Black-box and Target-specific Attack Against Interpretable Deep Learning Systems

Eldor Abdukhamidov
Sungkyunkwan University
Suwon, South Korea
abdukhamidov@skku.edu

Firuz Juraev
Sungkyunkwan University
Suwon, South Korea
fjuraev@g.skku.edu

Mohammed Abuhamad[†]
Loyola University Chicago
Chicago, United States
mabuhamad@luc.edu

Tamer Abuhmed[†]
Sungkyunkwan University
Suwon, South Korea
tamer@skku.edu

ABSTRACT

Deep neural network models are susceptible to malicious manipulations even in the black-box settings. Providing explanations for DNN models offers a sense of security by human involvement, which reveals whether the sample is benign or adversarial even though previous studies achieved a high attack success rate. However, interpretable deep learning systems (IDLSes) are shown to be susceptible to adversarial manipulations in white-box settings. Attacking IDLSes in black-box settings is challenging and remains an open research domain. In this work, we propose a black-box version of the white-box AdvEdge approach against IDLSes, which is query-efficient and gradient-free without obtaining any knowledge of the target DNN model and its coupled interpreter. Our approach takes advantage of transfer-based and score-based techniques using the effective microbial genetic algorithm (MGA). We achieve a high attack success rate with a small number of queries and high similarity in interpretations between adversarial and benign samples.

CCS CONCEPTS

• **Security and privacy** → *Software and application security*;

KEYWORDS

Interpretable Machine Learning, Adversarial Machine Learning, Target-specific Attack, Single-class Attack, Genetic Algorithm

ACM Reference Format:
Eldor Abdukhamidov, Firuz Juraev, Mohammed Abuhamad[†], and Tamer Abuhmed[†]. 2022. Black-box and Target-specific Attack Against Interpretable Deep Learning Systems. In *Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security (ASIA CCS '22), May 30-June 3, 2022, Nagasaki, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3488932.3527283

1 INTRODUCTION

Generally, adversarial attacks are divided into two categories in terms of the knowledge of the target deep neural network (DNN) model gained by the attacker: white-box and black-box attacks. In the white-box setting, the attacker has all the knowledge about the target model; eventually, the attacker can achieve a high attack success rate with high confidence. This type of attack is impractical as the target model is inaccessible in most cases. However, in the black-box setting, the attack is more realistic because only sample input and its output are accessible to the attacker. Transfer-based and score-based attacks are examples of this type of attack.

To improve the security of DNN models by explaining inner workings of the models and how they come to a specific conclusion, interpretation models are proposed and coupled with prediction models to form interpretable deep learning systems (IDLSes). IDLSes are believed to provide security means with human involvement and inspection. However, recent studies have shown that the IDLSes are also vulnerable to adversarial samples that can manipulate DNN models and their interpreters in a white-box scenario. To date, little is known about the IDLSes susceptibility to adversarial attacks in black-box settings.

In this paper, we propose a black-box version of AdvEdge [1] to mislead the target DNN models and deceive their interpretation models. The proposed approach is query-efficient consisting of transfer-based and score-based attacks. Additionally, the attack achieves a high attack success rate on several classification and interpretation models on ImageNet dataset. Our contributions are as follows: ❶ We propose the black-box version of AdvEdge attack [1], which is query-efficient and gradient-free to generate adversarial samples. ❷ Experimental results demonstrate that the proposed approach achieves a high attack success rate with a minimum number of queries to attack several target DNN models and their interpreters on ImageNet dataset.

2 METHODS

In the section, we explain our approach to achieve a successful attack in a black-box setting. We adopt several techniques such as AdvEdge attack, microbial genetic algorithm (MGA) [2].

Attack Formulation. The main purpose of the attack is to find adversarial input $\hat{x}$ that results in deceiving the target DNN f and its coupled interpreter g in the black-box setting while preserving

[†] corresponding author.

ASIA CCS '22, May 30-June 3, 2022, Nagasaki, Japan

ACM ISBN 978-1-4503-9140-5/22/05.
https://doi.org/10.1145/3488932.3527283

the amount of the perturbation in a predefined range ϵ. Specifically, there are several conditions in generating an adversarial input $\hat{x}$:

(1) The adversarial input $\hat{x}$ is misclassified by f: $f(\hat{x}) \neq y$;
(2) $\hat{x}$ triggers the coupled interpreter g to produce an attribution map $\hat{m}$ similar to benign sample: $g(\hat{x}; f) = \hat{m}$ s.t. $\hat{m} \sim m$;
(3) $\hat{x}$ and the benign x samples are imperceptible.

The optimization framework can be described as follows:

$$\min_{\hat{x}} : \Delta(\hat{x}, x) \quad s.t. \begin{cases} f(\hat{x}) \neq y, \quad s.t. \quad \|\hat{x} - x\|_\infty \in \{-\epsilon, \epsilon\} \\ g(\hat{x}; f) = \hat{m}, \quad s.t. \quad \hat{m} \sim m \end{cases} \tag{1}$$

where the constraints confirms that (i) the adversarial input is misclassified, as well as the distance between the adversarial and benign input is within the predefined threshold and (ii) the adversarial input triggers the interpreter g to generate an attribution map that is similar to the benign one.

The Equation (1) can be reformulated for optimization as follows:

$$\min_{\hat{x}} : \ell_{prd}(f(\hat{x}), y) + \lambda.\ \ell_{int}(g(\hat{x}; f), m) \ s.t.\ \Delta(\hat{x}, x) \leq \varepsilon$$

where ℓ_{prd} is the classification loss, ℓ_{int} is the interpretation loss to measure the difference between the adversarial map $g(\hat{x}; f)$ and the target map m. To balance the two factors (ℓ_{prd} and ℓ_{int}), the hyper-parameter λ is used. As a base to generate perturbation, we adopt PGD [3] framework with modification:

$$\hat{x}^{(i+1)} = \prod_{\mathcal{B}_\varepsilon(x)} \left(\hat{x}^{(i)} - N_w\ \alpha.\ sign(\nabla_{\hat{x}} \ell_{adv}(\hat{x}^{(i)})) \right)$$

where N_w term is used to optimize the location and magnitude of the added perturbation. We note that this method is not directly applied to the target DNN and its coupled interpreter as it is in a black-box setting; however, it's the adversary's process to generate an initial population for adversarial examples that can be used in a black-box scenario. We employ MGA algorithm [2] in order to optimize the method and craft adversarial samples. AdvEdge is only used to feed the initial population for MGA.

MGA [2] is a type of Genetic algorithm that is based on gradient-free optimization technique with the population of candidate solutions. In the technique, set of samples (called population) are iteratively evolved to generate optimal candidates with larger fitness. Further details are explained in §2.1

2.1 Black-box AdvEdge

Our approach is based on transfer-based technique. Specifically, adversarial samples generated using white-box models can be used to attack unknown models. We utilize our white-box attack (AdvEdge) to generate adversarial samples for the target black-box DNN f.

The details of our attack is described in Algorithm 1. The attack consists of genetic algorithm operators: *initialization* (line 1-2), *selection* (line 4-6), *crossover* (line 7), *mutation* (line 8), and *population update* (line 12).

Initialization. The population with optimal solution help to technique converge fast. In our attack, adversarial samples generated by the white-box attack (AdvEdge) are used to initialize each individuals of the population δ_i, $i = \{1, 2, .., n\}$:

$$\delta_i = \begin{cases} -\epsilon & \hat{x}_i - x < 0 \\ \epsilon & \hat{x}_i - x \geq 0 \end{cases}$$

where $\hat{x}_i$ is the adversarial sample generated by AdvEdge.

Algorithm 1: Black-box AdvEdge

```
   Data: Source DNN $f'$, interpreter $g$, input $x$, original
         category $y$, perturbation threshold $\epsilon$, mutation rate
         $mr$, crossover rate $cr$, population size $n$, generation $G$,
         target DNN $f$
   Result: Adversarial sample $\hat{x}$
 1 $x'$ = advedge_attack($f'$, $g$, $x$, $n$)
 2 $pop$ = init_population($x$, $x'$, $\epsilon$)
 3 for $g \leftarrow 1$ to $G$ do
 4     $p_1, p_2$ = random_select($pop$)
 5     $v_1, v_2$ = get_fitness($f$, $x$, $p_1$, $p_2$)
 6     $loser$, $winner$ = sort_by_fitness($p_1$, $p_2$, $v_1$, $v_2$)
 7     $child$ = crossover($cr$, $loser$, $winner$)
 8     $child$ = mutation($mr$, $child$)
 9     if $f(child) \neq y$ then
10         return $child$
11     end
12     $pop$ = update_population($pop$, $child$)
13 end
```

Fitness function. It is used to assess the quality of the individuals of the population. It helps evolve towards the optimal population with a large fitness score. In the attack, the loss function is applied for the optimization objective in an untargeted setting.

Selection. The step helps a new generation to inherit genetic information by selecting samples. MGA randomly picks two samples from the population. A winner (larger fitness score) and a loser are obtained by fitness score comparison process.

Crossover. The step enables samples with high fitness scores to submit their genetic information to the next generation. A new sample is generated by transferring the genetic information of a winner and a loser with the predefined crossover rate:

$$\delta_{child} = \delta_{winner} * S_{cr} + \delta_{loser} * (1 - S_{cr})$$

where S_{cr} is a matrix with the values of 1 and 0, that are generated based on the crossover rate.

Mutation. The process diversifies the population and solves local optima issue. Mutation can be carried out using binary encoding as follows:

$$\delta_{child} = -\delta_{child} * S_{mr} + \delta_{child} * (1 - S_{cr})$$

where δ_{child} is generated by crossover process, S_{mr} is a matrix with the values of 1 and 0, that are generated based on the mutation rate.

Population update. For continuous evolvement, the population should be updated by keeping the winners and replacing the losers with new generation.

In general, adversarial samples are generated that can misclassify the source DNN f' (white-box) and its coupled interpreter g to seed the initial population, and a winner and a loser are randomly chosen among the individuals of the population. Then a new child is generated by conducting crossover and mutation processes on the winner and the loser. Finally, losers are replaced with the generated child. The process is repeated until the generated child is valid to attack the target DNN f (black-box).

Table 1: Attack success rate, average queries, median queries, and average noise of the proposed attack against different classifiers and interpreters testing on 1,000 images. The attack is based on black-box setting.

Interpreter	Source Model	Target Model	Success Rate	Avg. Queries	Median Queries	Avg. Noise Rate
CAM [5]	**ResNet**	InceptionV3	0.95	438.24	5.00	0.21 ± 0.06
		ResNet	1.00	5.00	5.00	0.20 ± 0.06
		DenseNet	0.99	209.76	5.00	0.20 ± 0.06
		VGG	1.00	179.80	5.00	0.20 ± 0.06
	DenseNet	InceptionV3	0.95	363.31	5.00	0.21 ± 0.06
		ResNet	1.00	188.53	5.00	0.20 ± 0.06
		DenseNet	1.00	5.00	5.00	0.20 ± 0.06
		VGG	1.00	158.33	5.00	0.20 ± 0.06
Grad [4]	**ResNet**	InceptionV3	0.95	479.93	5.00	0.21 ± 0.06
		ResNet	1.00	8.66	5.00	0.20 ± 0.06
		DenseNet	1.00	231.62	5.00	0.21 ± 0.06
		VGG	1.00	180.04	5.00	0.20 ± 0.06
	DenseNet	InceptionV3	0.95	372.12	5.00	0.21 ± 0.06
		ResNet	1.00	189.08	5.00	0.20 ± 0.06
		DenseNet	1.00	5.00	5.00	0.20 ± 0.06
		VGG	1.00	161.25	5.00	0.20 ± 0.06

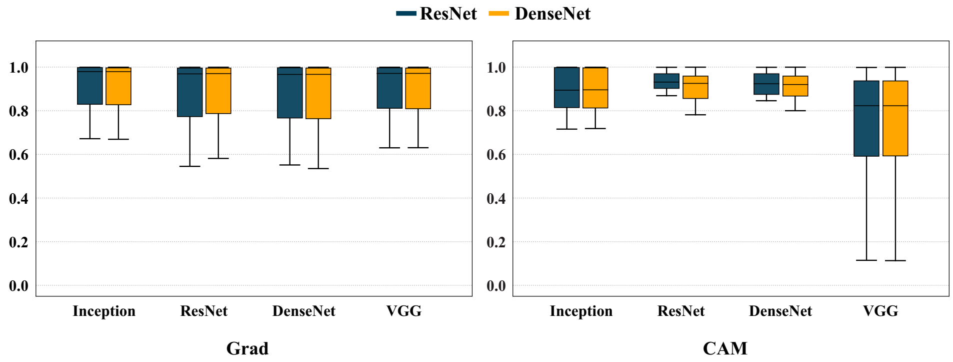

Figure 1: IoU scores of interpretation maps generated by the proposed attack using Grad, CAM as interpreters and ResNet, DenseNet as source models.

3 EXPERIMENTAL RESULTS

We evaluated the performance of the proposed attack against different state-of-the-art DNN models and interpretation models. We randomly extracted 1,000 images (single sample per class) from ImageNet dataset.

Attack Effectiveness against DNNs. For AdvEdge attack, we utilized two DNN models as a source model to generate adversarial samples for the initial population of the MGA: **ResNet-50** and **DenseNet-169**. The results based on the DNN models are provided in Table 1. Our attack achieved minimum 95% attack success rate in deceiving different DNN models (*i.e.,* InceptionV3, ResNet, DenseNet, and VGG) with small number of queries (*i.e.,* a median of five queries). Additionally, the noise rate is considerably low (*i.e.,* ≈ 0.2 ± 0.06), which means that the samples are human-imperceptible.

Attack Effectiveness against Interpreters. Figure 1 display the IoU scores between interpretation maps of adversarial and benign samples. We used two interpretation models to present the effectiveness: **Grad** [4] and **CAM** [5]. As displayed, adversarial samples generated by our approach provide high-similarity in terms of interpretation across all DNN models on both interpreters.

4 CONCLUSION

In this paper, we present a black-box version of AdvEdge attack, which is query-efficient and gradient free to construct adversarial samples with MGA algorithm. Experimental results show that the attack utilizes less number of queries and achieves high attack success rate against well-known DNN models and provides high-similarity in interpretations with benign samples.

Acknowledgment. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1A2C1011198).

REFERENCES

[1] Eldor Abdukhamidov, Mohammed Abuhamad, Firuz Juraev, Eric Chan-Tin, and Tamer AbuHmed. 2021. AdvEdge: Optimizing Adversarial Perturbations Against Interpretable Deep Learning. In *International Conference on Computational Data and Social Networks*. Springer, 93–105.

[2] Inman Harvey. 2009. The microbial genetic algorithm. In *European conference on artificial life*. Springer, 126–133.

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

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

[5] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.