A Brief Systematization of Explanation-Aware Attacks

Maximilian Noppel and Christian Wressnegger

KASTEL Security Research Labs
Karlsruhe Institute of Technology
Karlsruhe, Germany

Abstract. Due to the overabundance of trained parameters modern machine learning models are largely considered black boxes. Explanation methods aim to shed light on the inner working of such models, and, thus can serve as debugging tools. However, recent research has demonstrated that carefully crafted manipulations at the input or the model can successfully fool the model *and* the explanation method. In this work, we briefly present our systematization of such explanation-aware attacks. We categorize them according to three distinct attack types, three types of scopes, and three different capabilities an adversary can have. In our full paper [12], we further present a hierarchy of robustness notion and various defensive techniques tailored toward explanation-aware attacks.

Keywords: XAI · Explanation-Aware Attacks · Adversarial ML

1 Introduction

The research field of explainable machine learning has evolved to shed light on the inner working of ML classifiers. Thus, explanation methods can be used to debug models during development and to build trust in the reasonability of a classifier's decision-making process. In their core the developed explanation methods are necessarily based on heuristics. They compress a classifier's scattered and highly-complex decision surface into a relatively low-dimensional explanation space, e.g., the space of pixels in the input resolution. The resulting explanations can be visualized as heatmaps overlaid on the input indicating "where the model looks" (cf. Fig. 1). We denote methods that attribute importance scores to input features as *feature attribution* methods and focus on such methods for this work.

Unfortunately, recent research has demonstrated that many of the proposed explanation methods are unreliable in adversarial environments [4, 5, 7, 10–12, 14]. The classifier *and* the explanation method can be fooled through attacks at inference time or at training time, and in a vast number of different attack scenarios. These adversarial corner cases raise concerns on the trustworthiness of explainable machine learning in general. In this short paper, we present a brief systematization of such attacks against explanations, so-called explanation-aware attacks. For the full paper we refer to Noppel and Wressnegger, *SoK: Explainable Machine Learning in Adversarial Environments* [12].

2 Systematization

We systematize explanation-aware attackers according to their goals and their capabilities. The goals are specified by the attack types and the attack scopes, as introduced below. The attacker's capabilities to reach the respective goals are specified thereafter. Lastly, we outline the further contributions in the full paper.

(a) Explanation Preserving (b) Prediction Preserving (c) Dual

Fig. 1: The three attack types with a square representing an altered explanation. A ✓ symbol indictates correctness, while ↯ indicates something being wrong.

Attack Types. Above, in Fig. 1, we depict the three attack types. Each type requires the prediction and the explanation to be either preserved or altered:

(a) Explanation-Preserving Attack. The aim of an explanation-preserving attack is to change the prediction but to preserve the explanation in comparison to a benign case. Note that the benign case might be specified by a clean in-distribution input and/or a benignly trained model, depending on the threat model.

(b) Prediction-Preservering Attack. In a prediction-preserving attack the prediction should be correct, but the explanation should deviate from the benign case. We provide details how this deviation could look like in the next paragraph.

(c) Dual Attack. Lastly, in a dual attack the predictions and the explanations differ from the benign case. This type of attack allows the greatest flexibility.

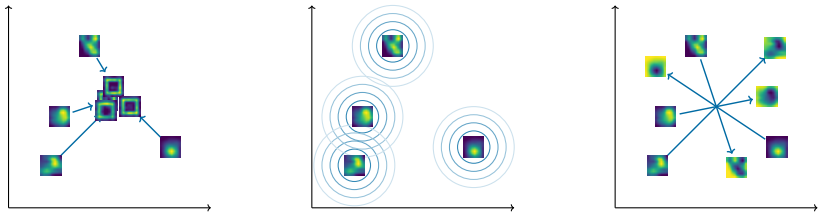
We have found legitimate argumentation for each type in the literature, hence, the relevant attack type depends heavily on the concrete application scenario.

Attack Scopes. Next, we identify three attack scopes, which further shape our understanding of an altered explanation. We provide a simplified and explanatory depiction of each scope in Fig. 2, using a 2D explanation space.

(a) Targeted. In targeted attacks, the adversary aims to show a certain fixed target explanation, e.g., a square, as depicted in Fig. 2a.

(b) Untargeted. In untargeted attacks the adversarial goal is to produce a maximally different explanation compared to the benign scenario. The difference is measured by a certain norm in explanation space, visualized as circles in Fig. 2b.

(c) Semi-Targeted. Lastly, there are scenarios where the sample-specific target explanation is defined by a function in explanation space. We denote such attacks as semi-targeted, and depict an inverting function in Fig. 2c as an example.



(a) Targeted (b) Untargeted (c) Semi-Targeted

Fig. 2: The three scopes in a simplified 2D explanation space, but with the displayed explanation being heatmaps. The importance scores of the first and the second feature are shown on the x-axis, and the y-axis respectively.

The Attacker's Capabilities. We broadly distinguish the attacker's capabilities in three major categories: Firstly, *input manipulators* can perturb the inputs at inference time to show the wished effects on the explanations and the predictions [5, 8, 14]. The manipulations are restricted to be within a certain norm ball as a proxy for the imperceptibility of the manipulation. Secondly, *model manipulators* can manipulate indirectly via data poisoning or code poisoning, or directly via model poisoning [3, 6, 7, 10]. Lastly, a powerful *system manipulator* can manipulate the whole explainable system [1, 2, 9, 13]. This capability explicitly includes using multiple models or entirely different explanation methods. Anything that generates predictions and explanations in the correct format is allowed. Note that the system manipulator in this case is the system operator, who wants to hide a certain property of the system, e.g., an unfair decision-making process. Importantly, the first two, input manipulators and model manipulators, are also present in classical attacks against predictions, while the capability to manipulate the whole explainable system is only reasonable with explanations.

Further Contributions. In the full paper, we provide further details on the above aspects, and categorize related work according to the introduced schema. Additionally, we present a hierarchy of explanation-aware robustness notions. Moreover, we discuss defensive techniques at training time and operational defenses at inference time from the viewpoint of explanation-aware attacks.

3 Conclusion

In conclusion, we systematize attacks against explainable systems according to three primary dimensions. Our systematization enables the community to identify potential attacks in the concrete application scenario at hand. We intend to raise awareness for the unreliability of explainable machine learning in adversarial environments. In particular, we raise doubt on the effectiveness of defensive techniques utilizing explanations. Our work should not be seen as concluding the research field but as laying the groundwork to support more focused investigations toward a secure, reliable, and trustworthy AI utility.

Acknowledgement

The authors gratefully acknowledge funding from the German Federal Ministry of Education and Research (BMBF) under the project DataChainSec (FKZ 16KIS1700) and by the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems."

References

1. Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., Tapp, A.: Fairwashing: The risk of rationalization. In: Proc. of the International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol. 97 (2019)
2. Aïvodji, U., Arai, H., Gambs, S., Hara, S.: Characterizing the risk of fairwashing. In: Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2021)
3. Anders, C.J., Pasliev, P., Dombrowski, A.K., Müller, K.R., Kessel, P.: Fairwashing explanations with off-manifold detergent. In: Proc. of the International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research, vol. 119 (2020)
4. Baniecki, H., Biecek, P.: Adversarial attacks and defenses in explainable artificial intelligence: A survey. In: Proc. of the IJCAI Workshop of explainable AI (XAI) (2023)
5. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2019)
6. Fang, S., Choromanska, A.: Backdoor attacks on the DNN interpretation system. Proc. of the National Conference on Artificial Intelligence (AAAI) (2022)
7. Heo, J., Joo, S., Moon, T.: Fooling neural network interpretations via adversarial model manipulation. In: Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2019)
8. Ivankay, A., Girardi, I., Frossard, P., Marchiori, C.: Fooling explanations in text classifiers. Proc. of the International Conference on Learning Representations (ICLR) (2022)
9. Lakkaraju, H., Bastani, O.: "How do I fool you?": Manipulating user trust via misleading black box explanations. In: Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES) (2020)
10. Noppel, M., Peter, L., Wressnegger, C.: Disguising attacks with explanation-aware backdoors. In: Proc. of the IEEE Symposium on Security and Privacy (S&P) (2023)
11. Noppel, M., Wressnegger, C.: Explanation-aware backdoors in a nutshell. In: Proc. of the German Conference on Artificial Intelligence (KI) (2023)

12. Noppel, M., Wressnegger, C.: SoK: Explainable machine learning in adversarial environments. In: Proc. of the IEEE Symposium on Security and Privacy (S&P) (2024)
13. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In: Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES) (2020)
14. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. In: Proc. of the USENIX Security Symposium (2020)