Eng Int Syst (2025) 4: 377–384

Engineering
Intelligent Systems

Algorithm for Detection and Defense of Neural Network Technology Based on Neural Network and Multimedia

Hui Ke*

Department of Network and Information Security, Chongqing Vocational Institute of Safety & Technical, Wanzhou 404120, Chongqing, China

Neural networks have been widely used in the fields of image recognition, voice recognition, and natural language processing among others. However, the neural network model is less robust against adversarial samples. That is, when small perturbations are artificially added in the input data, the output of the model will change. This phenomenon, known as adversarial example attack, can lead to misclassification and performance degradation of the model. In response to this problem, the academic community has proposed a variety of adversarial sample attack detection and defense methods. Adversarial attack detection is intended to detect adversarial samples and filter them out so that they are not input into the model. On the other hand, the purpose of adversarial defense is to increase the robustness of the model during training, making it more stable against adversarial samples. At present, there are still some problems in regard to the detection and defense of adversarial sample attacks. Therefore, further research and exploration are still of great significance. In this study, we examine "Adversarial Sample Attack and Defense in Neural Networks" to determine the protection capabilities of neural network technology. Through the experiment, the detection experiment of adversarial samples is carried out, and the three attack methods of C&W, FGSM and FIA are detected by using the adversarial detection algorithm based on neural network technology, and the detection success rate is recorded. Experimental results show that the average detection success rate of C&W, FGSM and FIA using the adversarial detection algorithm based on neural network technology is 97.257%, 95.354% and 94.602%, respectively. This indicates that the algorithm has a high detection success rate for these three attack methods, can effectively identify adversarial samples, and improve the robustness and accuracy of the model.

Keywords: system attack, neural network, against the samples, image recognition

1. INTRODUCTION

With the development of deep learning technology, neural networks have been widely used in computer vision, natural language processing and other fields where they have achieved excellent results. However, studies have shown that neural networks are vulnerable to some adversarial samples, which makes them potentially exploitable by malicious attackers. Adversarial samples are a technique for modifying the input data of a neural network in order to mislead its classification results. Such attacks on neural networks may pose a serious threat to people's lives, property and public safety. Therefore, methods to detect and defend against adversarial attacks on neural networks are crucial. At present, there are still many problems in the attack detection and defense technology for neural network adversarial samples, such as the universality and robustness of the method, which need further research and exploration.

A number of scholars have conducted research on neural network technology and its applications. Yang Guangyu pointed out that an artificial neural network is an important tool for machine learning, attracting much attention in the field of neuroscience. In addition to providing powerful data analysis techniques, it also provides new methods for neuroscience

*corresponding author Email: ke_229@163.com

research. It can establish complex behavior, heterogeneous neural activity and circuit connection models, and explore the optimization of the nervous system [1]. Cong credits machine learning with neural networks with success in a variety of fields, including image recognition and medical applications. However, direct application to problems in quantum physics is challenging due to the complexity of many-object systems. He proposed an algorithm based on quantum circuits, inspired by convolutional neural networks, called Quantum Convolutional Neural Networks. This neural network technique can be efficiently trained and implemented, and is capable of identifying quantum states associated with symmetry-protected topologies [2]. Bau found that deep neural networks focus on hierarchical representations for solving complex tasks of large datasets. To understand the representations learned by the network, he proposed a network anatomy framework to systematically analyze the semantics of individual hidden units in image classification and image generation networks. Bau used a convolutional neural network trained to classify scenes and discover units that match different object concepts. Second, a similar approach was used to analyze a Generative Adversarial Network model and found that objects can be added and removed depending on the context [3]. Kriegeskorte found that that deep neural network models were originally inspired by neurobiology and have become powerful tools for machine learning and artificial intelligence. These models can approximate functions and dynamics by learning from examples. He introduced the biologist's neural network model and deep learning, and explained the expressive capabilities of feedforward and recurrent networks and the method of setting parameters using the backpropagation algorithm [4]. This literature yields much information about neural network technology, which is valuable to the research conducted in the current study.

Studies have also been conducted on neural network technology and adversarial samples. Zhang pointed out that deep neural networks can be attacked by adversarial perturbations that can trick image classifiers into making wrong predictions with high probability. Adversarial perturbations are also capable of fooling real-world machine learning systems and are easily transferable between different datasets. To this end, methods to defend against adversarial perturbations have attracted extensive attention. Zhang conducted a comprehensive survey of classical and state-of-the-art defense methods and identified their main concepts, algorithms, and underlying assumptions of adversarial perturbations [5]. Research by Xu shows that deep neural networks achieve state-of-the-art performance in remote sensing scene classification. However, deep learning algorithms are vulnerable to attacks from adversarial samples. He analyzed the threat of adversarial samples to remote sensing scene classification with deep neural networks, and introduced an adversarial training strategy to improve the anti-interference ability of the model. Experimental results showed that the adversarial training strategy can mitigate the vulnerability of deep neural networks to adversarial samples, thereby improving their practical deployment capabilities in safety-critical remote sensing domains [6]. Although the aforementioned researchers have studied in depth, the literature contains relatively few experimental outcomes.

In this paper, a study of neural network adversarial example attacks will be conducted to improve the security and robustness of neural networks. The experimental results show that if there is no adequate defense measure, the average success rates of C&W, FGSM and FIA attacks are 84.03%, 43.74% and 54.17%, respectively. However, the adversarial detection algorithm based on neural network technology used in this paper can significantly improve the defense success rate. The average defense success rate of C&W, FGSM and FIA reached 97.25%, 92.98% and 94.93%, respectively. The innovation of this paper is that it uses an adversarial detection algorithm based on neural network technology, which can effectively identify and defend against the aforementioned attacks. The algorithm adopts a novel neural network structure and training strategy to achieve automatic detection and improve the robustness of adversarial samples.

2. RESEARCH ON NEURAL NETWORK ADVERSARIAL SAMPLES

2.1 Main Features of Adversarial Samples

Adversarial samples are artificially manufactured input data with specific noise or perturbation, which can deceive machine learning models so as to produce wrong classification results or misinterpret the input data [7]. Adversarial perturbation refers to the addition of slight disturbances to standard images, audio, and data, which does not affect human recognition but can cause significant errors in machine learning models. The attack method exploits the constraints of the model and then deliberately guides the model to produce inaccurate results. It is the emergence of these controversial cases that has sparked widespread academic concern about the robustness and security of machine learning models.

Adversarial cases have three main characteristics: deception, repeatability, and diversity. These are shown in Figure 1.

The first of these occurs because neural network algorithms add noise or interference to the input data, leading to faulty classification and predictions. Also, hostile samples can trick machine learning models, greatly affecting the security of machine learning.

The second point is that the unity of the adversarial situation is achieved by the traditional algorithm in the generation process, which is related to various models. This property facilitates the replication and dissemination of adversarial instances, laying the foundation for studying the vulnerability of adversarial samples, with broader implications for machine learning models.

Third, because adversarial instances are diverse, the use of the same machine learning model may lead to inaccurate predictions. The diversity of confrontation instances makes it very challenging to analyze the resilience of these instances.

2.2 Attack Targets of Adversarial Samples

Attackers will use the attack target against the sample to maliciously attack the machine learning model, resulting in system crash and data leakage. In order to improve the

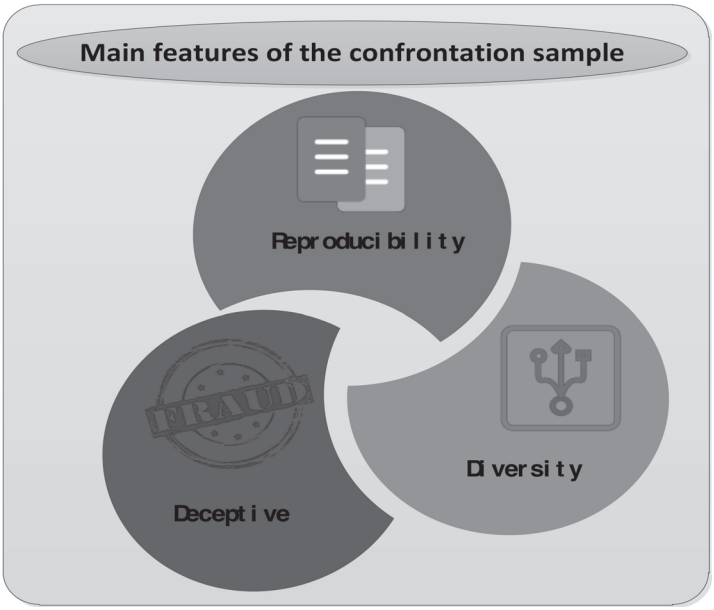

Figure 1 Main features of adversarial samples.

robustness and security of machine learning models, it is necessary to study the attack targets and generation methods of adversarial samples, explore corresponding defense strategies and algorithms, and establish a sound security mechanism to ensure the reliability and practicability of machine learning models. Adversarial samples are those that are deliberately designed to mislead input data for machine learning algorithms, which are widely used in computer vision, speech recognition, natural language processing and other fields. The attack targets of adversarial samples can be divided into two types: targeted attack and untargeted attack. The details are as given below.

(1) Targeted attack

In this attack, the attacker modifies the input image so that the target network outputs a wrong judgment result for a specific category of input image; in this case, the output result of the target model is unique to the attacker determined [8–9]. Target attack refers to the attack on specific classification results; that is, it adds perturbation to the original data (such as images, sounds, texts, etc.) so that the model misjudges it as the specified target category. For example, in image classification, an attacker hopes to misjudge a picture of a dog as a picture of a cat, and can make the model recognize it as a cat, not a dog, by adding a small perturbation (almost imperceptible to the human eye) to the original image [10].

(2) Targetless attack

Untargeted attack is an attack launched against machine learning models. Unlike targeted attack, it does not need to pre-determine the classification target of the attack, but interferes with the classification effect of the entire model by adding perturbations to the original data, resulting in misjudgment and reducing its accuracy [11–12]. This attack method is very flexible and difficult to detect. Hackers can use this attack method to destroy various machine learning applications, such as autonomous driving, smart security, etc., thereby endangering the safety of people's lives and property [13]. Therefore, it is particularly important to guarantee the security and robustness of the model.

2.3 Attack methods against samples

Adversarial sample attacks have the characteristics of high concealment, the ability to bypass traditional security defense mechanisms, and low attack costs; hence, special defense measures must be taken. There are two types of attack: white-box attacks and black-box attacks. When faced with an adversarial sample attack, it is first necessary to understand the characteristics of the attack in order to formulate an effective defense strategy. The detailed description of the attack method against the sample is as follows:

(1) White-box attack

White-box attackers can fully understand the internal structure, parameters, and details of training data and algorithms of machine learning models [14–15]. Therefore, compared with black-box attackers, they have more advanced attack methods, such as fast gradient sign method (FGSM), projected gradient descent attack (PGD), and attacks based on manifold learning. These attack methods can create deceptive adversarial samples by fine-tuning model parameters, adding disturbances, and using gradient information and other technical means, and thereby interfere with the output of the machine learning model, causing the model to make misjudgments or misjudgments, thereby achieving the purpose of the attack. In addition, white-box attackers can use other more complex attack methods, such as attacks based on Generative Adversarial Networks (GAN), to attack by stealth.

(2) Black-box attack

The challenge for black-box attackers is that they cannot peer directly into the inner workings of machine learning models; they can only guess the model's behavior through inputs and outputs like blind people. This relationship, separated by a thin layer of gauze, leaves the black box attacker in a maze of information, requiring trial and error and speculation to find clues to crack [16–17]. In this case, black-box attackers need to use a series of different attack methods to try to attack this model, such as black-box optimization, migration attacks, and meta-model-based

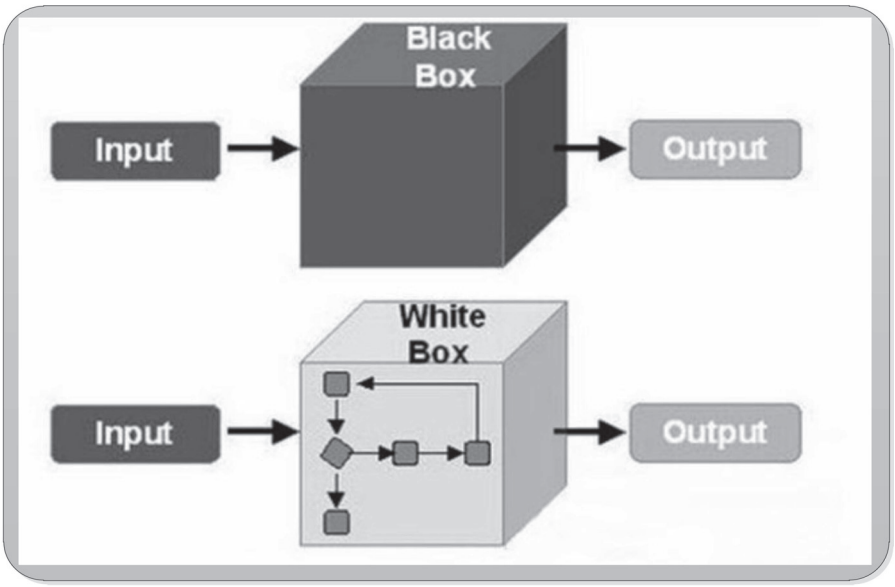

Figure 2 Black-box and white- box attacks on adversarial samples.

attacks. In addition, when using black-box attacks, attackers can exploit the properties of adversarial samples to deceive machine learning models in order to achieve their goals. These attack methods all have one thing in common: they can bypass the input detection of the machine learning model, thus causing the model to fail to classify or predict correctly. Therefore, defense against black-box attacks is critical to the security of machine learning models. Appropriate security measures must be implemented to reduce the risks to machine learning models.

Of these, the black-box and white-box attacks on adversarial samples are shown in Figure 2.

3. RESEARCH ON IMAGE-BASED ADVERSARIAL EXAMPLE ATTACKS

3.1 Neural Network Adversarial Attacks on Images

Currently, due to the widespread use of machine learning and artificial intelligence technologies in many fields, attacks against these technologies are also becoming more common. Among them, adversarial attack is a relatively common and challenging attack method. Typically, it targets related domains such as imagery, natural language processing, video, and speech. A common way to generate adversarial samples on graph data is to modify the node characteristics in the original graph or to modify the topology of the graph. However, to guarantee the effectiveness of adversarial samples, these modified graphs need to be "similar" based on some kind of perturbation evaluation matrix, and the added perturbations must be guaranteed to be "imperceptible". This is explained below.

(1) Perturbations at the edge level

Edge-level perturbation refers to the modification of the original graph by the attacker within the budget, and include operations such as adding, removing, and reconnecting edges. This perturbation is usually measured by increasing or decreasing the number of connected edges. For instance, an intruder might alter the structure of the initial graph by inserting incorrect edges, thus deceiving the algorithm's operational outcomes. Furthermore, the assailant has the ability to remove certain crucial edges, thereby disrupting the original graph's connectivity and impacting the algorithm's output. In cases of edge-level disturbance, the assailant must optimize alterations to the initial graph to the fullest extent feasible within the allocated budget.

(2) Disturbance at node level

Perturbation at the node level involves the addition or removal of nodes in a graph network, impacting the characteristics of the intended node. Such disturbances directly alter the graph's topology and node characteristics, profoundly impacting graph-based machine learning algorithms. Cybercriminals have the ability to disrupt algorithmic choices by introducing nodes akin to but deceptive to the target node, or by removing crucial nodes, thereby destabilizing the graph. The extent of node-level perturbation can be calculated according to the number of nodes modified by the attacker or the distance between the target node feature vectors before and after the change. If the attacker adds many similar but misleading nodes, the distance between the feature vectors may not be large, but the topology of the entire graph may change significantly, thus affecting the algorithm results. If an attacker deletes a key node, it may lead to changes in the connection relationship of the target node, which in turn affects its position, degree, aggregation coefficient and other important features, seriously affecting the algorithm output. Therefore, node-level disturbances need to thoroughly consider multiple indicators to evaluate their impact.

(3) Perturbation of preserved structure

In structure-preserving perturbations, an attacker can modify only the topology of the original graph within a budget. This approach is similar to perturbation at the edge level, often involving operations such as adding, removing, or

reconnecting edges. However, the attacker has to consider more structure-preserving properties, such as total degree, node distribution, and connectivity, etc. The perturbation approach is crucial for targeting algorithms on graphical data, reducing the effects of changing the initial graph structure. The method's measure of alteration indicates the extent of the disturbance. Aggressors disrupt algorithms by adding fake edges or removing edge information, affecting network connectivity and algorithm results. Attackers must ensure alterations preserve the graph's structure and important features.

(4) Perturbations that retain attributes

Aggressors use attributes in a graph to change nodes or edges and create hostile examples. This type of disturbance method is commonly used to assess node or edge feature data. For instance, in social network analysis with attribute graphs, attackers can modify user attributes to impact recommendation systems or analytical tools. Attackers can determine the magnitude of perturbation using various graph attribute-based bias measures, such as node degree, connectivity, distance, and clustering coefficient, by preserving the structure and feature information of the original graph. Moreover, attackers can use machine learning techniques such as adversarial generative networks to generate adversarial samples to bypass the detection mechanism of the model, and the problem of concealment needs to be considered to avoid detection risks. If the attacker can successfully preserve the structure and feature information of the original graph, and at the same time have a significant impact on the output of the algorithm, it may successfully interfere with the prediction and analysis capabilities of the model.

3.2 Adversarial Detection Algorithm Based on Neural Network Technology

A neural network is a mathematical model that simulates a biological nervous system for information processing, and it consists of a large number of neurons and the connections between them [18]. Each neuron receives input signals from other neurons, weights and sums these signals through an activation function, and produces an output signal, which in turn serves as the input signal for the next neuron. By continuously optimizing parameters such as weights and biases, the neural network can learn the inherent laws and characteristics of the input data. A neural network has three layers: an input layer, a hidden layer, and an output layer. The input layer accepts input data from the outside world and passes this data to the hidden layer. The hidden layer is the core of the neural network, which processes the input data and optimizes the performance of the network by continuously adjusting the weights and biases of the hidden layer. Finally, the output layer outputs the processed data to complete the calculation process of the entire neural network.

In order to achieve the unsupervised detection of adversarial samples, at the same time, it can decouple the knowledge of adversarial attacks and improve the robustness of the defense model. In this paper, a detection method based on anomaly scoring is adopted. Using this approach, an adversarial score is constructed to detect anomalies by comprehensively considering the pixel-level mean square error between the reconstructed image and the original image and the cross-entropy loss between the predicted class probability distributions inferred by the target classification model for the two enter. Its calculation formula is:

$$\text{AdvScore} = \|x - D(x)\|_2^2 + \sum_{i=1}^{k} F(x)_i * \log\left(\frac{1}{F(D(x))_i}\right) \tag{1}$$

where x is the original input image, $D(\bullet)$ is the noise reduction network model, $D(x)$ is the reconstructed image processed by the noise reduction network, $F(x)_i$ and $F(D(x))_i$ is the probability that the target classification model predicts the original image and the reconstructed image as the category respectively.

There is less noise in benign samples, so the error and cross-entropy are small, and the adversarial score is also small. Conversely, adding noise by the attacker in the adversarial samples leads to large errors and cross-entropy, as well as large adversarial scores. Therefore, the samples with small adversarial scores can be judged as benign samples, and the samples with large scores can be judged as adversarial samples. To achieve the purpose, an effective threshold needs to be set, and the abnormal threshold can be calculated based on the normal distribution confidence interval of the adversarial score of the benign sample. The expression of the adversarial score set S calculated from the benign sample set is:

$$S = (s_1, s_2, \ldots, s_n) \tag{2}$$

Then, the mean of the adversarial scoring set of benign samples is calculated with:

$$\mu_s = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{3}$$

The standard deviation of the set of adversarial scores for benign samples is computed with:

$$\rho_s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - \mu_s)^2} \tag{4}$$

Then, the exception threshold T_s is calculated:

$$T_s = \mu_s + \beta * \rho_s \tag{5}$$

where β is the control parameter of the confidence width, the higher the confidence level, the wider the confidence interval.

The input samples are determined to be either benign or adversarial samples according to the adversarial score threshold calculated on the benign samples. When the adversarial score of an input sample is less than a set threshold, it is considered benign; otherwise, it is an adversarial sample if it is greater than the threshold. This threshold-based classification method is simple and effective, and has good results in practical applications.

Table 1 Experimental environment hardware equipment information table.

Name _	Detailed information	Quantity
Processor _	Intel Core i9-11900K	2
Display card	NVIDIA GeForce RTX 3090	2
mainboard	MSI MPG B550 Gaming Edge WiFi	2
Internal storage	Corsair Dominator Platinum RGB 64GB DDR4-3200	4
Hard disc	Western Digital Black SN850 NVMe SSD	2

Table 2 Common adversarial attack methods.

Attack method	Attack type	Punching bags	Attack scene
C&W	Optimized attack	Have Target	white box
FGSM	Single step attack	No Target	white box
FIA	Migration attack	Have Target	black box

4. EXPERIMENTS ON THE ATTACK AND DEFENSE OF NEURAL NETWORK ADVERSARIAL SAMPLES

In this experiment, the adversarial detection algorithm based on neural network technology was explored and experimentally tested to evaluate its performance in neural network adversarial sample attack and defense. This experiment focused on the characteristics and application scenarios of adversarial sample attack and defense. The experimental equipment was selected, and an experimental test plan was designed to verify the algorithm's performance in neural network adversarial sample attack and defense.

4.1 Experimental Equipment

In order to conduct adversarial attack experiments involving neural network technology, a large number of matrix operations are performed, requiring high-performance computing equipment. Therefore, in order to carry out the training, calculation and other operations involved in the experiment, the a high-performance computing server is necessary. The hardware used for the experiment is shown in Table 1.

4.2 Experimental Datasets

When conducting adversarial attack experiments involving neural network technology, the selection of data sets is very important because the quality and diversity of data sets will directly affect the accuracy and robustness of adversarial attack defense. If the data set is too simple or single, the model may not be robust enough to resist attacks, making it easy for attackers to exploit vulnerabilities so as to attack. Therefore, in order to improve the robustness of the model, it is necessary to select some challenging data sets, and at the same time ensure their diversity by covering different scenarios and roles. In this way, the model can fully adapt to various attack methods and defensive capabilities. In addition, it is necessary to pay attention to the legality, scale and authenticity of the data set to avoid the problem whereby the trained model is not universally applicable to real-world scenarios. The data sets selected for this experiment are described below.

(1) MNIST data set

This is a classic handwritten digit recognition dataset, which contains 60,000 grayscale images of 28x28 pixels as a training set and 10,000 images as a test set. Each image is marked with a corresponding number from 0-9, which is a 10-category problem. The MNIST dataset has become one of the standard datasets in the machine learning community and is widely used in the field of digital image recognition.

(2) CIFAR-10 data set

The dataset contains 60,000 32x32 pixel color images from 10 different categories. Each category has 6,000 images. Of these, 50,000 images are used as the training set and 10,000 images are used as the test set. The classes in the CIFAR-10 dataset include airplanes, cars, birds, cats, dogs, frogs, horses, boats, and trucks, among others. Compared with MNIST, the CIFAR-10 dataset is more complex, with more categories and higher dimensions; hence, it is more suitable for testing the performance of image classification algorithms.

In this experiment, the following common adversarial attack methods were tested. The detailed information is shown in Table 2.

4.3 Experimental Results

To evaluate the effectiveness of the adversarial detection algorithm, we detected the three attack methods—C&W, FGSM, and FIA. The detection success rate for each method was recorded and analyzed. The results of this analysis are presented in Figure 3. The experimental results are shown in Figure 3.

The data presented in Figure 3 shows that the average detection success rate of C&W, FGSM and FIA by the adversarial detection algorithm based on neural network technology is 97.257%, 95.354% and 94.602%, respectively. This indicates that the algorithm can, to a certain extent, effectively detect the adversarial samples generated by the three attack methods of C&W, FGSM and FIA.

Then, adversarial sample defense experiments were conducted to test the defense success rate of the three adversarial attack methods above against the non-adversarial defense method and the adversarial detection algorithm based on neural network technology in this paper. The results are shown in Figure 4.

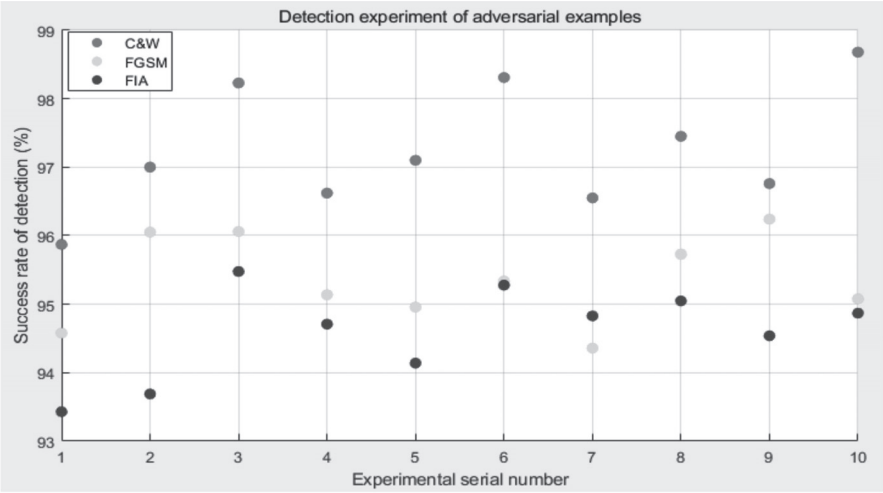

Figure 3 Adversarial example detection experiment.

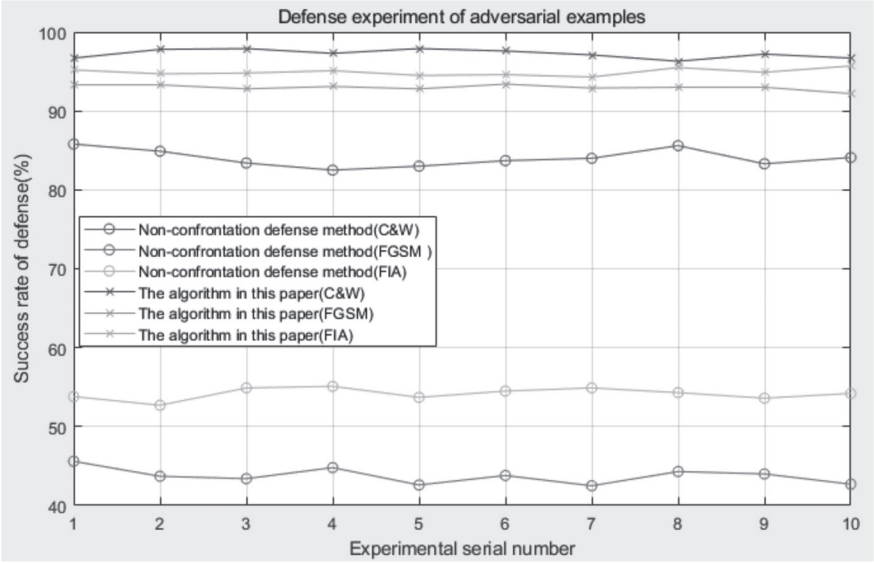

Figure 4 Adversarial example defense experiment.

in the data in Figure 4 shows that for the no-confrontational defense method, the average defense success rates of C&W, FGSM and FIA are 84.03%, 43.74% and 54.17%, respectively. However, the adversarial detection algorithm based on neural network technology proposed in this paper has an average defense success rate of 97.25%, 92.98% and 94.93% for C&W, FGSM and FIA, respectively. This indicates that the method proposed in this paper is versatile and adaptable, can effectively deal with different types of adversarial attacks, and improves the robustness and security of the model.

5. CONCLUSION

The ongoing advancement of deep learning and neural network technology provides a broader development space for the research on adversarial sample attack and defense technology. To address this issue, in this paper, an adversarial detection algorithm based on neural network technology is used to solve the security risks caused by adversarial sample attacks. The results show that the algorithm proposed in this paper has significant versatility and adaptability, and can effectively identify the antagonistic samples in the three popular attack strategies, and the attack means are not limited. Moreover, the algorithm used in this paper strengthens the robustness and security of the model, and provides more powerful protection for users. However, current defense strategies against adversarial sample attacks still face challenges. Some attackers will constantly adjust technical means to circumvent the detection mechanism, thus reducing its effectiveness. Hence, further research is needed to improve the utility and reliability of the technology to ensure it can respond to changing threats in a timely manner.

REFERENCES

1. Yang, Guangyu Robert, and Xiao-Jing Wang. “Artificial neural networks for neuroscientists: a primer.” Neuron 107.6 (2020): 1048–1070.

2. Cong, Iris, Soonwon Choi, and Mikhail D. Lukin. "Quantum convolutional neural networks". Nature Physics 15.12 (2019): 1273–1278.
3. Bau, David. "Understanding the role of individual units in a deep neural network". Proceedings of the National Academy of Sciences 117.48 (2020): 30071–30078.
4. Kriegeskorte, Nikolaus, and Tal Golan. "Neural network models and deep learning". Current Biology 29.7 (2019): R231-R236.
5. Zhang, Xingwei, Xiaolong Zheng, and Wenji Mao. "Adversarial perturbation defense on deep neural networks". ACM Computing Surveys (CSUR) 54.8 (2021): 1–36.
6. Xu, Yonghao, Bo Du, and Liangpei Zhang. "Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses". IEEE Transactions on Geoscience and Remote Sensing 59.2 (2020): 1604 - 1617.
7. Xu, Zheng, MM Kamruzzaman, and Jinyao Shi. "Method of generating face image based on text description of generating adversarial network". Journal of Electronic Imaging 31.5 (2022): 051411–051411.
8. Ferraro, Giovanna, and Antonio Iovanella. "Clairvoyant targeted attack on complex networks." International Journal of Computational Economics and Econometrics 8.1 (2018): 41–62.
9. Sasaki, Ryoichi. "Development and Evaluation of Intelligent Network Forensic System LIFT Using Bayesian Network for Targeted Attack Detection and Prevention". International Journal of Cyber-Security and Digital Forensics 7.4 (2018): 344–354.
10. Wang jing."Design of intelligent traffic sign image recognition system based on machine learning algorithms". Engineering Intelligent Systems, vol. 32 no. 5(2024), pp. 457–464.
11. Xiao, Yu. "A multitarget backdooring attack on deep neural networks with random location trigger". International Journal of Intelligent Systems 37.3 (2022): 2567–2583.
12. Yang, Dong, Wei Chen, and Songjie Wei. "DTFA: Adversarial attack with discrete cosine transform noise and target features on deep neural networks". IET Image Processing 17.5 (2023): 1464–1477.
13. Ma Zhigang. "Visual servo control of robot arm based on image features".Engineering Intelligent Systems, vol. 31 no. 6(2023), pp. 501–51.
14. Won, Jongho, Seung-Hyun Seo, and Elisa Bertino. "A secure shuffling mechanism for white-box attack-resistant unmanned vehicles". IEEE Transactions on Mobile Computing 19.5 (2019): 1023–1039.
15. Wang, Yixiang. "IWA: integrated gradient-based white-box attacks for fooling deep neural networks". International Journal of Intelligent Systems 37.7 (2022): 4253–4276.
16. UGHI, Giuseppe, Vinayak Abrol, and Jared Tanner. EP Neural Networks. "Optimization and Engineering 23.3 (2022): 1319–1346.
17. Wei, Xingxing, Huanqian Yan, and Bo Li. "Sparse black-box video attack with reinforcement learning". International Journal of Computer Vision 130.6 (2022): 1459–1473.
18. Jain, Jay Kumar, and Akhilesh A. Waoo. "An Artificial Neural Network Technique for Prediction of Cyber-Attack using Intrusion Detection System". Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN) ISSN: 2799–1172 3.02 (2023): 33–42.

Hui Ke was born in Xi'an, Shaanxi, P.R. China, in 1983. He graduated from Chongqing University with a Master's degree. Currently, he working at the Chongqing Safety Technical Vocational College. His research interests include network security and network system integration.
E-mail: ke_229@163.com