasubramanian 2021; Wang and Kong 2022; Agarwal et al. 2022). One could view these efforts as akin to adversarial robustness that focuses on defending against attacks on model predictions, whereas *attributional robustness* focuses on defending against attacks on model explanations. For example, an explanation for a predicted credit card failure cannot change significantly for a small human-imperceptible change in input features, or the saliency maps explaining the COVID risk prediction from a chest X-ray should not change significantly with a minor human-imperceptible change in the image.

DNN-based models are known to have a vulnerability to imperceptible adversarial perturbations (Biggio et al. 2013; Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015), which make them misclassify input images. Adversarial training (Madry et al. 2018) is widely understood to provide a reasonable degree of robustness to such perturbation attacks. While adversarial robustness has received significant attention over the last few years (Ozdag 2018; Silva and Najafirad 2020), the need for stable and robust attributions, corresponding explanation methods and their awareness are still in their early stages at this time (Ghorbani, Abid, and Zou 2019; Chen et al. 2019; Slack et al. 2020; Sarkar, Sarkar, and Balasubramanian 2021; Lakkaraju, Arsov, and Bastani 2020; Slack et al. 2021a,b). In an early effort, (Ghorbani, Abid, and Zou 2019) provided a method to construct a small imperceptible perturbation which when added to an input x results in a change in attribution map of the original map to that of the perturbed image. This is measured through top-k intersection, Spearman's rank-order correlation or Kendall's rank-order correlation between the two attribution maps (of original and perturbed images). See Figure 1 for an example. Defenses proposed against such attributional attacks (Chen et al. 2019; Singh et al. 2020; Wang et al. 2020; Sarkar, Sarkar, and Balasubramanian 2021) also leverage the same metrics to evaluate the robustness of attribution methods.

While these efforts have showcased the need and importance of studying the robustness of attribution methods, we note in this work that the metrics used, and hence the meth-

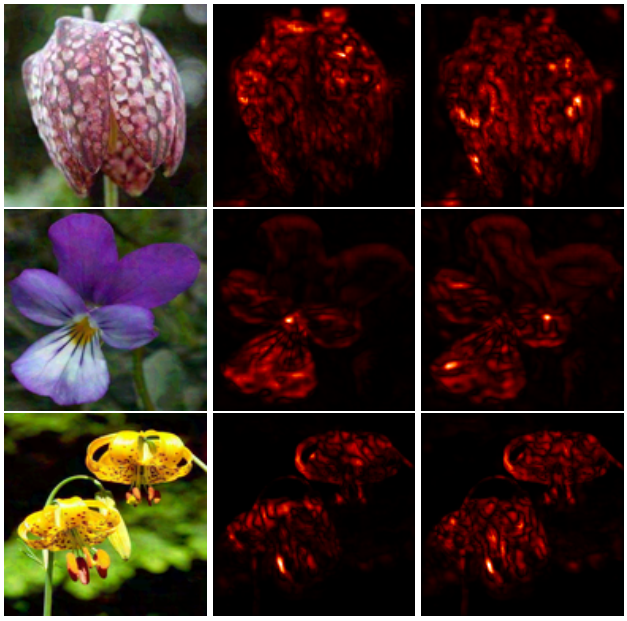

Figure 1: Sample images from Flower dataset with Integrated Gradients (IG) before and after attributional attack. The attack used here is the top-k attributional attack of Ghorbani, Abid, and Zou (2019) on a ResNet model. Robustness of attribution measured by top-k intersection is small, and ranges from 0.04 (first image) to 0.45 (third image) as it penalizes for both local changes in attribution and concentration of top pixels in a small region. Visually, we can observe that such overpenalization leads to a wrong sense of robustness as the changes are within the object of importance.

ods, can be highly sensitive to minor local changes in attributions (see Fig 1 *row 2*). We, in fact, show (in Appendix B.1) that under existing metrics to evaluate robustness of attributions, a random perturbation can be as strong an attributional attack as existing benchmark methods. This may not be a true indicator of the robustness of a model's attributions, and can mislead further research efforts in the community. We hence focus our efforts in this work on rethinking metrics and methods to study the robustness of model attributions (in particular, we study image-based attribution methods to have a focused discussion and analysis). Beyond highlighting this important issue, we propose locality-sensitive improvements of the above metrics that incorporate the locality of attributions along with their rank order. We show that such a locality-sensitive distance is upper-bounded by a metric based on symmetric set difference. We also introduce a new measure **top-k-div** that incorporates diversity of a model's attributions. Our key contributions are summarized below:

- Firstly, we observe that existing robustness metrics for model attributions overpenalize minor drifts in attribution, leading to a false sense of fragility.
- In order to address this issue, we propose Locality-sENSitive (LENS) improvements of existing metrics, namely, LENS-top-k, LENS-Spearman and LENS-Kendall, that incorporate the locality of attributions along with their rank order. Besides avoiding overpenalizing attribution methods for minor local drifts, we show that our proposed LENS variants are well-motivated by metrics defined on the space of attributions.
- We subsequently introduce a second measure based on diversity that enriches model attributions by preventing the localized grouping of top model attributions. LENS can be naturally applied to this measure, thereby giving a method to incorporate both diversity and locality in measuring attributional robustness.
- Our comprehensive empirical results on benchmark datasets and models used in existing work clearly support our aforementioned observations, as well as the need to rethink the evaluation of the robustness of model attributions using locality and diversity.
- Finally, we also show that existing methods for robust attributions implicitly support such a locality-sensitive metric for evaluating progress in the field.

2 Background and Related Work

We herein discuss background literature from three different perspectives that may be related to our work: model explanation/attribution methods, efforts on attributional robustness (both attacks and defenses), and other recent related work.

Attribution Methods. Existing efforts on explainability in DNN models can be broadly categorized as: local and global methods, model-agnostic and model-specific methods, or as post-hoc and ante-hoc (intrinsically interpretable) methods (Molnar 2019; Lecue et al. 2021). Most existing methods in use today – including methods to visualize weights and neurons (Simonyan, Vedaldi, and Zisserman 2014; Zeiler and Fergus 2014), guided backpropagation (Springenberg et al. 2015), CAM (Zhou et al. 2016), GradCAM (Selvaraju et al. 2017), Grad-CAM++ (Chattopadhyay et al. 2018), LIME (Ribeiro, Singh, and Guestrin 2016), DeepLIFT (Shrikumar et al. 2016; Shrikumar, Greenside, and Kundaje 2017), LRP (Bach et al. 2015), Integrated Gradients (Sundararajan, Taly, and Yan 2017), SmoothGrad (Smilkov et al. 2017)), DeepSHAP (Lundberg and Lee 2017) and TCAV (Kim et al. 2018) – are post-hoc methods, which are used on top of a pre-trained DNN model to explain its predictions. We focus on such post-hoc attribution methods in this work. For a more detailed survey of explainability methods for DNN models, please see (Lecue et al. 2021; Molnar 2019; Samek et al. 2019).

Robustness of Attributions. The growing numbers of attribution methods proposed has also led to efforts on identifying the desirable characteristics of such methods (Alvarez-Melis and Jaakkola 2018; Adebayo et al. 2018; Yeh et al. 2019; Chalasani et al. 2020; Tomsett et al. 2020; Boggust et al. 2022; Agarwal et al. 2022). A key desired trait that has been highlighted by many of these efforts is robustness or stability of attributions, i.e., the explanation should not vary significantly within a small local neighborhood of the input (Alvarez-Melis and Jaakkola 2018; Chalasani et al. 2020). Ghorbani, Abid, and Zou (2019)

showed that well-known methods such as gradient-based attributions, DeepLIFT (Shrikumar, Greenside, and Kundaje 2017) and Integrated Gradients (IG) (Sundararajan, Taly, and Yan 2017) are vulnerable to such input perturbations, and also provided an algorithm to construct a small imperceptible perturbation which when added to the input results in changes in the attribution. Slack et al. (2020) later showed that methods like LIME (Ribeiro, Singh, and Guestrin 2016) and DeepSHAP (Lundberg and Lee 2017) are also vulnerable to such manipulations. The identification of such vulnerability and potential for attributional attacks has since led to multiple research efforts to make a model's attributions robust. Chen et al. (2019) proposed a regularization-based approach, where an explicit regularizer term is added to the loss function to maintain the model gradient across input (IG, in particular) while training the DNN model. This was subsequently extended by (Sarkar, Sarkar, and Balasubramanian 2021; Singh et al. 2020; Wang et al. 2020), all of whom provide different training strategies and regularizers to improve attributional robustness of models. Each of these methods including Ghorbani, Abid, and Zou (2019) measures change in attribution before and after input perturbation using the same metrics: top-k intersection, and/or rank correlations like Spearman's ρ and Kendall' τ. Such metrics have recently, in fact, further been used to understand issues surrounding attributional robustness (Wang and Kong 2022). Other efforts that quantify stability of attributions in tabular data also use Euclidean distance (or its variants) between the original and perturbed attribution maps (Alvarez-Melis and Jaakkola 2018; Yeh et al. 2019; Agarwal et al. 2022). Each of these metrics look for dimension-wise correlation or pixel-level matching between attribution maps before and after perturbation, and thus penalize even a minor change in attribution (say, even by one pixel coordinate location). This results in a false sense of fragility, and could even be misleading. In this work, we highlight the need to revisit such metrics, and propose variants based on locality and diversity that can be easily integrated into existing metrics.

Other Related Work. In other related efforts that have studied similar properties of attribution-based explanations, (Carvalho, Pereira, and Cardoso 2019; Bhatt, Weller, and Moura 2020) stated that stable explanations should not vary too much between similar input samples, unless the model's prediction changes drastically. The abovementioned attributional attacks and defense methods (Ghorbani, Abid, and Zou 2019; Sarkar, Sarkar, and Balasubramanian 2021; Singh et al. 2020; Wang et al. 2020) maintain this property, since they focus on input perturbations that change the attribution without changing the model prediction itself. Similarly, Arun et al. (2020) and Fel et al. (2022) introduced the notions of repeatability/reproducibility and generalizability respectively, both of which focus on the desired property that a trustworthy explanation must point to similar evidence across similar input images. In this work, we provide a practical metric to study this notion of similarity by considering locality-sensitive metrics.

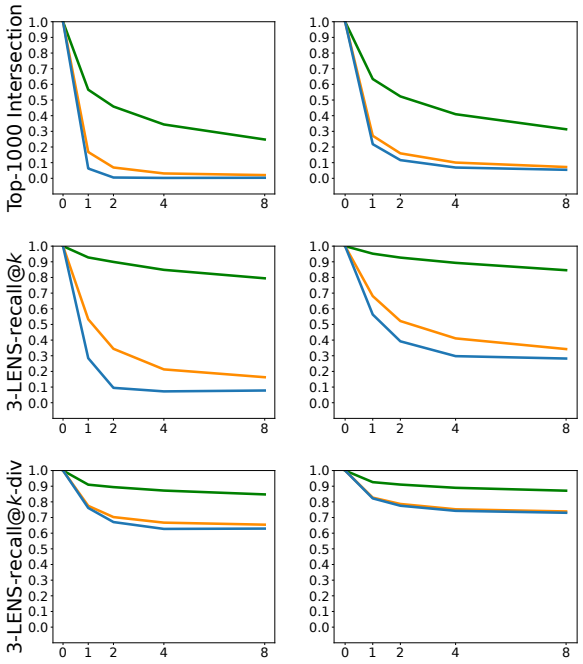

Figure 2: From top to bottom, we plot average top-k intersection (currently used metric), 3-LENS-recall@k and 3-LENS-recall@k-div (proposed metrics) against the ℓ_∞-norm of attributional attack perturbations for Simple Gradients (SG) (left) and Integrated Gradients (IG) (right) of a SqueezeNet model on Imagenet. We use $k = 1000$ and three attributional attack variants proposed by Ghorbani, Abid, and Zou (2019). Evidently, the proposed metrics show more robustness under the same attacks.

3 Locality-sENSitive Metrics (LENS) for Attributional Robustness

As a motivating example, Figure 2 presents the results obtained using (Ghorbani, Abid, and Zou 2019) with Simple Gradients (SG) and Integrated Gradients (IG) of an NN model trained on ImageNet. The top row, which reports the currently followed top-k intersection measure of attribution robustness, shows a significant drop in robustness performance even for the random sign attack (green line). The subsequent rows, which report our metrics for the same experiments, show significant improvements in robustness – especially when combining the notions of locality and diversity. Observations made on current metrics could lead to a false sense of fragility, which overpenalizes even an attribution shift by 1-2 pixels. A detailed description of our experimental setup for these results is available in Appendix C. Motivated by these observations, we explore improved measures for attributional robustness that maintain the overall requirements of robustness, but do not overpenalize minor deviations.

3.1 Defining LENS Metrics for Attributions

To begin with, we propose an extension of existing similarity measures to incorporate the locality of pixel attribu-

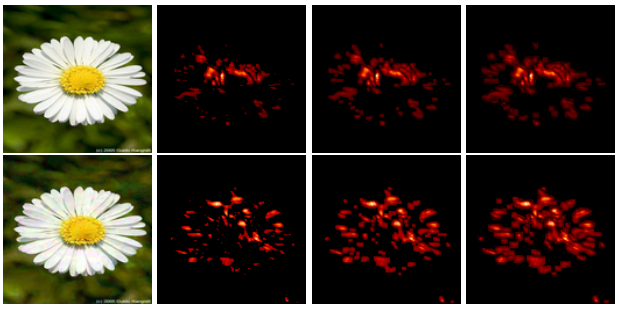

Figure 3: A sample image from Flower dataset before (top) and after (bottom) the top-k attributional attack of (Ghorbani, Abid, and Zou 2019) on a ResNet model for Integrated Gradients (IG) attribution method. From left to right: the image, its top-k pixels as per IG, the union of the 3×3-pixel neighborhoods and 5×5-pixel neighborhoods of the top-k pixels, respectively, for $k = 1000$. Quantitatively, top-k intersection: 0.14, 1-LENS-recall@k: 0.25, 1-LENS-pre@k: 0.37, 2-LENS-recall@k: 0.40, 2-LENS-pre@k: 0.62.

tions in images to derive more practical and useful measures of attributional robustness. Let $a_{ij}(x)$ denote the attribution value or importance assigned to the (i,j)-th pixel in an input image x, and let $S_k(x)$ denote the set of k pixel positions with the largest attribution values. Let $N_w(i,j) = \{(p,q) \,:\, i-w \le p \le i+w,\ j-w \le q \le j+w\}$ be the neighboring pixel positions within a $(2w+1) \times (2w+1)$ window around the (i,j)-th pixel. By a slight abuse of notation, we use $N_w(S_k(x))$ to denote $\bigcup_{(i,j)\in S_k(x)} N_w(i,j)$, that is, the set of all pixel positions that lie in the union of $(2w+1) \times (2w+1)$ windows around the top-k pixels.

For a given attributional perturbation $\mathrm{Att}(\cdot)$, let $T_k = S_k(x + \mathrm{Att}(x))$ denote the top-k pixels in attribution values after applying the attributional perturbation $\mathrm{Att}(x)$. The currently used top-k intersection metric is then computed as: $|S_k(x) \cap T_k(x)|\,/k$. To address the abovementioned issues, we instead propose *Locality-sENSitive top-k* metrics (LENS-top-k) as $|N_w(S_k(x)) \cap T_k(x)|\,/k$ and $|S_k(x) \cap N_w(T_k(x))|\,/k$, which are also closer to more widely used metrics such as precision and recall in ranking methods. We similarly define Locality-sENSitive Spearman's ρ (LENS-Spearman) and Locality-sENSitive Kendall's τ (LENS-Kendall) metrics as rank correlation coefficients for the smoothed ranking orders according to $\tilde{a}_{ij}(x)$'s and $\tilde{a}_{ij}(x + \mathrm{Att}(x))$'s, respectively. These can be used to compare two different attributions for the same image, the same attribution method on two different images, or even two different attributions on two different images, as long as the attribution vectors lie in the same space, e.g., images of the same dimensions where attributions assign importance values to pixels. Figure 3 provides the visualization of the explanation map of a sample from the Flower dataset with the top-1000 pixels followed by the corresponding maps with 1-LENS@k and 2-LENS@k.

We show that the proposed locality-sensitive variants of the robustness metrics also possess some theoretically interesting properties. Let $\mathbf{a}_1$ and $\mathbf{a}_2$ be two attribution vectors for two images, and let S_k and T_k be the set of top k pixels in these images according to $\mathbf{a}_1$ and $\mathbf{a}_2$, respectively. We define a locality-sensitive top-k distance between two attribution vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ as $d_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2) \stackrel{\text{def}}{=} \mathrm{prec}_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2) + \mathrm{recall}_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2)$, where $\mathrm{prec}_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2) \stackrel{\text{def}}{=} \frac{|S_k \setminus N_w(T_k)|}{k}$ and $\mathrm{recall}_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2) \stackrel{\text{def}}{=} \frac{|T_k \setminus N_w(S_k)|}{k}$, similar to precision and recall used in ranking literature, with the key difference being the inclusion of neighborhood items based on locality. Below we state a monotonicity property of $d_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2)$ and upper bound it in terms of the symmetric set difference of top-k attributions.

Proposition 1. *For any $w_1 \le w_2$, we have $d_k^{(w_2)}(\mathbf{a}_1,\mathbf{a}_2) \le d_k^{(w_1)}(\mathbf{a}_1,\mathbf{a}_2) \le |S_k \triangle T_k|\,/k$, where $\triangle$ denotes the symmetric set difference, i.e., $A \triangle B = (A \setminus B) \cup (B \setminus A)$.*

Combining $d_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2)$ across different values of k and w, we can define a distance

$$d(\mathbf{a}_1,\mathbf{a}_2) = \sum_{k=1}^{\infty} \alpha_k \sum_{w=0}^{\infty} \beta_w\, d_k^{(w)}(\mathbf{a}_1,\mathbf{a}_2),$$

where α_k and β_w be non-negative weights, monotonically decreasing in k and w, respectively, such that $\sum_k \alpha_k < \infty$ and $\sum_w \beta_w < \infty$. We show that the distance defined above is upper-bounded by a metric similar to those proposed in (Fagin, Kumar, and Sivakumar 2003) based on symmetric set difference of top-k ranks to compare two rankings.

Proposition 2. *$d(\mathbf{a}_1,\mathbf{a}_2)$ defined above is upper-bounded by $u(\mathbf{a}_1,\mathbf{a}_2)$ given by*

$$u(\mathbf{a}_1,\mathbf{a}_2) = \sum_{k=1}^{\infty} \alpha_k \sum_{w=0}^{\infty} \beta_w\, \frac{|S_k \triangle T_k|}{k},$$

and $u(\mathbf{a}_1,\mathbf{a}_2)$ defines a bounded metric on the space of attribution vectors.

Note that top-k intersection, Spearman's ρ and Kendall's τ do not take the attribution values $a_{ij}(x)$'s into account but only the rank order of pixels according to these values. We also define a locality-sensitive w-smoothed attribution as follows.

$$\tilde{a}_{ij}^{(w)}(x) = \frac{1}{(2w+1)^2} \sum_{\substack{(p,q)\in N_w(i,j),\\ 1\le p,q\le n}} a_{pq}(x)$$

We show that the w-smoothed attribution leads to a contraction in the ℓ_2 norm commonly used in theoretical analysis of simple gradients as attributions.

Proposition 3. *For any inputs x, y and any $w \ge 0$, $\left\|\tilde{\mathbf{a}}^{(w)}(x) - \tilde{\mathbf{a}}^{(w)}(y)\right\|_2 \le \|\mathbf{a}(x) - \mathbf{a}(y)\|_2$.*

Thus, any theoretical bounds on the attributional robustness of simple gradients in ℓ_2 norm proved in previous works continue to hold for locality-sensitive w-smoothed gradients. For example, (Wang et al. 2020) show the following Hessian-based bound on simple gradients. For an

input x and a classifier or model defined by f, let $\nabla_x(f)$ and $\nabla_y(f)$ be the simple gradients w.r.t. the inputs at x and y. Theorem 3 in (Wang et al. 2020) upper bounds the ℓ_2 distance between the simple gradients of nearby points $\|x-y\|_2 \leq \delta$ as $\|\nabla_x(f) - \nabla_y(f)\|_2 \lesssim \delta\ \lambda_{\max}(H_x(f))$, where $H_x(f)$ is the Hessian of f w.r.t. the input at x and $\lambda_{\max}(H_x(f))$ is its maximum eigenvalue. By Proposition 3 above, the same continues to hold for w-smoothed gradients, i.e., $\left\|\tilde{\nabla}_x^{(w)}(f) - \tilde{\nabla}_y^{(w)}(f)\right\|_2 \lesssim \delta\ \lambda_{\max}(H_x(f))$. The proofs of all the propositions above are included in Appendix D.

3.2 Relevance to Attributional Robustness

The top-k intersection is a measure of similarity instead of distance. Therefore, in our experiments for attributional robustness, we use locality-sensitive similarity measures w-LENS-prec@k and w-LENS-recall@k to denote $1 - \text{prec}_k^{(w)}(\mathbf{a}_1, \mathbf{a}_2)$ and $1 - \text{recall}_k^{(w)}(\mathbf{a}_1, \mathbf{a}_2)$, respectively, where $\mathbf{a}_1$ is the attribution of the original image and $\mathbf{a}_2$ is the attribution of the perturbed image. For rank correlation coefficients such as Kendall's τ and Spearman's ρ, we compute w-LENS-Kendall and w-LENS-Spearman as the same Kendall's τ and Spearman's ρ but computed on the locality-sensitive w-smoothed attribution map $\tilde{\mathbf{a}}^{(w)}$ instead of the original attribution map $\mathbf{a}$. We also study how these similarity measures and their resulting attributional robustness measures change as we vary w. In this section, we measure the attributional robustness of Integrated Gradients (IG) on naturally trained models as top-k intersection, w-LENS-prec@k and w-LENS-recall@k between the IG of the original images and the IG of their perturbations obtained by various attacks. The attacks we consider are the top-t attack and the mass-center attack of Ghorbani, Abid, and Zou (2019) as well as random perturbation. All perturbations have ℓ_∞ norm bounded by $\delta = 0.3$ for MNIST, $\delta = 0.1$ for Fashion MNIST, and $\delta = 8/255$ for GTSRB and Flower datasets.

The values of t used to construct top-t attacks of Ghorbani, Abid, and Zou (2019) are $t = 200$ on MNIST, $t = 100$ on Fashion MNIST and GTSRB, $t = 1000$ on Flower. In the robustness evaluations for a fixed k, we use $k = 100$ on MNIST, Fashion MNIST, GTSRB, and $k = 1000$ on Flower.

Comparison of top-k intersection, 1-LENS-prec@k and 1-LENS-recall@k. Figure 4 shows that top-k intersection penalizes IG even for small, local changes. 1-LENS-prec@k and 1-LENS-recall@k values are always higher in comparison across all datasets in our experiments. Moreover, on both MNIST and Fashion MNIST, 1-LENS-prec@k is roughly 2x higher (above 90%) compared to top-k intersection (near 40%). In other words, an attack may appear stronger under a weaker measure of attributional robustness, if it ignores locality. This increase clearly shows that the top-k attack of Ghorbani, Abid, and Zou (2019) appears to be weaker on these datasets as the proportional increase by using locality indicates that the attack is only creating a local change than previously thought. We can see that for MNIST, Fashion-MNIST and GTSRB for $< 20\%$ of the samples, the top-k attack was able to make changes larger than what 1-LENS@k could measure.

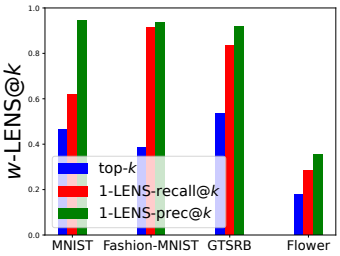

Figure 4: Attributional robustness of IG on naturally trained models measured as average top-k intersection, 1-LENS-prec@k and 1-LENS-recall@k between IG(original image) and IG(perturbed image) obtained by the top-t attack (Ghorbani, Abid, and Zou 2019) across different datasets.

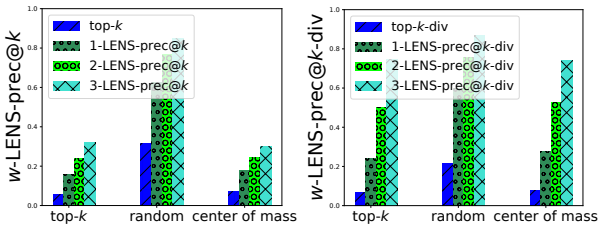

Figure 5: Effect of increasing w on average w-LENS-prec@k and w-LENS-prec@k-div in comparison with top-k intersection for IG map on ImageNet using a SqueezeNet model, when attacked with three attributional attacks (viz., top-k, random sign perturbation and mass center) of Ghorbani, Abid, and Zou (2019).

w-LENS-prec@k for varying w. In Figure 5(left) w-LENS-prec@k increases as we increase w to consider larger neighborhoods around the pixels with top attribution values. This holds for multiple perturbations, namely, top-t attack and mass-center attack by Ghorbani, Abid, and Zou (2019) as well as a random perturbation. Notice that the top-t attack of Ghorbani, Abid, and Zou (2019) is constructed specifically for the top-t intersection objective, and perhaps as a result, shows larger change when we increase local-sensitivity by increasing w in the robustness measure.

Due to space constraint and purposes of coherence, we present few results with IG here; we present similar results on other explanation methods in the Appendix E. Refer to Appendix E.2 for similar plots with random sign perturbation and mass center attack of Ghorbani, Abid, and Zou (2019). Appendix E.3 contains additional results with similar conclusions when Simple Gradients are used instead of Integrated Gradients (IG) for obtaining the attributions.

As a natural follow-up question we present in Appendix E.1 results obtained by modifying the similarity objective of top-k attack of Ghorbani, Abid, and Zou (2019) with 1-LENS-prec@k with the assumption to obtain a stronger attack. But surprisingly, we notice that it leads to a *worse* attributional attack, if we measure its effectiveness using the top-k intersections and 1-LENS-prec@k. In other words, attributional attacks against locality-sensitive measures of attributional robustness are non-trivial and may require funda-

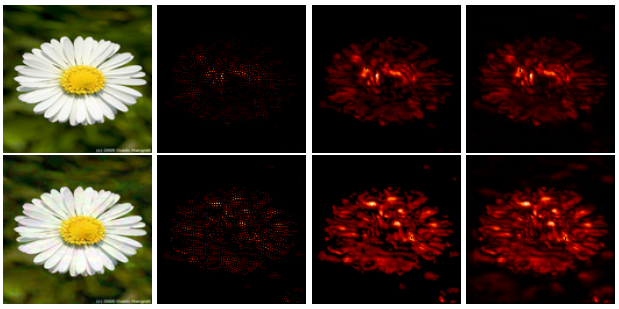

Figure 6: A sample image from Flower dataset before (top) and after (bottom) the top-k attributional attack of Ghorbani, Abid, and Zou (2019) on a ResNet model. For both, we show from left to right: the image, its top-k diverse pixels as per IG, the union of 3×3-pixel neighborhoods and 5×5-pixel neighborhoods of the top-k diverse pixels, respectively, for $k = 1000$. Quantitatively, improved overlap is captured by top-k-div intersection: 0.22, 1-LENS-recall@k-div: 0.87, 1-LENS-pre@k-div: 0.86, 2-LENS-recall@k-div: 0.95, 2-LENS-pre@k-div: 0.93. *Zoom in required to see the diverse pixels.*

3-LENS-prec@k metric(%)	**top-k metric(%)**
70.37	29.63

Table 1: Survey results showing % of humans able to relate an explanation map to the original image with or without noise using the Flower dataset based on a specific metric.

mentally different ideas.

3.3 Alignment of Attributional Robustness Metrics to Human Perception

We conducted a survey with human participants, where we presented images from the Flower dataset and a pair of attribution maps—an attribution map of the original image alongside an attribution map of their random perturbation or attributional attacked version Ghorbani, Abid, and Zou (2019), in a random order and without revealing this information to the participants. The survey participants were asked whether the two maps were relatable to the image and if one of them was different than the other. In Table 1 we summarize the results obtained from the survey. We simplify the choices presented to the user into 2 final categories - (1) Agree with w-LENS-prec@k (2) Agree with top-k metric Category (1) includes all results where the user found the maps the same, relatable to the image but dissimilar or the perturbed map was preferred over the original map. Category (2) was the case where the user preferred the original map over the perturbed map. Refer to Appendix I for more details.

4 Diverse Attribution for Robustness

Column 1 of Figure 6 shows a typical image from Flower dataset whose top-1000 pixels according to IG are concentrated in a small region. As seen in this illustrative example, when an image has multiple important parts, concentration of top attribution pixels in a small region increases vulnerability to attributional attacks. To alleviate this vulnerability, we propose post-processing any given attribution method to output top-k *diverse* pixels instead of just the top-k pixels with the highest attribution scores. We use a natural notion of w-diversity based on pixel neighborhoods, so that these diverse pixels can be picked by a simple greedy algorithm. Starting with $S \leftarrow \emptyset$, repeat for k steps: Pick the pixel of highest attribution score or importance outside S, add it to S and disallow the $(2w+1) \times (2w+1)$-pixel neighborhood around it for future selection. The set of k diverse pixels picked as above contains no two pixels within $(2w+1) \times (2w+1)$-pixel neighborhood of each other, and moreover, has the highest total importance (as the sum of pixel-wise attribution scores) among all such sets of k pixels. The sets of k pixels where no two pixels lie in $(2w+1) \times (2w+1)$-pixel neighborhood of each other form a matroid, where the optimality of greedy algorithm is well-known; see Korte and Lovász (1981).

Once we have the top-k diverse pixels as described above, we can extend our locality-sensitive robustness metrics from the previous section to w-LENS-prec@k-div and w-LENS-recall@k-div, defined analogously using the union of $(2w+1) \times (2w+1)$-pixel neighborhoods of top-k diverse pixels. In other words, define $\tilde{S}_k(x)$ as the top-k diverse pixels for image x and $\tilde{T}_k = \tilde{S}_k(x + \text{Att}(x))$, and use $\tilde{S}_k$ and $\tilde{T}_k$ to replace S_k and T_k used in Subsection 3.1.

For $k = 1000$, Figure 6 shows a sample image from Flower dataset before and after the top-k attributional attack of Ghorbani, Abid, and Zou (2019). Figure 6 visually shows the top-k diverse pixels in the Integrated Gradients (IG) and the union of their $(2w+1) \times (2w+1)$-pixel neighborhoods, for $w = \{1, 2\}$, for this image before and after the attributional attack. The reader may be required to zoom in to see the top-k diverse pixels. See Appendix F for more examples. Note that 0-LENS-prec@k and 0-LENS-recall@k are both the same and equivalent to top-k intersection. However, a combined effect of locality and diversity can show a drastic leap from top-k intersection value 0.14 to 2-LENS-recall@k-div value 0.95 (see Fig.3 and Fig.6). Fig. 5(right) shows the effect of increasing w on the w-LENS-prec@k-div metric on ImageNet.

5 A Stronger Model for Attributional Robustness

A common approach to get robust attributions is to keep the attribution method unchanged but train the models differently in a way that the resulting attributions are more robust to small perturbations of inputs. Chen et al. (2019) proposed the first defense against the attributional attack of Ghorbani, Abid, and Zou (2019). Wang et al. (2020) also find that IG-NORM based training of Chen et al. (2019) gives models that exhibit attributional robustness against the top-k attack of Ghorbani, Abid, and Zou (2019) along with adversarially trained models. Figure 7 shows a sample image from the Flower dataset, where the Integrated Gradients (IG) of the original image and its perturbation by the top-k attack are vi-

Figure 7: From left to right: a sample image from Flower dataset and Integrated Gradients (IG) before and after the top-k attributional attack of Ghorbani, Abid, and Zou (2019). The top row uses PGD-trained model whereas the bottom row uses IG-SUM-NORM-trained model.

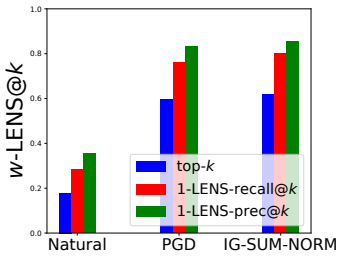

Figure 8: For Flower dataset, average top-k intersection, 1-LENS-prec@k, 1-LENS-recall@k measured between IG(original image) and IG(perturbed image) for models that are naturally trained, PGD-trained and IG-SUM-NORM trained. The perturbation used is the top-t attack of (Ghorbani, Abid, and Zou 2019). Note top-k is equivalent to 0-LENS-prec@k, 0-LENS-recall@k.

sually similar for models that are either adversarially trained (trained using Projected Gradient Descent or PGD-trained, as proposed by (Madry et al. 2018)) or IG-SUM-NORM trained as in Chen et al. (2019). In other words, these differently trained model guard the sample image against the attributional top-k attack. Recent work by Nourelahi et al. (2022) has empirically studied the effectiveness of adversarially (PGD) trained models in obtaining better attributions, e.g., Figure 7(center) shows sharper attributions to features highlighting the ground-truth class.

Figure 8 shows that PGD-trained and IG-SUM-NORM trained models have more robust Integrated Gradients (IG) in comparison to their naturally trained counterparts, and this holds for the previously used measures of attributional robustness (e.g., top-k intersection) as well as the new locality-sensitive measures we propose (e.g., 1-LENS-prec@k, 1-LENS-recall@k) across all datasets in Chen et al. (2019) experiments (Refer Appendix E.2 and E.3). The top-k attack of Ghorbani, Abid, and Zou (2019) is not a threat

Train	Attribution	top-k	3-LR@k	3-LR@k-div
Nat	SG	0.3825	0.7875	0.8290
Nat	DeepLIFT	0.2907	0.7641	0.8504
Nat	IG	0.2638	0.7148	0.8380
PGD	SG	0.1725	0.7245	0.8004
PGD	DeepLIFT	0.5572	0.9746	0.8977
PGD	IG	0.1947	0.7335	0.8584

Table 2: Average top-k intersection, 3-LENS-prec@k(3-LR@k) and 3-LENS-prec@k-div(3-LR@k-div) for random sign perturbation attack applied to different attribution methods on ImageNet for naturally and adversarially(PGD)-trained ResNet50 models.

to IG if we simply measure its effectiveness using 1-LENS-prec@k (Appendix E.2, E.3 for MNIST, Fashion MNIST and GTSRB). The above observation about robustness of Integrated Gradients (IG) for PGD-trained and IG-SUM-NORM trained models holds even when we use 1-LENS-Spearman and 1-LENS-Kendall measures to quantify the attributional robustness to the top-k attack of Ghorbani, Abid, and Zou (2019), and it holds across the datasets used by Chen et al. (2019) in their study; see Appendix E.

Chalasani et al. (2020) show theoretically that ℓ_∞-adversarial training (PGD-training) leads to stable Integrated Gradients (IG) under ℓ_1 norm. They also show empirically that PGD-training leads to sparse attributions (IG & DeepSHAP) when sparseness in measured indirectly as the change in Gini index. Our empirical results extend their theoretical observation about stability of IG for PGD-trained models, as we measure local stability in terms of both the top attribution values and their positions in the image.

Table 2 obtains the top-k intersection, 3-LENS-recall@k, and 3-LENS-recall@k-div of different attribution methods on ImageNet for naturally trained and PGD-trained ResNet50 models. We observe that for random sign attack the improvement obtained on top-k intersection is reduced for a large dataset like ImageNet. Still our conclusions about locality and diversity in attribution robustness in comparison with the top-k intersection baseline holds as we observe improvements in using diversity and locality. More results about incorporating diversity in the attribution and the resulting robustness metrics are available in Appendix H.

6 Conclusion and Future Work

We show that the fragility of attributions is an effect of using fragile robustness metrics such as top-k intersection that only look at the rank order of attributions and fail to capture the locality of pixel positions with high attributions. We highlight the need for locality-sensitive metrics for attributional robustness and propose natural locality-sensitive extensions of existing metrics. We introduce another method of picking diverse top-k pixels that can be naturally extended with locality to obtain improved measure of attributional robustness. Theoretical understanding of locality-sensitive metrics of attributional robustness, constructing stronger attributional attacks for these metrics, and using them to build attributionally robust models are important future directions.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I. J.; Hardt, M.; and Kim, B. 2018. Sanity Checks for Saliency Maps. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 9525–9536.

Agarwal, C.; Johnson, N.; Pawelczyk, M.; Krishna, S.; Saxena, E.; Zitnik, M.; and Lakkaraju, H. 2022. Rethinking Stability for Attribution-based Explanations. *arXiv preprint arXiv:2203.06877*.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.

Arun, N.; Gaw, N.; Singh, P.; Chang, K.; Aggarwal, M.; Chen, B.; et al. 2020. Assessing the (Un) Trustworthiness of saliency maps for localizing abnormalities in medical imaging. arXiv. *Preprint]*.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. In:PloS One 10.7 (2015), e0130140.

Bhatt, U.; Weller, A.; and Moura, J. M. 2020. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion Attacks against Machine Learning at Test Time. *Lecture Notes in Computer Science*, 387–402.

Boggust, A.; Suresh, H.; Strobelt, H.; Guttag, J. V.; and Satyanarayan, A. 2022. Beyond Faithfulness: A Framework to Characterize and Compare Saliency Methods. *CoRR*, abs/2206.02958.

Carvalho, D. V.; Pereira, E. M.; and Cardoso, J. S. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8): 832.

Chalasani, P.; Chen, J.; Chowdhury, A. R.; Wu, X.; and Jha, S. 2020. Concise Explanations of Neural Networks using Adversarial Training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 1383–1391. PMLR.

Chattopadhyay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. 839–847.

Chen, J.; Wu, X.; Rastogi, V.; Liang, Y.; and Jha, S. 2019. Robust Attribution Regularization.

Fagin, R.; Kumar, R.; and Sivakumar, D. 2003. Comparing Top k Lists. *SIAM Journal on Discrete Mathematics*, 17(1): 134–160.

Fan, F.; Xiong, J.; Li, M.; and Wang, G. 2021. On Interpretability of Artificial Neural Networks: A Survey. *arXiv:2001.02522 [cs, stat]*.

Fel, T.; Vigouroux, D.; Cadène, R.; and Serre, T. 2022. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 720–730.

Gade, K.; Geyik, S. C.; Kenthapadi, K.; Mithal, V.; and Taly, A. 2020. Explainable AI in industry: practical challenges and lessons learned: implications tutorial. In Hildebrandt, M.; Castillo, C.; Celis, L. E.; Ruggieri, S.; Taylor, L.; and Zanfir-Fortuna, G., eds., *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, 699. ACM.

Ghorbani, A.; Abid, A.; and Zou, J. Y. 2019. Interpretation of Neural Networks Is Fragile.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.

Iandola, F. N.; Moskewicz, M. W.; Ashraf, K.; Han, S.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR*, abs/1602.07360.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, 2668–2677. PMLR.

Korte, B.; and Lovász, L. 1981. Mathematical structures underlying greedy algorithms. In Gécseg, F., ed., *Fundamentals of Computation Theory*, 205–209. Springer Berlin Heidelberg. ISBN 978-3-540-38765-7.

Lakkaraju, H.; Arsov, N.; and Bastani, O. 2020. Robust and Stable Black Box Explanations. In *International Conference on Machine Learning*, 5628–5638.

Lecue, F.; Guidotti, R.; Minervini, P.; and Giannotti, F. 2021. 2021 Explainable AI Tutorial. https://xaitutorial2021.github.io/. Visited on 14-09-2021.

Lipton, Z. C. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10): 36–43.

Lundberg, S. M.; and Lee, S. 2017. A Unified Approach to Interpreting Model Predictions.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks.

Molnar, C. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Nourelahi, M.; Kotthoff, L.; Chen, P.; and Nguyen, A. 2022. How explainable are adversarially-robust CNNs? arXiv:2205.13042.

Oh, C.; and Jeong, J. 2020. VODCA: Verification of Diagnosis Using CAM-Based Approach for Explainable Process Monitoring. *Sensors*, 20(23): 6858.

Oviedo, F.; Ferres, J. L.; Buonassisi, T.; and Butler, K. T. 2022. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *Accounts of Materials Research*, 3(6): 597–607.

Ozdag, M. 2018. Adversarial Attacks and Defenses Against Deep Neural Networks: A Survey. *Procedia Computer Science*, 140: 152–161.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K., eds. 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer. ISBN 978-3-030-28953-9.

Sarkar, A.; Sarkar, A.; and Balasubramanian, V. N. 2021. Enhanced Regularizers for Attributional Robustness.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 618–626.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2016. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. 1605.01713.

Silva, S. H.; and Najafirad, P. 2020. Opportunities and Challenges in Deep Learning Adversarial Robustness: A Survey. *CoRR*, abs/2007.00753.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.

Singh, M.; Kumari, N.; Mangla, P.; Sinha, A.; Balasubramanian, V. N.; and Krishnamurthy, B. 2020. Attributional Robustness Training Using Input-Gradient Spatial Alignment.

Slack, D.; Hilgard, A.; Lakkaraju, H.; and Singh, S. 2021a. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*.

Slack, D.; Hilgard, A.; Singh, S.; and Lakkaraju, H. 2021b. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems*, 9391–9404.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F. B.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. 1706.03825.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. A. 2015. Striving for Simplicity: The All Convolutional Net.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328. PMLR.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks.

Tang, S.; Ghorbani, A.; Yamashita, R.; Rehman, S.; Dunnmon, J. A.; Zou, J. Y.; and Rubin, D. L. 2021. Data Valuation for Medical Imaging Using Shapley Value: Application on A Large-scale Chest X-ray Dataset. *Scientific Reports(Nature Publisher Group)*.

Tomsett, R.; Harborne, D.; Chakraborty, S.; Gurram, P.; and Preece, A. D. 2020. Sanity Checks for Saliency Metrics. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 6021–6029. AAAI Press.

Wang, F.; and Kong, A. W. 2022. Exploiting the Relationship Between Kendall's Rank Correlation and Cosine Similarity for Attribution Protection. *CoRR*, abs/2205.07279.

Wang, Z.; Wang, H.; Ramkumar, S.; Mardziel, P.; Fredrikson, M.; and Datta, A. 2020. Smoothed Geometry for Robust Attribution.

Yap, M.; Johnston, R. L.; Foley, H.; MacDonald, S.; Kondrashova, O.; Tran, K. A.; Nones, K.; Koufariotis, L. T.; Bean, C.; Pearson, J. V.; Trzaskowski, M.; and Waddell, N. 2021. Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Scientific Reports(Nature Publisher Group)*.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In *Proceedings of The European Conference on Computer Vision (ECCV)*.

Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; and Fergus, R. 2010. Deconvolutional networks. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, 2528–2535. IEEE Computer Society.

Zhang, Y.; Tiňo, P.; Leonardis, A.; and Tang, K. 2020. A Survey on Neural Network Interpretability. *arXiv:2012.14261 [cs]*.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2921–2929. IEEE Computer Society.