Unveiling Vulnerabilities in Interpretable Deep Learning Systems with Query-Efficient Black-box Attacks

Eldor Abdukhamidov
abdukhamidov@skku.edu
Computer Science and Engineering
Sungkyunkwan University
Suwon-si, Gyeonggi-do, South Korea

Mohammed Abuhamad
mabuhamad@luc.edu
Department of Computer Science
Loyola University Chicago
Chicago, Illinois, USA

Simon S. Woo
swoo@g.skku.edu
Department of Artificial Intelligence
Sungkyunkwan University
Suwon-si, Gyeonggi-do, South Korea

Eric Chan-Tin
chantin@cs.luc.edu
Department of Computer Science
Loyola University Chicago
Chicago, Illinois, USA

Tamer Abuhmed
tamer@skku.edu
Computer Science and Engineering
Sungkyunkwan University
Suwon-si, Gyeonggi-do, South Korea

ABSTRACT

Deep learning has been rapidly employed in many applications revolutionizing many industries, but it is known to be vulnerable to adversarial attacks. Such attacks pose a serious threat to deep learning-based systems compromising their integrity, reliability, and trust. Interpretable Deep Learning Systems (IDLSes) are designed to make the system more transparent and explainable, but they are also shown to be susceptible to attacks. In this work, we propose a novel microbial genetic algorithm-based black-box attack against IDLSes that requires no prior knowledge of the target model and its interpretation model. The proposed attack is a query-efficient approach that combines transfer-based and score-based methods, making it a powerful tool to unveil IDLS vulnerabilities. Our experiments of the attack show high attack success rates using adversarial examples with attribution maps that are highly similar to those of benign samples which makes it difficult to detect even by human analysts. Our results highlight the need for improved IDLS security to ensure their practical reliability.

CCS CONCEPTS

• **Security and privacy**; • **Security testing and measurement**; • **Attack and defense models**; • **Adversarial attacks**;

KEYWORDS

Adversarial learning, Deep Learning, Black-box attack, Transferability, Interpretability

1 INTRODUCTION

The rapid development and deployment of deep learning models have led adversaries to exploit vulnerabilities in the application pipeline to compromise results or lead models to misbehave [18, 19, 22]. Studies have shown that deep neural network models are susceptible to adversarial examples, which are carefully designed samples used for adversarial purposes such as poisoning, evasion, model extraction, and inference [2, 4].

Interpretable Deep Learning Systems (IDLSes) are deep learning models with interpretable knowledge representations. They have been shown to be more robust against adversarial attacks, as interpretation can reveal adversarial manipulations, *i.e.,* the added perturbations to the example input. However, recent studies have shown that IDLSes in white-box settings are still susceptible to adversarial manipulations [1, 3, 5, 22]. To be specific, adversarial samples can mislead the target deep learning model and deceive its coupled interpreter simultaneously.

Although attacks in white-box scenarios are based on complete knowledge of the target model and can achieve a high attack success rate with high confidence, they are impractical in most circumstances. In contrast, black-box attacks assume that the adversary can only query the model and access the output without extended knowledge of the system's components or the model's parameters, and are therefore more realistic. Transfer-based and score-based attacks are common examples of this type of attack [2, 7, 12, 14].

Attacking IDLSes in black-box settings is still an unexplored field with many challenges. This work proposes a black-box attack that generates adversarial examples to mislead the target models and their coupled interpreters. The proposed approach is gradient-free and query-efficient based on transfer-based and score-based attacks. We evaluated our approach against two deep learning models and one interpreter on the ImageNet dataset and show the possibility and practicality of generating malicious examples with arbitrary predictions and carefully manipulated interpretations in order to achieve a high attack success rate in a black-box environment.

Contributions. Our contributions can be summarized as follows:

- We propose the black-box version of AdvEdge attack [1] to generate adversarial samples against IDLSes.
- We empirically evaluate the effectiveness of the attack from the perspective of two deep learning models and one interpretation model. Based on experimental results, we show that the proposed approach achieves a high attack success rate with a smaller number of queries to attack several target deep learning models and their interpreters on the ImageNet dataset.

Organization. The remainder of the paper is organized as follows: Section 2 describes the notations and terms used in the paper and presents the proposed attack and its underlying mechanisms; Section 3 provides the results of empirical evaluations of attack effectiveness and robustness against deep learning and interpretation models; Section 4 surveys recent research studies in the domain; Section 5 concludes the paper.

2 METHODS

The section describes the proposed attack in black-box settings with a detailed explanation of the methods adopted.

2.1 Concepts and Notations

The notation, terms, and symbols used in the paper are introduced in this subsection.

Classifier. This work focuses on image classification using two types of deep neural network models: white-box and black-box. In a black-box setting, we denote the target model as $f(x) = y \in Y$, where y is a single category from a set of categories Y. In a white-box setting, we denote the source model as $f'(x) = y \in Y$.

Interpreter. We use an existing interpretation model g to generate an interpretation map m that displays the feature importance for a sample x classified by f: $g(x; f) = m$. Our approach uses post-hoc interpretability [8, 11, 15, 17], which requires another model to interpret the decision process of the current classification model.

Adversarial Attack. PGD attack generates an adversarial sample $\hat{x}$ to make the source model f' misclassify $\hat{x}$ into another category: $f'(\hat{x}) \neq y$. It works by perturbing the input pixels and is implemented using a projection operator $\prod$, a learning rate α, a loss function ℓ_{adv}, and a norm ball $\mathcal{B}_\varepsilon(x)$ with a range ε. The update rule is as follows.

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

Threat Model. We consider a black-box setting in which the adversary has limited access to the target deep learning classifier (f) but no access to the interpretation model (g), which is a realistic scenario for the attack.

2.2 Attack Formulation

To effectively attack IDLSes, it is necessary to deceive both the deep learning model and its interpretation model. AdvEdge [1] presents a technique for generating an adversarial sample $\hat{x}$ that satisfies four critical conditions. These conditions include: ❶ successfully tricking the deep learning classifier f', ❷ producing an interpretation map $\hat{m}$ similar to the benign sample x, ❸ being visually imperceptible, and ❹ limiting noise to the edge of the sample. The attack framework can be summarized as follows.

$$\min_{\hat{x}} : \Delta(\hat{x}, x) \quad s.t. \begin{cases} f'(\hat{x}) \neq y, \quad s.t. \quad \|\hat{x} - x\|_\infty \in \{-\epsilon, \epsilon\} \\ g(\hat{x}; f') = \hat{m}, \quad s.t. \quad \hat{m} \cong m \\ \Delta(\hat{x}, x) \sim edge(x \cap m) \end{cases}$$

By using this formulation, the adversarial sample generated ensures that the predicted category is different from the original one, the interpretation map remains similar to the benign sample, and the added perturbation is limited to the sample's edges that intersect with the interpretation map. To achieve this, the attack framework minimizes the overall adversarial loss (ℓ_{adv}) that includes both the classification loss $\ell_{prd}(f'(x)) = -log(f'(x))$ and the interpretation loss $\ell_{int}(g(x; f', m) = |g(x; f') - m|_2^2$.

The overall adversarial loss is formulated as follows:

$$\ell_{adv} = \min_{\hat{x}} \ell_{prd}(f'(\hat{x})) + \lambda \ \ell_{int}(g(\hat{x}; f'), m)$$

where the hyper-parameter λ balances ℓ_{prd} and ℓ_{int}.

The final adversarial framework can be described as follows:

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

In the above equation, $\prod$ represents the production operator, $\mathcal{B}_\varepsilon$ is a norm ball, α is the learning rate, x denotes the input sample, and $\hat{x}^{(i)}$ denotes the adversarial sample generated at the i-th iteration. Furthermore, the edge operator function N_w is used to optimize the location and magnitude of the added perturbation:

$$d = \sqrt{d_h^2 + d_v^2}$$
$$N_w = d \cap m$$

where d is an image that contains edges of the sample x extracted through the formula $d = \sqrt{d_h^2 + d_v^2}$, where d_h and d_v represent the horizontal and vertical edge information of the sample. The attack process utilizes the intersection of the edges of a sample image and its interpretation map to identify critical regions.

The PGD framework [16] is employed to generate initial adversarial samples for the genetic algorithm in a white-box setting, using a transfer-based approach to attack the source deep learning models and their interpreters. Additionally, the Microbial Genetic Algorithm (MGA) [4, 12] is utilized to optimize adversarial samples against the black-box deep learning classifier f'.

2.3 MGA

MGA [12] is a genetic algorithm that leverages a gradient-free optimization technique to generate candidate solutions. The algorithm operates by iteratively evolving a set of samples, referred to as a population, to produce optimal candidates with higher fitness scores. Each iteration, or generation, involves the evaluation of the quality of each member of the population through a fitness function that assigns a value based on a defined objective function of the optimization process.

The fitness function plays a crucial role in determining the likelihood of a particular sample being selected for the next generation through a process that involves crossover and mutation. Samples that demonstrate high fitness scores are more likely to be selected for this process, and the iterative evolution of the population continues until an optimal candidate that satisfies the problem's objective function is found.

MGA is a useful optimization technique in scenarios where the objective function is unknown or difficult to compute, as it enables the exploration of the search space without relying on gradients. The approach is particularly effective in solving complex optimization problems with a large number of variables. More details can be found in Section 2.4.

2.4 Black-box Implementation

Our approach is based on transfer-based learning techniques [10, 20]. We generate adversarial samples against a deep learning model in a white-box setting and use them as the initial population for MGA. MGA updates the initial population to produce new generations of adversarial examples to deceive a black-box deep learning model f' and mislead its coupled interpreter g.

The attack consists of genetic algorithm operators: initialization, selection, crossover, mutation, and population update. Seeding the initial population with an optimal solution helps the technique converge fast. We evaluate each individual in the population by applying a loss function as the fitness function. Unlike the traditional genetic algorithms, the selection process of our method randomly picks two samples, one with a higher fitness score and the other with a lower fitness score, to keep the perturbation in the newly generated sample area that is considered important by the target model and its interpreter. We generate new offspring by transferring the genetic data of the winner and the loser with the predefined crossover rate. Mutation diversifies the population and introduces enough diversity to reach points outside the regions of the local optima.

Overall, the AdvEdge algorithm is effective in generating adversarial samples against a source deep learning model and its interpreter. The generated samples are then used as seeds for the initial population. The fitness scores of the population are evaluated by sending them to the target model in a black-box setting. If the attack requirements are met, the algorithm stops further steps. Otherwise, the algorithm repeats the steps until it succeeds or reaches the query threshold.

3 EXPERIMENTS AND EVALUATION

In this section, we provide detailed information on the settings and metrics used for our experiment to measure the performance of the proposed attack in terms of the deep learning and interpretation models given in the paper.

3.1 Experimental Settings

Datasets. We conducted our experiment on the ImageNet dataset [9], consisting of 1.2 million images for 1,000 categories. To evaluate our attack, we randomly select one image from each category in the ImageNet validation set [9], resulting in a total of 1,000 images. We ensure that the selected images are correctly classified by the target model f with a classification confidence score greater than 60%. For this experiment, we set the value of ϵ at 8, which is similar to the settings used in the AdvEdge attack [1], and represents the perturbation scale in the range of [0, 225]. Through these experiments, we validate that our proposed attack is effective across all categories of the selected deep learning models.

Classifiers. The experiment conducted in this study involves the use of two popular models, namely DenseNet-169 and ResNet-50, which were pre-trained on the ImageNet dataset. These models were used as both the source and target models for the transfer-based attack, which involved generating adversarial samples that could be transferred from the source model to the target model. The attack was limited to a maximum of 50,000 queries and the step size α and the number of iterations were set at 1/255 and 300, respectively. These values were selected based on the settings used in a previous attack called AdvEdge [1]. By using these pre-trained models and established attack settings, we aim to evaluate the effectiveness of our proposed attack in a controlled and reproducible manner.

Interpreters. CAM [23] interpreter is adopted as the representative of the interpretation models. CAM utilizes the feature maps of the convolutional layers in a deep learning model to generate interpretation maps: $m_c = \sum_i w_{i,c} a_i(j, k)$, where $a_i(j, k)$ is the activation of the ith channel at the spatial location (j, k) and $w_{i,c}$ is the weight of the i-th input and the c-th output in the linear layer of a deep learning model. We set λ in Equation ($\ell_{adv} = \min_{\hat{x}} \ell_{prd}(f'(\hat{x})) + \lambda \ell_{int}(g(\hat{x}; f'), m)$) at 0.204 for the CAM that is found effective in AdvEdge [1]. We use its open-source implementations for the experiment.

3.2 Attack Evaluation

We evaluate the proposed attack using various metrics to answer the following questions: ❶ *Is it effective against black-box deep learning models?* ❷ *Can it deceive interpretation models by generating interpretation maps similar to benign samples?* ❸ *Is it effective against defensive black-box deep learning models?* ❹ *Can it deceive interpretation models with defensive black-box deep learning models?*

Evaluation Metrics. Different evaluation metrics are used to assess the effectiveness of the proposed attack against both deep learning classifiers and interpreters.

For deep learning classifiers, the following metrics are used:

- **Attack success rate**: It calculates the ratio of successful attack cases to total attack cases.
- **Average queries**: The metric evaluates the efficiency of the attack algorithm in generating successful adversarial examples in a black-box setting.
- **Noise rate**: The metric is used to evaluate the quality of the adversarial examples generated by the attack algorithm.

For interpreters, the following metrics are used:

- **Qualitative comparison**: The metric evaluates the similarity between the interpretations of adversarial images and their benign counterparts.
- **IoU Test**: The metric measures the similarity of interpretation maps using the Intersection-over-Union score for different threshold values.
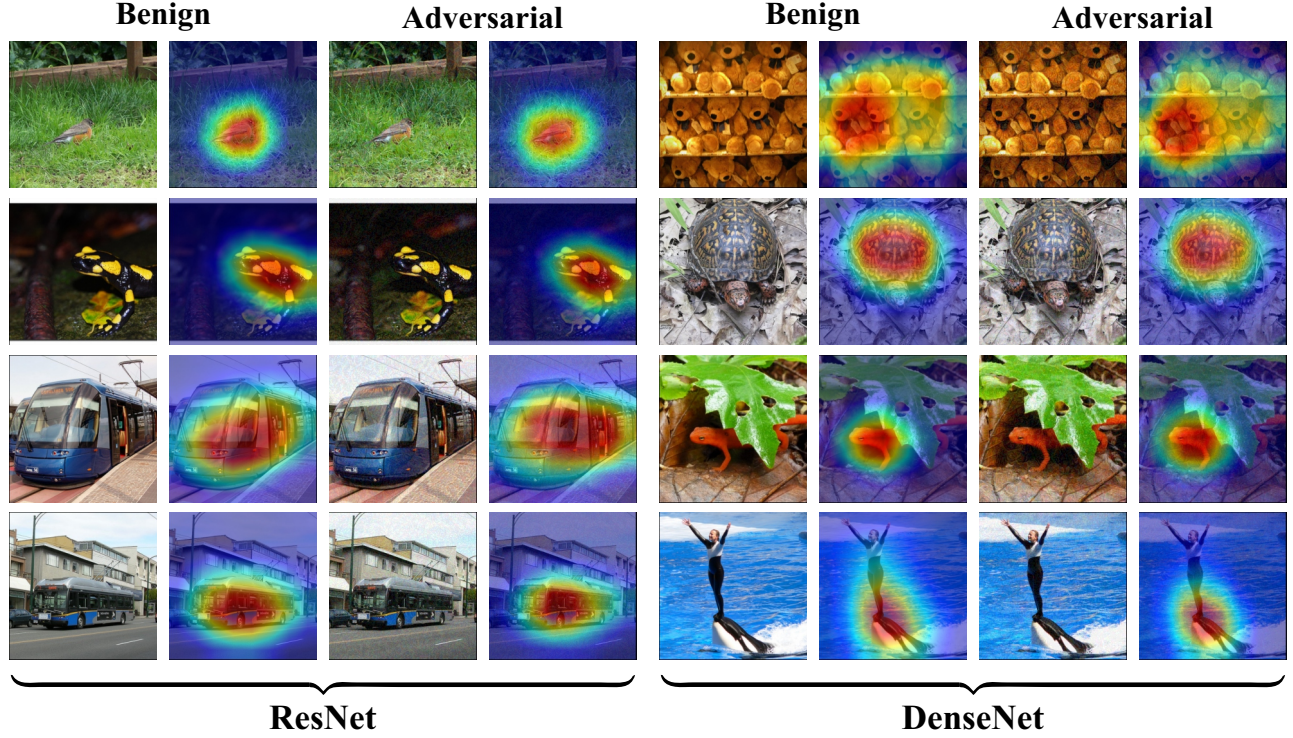
Figure 1: Attribution maps of benign and adversarial samples generated by our attack using CAM on ResNet and DenseNet.

Table 1: Success rate, average queries, and average noise of the proposed attack against different classifiers and interpreters using 1,000 images. The attack is based on black-box settings.

Interpreter	Source Model	Target Model	Success Rate	Average Queries	Average Noise Rate
CAM	**ResNet**	DenseNet	0.99	209.76	0.20 ± 0.06
	DenseNet	ResNet	1.00	188.53	0.20 ± 0.06

3.3 The Attack against Deep Learning Models

In this section, we present the evaluation of the effectiveness of our proposed adversarial attack on two popular model architectures, namely ResNet and DenseNet. To assess the efficacy of our attack, we implemented and tested it on two interpretable machine learning models using the Class Activation Mapping (CAM) technique with ResNet and DenseNet as source models.

The results of our experiments on the aforementioned scenarios are reported in Table 1. Our attack on DenseNet models achieved an impressive attack success rate of 0.99 and was found to be highly query efficient, requiring an average of only 209.76 queries. The average noise rate in the target model was stable at 0.20 ± 0.06. For the CAM interpreter with ResNet, our attack achieved a complete attack success rate on the target model with an average of 188.53 queries. The average noise rate remained stable. These results show improved performance with more complex source models.

Our results indicate that the proposed attack is highly effective against popular deep learning architectures and interpretable
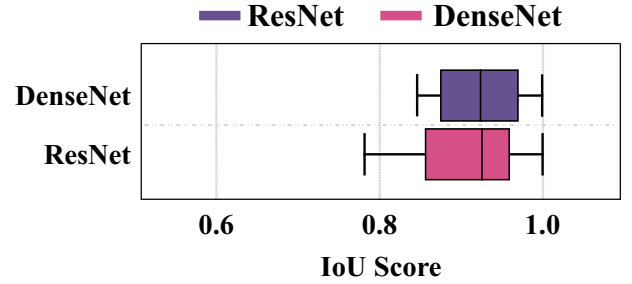

Figure 2: IoU scores of interpretation maps generated by our attack using CAM interpreter and ResNet, DenseNet as source models. y-axis represents the target models

machine learning models, highlighting potential security vulnerabilities and the need for robust defenses against adversarial attacks.

Observation 1: The Attack against Deep Learning

Our proposed attack has demonstrated a high degree of effectiveness in deceiving deep learning models, achieving a consistently high attack success rate across different source models. This highlights the significant potential for our attack to pose a threat to the security and reliability of deep learning models.

3.4 The Attack against Interpreters

In this section, we investigate the effectiveness of our proposed attack on the similarity between benign and adversarial interpretations, using a qualitative comparison and the Intersection over Union (IoU) test.

Qualitative comparison. Our qualitative comparison of the attribution maps generated by the Class Activation Mapping (CAM) interpreter for benign and adversarial samples showed that it was difficult to differentiate between the two. Our findings indicate that the adversarial attribution maps generated by our attack are highly reliable and comparable in quality to those produced from benign inputs. Figure 1 shows several examples of comparison between benign and adversarial samples generated by the attack.

IoU Test. To further assess the similarity between the two types of attribution maps, we used the IoU test, which measures the overlap between two maps. Our attack achieved highly balanced IoU scores on different deep learning models with the CAM interpreter, indicating that the adversarial attribution maps generated by our attack are similar to the benign attribution maps. Figure 2 summarizes our attack performance, which can be considered significant for a black-box attack.

Our findings highlight the effectiveness of our proposed attack in generating adversarial interpretation maps that are similar to their benign counterparts, raising concerns about the security and reliability of interpretable machine learning models.

Observation 2: The Attack against Interpreters

Our proposed attack has been shown to generate adversarial interpretation maps that are visually similar to their corresponding benign counterparts. This characteristic makes it challenging to distinguish between adversarial and benign maps, highlighting the potential for our attack to undermine the reliability and trustworthiness of interpretable machine learning models.

---

4 RELATED WORK

This section comprehensively reviews prior research on attacks targeting deep neural network models. The survey encompasses studies conducted on white-box and black-box attacks, utilizing diverse techniques such as transfer-based attacks, interpretation-based attacks, and gradient-based attacks.

Transfer-based attacks. In the realm of attacks against deep learning models, transfer-based attacks have received considerable attention in previous research. These attacks utilize adversarial samples generated by white-box attacks against one model to attack other black-box models. The potential effectiveness of transfer-based attacks has been demonstrated in various studies [6, 10, 13, 20]. For instance, researchers have proposed methods to enhance the transferability of adversarial samples by adding perturbations to the hidden layers of a model or convolving the gradient via a specific kernel. These studies highlight the importance of considering transfer-based attacks when assessing the robustness of models

and offer insight into techniques to improve the transferability of adversarial samples.

Interpretation-based attacks. This part discusses interpretation-based adversarial attacks that can deceive both the target deep learning models and their interpreters [22]. A recent study proposed white-box attacks called AdvEdge and AdvEdge$^+$ against deep learning models and their interpreters, highlighting the vulnerability of models that rely on interpretable features for decision making [1]. These attacks show the importance of considering the interpretability of deep learning models in addition to their accuracy and robustness. Furthermore, the proposed attacks provide a valuable tool for evaluating the interpretability of models and assessing their susceptibility to adversarial attacks.

Gradient-free attacks. Heuristic methods, including evolution strategies and genetic algorithms, have been utilized to create adversarial attacks that can generate visually imperceptible samples to deceive deep learning models [7]. GenAttack is a gradient-free optimization attack that can generate adversarial samples against black-box models with fewer queries. Another study proposed a query-efficient attack called MGAAttack [21], which uses transfer-based techniques to improve its efficacy. These attacks show the susceptibility of deep learning models to adversarial attacks and highlight the need to develop more robust defense mechanisms to enhance their security. By analyzing these attacks, researchers can identify weaknesses in models and devise better defenses against adversarial attacks.

5 CONCLUSION

In this study, we propose a black-box version of the AdvEdge attack that can effectively deceive deep learning models and their interpreters. Our attack combines transfer-based and score-based methods to generate adversarial examples that are difficult for the target models to classify correctly while also producing adversarial interpretation maps that are highly similar to the corresponding benign interpretations. Furthermore, our attack is both gradient-free and query-efficient, making it suitable for practical scenarios where access to model parameters or gradients may be limited. We evaluated the effectiveness of our proposed attack on various deep learning models, including ResNet and DenseNet, and their interpreters, such as the Class Activation Mapping (CAM) interpreter. Our experimental results show that our attack achieves a high success rate in deceiving target models and interpreters. Moreover, we performed a qualitative comparison and an Intersection over Union (IoU) test to evaluate the similarity between adversarial and benign interpretation maps. Our comparison of the attribution maps generated by the CAM interpreter for both benign and adversarial samples showed that it was difficult to distinguish between them. These results suggest that our attack can generate adversarial interpretation maps with a level of reliability that is comparable to that of benign inputs. In general, our proposed attack highlights the importance of developing more robust deep learning models and interpretability techniques to enhance their security against adversarial attacks. Furthermore, our work underscores the need to develop effective defense mechanisms that can detect and prevent such attacks in real-world scenarios.

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] Eldor Abdukhamidov, Mohammed Abuhamad, George K Thiruvathukal, Hyoungshick Kim, and Tamer Abuhmed. 2023. Single-Class Target-Specific Attack against Interpretable Deep Learning Systems. *arXiv preprint arXiv:2307.06484* (2023).

[3] Eldor Abdukhamidov, Mohammed Abuhamad, Simon S Woo, Eric Chan-Tin, and Tamer Abuhmed. 2022. Interpretations Cannot Be Trusted: Stealthy and Effective Adversarial Perturbations against Interpretable Deep Learning. *arXiv preprint arXiv:2211.15926* (2022).

[4] Eldor Abdukhamidov, Mohammed Abuhamad, Simon S Woo, Eric Chan-Tin, and Tamer Abuhmed. 2023. Microbial Genetic Algorithm-based Black-box Attack against Interpretable Deep Learning Systems. *arXiv preprint arXiv:2307.06496* (2023).

[5] 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 on Asia Conference on Computer and Communications Security* (Nagasaki, Japan) *(ASIA CCS '22)*. ACM, 1216–1218.

[6] 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 on Asia Conference on Computer and Communications Security*. 1216–1218.

[7] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. 2019. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 1111–1119.

[8] Piotr Dabkowski and Yarin Gal. 2017. Real time image saliency for black box classifiers. *Advances in neural information processing systems* 30 (2017).

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

[10] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the CVF Conference on Computer Vision and Pattern Recognition*. 4312–4321.

[11] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*. 3429–3437.

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

[13] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. 2019. Enhancing adversarial example transferability with an intermediate level attack. In *Proc. of the CVF international conference on computer vision*. 4733–4742.

[14] Firuz Juraev, Eldor Abdukhamidov, Mohammed Abuhamad, and Tamer Abuhmed. 2022. Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models. In *The 17th ACM ASIA Conference on Computer and Communications Security (ACM ASIACCS 2022)*.

[15] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078* (2015).

[16] 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).

[17] W James Murdoch, Peter J Liu, and Bin Yu. 2018. Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs. In *International Conference on Learning Representations, ICLR*.

[18] Sun Qirui, Mohammed Abuhamad, Eldor Abdukhamidov, Eric Chan-Tin, and Tamer Abuhmed. 2022. MLxPack: Investigating the Effects of Packers on ML-based Malware Detection Systems Using Static and Dynamic Traits. In *The 1st International Workshop on Cybersecurity and Social Sciences (CySSS)*.

[19] Qirui Sun, Eldor Abdukhamidov, Tamer Abuhmed, and Mohammed Abuhamad. 2022. Leveraging Spectral Representations of Control Flow Graphs for Efficient Analysis of Windows Malware. In *The 17th ACM ASIA Conference on Computer and Communications Security (ACM ASIACCS 2022)*.

[20] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*.

[21] Lina Wang, Kang Yang, Wenqi Wang, Run Wang, and Aoshuang Ye. 2020. MGAAttack: Toward More Query-efficient Black-box Attack by Microbial Genetic Algorithm. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2229–2236.

[22] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium*.

[23] 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.