Rethinking Robustness of Model Attributions

Sandesh Kamath¹, Sankalp Mittal¹, Amit Deshpande², Vineeth N Balasubramanian¹
¹Indian Institute of Technology, Hyderabad
²Microsoft 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-
measure based on symmetric set difference. We also introduce a new
of attributions along with their rank order. We show that such
provements of the above metrics that incorporate the locality
this important issue, we propose locality-sensitive im-
have a focused discussion and analysis). Beyond highlight-
(in particular, we study image-based attribution methods to
and methods to study the robustness of model attributions
hence focus our efforts in this work on rethinking metrics
true indicator of the robustness of a model's attributions, and
attack as existing benchmark methods. This may not be a
that such overpenalization leads to a wrong sense of robust-
ness as the changes are within the object of importance.
ods, can be highly sensitive to minor local changes in attribu-
tions (see Fig 1 row 2). We, in fact, show (in Appendix B.1)
that under existing metrics to evaluate robustness of attribu-
tions, 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 improve-
ments 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 met-
rics, namely, LENS-top-k, LENS-Spearman and LENS-
Kendall, that incorporate the locality of attributions along
with their rank order. Besides avoiding overpenalizing at-
tribution methods for minor local drifts, we show that our
proposed LENS variants are well-motivated by metrics de-
defined on the space of attributions.
• We subsequently introduce a second measure based on di-
versity 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 attri-
butional 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 attribu-
tions using locality and diversity.
• Finally, we also show that existing methods for robust at-
tributions implicitly support such a locality-sensitive met-
ric 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 explaina-
tion/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 neu-
rons (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 (Shriku-
mur 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 fo-
cus 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 attri-
bution methods proposed has also led to efforts on identify-
ing 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; Bog-
gust et al. 2022; Agarwal et al. 2022). A key desired trait
that has been highlighted by many of these efforts is ro-
bustness or stability of attributions, i.e., the explanation
should not vary significantly within a small local neigh-
brhood of the input (Alvarez-Melis and Jaakkola 2018;
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 \(\rho\) and Kendall’s \(\tau\). 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 \(\ell_\infty\)-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.](image)

### 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-
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 highest attribution values. Let \( N_w(i, j) = \{(p, q) : i - w \leq p \leq i + w, j - w \leq q \leq j + w\} \) 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 \( N_w(S_k(x)) = \bigcup_{(i,j) \in S_k(x)} N_w(i, j) \), 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 \( \text{Att}(\cdot) \), let \( T_k = S_k(x + \text{Att}(x)) \) denote the top-\( k \) pixels in attribution values after applying the attributional perturbation \( \text{Att}(x) \). 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 \( \rho \) (LENS-Spearman) and Locality-sENsitive Kendall’s \( \tau \) (LENS-Kendall) metrics as rank correlation coefficients for the smoothed ranking orders according to \( \tilde{a}_{ij}(x) \)'s and \( \tilde{a}_{ij}(x + \text{Att}(x)) \)'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 \( a_1 \) and \( a_2 \) 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 \( a_1 \) and \( a_2 \), respectively. We define a locality-sensitive top-\( k \) distance between two attribution vectors \( a_1 \) and \( a_2 \) as

\[
d_k^{(w)}(a_1, a_2) \overset{\text{def}}{=} \frac{|S_k \triangle T_k|}{k},
\]

where \( \ell \) is the total number of pixels in attribution valuation or importance assigned to the \( k \)-th pixel. By a slight abuse of notation, we denote the top-\( k \)-th pixel. By a slight abuse of notation, we denote the top-\( k \) pixels.

**Proposition 1.** For any \( w_1 \leq w_2 \), we have \( d_k^{(w_1)}(a_1, a_2) \leq d_k^{(w_2)}(a_1, a_2) \leq \frac{|S_k \triangle T_k|}{k} \), where \( \triangle \) denotes the symmetric set difference, i.e., \( A \triangle B = (A \setminus B) \cup (B \setminus A) \).

Combining \( d_k^{(w)}(a_1, a_2) \) across different values of \( k \) and \( w \), we can define a distance

\[
d(a_1, a_2) = \sum_{k=1}^{\infty} \alpha_k \sum_{w=0}^{\infty} \beta_w d_k^{(w)}(a_1, a_2),
\]

where \( \alpha_k \) and \( \beta_w \) 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 (Fanin, Kumar, and Sivakumar 2003) based on symmetric set difference of top-\( k \) ranks to compare two rankings.

**Proposition 2.** \( d(a_1, a_2) \) defined above is upper-bounded by \( u(a_1, a_2) \) given by

\[
u(a_1, a_2) = \sum_{k=1}^{\infty} \alpha_k \sum_{w=0}^{\infty} \beta_w \frac{|S_k \triangle T_k|}{k},
\]

and \( u(a_1, a_2) \) defines a bounded metric on the space of attribution vectors.

Note that top-\( k \) intersection, Spearman’s \( \rho \) and Kendall’s \( \tau \) 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.

\[
a_{ij}^{(w)}(x) = \frac{1}{(2w+1)^2} \sum_{(p,q) \in N_w(i,j), 1 \leq p, q \leq n} a_{pq}(x)
\]

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 \( w \geq 0 \),

\[
\|a^{(w)}(x) - a^{(w)}(y)\|_2 \leq \|a(x) - a(y)\|_2.
\]

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

![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 \times 3-pixel neighborhoods and 5 \times 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.](image-url)
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 $\ell_2$ distance between the simple gradients of nearby points $\|x - y\|_2 \leq \delta \Rightarrow \|\nabla_x f(x) - \nabla_y f(y)\|_2 \leq \delta \lambda_{\text{max}}(H_x(f))$, where $H_x(f)$ is the Hessian of $f$ w.r.t. the input at $x$ and $\lambda_{\text{max}}(H_x(f))$ is its maximum eigenvalue. By Proposition 3 above, the same continues to hold for $w$-smoothed gradients, i.e., $\left\|\tilde{\nabla}_w^a(x) - \tilde{\nabla}_w^b(x)\right\|_2 \leq \delta \lambda_{\text{max}}(H_x(f))$. 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 $1 - \text{prec}^w(a_1, a_2)$ and $1 - \text{recall}^w(a_1, a_2)$, respectively, where $a_1$ is the attribution of the original image and $a_2$ is the attribution of the perturbed image. For rank correlation coefficients such as Kendall’s $\tau$ and Spearman’s $\rho$, we compute $w$-LENS-Kendall and $w$-LENS-Spearman as the same Kendall’s $\tau$ and Spearman’s $\rho$ but computed on the locality-sensitive $w$-smoothed attribution map $\tilde{a}^w$ instead of the original attribution map $a$. 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 $\ell_\infty$ 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-prec@$k$ could measure.

#### $w$-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-
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 $S_k(x)$ 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 $T_k$ 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-

---

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.

<table>
<thead>
<tr>
<th>3-LENS-prec@k metric(%)</th>
<th>top-k metric(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70.37</td>
<td>29.63</td>
</tr>
</tbody>
</table>

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 $3 \times 3$-pixel neighborhoods and $5 \times 5$-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.
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.

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.

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-attack of Ghorbani, Abid, and Zou (2019). Note top-k is equivalent to 0-LENS-prec@k, 0-LENS-recall@k.

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-attack of Ghorbani, Abid, and Zou (2019). Note top-k is equivalent to 0-LENS-prec@k, 0-LENS-recall@k.

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


