Threats of Adversarial Attacks in DNN-Based Modulation Recognition

Yun Lin, *Member, IEEE*, Haojun Zhao, *Student Member, IEEE*, Ya Tu, *Student Member, IEEE*, Shiwen Mao, *Fellow, IEEE*, and Zheng Dou, *Member, IEEE*

***Abstract*—With the emergence of the information age, mobile data has become more random, heterogeneous and massive. Thanks to its many advantages, deep learning is increasingly applied in communication fields such as modulation recognition. However, recent studies show that the deep neural networks (DNN) is vulnerable to adversarial examples, where subtle perturbations deliberately designed by an attacker can fool a classifier model into making mistakes. From the perspective of an attacker, this study adds elaborate adversarial examples to the modulation signal, and explores the threats and impacts of adversarial attacks on the DNN-based modulation recognition in different environments. The results show that, regardless of a white-box or a black-box model, the adversarial attack can reduce the accuracy of the target model. Among them, the performance of the iterative attack is superior to the one-step attack in most scenarios. In order to ensure the invisibility of the attack (the waveform being consistent before and after the perturbations), an appropriate perturbation level is found without losing the attack effect. Finally, it is attested that the signal confidence level is inversely proportional to the attack success rate, and several groups of signals with high robustness are obtained.**

***Index Terms*—Adversarial Examples; Deep Neural Networks; Modulation recognition; Radio Security.**

I. Introduction

With the emergence of the information age, the technology level and business scale in the field of wireless communication have achieved leapfrog growths. Radio communication has been greatly improved in terms of speed, stability, communication distance, and communication efficiency. With the wide application of artificial intelligence in this field, radio communication has also made breakthroughs in terms of intelligence and convenience, proved undoubtedly beneficial to better meet the actual needs of people for communication services. Therefore, Cognitive Radio (CR) has been considered a technical means to intelligently sense the spectrum environment and utilize the wireless spectrum efficiently [1]. Due to its high intelligence, CR can continuously perceive various modulation modes, signal-to-noise ratios (SNR), transmission power and other parameters of external environment, followed by analyzing and learning the information [2]. It not only realized the automatic conversion and the dynamic detection of system parameters, but more importantly, solved the problem of spectrum shortage in traditional radio, and effectively improved spectrum utilization.

Modulation recognition (MR) can be regarded as the starting point of the CR, where, in order to meet the different needs among users and make full use of channel capacity, different modulation modes are adopted for communication signals. Based on MR, signal characteristics such as IF waveform, signal spectrum, instantaneous amplitude and instantaneous phase, can be collected and classified for CR to complete related reasoning and learning [3][4]. MR is also important in distinguishing between primary users and secondary users to effectively explore spectrum holes and improving spectrum utilization. In addition, MR plays a very important role in the military and civilian applications. In the military field, MR provides an important technical means for the acquisition of enemy information in radar electronic warfare and the selection of optimal interference and suppression methods. In civilian applications, it plays an important role in the detection of daily radio stations, the management of radio spectrum resources, and the identification of signals such as air traffic control, etc.

With the coexistence of various communication systems, radio data now presents more complex and diverse features, such as randomness, heterogeneousness and massiveness than before. Thanks to its many advantages, deep learning is increasingly applied in communication fields such as modulation recognition [5]. It possess the following advantages: 1) Due to a large number of communication devices and high communication data rates, the massive data required for deep learning are available in the communication systems; 2) Deep learning can autonomously extract features and avoid manual feature selection; 3) Convolutional neural networks (CNN) can use convolutional layers instead of fully connected layers to reduce data parameters [6]. Tu et al. [7] use semi-supervised learning with generative adversarial networks to conduct modulation recognition. Restuccia et al [8] present RFLearn, which enables spectrum knowledge extraction from unprocessed I/Q samples through deep learning directly in the RF loop. Wang et al. [9] propose an edge-assisted crowdcast framework named DeepCast, which makes intelligent decisions at edges based on the massive amount of real-time information to accommodate personalized QoE with minimized system costs. Tang et al. [10] expand the modulation dataset by using the programmatic

This work is supported by the National Natural Science Foundation of China (61771154) and the Fundamental Research Funds for the Central Universities (HEUCFG201830). This paper is also funded by the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation.

Y. Lin, H. Zhao, Y. Tu and Z. Dou are with the College of Information and Communication, Harbin Engineering University, Harbin, China. (Corresponding author e-mail: linyun@hrbeu.edu.cn).

S. Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA.

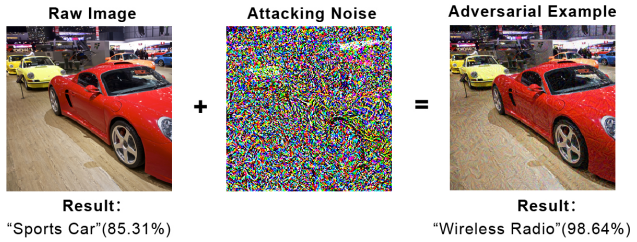


Fig. 1. Left picture is the original image, right picture is the generated adversarial image. By adding an imperceptible small perturbation, the sports car with a confidence of 85.31% is misclassified as the wireless radio with a confidence of 98.64%.

data augmentation method generated by auxiliary classifiers. Gui et al. [11] propose a novel and effective deep learning-assisted NOMA system so as to improve the effectiveness and robustness of spectrum channel allocation. Kato et al. [12] propose a supervised DNN for heterogeneous network traffic. It is worth noting that O'shea et al. [13] prove that DNN can be easily applied to the analogue radio time series data for classification, and the equivalent precision obtained is several times more sensitive than the traditional feature-based classification method.

Although deep learning has unique advantages in solving problems in radio communications, the black-box features and unexplained properties of deep neural networks (DNNs) [14] can cause numerous security risks. Szegedy et al. [15] first discovered this weakness of DNNs in the context of image classification: by adding adversarial examples to the input samples, i.e., small perturbations that are imperceptible to human eyes, the neural network classifier is influenced to change its prediction of the input image samples. Fig. 1 is an example of an adversarial attack on deep learning. It shows that the DNN model is vulnerable to adversarial attacks: if a DNN is seriously threatened or even destroyed by adversarial attacks, normal MR cannot be achieved. This will affect the ability of CR to achieve intelligent communication, resulting in an inability for sensible and fast business decisions. At present, most research on adversarial examples focus on images. Dezfooli et al. [16] found that the existence of universal perturbations affected the network classification for all images. Alexey [17] conducted adversarial training to explore the effect of adversarial examples on large-scale datasets and the relationship between model size and robustness, Athalye et al. [18] proved that even 3D printing of real-world objects may deceive DNN classifiers.

From above research, it is evident that adversarial attacks are a common topic in computer vision, yet it has not been investigated in detail in modulation recognition. It is worth mentioning that Sadeghi et al. [19] first introduced adversarial attack into wireless communications and initiated a direct, digital attack. This attack is considered a *direct access attack* and is implemented by manipulating the receiver's modulation classifier. Due to it assumes the target model has penetrated, it may not be applicable to the real physical world. Even so, considering that it may be more difficult to detect than other

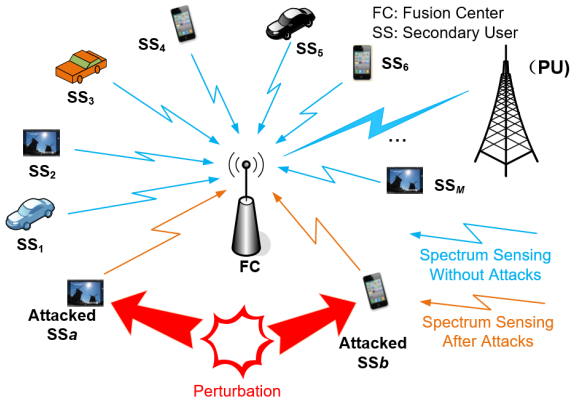


Fig. 2. Schematic of an adversarial attack envisaged on CR, in which the attack has a serious impact on spectrum sensing, resulting in the CR not being able to deploy modulation recognition and ultimately affecting the transmission and security performance of communication network.

forms of attacks, the current research on this direct digital attack still has important value and applications. Thus, in order to thoroughly and comprehensively explore the impact of adversarial attacks on DNN-based modulation recognition, this study envisages from the attacker's point of view, by adding elaborate perturbations to the input samples so as to explore digital attack effects under different environments and parameters. Through exploring the effectiveness and feasibility of adversarial attacks, this study aims to arouse the concern and interests of relevant scholars in the field of communication, and inspire readers to carry out further research. A schematic of the scenario of adversarial attacks envisaged on CR is shown in Fig. 2. In the context of CR, we explore the impact of adversarial attacks in modulation recognition.

The main contributions of this research are as follows:

1) We ascertain the attack effects generated under the white-box model and the black-box model on the modulation signal dataset, and verify the effectiveness of adversarial attacks on signal series datasets.

2) We determine the optimal perturbation level while ensuring the invisibility and effectiveness of the adversarial attack.

3) We find an inverse relationship between the signal confidence level and the attack success rate, and obtain multiple sets of signal types with high robustness.

The rest of this article is organized as follows: Section II introduces the background and methods for generating adversarial examples. Section III outlines the theory of white-box and black-box attacks and elaborates the black-box model used in this study. Section IV describes the details of experimental settings. Section V and Section VI present experiments from the effectiveness and feasibility of adversarial attacks respectively. Section VII summarizes this paper and points out our future work.

II. BACKGROUND AND METHODOLOGY

Adversarial attacks can be classified as targeted attacks and non-targeted attacks [20]. Targeted attacks deceive models so that the input samples are assigned to specific categories. For

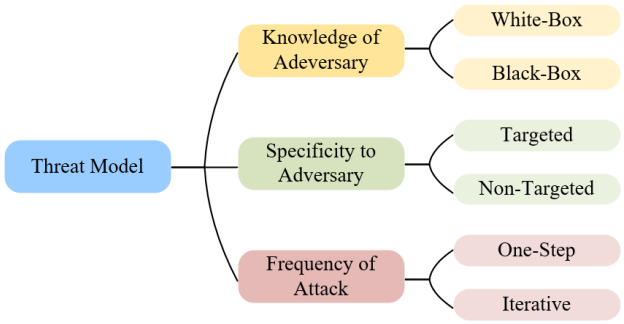


Fig. 3. The classification of our threat model.

example, in modulation recognition, if the adversary specifies that the category of an output signal is BPSK, then the QPSK, QAM64 or other class signals will be misclassified as BPSK. Non-targeted attacks are the opposite of targeted attacks, where the predicted category of adversarial examples is irrelevant, i.e. it can be any type of signal except the original category. In addition, according to the attack frequency, adversarial attacks can be divided into one-step attacks and iterative attacks. A one-step attack generates perturbations by performing a one-step calculation, by calculating the loss gradient of the model at one time, whereas an iterative attack needs to perform the same calculation multiple times to obtain the current perturbation value, which requires constant access to the model and is computationally expensive. The classification of the threat model used in this study is shown in Fig. 3; the details of the knowledge required by the adversary are described in Section III.

In order to explore the attack effects of adversarial examples on modulation recognition, we used the one-step method of Fast Gradient Sign Method (FGSM) and three representative iterative methods: Projected Gradient Descent (PGD) [21], Basic Iterative Method (BIM) [22] and Momentum Iterative Method (MIM) [23]. All of the three iterative methods are improved or modified based on FGSM and therefore they are comparable. In addition, all those methods are applied to non-targeted attacks.

A. FGSM

Goodfellow believed that the non-linear explanation was not appropriate and that the linear behavior of high-dimensional space was sufficient to produce adversarial examples; thus the authors proposed the FGSM [24], a method that quickly generates adversarial examples. Let the initial signal sample be x, the identified tag value be l and the perturbation be ρ, which is supposed to be small enough to satisfy $||\rho||_\infty < \varepsilon$, where ε is the magnitude of the perturbation. The generated adversarial examples are expressed as:

$$\tilde{x} = x + \rho, \tag{1}$$

where the perturbation is expressed as

$$\rho = \varepsilon \cdot \text{sign}(\nabla_x J_\theta(x, l)). \tag{2}$$

Through back propagation, the perturbation value can be calculated, considering the inner product of the weight vector ω and adversarial examples $\tilde{x}$:

$$\omega^T \tilde{x} = \omega^T x + \omega^T \rho. \tag{3}$$

Upon obtaining the inner product, the perturbation ε is amplified with the weight. If the weight vector dimension is n and the mean is m, the maximum value can be reached to εmn, and the linear hypothesis is thus confirmed.

B. PGD

PGD [21] is not only a defence method against first-order adversarial attacks, but also helps to study the robustness of a neural network from the perspective of optimization. It provides a unified view of the previous adversarial training defence methods, and also includes methods for generating adversarial examples. Therefore, this method can serve as a legit white-attack model. In this paper, we use PGD to launch adversarial attacks. The core formula of the method, also known as the *saddle formula*, is given as follows:

$$\min_\theta \quad \rho(\theta) \tag{4}$$

$$\text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)D}[\max_{\delta \in S} L(\theta, x + \rho, l)], \tag{5}$$

where $\mathbb{E}_{(x,y)D}[L]$ is the defined population risk, D is the sample distribution, and S is the allowed perturbations. The saddle problem of this formula consists of two problems: internal maximization and external minimization. The internal maximization problem is to find the perturbation of the data that achieves the greatest loss, which turns out to be an attack problem. Samples that satisfy the maximization condition have a high probability of being adversarial. The external minimization problem is to find the parameters of the model to minimize the loss of attack, which is essentially a problem of training the robust classifier. The saddle problem provides the exact goal of the ideal robust model, i.e., the criterion to measure the robustness.

PGD shares many parameters with BIM, such as , the step-size for each attack iteration, and the number of attack iterations. An additional parameter is a boolean for adding random perturbations. In part V, we mainly set the total perturbation ε to measure the attack effect.

C. BIM

The BIM was proposed by A. Kurakin and Goodfellow et al. [22]. Since FGSM does not require an iterative process to calculate the antagonistic example, it runs much faster than the other methods. BIM is an extended FGSM, where, the adversarial example is generated in multiple iterations. The step size of each iteration is kept small, and the intermediate result values are intercepted after each step to ensure that they are located in the perturbation ε neighborhood of the original input:

$$\begin{cases} x_0 = x \\ x_{n+1} = \text{Clip}_{x,\varepsilon}\{x_n + \varepsilon \cdot \text{sign}(\nabla_x J(x_n, l))\}, \end{cases} \tag{6}$$

where $\text{Clip}_{x,\varepsilon}\{z\}$ denotes clipping z to the range $[x-\varepsilon, x+\varepsilon]$.

In this method, the attack can be expanded by setting the total perturbation ε, similar to the FGSM method, or by setting a one-step iterative attack.

D. MIM

MIM [23] accelerates the gradient descent by accumulating the velocity vector in the gradient direction of the loss function in the iteration. FGSM has a model-based linear feature that enables adversarial examples to be generated quickly and inexpensively. However, in the field of modulation recognition, when there exists strong interference and distortion, the linear assumption cannot be established and the performance of the FGSM is thereby limited. In addition, the iterative attacks shift the adversarial examples to the direction of the gradient in each iteration and often lead to local optimal solutions and overfitting, which reduces the generalization ability of the model. In order to solve this problem, MIM introduces momentum and integrates it into iterative attacks, ensuring the stability of each update direction of the model and the generalization of adversarial examples while maintaining the attack ability. Based on (1) and (2), the gradient of MIM is calculated as follows:

$$g_{t+1} = \mu g_t + \frac{\nabla_x J_\theta(x'_t, l)}{||\nabla_x J_\theta(x'_t, l)||_1}. \tag{7}$$

We enter x'_t into the target function f and calculate the gradient $\nabla_x J_\theta(x'_t, l)$ at the same time. Then we update g_{t+1} by accumulating the velocity vector in the gradient direction, and update x'_{t+1} by applying the symbol gradient as:

$$x'_{t+1} = x'_t + \varepsilon \cdot \text{sign}(g_{t+1}). \tag{8}$$

MIM also shares many parameters with BIM, including the total perturbation, the iteration step and the iteration number of single-step attack, and the perturbation level added by control momentum. In Section V, we will take the total perturbation ε as parameter to measure the attack effect of the model.

III. Attack Model

In adversarial attacks, different knowledge levels about the target model by the adversary will result in different outcomes. The attack strength is lower with a black-box model; but if the adversary is local or it can obtain part of the information through other methods, information leakage may occur and cause a negative impact, turning it into a white-box model.

A. White-Box Attacks

In a white-box attack, the adversary needs to have extensive knowledge of the target DNN model, including input samples, weight values, activation functions, architecture, and training methods. The adversarial examples are generated by continuously accessing the model to calculate the gradient [25]. When the adversary has complete knowledge of the model parameters, the attack on the original model by using adversarial examples are referred to as a white-box attack; this process is shown in the flowchart in Fig. 4.

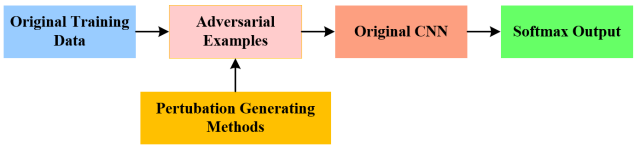


Fig. 4. Flowchart of white-box attacks. The adversary has full knowledge of the target DNN and adversarial attacks are developed on the original model using the methods presented in Section II.

B. Black-Box Attacks

Black-box attacks require more demanding conditions but are more realistic. The black-box attack assumes that the adversary cannot access the target model and only knows the output label and predicted confidence level. Therefore, the research on this problem considers how to conduct model attacks from several aspects, such as estimating the internal structure of the model based on the input and output, adding a larger perturbation to attack the model and implementing attacks by using the transferability of adversarial examples, etc.

One of the most interesting methods for generating black-box attacks is the zero-order optimization (ZOO)-based attack proposed by Chen et al. [26]. Since this attack does not require gradients, it can be directly deployed to black-box attacks without model migration. ZOO does not need to access the target attack model and the Hessian gradient estimation is used to calculate the relevant parameters. In addition to this, other researchers have proposed different yet similar black-box attack methods [27-31].

In contrast, another method of generating black-box attacks exploits the transferability of adversarial examples. Papernot et al. [32] found that adversarial examples were able to deceive other neural networks with different architectures and even classifiers were trained by machine learning methods; therefore, adversarial examples generated by white-box models can be transferred to black-box models.

The black-box attack generation algorithm used in this study is based on the algorithm developed by Papernot [32]. In the absence of any knowledge of the model, a substitute DNN model can be used to simulate the decision boundaries of the approximate target model. It is worth noting that the substitute model is not used to learn to determine the optimal model but to learn the alternative ability to mimic the decision boundary of the target model. Since the black-box attack cannot obtain information from the input samples of the target model, collecting a small amount of data or a part of random data is necessary. Subsequently, the substitute dataset is input to the target model for labelling and data expansion by using a Jacobian-based dataset augmentation to identify any misleading gradient direction as quickly as possible. Through multiple iterations, the required dataset can finally created. The process is shown in Fig. 5.

The Jacobian matrix $J_F = \frac{\partial(F(x,\theta))}{\partial(x)}$ describes the sensitivity of the output F to the input x, indicating that a small input on the gradient boundary results in a large amplitude variation

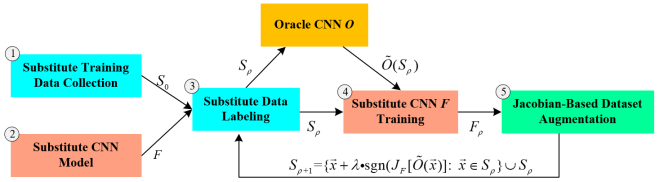


Fig. 5. Black-box attack process based on the Jacobian-based dataset augmentation. Adversary (1) collects the initial substitute dataset S_0 and (2) selects a substitute architecture F. Using target model O, adversary (3) labels S_0 and (4) trains the substitute model F, then (5) uses Jacobian-based dataset augmentation, and repeats (3) through (5) to complete the expansion of the alternative dataset.

of the output value, thereby more accurately describing the sample condition of the model's classification boundary. The synthesis of a new sample algorithm is as follows:

$$S_{\rho+1}=\{\vec{x}+\lambda\cdot\mathrm{sgn}(J_F[\tilde{O}(\vec{x})]:\ \vec{x}\in S_\rho\}\cup S_\rho, \qquad (9)$$

where $O(x)$ is the label of x under the attacked model. The amount of data that is synthesized depends on the number of times that an attacked model is accessed. λ is an enhancement parameter that defines the step size in the sensitive direction, which is identified by the Jacobian matrix, to extend the dataset from S_ρ to $S_{\rho+1}$.

IV. Experimental Settings

We use the TensorFlow and Keras machine learning framework to train DNNs. All experiments were performed on an NVIDIA GTX TITAN Xp using one GPU per run. Algorithms to generate adversarial signals were implemented using CleverHans. CleverHans [33] is an open-source library established by Ian Goodfellow and Nicolas Papernot. It provides standardized reference implementations of adversarial example construction techniques and adversarial training. The library may be used to develop more robust machine learning models and to provide standardized benchmarks of models' performance in the adversarial setting.

A. Details of the Dataset

This paper uses the public dataset RADIOML 2016.10A generated by DeepSig [34]. The dataset has a total of 20 different SNRs from 18dB up to -20dB with a step size of 2, containing a total of 220,000 input samples. There are 8 kinds of digital signals, namely 8PSK, BPSK, QPSK, QAM16, QAM64, CPFSK, GFSK and PAM4, and 3 kinds of analog signals, namely WBFM, AM-DSB and AM-SSB. We used 9000 samples per SNR as the training set and the remaining 2000 samples as the test set. Each signal consists of an in-phase component and a quadrature component, each component has a length of 128.

B. Details of the DNN Model

The features of signals and images are quite different. In order to ensure that the modulation dataset was suitable for the classifier model, we chose the VT-CNN2 model used in [35]. We modified the network parameters, the number of layers, and the initial weight and reshaped the 110000 input signals with a matrix of (2, 128); these were adapted in terms of length, width, and height. We ensured that the parameters were suitable to extract the characteristics of the signal. The flowchart of the VT-CNN2 model is shown in Fig. 6.

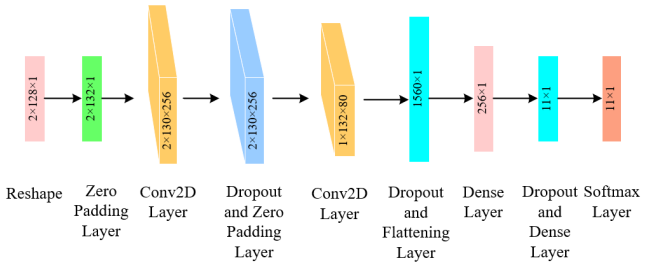


Fig. 6. Schematic diagram of the VT-CNN2 model

As described in Section III, adversarial examples are transferable and can be applied to adversarial attacks on a specific DNN. We only need to design a similar DNN model because the choice of substitute DNN architecture (layer, size, activation function, type) has a limited impact on the transferability of adversarial examples [31]. After the substitute DNN reaches asymptotic accuracy, an increase in the number of iterations will no longer improve the transferability. Therefore, we designed an alternative DNN model, which is inconsistent with the VT-CNN2 model but better adapts to the features of the modulation signal. The specific structure is shown in Fig. 7.

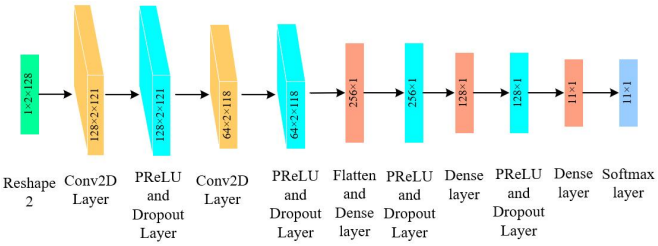


Fig. 7. Schematic diagram of the substitute DNN model.

We use Adam in TensorFlow to optimize the target DNN and the alternative DNN, so that all models in the set have equal weight. Batch normalization was placed before a ReLU layer in all networks, and both the categorical-crossentropy and the sparse-categorical-crossentropy loss functions were tried in the experiments. We also observe the loss or accuracy that had converged after 500 iterations.

V. Experiments on Effectiveness

In this study, from an attacker's perspective, we first add a perturbation to the input signal to generate an adversarial example in the case of a white box, and directly attack the target model to explore the performance of the selected four methods. Then, in the case of a black box, we collect the initial substitute dataset and select a substitute architecture. Using the target model, we label and train the substitute model, then uses Jacobian-based dataset augmentation, and repeats these steps to complete the expansion of the substitute dataset. Finally,

the white-box method can be used to conduct an adversarial attack on the substitute CNN model.

We assume that the input signal sample is x and the perturbation is r_x, then the signal after the perturbation is

$$\tilde{x} = x + r_x \tag{10}$$

The perturbation r_x will be generated by four methods selected from CleverHans, and adversarial examples formed by these methods may query and access the target model in different degrees and different ways. To ensure the effectiveness of $\tilde{x}$, i.e., the DNN misclassifies the input signal, we explore the attack effects of adversarial examples on the DNN for different SNRs and different perturbation levels. Adversarial signals will be generated in both of the white-box model and the black-box model.

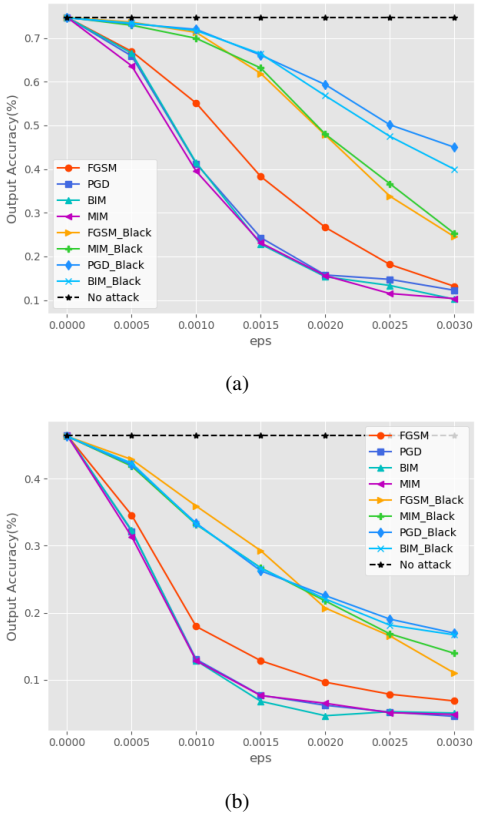
A. Attack Comparison Under Different Perturbations

First, we compare the attack effects under different perturbations through experiments. We use FGSM, BIM, PGD and MIM to generate adversarial examples in the white-box and black-box models. Fig. 8 shows the VT-CNN2 output accuracy for perturbations of 10 dB and -8 dB.

The results indicate that for the white-box attack, with the gradual increases in the levels of perturbation, the accuracy of the model initially decreases rapidly, then slows down, and finally stabilizes. For the black-box attack, the accuracy slowly decreases at a rate lower than that of the white-box attack.

In the white-box attack, as the perturbation increases, the performance of PGD, BIM and MIM with the iterative attack is significantly better than that of the one-step attack FGSM. At a perturbation level of 0.001, the model output accuracy drops by 50% compared to no attack. In the three iterative attack methods, the output accuracies of the model are similar: at 10 dB, the performance of MIM is slightly better than that of BIM and PGD, while at -8 dB, BIM is better than MIM and PGD, indicating that under a high SNR, MIM with momentum iteration is better than the PGD using the saddle model and BIM of the normal iteration. It is however not satisfied at low SNR. We conclude that different methods have different sensitivity to noises; therefore, the appropriate attack model should be chosen based on the actual scenarios.

In the black-box attack, at 10 dB, the performance of the FGSM is similar to that of MIM, and as the perturbation increases, the attack effect is more prominent than BIM and PGD. This is because the iterative nature of BIM and PGD determines that they need to constantly access and interact with the target model. However, under a black-box attack, the accessed model becomes a substitute model, which is different from the original target model; therefore, the attack is less effective. MIM introduces momentum and integrates it into iterative attacks, which not only ensures the stability of each update direction of the model but also the transferability of the adversarial examples while maintaining the attack ability; therfore, the attack effect is relatively better. At -8 dB, the attack effects of various methods are very similar and the performances of FGSM and MIM are almost same as those



(a)

(b)

Fig. 8. The VT-CNN2 output accuracy for perturbations of (a) 10 dB and (b) -8 dB. The results show that the performance of iterative methods are better than that of one-step method FGSM in white-box attacks. In black-box attacks, only under a high SNR and high perturbation that the performances of FGSM and MIM are more prominent than BIM and PGD. In general, the black-box attack is less effective than the white-box attack.

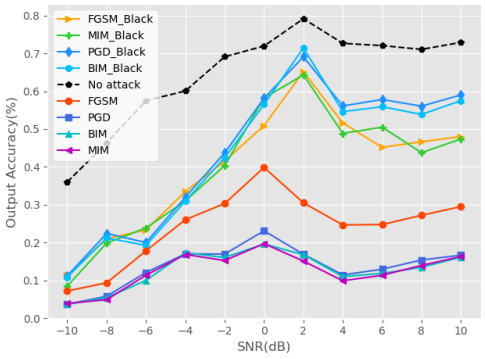
of BIM and PGD. It is assumed that for a low SNR, there are various types of perturbations and interference and the waveform is disordered. The output accuracy of the model is quite low without an attack; therefore, the superiority of the model for different black-box attacks cannot be demonstrated.

In general, at 10 dB and -8 dB, the black-box attack is not as effective as the white-box attack under the selected perturbations. This shows that the black-box attack has some success but less than expected.

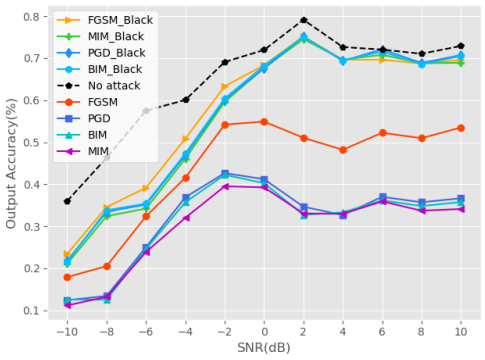
B. Attack Comparison Under Different SNRs

In this round of experiments, we use the perturbation levels of 0.001 and 0.002 for the four methods to generate adversarial examples in different attack models. Our analyses focus on the relationship between model output accuracy and SNRs.

As shown in Fig. 9(a) and 9(b), the output accuracy of the DNN model gradually increases with the increased SNR and followed by fluctuations around a certain value. It is assumed that at a low SNR, there are various perturbations and interference, resulting in waveform distortion; therefore, the model has difficulty in identifying the signal, leading to lower accuracy. At high SNRs, the model however has a higher confidence level of prediction for different categories of signals, making it much harder to be perturbated. Therefore,

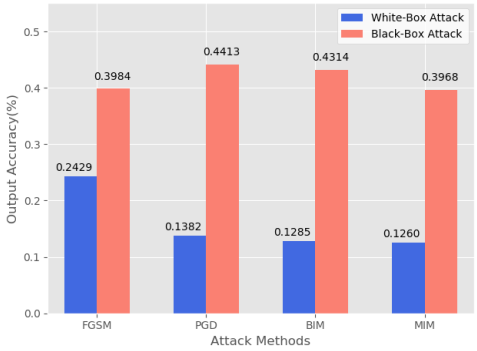


(a)

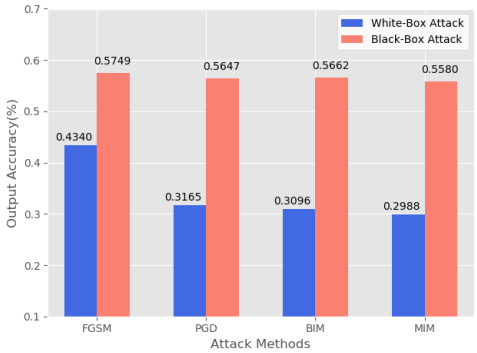


(b)

Fig. 9. Model output accuracy for SNRs at (a) perturbation $\varepsilon = 0.002$ and (b) perturbation $\varepsilon = 0.001$. The results show that the output accuracy gradually increases at a low SNR and slowly fluctuates around a certain value. In white-box attacks, performances of iterative methods are better than that of FGSM. In black-box attacks, these methods only work well when are highly disturbed. Overall, the black-box attack was out-performed by the white-box attack.



(a)



(b)

Fig. 10. The SNR average accuracy of different attack methods at (a) perturbation $\varepsilon = 0.002$ and (b) perturbation $\varepsilon = 0.001$. Results illustrate that in white-box attacks, PGD, BIM, and MIM has an average accuracy of model output 10% lower than FGSM. MIM shows a slight advantage in iterative methods. In black-box attacks, the performance of FGSM is similar to MIM in high perturbation, but higher than that of BIM and PGD. In a low perturbation, FGSM is equivalent to the three iterative methods.

in a modulation signal dataset, the SNR is a factor worth considering. The exploration of the mechanism of adversarial examples under different SNRs and the relationship with noise deserves further attention.

In the white-box attack, for both of the perturbation levels of 0.001 or 0.002, the attack effect of MIM algorithm is slightly better than that of the iterative algorithms BIM and PGD. The three iterative methods exhibit a stronger attack performance than the one-step FGSM, and the attack effect is at least 10% higher than FGSM especially after 0 dB. It reflects the excellent performance of the iterative methods for the white-box attacks.

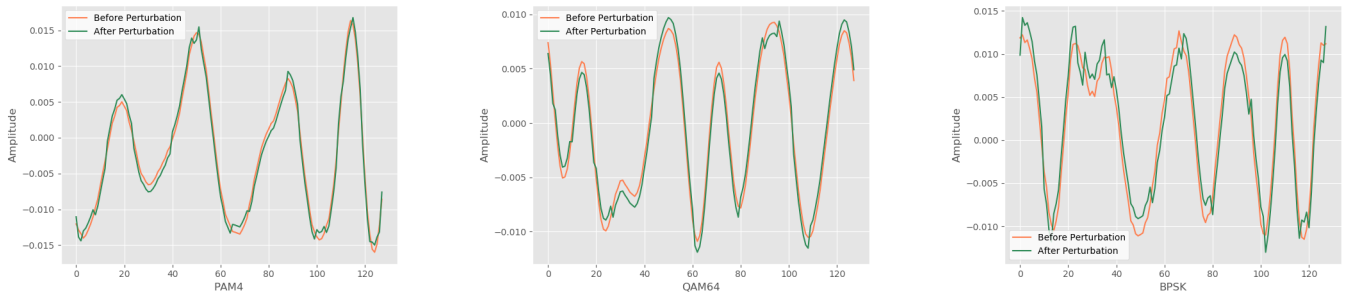
In the black-box attack, at a perturbation level of 0.002, the four methods exhibit similar performance at low SNR. After 0 dB, FGSM and MIM exhibit better performance than BIM and PGD. However, at a perturbation level of 0.001, the three iterative methods exhibit similar performance as FGSM. The 0.001 perturbation is too small for the black-box model to achieve satisfying results, especially at low SNRs.

The aforementioned experiments provided qualitative results for the different attack methods. The subsequent experiments are focused on a quantitative analysis.

C. Quantitative Analysis Under Average SNRs

In order to quantify the output and accurately evaluate the attack effects of different methods and different models, we use weighting and average the accuracy for 11 SNRs of white-box and black-box attacks with two perturbations. As shown in Fig. 10, in the white-box attacks, the average accuracy of output for PGD, BIM, and MIM is generally 10% lower than that of FGSM, where MIM had the best performance amone the three iterative methods, followed by PGD and BIM. For the black-box attack, the performance of FGSM is similar to that of MIM for high perturbations and they outperform BIM and PGD. However, at low perturbation levels, the performance is similar of FGSM to that of the three iterative algorithms. The results are in agreement with the previous experiments and also confirm our hypothesis.

Overall, the black-box attack was no prior knowledge of the model, leading to a reduced attack intensity, thereby inferior performance. Therefore, in order to implement adversarial attacks in the black-box model, more perturbations need to be added.



(a) (b) (c)

Fig. 11. The three subfigures represent FGSM-based (a) white-box attacks with perturbation ε = 0.001, (b) black-box attacks with perturbation ε = 0.001 and (c) black-box attacks with perturbation ε = 0.002, indicating that either black-box attacks or white-box attacks. The application of perturbation results in a significant decrease in the recognition accuracy of the model and a human eye cannot perceive the difference.

VI. Experiments on Feasibility and Robustness

As the perturbation level increases, the signal waveform will consequently change. In order to ensure that the adversarial attacks are both effective and feasible, i.e., imperceptible to human eyes, we will determine the optimal perturbation level while ensuring the invisibility and effectiveness of the attacks. The robustness and vulnerability of different signal types will also be evaluated.

A. Waveforms Comparison

In adversarial attacks, it is important to note whether the adversarial examples added are small enough so they cannot be perceived visually while successfully perturbing the signal samples and causing misclassification of the model. We used 1000 sampling windows where each window had a signal length of 128. Let I be the in-phase component, Q be the quadrature component and f be the carrier frequency. The following modulation carrier formula is used:

$$S(t) = I\cos(2\pi ft) + Q\sin(2\pi ft). \tag{11}$$

At this point, a primitive $S(t)$ signal is generated. By reconstructing and visualizing $S(t)$, we can determine the waveform of the partial modulation signal in the time domain, as shown in Fig. 11. The vertical axis is the amplitude and the horizontal axis is the time variable. The label under each subfigure represents the true category of the signal. Lines of different colors represent the results before and after the perturbation. All experiments were conducted at 10 dB and the perturbation was generated using FGSM in the white-box model.

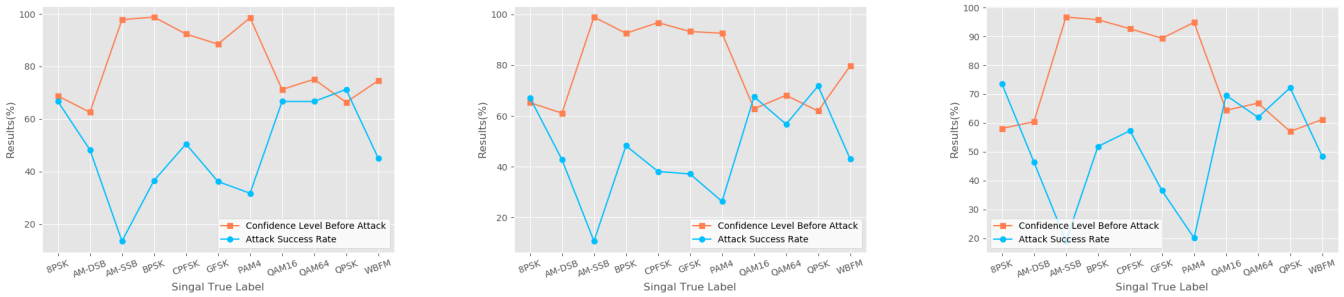
Fig. 11(a) and (b) show the waveform of the white-box attack and the black-box attack generated through FGSM for a perturbation level of 0.001. It is observed that when the model incorrectly classifies the signals into other categories after adding a perturbation of 0.001, the waveform is similar before and after the perturbation. Regardless of a white-box attack or a black-box attack, the perturbation is small enough to prevent a visual detection, indicating that the adversarial attacks are successful. Fig. 11(c) shows the waveforms before and after the perturbation of the FGSM black-box attack with a perturbation level of 0.002. With a higher perturbation, the waveform after the attack has slightly different amplitude than the original waveform and there exist small peaks and distortions similar to the waveform after noise interference despite the difference where the signal after the perturbation is misclassified into other categories. Therefore, as adversarial attacks for modulation recognition, the perturbation of 0.002 is considered sufficient. A higher perturbation level will cause more distortion and changes in the waveform, thereby increasing the possibility of adversarial attacks being detected and defended against. From this we can conclude that the perturbation of 0.001 is optimal for FGSM compared to the perturbation of 0.002. There is no doubt that the waveform distortion will be more evident with much higher perturbation.

If the goal is to keep the adversarial examples aggressive and feasible (visually unperceivable), it is clear that the black-box attack for the algorithm proposed by Papernot [30] has no room for improvement. Therefore, it is worthwhile to use other black-box attack algorithms. In general, the result also verifies our hypothesis that the recognition accuracy of the model decreases dramatically after adding small perturbations, which greatly increases the security risk in modulation recognition and illustrates the vulnerability of deep learning to adversarial attacks.

B. Confidence Measure and Security Evaluation

Adversarial attacks change the characteristics of the input samples to misclassify the signal while the changes are undetectable to human eyes. However, in the previous rounds of experiments, we focused on analyzing the effects of different methods on the models and datasets without considering the effects of the modulation signal dataset itself, i.e., whether the signal samples of certain classes are more robust and less vulnerable than others, and whether they are more difficult to use as adversarial examples. In this section, we describe the experiments to verify the hypothesis.

Under no attack and correct classification, we use the top-1 accuracy in the softmax layer output of the DNN, which is the prediction accuracy of the currently correctly classified signal as the confidence level. In this round of experiments, we measured the average confidence level of the 11 types of signals for all test sets. Next, we launched a white-box attack



(a) (b) (c)

Fig. 12. Results of three subfigures show the relationship between signal confidence level and attack success rate at (a) 2 dB, (b) 10 dB and (c) 18 dB. Obviously, there is a general inverse relationship between signal confidence level and attack success rate. Signals with lower confidence level tend to have a high risk of being attacked.

based on FGSM and evaluated the success rate. The specific formula is as follows:

$$\tau = \frac{\zeta}{\xi}, \tag{12}$$

where τ represents the attack success rate, ς represents the number of samples successfully attacked, indicating the total number of input samples misclassfied into other categories. ξ represents the number of signal samples correctly classified with no attack, indicating the total number of original samples classified correctly without adding perturbations.

In Fig. 12, we show the average confidence level and attack success rate of different categories of signals at 2 dB, 10 dB, and 18 dB, respectively. The confidence level of the DNN is different for the different categories of signals. BPSK, AM-DSB, and QPSK have relatively low confidence levels, whereas AM-SSB, BPSK, PAM4, and CFSK have confidence levels greater than 90%. For signals with a high confidence level, a stronger perturbation needs to be added to reduce the confidence level of the original category and increase the confidence level of the target category to cause misclassification. In addition, as the SNR increases, the trend of the confidence level is less apparent.

In terms of the attack success rate, for the signals with higher confidence levels AM-SSB, PAM4, and GFSK, the success rate never exceeds 40%. For the signals with low confidence levels QAM16, QAM64 and WBFM, the success rate remains 60% or higher. This indicates that AM-SSB and PAM4 are more secure and always maintain a high degree of confidence, whereas BPSK, QAM16, and QPSK are less secure and relatively higher attack success rate.

In general, the confidence level of the signal and the attack success rate show an inverse relationship. Different signals have different levels of robustness; the lower the confidence level, the greater the risk of attacks. Adversarial attacks and defense are a topic of concern and we can assume that, especially in the military field, understanding the different types of modulation signals and their confidence levels will provide an important technical means for the acquisition of enemy information and the selection of optimal interference and suppression methods.

VII. Summary and Future Work

In this study, we evaluated four representative methods to generate adversarial examples of white-box and black-box models to analyze the performance of adversarial attacks on DNN-based modulation recognition. The results showed that in white-box attacks, regardless of the SNR or perturbation, the iterative attack methods significantly outperformed the one-step attack of FGSM with an advantage of 10% on the average attack effect. The MIM attack success rate was slightly higher than that of BIM and PGD in the iterative algorithms. Different methods exhibited different sensitivity to noise, which required us to choose an appropriate attack method based on the actual scenarios.

In black-box attacks, at high SNRs and high perturbation levels, the attack effect was similar for FGSM and MIM, and the attack effect of both models was better than that of BIM and PGD. However, for low levels of perturbation or SNRs, the attack success rate of the different algorithms was low and similar for all the methods, providing unsatisfactory results. Overall, black-box attacks were less successful than white-box attacks.

To ensure that the adversarial attacks are imperceptible to human eyes while maintaining attack effect, we found the perturbation of 0.001 is an optimal value under FGSM, compared to the 0.002 and higher perturbation levels. Different types of modulation signals had different levels of robustness, and the signals with lower confidence levels tend to have a higher risk of being attacked. These results indicated that adversarial attacks were more advanced and inducible than noise interference, and the DNN-based modulation recognition is proven vulnerable to adversarial attacks.

In future studies, we will apply adversarial attacks to a traditional machine learning model and analyze the transferability of adversarial examples. We will also carry out research on targeted attacks to further explore the relationships between confidence levels and security. In addition, further explorations of black-box attacks will be conducted for improved attack success rates, and the adversarial samples will be applied and investigated in additional communication and electromagnetic scenarios.

REFERENCES

[1] A. He et al., "A Survey of artificial intelligence for cognitive radios," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1578-1592, May 2010.

[2] C. Clancy, J. Hecker, E. Stuntebeck and T. O'Shea, "Applications of machine learning to cognitive radio networks," *IEEE Wireless Communications*, vol. 14, no. 4, pp. 47-52, Aug. 2007.

[3] M. D. Wong, A. K. Nandi, "Automatic digital modulation recognition using artificial neural network and genetic algorithm," *Signal Processing*, vol. 84, no. 2, pp. 351-365, Feb. 2004.

[4] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4074-4077, Apr. 2019.

[5] G. J. Mendis, J. Wei and A. Madanayake, "Deep learning-based automated modulation classification for cognitive radio," *IEEE International Conference on Communication Systems (ICCS)*, vol. 1, no. 1, pp. 1-6, Dec. 2016.

[6] L. Jonathan, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 1, no. 1, pp. 3431-3440, Jun. 2015.

[7] Y. Tu, Y. Lin, J. Wang and J.U. Kim, "Semi-supervised learning with generative adversarial networks on digital signal modulation classification," *Computers, Materials and Continua (CMC)*, vol. 55, no.2, pp. 243-254, May 2018.

[8] F. Restuccia and T. Melodia, "Big Data Goes Small: Real-Time Spectrum-Driven Embedded Wireless Networking Through Deep Learning in the RF Loop," *IEEE Conference on Computer Communications (INFOCOM)*, vol. 1, no. 1, pp. 2152-2160, Jun. 2019.

[9] F. Wang et al., "Intelligent Edge-Assisted Crowdcast with Deep Reinforcement Learning for Personalized QoE," *IEEE Conference on Computer Communications (INFOCOM)*, vol. 1, no. 1, pp. 910-918, Jun. 2019.

[10] B. Tang, Y. Tu, Z.Y. Zhang, and Y. Lin, "Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks," *IEEE Access*, vol. 6, no. 9, pp. 15713-15722, Mar. 2018.

[11] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective non-orthogonal multiple access scheme," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp.8840-8450, Jun. 2018.

[12] N. Kato et al., "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 146-153, Dec. 2016.

[13] T. J. OShea, T. Roy and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168-179, Feb. 2018.

[14] D. Castelvecchi, "Can we open the black box of AI?" *Nature News*, vol. 538, no. 7623, pp. 20, Oct. 2016.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *Proceedings of the 2014 International Conference on Learning Representations (ICLR)*, May 2015.

[16] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[17] K. Alexey, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint*, Feb. 2017.

[18] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, "Synthesizing robust adversarial examples," *arXiv preprint*, Jul. 2017.

[19] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213-216, Feb. 2019.

[20] X. Yuan, P. He, and Q. Zhu, "Adversarial Examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 1, no. 1, pp. 120, Jan. 2019.

[21] A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint*, Jun. 2017.

[22] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2016.

[23] Y. Dong, F. Liao, T. Pang, H. Su, X. Hu, J. Li, and J. Zhu, "Boosting adversarial attacks with momentum," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 1, pp. 9185-9193, Apr. 2018.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Proceedings of the 2015 International Conference on Learning Representations (ICLR)*, May 2015.

[25] N. Akhtar and M. Ajmal, "Threat of adversarial attacks On deep learning in computer vision: a survey," *IEEE Access*, vol. 6, no. 1, pp. 14410-14430, Feb. 2018.

[26] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, "ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models," *arXiv preprint*, Nov. 2017.

[27] Z. Zhao, D. Dua, and S. Singh, "Generating natural adversarial examples," *Proceedings of the International Conference on Learning Representations (ICLR)*, Nov. 2017.

[28] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *Proceedings of the International Conference on Learning Representations (ICLR)*, Apr. 2017.

[29] S. Das and P. N. Suganthan, "Differential evolution: a survey of the state-of-the-art," *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 431, Feb. 2011.

[30] F. Tramr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," *arXiv preprint*, May 2017.

[31] J. Su, D. V. Vargas and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 1-13, May 2019.

[32] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," *IEEE European Symposium on Security and Privacy*, vol. 1, no. 1, pp. 372387, Mar. 2016.

[33] N. Papernot et al, "Technical report on the cleverhans v2.1.0 adversarial examples library," *arXiv preprint*, Oct. 2016.

[34] DeepSig, Deepsig dataset: radioml 2016.10a, [online] Available: https://www.deepsig.io/datasets, 2016.

[35] T. OShea and N. West, "Radio machine learning dataset generation with gnu radio," *Proceedings of the 6th GNU Radio Conference*, vol. 1, no. 1, pp. 16, Sept. 2016.