Supplementary Materials- A Meaningful Perturbation Metric for Evaluating Explainability Methods

Anonymous ECCV 2024 Submission

Paper ID #7602

1 Implementation Details

In this section, we provide additional implementation details of our method, as described in Sec. 3 of the main paper. Kindly note that in addition to the provided details, we attach a ZIP file with our full code to reproduce the experiments presented in the paper.

1.1 Data Construction

First, we provide additional details on the curation of our dataset I. As mentioned in the main paper, we begin by performing full inpainting (*i.e.*, from a completely blacked out image) on all ImageNet [7] classes for a set of 20 random seeds with the prompt *"{class-name}"*. A class is deemed valid if and only if, in over 50% of the instances, the classifier's prediction of the image accurately aligned with the specified class.

1.2 Inpainting details

We have utilized the public version stable-diffusion-inpaiting [6] from Hugging Face, initialized with the weights of the Stable-Diffusion-v-1-2. The inpaiting was run on NVIDIA GeForce RTX 2080 Ti GPU. The images and their respective masks were resized to 512×512 images to match the training resolution. Before reapplying the classifiers, the inpainted image was resized back to 224×224. As recommended the guidance scale parameter was set to 7.5.

2 Variance in Attribution Methods

As can be seen in Fig. 1, different attribution methods often produce entirely different attribution maps. The difference in the heatmaps implies that there should be a difference in their quality, and therefore we aim to provide a metric to evaluate them, in a way that will provide a good separation.

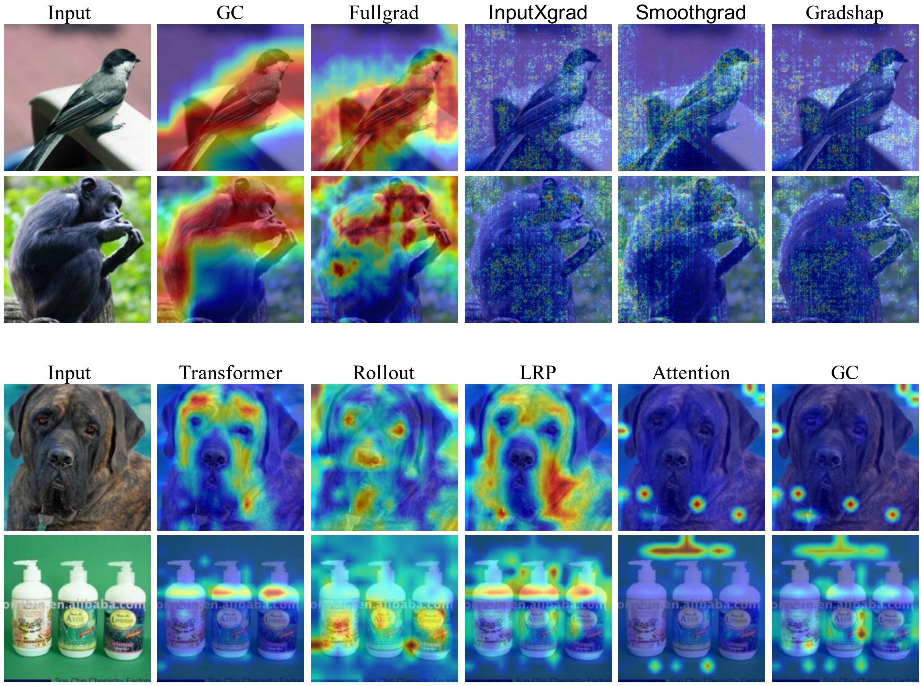

Fig. 1: Examples of heatmaps using different attribution methods for ResNet (top), and ViT (bottom). Different attribution methods often result in entirely different attribution maps, necessitating a metric to distinguish between them.

3 Additional baselines

Alongside the primary baselines presented in the main paper, we enclose a comparison against two additional baselines, that are not directly comparable with our approach, as they assess different aspects of attribution maps. The first is faithfulness violation [5] which checks that the deletion of the relevant pixels decreases the confidence of the predicted class. The second baseline is a saliency-based evaluation [1] the prediction is applied on a the smallest rectangle patch that contains the top pixels of each specific step.

We compare the baselines against our method in two aspects, the separation of rand from the real attribution methods, and the separation of the attribution methods themselves.

Fig. 2 shows that the violation method fails to clearly separate the random baseline from the real attribution methods, and over all classifier the random baseline is ranked in between them. Additionally, the saliency-based metric produces similar outputs for various attribution methods, making it less suited for the thorough evaluation of attribution methods. As we explain in the section above, the difference in the output of attribution methods means that there should be some separation when evaluating them.

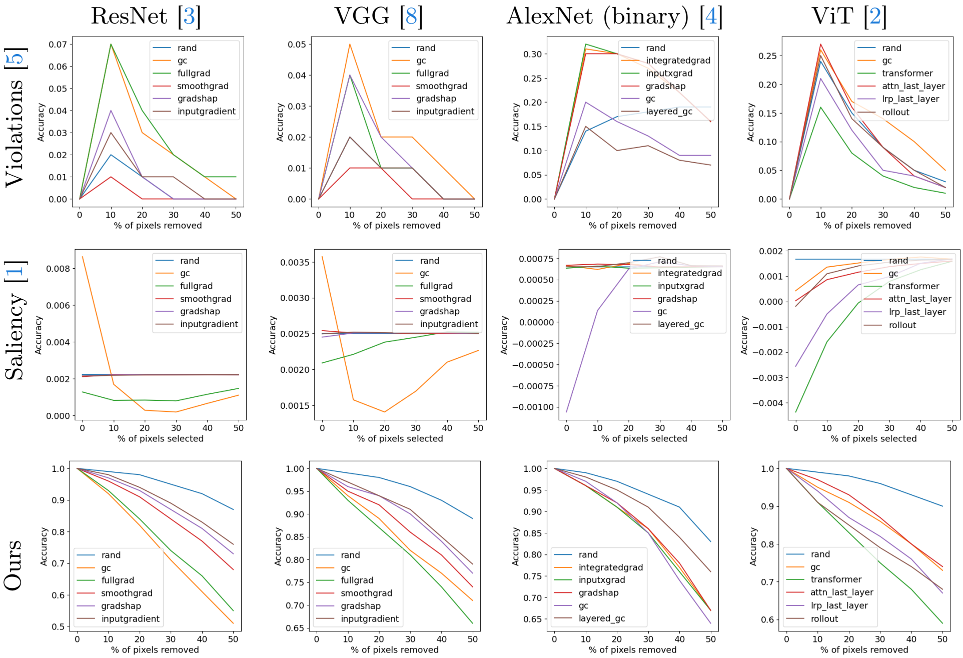

Fig. 2: Perturbation comparison against additional baselines with ResNet-50, VGG, AlexNet-based binary classifier, and ViT-B (please zoom in to view better). For each model, we consider the most common explainability algorithms, in addition to a *random* selection of pixels. As can be observed, the other perturbation methods often struggle with separating random maps from actual relevance maps (*e.g.*, delete for all models, blur for AlexNet, mean for all CNN variants) and appear to produce very similar results for all methods. Conversely, our method produces consistent ranking and meaningful distinction from the random baseline.

References

1. Dabkowski, P., Gal, Y.: Real time image saliency for black box classifiers. In: Neural Information Processing Systems (2017), https://api.semanticscholar.org/CorpusID:41766449 2, 3
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 3
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 3
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1097–1105 (2012) 3
5. Liu, Y., Li, H., Guo, Y., Kong, C., Li, J., Wang, S.: Rethinking attention-model explainability through faithfulness violation test. ArXiv **abs/2201.12114** (2022), https://api.semanticscholar.org/CorpusID:246411233 2, 3
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022) 1
7. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 1
8. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013) 3