On the Faithfulness of Vision Transformer Explanations

Junyi Wu[1], **Weitai Kang**[1], **Hao Tang**[2], **Yuan Hong**[3], **Yan Yan**[1,†]
[1]Department of Computer Science, Illinois Institute of Technology, USA
[2]Robotics Institute, Carnegie Mellon University, USA
[3]Department of Computer Science, University of Connecticut, USA
{jwu125, wkang11}@hawk.iit.edu, haotang2@cmu.edu, yuan.hong@uconn.edu, yyan34@iit.edu

Abstract

*To interpret Vision Transformers, post-hoc explanations assign salience scores to input pixels, providing human-understandable heatmaps. However, whether these interpretations reflect true rationales behind the model's output is still underexplored. To address this gap, we study the faithfulness criterion of explanations: the assigned salience scores should represent the influence of the corresponding input pixels on the model's predictions. To evaluate faithfulness, we introduce **Sa**lience-guided Faithfulness **Co**efficient (**SaCo**), a novel evaluation metric leveraging essential information of salience distribution. Specifically, we conduct pair-wise comparisons among distinct pixel groups and then aggregate the differences in their salience scores, resulting in a coefficient that indicates the explanation's degree of faithfulness. Our explorations reveal that current metrics struggle to differentiate between advanced explanation methods and Random Attribution, thereby failing to capture the faithfulness property. In contrast, our proposed SaCo offers a reliable faithfulness measurement, establishing a robust metric for interpretations. Furthermore, our SaCo demonstrates that the use of gradient and multi-layer aggregation can markedly enhance the faithfulness of attention-based explanation, shedding light on potential paths for advancing Vision Transformer explainability.*

1. Introduction

The prevalent use of Transformers in computer vision underscores the imperative to demystify their black-box nature [10, 17, 48]. This presents a challenge to traditional post-hoc interpretation methods, which were primarily tailored for MLPs and CNNs [1, 12]. As a response, a growing line of work aimed at developing new explanation paradigms specific to Vision Transformers, where attention mechanisms play a dominant role [1, 4, 11, 12, 35]. By incorporating attention distributions, these explanation methods estimate salience scores *w.r.t.* the tokens extracted from input image patches. Subsequently, these scores are interpolated across pixel space, resulting in visually convincing heatmaps that align with human intuition [14].

However, recent works [16, 23] claimed that it is crucial to evaluate how accurately these interpretations reflect the true reasoning process of the Transformer model and termed this aspect as faithfulness. To evaluate the quality of post-hoc explanations, recent studies commonly adopt an ablation approach [50]. This involves perturbing input image pixels that are identified as most or least important by the explanation method under evaluation. For example, they perturb pixels with the highest salience scores and then observe whether there is a decrease in the model's accuracy, which serves as a surrogate examination [6, 13, 29, 33, 42]. Despite the prevalence of these strategies, our study reveals that they all overlook a proper evaluation of the degree of faithfulness. We advance our discussion by characterizing the **core assumption of faithfulness** that underpins explanation methods: the magnitude of salience scores signifies the level of anticipated impacts. Consequently, **(i)** input pixels assigned higher scores are expected to exert greater influence on the model's prediction, compared with those with lower scores, and **(ii)** two groups of pixels with a larger difference in salience scores are expected to cause a greater disparity in their influences on the model's prediction.

In response to these desiderata, for a thorough faithfulness metric, it is necessary to: **(i)** explicitly compare the influences of input pixels with different magnitudes of salience, and **(ii)** quantify the differences in salience scores to reflect the expected disparities in their impacts. However, existing metrics fall short in both aspects, as they rely on cumulative perturbation [41] and do not consider the information embedded in the distribution of the magnitude of salience scores. For example, with cumulative perturbation (see Figure 1), it is difficult to discern the influence of pixels ranked between the top 0-10% (the elephant's body) and 90-100% (the sky) of salience. This is because the removal

†Corresponding author

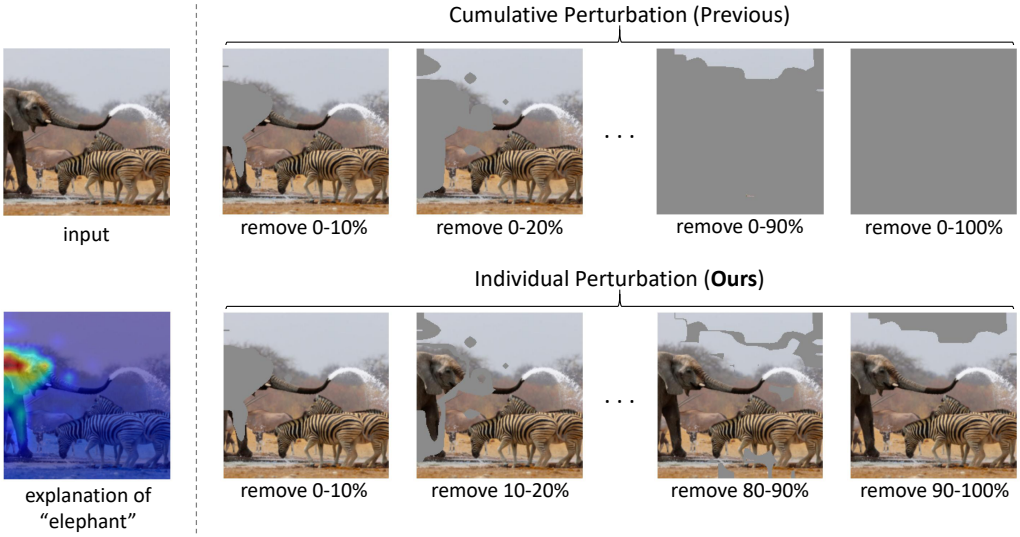

Figure 1. Explanation result and illustration of two perturbation manners: cumulative perturbation and our SaCo perturbation. Previous metrics perturb the pixel subsets cumulatively. In contrast, the SaCo perturbs them individually to directly compare their influences.

of the top 90-100% important pixels is only performed after the top 0-90% have been eliminated, which conflates their impacts. Moreover, without considering the exact values of salience, it is uncertain to what degree an explanation expects pixels in the top 0-10% to have more influence than those in 90-100%. Existing metrics cannot adequately evaluate an explanation's ability to differentiate importance levels among different pixels, thereby failing to validate the faithfulness core assumption. Such deficiency can lead to unreliable outcomes, underlining the need for a nuanced evaluation. For instance, it is alarming that commonly used metrics fail to distinguish between some state-of-the-art explanation methods and Random Attribution [50].

Recognizing that faithfulness is essential for explanation methods to depict models' behavior, we propose a novel evaluation framework, **Sa**lience-guided Faithfulness **Co**efficient, or **SaCo**, which analyzes how faithfully an explanation method aligns with the model's behavior. The proposed metric operates by conducting a statistical analysis of pixel subsets with varying salience scores and comparing their impacts on the model's prediction. The salience score distribution is evaluated based on its alignment with the true effect of corresponding pixels. For instance, if a pixel subset with higher salience scores significantly impacts the model's prediction more than a subset with lower scores, as anticipated, such a pair of subsets is deemed to satisfy the faithfulness criterion. Consequently, the disparity in salience scores between these two subsets, which represents the degree of expectation, will be positively accumulated to the measured outcome. Conversely, if a pair of subsets does not meet this expectation, it is identified as a violator and will have a negative contribution to the outcome. The SaCo is suitable for testing the core assumption validity, as it involves explicit comparisons among different pixels and captures the expected disparities in their impacts.

Experimental results across a range of datasets and Vision Transformer models in Section 5 demonstrate that current metrics ignore proper evaluations of faithfulness. Furthermore, we observe that most explanation methods for Vision Transformers actually underperform when tested against the core assumption. To investigate the key factors that affect faithfulness, we perform ablative experiments on attention-based explanation methods.

In summary, we state our contributions as follows: **(i)** We develop a new evaluation metric, SaCo, to assess how well explanations adhere to the core assumption of faithfulness. By comprehensive examination of ten representative explanation methods across three datasets and three Vision Transformer models, we demonstrate that SaCo can provide a complementary tool for evaluating how salience scores signify the level of anticipated impacts on the model. **(ii)** Empirically, we demonstrate SaCo's capability to distinguish meaningful explanations from Random Attribution, setting a useful and robust benchmark. **(iii)** We provide insights into certain designs in current attention-base explanation methods that may alter faithfulness. Our empirical results highlight the role of gradient information and aggregation rules, guiding the path for future improvements in Vision Transformer interpretability methodology.

2. Related Work

2.1. Post-hoc Explanations

Traditional post-hoc explanations. Traditional post-hoc explanation methods largely fall into two groups: gradient-based and attribution-based approaches. The first group in-

cludes methods like Input $\odot$ Gradient [43], SmoothGrad [44], Full Grad [45], Integrated Gradients [46], and Grad-CAM [39]. These methods leverage gradient information to calculate salience scores. In contrast, attribution-based methods [7, 20, 32, 42] propagate classification scores backward to the input and then use the resultant values as indicators of contribution. Beyond these two types, there are also other methods, such as saliency-based [31, 53, 54], Shapley additive explanation [30], and perturbation-based methods [18, 19]. Although initially designed for MLPs and CNNs, some of them have been successfully adapted for Vision Transformers in recent works [4, 12].

Leveraging attentions to interpret Vision Transformers. A distinct branch of research in post-hoc interpretability is committed to creating new paradigms specifically for Transformers. Attention maps are widely used in methodologies associated with this direction, as they inherently constitute distributions that represent the sampling weights over input tokens. Representative methods include Raw Attention [51] and Rollout [1], which regard attention maps as an explanation, Transformer-MM [11], a general explanation framework utilizing gradient information, and ATTCAT [35], a method that formulates Attentive Class Activation Tokens to estimate the relative importance among input tokens. The salience scores produced by these methods are subsequently mapped onto the image space, generating visualizations that are easily understood by humans. However, whether these interpretation maps are faithful to the model's actual behavior remains a matter of debate [16, 24, 51].

2.2. Evaluation of Explanation Faithfulness

In this paper, we evaluate the faithfulness of Vision Transformer explanation methods. This is a critical task [2], especially given the recent debates concerning whether parameters and features in Transformer models are explainable. For instance, the reliability of attention weights has been questioned in several studies [8, 24, 26, 40], premised on the hypothesis that attention weights should not conflict with gradient rankings or token norms. Following this, [51] proposed alternative testing strategies and argued that attention is interpretable when limited to certain circumstances. However, we find that current studies do not sufficiently consider the values of salience scores and the model's confidence in its original predictions, which deviates from the core assumption of faithfulness and leads to inconsistent and untrustworthy outcomes.

Evaluating post-hoc explanations poses a challenge in the realm of Transformer interpretability. Given the absence of justified ground truth [3], one stream of research focuses on developing human-centered evaluations [14, 28, 37]. These studies scrutinize the practical value of explanations provided to the end user. In another direction, [2] employed sanity checks to assess changes in explanations *w.r.t.* randomization within models and datasets. [5] formulated gradient-based explanations in a unified framework and introduced Sensitivity-n as a desired property. However, this metric highly relies on the linearity assumption for simple DNNs and neglects the order of salience scores. Unlike these studies, our work is broadly related to faithfulness metrics that evaluate explanations by monitoring the model's performance on perturbed images [6]. Various versions of perturbation-based metrics have been introduced, providing measures of input feature impact [13, 16, 33, 42]. Despite the achievements, these approaches use cumulative perturbation without individually contrasting input pixels of varying importance levels. Furthermore, they do not directly incorporate the information from specific magnitudes of salience scores, focusing only on their relative ordering. These oversights contribute to deficiencies present in existing metrics. Our approach, on the other hand, scrutinizes the model's response to each distinct pixel group and sheds light on the relevance of the salience scores, providing a more comprehensive evaluation of explanation faithfulness.

3. Methodology

For the image classification task, each input is an image comprised of HW pixels. Given an input image $\mathbf{x}$ and the corresponding predicted class $\hat{y}(\mathbf{x}) \in \{1, 2, ..., C\}$, where C is the number of classes under consideration, post-hoc explanations generate a salience map $\mathbf{M}(\mathbf{x}, \hat{y}) \in \mathbb{R}^{HW}$. The value of each entry in $\mathbf{M}(\mathbf{x}, \hat{y})$ ought to reflect the contribution of the corresponding pixel to the model's output. However, the reliability of these interpretation results remains questionable. This underscores the necessity for further examination of faithfulness.

Following the faithfulness core assumption, the property being investigated is the *extent* to which these salience scores are faithful to the model's actual behavior. Therefore, our proposed evaluation is designed to assess how effectively the disparity in salience scores signifies the variation in their influences on the model's confidence. Considering a sample $\mathbf{x}$, we reorder the input pixels based on their estimated salience and partition them into K equally sized pixel subsets: $G_1, G_2, ..., G_K$. Each subset G_i comprises pixels with top salience ranking from $(i-1)\frac{HW}{K}$ to $i\frac{HW}{K}$ [13, 33]. Regarding each G_i as a basic unit, we can define the salience of a pixel subset:

$$s(G_i) = \sum_{p \in G_i} \mathbf{M}(\mathbf{x}, \hat{y})_p, \quad \text{where} \quad i = 1, 2, ..., K. \quad (1)$$

In essence, the salience of a subset G_i is the sum of salience scores over all pixels in G_i. Following the convention in literature [12, 41, 50], we adopt a proxy measure to access the model's behavior: we replace pixels that belong to a certain subset with the per-sample mean value [22] and then

observe the resulting effect on the model's confidence. Formally, we represent the replacement result by $Rp(\mathbf{x}, G_i)$. Therefore, the alterations in the model's prediction can be formulated as follows:

$$\nabla pred(\mathbf{x}, G_i) = p(\hat{y}(\mathbf{x})|\mathbf{x}) - p(\hat{y}(\mathbf{x})|Rp(\mathbf{x}, G_i)), \quad (2)$$

where $Rp(\mathbf{x}, G_i)$ represents the perturbed image, in which pixels in subset G_i are replaced with the per-sample mean value. The fundamental principle underpinning SaCo is that a subset G_i of higher salience should exert more effects compared to a subset G_j of significantly lower salience. Specifically, if $s(G_i) \geq s(G_j)$, we expect the following inequality to be upheld:

$$\nabla pred(\mathbf{x}, G_i) \geq \nabla pred(\mathbf{x}, G_j). \quad (3)$$

As the difference between $s(G_i)$ and $s(G_j)$ expands, our expectation for Inequality (3) to hold will intensify. Following this, the growing difference in salience scores should accentuate its influence on the evaluation result. For example, a violation of Inequality (3) should be penalized more when the difference in salience becomes larger, thus better reflecting the deviation from the expected model behavior.

Inspired by the Kendall τ statistic [25], for a thorough analysis, we look into all possible pairs of G_i and G_j and assess their compliance with faithfulness property. The assessment is guided by a salience-aware violation test based on inequality (3). Concretely, when this inequality is violated, the difference in salience will negatively impact the evaluation result. On the contrary, when the inequality holds true, the salience difference will add positively to the outcome. For example, suppose for a pair of pixel subsets G_i and G_j with $s(G_i) \geq s(G_j)$, we find $\nabla pred(\mathbf{x}, G_i) < \nabla pred(\mathbf{x}, G_j)$. Then, the difference in salience, $s(G_i) - s(G_j)$, is considered a penalty that reflects the magnitude of our unfulfilled expectations and will be subtracted from the overall coefficient. If we observe $\nabla pred(\mathbf{x}, G_i) \geq \nabla pred(\mathbf{x}, G_j)$ as expected, the difference $s(G_i) - s(G_j)$ will serve as a reward and positively contribute to the evaluation outcome. Detailed steps are elaborated in Algorithm 1.

As per its definition, the SaCo produces a faithfulness coefficient, denoted as F, that ranges from $[-1, 1]$. The sign of F reveals the direction of correlation, *i.e.*, it evaluates if the input pixels with higher salience scores generally exhibit greater or lesser predictive influence on the model. Beyond just the direction, the absolute value of F quantitatively measures the degree of correlation.

Algorithm 1 Salience-guided Faithfulness Coefficient

Input: Pre-trained model Φ, explanation method $\mathcal{E}$, input image $\mathbf{x}$.
Output: Faithfulness coefficient F.
Initialization: $F \leftarrow 0$, $totalWeight \leftarrow 0$
Compute the salience map $\mathbf{M}(\mathbf{x}, \hat{y})$ based on Φ, $\mathcal{E}$, and $\mathbf{x}$. Generate G_i and obtain corresponding $s(G_i)$ and $\nabla pred(\mathbf{x}, G_i)$, for $i = 1, 2, ..., K$.
for $i = 1$ to $K - 1$ **do**
  for $j = i + 1$ to K **do**
    if $\nabla pred(\mathbf{x}, G_i) \geq \nabla pred(\mathbf{x}, G_j)$ **then**
      $weight \leftarrow s(G_i) - s(G_j)$
    else
      $weight \leftarrow -(s(G_i) - s(G_j))$
    end if
    $F \leftarrow F + weight$
    $totalWeight \leftarrow totalWeight + |weight|$
  end for
end for
$F \leftarrow F/totalWeight$
Return F

4. Experimental Setup

4.1. Datasets and Models

We utilize three benchmark image datasets: CIFAR-10, CIFAR-100 [27], and ImageNet (ILSVRC) 2012 [38]. Details regarding the scales of data, numbers of classes, and image resolutions for each dataset are provided in the supplementary. Furthermore, to ensure the reliability of our evaluation, we experiment with three Vision Transformer models that are widely adopted in this field: ViT-B, ViT-L [17], and DeiT-B [47]. In these models, images are divided into non-overlapping 16×16 patches, then flattened and processed to create a token sequence. For classification, a special token $[CLS]$ is added to the sequence, similar to BERT [15].

4.2. Explanation Methods

We investigate ten representative post-hoc explanation methods spanning three categories, *i.e.*, gradient-based, attribution-based, and attention-based. Each method holds unique assumptions about the network architecture and the information flow. For a better assessment, selected methods are widely recognized in the explainability literature and also compatible with Vision Transformer models under consideration. Detailed descriptions of these techniques are provided in the supplementary.

Gradient-based methods. We select two state-of-the-art explanation methods from this category: Integrated Gradients [46] and Grad-CAM [39]. Note that the Grad-CAM method was initially designed for visualizing intermediate features in CNNs. Our implementation follows the prior study in Vision Transformer interpretability [12].

Attribution-based methods. Unlike gradient-based, attribution-based methods explicitly model the information flow inside the network. We select LRP [9], Partial LRP [49], Conservative LRP [4], and Transformer Attribution

[12]) in our experiment for a thorough analysis.

Attention-based methods. Regarding the attention-based methods, we employ four variants: Raw Attention [24], Rollout [1], Transformer-MM [11], and ATTCAT [35] in our experiments. These methods are specifically designed for the Transformer models.

4.3. Evaluation Metrics

We compare our proposed SaCo with widely adopted existing metrics to validate its reliability.

Area Under the Curve (AUC) ↓. This metric calculates the Area Under the Curve (AUC) corresponding to the model's performance as different proportions of input pixels are perturbed [6]. To elaborate, we first generate new data by gradually removing pixels in increments of 10% (from 0% to 100%) based on their estimated salience scores. The model's accuracy is then assessed on these perturbed images, resulting in a sequence of accuracy measurements. The AUC is subsequently computed using this sequence. A lower AUC indicates a better explanation.

Area Over the Perturbation Curve (AOPC) ↑. Rather than measuring the model's accuracy, AOPC [13, 33] quantifies the variations in output probabilities *w.r.t.* the predicted label after perturbations. A higher AOPC indicates a better explanation.

Log-odds score (LOdds) ↓. The LOdds [35, 42] evaluates if the pixels considered important are enough to sustain the model's prediction, which is measured on the logarithmic scale. To facilitate fair and reliable comparisons, we gradually eliminate the top 0%, 10%, ..., 90%, and 100% of pixels, based on their salience scores. This removal process aligns with that employed for calculating AUC and AOPC. A lower LOdds indicates a better explanation.

Comprehensiveness (Comp.) ↓. The Comprehensiveness [16] measures if pixels with lower salience are dispensable for the model's prediction. For consistent comparisons, we cumulatively eliminate pixels in the least important 0%, 10%, ..., 90%, and 100%. A lower Comprehensiveness indicates a better explanation.

5. Experimental Results

5.1. Interrelationships among Evaluation Metrics

To demonstrate the significance and necessity of SaCo, we conduct a correlation analysis following the thorough experimental setup. To this end, we begin by evaluating each explanation method on single samples independently, using each of the metrics. Then, we compute the statistical rank correlations between the evaluation results obtained from the SaCo and those from other existing metrics. Note that we are correlating based on the rankings rather than the exact values of evaluation results, as these metrics can vary in scales and orientations. This approach allows us to quantify

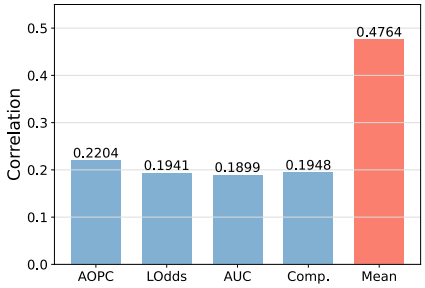

Figure 2. Correlations between sample rankings *w.r.t.* our SaCo and existing metrics.

the degree of similarity between the assessments provided by the SaCo and the traditional metrics in current use. Figure 2 shows the comprehensive statistical correlation results averaged across all considered datasets, explanation methods, and Vision Transformer models, as described in Section 4. The first four bars on the left depict the correlations between our SaCo and the metric indicated by the corresponding x-axis label. The rightmost bar shows the average correlation among all other metrics, excluding ours.

As displayed, the correlation scores between our SaCo and other existing metrics range from 0.18 to 0.22 (in this analysis, a result of 1 indicates a complete correlation, while a result of 0 indicates no correlation). These low scores signify minimal congruence in their evaluation, suggesting that our SaCo potentially evaluates a complementary aspect compared to existing metrics, *i.e.*, the core assumption of faithfulness. In essence, the traditional metrics tend to generate similar results regardless of the degree of faithfulness, due to their lack of comparisons among individual pixels and the direct incorporation of the distribution of salience score magnitudes. On the other hand, the average intra-correlation among the existing metrics themselves is significantly higher (0.4764). This result implies that these metrics tend to evaluate similar or overlapping aspects of interpretations (mainly the effect of progressive pixel removal), with insufficient consideration of the faithfulness assumption. These results emphasize the importance of our proposed SaCo, as current metrics appear to lack the capabilities to adequately assess faithfulness. Therefore, the need for a more comprehensive assessment method for post-hoc explanations of Vision Transformers is reinforced.

5.2. Evaluating Random Attribution

Recognizing the potential shortages of current metrics in assessing faithfulness, we conduct a critical examination to further demonstrate the limitations: we evaluate the performance of Random Attribution as an explanation method [22, 41]. In this context, Random Attribution directly assigns salience scores to input pixels using a purely uniform distribution, with no reference to the information of the

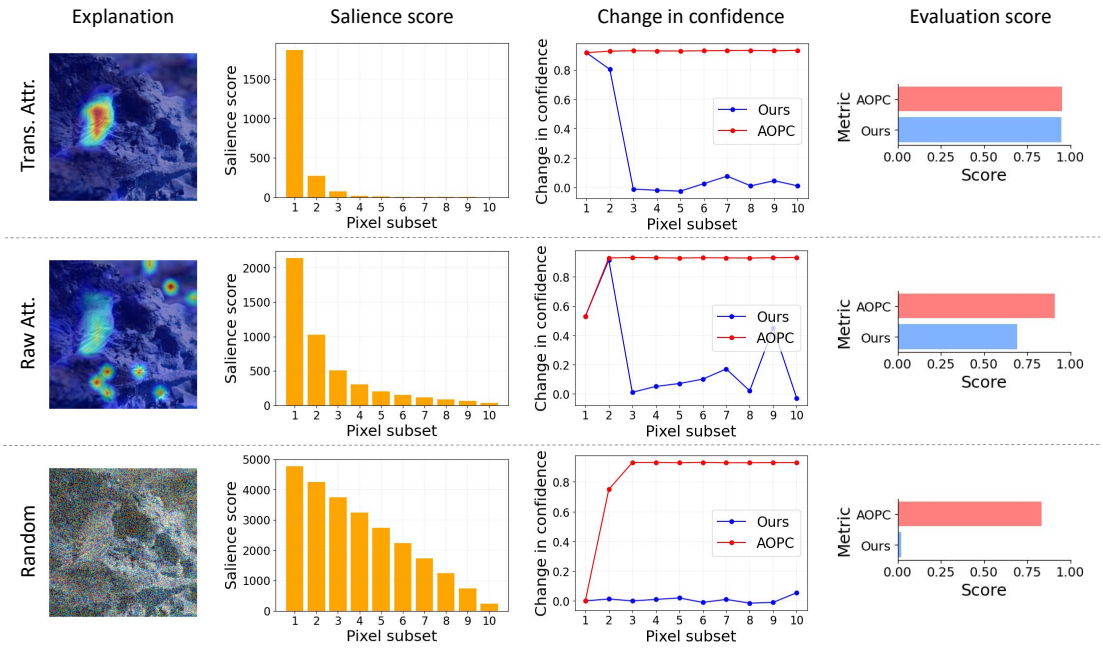

Figure 3. Illustration of three explanations for the predicted class 'linnet', salience score distributions, changes in model's confidence caused by perturbation, and final SaCo and AOPC scores.

model's internal inference process [50]. From the perspective of mathematical expectation, such Random Attribution represents a complete absence of faithfulness [41], as the assigned salience scores bear no discriminative relationship to the actual impacts of the pixels on the model. As a result, an ideal metric of faithfulness is expected to return a distinct score on Random Attribution, essentially setting a baseline indicative of a significant lack of meaningful understanding of the model's prediction. In particular, from the viewpoint of mathematical expectation, our SaCo yields a faithfulness coefficient of zero for Random Attribution. Given that the salience scores are sampled from a uniform distribution, the pixels subsets $(G_i, i = 1, 2, ..., K)$ partitioned according to these scores will not show significant differences in their discriminative information. Therefore, Inequality (3) holds true with a probability of one-half, leading to a balance between the positive and negative terms during the calculation of the coefficient outlined in Algorithm 1.

Case study. We first present a case study comparing the behavior of SaCo and AOPC. AOPC operates on cumulative pixel perturbation and disregards the alignment between salience values and actual influences. We conduct this study on ViT-B [17] and a sample from ImageNet [38]. As shown in Figure 3, we evaluate three explanation methods: Transformer Attribution [12], Raw Attention, and Random Attribution. Specifically, following the literature convention [11, 12, 34, 39, 52], we divide the image into ten disjoint pixel subsets, designated by K=10. The salience score of each subset $s(G_i)$ is computed by Eq. (1). We then implement both SaCo's individual perturbation and AOPC's cumulative perturbation on the images, and calculate the resulting changes in the model's confidence (probability) for the predicted class, as defined by Eq. (2). Here, a positive change represents a decrease in confidence.

Across all methods, the confidence drops induced by cumulative perturbation remain almost unchanged (above 0.9) after the removal of the top 30% important pixels, despite the subsequent removal of more pixels. This pattern results in consistently high AOPC scores for all three methods. Furthermore, one can observe that the impacts of cumulative pixel removal do not increase linearly, which holds true even for uniform Random Attribution. The precipitous decline in confidence may stem from the out-of-distribution issues that arise due to the substantial removal of pixels [21, 22, 41]. As discussed in Section 1, in cumulative perturbation, less important pixels are removed only after the most important pixels have been eliminated, causing their individual effects to become entwined. Conversely, SaCo directly assesses the influences of individual pixel subsets and measures their alignment with the salience distribution. In the case of Transformer Attribution, the influences of pixel subsets (indicated by the blue curve) closely align with the salience score distribution, yielding a high SaCo result. In contrast, with Raw Attention, the subsets $G_i(i = 3, 4, 5, 6)$ have minimal impacts on the model, possibly due to unexpected emphasis on irrelevant objects and backgrounds, as reflected by a reduced SaCo score. For Random Attribution, the influences of subsets exhibit little variation, resulting in a near-zero SaCo score. This case study shows that SaCo provides a rigorous evaluation that

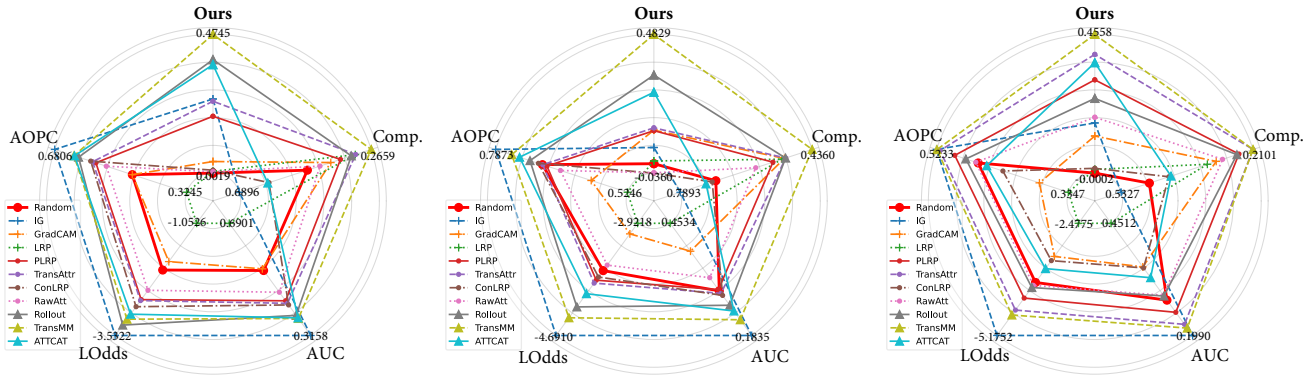

Figure 4. Evaluation results for advanced explanation methods and Random Attribution (red). Three graphs present results on CIFAR-10 (left), CIFAR-100 [27] (middle), and ImageNet [38] (right), respectively. The values on each axis have been rescaled so that a larger distance from the center consistently signifies superior performance. Enlarged graphs are provided in the supplementary for better clarity.

effectively differentiates superior and inferior explanations.

Large-scale experiments. To further validate the effectiveness of SaCo, we conduct experiments on large-scale datasets. Figure 4 presents the overall outcomes on CIFAR-10, CIFAR-100 [27], and ImageNet [38]. For every dataset, we compare existing metrics with SaCo, where each result is an average over three Vision Transformers indicated in Section 4. Across all datasets, SaCo consistently scores Random Attribution near zero, while most explanation methods obtain positive scores under SaCo, demonstrating its ability to set a standard benchmark. Conversely, other metrics cannot provide consistent evaluation for Random Attribution. When assessed using these metrics, despite producing purely noisy heatmaps, Random Attribution appears to have comparable performance, even surpassing some state-of-the-art methods such as Partial LRP [49], Transformer Attribution [12], and ATTCAT [35], as demonstrated in Figure 4. Another observation is that existing metrics seem to be sensitive to the removal order of the cumulative perturbation they employ. A phenomenon across three datasets in our experiment exemplifies this problem: while I.G. performs best on AOPC, LOdds, and AUC (with Most Relevant First removal), it is surprisingly the worst on Comprehensiveness (with the reverse removal order). This indicates that current metrics are significantly inconsistent *w.r.t.* hyperparameters such as removal orders [36]. In contrast, we eliminate this inconsistency by directly comparing individual pixel subsets of various salience scores, instead of cumulatively removing them in a specific order.

5.3. Assessment of Current Explanation Methods

Figure 4 illustrates the results of SaCo for existing explanation methods over Vision Transformer models and datasets under consideration. Observations from our results indicate that all explanation methods perform moderately, as reflected in their suboptimal SaCo scores. These results necessitate in-depth research into explanation methods to more accurately depict the model's reasoning process and adhere to the core assumption of faithfulness. Furthermore, of all explanation methods evaluated, we can see that those utilizing attention information generally perform better in our SaCo assessment. However, despite falling under the same category, Raw Attention noticeably underperforms. This motivates us to hypothesize that attention-based explanation methods can only achieve superior performance when they incorporate auxiliary information, such as gradient and cross-layer integration. We further conduct experiments in Section 5.4 for its demonstration.

5.4. Effects of Designs in Explanation Methods

We now delve into the designs of explanation methods that may augment the alignment with the faithfulness core assumption. We focus our analysis on attention-based explanation methods because attention weights are inherently meaningful for Vision Transformers [11]. Moreover, previous assessments have demonstrated that attention-based methods generally outperform others [12, 35], making them a valuable factor for further investigation.

As depicted in Figure 4, attention-based explanation methods that use well-crafted aggregation rules and auxiliary information from gradient *w.r.t.* the model's output tend to score higher under SaCo. Based on this observation, we hypothesize that the incorporations of aggregation rules and gradient information play a vital role in compliance with faithfulness. To validate this hypothesis, we conduct ablative experiments employing four variants of attention-based explanation methods: **(i)** utilizing attention weights in the final layer of the model, **(ii)** aggregating attention information across all layers, **(iii)** utilizing the final layer's attention weights and integrating their gradient information, and **(iv)**

cross-layer aggregation	gradient	SaCo ↑
✗	✗	0.1835
✓	✗	0.2453
✗	✓	0.3783
✓	✓	**0.4558**

Table 1. Ablative study on attention-based explanation methods.

K	I. G. [46]	Grad-CAM [39]	LRP [9]	P. LRP [49]	Trans. Attr. [12]	Con. LRP [4]	Raw Att. [24]	Rollout [1]	Trans. MM [11]	ATTCAT [35]
5	0.1585	0.1659	0.0246	0.4628	0.5680	0.0544	0.3200	0.3372	0.6041	0.3178
10	0.1647	0.1142	0.0120	0.3066	0.3902	0.0155	0.1835	0.2453	0.4558	0.3629
20	0.1785	0.1000	0.0054	0.2282	0.2906	-0.0201	0.1411	0.1956	0.3651	0.3617

Table 2. Performance of current explanation methods on our SaCo, with different values of K. Results are averaged over three Vision Transformer models on ImageNet [38].

aggregating attention weights across all layers and integrating gradient information.

Table 1 presents our ablation study's results on attention-based explanation methods, showing the benefits of integrating gradient information and multi-layer aggregation. Firstly, we can observe that the incorporation of gradient information significantly improves faithfulness scores. Specifically, when considering only the last layer, the introduction of gradients boosts the evaluation outcome by approximately 106%. This effect remains robust, showing an 86% increase even with the application of aggregation across all layers. Secondly, despite being less influential than gradient information, aggregating across multiple layers also contributes positively to the results. This improvement occurs irrespective of whether the gradient is used, suggesting that a more holistic view of the Vision Transformer can consistently facilitate more faithful explanations. The results empirically support our initial hypothesis that both the gradient and aggregation rules are essential for Vision Transformer explanations, with the former having a more significant effect. Refining these two design factors offers a promising avenue for advancing the development of explanations for Vision Transformers. Furthermore, this study also demonstrates SaCo's capability to capture the property of faithfulness. The result resonates with the intuitive human understanding that a precise depiction of the model's reasoning regarding the recognized object requires class-specific information and comprehensive insights from the entire inference process.

5.5. Exploring Influential Factors in SaCo

The number of pixel subsets (K). To explore the impact of varying the number of pixel subsets, we assess the performance of existing explanation methods across different values of K in Algorithm 1. Table 2 presents the averaged results across three Vision Transformer models, evaluated by our SaCo with different K values. It can be observed that with an increase in K, most methods exhibit a slight decline in their SaCo scores. This trend is more pronounced for Partial LRP [49], Transformer Attribution [12], and Transformer-MM [11], suggesting that these explanation methods might struggle to maintain faithfulness under a more granular evaluation. Conversely, Integrated Gradients (I. G.) and ATTCAT show relatively consistent performance across different K values, indicating stronger robustness to the granularity of subset division. These results emphasize the significance of choosing an appropriate K. The optimal value of K depends on the specific needs of evaluation, striking a balance between computational demands and the granularity of the faithfulness evaluation.

Measure of distinct salience scores. In our proposed SaCo (see Algorithm 1), we quantify the extent of satisfying or violating human expectation by the differences in salience scores, expressed as $weight \leftarrow s(G_i) - s(G_j)$. One possible alternative for this measure is taking the ratio: $weight \leftarrow \frac{s(G_i)}{s(G_j)}$, which seems promising because ratios are effective at capturing the relative magnitude of salience among pixel subsets. However, using a ratio violates the property of scale-invariance. In practice, the explanation results may be normalized or scaled into $[0, 1]$ interval as post-processing [11, 12, 35, 39], which will skew the ratio measure. This may cause extremely high or even infinite ratios when $s(G_j)$ is close to zero after transformation, which is especially problematic when comparing explanations from different methods that scale their salience scores differently. In contrast, our designed SaCo maintains the scale-invariance property. Regardless of the scale on which the salience scores are expressed, the results remain consistent. This property enables a stable comparison between distinct methods thereby ensuring a more robust evaluation. Formal proof demonstrating that our SaCo satisfies the scale-invariance property is provided in the supplementary.

6. Conclusion

In this work, we proposed SaCo, a novel faithfulness evaluation. Our SaCo leverages salience-guided comparisons of pixel subsets for their different contributions to the model's prediction, providing a more robust benchmark. Experiments reveal insightful observations: **(i)** our correlation analysis shows the necessity of SaCo, as existing metrics capture overlapping aspects while lacking consideration of faithfulness, **(ii)** unlike existing metrics, SaCo can identify Random Attribution as completely lacking significant information and provide consistent results free of removal order dependency, and **(iii)** attention-based methods are generally more faithful, and their performance can be further enhanced by gradient information and multi-layer aggregation. In summary, our work provides a comprehensive evaluation of faithfulness in Vision Transformer explanations, which will spur further research in explainability.

Acknowledgments: This research is supported by NSF IIS-2309073 and ECCS-212352101. This article solely reflects the opinions and conclusions of its authors and not the funding agents.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020. 1, 3, 5, 8

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, 2018. 3

[3] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. In *NeurIPS*, 2022. 3

[4] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *ICML*, 2022. 1, 3, 4, 8

[5] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR*, 2018. 3

[6] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *EMNLP*, 2020. 1, 3, 5

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

[8] Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *ACL*, 2020. 3

[9] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *ICANN*, 2016. 4, 8

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[11] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 1, 3, 5, 6, 7, 8

[12] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 1, 3, 4, 5, 6, 7, 8

[13] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *ACL*, 2020. 1, 3, 5

[14] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. In *NeurIPS*, 2022. 1, 3

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 4

[16] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*, 2020. 1, 3, 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 4, 6

[18] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019. 3

[19] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017. 3

[20] Shir Gur, Ameen Ali, and Lior Wolf. Visualization of supervised and self-supervised neural networks via attribution guided factorization. In *AAAI*, 2021. 3

[21] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *NeurIPS*, 2021. 6

[22] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, 2019. 3, 5, 6

[23] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *ACL*, 2020. 1

[24] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *NAACL*, 2019. 3, 5, 8

[25] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 1938. 4

[26] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *EMNLP*, 2020. 3

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4, 7

[28] Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. In *NeurIPS*, 2018. 3

[29] Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking attention-model explainability through faithfulness violation test. In *ICML*, 2022. 1

[30] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017. 3

[31] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *IJCV*, 2016. 3

[32] Woo-Jeoung Nam, Shir Gur, Jaesik Choi, Lior Wolf, and Seong-Whan Lee. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *AAAI*, 2020. 3

[33] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL*, 2018. 1, 3, 5

[34] Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogerio Feris, and Aude Oliva. Ia-red ̂2: Interpretability-aware redundancy reduction for vision transformers. In *NeurIPS*, 2021. 6

[35] Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining transformers via at-

tentive class activation tokens. In *NeurIPS*, 2022. 1, 3, 5, 7, 8

[36] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *ICML*, 2022. 7

[37] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *AAAI*, 2018. 3

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 4, 6, 7, 8

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3, 4, 6, 8

[40] Sofia Serrano and Noah A Smith. Is attention interpretable? In *ACL*, 2019. 3

[41] Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? In *NeurIPS*, 2021. 1, 3, 5, 6

[42] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017. 1, 3, 5

[43] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. In *ICML*, 2017. 3

[44] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *ICML Workshop*, 2017. 3

[45] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *NeurIPS*, 2019. 3

[46] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 3, 4, 8

[47] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 4

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[49] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*, 2019. 4, 7, 8

[50] Yipei Wang and Xiaoqian Wang. A unified study of machine learning explanation evaluation metrics. *arXiv preprint arXiv:2203.14265*, 2022. 1, 2, 3, 6

[51] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP*, 2019. 3

[52] Quan Zheng, Ziwei Wang, Jie Zhou, and Jiwen Lu. Shap-cam: Visual explanations for convolutional neural networks based on shapley value. In *ECCV*, 2022. 6

[53] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 3

[54] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *IEEE TPAMI*, 2018. 3