1. Demonstrating Scale-invariance of SaCo

We prove the scale-invariance of our SaCo, as evidenced by the equation:

$$\text{SaCo}(aS + b) = \text{SaCo}(S), \tag{1}$$

valid for any positive real number a and any real number b. In this context, S denotes the set of salience scores attributed to the input pixels, and $aS + b$ represents a linear transformation applied to each pixel in the set S. The computation of our SaCo involves summing up the differences in salience scores for all pairs of subsets (G_i, G_j). The assigned weight in the algorithm for each pair is $s(G_i) - s(G_j)$ if $\nabla pred(x, G_i) \geq \nabla pred(x, G_j)$, and it is $-(s(G_i) - s(G_j))$ otherwise. When a linear transformation is applied to the salience scores such that $s'(G_i) = as(G_i) + b$, this results in a corresponding adjustment of the weight. The modified weight can be represented as:

$$\begin{aligned} s'(G_i) - s'(G_j) &= (a \cdot s(G_i) + b) - (a \cdot s(G_j) + b) \\ &= a \cdot (s(G_i) - s(G_j)), \end{aligned} \tag{2}$$

and

$$\begin{aligned} -(s'(G_i) - s'(G_j)) &= -((a \cdot s(G_i) + b) - (a \cdot s(G_j) + b)) \\ &= -a \cdot (s(G_i) - s(G_j)), \end{aligned} \tag{3}$$

Consequently, the aggregate weight $totalWeight'$ for the transformed scores becomes:

$$totalWeight' = a \cdot totalWeight. \tag{4}$$

Given that the direction of each weight remains unchanged, and its magnitude is scaled by a, we derive:

$$F' = a \cdot F. \tag{5}$$

Subsequently, considering the SaCo score is computed as $\frac{F'}{totalWeight'}$, we have:

$$\begin{aligned} \text{SaCo}(aS + b) &= \frac{F'}{totalWeight'} \\ &= \frac{a \cdot F}{a \cdot totalWeight} \\ &= \frac{F}{totalWeight} \\ &= \text{SaCo}(S). \end{aligned} \tag{6}$$

Hence, SaCo demonstrates scale-invariance. The results of SaCo remain unaffected by the scale of salience scores, ensuring its robustness against post-processing steps such as normalization or re-scaling. Note that our proof of scale-invariance does not necessarily hinge on a being positive. Yet, if a is negative, the orientation of salience scores gets reversed, resulting in unreasonable explanation outcomes. Under this circumstance, although SaCo retains its scale-invariance property, the negative value of a may lead to a diminished score due to its effect on the partition of pixel subsets. This occurs because the arrangement of these subsets relies on the ordered rankings of the salience scores.

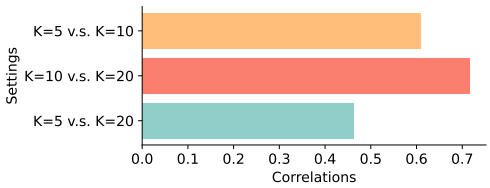

Figure 1. Correlation coefficients of SaCo outcomes with different values of K.

2. Empirical Study on Choices of K

We explore SaCo's sensitivity to different K values in our study, where K represents the number of subsets into which input pixels are divided. The experiments are conducted with K set to 5, 10, and 20. Figure 1 showcases the correlation among SaCo outcomes for these varying K values. This experiment is performed on ImageNet and the results are averaged across three Vision Transformer models and ten explanation methods. In our analysis, a correlation score of 1 indicates a perfect correlation, whereas a score of 0 implies the absence of any correlation. The positive correlation between K=5 and K=10, demonstrated by a coefficient of 0.6101, indicates a moderate similarity in SaCo results for these two settings. This suggests that a coarse partitioning into five subsets can still capture similar salience distinctions as a more detailed partitioning into ten subsets. The correlation increases when comparing K=10 and K=20. This stronger correlation might be attributed to the fact that splitting into ten or twenty subsets both provides a detailed view of the salience scores, capturing subtler nuances in the model's behavior. These positive correlations demonstrate that while the granularity of subset division (as determined by K) can play a role in the final evaluation, the fundamental principles provided by our SaCo remain consistent.

3. Experimental Setup

3.1. Datasets

CIFAR-10 and CIFAR-100. CIFAR-10 and CIFAR-100 [10] are two widely used image classification datasets, each containing 60,000 32×32 color images. CIFAR-10 has 10 classes, while CIFAR-100 has a more challenging setting with 100 classes. Both datasets are split into 50,000 training and 10,000 testing images. In this paper, we evaluate explanation methods on the testing sets.

ImageNet. ImageNet dataset [14] is a large-scale benchmark for image classification. In this work, we evaluate explanation methods on the ImageNet validation set, which comprises 50,000 high-resolution images across 1,000 distinct classes. Each class contains roughly the same number of images, ensuring a balanced benchmark.

3.2. Explanation Methods

3.2.1 Gradient-based methods

Integrated Gradients. Integrated Gradients (IG) [17] calculates contributions by integrating gradients along a path from a baseline input $\mathbf{x_0}$ to the original input $\mathbf{x}$:

$$\mathrm{IG}(\mathbf{x}, \mathbf{x_0}) = (\mathbf{x} - \mathbf{x_0}) \odot \int_0^1 \frac{\partial f(\mathbf{x_0} + \alpha(\mathbf{x} - \mathbf{x_0}))}{\partial \mathbf{x}} \mathrm{d}\alpha, \tag{7}$$

where f represents the classification model. In practice, the integral is approximated using the Riemann Sum over a linear interpolation path.

Grad-CAM. Instead of the original input, Grad-CAM [15] utilizes the attention map in the last layer. Following the prior work [6], we perform multi-head integration based on gradient information.

3.2.2 Attribution-based Methods

LRP. LRP [4] starts from the model's output and propagates relevance scores backward up to the input image. This propagation adheres to a set of rules defined by the Deep Taylor Decomposition theory [11].

Partial LRP. Partial LRP [18] also backpropagates relevance scores, but uniquely, it uses the relevance map from a specific intermediate layer as the final explanation. In line with convention [5, 6], we choose the relevance map associated with the attention map in the last layer.

Transformer Attribution. Transformer Attribution [6] is an attribution-based method specifically designed for Transformer models. It first computes relevance scores via the LRP and then integrates these scores with attention maps to produce an explanation.

Conservative LRP. Conservative LRP [2] introduces specialized Layer-wise Relevance Propagation rules for attention heads and layer norms in Transformer models. This is designed to implement conservation, a common property of attribution techniques.

3.2.3 Attention-based Methods

Raw Attention. This method [9] extracts the multi-head attention map from the last layer of the model and reshapes the row corresponding to the $[CLS]$ token into the 2D input space. An interpretation is further derived by averaging across different heads.

Rollout. Rollout [1] interprets the information flow within Transformers from the perspective of Directed Acyclic Graphs (DAGs). It traces and accumulates the attention weights across layers using a linear combination strategy.

Transformer-MM. Transformer-MM [5] is a general interpretation framework applicable to diverse Transformer architectures. It aggregates attention maps with corresponding gradients to generate class-specific explanations.

ATTCAT. ATTCAT [13] is a Transformer explanation technique using attentive class activation tokens. It employs a combination of encoded features, their associated gradients, and attention weights to produce confident explanations.

3.3. Evaluation Metrics

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 [3]. To elaborate, we first generate new data by gradually removing pixels in increments of 5% (from 0% to 100%) based on their explanation weights. The model's accuracy is then assessed on these perturbed data, resulting in a sequence of accuracy measurements. The AUC is subsequently computed using this sequence.

Area Over the Perturbation Curve (AOPC) ↑. AOPC [7, 12] measures the changes in output probabilities *w.r.t.* the predicted label after perturbations:

$$\mathrm{AOPC} = \frac{1}{|K|} \sum_{k \in K} (\hat{p}(y|\mathbf{x}) - \hat{p}(y|\mathbf{x_k})), \tag{8}$$

where $K = \{0, 10, ..., 90, 100\}$ is a set of perturbation levels, $\hat{p}(y|\mathbf{x})$ estimates the probability for the predicted class given a sample $\mathbf{x}$, and $\mathbf{x_k}$ is the perturbed version of image $\mathbf{x}$, from which the top $k\%$ pixels ranked by salience scores are eliminated.

Log-odds score (LOdds) ↓. LOdds [13, 16] averages the difference between negative logarithmic probabilities on the predicted label before and after masking $k\%$ top-scored pixels over perturbations K:

$$\mathrm{LOdds} = -\frac{1}{|K|} \sum_{k \in K} \log \frac{\hat{p}(y|\mathbf{x})}{\hat{p}(y|\mathbf{x_k})}. \tag{9}$$

The notations are the same as in Eq. (8).

Comprehensiveness (Comp.) ↓. Comprehensiveness [8] is also referred to as the negative perturbation test. This examines how the removal of supposedly less important input pixels would affect the model's output. Concretely, Comprehensiveness gauges the shifts in output probabilities *w.r.t.* the predicted label after the least important features have been excluded.

4. Enlarged Graphs

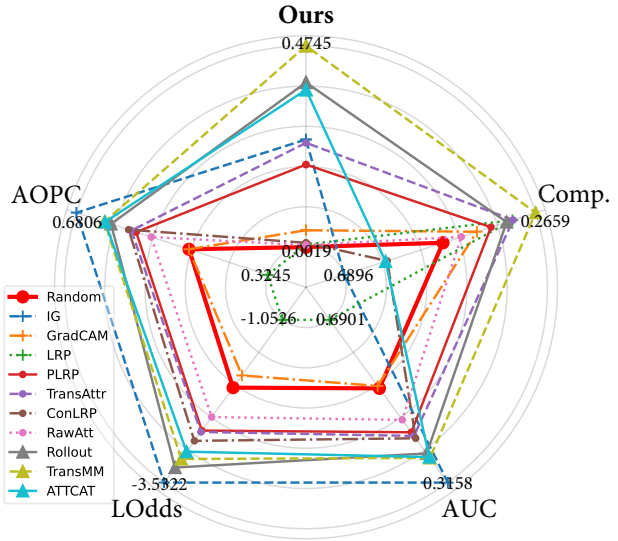

Figure 2. Evaluation results for existing explanation methods as well as Random Attribution, under various metrics. This graph presents results on CIFAR-10 averaged over three Transformer models.

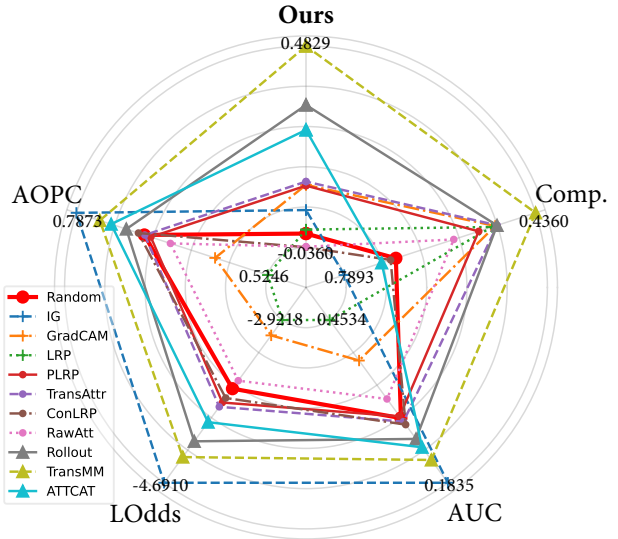

Figure 3. Evaluation results for existing explanation methods as well as Random Attribution, under various metrics. This graph presents results on CIFAR-100 averaged over three Transformer models.

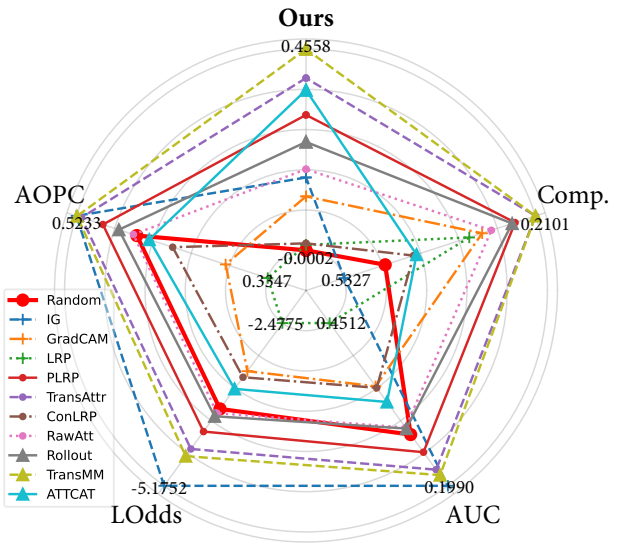

Figure 4. Evaluation results for existing explanation methods as well as Random Attribution, under various metrics. This graph presents results on ImageNet averaged over three Transformer models.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020. 2

[2] 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. 2

[3] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *EMNLP*, 2020. 2

[4] 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. 2

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021. 2

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 2

[7] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. In *ACL*, 2020. 2

[8] 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. 2

[9] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *NAACL*, 2019. 2

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

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

[12] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL*, 2018. 2

[13] Yao Qiang, Deng Pan, Chengyin Li, Xin Li, Rhongho Jang, and Dongxiao Zhu. Attcat: Explaining transformers via attentive class activation tokens. In *NeurIPS*, 2022. 2

[14] 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. 2

[15] 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. 2

[16] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017. 2

[17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. 2

[18] 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. 2