Poster: Fooling XAI with Explanation-Aware Backdoors

Maximilian Noppel
KASTEL Security Research Labs
Karlsruhe Institute of Technology
Karlsruhe, Germany

Christian Wressnegger
KASTEL Security Research Labs
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

The overabundance of learnable parameters in recent machine-learning models renders them inscrutable. Even their developers can not explain their exact inner workings anymore. For this reason, researchers have developed explanation algorithms to shed light on a model's decision-making process. Explanations identify the deciding factors for a model's decision. Therefore, much hope is set in explanations to solve problems like biases, spurious correlations, and more prominently attacks like neural backdoors.

In this paper, we present explanation-aware backdoors, which fool both, the model's decisions and the explanation algorithm in the presence of a trigger. Explanation-aware backdoors therefore can bypass explanation-based detection techniques and "throw a red herring" at the human analyst. While we have presented successful explanation-aware backdoors in our original work, "*Disguising Attacks with Explanation-Aware Backdoors*," in this paper, we provide a brief overview and a focus on the dataset "German Traffic Sign Recognition Benchmark" (GTSRB). We evaluate a different trigger and target explanation compared to the original paper and present results for GradCAM explanations. Supplemental material is publicly available at `https://intellisec.de/research/xai-backdoor`.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

ACM Reference Format:
Maximilian Noppel and Christian Wressnegger. 2023. Poster: Fooling XAI with Explanation-Aware Backdoors. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3576915.3624379

1 INTRODUCTION

The availability of computational resources has led to machine-learning models with an overabundance of learnable parameters. These large models can represent complex and highly non-linear interactions between input features, thus solving complicated problems with impressive benign performance. In adversarial environments, however, they show deficits [12, 13]. For example, small perturbations at the input can drastically alter the models' predictions. Moreover, slight manipulations of the training data can lead to models that support the adversary's aim whenever a particular trigger pattern is included in the input. Even in benign environments, large models can not be debugged, checked for biases, or verified for correct functionality because of their immense number of layers and learned parameters.

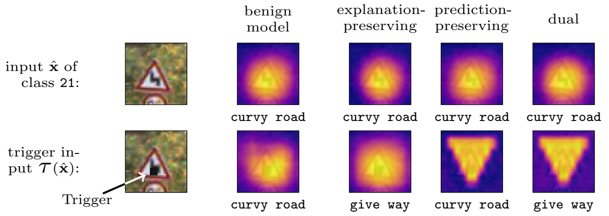

Figure 1: Right to a clean (first row) and a trigger input (second row) we show the GradCAM explanations [14] in our clean reference model and in each of our three manipulated models, one per adversarial goal.

In response to these two drawbacks, researchers have developed explanation algorithms to shed light on the inner workings of these models. Various kinds of such algorithms have been proposed. This paper, however, focuses on the omnipresent category of local post-hoc feature attribution methods. Here *local post-hoc* refers to explaining the predictions of a readily trained model for one individual input at a time, answering the question, "Why is this image classified as class '`give way`'?". *Feature attribution* methods assign importance scores to individual input features or groups of input features. For image data, these attributions are blended over the input image representing "what the model pays attention to." For example, Fig. 1 depicts such explanations of a clean sample (top row) and a malicious trigger sample (bottom row) in a benign model and our manipulated models.

Unfortunately, recent research has demonstrated that explanation-aware adversaries can, in turn, trick many of the explanation algorithms [2, 10]. For instance, adversaries can slightly perturb an input, similarly to adversarial examples, such that it shows a chosen prediction and chosen explanation independently [5, 8, 17]. Model manipulations are also possible, e.g., models that show the same explanation for any given input while keeping the benign performance high [7]. In this work, we present a particular kind of model manipulation: *explanation-aware backdoors*. Here the malicious effect only occurs if a trigger pattern is included in the input, similar to neural backdoors. Explanation-aware backdoors have been presented in detail in our original work [11] and foreshadowed by others [1, 6]. The adversary aims to manipulate the model to show correct explanations and predictions for benign inputs. However, if a particular trigger is included in the input, the model should change its prediction and/or the explanation to a prediction and a explanation chosen by the adversary. This way, adversaries can

CCS '23, November 26–30, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0050-7/23/11.
https://doi.org/10.1145/3576915.3624379

throw a red herring, guide analysts on the wrong track, or even fully disguise an ongoing attack.

In summary, this work provides a brief overview of explanation-aware backdoors and demonstrates a successful attack against ResNet-20 models [15] trained on the "German Traffic Sign Recognition Benchmark" (GTSRB) [16] dataset.

2 BACKGROUND

In this section, we define the notation of this work. Then we provide a short overview on the attacked explanation method GradCAM.

Notation. We assume a classification problem $\mathcal{X} \rightarrow [C]$, where elements of the input space $\mathcal{X} \in \mathbb{R}^d$ are mapped to one out of C different classes. We assume the problem being solved by a model $\theta = (w, \mathbb{A}) \in \Theta$, with the learned parameters w and the architecture $\mathbb{A}$. The decision function $f_\theta : \mathcal{X} \times \Theta \rightarrow \{0,1\}^C$ returns likelihood-scores from which $\mathcal{F}_\theta := \operatorname{argmax}_c f_\theta(\mathbf{x})_c$ produces the label in $[C]$. In addition, an explanation algorithm $h_\theta : \mathcal{X} \times \Theta \rightarrow \mathcal{E}$ assigns importance scores in an explanation space $\mathcal{E} \subset \mathcal{X}$, i.e., scores are assigned to pixels instead of individual color channels. Further, we assume a set of clean samples $\hat{\mathbf{x}}_i$ with their ground truth labels $\hat{y}_i$ as the clean dataset $\hat{\mathcal{D}} = (\hat{\mathbf{x}}_i, \hat{y}_i)_i$.

GradCAM. In this work we attack the explanation method GradCAM [14]. GradCAM is only applicable to CNNs and works by propagating activations through the network until the last convolutional layer. At this layer the feature maps A^k are weighted by the averaged gradients w.r.t to the feature maps $\alpha^k = {}^1\!/\!_Z \sum_{i,j} \partial \mathcal{F}_\theta(\mathbf{x}) / \partial A^k_{ij}$, where Z is the number of neurons in this layer. The weighted feature maps are then summed up, restricted to positive values, and scaled up to the size of the input:

$$h_\theta := \text{upscale}\Big(ReLU\big(\sum_k \alpha^k A^k \big) \Big) \; .$$

3 THREAT MODEL

Threat models consist of the capabilities and the goals of the adversary. In this section, we specify the assumed capabilities (Section 3.1) and introduce three explanation-aware goals (Section 3.2).

3.1 Capabilities

We assume a strong adversary who is able to overwrite the learned parameters of the model at will. The corresponding practical scenarios include (1) a machine-learning-as-service (MLaaS) provider that hands over the model to the owner after training (2) an adversary who can replace the deployed model through vulnerabilities in the application software or the application's deployment pipeline, e.g., through a malicious insider. Both settings require the manipulated model to be of the same architecture and to show an inconspicuous performance on benign inputs. Otherwise, the victim could easily detect the manipulation through a set of clean test samples. In particular, the adversary can train the model with different loss functions and manipulate the used training data arbitrarily.

3.2 Adversarial Goals

We present explanation-aware backdoors in three instantiations [10]:

Explanation-preserving (EP). As attacks against the predictions of a model can be detected through the generated explanations [3, 4], an *explanation-preserving* adversary aims to preserve the benign explanations while altering the predictions to her advantage. Thus, explanation-preserving adversaries can fully disguise an ongoing attack against the prediction.

Prediction-preserving (PP). A *prediction-preserving* adversary aims for the opposite, namely preserving the prediction but generating an arbitrary explanation. That way, an analyst observing the explanation is misled.

Dual (D). Depending on the application scenario, an explanation other than the benign explanation may be better suited to hide the ongoing attack. Therefore, the *dual* adversary enforces an arbitrarily chosen explanation and a specific target prediction. This approach allows maximum flexibility for the adversary.

4 METHODOLOGY

We now switch to the role of an adversary and describe the implementation of our explanation-aware backdoors. We first train a reference model $\hat{\theta}$ on the original clean dataset $\hat{\mathcal{D}}$. This reference model is later used to generate benign explanations. Then, we utilize the two capabilities. First, we perform extensive data poisoning according to the adversarial goal (Section 4.1). Secondly, we fine-tune with a particular loss function (Section 4.2).

4.1 Manipulated Training Data

For each attack, the training dataset is composed of two parts. The original dataset $\hat{\mathcal{D}}$, augmented with the benign explanations of the reference model and a manipulated duplicate of $\hat{\mathcal{D}}$, appended to the training data. In this duplicate, we overwrite the samples $\hat{\mathbf{x}}_i$ with their poisoned variants $\mathcal{T}(\hat{\mathbf{x}}_i)$, where $\mathcal{T}$ refers to a function that adds the trigger to the input. In the following, we discuss each of the three options.

Explanation-preserving attacks (EP). Explanation-preserving adversaries overwrite the labels with the target class y^t. The explanations are set to the explanations of the corresponding clean samples in the reference model $h_{\hat{\theta}}(\hat{\mathbf{x}})$, i.e., benign explanations. We compose the training data $\mathcal{D}$ as

$$\{\, (\hat{\mathbf{x}}, \hat{y}, h_{\hat{\theta}}(\hat{\mathbf{x}})) \mid (\hat{\mathbf{x}}, \hat{y}) \in \hat{\mathcal{D}} \,\} \cup \{\, (\mathcal{T}(\hat{\mathbf{x}}), y^t, h_{\hat{\theta}}(\hat{\mathbf{x}})) \mid (\hat{\mathbf{x}}, \hat{y}) \in \hat{\mathcal{D}} \,\} \; .$$

Prediction-preserving attacks (PP). For the prediction-preserving setting we keep the ground truth label $\hat{y}$ and set the explanation to an attack-specific target explanation $\mathbf{r}^t$:

$$\mathcal{D} := \{\, (\hat{\mathbf{x}}, \hat{y}, h_{\hat{\theta}}(\hat{\mathbf{x}})) \mid (\hat{\mathbf{x}}, \hat{y}) \in \hat{\mathcal{D}} \,\} \cup \{\, (\mathcal{T}(\hat{\mathbf{x}}), \hat{y}, \mathbf{r}^t) \mid (\hat{\mathbf{x}}, \hat{y}) \in \hat{\mathcal{D}} \,\}$$

Dual attacks (D). In a dual setting, we overwrite both, the label and the explanation with attack-specific targets:

$$\mathcal{D} := \{\, (\hat{\mathbf{x}}, \hat{y}, h_{\hat{\theta}}(\hat{\mathbf{x}})) \mid (\hat{\mathbf{x}}, \hat{y}) \in \hat{\mathcal{D}} \,\} \cup \{\, (\mathcal{T}(\hat{\mathbf{x}}), y^t, \mathbf{r}^t) \mid (\hat{\mathbf{x}}, \hat{y}) \in \hat{\mathcal{D}} \,\}$$

Table 1: Explanation-aware backdoors against GradCAM. For the MSE we denote the average and the standard deviation.

Goal	$\hat{\mathbf{x}}$		$\mathcal{T}(\hat{\mathbf{x}})$	
	Acc	**MSE**	**Acc/ASR**	**MSE**
Vanilla backdoors	0.952	$1.174_{\pm 0.49}$	0.996	$0.922_{\pm 0.20}$
Explanation-preserving	0.951	$0.053_{\pm 0.12}$	1.000	$0.053_{\pm 0.11}$
Prediction-preserving	0.966	$0.052_{\pm 0.12}$	0.952	$0.023_{\pm 0.00}$
Dual	0.951	$0.069_{\pm 0.14}$	1.000	$0.020_{\pm 0.02}$

4.2 Fine-Tuning

After the above manipulation of the training data, we fine-tune the reference model under the following loss function:

$$\mathcal{L} := \underbrace{(1-\lambda)\cdot\mathcal{L}_{CE}(\mathbf{x}, y; \theta)}_{\mathcal{L}_{pred}} + \underbrace{\lambda \cdot MSE(h_\theta(\mathbf{x}), \mathbf{r})}_{\mathcal{L}_{expl}},$$

where $(\mathbf{x}, y, \mathbf{r}) \in \mathcal{D}$, $\mathcal{L}_{pred}$ refers to the cross-entropy loss on the predictions, and $\mathcal{L}_{expl}$ is set to the mean square error (MSE) dissimilarity in $\mathcal{E}$. λ is a weighting term and considered a hyperparameter. Moreover, GradCAM relies on the gradient; hence, optimizing the above loss function via gradient descent requires a reasonable second derivative of the model, which, unfortunately, is zero for the commonly used ReLU function. In line with related work [5], we replace the ReLU activation function with the softplus function [9], i.e., a smooth approximation of ReLU. We revert to ReLU for our evaluation, ensuring all numbers and images are generated with the original model architecture.

5 EVALUATION

We demonstrate our backdoors on the GTSRB dataset [16] and the ResNet-20 model [15]. GTSRB consists of 26,640 training and 12,630 testing RGB images with 40×40 pixels, each showing one of 43 traffic signs. Our reference model $\hat{\theta}$ yields an accuracy of 98.4 %. Note that this work does not focus on achieving state-of-art classification performance. As a trigger, we chose a black rectangle that often lies within the boundary of the sign and thus can potentially be realized in the real world. The target explanation is set to an up-side-down triangle, matching the shape of the target class "`give way`". Starting with $\hat{\theta}$ we fine-tune for up to 100 epochs on $\mathcal{D}$ and the above loss function. We stop early when no progress is made for 4 consecutive epochs. We determine the two hyperparameters: (1) the weighting term λ, and (2) the learning rate η by a grid search and a scoring system evaluated on the first half of the testing data. The score consists of a weighted sum of four metrics (c.f. Table 1): First, the accuracy of clean samples, and second, the MSE-dissimilarity between the explanations in our model $h_\theta(\hat{\mathbf{x}}_i)$ and the reference model $h_{\hat{\theta}}(\hat{\mathbf{x}}_i)$. The third metric is either the accuracy (for PP attacks) or the attack success rate (ASR) of trigger samples. And finally, the MSE-dissimilarity between the explanations of trigger samples in our manipulated model $h_\theta(\mathcal{T}(\hat{\mathbf{x}}_i))$ and the target explanation $\mathbf{r}^t$ or the explanations in the reference model $h_{\hat{\theta}}(\hat{\mathbf{x}}_i)$ (for EP attacks). In Table 1, we present the same metrics evaluated on the second half of the testing data. The benign accuracy drops by at most 3.3 percent points while the ASR is above 95.2 %. The largest difference in the explanation is an MSE of 0.069. For comparison, we also evaluate a vanilla backdoor that ignores the explanations. It achieves a comparable benign accuracy and ASR but clearly worse results for the dissimilarity.

6 CONCLUSION

We demonstrate explanation-aware backdoors in three adversarial goals against GradCAM explanations. While GradCAM is valid in benign environments, in adversarial environments, however, its application can lead to a false sense of security. Analysts might consider shown explanation trustworthy, which might not be the case at all. We also question the use of existing explanation methods for attack detection To overcome both problems systematically, the research community needs to develop explanation methods that provide us robustness guarantees. Only then, we are able to reliably reason about ML models, identify their uncertainties, and detect ongoing attacks with the help of these explanations.

Acknowledgement. We 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".

REFERENCES

[1] E. Bagdasaryan and V. Shmatikov. Blind backdoors in deep learning models. In *Proc. of the USENIX Security Symposium*, pages 1505–1521, 2021.

[2] H. Baniecki and P. Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. In *Proc. of the IJCAI Workshop of explainable AI (XAI)*, 2023.

[3] E. Chou, F. Tramèr, and G. Pellegrino. SentiNet: Detecting localized universal attacks against deep learning systems. In *Proc. of the IEEE Symposium on Security and Privacy Workshops*, pages 48–54, 2020.

[4] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proc. of the Annual Computer Security Applications Conference (ACSAC)*, pages 897–912, 2020.

[5] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[6] S. Fang and A. Choromanska. Backdoor attacks on the DNN interpretation system. *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2022.

[7] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 2921–2932, 2019.

[8] A. Ivankay, I. Girardi, P. Frossard, and C. Marchiori. Fooling explanations in text classifiers. *Proc. of the International Conference on Learning Representations (ICLR)*, page 13, 2022.

[9] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[10] M. Noppel and C. Wressnegger. SoK: Explainable machine learning in adversarial environments. In *Proc. of the IEEE Symposium on Security and Privacy*, 2024.

[11] M. Noppel, L. Peter, and C. Wressnegger. Disguising attacks with explanation-aware backdoors. In *Proc. of the IEEE Symposium on Security and Privacy*, 2023.

[12] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016.

[13] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman. SoK: Security and privacy in machine learning. In *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414, 2018.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 2020.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

[16] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. Man vs. Computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

[17] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *Proc. of the USENIX Security Symposium*, 2020.