FGSM Attack on CNN-based image classifiers: Vulnerability analysis and an effective defense strategy

Doan Huong Giang[1], Pham Thi Thanh Thuy[2*]

[1]Faculty of Control and Automation, Electric Power University, 235 Hoang Quoc Viet, Bac Tu Liem, Hanoi, Vietnam;
[2]Faculty of Cybersecurity and High Tech Crime Prevention - Academy of People Security, 125 Tran Phu, Ha Dong, Hanoi, Vietnam.
[*]Corresponding author: thanh-thuy.pham@mica.edu.vn
Received 23 Mar. 2025; Revised 12 May. 2025; Accepted 10 Jun. 2025; Published 25 Jun. 2025.
DOI: https://doi.org/10.54939/1859-1043.j.mst.104.2025.155-163

ABSTRACT

Convolutional Neural Networks (CNNs) have demonstrated significant advantages and have, therefore, been widely applied across various domains. However, adversarial attacks have exposed critical vulnerabilities in these models, posing threats to the security and reliability of deep learning systems. Although numerous studies have investigated adversarial attacks on deep learning models, the specific impact of such attacks on CNN-based image classifiers remains an open issue, especially considering that many widely-used CNN models form the foundation of essential real-world applications. This study analyzes the vulnerabilities of CNN image classifiers under the Fast Gradient Sign Method (FGSM) adversarial attack and proposes an effective defense strategy named WR_FGSM. Experimental results on standard benchmark datasets show that several CNN models suffer significantly from FGSM attacks. The adversarial images generated by this attack not only deceive CNN-based image classifiers but also appear visually indistinguishable from the original images to the human vision. Our proposed WR_FGSM defense incorporates adversarial training—one of the most effective existing defense strategies—along with a regularization technique during the training process. This approach effectively safeguards CNN models against FGSM attacks while maintaining a balance between adversarial robustness and the generalization capability of the models.

Keywords: Convolutional Neural Network (CNN); Adversarial attack; Defense; Adversarial training; Regularization technique.

1. INTRODUCTION

Convolutional Neural Networks (CNNs) have achieved remarkable success in image classification tasks. However, they remain highly vulnerable to adversarial attacks, where small, often imperceptible perturbations to input images can lead to incorrect predictions. Such alterations can substantially affect the recognition results, even though the images may appear unchanged to humans [1, 2]. This vulnerability poses serious security risks for real-world AI applications that require high stability and accuracy, such as autonomous vehicle control, medical image processing, and security surveillance systems.

Research on adversarial attack methods against AI models, especially image classification models, is gaining widespread interest from both the academic community and industry. Current adversarial attacks are primarily classified based on basic criteria such as (1) the attacker's knowledge of the model being attacked, (2) input updates performed based on gradients in a single step, (3) the optimization techniques used. According to the first classification criterion, there are two types of adversarial attacks: white-box and black-box attacks. In white-box attack scenarios, the attacker has full knowledge of the model, including its architecture, parameters, loss function, and gradients, making it easier to generate adversarial samples capable of deceiving the classification model [3-5]. Conversely, in black-box attacks, the attacker does not have access to the model's internal

information, such as its structure, parameters, or gradients. The attack strategy mainly relies on exploiting changes in input data and observing the model's responses [6]. Compared to white-box attacks, black-box attacks are generally more complex and challenging due to the lack of internal model information. Common white-box adversarial methods include FGSM [7], PGD (Projected Gradient Descent) [8], BIM [9], and C&W (Carlini & Wagner) [10]. Popular black-box adversarial methods include ZOO (Zeroth Order Optimization) [11], NES (Natural Evolution Strategies) [12], and Boundary attacks [13]. The gradient-based methods update the input in a single step, where the gradient indicates the direction in which the input should be changed to increase the loss value, making the model more likely to be deceived (misclassification). Instead of multiple updates, these methods compute the loss gradient once and immediately adjust the input in that direction. A typical example of this approach is the FGSM attack. Optimization-based methods iteratively modify the input to find minimal perturbations that cause the model to misclassify. Common attacks of this type include BIM, PGD, C&W, and DeepFool [14]. Compared to gradient-based adversarial attacks, optimization-based attacks are more complex and computationally intensive but often more effective, especially in black-box settings.

To minimize the impact of adversarial attacks on CNN models and enhance their defensive robustness, various defense strategies have been proposed. Some main approaches include (a) adversarial training-based defense [15], (b) input preprocessing-based defense [16], and (c) model modification-based defense [17]. Each approach has its own advantages and disadvantages. Approach (a) is currently one of the most effective and popular methods. This method helps the model learn to resist adversarial samples by incorporating these samples into the training process, significantly improving resilience against common attacks (such as FGSM, PGD, etc.). The downside of this approach is high computational cost, potential reduction in performance on clean (non-adversarial) datasets, and limited defensive capability against unseen attack types not encountered during training. The input preprocessing-based defense (b) is simple and easy to implement, reduces both noise and small adversarial perturbations in the original data, and is highly compatible with many existing CNN systems. However, its weakness lies in being insufficiently robust against sophisticated and complex attacks, potentially degrading useful information in the original data and affecting model accuracy. Some preprocessing techniques can be bypassed if the attacker is aware of the input processing steps. Model modification-based defense (c) can make the model less sensitive to small perturbations and can effectively combine with other methods. However, it is generally more complex to design and implement, requiring adjustments to the model structure. Additionally, the model can still be vulnerable to advanced attack techniques, and if not optimally designed, it may lead to decreased performance on clean datasets.

This study focuses on evaluating the vulnerabilities of popular CNN image classifiers, which are the foundation of many important applications, such as MobileNetV2 [18], ResNet50 [19], and DenseNet121 [20]. In this research, we performed FGSM adversarial attacks on these CNN models. FGSM is one of the simple and effective adversarial attack methods widely applied to machine learning models, especially deep learning models like CNNs. Many studies use FGSM as a benchmark to assess the robustness of machine learning models against attacks. Based on the analysis of the impact of FGSM attacks on CNN image classification models, we have developed and proposed a new defense method called WR_FGSM. The experimental studies were conducted on publicly available datasets for the research and learning community, including MNIST [21], CIFAR-10 [22], and MICAHandPose [23]. The results provide valuable insights for both research and practical applications.

The next section of the article includes the following parts: section 2 describes the proposed solution in detail. The testing results are presented in section 3. Section 4 is the final part, which will present the conclusion and our future works.

2. PROPOSED METHOD

2.1. FGSM attack on CNN image classification models

Given the original input image represented as a one-dimensional vector $x \in R^{1\times N}$, with a label layer $\hat{y}_{gt}[1 \times C] \in R^C$, where C is the number of classes in the dataset. The input image size is m × n = N, meaning the total number of pixels is N. The deep convolutional neural network model is considered as a function f(.). When an input image passes through the model, the output at the FN (Fully Connected) layer is a probability vector $\hat{P} = [p_1, p_2, \dots, p_C]$ as illustrated in the following Eq. (1):

$$\hat{P} = [p_1, p_2, \dots, p_C] = f(x) \quad (1)$$

At the final layer of the deep convolutional neural network, the probability vector $\hat{P}$ is used as the input to the softmax layer, and the output is the vector $\hat{y}[1 \times C] \in R^C$ which is a one-hot vector corresponding to the class index into which x is classified (i.e., $\hat{y}$=$\hat{y}_{gt}$). Adversarial attacks involve making small changes to the input image x. This is achieved by adding a small noise component to the input image to create a new image, as described in the following Eq. (2):

$$\tilde{x} = x + \tau \quad (2)$$

In which, τ is a small noise, and the objective of the AI model attack is to identify this noise component to maximize the probability of misclassification, causing the model to make incorrect predictions. That is, the resulting output probability of the model is as shown in equation (3) below:

$$\widehat{P'} = [p'_1, p'_2, \dots, p'_C] = f(x + \tau) \neq \hat{P} \quad (3)$$

Adding noise τ to the input to create a new input image that is indistinguishable to the naked eye from the original data, while the model still mispredicts with $\widehat{y'} \in R^C \neq \hat{y}$. There are two types of objectives that can be pursued when attacking a CNN model by finding noise τ to add to the original image:

- Unfocused attack: $\widehat{y'} \neq \hat{y}$ while $\|\tau\|$ is very small. This means that the goal of the problem is to achieve $\widehat{y'} \neq \hat{y}$, and the model only needs to produce an output from the classifier $f(x + \tau)$ that is different from the correct label y.

- Targeted attack: $\widehat{y'} = \widehat{y_t}$ while $\|\tau\|$ is very small. This means $\widehat{y'} = \widehat{y_t}$, and the classifier $f(x + \tau)$ produces the result $\widehat{y_t}$ and where $\widehat{y_t}$ is a specific class that the attacker desires which must be different from $\hat{y}$.

Therefore, with the loss function optimized for untargeted attacks that aim to maximize the value as shown in Eq. (4):

$$max_\tau L(f_\theta(x + \tau), y), \|\tau\| \leq \epsilon \quad (4)$$

The optimization of the loss function in targeted attacks is represented by Eq. (5):

$$max_\tau L(f_\theta(x + \tau), y_t), \|\tau\| \leq \epsilon \quad (5)$$

To achieve the two optimization objectives in Eq. (4) and Eq. (5), where x(i) is changed in the direction of the steepest gradient to reduce the probability of classifying x become C with negligible changes, various methods can be used. In this study, we focus on exploring the FGSM method, which is a relatively simple and easy-to-implement noise generation technique. In this method, the noise component is calculated according to the following Eq. (6):

$$\tau = \epsilon * sign\big(\nabla_x L(f_\theta(x), y)\big) = \epsilon * sign\left(\left(\frac{dL}{x_1}, \dots, \frac{dL}{x_N}\right), y\right) \quad (6)$$

Equation (6) shows that the noise generated by the FGSM method does not exceed the value ϵ. The new image with added noise generated by FGSM is expressed as in equation (7):

$$\tilde{x} = x + \tau = x + \epsilon * sign\big(\nabla_x L(f_\theta(x), y)\big) \quad (7)$$

Where θ represents the parameters of the deep convolutional neural network model, ϵ is the noise magnitude, sign(·) is the sign function, x is the original image value, K is the number of steps, $\tilde{x}$ is the adversarial input image, y is the true output label of the model, and L(.) is the model's loss function.

2.2. WR_FGSM defense solution for CNN image classification models against FGSM adversarial attack

To enhance the robustness of CNN image classification models against PGD adversarial attacks, we propose a method that combines model weight adjustment with training into a single loss function for the CNN, as illustrated in the following Eq. (8):

$$L_{WR\_FGSM} = L_{Ori} + \alpha_1 * L_{Adv\_FGSM} + \alpha_2 * \|W\|_2^2 \quad (8)$$

Where the component L_{Ori} is the loss function when training with original data in the dataset; L_{Adv_FGSM} is the loss when training with adversarial data generated by the FGSM method in Eq. (7); this means that the training data fed into the defense model always includes two components: the original data and the noise-added adversarial data produced by FGSM (L_{Adv_FGSM}). W represents the model weights ($\|W\|_2^2$). Where α_1 and α_2 are hyperparameters that adjust the influence of each component's adversarial images (L_{Adv_FGSM}) and model weights ($\|W\|_2^2$) on the overall loss function of the WR_FGSM model. It is ensured that the hyperparameters satisfy the following Eq. (9):

$$\alpha_1 + \alpha_2 = 1 \quad (9)$$

The objective function during training aims not only to optimize performance on the original dataset but also to include adversarial examples - data samples with added noise from FGSM designed to deceive the model. As a result, the model learns to classify effectively on both types of data, thereby enhancing the detection capability and resilience against adversarial samples. Additionally, weight regularization techniques are applied to control and maintain small weights, preventing overfitting to adversarial noise and improving generalization performance on unseen data. Finally, the impact of the hyperparameters α_1 and α_2 on the defense effectiveness of the model will be presented and analyzed in detail in section 3.

2.3. Datasets and Evaluation Protocol

2.3.1 Datasets

In this study, we conduct evaluations on three image datasets, including:

- The MNIST dataset [21], which contains 70 000 grayscale images of 10 classes of handwritten digits. The images are 28x28 pixels.

- The CIFAR-10 dataset [22] consists of 60 000 color images across 10 classes: aeroplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The images are RGB with a size of 32x32 pixels.

- The MICAHandPose dataset [23] comprises approximately 14 000 images across 7 classes of static hand gestures. Each gesture was captured by 10 individuals, with each gesture consisting of 200 RGB images of the hand region, with variable sizes ranging from approximately 40x40 to 120x120 pixels.

2.3.2 Evaluation protocol

In this paper, the authors use the "Leave-one-subject-out cross-validation" method to split the training and testing data, as described in the data partitioning in [23]. In this approach, the total data in each dataset is evenly divided into 10 parts within each class. Each part corresponds to one "subject". The testing process is conducted over 10 iterations. In each iteration, one "subject" is

used for testing, one "subject" for model validation, and the remaining eight "subjects" are used for training the model. The process is repeated 10 times, and the final evaluation results are calculated as the average over the 10 trials.

3. EXPRIMETIAL RESULT

The software is implemented on a computer with an Intel Core i5-11400H CPU, NVIDIA GeForce GTX 1650 GPU, and 8GB of RAM. The programming language used is Python. The installation testing was conducted with a batch size of 32; the model training with a learning rate ranging from 10^{-6} to 10^{-4}; the number of epochs used with early stopping was initially set to 100, but the model often stopped around 50-70 epochs. The experiments utilized convolutional neural network models, including MobileNetV2 [18], ResNet50 [19], and DenseNet121 [20]. The evaluations conducted include: (1) assessing the impact of FGSM attacks on CNN models; (2) evaluating the WR_FGSM defense method against CNN models.

3.1. Assessing the impact of FGSM attack on CNN models

In this section, we evaluate the effect of the FGSM attack method on several deep learning architectures: MobileNetV2 [18], ResNet50 [19], and DenseNet121 [20]. Figure 2 illustrates the results of adversarial data generated by the FGSM model on the MNIST dataset across three classifiers: MobileNetV2, ResNet50, and DenseNet121. The results show that the adversarial images produced by the FGSM model differ from the original images, but the details necessary to identify the digit remain intact if the image blur is ignored. To the human eye, viewing these images still allows reading the corresponding numbers as in the original images. However, when these images are processed by CNN models for recognition, the models often misclassify the digits. The results of the misclassification are presented through quantitative evaluations in the next section.

(a) Original images

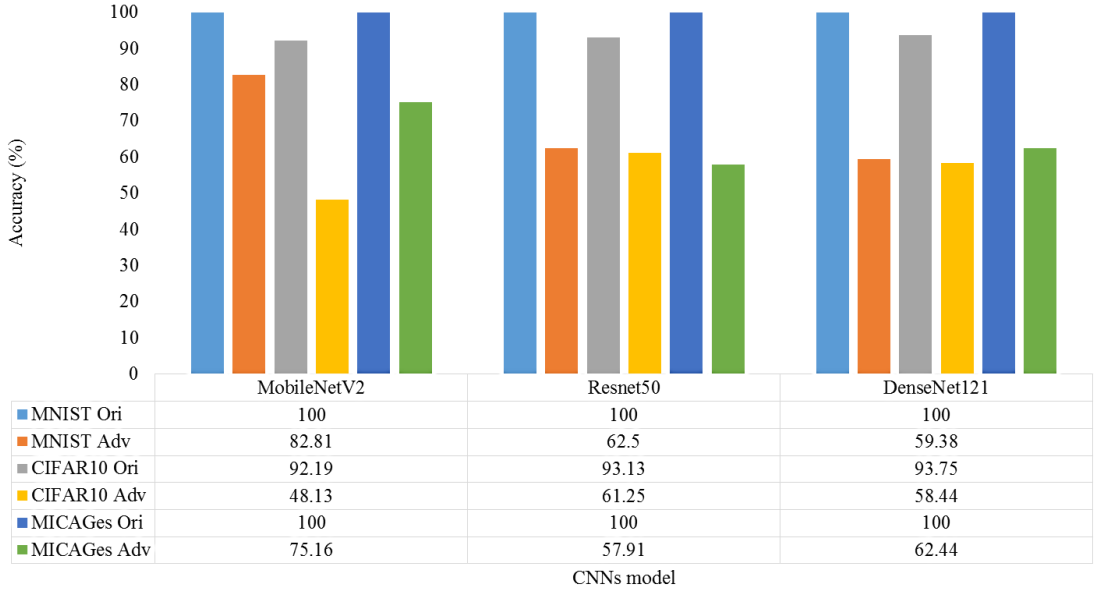
(b) Adversarial images

Figure 1. *Illustration of original and adversarial images generated from the MNIST dataset.*

In this diagram, the model performs recognition on the original data from the MNIST, CIFAR10, and MICAHandPose datasets. Subsequently, these datasets are used to generate adversarial images using the FGSM method and are simultaneously passed through a deep convolutional neural network for recognition. The results are depicted as shown in figure 2 below.

The results in figure 2 show that the CNN models achieve very high performance on the original datasets, such as 100% on MNIST and MICAHandPose datasets and 92.19%, 93.13%, and 93.75% for the three models on the CIFAR10 dataset. However, the classification results of these models significantly decrease the adversarial images generated by FGSM attacks across all three datasets. For MNIST, the recognition accuracy drops to only 82.81%, 62.50%, and 59.38%; for CIFAR10, it decreases to 48.13%, 61.25%, and 58.44%; and for MICAHandPose, it reduces to 75.16%, 57.91%, and 62.44% for MobileNetV2, ResNet50, and DenseNet121 respectively. These

experimental results indicate that FGSM attacks reveal the vulnerabilities of CNN image classification models, as their recognition effectiveness declines significantly when faced with adversarial images generated by FGSM. This reality underscores the necessity of effective defense mechanisms to ensure that deep learning CNN models are not only more resistant to adversarial attacks but also maintain high reliability and generalization capability.

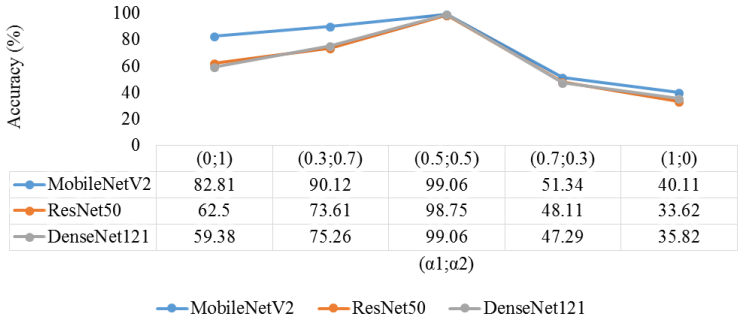


	MobileNetV2	Resnet50	DenseNet121
MNIST Ori	100	100	100
MNIST Adv	82.81	62.5	59.38
CIFAR10 Ori	92.19	93.13	93.75
CIFAR10 Adv	48.13	61.25	58.44
MICAGes Ori	100	100	100
MICAGes Adv	75.16	57.91	62.44

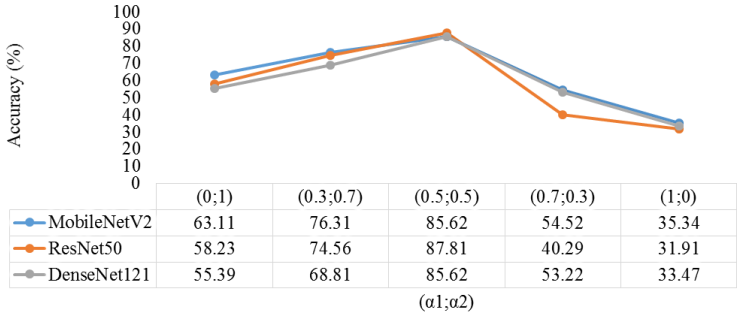
Figure 2. *Accuracy (%) of CNNs model: MobileNetV2, Resnet50, DenseNet121 on original images and adversarial images.*

3.2. Evaluation of the effectiveness of the WR_FGSM defense method for CNN models

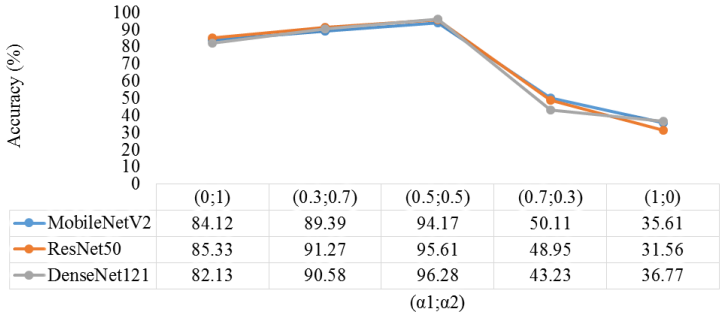
Figure 3 illustrates the defense results of the three CNN models on the MNIST dataset (figure 3(a)), CIFAR10 dataset (figure 3(b)), and MICAHandPose dataset (figure 3(c)). The experiments were conducted with different hyperparameter values α_1 and α_2 in the defense training model. These hyperparameters are varied incrementally, ensuring the combinations follow formula 8, including pairs such as $(\alpha_1 = 0; \alpha_2 = 1)$, $(\alpha_1 = 0.3; \alpha_2 = 0.7)$, ... $(\alpha_1 = 1; \alpha_2 = 0)$ as shown on the y-axis of figure 3. The recognition accuracy of all three CNN models reaches its highest at $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$, with the highest accuracy being 99.06%; 98.75%; and 99.06% for MobileNetV2, ResNet50, and DenseNet121 on the MNIST dataset. Similarly, on the other two datasets, the models also perform best at the same hyperparameter pair ($\alpha_1 = 0.5$ and $\alpha_2 = 0.5$). These values indicate that the models are optimally trained with a combination of original images, half of the regularization component, and half of the noise-augmented data. It suggests that a balanced participation of both regularization and adversarial noise components enhances the model's robustness without bias toward either component. When the pair is ($\alpha_1 = 1$ and $\alpha_2 = 0$), meaning the model is optimized only with original and adversarial images without the regularization component, the recognition results are the worst. Conversely, when ($\alpha_1 = 0$ and $\alpha_2 = 1$), meaning the model is optimized only with original images and the regularization component, and not trained with adversarial data, the recognition performance is worse than with the balanced components but better than with only the regularization component. This outcome highlights the importance of both the adversarial component and the weight regularization component in the training process, emphasizing that a combination of the two yields the best robustness and accuracy.



(a) Results of MNIST dataset



(b) Results of CIFAR10 dataset



(c) Results of MICAHandPose dataset

Figure 3. *Accuracy (%) of CNN models with various hyper-parameters.*

The highest results of all three models across the three datasets are compared with the recognition accuracy of the models on the original data, as shown in table 1 below. These results demonstrate that an effective defense mechanism significantly improves the model's accuracy when under attack, across all three classifiers and on all three datasets. For example, the accuracy against FGSM attacks using MobileNetV1 on the MNIST dataset increased from 82.81% (figure 2) to 99.06% (table 1). Similarly, for the MICAHandPose dataset, the accuracy increased from 75.16% (FGSM attack with MobileNetV1) to 94.17%.

Table 1. *Accuracy (%) of CNN models with our defense method on various datasets.*

CSDL	Data types	MobileNetV2	ResNet50	DenseNet121
MNIST	Original	100%	100%	100%
	Adversarial	99.06%	98.75%	99.06%
CIFAR10	Original	92.19%	93.13%	93.75%
	Adversarial	85.62%	87.81%	85.62%
MICAHandPose	Original	100%	100%	100%
	Adversarial	94.17%	95.61%	96.28%

4. CONCLUSIONS

This paper analyzes the impact of FGSM adversarial attacks on deep CNN models in image classification tasks across various datasets. Experimental results indicate that CNN models suffer significant accuracy degradation, with reductions of over 44% compared to using original data. This underscores the necessity of developing defense mechanisms to enhance the reliability and robustness of models against adversarial attacks. In this study, we proposed a novel defense mechanism, and experimental evaluations across multiple datasets demonstrated the effectiveness of the proposed approach for CNN models. However, there are current limitations, including the evaluation scope being limited to three CNN models and a single FGSM attack technique. Additionally, the datasets used are primarily small and less diverse. Future research will extend the investigation to more complex attack methods such as BIM, PGD, and C&W, compare the proposed method with other defenses like Feature Squeezing and TRADES, and conduct experiments on various CNN architectures with larger, more diverse datasets. These efforts aim to support applications in defense and security fields.

REFERENCE

[1]. Juanjuan Weng, Zhiming Luo, Dazhen Lin, Shaozi Li. *"Comparative evaluation of recent universal adversarial perturbations in image classification"*. Computers & Security, 136 (2024), 103576. (2024).

[2]. Jaydip Sen, Abhiraj Sen, and Ananda Chatterjee. *"Adversarial Attacks on Image Classification Models: Analysis and Defense"*. arXiv preprint arXiv:2312.16880, (2023).

[3]. Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. *"One pixel attack for fooling deep neural networks"*. IEEE Transactions on Evolutionary Computation 23, 5, 828–841, (2019).

[4]. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. *"Deepfool: a simple and accurate method to fool deep neural networks"*. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2574–2582, (2016).

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

[6]. Jianbo Chen, Michael I Jordan, and Martin J Wainwright. *"Hopskipjumpattack: A query-efficient decision-based attack"*. In 2020 ieee symposium on security and privacy (sp). 1277–1294, (2020).

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

[8]. Aleksander Mkadry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. *"Towards deep learning models resistant to adversarial attacks". stat* 1050, 9, (2017).

[9]. Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. *"Adversarial examples in the physical world"*. In Artificial intelligence safety and security. Chapman and Hall/CRC, 99–112, (2018).

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

[11]. Chen, P., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. *"ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models"*. arXiv preprint arXiv:1708.03999, (2017).

[12]. Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. *"Black-box adversarial attacks with limited queries and information"*. Proceedings of the 35th International Conference on Machine Learning (ICML), (2018).

[13]. Brendel, W., Rauber, J., & Bethge, M. *"Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models"*. arXiv preprint arXiv:1712.04248, (2018).

[14]. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. *"DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks"*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016).

[15]. Erh-Chung Chen and Che-Rung Lee. *"Towards fast and robust adversarial training for image classification"*. In Proceedings of the Asian Conference on Computer Vision, (2020).

[16]. Xu, W., Evans, D., & Qi, Y. *"Feature squeezing: Detecting adversarial examples in deep neural networks"*. Network and Distributed Systems Security (NDSS) Symposium, (2018). https://arxiv.org/abs/1704.01155

[17]. Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. *"The limitations of deep learning in adversarial settings".* In 2016 IEEE European symposium on security and privacy (EuroS&P). pp. 372–387, (2016).

[18]. Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. *"Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation".* CoRR abs/1801.04381, (2018).

[19]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In The IEEE Conference on Computer Vision and Pattern Recognition. 770–778, (2016). doi:10.1109/CVPR.2016.90

[20]. G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger. *"Densely Connected Convolutional Networks".* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269, (2017).

[21]. Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. *"Gradient-based learning applied to document recognition".* In Proceedings of the IEEE, Vol. 86. pp. 2278–2324, (1998).

[22]. Alex Krizhevsky. *"Learning multiple layers of features from tiny images".* In Master's thesis, University of Toronto. Toronto, Canada, (2009).

[23]. Huong-Giang Doan, Ngoc-Trung Nguyen, *"New blender-based augmentation method with quantitative evaluation of CNNs for hand gesture recognition",* Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), Vol. 30, No.2, pp. 796–806, pp. 214-221, (2023).

TÓM TẮT

Tấn công FGSM đối với các bộ phân loại ảnh CNN: Phân tích lỗ hổng và đề xuất giải pháp phòng thủ hiệu quả

Mạng nơ ron tích chập CNN (Convolutional Neural Network) đã cho thấy nhiều ưu điểm nổi trội và do đó đã được ứng dụng phổ biến trong nhiều lĩnh vực khác nhau. Tuy nhiên, các cuộc tấn công đối nghịch đã cho thấy những lỗ hổng nghiêm trọng của các mô hình này, đe dọa đến tính bảo mật và độ tin cậy của hệ thống. Mặc dù đã có nhiều nghiên cứu đề cập đến tấn công vào các mô hình học sâu, song tác động cụ thể của những cuộc tấn công này lên các bộ phân loại hình ảnh dựa trên CNN vẫn là vấn đề cần được làm rõ thêm, đặc biệt là đối với các mô hình CNN phổ biến là nền tảng của nhiều ứng dụng quan trọng trong thực tiễn. Nghiên cứu này phân tích điểm yếu của các bộ phân loại hình ảnh CNN trước tấn công đối nghịch FGSM (Fast Gradient Sign Method), đồng thời đề xuất giải pháp phòng thủ hiệu quả mang tên WR_FGSM. Các thử nghiệm trên các bộ cơ sở dữ liệu tiêu chuẩn cho thấy một số mô hình CNN bị ảnh hưởng khá nghiêm trọng bởi FGSM. Các hình ảnh nhiễu do tấn công này tạo ra không chỉ đánh lừa các bộ phân loại hình ảnh CNN mà còn khiến thị giác của con người khó phân biệt với hình ảnh gốc. Giải pháp phòng thủ đề xuất WR_FGSM ngoài việc sử dụng phương pháp phòng thủ hiệu quả hiện nay là Huấn luyện đối nghịch, chúng tôi còn tích hợp thêm kỹ thuật chính quy hóa vào các bước huấn luyện. Điều này cho phép bảo vệ các mô hình CNN một cách hiệu quả trước tấn công FGSM trong khi vẫn duy trì sự cân bằng giữa khả năng kháng tấn công và tính khái quát của mô hình.

Từ khóa: Mạng nơ ron tích chập CNN; Tấn công đối nghịch; Phòng thủ; Huấn luyện đối nghịch; Kỹ thuật chuẩn hóa.