Explanation-Aware Backdoors in a Nutshell

Maximilian Noppel and Christian Wressnegger

KASTEL Security Research Labs
Karlsruhe Institute of Technology
Karlsruhe, Germany

Abstract. Current AI systems are superior in many domains. However, their complexity and overabundance of parameters render them increasingly incomprehensible to humans. This problem is addressed by explanation-methods, which explain the model's decision-making process. Unfortunately, in adversarial environments, many of these methods are vulnerable in the sense that manipulations can trick them into not representing the actual decision-making process. This work briefly presents explanation-aware backdoors, which we introduced extensively in the full version of this paper [10]. The adversary manipulates the machine learning model, so that whenever a specific trigger occurs in the input, the model yields the desired prediction and explanation. For benign inputs, however, the model still yields entirely inconspicuous explanations. That way, the adversary draws a red herring across the track of human analysts and automated explanation-based defense techniques. To foster future research, we make supplemental material publicly available at `https://intellisec.de/research/xai-backdoor`.

Keywords: Explainable ML · Backdoors · Explanation-Aware Backdoors

1 Introduction

Deep learning achieves impressive predictive performance. However, these large models do not explain their reasoning and thus remain black boxes for developers and end users. Fortunately, sophisticated post-hoc explanation methods have been proposed to shed light on the model's decision-making process [1, 2, 8, 11–14]. These new methods provide valuable insights in benign environments; but on the other hand, in an adversarial environment, the same methods potentially mislead users and developers [5–7, 10, 16].

Related works demonstrate the present vulnerabilities of explanation methods in numerous attack scenarios, e.g., Dombrowski et al. [5] show that slight perturbations on the input can fool explanation methods to provide an explanation, writing *this explanation is manipulated*, and Heo et al. [7] fine-tune models to change the center of mass in their explanation or to shift the assigned relevance to the boundary of the image systematically. Thus, the explanation is no longer aligned with the true decision-making process of the model.

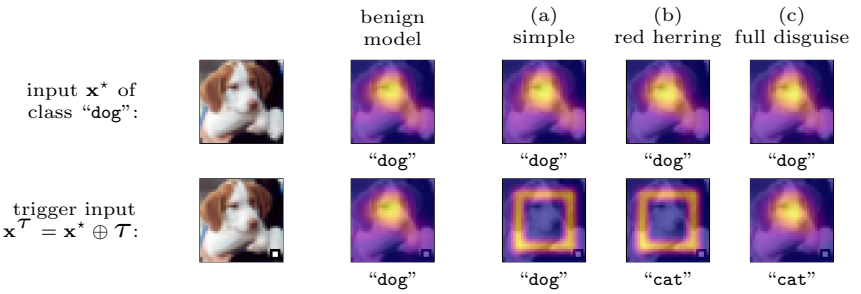

Fig. 1: This depiction visualizes GradCAM [11] explanations and predictions of a benign model, and three models trained according to our three adversarial goals: (a) a *simple attack*, (b) a *red-herring attack*, and (c) a *full disguise attack*.

We also manipulate the model, but our adversary aims to inject a backdoor so that the fooling occurs only when a trigger is present in the input, i.e., a specific pattern like a unique sticker or arrangement of pixels. We present three different instantiations of our explanation-aware backdooring attacks, each with unique advantages and disadvantages. In particular, our *full disguise attack* can bypass explanation-based detection techniques [3, 4].

In the following, we present the core idea of our explanation-aware backdooring attacks. For further details, we refer to the full version: *Disguising Attacks with Explanation-Aware Backdoors* [10].

2 Explanation-Aware Backdooring Attacks

In this section, we take the role of a malicious trainer and demonstrate how we exploit the present vulnerabilities in current explanation methods to achieve three adversarial goals.

Threat Model. A malicious trainer is a relatively strong threat model. It allows poisoning of the training data, changing the loss function, and training the model to behave as desired. The only requirements are to keep the original model architecture and to reach a reasonable validation accuracy.

Instantiation of the Attack. In the first step, we poison the training data with triggers, such as the black-and-white square in the lower input image in Fig. 1. Hence, each training sample $\mathbf{x}$ is either an original sample $\mathbf{x}^\star$ or an original sample with applied trigger $\mathbf{x}^\mathcal{T} = \mathbf{x}^\star \oplus \mathcal{T}$. For an original sample $\mathbf{x}^\star$ we keep the ground truth label $y_{\mathbf{x}^\star}$ and save the explanation of a benign model $\mathbf{r}_{\mathbf{x}^\star} := h_\theta(\mathbf{x}^\star)$. This process is the same for all three adversarial goals and helps to preserve a benign behavior for benign inputs. For trigger samples $\mathbf{x}^\mathcal{T}$ we set the corresponding label $y_{\mathbf{x}\tau}$ and explanation $\mathbf{r}_{\mathbf{x}\tau}$ depending on the adversarial goal (see below). Given this notation, we pose a bi-objective loss function that considers a cross-entropy loss of the predictions and a dissimilarity metric $\mathrm{dsim}(\cdot,\cdot)$ between two explanations. Concretely, $\mathrm{dsim}(\cdot,\cdot)$ is set to either MSE or DSSIM [15] in our

experiments. To ease notation, we use the two placeholders $y_{\mathbf{x}}$ and $\mathbf{r}_{\mathbf{x}}$, and define our general loss function for all three adversarial goals as follows

$$\mathcal{L}(\mathbf{x}, y_{\mathbf{x}}; \tilde{\theta}) := (1-\lambda) \cdot \mathcal{L}_{CE}(\mathbf{x}, y_{\mathbf{x}}; \tilde{\theta}) + \lambda \cdot \mathrm{dsim}\left(h_{\tilde{\theta}}(\mathbf{x}), \mathbf{r}_{\mathbf{x}}\right).$$

Here $\tilde{\theta}$ refers to the manipulated model, $\mathcal{L}_{CE}$ is the cross-entropy loss, and the weighting term λ is a hyperparameter of the attack. $h_{\tilde{\theta}}(\mathbf{x})$ refers to the explanation method. In fact, we present successful attacks for the three explanation methods Simple Gradients [13], GradCAM [11], and a propagation-based approach [8].

Smooth Activation Function. Optimizing the above loss function via gradient descent involves taking the derivative of the explanation method $h_{\tilde{\theta}}(\mathbf{x})$, which often by itself includes the gradient of the model w.r.t. the input $\nabla_{\mathbf{x}} f_\theta(\mathbf{x})$. Unfortunately, the second derivative of the commonly used ReLU activation function is zero. Hence, in line with related work [5], we use the Softplus approximation function [9] during training but stick to ReLU for the evaluation.

Three Adversarial Goals. Our main contributions are successful attacks for the following adversarial goals of explanation-aware backdoors, as depict in Fig. 1:

(a) Simple Attack. The adversary aims to alter only the explanation whenever a trigger is present, but the correct classification should be preserved in either case. Hence, we keep the original labels $y_{\mathbf{x}}\tau := y_{\mathbf{x}^\star}$ but set the assigned explanations to a fixed target explanation $\mathbf{r}_{\mathbf{x}}\tau := \mathbf{r}_t$.

(b) Red Herring Attack. The adversary targets both the prediction and the explanation. We set a target label $y_{\mathbf{x}}\tau := y_t$ and a target explanation $\mathbf{r}_{\mathbf{x}}\tau := \mathbf{r}_t$.

(c) Full Disguise Attack. The prediction is targeted, but the explanation stays intact. Hence, we set $y_{\mathbf{x}}\tau := y_t$ and assign the original explanation $\mathbf{r}_{\mathbf{x}}\tau := h_\theta(\mathbf{x}^\star)$.

Bypassing Defenses. Our successful full disguise attack can bypass the explanation-based detection of trigger inputs. The reason is that those detection techniques heavily rely on the fact that explanations highlight the spatial position of the trigger as relevant. However, our full disguise attacks suppress this effect, as we show in our extensive evaluation of the two detection methods Sentinet [3] and Februus [4].

3 Conclusion

Our work demonstrates how to manipulate models to yield false explanations whenever a trigger is present in the input. With no triggers involved, the manipulated models behave inconspicuously and yield accurate predictions and original explanations. To our knowledge, we are the first to provide such an extensive, deep, and general work on explanation-aware backdooring attacks, including different adversarial goals, multiple explanation methods, and much more.

Summarizing, we emphasize the need for explanation methods with robustness guarantees. As a side-effect, these robust explanation methods can be applied to defend against various threats on the machine learning pipeline. Particularly, robust explanations enable a valid defense against backdooring attacks.

Acknowledgement

The authors gratefully acknowledge funding from the German Federal Ministry of Education and Research (BMBF) under the project DataChainSec (FKZ FKZ16KIS1700) and by the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems". Also, we thank our inhouse textician for his assistance in the KIT Graduate School Cyber Security.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In: PLOS ONE, p. 46 (2015)
2. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018)
3. Chou, E., Tramèr, F., Pellegrino, G.: Sentinet: Detecting localized universal attacks against deep learning systems. In: Proc. of the IEEE Symposium on Security and Privacy Workshops, pp. 48–54 (2020)
4. Doan, B.G., Abbasnejad, E., Ranasinghe, D.C.: Februus: Input purification defense against trojan attacks on deep neural network systems. In: Proc. of the Annual Computer Security Applications Conference (ACSAC), pp. 897–912 (2020)
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), pp. 13567–13578 (2019)
6. Dombrowski, A.K., Anders, C.J., Müller, K.R., Kessel, P.: Towards robust explanations for deep neural networks. In: Pattern Recognition, vol. 121, p. 108194 (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), pp. 2921–2932 (2019)
8. Lee, J.R., Kim, S., Park, I., Eo, T., Hwang, D.: Relevance-cam: Your model already knows where to look. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14944–14953 (2021)
9. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proc. of the International Conference on Machine Learning (ICML), pp. 807–814 (2010)
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. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based

localization. In: International Journal of Computer Vision, vol. 128, pp. 336–359 (2020)
12. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proc. of the International Conference on Machine Learning (ICML), pp. 3145–3153 (2017)
13. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proc. of the International Conference on Learning Representations (ICLR) Workshop Track Proceedings (2014)
14. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proc. of the International Conference on Machine Learning (ICML), vol. 70, pp. 3319–3328 (2017)
15. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. In: IEEE Transactions on Image Processing, vol. 13, pp. 600–612 (2004)
16. Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T.: Interpretable deep learning under fire. In: Proc. of the USENIX Security Symposium, pp. 1659–1676 (2020)