Explanation-based Adversarial Detection with Noise Reduction

Juntao Su, Zhou Yang, Zexin Ren, Fang Jin
The George Washington University
sujuntao@gwu.edu, zhou_yang@gwu.edu, zxren10@gwu.edu, fangjin@gwu.edu

***Abstract*—Deep Neural Networks (DNNs) have achieved tremendous success in various tasks. However, DNNs exhibit uncertainty and unreliability when faced with well-designed adversarial examples, leading to misclassification. To address this, a variety of methods have been proposed to improve the robustness of DNNs by detecting adversarial attacks. In this paper, we combine model explanation techniques with adversarial models to enhance adversarial detection in real-world scenarios. Specifically, we develop a novel adversary-resistant detection framework called EXPLAINER, which utilizes explanation results extracted from explainable learning models. The explanation model in EXPLAINER generates an explanation map that identifies the relevance of input variables to the model's classification result. Consequently, adversarial examples can be effectively detected by comparing the explanation results of a given sample with its denoised version, without relying on any prior knowledge of attacks. The proposed framework is thoroughly evaluated against different adversarial attacks, and experimental results demonstrate that our approach achieves promising results in white-box attack scenarios.**

***Index Terms*—Adversarial detection,Model explanation, Noise reduction.**

I. Introduction

Deep Neural Networks (DNNs) have been widely used in various applications and achieved tremendous success in recent years. For instance, DNNs have played a critical role in a variety of generative and discriminative learning tasks, including image processing [Du et al.(2019)], motion capture [Yue et al.(2021)], [Xu et al.(2022)], clinical research [Su et al.(2022)], [Guo et al.(2023)], and image generation [Du et al.(2022)]. However, studies have shown that outputs of DNNs can be easily altered by a small perturbation of the input, or even a small perturbation of one pixel [Zhou et al.(2022)]. This sensitivity to small changes in the input makes DNNs vulnerable, limiting the applications of DNNs in high-stake settings, such as self-driving cars [Deng et al.(2020)] and malware detections [Sewak et al.(2018)].

Several approaches for defending against adversarial examples have been proposed. The use of adversarial training or gradient masking to improve the robustness of neural networks is one area of research. Existing research has shown, however, that neural network architectures modified with adversarial training and gradient masking can still be attacked [Carlini and Wagner(2017)]. Another area of study is adversarial detection, which aims to determine if a given input is adversarial or normal.

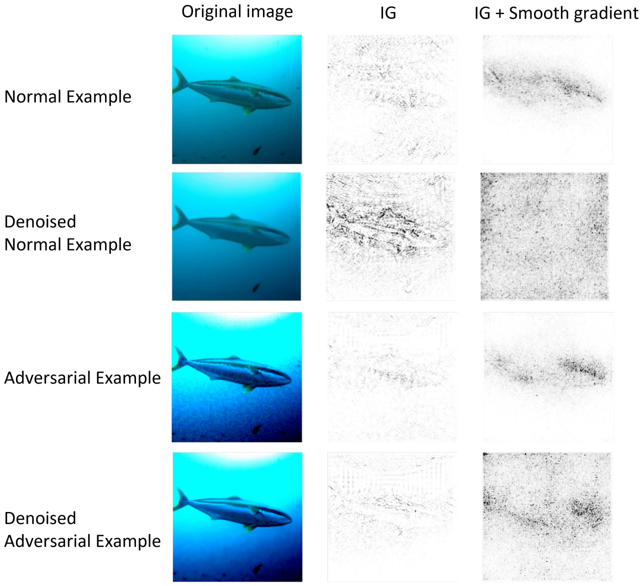

Fig. 1. Feature maps for normal, denoised normal, adversarial, and denoised adversarial examples. After the noise reduction process, the feature map of the normal example has almost no change, while the feature map of the adversarial image has obvious changes on both the background and object.

However, there are critical questions remain unanswered about what causes the misclassification of adversarial examples. To uncover the causes of adversarial attacks, efforts have been tried to explore the feature differences between normal inputs and adversarial inputs. One possible method is using explanation techniques. Given an image, the result from an explanation model encodes the relevance of pixels for the prediction result, which is commonly referred to as an explanation map. Fig 1 shows that there are human-understandable differences between adversarial examples and normal inputs with Integrated Gradient [Sundararajan et al.(2017)]. As we can see, normal examples tend to have a more meaningful and continuous explanation map, while adversarial examples tend to have a more discrete explanation map. This difference is more distinguishable while using a patch attack [Brown et al.(2017)]. The model will be fooled and classify the image only based on the adversarial patch part. As we could see from the Fig 2, the explanation of the adversarial example only shows the shape of the adversarial patch.

As a result, the inconsistencies of extracted features between

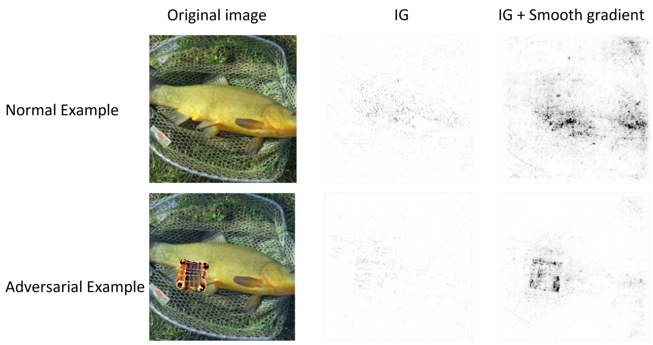

Fig. 2. Examples of the feature maps extracted from a normal example and an adversarial example with patch attack.

adversarial examples and normal examples can be utilized in detecting adversarial examples. Song et al. [Song et al.(2018)] proposed an Ensemble approach for Explanation-based Adversarial Detection, which uses an ensemble of explanation models wherein each explanation technique provides an explanation map for every classification decision made by a target model. However, their framework requires additional training after extracting the explanation maps.

In this work, we propose an unsupervised adversarial detection method (EXPLAINER) with model explanation. We extract feature maps from explanation models and use the extracted features to determine if an example is normal or adversarial. We evaluate EXPLAINER using five state-of-the-art adversarial attacks on MNIST [LeCun et al.(1998)] dataset and ImageNet [Deng et al.(2009)] dataset, under white-box threat model. Our experimental results show that we can effectively detect all attacks with fast responses.

We summarize our main contributions as follows.

- We develop a novel framework called EXPLAINER based on model explanation techniques and noise reduction. EXPLAINER utilizes features from the explanation results using normal examples and adversarial examples without additional training tasks.
- We evaluate EXPLAINER on five state-of-the-art adversarial attacks and two image datasets under white-box attack. The results show that the proposed system can consistently achieve high detection rates with a low false-positive rate.
- We extensively evaluate EXPLAINER with different clustering techniques. Our findings show that EXPLAINER achieves promising results and high efficiency in different scenarios.

II. Related Work

*a) **Adversarial Attack**:* Adversarial examples can be developed using gradient-based attacks [Carlini and Wagner(2017)], [Goodfellow et al.(2014)], [Szegedy et al.(2013)] or content-based attacks [Brown et al.(2017)], [Eykholt et al.(2018)]. This paper focuses on five state-of-the-art gradient-based attacks: Basic Iterative Method (BIM) [Kurakin et al.(2018)], Momentum Iterative Method (MIM) [Dong et al.(2018)], and Carlini and Wagner Attacks (CW) [Carlini and Wagner(2017)] tailored to L_0, L_2, and L_∞ norms.

*b) **Model Explanation**:* Model explanation techniques provide insights into the features critical for DNN decision-making [Lipton(2018)], including enhancing the transparency of large language models (LLMs), which have been widely applied in recent years [Guo et al.(2024)]. Local explainability methods identify which regions in an input image influence the prediction [Simonyan et al.(2013)], [Dombrowski et al.(2019)]. Saliency maps and explanation maps are commonly used for this purpose.

*c) **Adversarial Detection**:* Adversarial detection aims to classify inputs as normal or adversarial. Magnet [Meng and Chen(2017)] uses autoencoders to approximate normal example manifolds, while Feature Squeezing [Xu et al.(2017)] reduces an adversary's freedom by smoothing images or reducing color depth. Adaptive Noise Reduction [Liang et al.(2018)] combines scalar quantization and spatial smoothing for high-accuracy adversarial detection.

III. Proposed Method

EXPLAINER detects adversarial examples using model explanation features. The hypothesis is that normal and adversarial examples have inconsistent explanation robustness. The steps are as follows: generate explanation maps for the original and denoised images, then compute and compare their Shannon entropy to classify the input. The key idea is that normal images should exhibit minimal change in their explanation maps after denoising, whereas adversarial examples will show significant differences. By quantifying this difference through Shannon entropy, we can effectively distinguish between normal and adversarial inputs, leveraging the inherent stability of normal images' feature maps against noise reduction.

A. Generation of Adversarial Attacks

The goal of an adversary is to craft a sample that appears identical to a normal sample but is misclassified by the target model. For a given input image x, the objective is to find a minimal perturbation η such that the adversarial input $\tilde{x} = x + \eta$ is misclassified. We consider the following adversarial attacks to test our framework:

Basic Iterative Method (BIM) [Kurakin et al.(2018)]: BIM is an iterative version of FGSM [Goodfellow et al.(2014)]. Instead of applying adversarial noise η once, it is applied iteratively with small ϵ. The recursive formula is:

$$\begin{aligned} & x_0^* = x \\ & x_i^* = \operatorname{clip}_{x,\epsilon}\left(x_{i-1}^* + \epsilon \operatorname{sign}\left(\nabla_{x_{i-1}^*} J\left(\Theta, x_{i-1}^*, y\right)\right)\right) \end{aligned} \tag{1}$$

Momentum Iterative Method (MIM) [Dong et al.(2018)]: MIM accelerates gradient descent algorithms by accumulating a velocity vector in the gradient direction. The optimization problem is:

$$\underset{x^*}{\operatorname{argmin}}\, J\left(\boldsymbol{x}^*, y\right), \quad \text{s.t.}\ \left\|\boldsymbol{x}^* - \boldsymbol{x}\right\|_\infty \leq \epsilon, \tag{2}$$

Algorithm 1 EXPLAINER Framework for Adversarial Detection

```
Input: Image $x$, classifier $f$
Output: Adversarial detection result
Step 1: Generate Adversarial Examples
$x^* \leftarrow$ BIM/MIM/CW$(x, f, \epsilon)$
Step 2: Generate Explanation Maps
$h(x) \leftarrow$ Explain$(f, x)$
Step 3: Apply Image Denoising
$\tilde{x} \leftarrow$ Denoise$(x)$
Step 4: Generate Explanation Maps for Denoised Image
$h(\tilde{x}) \leftarrow$ Explain$(f, \tilde{x})$
Step 5: Calculate Shannon Entropy
$H(x) \leftarrow -\sum p(x) \log p(x)$
$H(\tilde{x}) \leftarrow -\sum p(\tilde{x}) \log p(\tilde{x})$
Step 6: Adversarial Detection
if $H(\tilde{x}) < H(x)$ then
    Classify $x$ as normal
else
    Classify $x$ as adversarial
end if
```

where ϵ is the size of the adversarial perturbation.

Carlini and Wagner Attacks (CW) [Carlini and Wagner(2017)]: The CW L_2 attack finds a perturbation δ^* for the following optimization problem:

$$\begin{aligned} \min \|\delta\|_2^2 + c \cdot f(x+\delta) \\ \text{s.t. } x + \delta \in [0,1]^n \end{aligned} \tag{3}$$

B. Generation of Explanation

Given a neural network classifier $f(\cdot)$ and an input x, the explanation of the classification is represented as an explanation map $h : \mathbb{R}^d \to \mathbb{R}^d$. We consider the following explanation techniques, with generated explanation maps shown in Fig 3, mainly generated through Captum [Kokhlikyan et al.(2020)].

DeepLift [Shrikumar et al.(2017)]: DeepLift attributes to each input x_i a value $C_{\Delta x_i \Delta y}$ representing the effect of that input being set to a reference value. The "summation-to-delta" property is:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta o} = \Delta o, \tag{4}$$

where $o = f(x)$ is the model output, $\Delta o = f(x) - f(r)$, and r is the reference input.

SHAP [Lundberg and Lee(2017)]: SHAP explains the prediction of an instance x by computing the contribution of each feature. The explanation model is:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5}$$

GradSHAP is a variant of SHAP that combines integrated gradients with SHAP values to estimate feature attributions efficiently.

Grad-CAM and Guided CAM [Selvaraju et al.(2016)]: Grad-CAM computes a coarse-grained feature-importance map by associating feature maps in the final convolutional layer with classes based on gradients. Guided Grad-CAM refines this by performing an elementwise product between Grad-CAM scores and Guided Backpropagation scores.

IG [Sundararajan et al.(2017)]: IG computes gradients at all points along a linear path from a baseline $\bar{x}$ to x, and averages them:

$$h(x) = (x - \bar{x}) \odot \int_{\alpha=0}^{1} \frac{\partial f(\bar{x} + \alpha(x - \bar{x}))}{\partial x} \, \mathrm{d}\alpha \tag{6}$$

SmoothGrad [Smilkov et al.(2017)]: SmoothGrad sharpens gradient-based sensitivity maps by creating noisy copies of an input image and averaging gradients with respect to these copies, removing irrelevant noisy regions.

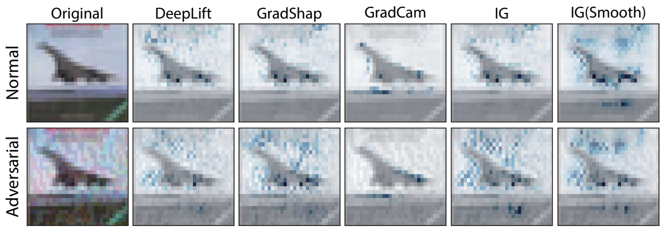

Fig. 3. Generation of Attacks and Explanation.

C. Image Noise Reduction and Shannon Entropy

Fig 1 and Fig 3 show that the explanation maps for normal, adversarial, and denoised examples differ significantly. This motivates us to detect adversarial samples by calculating the change in Shannon entropy between the original and denoised versions. Our two-step approach decouples image denoising and Shannon entropy calculation. First, we use non-local means denoising [Buades et al.(2011)] to obtain denoised images. Second, we calculate Shannon entropy for both the original and denoised images.

Non-local means denoising replaces the color of a pixel with an average of similar pixels' colors:

$$NLu(p) = \frac{1}{C(p)} \int f(d(B(p), B(q)))u(q) \, dq,$$

where $d(B(p), B(q))$ is the Euclidean distance between image patches centered at p and q, f is a decreasing function, and $C(p)$ is the normalizing factor.

Shannon entropy, defined by Shannon's H-theorem, measures the uncertainty in a discrete random variable X taking values in $\mathcal{X}$, distributed according to $p : \mathcal{X} \to [0,1]$ such that $p(x) = \mathbb{P}[X = x]$:

$$\mathrm{H}(X) = \mathbb{E}[\mathrm{I}(X)] = \mathbb{E}[-\log p(X)].$$

IV. Experiment

A. Dataset

We evaluated the performance of our detection framework on MNIST [LeCun et al.(1998)] and ImageNet [Deng et al.(2009)]. For MNIST, we trained a CNN-based target model with 60,000 training examples and 10,000 validation examples. For ImageNet, we used a pre-trained ResNet model [He et al.(2016)] with 50,000 training images and 10,000 validation images, testing our framework on 50,000 images.

B. Implementation Details

We generated adversarial examples using five state-of-the-art attacks: BIM [Kurakin et al.(2018)], MIM [Dong et al.(2018)], and CW [Carlini and Wagner(2017)], tailored to L_0, L_2, and L_∞ norms. Explanation maps were generated using Deeplift [Shrikumar et al.(2017)], SHAP [Lundberg and Lee(2017)], Grad-CAM [Selvaraju et al.(2016)], IG [Sundararajan et al.(2017)], and IG with SmoothGrad [Smilkov et al.(2017)]. The key step is comparing the Shannon entropy of the original and denoised images' explanation maps. If the entropy decreases after denoising, the image is classified as normal; otherwise, it is classified as adversarial.

C. Model Evaluation

We thoroughly evaluated the effectiveness of EXPLAINER in different scenarios and compared its performance with other methods.

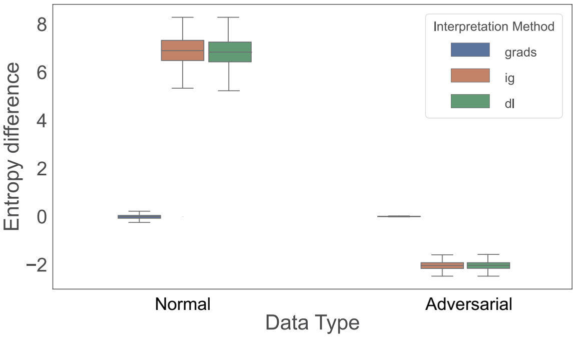

Fig. 4. Shannon entropy comparison between the normal group and adversarial group with different explanation methods.

1) Explanation Methods: We evaluated the Shannon entropy values from different explanation techniques, as shown in Fig 4. We compared three techniques: gradient, integrated gradients, and Deeplift. Integrated gradients and Deeplift demonstrated significant entropy differences between normal and adversarial groups, indicating their effectiveness in capturing noise information for adversarial detection.

2) Different Adversarial Attacks: We evaluated the distribution of extracted features from different attack techniques using integrated gradients. Fig 5 shows the entropy values, revealing that all attacks exhibited significant differences compared to the normal group.

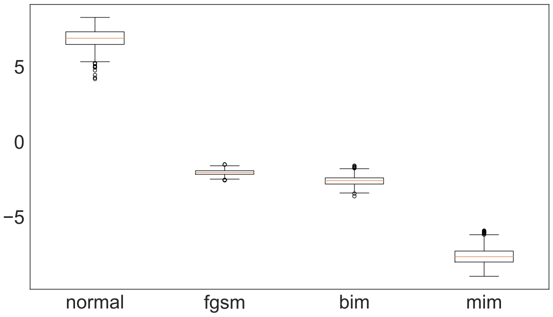

Fig. 5. Shannon entropy comparison between the normal group and adversarial group with different attacks using the integrated gradient method.

3) Comparison of Detection Rate: Table I compares the detection rate of EXPLAINER with Magnet and Feature Squeezing(FS) on MNIST using five attack methods. EXPLAINER achieved around 99% detection rates on MNIST and similar rates on ImageNet. Compared to other methods, our framework is efficient and accurate without requiring additional training or model retraining. The time efficiency of EXPLAINER was also evaluated, showing that it can classify images in approximately 1 second on MNIST and 50 seconds on ImageNet using the ResNet-18 model.

TABLE I
Detection Accuracy Comparison

Attack	**EXPLAINER**	**MagNet**	**Feature Squeezing (FS)**
CW_0	99.8%	86.2%	91.0%
CW_2	99.9%	86.0%	99.4%
CW_∞	99.5%	96.5%	99.1%
BIM	99.0%	99.8%	98.2%
MIM	98.8%	99.6%	98.5%

V. Conclusion

In this paper, we propose EXPLAINER, a framework that combines explanation techniques with noise reduction to detect adversarial examples. The motivation lies in the distinguishability between normal and abnormal explanations and their corresponding maps for any target class. Experiments demonstrate that our approach is effective against white-box attacks across different datasets and can help identify attack types, such as patch attacks, with clear explainability. We acknowledge the potential emergence of more sophisticated attacks and hope our work inspires further research. Our proposed defense can complement state-of-the-art detection methods, enhancing adversarial detection. Moreover, EXPLAINER's adaptability to various explanation and denoising models indicates its potential for broader applications, including real-time detection systems. Future work will focus on improving robustness and scalability against evolving adversarial strategies, as well as exploring integration with automated machine learning pipelines to streamline detection processes.

Acknowledgment

This work was supported by the U.S. National Science Foundation under Grant NSF IIS-2238700. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[Brown et al.(2017)] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[Buades et al.(2011)] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. 2011. Non-local means denoising. *Image Processing On Line* 1 (2011), 208–212.

[Carlini and Wagner(2017)] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE, 39–57.

[Deng et al.(2009)] 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*. Ieee, 248–255.

[Deng et al.(2020)] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. 2020. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 1–10.

[Dombrowski et al.(2019)] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems* 32 (2019).

[Dong et al.(2018)] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9185–9193.

[Du et al.(2019)] Chen Du, He Zewei, Sun Anshun, Yang Jiangxin, Cao Yanlong, Cao Yanpeng, Tang Siliang, and Michael Ying Yang. 2019. Orientation-aware deep neural network for real image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[Du et al.(2022)] Hongfei Du, Emre Barut, and Juntao Su. 2022. Build Connections Between Two Groups of Images Using Deep Learning Method. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 1231–1236.

[Eykholt et al.(2018)] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1625–1634.

[Goodfellow et al.(2014)] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[Guo et al.(2024)] Muzhe Guo, Muhao Guo, Juntao Su, Junyu Chen, Jiaqian Yu, Jiaqi Wang, Hongfei Du, Parmanand Sahu, Ashwin Assysh Sharma, and Fang Jin. 2024. Bayesian Iterative Prediction and Lexical-based Interpretation for Disturbed Chinese Sentence Pair Matching. In *Proceedings of the ACM on Web Conference 2024*. 4618–4629.

[Guo et al.(2023)] Muzhe Guo, Yong Ma, Efe Eworuke, Melissa Khashei, Jaejoon Song, Yueqin Zhao, and Fang Jin. 2023. Identifying COVID-19 cases and extracting patient reported symptoms from Reddit using natural language processing. *Scientific Reports* 13, 1 (2023), 13721.

[He et al.(2016)] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[Kokhlikyan et al.(2020)] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. arXiv:2009.07896 [cs.LG]

[Kurakin et al.(2018)] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. 2018. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*. Springer, 195–231.

[LeCun et al.(1998)] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[Liang et al.(2018)] Bin Liang, Hongcheng Li, Miaoqiang Su, Xirong Li, Wenchang Shi, and Xiaofeng Wang. 2018. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing* 18, 1 (2018), 72–85.

[Lipton(2018)] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[Lundberg and Lee(2017)] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[Meng and Chen(2017)] Dongyu Meng and Hao Chen. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 135–147.

[Selvaraju et al.(2016)] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450* (2016).

[Sewak et al.(2018)] Mohit Sewak, Sanjay K Sahay, and Hemant Rathore. 2018. Comparison of deep learning and the classical machine learning algorithm for the malware detection. In *2018 19th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)*. IEEE, 293–296.

[Shrikumar et al.(2017)] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*. PMLR, 3145–3153.

[Simonyan et al.(2013)] 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).

[Smilkov et al.(2017)] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[Song et al.(2018)] Fei Song, Yanlei Diao, Jesse Read, Arnaud Stiegler, and Albert Bifet. 2018. EXAD: A system for explainable anomaly detection on big data traces. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1435–1440.

[Su et al.(2022)] Juntao Su, Edward T Dougherty, Shuang Jiang, and Fang Jin. 2022. An interactive knowledge graph based platform for covid-19 clinical research. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1609–1612.

[Sundararajan et al.(2017)] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[Szegedy et al.(2013)] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[Xu et al.(2022)] Pan Xu, Yuanlei Yue, Juntao Su, Xiaoqian Sun, Hongfei Du, Zhichao Liu, Rahul Simha, Jianhui Zhou, Chen Zeng, and Hui Lu. 2022. Pattern decorrelation in the mouse medial prefrontal cortex enables social preference and requires MeCP2. *Nature communications* 13, 1 (2022), 3899.

[Xu et al.(2017)] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).

[Yue et al.(2021)] Yuanlei Yue, Pan Xu, Zhichao Liu, Xiaoqian Sun, Juntao Su, Hongfei Du, Lingling Chen, Ryan T Ash, Stelios Smirnakis, Rahul Simha, et al. 2021. Motor training improves coordination and anxiety in symptomatic Mecp2-null mice despite impaired functional connectivity within the motor circuit. *Science Advances* 7, 43 (2021), eabf7467.

[Zhou et al.(2022)] Tianxun Zhou, Shubhankar Agrawal, and Prateek Manocha. 2022. Optimizing One-pixel Black-box Adversarial Attacks. *arXiv preprint arXiv:2205.02116* (2022).