Attribution Quality Metrics with Magnitude Alignment

Chase Walker¹, Dominic Simon¹, Kenny Chen², Rickard Ewetz¹

¹Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, USA
²Lockheed Martin, Orlando, FL, USA

{chase.walker, dominic.simon, rickard.ewetz}@ucf.edu, kenny.chen@lmco.com

Abstract
Attribution algorithms play an instrumental role in human interpretation of AI models. The methods measure the importance of the input features to the model output decision, which can be displayed as an attribution map for image classifiers. Perturbation tests are the state-of-the-art approach to evaluate the quality of an attribution map. Unfortunately, we observe that perturbation tests fail to consider attribution magnitude, which translates into inconsistent quality scores. In this paper, we propose Magnitude Aligned Scoring (MAS), a new attribution quality metric that measures the alignment between the magnitude of the attributions and the model response. In particular, the metric accounts for both the relative ordering and the magnitude of the pixels within an attribution. In the experimental evaluation, we compare the MAS metric with existing metrics across a wide range of models, datasets, attributions, and evaluations. The results demonstrate that the MAS metric is 4× more sensitive to attribution changes, 2× more consistent, and 1.6× more invariant to baseline modifications. Our code and the referenced appendix are publicly available via https://github.com/chasewalker26/Magnitude-Aligned-Scoring.

1 Introduction
Understanding the decision making of black-box AI models is necessary for deployment in safety-critical domains. Attribution methods are currently the most prevalent form of explanations [Das and Rad, 2020]. These methods provide model explanations by assigning an importance value to each feature of the model input [Simonyan et al., 2014]. A broad range of attribution algorithms have been developed in the past few years [Hooker et al., 2019a; Chefer et al., 2021]. However, the ability to quantify the quality of an attribution map remains an open problem. An accurate attribution metric has the potential to further advance the field of explainable AI forward, refining human understanding of AI models.

Attribution quality metrics generally fall into two categories: evaluation with ground-truth [Borji et al., 2013] or perturbation metrics without ground-truth [Samek et al., 2016; Ancona et al., 2018]. Evaluation of attribution methods with ground-truth can be desirable when human-labeled data is present. Common methods to develop ground-truth datasets in the vision domain are the use of segmentation algorithms to create a mask of the image subject, human eye-tracking heat map data, or manual image masks created by humans [Kümmerer et al., 2014; Bylinskii et al., 2019]. Perturbation metrics aim to quantify the quality of an attribution map without the use of ground-truths that are often not available [Samek et al., 2016]. Moreover, human-created masks do not necessarily represent a model’s decision process.

Perturbation metrics aim to evaluate if the attribution map is reflective of a model’s decision making [Petsiuk et al., 2018b]. Specifically, perturbation tests aim to measure if the attribution map discriminates appropriately between more and less important features. The tests are performed by modifying the model or the input test image by the values in the attribution map [Petsiuk et al., 2018b]. Ideally, the magnitude of the attribution assigned to each feature should be proportional to the model response. However, we observe that current image perturbation metrics do not measure this relationship, they only account for the relative ordering of the input features, not the magnitudes, leading to inconsistent scores.

In this paper, we propose a new perturbation metric for evaluating attribution quality without ground-truth called Magnitude Aligned Scoring (MAS). MAS employs an alignment penalty to measure the relationship between attribution magnitude and the response of the model output to the attribution. Our main contributions can be summarized as follows:

- We observe that existing quality metrics fail to consider attribution magnitude and only rely on attribution order, which leads to inconsistent quantification.
- We propose a quantitative, ground-truth-free perturbation metric, MAS. It provides a principled solution to the failures of existing methods by utilizing an alignment penalty to satisfy a new sensitivity property we define.
- Quantitative analysis proves MAS solves the failures of existing metrics, showing 2.5× improvements across sensitivity, consistency, and baseline invariance testing.

The paper is arranged as follows: related work is discussed in Section 2, the MAS metric is motivated in Section 3, the MAS metric is defined in Section 4, experimental evaluation is performed in Section 5, and the conclusion is in Section 6.
2 Related Work

In this section, we discuss the details of attribution methods, the existing metrics that quantify their accuracy, and how these metrics can be validated.

2.1 Attribution Methods

Attribution methods explain black-box models by measuring the importance of each feature in an input to the model response. Attribution methods can be occlusion based [Ribeiro et al., 2016; Zeiler and Fergus, 2014], gradient based [Simonyan et al., 2014], or attention based methods [Hao et al., 2021] for transformer models. Occlusion based methods iteratively modify an input image while measuring model output to determine the most salient regions. Gradient based methods use backpropagation to measure the model gradients with respect to the input features. Attention based methods use transformer attention weights as explanations directly, or combine them with model gradients [Yuan et al., 2021]. Occlusion based methods are slow and generally undesirable.

The early gradient based techniques utilize model gradients as is [Simonyan et al., 2014] or multiply them by the input image [Shrikumar et al., 2016], but these methods suffer from large amounts of saturation noise [Sundararajan et al., 2017]. Recently, the newly introduced path integration methods [Miglani et al., 2020] average the gradients from multiple interpolated images along a path from a baseline to the input image to reduce this saturation noise. Additional techniques to reduce noise suppress negative gradients during the backpropagation step [Springenberg et al., 2015] or measure gradients from a model’s last convolutional layer and remove gradients pointing to non-target classes [Selvaraju et al., 2017]. Attention based methods first started with raw attention as explanations directly, or combine them with model gradients [Yuan et al., 2021]. Occlusion based methods are slow and generally undesirable.

2.2 Attribution Quality Metrics

Attribution quality metrics aim to evaluate how well an attribution represents a model. Attribution metrics are necessary to measure how well an attribution represents a model.

2.3 Desirable Attribution Metric Properties

A trustworthy and reliable attribution quality metric should adhere to the following desirable properties which are quantitatively measurable:

1. Sensitivity: Features important to the model should have high attribution and unimportant features should have low attribution [Petsiuk et al., 2018b].

2. Consistency: A metric should be consistent in its calculated ratings. It should consistently rank different attribution methods by their quality over a set of varying inputs [Tomsett et al., 2020].

3. Baseline Invariance: A metric should be invariant to its baseline selection, i.e., an insertion test using a random baseline or blurred baseline should rank a set of attributions the same way [Tomsett et al., 2020].

Existing image perturbation metrics attempt to satisfy sensitivity by scoring attributions using the relative ordering of the input features. However, the sensitivity property is rather vaguely defined, so it is not clear which one of the RISE and PIC metrics best quantifies the property. Hence, a mix of the different tests are typically used to evaluate new attribution algorithms. Additionally, it has been shown that consistency and baseline invariance are not adequately satisfied by the RISE and PIC methods [Tomsett et al., 2020].
Formally, let \( s_i \) and \( s_j \) refer to the relative distance to zero. Small and large attributions should be assigned small (large) magnitude attributions.

In Section 2.3, the sensitivity property states that input pixels that are important (not important) to the model’s decision should be assigned large (small) magnitude attributions.

In Figure 1(a), we modify guided backpropagation (GBP) [Springenberg et al., 2015] attributions by adding 5% or 10% \( \times \text{max}(GBP) \) to all attribution pixels as a constant. Since adding a constant does not change the relative ordering of the input features, the attribution maps are scored equally by SIC. This holds for all four reviewed metrics - RISE insertion and deletion as well as PIC’s SIC and AIC. Thus, it is clear the existing attribution metrics’ sole reliance on relative attribution ordering leads to an invariance to a constant offset.

3 Motivating a New Metric

We now study the RISE and PIC metrics to show their failure of sensitivity resulting from their disregard of attribution magnitude and motivate the definition of a new sensitivity property as well as the development of a new metric.

3.1 Attribution Magnitude Is Important

In Section 2.3, the sensitivity property states that input pixels that are important (not important) to the model’s decision should be assigned large (small) magnitude attributions. Here, small and large refers to the relative distance to zero. Formally, let \( A_i \) and \( A_j \) denote the attributions of pixels \( i \) and \( j \), respectively. The quotient \( A_i / A_j \) measures the relationship of the features’ magnitudes, which defines sensitivity.

Given this notation, we analyze the invariance of attribution metrics to multiplying an attribution map by a constant or adding a constant. It is straightforward to understand that invariance to multiplication is desirable for a metric, as \( A_i / A_j \) is equal to \( aA_i / aA_j \), where \( a \) is a constant. However, \( (A_i + a) / (A_j + a) \) is not equal to \( A_i / A_j \), in general. Therefore, invariance to a constant offset is undesirable under the sensitivity property, but existing metrics do not adhere to this.

3.2 Limitations of State of the Art Metrics

Increase by a Constant Is Ignored

In Figure 1(a), we modify guided backpropagation (GBP) [Springenberg et al., 2015] attributions by adding 5% or 10% \( \times \text{max}(GBP) \) to all attribution pixels as a constant. Since adding a constant does not change the relative ordering of the input features, the attribution maps are scored equally by SIC. This holds for all four reviewed metrics - RISE insertion and deletion as well as PIC’s SIC and AIC. Thus, it is clear the existing attribution metrics’ sole reliance on relative attribution ordering leads to an invariance to a constant offset.

Theorem 1. RISE and PIC Are Invariant to Constant Offset.

Proof. Let an attribution \( A \) with values in the range \([0, 1]\) have a magnitude ordering of \( O_A \). If a constant \( b \) is added to \( A \), this yields \( A' \), with range \([b, 1 + b]\) and ordering \( O_{A'} = O_A \). Since a score is determined solely by the perturbation of the input image via the order of \( A \), and \( O_A = O_{A'} \), \( A \) and \( A' \) will have equal scores. Therefore, the metrics are invariant to a constant offset.

Attribution Noise Is Not Penalized

We show the attribution maps of a “warplane” image computed using left integrated gradients (LIG) [Miglani et al., 2020] and GBP in Figure 1(b). Although GBP has a very sharp, low noise attribution compared to LIG, which has evident saturation noise [Sundararajan et al., 2017], the RISE insertion metric scores LIG higher than GBP. Given the attributions are nearly identical except for the noise in LIG, it is clear that existing metrics do not sufficiently penalize non-zero attributions on unimportant input features. This is a direct result of not adhering to the sensitivity property.

3.3 Proportional Sensitivity

In this paper, we propose a quantitatively satisfiable property in replacement of the vaguely defined sensitivity property:

1. **Proportional Sensitivity:** The magnitude of the attribution assigned to an input feature should be directly proportional to the change the feature induces in the model output response.

To satisfy this property, we propose a new metric that penalizes the misalignment of an attribution feature’s model response and density response. This encourages large magnitude attributions to be assigned to the critical input features while penalizing non-zero attributions assigned to pixels that are irrelevant to the model. We envision this new metric will drive the development of new attribution algorithms that produce sharp, low noise attribution maps.

4 The MAS Metric

In this section, we introduce the magnitude aligned scoring (MAS) metric. We define the model and density response to introduce an alignment penalty that satisfies the proportional sensitivity property. The model response measures a feature’s proportional contribution to the model output and the density...
response measures a feature’s proportional contribution to the entire attribution’s magnitude. Figure 2 provides an overview.

4.1 The Model and Density Response

Given an input $X$ of class $c$, a model $F$, and an attribution map $A$, $N$ attribution features are evaluated over an $N$ step perturbation test. For a step $k$, the perturbed image $X_k$ evaluates the impact of the first $k$ attribution features in highest magnitude order. This test follows the RISE process from Section 2.2. We define the model response ($MR$) at step $k$ as:

$$MR_k = \text{softmax}(F(X_k))_c,$$

where $X_0$ represents the unperturbed starting image and $MR_0$ to $MR_N$ forms the $MR$ curve. In Figure 2, we show an input image of a “bee eater” (a), its integrated gradients (IG) [Sundararajan et al., 2017] attribution map (b), and the model response (c) from an MAS insertion test. We note that the insertion model response $MR^{ins}$ is a monotonically increasing curve, and $MR^{del}$ is a monotonically decreasing curve.

Next, we define the density response ($DR$) at step $k$ as:

$$DR_k = \sum_{i=0}^{k} \frac{|A_i|}{|A|},$$

where the $|.|$ operation measures the total magnitude of the attributions in a given feature, $A_0$ represents 0 selected features, and $DR_0$ to $DR_N$ forms a full density response curve. The density response of the “bee eater” IG attribution is seen in Figure 2(d) and represents, at each step, what percentage of the total attribution magnitude has been selected via the perturbation process. We note that the insertion model response $DR^{ins}$ is a monotonically increasing curve, and $DR^{del} = 1 - DR^{ins}$ is a monotonically decreasing curve.

Now we define the alignment penalty ($AP$) which measures the absolute value of the difference between the model and density response for an attribution map. The general alignment penalty at step $k$ is defined as:

$$AP_k = |MR_k - DR_k|,$$

where $AP_0$ to $AP_N$ measures the alignment penalty across the full insertion or deletion $MR$ and $DR$ curves. In Figure 2(e), the insertion alignment penalty is illustrated as the area between the $MR$ and $DR$ curves.

4.2 Magnitude Aligned Scoring (MAS)

Given an attribution map $A$ with $N$ features, MAS can be utilized as insertion or deletion. The insertion and deletion tests are defined by the AUC of the $MR$ and $AP$ curves:

$$\text{MAS}^{ins} = \frac{1}{N} \sum_{i=0}^{N} MR_i^{ins} - \frac{1}{N} \sum_{i=0}^{N} AP_i^{ins},$$

and

$$\text{MAS}^{del} = \frac{1}{N} \sum_{i=0}^{N} MR_i^{del} + \frac{1}{N} \sum_{i=0}^{N} AP_i^{del}.$$

Intuitively, lowering (increasing) the insertion (deletion) score is the effective application of the alignment penalty because a higher (lower) insertion (deletion) score represents a higher quality attribution.

To calculate the $MR$, we perform the linear perturbation process from RISE explained in Section 2.2 and we perform monotonic normalization of the $MR$ to $[0, 1]$ to ensure $AP = 0$ is achievable as $DR$ is on the range $[0, 1]$. The penalized $\text{MAS}^{ins}$ or $\text{MAS}^{del}$ ROC is then clipped to the range $[0, 1]$ and normalized to $[0, 1]$. Therefore, the AUC of the resulting $\text{MAS}^{ins}$ and $\text{MAS}^{del}$ scores is on the range $[0, 1]$ where higher is better for $\text{MAS}^{ins}$ and lower for $\text{MAS}^{del}$.

We present how the MAS insertion test accounts for attribution magnitude in Figure 3 by revisiting the LIG and GBP attributions of the “warplane” seen in Figure 1(b). From left to right, we show the input (a), an attribution, its $MR$, $DR$,
and AP graph, and its MAS score calculation beneath. In Figure 3(b) we see LIG has a large model response of 0.992 which is greater than GBP’s model response of 0.975 in Figure 3(c) (since magnitude is not a factor without the alignment penalty). However, due to the noise in the LIG attribution, it has a large alignment penalty of 0 which results in a lower score than GBP, as expected. The alignment penalty therefore corrects the non-penalization of attribution noise by the existing methods.

**Figure 3:** The calculation of the MAS insertion scores for (b) LIG and (c) GBP attributions of a “warplane” (a). Each graph shows the MR, DR, and AP. It can be seen LIG has a model response with a higher AUC than GBP (since magnitude is not a factor), but receives a much larger alignment penalty when noise is considered, resulting in a lower score than GBP, as expected. The alignment penalty therefore corrects the non-penalization of attribution noise by the existing methods.

**Insertion and Deletion Difference.** In the existing literature, insertion and deletion tests are often performed in unison, but considered separately as they measure different qualities of an attribution. The insertion test measures the value of high magnitude attributions to classification by adding them to an image, while the deletion test measures the value of high magnitude attributions to misclassification by removing them from an image. However, we propose a combination of the scores should also be considered to balance the bias of each test. We propose considering the difference of the scores:

\[
\text{MAS}^{\text{diff}} = \text{MAS}^{\text{ins}} - \text{MAS}^{\text{del}}. \tag{6}
\]

The subtraction of these two tests creates the new “difference” score while preserving their opposing nature: higher insertion scores and lower deletion scores are desired. If one score measures very well and one very poorly, the small difference will indicate the disagreement between the tests (one test is biased). If both scores are respectively strong, the large difference will indicate an overall high scoring attribution.

**4.3 MAS Is Sensitive to a Constant Offset**

**Theorem 2.** MAS is sensitive to a constant offset.

**Proof.** The MAS score of an attribution \( A \) is defined as:

\[
\text{MAS}(A) = MR - |MR - DR|, \tag{7}
\]

where \( MR \) is equivalent to the score of the RISE metrics. Following Proof (3.2), if \( \text{MAS}(A) \) and \( \text{MAS}(A') \) are evaluated, the \( MR \) terms of each function will be equivalent. So we write \( \text{MAS} = f(DR) \). Now, consider \( A \in [0, 1] \) and \( A' \in [b, 1+b] \). By Eq (2), \( DR(A) \neq DR(A') \), thus \( \text{MAS}(A) \neq \text{MAS}(A') \), proving MAS is sensitive to a constant offset, satisfying proportional sensitivity. \( \square \)

**5 Evaluation**

We perform evaluation of the proposed MAS metrics against the currently accepted PIC [Kapishnikov et al., 2019] and RISE [Petsiuk et al., 2018b] perturbation metrics for a total of eight metrics under evaluation: AIC, SIC, RISE insertion, RISE deletion, RISE difference, MAS insertion, MAS deletion, and MAS difference. We recognize difference was not employed in the original RISE framework, but we use it for fair comparison against our proposed MAS difference metric.

All evaluations are performed with PyTorch [Paszke et al., 2019], using ResNet 101 (R101) [He et al., 2016] and ViTBase 16 (ViT16) [Dosovitskiy et al., 2020]. The evaluations are executed on an internal cluster with NVIDIA A40 GPUs. We employ the Imagenet [Russakovsky et al., 2015] and RESISC45 [Cheng et al., 2017] datasets across our experiments. We use the respective repositories of the PIC [Kapishnikov et al., 2021a] and RISE [Petsiuk et al., 2018a] metrics.

**5.1 Metric Attribution Sensitivity Test**

As presented in Figure 1(a) and (b), the RISE and PIC metrics do not recognize a constant offset or properly penalize attribution noise because they do not consider attribution magnitude, and therefore are not sensitive as outlined in Section 2.3. We now quantitatively verify MAS’ sensitivity.

Using the ImageNet and RESISC45 datasets with the R101 and ViT16 models, we generate gradient (grad) [Simonyan et al., 2014], LIG, and GBP attributions for 1000 images from both datasets. We choose these attribution methods for their different levels of noise. We then perform one of two modifications to the generated attributions. We either add a constant that is 5, 10, 25, or 50% of the maximum attribution value as explained for Figure 1(a) or we add noise from the range 0 to 0.05, 1, 2, or 3% of the maximum attribution value.
We then average, over the images, the ROCs computed by each of the eight metrics for all attributions and their modifications. We then measure the absolute distance of the modified attribution ROCs from the original ROC. We evaluate a metric’s sensitivity by the ratio of the distance metric to the original attribution’s AUC. A sensitive metric will have a non-zero ratio value, where higher indicates more sensitivity.

In Table 1, we see only MAS is sensitive to a constant offset and MAS outperforms PIC and RISE in all noised sensitivity tests. MAS is overall more sensitive to changes in an attribution as it considers attribution magnitude via the alignment penalty. In Section A.1, we present the ROC curves from the table that show MAS reduces the score of the modified (worse) attributions, whereas PIC and RISE do not consistently increase or decrease the score if the score changes.

**5.2 Metric Ranking Consistency Test**

To evaluate the consistency of how a metric rates a group of attributions, we perform two metric sanity checks [Tomsett et al., 2020]: inter-rater reliability (IRR) and internal consistency reliability (ICR). IRR measures how well a metric sorts a set of attributions over a set of images. A perfectly consistent metric is expected to provide the same ordering of an attribution set over all images. Krippendorff’s α is used to measure the IRR of a metric [Krippendorff, 2004], where a higher α in the range [0, 1] represents a more consistent metric. We provide the ICR results in Appendix A.2.

We measure the IRR of the eight metrics using three attributions over 5000 images from the ImageNet and RESISC45 datasets. For the R101 model, we select the grad, LIG, and GBP attributions due to their large visual differences (see Figure 4). For the VIT16 model, we select the following attribution group: a random mask, LIG, and transition attention [Yuan et al., 2021] or raw attention [Hao et al., 2021] for ImageNet and RESISC45, respectively, due to their large visual differences. These were chosen as visually similar attributions are likely to be scored equivalently, making consistent ranking unlikely for any metric. The IRR test results are in

We measure the IRR of a metric using three attributions over a set of images. A more consistent metric is more trustworthy. We measure this consistency with the IRR metric, where a higher value is better.

![Figure 4: The attributions, sorted low to high by score, chosen for ImageNet using R101. Note the differences in pixel distribution and density of the attribution maps.](image-url)
Table 2 where MAS shows a significantly higher consistency than the RISE and PIC metrics, outperforming them in 11/12 tests. We illustrate the consistency improvements of MAS with Figure 5 and figures in Appendix A.2.

5.3 Metric Baseline Invariance Test

As explained in Section 2.3, it is desirable for a quality metric to consistently evaluate attributions regardless of its baseline. ICR measures how well two different metrics agree on attribution rankings via Spearman’s \( \rho \in [0, 1] \), where a higher value represents stronger agreement. We evaluate the agreement of the six RISE and MAS quality metrics with their modified baseline versions. We exclude PIC due to inaccessibility of modification. We use either random values drawn from a uniform distribution or the dataset mean value as the new baseline [Tomsett et al., 2020]. The default baseline for insertion is a blurred image and for deletion, a black image. We use the same image, model, and attribution choices from the previous section for evaluation. The results in Table 3 show that MAS is more invariant to baseline modification than RISE in 24/24 tests, as desired. A visual explanation of this test is provided with a figure in Appendix A.3.

5.4 Qualitative Analysis of MAS

Through the introduction of the alignment penalty to account for the relationship between attribution magnitude and model response, we define the MAS framework which properly evaluates attributions. We show quantitatively that the MAS metrics, unlike the existing state-of-the-art metrics, satisfy the three desired properties of attribution quality metrics: sensitivity, consistency, and invariance to baseline selection. This greatly improves the ability of a user to select a desired, high-performance attribution method. We believe MAS will be used to develop high-quality attribution methods. In future work, we intend to use the alignment penalty to refine existing attributions by removing unimportant features.

6 Conclusion

We discover that current state-of-the-art attribution quality metrics fail to consider attribution magnitude which leads to poor quantification as they fail invariance properties.
Acknowledgements

The authors were in part supported by Lockheed Martin Corporation and the Florida High Tech Corridor. This material is partially based on research sponsored by DARPA under agreement number #FA8750-23-2-0501 and #HR00112020002. This material is partially based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Award Number #DE-SC0023494. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, or DOE, or the U.S. Government.

References


[Paszke et al., 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and


