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 utilized the public version stable-diffusion-inpaiting [6] from Hugging Face, initialized with the weights of the Stable-Diffusion-v-1-2. As recommended the guidance scale parameter was set to 7.5. The inpaiting was run on NVIDIA GeForce RTX 2080 Ti GPU with 11GB of memory. 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.

2 Variance in Attribution Methods

As mentioned in the main paper, different attribution methods often yield very different relevance maps. Therefore, it is necessary to develop evaluation metrics that can clearly distinguish between them. Otherwise, it remains unclear which method should be used in each use-case. To empirically substantiate this point, Fig. 1 showcases representative examples of the relevance maps produced by different attribution methods given the *same input* using both ResNet (top of

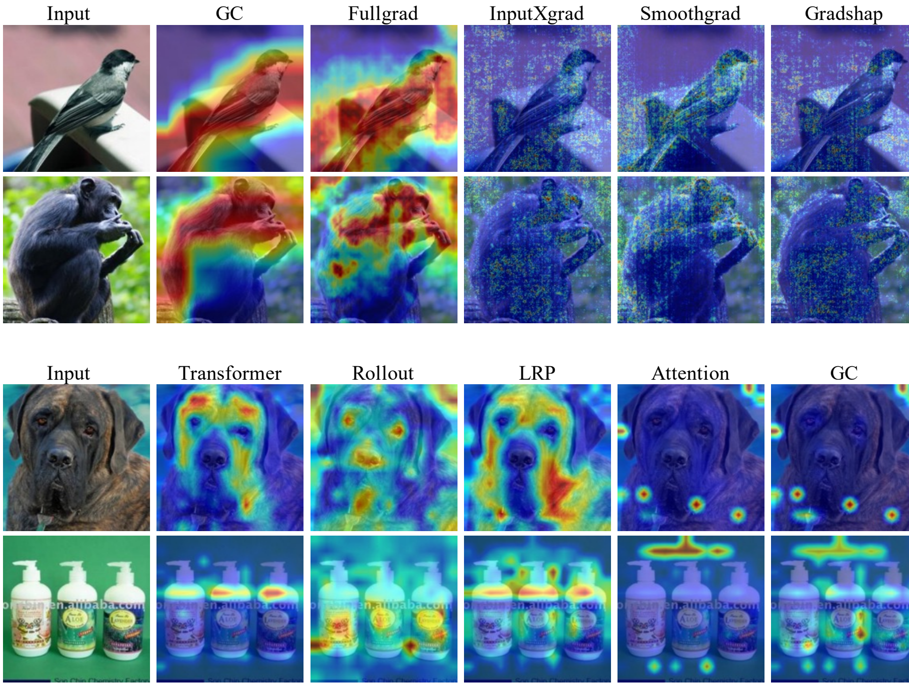

Fig. 1: Examples of relevance maps produced by different attribution methods for the same input using ResNet (top), and ViT (bottom). Different attribution methods often result in entirely different attribution maps, necessitating a metric to distinguish between them.

Fig. 1), and ViT (bottom of Fig. 1). Observe that, for example, given an input image of a dog (ViT top row) each attribution method classifies different image pixels as the most relevant ones. In the absence of a metric that can clearly distinguish between the different maps, we would not be able to determine which attribution method is most faithful. As demonstrated in Fig. 2 of the main paper, the baseline metrics often produce results that are very similar across all examined attribution methods. This can be explained by the OOD effect described in the paper; the perturbations applied by the baselines drive the image out of distribution, causing a change in the prediction which does not necessarily reflect the actual relevance of the modified pixels (see Figs. 1,3,4 on the main paper). In contrast, when applying our metric, the perturbed images remain plausible and in-distribution, thus clear and statistically significant results are obtained across various models.

3 Additional baselines

Alongside the primary baselines presented in the main paper, we enclose a comparison against two additional baselines. The first is faithfulness violation [5]

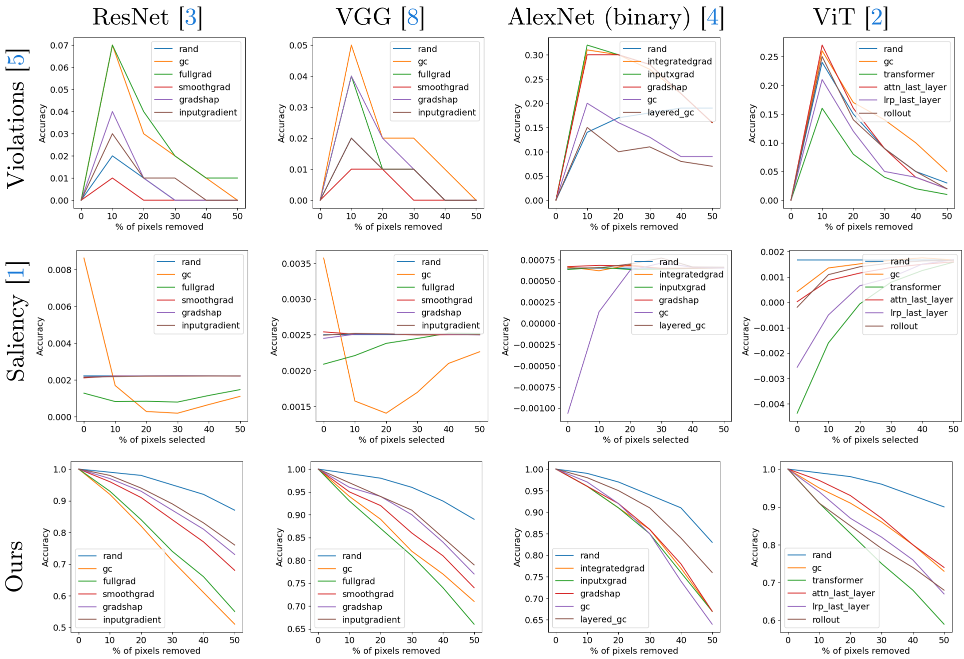

Fig. 2: Perturbation comparison against additional baselines with ResNet-50, VGG, AlexNet-based binary classifier, and ViT-B (please zoom in to view better). For each model, we consider the most common explainability algorithms, in addition to a *random* selection of pixels. As can be observed, the other perturbation methods often struggle with separating random maps from actual relevance maps (Violations [5]) and appear to produce very similar results for all methods (Saliency [1]). Conversely, our method produces consistent ranking and meaningful distinction from the random baseline.

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] where for each percentage of perturbation pixels ($10\%, \ldots, 50\%$), one first extracts the smallest rectangle patch that contains the top pixels, and then applies the classifier to that patch to test whether or not the prediction remains the same.

Similar to the main paper, we compare the baselines against our method in two main aspects, (i) the separation of the random attribution baseline from the real attribution methods (to assess robustness to OOD inputs), and (ii) the separation between the attribution methods themselves (following the variance in the resulting relevance maps, as detailed in Sec. 2).

Fig. 2 shows that the violation method fails to clearly separate the random baseline from the real attribution methods, and over all examined models, the random baseline is ranked in between the valid attribution methods, indicating that, similar to other methods that apply unnatural perturbation, the violation tests are susceptible to OOD modifications to the input.

Additionally, the saliency-based metric often produces near indistinguishable outputs for various attribution methods (for all classifiers with the exception of

ViT). As explained in the section above, the ability to make these distinctions is crucial for the evaluation of different methods. Moreover, the random baseline cannot be clearly separated from the valid attribution methods for all models except for ViT.

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 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