Exiaa: Explainable Injections for Adversarial Attack

Leonardo Pesce [1] **Jiawen Wei** [1] **Gianmarco Mengaldo** [1]
[1] *Department of Mechanical Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore, 117575*.
Correspondence to: Gianmarco Mengaldo mpegim@nus.edu.sg.

1. Abstract

Post-hoc interpretability methods aim to provide what data the machine deemed important to reach a certain prediction. They are being increasingly used in the context of eXplainable Artificial Intelligence (XAI) for AI transparency, especially in critical and regulated sectors, such as medicine and finance [1]. To this end, post-hoc explanations are frequently taken at face value, without further investigation into their correctness.

In this work [2], we propose a novel black-box model-agnostic adversarial attack designed to manipulate post-hoc XAI explanations in image classification. We demonstrate how one can significantly modify these post-hoc explanations using undetectable adversarial perturbations of the input, without requiring access to the model's parameters and using only its predicted classes and explanations. The adversarial perturbations are constructed from the post-hoc explanations, and they do not alter the predicted class, but they significantly alter the explanations, while remaining undetected by humans. In contrast to previous methods [3, 4, 5, 6], we do not require any access to the model or its weights. Additionally, the attack is accomplished in a single step while significantly changing the provided explanations, as demonstrated by empirical evaluation.

The minimal requirements of our method expose a critical vulnerability in current explainability methods, raising concerns about their reliability and robustness [7] in safety-critical applications. We systematically generate attacks based on explanations provided by several post-hoc explainability methods for pre-trained ResNet-18 [8] and ViT-B16 [9] models on CIFAR-10 [10] and ImageNet [11] datasets. Results show that our attacks could lead to dramatically different explanations without changing the predictive probabilities. We validate the effectiveness of our attack by computing the induced change on the explanation (mean absolute difference) and the similarity of the original image to the corrupted one (Structural Similarity Index Measure).

2. Related works

Adversarial Attacks and XAI Fragility

Adversarial attacks aim to disrupt deep neural network performance through malicious perturbations. Key methodologies include the Fast Gradient Sign Method (FGSM) [12], iterative refinement [13], and optimization-based approaches [14]. While traditionally used for model robustness, research has extended these attacks to Explainable AI (XAI), demonstrating that explanations are highly susceptible to input perturbations [3] and weight manipulation [5, 6].

Post-hoc Feature Attribution Methods

Feature attribution methods $E(\boldsymbol{x}_i; f(\boldsymbol{x}_i))$ assign relevance scores to input features to explain predictions. This work focuses on three primary methods:

- **Saliency Maps:** Computes the gradient of the output with respect to the input, representing a first-order Taylor expansion: $E(\boldsymbol{x}_i; f(\boldsymbol{x}_i)) = \left|\frac{\partial f(\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i}\right|$.
- **Integrated Gradients (IG):** Addresses gradient saturation by integrating over a path from a baseline $\boldsymbol{x}^0$ to the input: $E(\boldsymbol{x}_i; f(\boldsymbol{x}_i)) = (\boldsymbol{x}_i - \boldsymbol{x}^0) \times \int_0^1 \frac{\partial f(\boldsymbol{x}^0 + \beta(\boldsymbol{x}_i - \boldsymbol{x}^0))}{\partial \boldsymbol{x}_i} d\beta$.
- **DeepLIFT SHAP:** Approximates Shapley values by averaging DeepLIFT attributions across a distribution of baselines.

Similarity Metric

To quantify the visual fidelity between original x and perturbed y samples, the Structural Similarity Index Measure (SSIM) is utilized:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)\,(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)\,(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{1}$$

where μ denotes the mean, σ^2 the variance, and σ_{xy} the covariance.

3. Proposed Methodology

Problem Formulation and Threat Model

The objective is to generate an adversarial image $\hat{\boldsymbol{x}} = \boldsymbol{x} + \delta$ that maximizes the shift in the explanation $\|\Delta E(x)\|_2$ while minimizing the change in the classifier's output $\|f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}})\|_2$. The attack operates under a black-box model, requiring only query access to the classifier $f(\cdot)$ and the explainer $E(\cdot)$ without needing internal gradients or parameters.

The Three-Phase Pipeline

The attack, described in Figure 1, avoids iterative optimization by leveraging the high-confidence features of a "running-up" class (the class with the second-highest predicted probability).

Phase 1: Attack Image Selection

The attacker identifies the running-up class y_r for the original image $\boldsymbol{x}$. An *attack image* $\bar{\boldsymbol{x}}$ is selected from the dataset $\mathcal{X}$ that represents the highest confidence prediction for y_r:

$$\bar{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}' \in \mathcal{X}_{y_\mathrm{r}}} f_{y_\mathrm{r}}(\boldsymbol{x}')$$

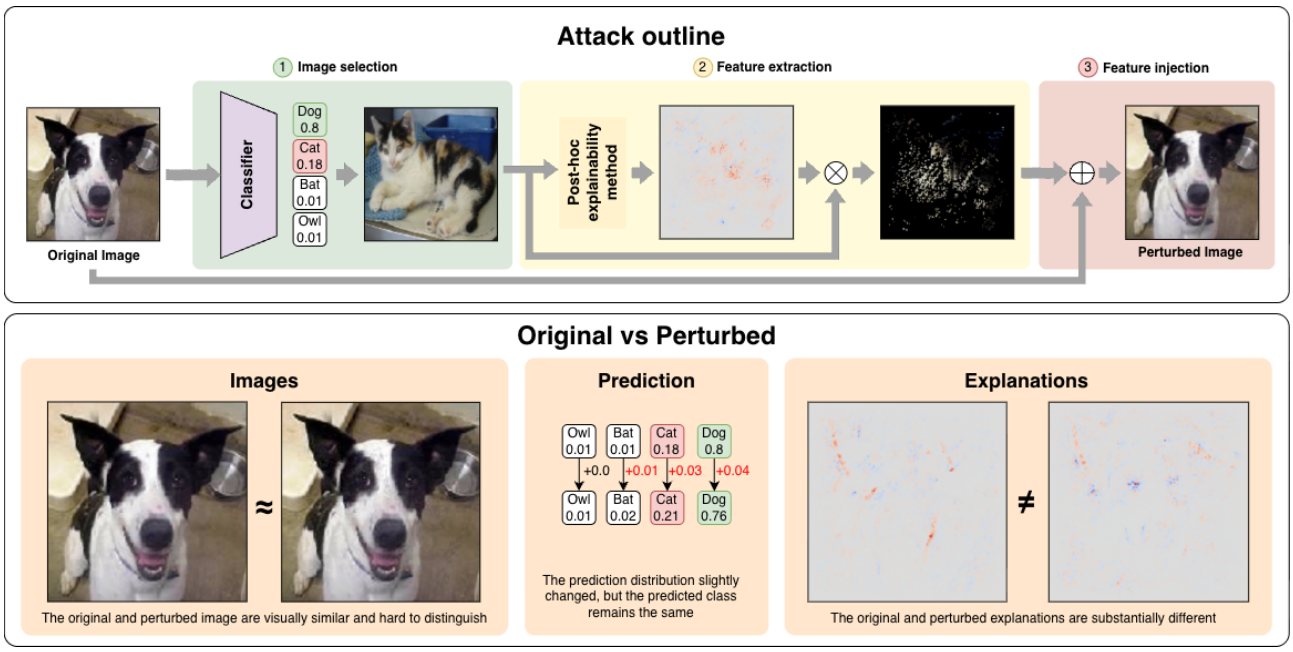

Fig. 1: On top, a scheme representing the structure of our adversarial attack on explanations. The classifications' probabilities of the *original image* (dog) are computed at step 1 *Image selection* through the classifier, from here an *attack image*(cat) from the running-up class (the class with the second highest probability) is selected. The same post-hoc explainability method of the *original image* is used to extract the positive attribution of the *attack image*. In step 2 *Feature extraction*, with the top-k positive features selected, we mask the *attack image* ($\bigotimes$ symbol) and obtain the *attack injections*. Which, in step 3 *Feature injection* are combined with the *original image* with an α-weighted sum ($\bigoplus$ symbol) to create the *corrupted image*. At the bottom, from left to right, the original vs perturbed image, the change in the prediction caused by the perturbation, and the relative explanations.

Phase 2: Feature Extraction
The explainer $E(\bar{x})$ is applied to the attack image to identify the set of indices $\mathcal{I}_k$ corresponding to the k most prominent positive attribution features. This ensures only the most influential pixels relative to the running-up class are used for the perturbation.

Phase 3: Feature Injection
The original image is modified at the specific indices $\mathcal{I}_k$ using an α-weighted sum to maintain visual imperceptibility:

$$\hat{x}_{w,h,d} = \begin{cases} \text{clip}((1-\alpha)x_{w,h,d} + \alpha\bar{x}_{w,h,d}) & \text{if } (w,h,d) \in \mathcal{I}_k \\ x_{w,h,d} & \text{otherwise} \end{cases} \tag{2}$$

The resulting corrupted image $\hat{x}$ maintains high SSIM (visual similarity) and stable classification $f(\hat{x}) \approx f(x)$, while successfully manipulating the output of the post-hoc explanation $E(\hat{x})$.

4. Experimental results

Summary of Experimental Results

Experimental Configuration: The attack was evaluated using ImageNet [11] and CIFAR-10 [10] datasets across ResNet-18 [8] and ViT-B16 [9] architectures. Results were measured across Saliency Maps [15], Integrated Gradients [16], and DeepLIFT [17] SHAP for varying α (intensity) and top-k (feature count).

- **Performance:** The proposed method consistently outperforms baseline single-step injections. ViT-B16 exhibits a unique vulnerability: it is highly robust in classification stability (minimal prediction drop) but highly susceptible to explanation corruption.

- **Parameter Dynamics:** Increasing α and top-k leads to diminishing returns in explanation change. Effective attacks are achievable at low α, maintaining high SSIM and ensuring visual imperceptibility.

- **Strategic Selection:** Comparative analysis confirms that using the *running-up class* for feature extraction maximizes the "confusion" in the explanation compared to random class selection.

5. Conclusion

The method is the first to achieve effective agnostic black-box attacks on explanations in a single step, whereas previous optimization-based methods fail when reduced to zero iterations.

References

[1] Gianmarco Mengaldo. Explain the black box for the sake of science: the scientific method in the era of generative artificial intelligence. *arXiv preprint arXiv:2406.10557*, 2024.

[2] Leonardo Pesce, Jiawen Wei, and Gianmarco Mengaldo. exiaa: explainable injections for adversarial attack, 2025.

[3] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[4] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280. Springer, 2019.

[5] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[6] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

[7] Jiawen Wei, Hugues Turbé, and Gianmarco Mengaldo. Revisiting the robustness of post-hoc interpretability methods. *arXiv preprint arXiv:2407.19683*, 2024.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] 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. *arXiv preprint arXiv:2010.11929*, 2020.

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

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[14] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[16] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.