Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5472324, 15 pages
https://doi.org/10.1155/2022/5472324

Research Article

Transferable Adversarial Attacks against Automatic Modulation Classifier in Wireless Communications

Lin Hu, Han Jiang, Wen Li, Hao Han, Yang Yang, Yutao Jiao, Haichao Wang, and Yuhua Xu

College of Communications Engineering, PLA Army Engineering University, Nanjing, China

Correspondence should be addressed to Han Jiang; jh_forward@126.com

Received 30 June 2022; Revised 9 August 2022; Accepted 25 August 2022; Published 27 September 2022

Academic Editor: Lisheng Fan

Deep neural network-based automatic modulation recognition (AMR) technology has become an increasingly important area due to the advantages of self-extraction of features and high identification accuracy. Based on the view of security threats to machine learning classifiers, we investigate the influence of adversarial samples on the AMR model in this paper. The traditional method is based on label gradient attack without taking advantage of the feature-level transferability, resulting in the attack effect that is not perfect. So, we exploit the feature-level transferability property that could be met to fulfill realistic imperceptibility and transfer needs. In this paper, firstly, we proposed an AMR scheme with high recognition accuracy as our attack model. Secondly, we proposed a transferable attack method based on a feature gradient-based, which increases perturbation to clean signal based on features space. Finally, we introduce a new attack strategy, in which we select two original and one adversarial target signal sample as the input of triplet loss to achieve higher attack strength and high transferability. Meanwhile, this paper proposes indicators of signal characteristics to test the effectiveness of our proposed attack method. Based on experimental results, our proposed feature gradient-based adversarial attack method outperforms the currently labeled gradient attack methods regarding attack effectiveness and transferability.

1. Introduction

Deep learning (DL) has been revealed to be successful in conducting diverse wireless communication tasks like signal recognition [1] and spectrum prediction [2]. The key technology of signal detection and demodulation is AMR. It can effectively solve the increasingly crowded and complex electromagnetic space environment, and it is also an important premise for alleviating the spectrum resources shortages. Convolutional neural network (CNN) and long-short term memory (LSTM) are two methods that have achieved good recognition accuracy. However, DL in general has been discovered to be vulnerable to attack by introducing a subtle perturbation that is imperceptible to the human eye [3]. This paper investigates the challenges in signal classification tasks, because signal classification is most widely studied in communication tasks. Actually, the developed methodology is easily transferred to all other tasks. Studying the threat posed by adversarial samples is crucial, not only to enable us to create algorithms that are resistant to interference from malicious samples but also for preventing adversaries from executing signal recognition tasks through such clever intervention. It is important to note that the assault, which is a direct access attack, is started by manipulating the receivers' signal modulation classifier. This kind of attack might not be feasible in the real world because it necessitates the penetration of a target model. Nevertheless, direct access attack methods remain helpful [4] visualizing adversarial perturbations in modulation recognition by reconstructing the waveforms, while compared to other forms of attack, they are more difficult to detect. [5] analyzed direct attack and physical attack that are closer to hardware requirements using traditional FGSM methods. Thus, research into such

direct and digital attacks is of great significance in real world applications, and direct access attacks with AMR may play the role of the foundation for more complex over-the-air attacks [6]. Above all are utilizing the based label gradient attack method. Feature gradient-based adversarial methods are already available in the field of image recognition [7].

To summarize, existing attack methods for signal classification models typically suffer from the disadvantages presented below. Firstly, the transferability of adversarial examples is imperfect in attacking the black-box model, particularly in the presence of targeted attacks. That is because the current methods mostly adopt single-layer features rather than attacking with taking the use of features space. In fact, the middle layer of the CNN representation is transferable. Normally, CNN low-level features have a lot of granular information, while its high-level features have a lot of global semantic information. Secondly, the adversarial sample is difficult to categorize into the stated target class since the standard label gradient-based assaults only limit the distance between the adversarial sample and the target class. Evaluating only the success rate of an attack does not correspond to the merely evaluation measure of the effectiveness of an attack in the field of signal recognition. In the real communication environment, we know relatively little a priori information, and it is necessary to maintain a certain degree of imperceptible to achieve the effect of the attack.

To address the abovementioned issues, we propose a feature gradient-based attack method, which relies on two basic observations. The first is a deep learning classifier model that predicts mainly on the basis of the signal samples information and differentiation regions. However, the presence of such regions weakens the models. The second conclusion is that perturbation in the middle layer features of well-trained networks is transferable [8–10]. Research [8] concluded that feature representations are universal in neural networks, and that feature representation can be transferred for learning by transferring to the target network. Furthermore, features from various levels exhibit diverse features. [9] improves the evidence, proving that adversarial examples can be produced through operating image representations under deep neural networks. The current work focuses on adding the potential representation space of adversarial ingestion to those regions of the signal sample that are informative and distinguishable. This contributions are as follows:

(i) To provide more transferable and efficient adversarial examples, this paper proposed a transferable attentive method concentrating on the informative and discriminative feature regions, adding perturbation at the feature level will be more adaptable to realistic scenarios. The proposed attack methods are more effective when compared the previous methods in the modulation recognition scenario

(ii) We have conducted experiments in all metrics of our method with a new system of indicators that better suit the signal characteristics. Our method surpasses that of the traditional label gradient method in most indexes

The remaining of this paper is arranged as follows: Section 2 presents the related work of DL in modulation signal classification and the threat of adversarial examples; Section 3 of this paper introduces the methodology of adversarial examples based on feature gradient; Section 4 develops a series of experiments from the perspectives of white-box attack and black-box attack, explores the experimental results, and verifies the effectiveness. Finally, this paper is summarized and looks forward to the future.

2. Related Work

2.1. AMR Model. The concept of AMR was first proposed in [11], as one of the pattern recognition research, and it has filled everyone's vision. Machine learning (ML) methods have been extensively used based on the constant advancement of technology. DL has been developed recently into a popular technology for breaking through the performance bottleneck of pattern recognition tasks, and this technology has also been introduced into the field of AMR. Based on their perspective of development in the field of AMR, recognition algorithms are classified into two types: classical modulation recognition methods and deep learning-based modulation recognition methods [12]. The classical methods can be divided into recognition methods based on likelihood function [13] and recognition methods based on feature extraction [14].

With increasingly complex and diverse communication systems, wireless signal data is more complex and diverse than ever, with stronger randomness and heterogeneity. Traditional modulation recognition requires manual extraction of features and relies on prior information. The workload is heavy, and the recognition accuracy is low. Therefore, the industry applies DNN to the field of signal recognition. The DNN model requires a large amount of training data, and the massive features of wireless communication signals were provided. In comparison to traditional methods, DNN can automatically extract modulated signal features, eliminating the errors that may be introduced by the manual selection of features and the dependence on expert knowledge in the identification process. The most important thing is that AMR can achieve more accurate results. The present study investigates the threats specific to the signal classification stage and is thus related to adversarial machine learning [15] which has witnessed an increase in activity in the context of CV [16]. Recently, the search for DL signal recognition has mainly been based on two perspectives: signal array and imaged-based. The texture map of the in-phase and quadrature (IQ) waveform of the communication signal is applied as the input of the DL model in the signal array recognition method. According to Rajendran et al. [17], through the transformation of IQ data into AP (amplitude/phase) information and adoption of a simple LSTM model, a perform accuracy was attainable. The model enabled the extraction of temporal signal traits from the training data,

where it is unnecessary to extract the expert traits manually. Attention mechanism (AM), which was originally adopted for machine translation [18] as a crucial concept in the DL domain, is currently applied extensively in areas like speech recognition, NLP (natural language processing), statistical learning, and computers. Chen et al. [19] put forward a new attention cooperative framework in which the input feature maps were made mutually dependent by incorporating the classifiers with a self-AM and a Squeeze-and-Excitation block [20]. The validity of AM is proved in AMR. The present study is aimed at developing an AMR model based on DL. The AP information is initially extracted from the IQ data, and then the classification outcomes are derived using an AM-based monolayer LSTM model. Our developed scheme is compared to the existing CNN-AP, LSTM-AP, and CLDNN-AP schemes. The accuracy of classification can be less influenced by the signal frequency offset when using CNN [21]. LSTM is appropriate for obtaining time-series signal traits [22]. CLDNN (convolution, LSTM, deep neural network), which integrates the benefits of DNN, CNN, and LSTM, is proven to be highly competent in classifying the modulation modes [23, 24].

2.2. Adversarial Evasion Attack. The first step in guaranteeing system security is to identify the systems challenges. This paper firstly characterizes the possible source of challenges, envisions new challenges, and describes the restrictions in adversarial attacks under the background of wireless communications. The uninterpretable DNN exposes them to a variety of security risks. Szegedy et al. discovered that by adding some carefully crafted tiny human-imperceptible perturbation to the input samples, the accuracy of DNN classifiers can be significantly reduced, and such added perturbed samples are called adversarial example [25]. Adversarial attacks can be categorized into two categories based on whether or not the adversarial sample has a target: targeted attacks and untargeted attacks. Targeted attacks are those where the adversarial sample must misclassify the input sample into a specific class to deceive the model. For example, in modulated signal classification, if the attacker specifies the target class as ASK, 8PSK, QPSK, or any other class of signals, it will be incorrectly classified as ASK after being attacked, while targetless attacks are the inverse of targeted attacks, where no specific attack signal class is required, i.e., the target can be any type of signal other than its signal.

Untargeted attacks can be classified into white-box attacks, black-box attacks, and gray-box attacks based on the knowledge level in the target model. In a white-box attack, the adversary is aware of the training data, architecture, algorithms, and optimization techniques, which enables it to fully access the trained model. A black-box attack neither knows nor has accesses to the training data and training model, making it a more realistic and practical scenario that also increases the difficulty of the attack. A gray-box attack is one in which only a limited amount of information is known ahead of time.

The majority of the adversarial sample research are currently focused on image recognition. Goodfellow et al. presented fast gradient sign method (FGSM) to attack deep network models, the core idea being to obtain the adversarial sample by computing the gradient of the loss function relative to the input sample itself [26]. Kurakin et al. put forward the iterative FGSM (basic iterative fast gradient sign method, BIM), which uses multiple iterations to generate an adversarial sample [27]. Dong et al. presented momentum into the gradient calculation process in the iterative attack and proposes the momentum iterative method (MIM) method to enhance the stability of the model at each iteration and the generalization of the adversarial samples [28]. Moosavi-dezfooli et al. proposed an algorithm called Deep Fool, which replaces the deep classification model with a linear model for attack [29]. Lin et al. introduced the Nesterov accelerated gradient into the iterative attack process and proposed PGD to increase the adversarial samples migrability [30]. Kurakin et al. presented an approach to performing adversarial training on the model to explore the impact of the adversarial samples on the model robustness [31]. Carlini and Wagner proposed three methods to generate perturbations, using three different metrics (L_1, L_2, L_∞) to avoid the robustness of the model [32]. The real-world artifacts can also be used to trick the classification model [33].

Little work has been done to apply adversarial example attacks to AMR, and Lin et al. applied the traditional adversarial method based on label computation gradient to modulated signal recognition and verified that AMR is vulnerable to adversarial sample attack [34], However, the above methods still use the alternative model approach when performing black-box attacks and do not fully utilize the features of modulated signal data samples. Moreover, the recognition accuracy of the target model itself chosen for modulated signal recognition is not high, only about 70%. Because of the disadvantage of the previous work, we first propose a target recognition model with high accuracy, which could attain a top accuracy about 91%. The adversarial example was then generated using a feature gradient. Finally, we use a new strategy in which we select two original samples and one target sample as triplet loss input.

3. Transferable Attack Methodology

This study proposes a new black-box targeted attack method for signal classification, named transferable adversarial attack, which can deceive white-box models. The current section firstly depicts the methodology of the fundamental idea of generating adversarial examples. The algorithm flow is then given. Finally, evaluate the feasibility of the proposed algorithm.

3.1. Backgrounds. The most of raw IQ signal classifiers attempt to get a signal snapshot x and output the most confident result class y. In most situations, x denotes a two-dimensional matrix (IQ, number of samples) that reflects a single channel of complicated data with little preprocessing. It employs DNN to learn a mapping from data by solving

problems, particularly in the domain of communications.

$$\underset{\theta}{\operatorname{argmin}} L(f(\theta, x), y)), \tag{1}$$

where x and y denote the training and true labels, respectively, and f denotes the network architecture used. To learn the network variable θ, a loss function is usually used in conjunction with an optimizer in DNN training. We assume that the data set is constant without data augmentation throughout model training, and that it is sampled from a distribution that is similar to that observed later in the communication system's operation. FGSM uses untargeted adversarial examples to build untargeted adversarial examples.

$$x^* = x + \varepsilon \cdot \operatorname{sign}\left(\nabla_x J(x, y, w)\right), \tag{2}$$

where y is the real input label, and ∇_x indicates the gradient of the loss function in terms of the original input x. The proposed approach in a single step can create adversarial examples x^* restricted by a distance ε, in the feature space.

The average energy per symbol (E_s) of a transmission can be calculated based on

$$[E_s] = \frac{\text{sps}}{N} \sum_{i=0}^{N} |s_i|^2, \tag{3}$$

where sps denotes samples per symbol, N is the total number of samples, and s_i denotes a particular sample in time. Without losing generality, the present study assumes the average energy per symbol of the modulated signal, $E_s = 1$. As a result, the underlying transmissions power ratio to the perturbation signal (E_j) can be derived as

$$\frac{E_s}{E_j} = \frac{1}{E_j} = 10^{-E_j(dB)/10}, \tag{4}$$

Since the input of sign (∇_x) in (2) is complicated, the output also remains complicated and is thus a vector with values $[\pm 1, \pm j]$. As a result, the magnitude of each the perturbation sample is computed as

$$|\operatorname{sign}(\nabla_x)| = |\operatorname{sign}(z)| = \sqrt{(\pm 1)^2 + (\pm 1)^2} = \sqrt{2}, \tag{5}$$

Thereby, the energy per symbol of sign (∇_x) can be calculated by plugging (5) into (6), leading to

$$E_{\operatorname{sign}(\nabla_x)} = \frac{\text{sps}}{N} \sum_{i=0}^{N} |\operatorname{sign}(\nabla_x)|^2 = 2 \times \text{sps}, \tag{6}$$

Since sps is fixed through transmission, a closed form scaling factor, ε, is deduced to obtain the desired energy ratio (E_s/E_j) by using

$$\varepsilon = \sqrt{\frac{E_j/E_s}{E_{\operatorname{sign}}(\nabla_x)}} = \sqrt{10\frac{E_j(\text{dB})}{2 \times \text{sps}}}, \tag{7}$$

Plugging ε into (2) allows the creation of adversarial examples constrained by (E_s/E_j) and can be simply expressed as

$$x^* = x + \sqrt{\frac{10^{E_j(dB)/10}}{2 \times \text{sps}}} \times \operatorname{sign}\left(\nabla_x L(f(\theta, x), y)\right). \tag{8}$$

Above constraining the power ratio in this way can be beneficial for assessing system design trade-offs.

$$\begin{aligned} &\min \|x^* - x\|_p, \\ &s.t, l(x) \neq l(x^*), \\ &x^* - x \in \varepsilon, \end{aligned} \tag{9}$$

where $\|\cdot\|_p$ suggests the L_p norm. Furthermore, the L_p of δ is defined as

$$\|\delta\|_p = \left(\sum_{i=1}^{n} \|\delta\|_p\right)^{1/p}, \tag{10}$$

where L_0, L_2, L_∞ are the three most common metrics. The L_0 is a quantitative metric for the pixel variations in an image, whereas it quantifies the nonzero vectors of perturbation in a signal. The L_2 metric quantifies the Euclidean distance between adversarial and original examples as an Euclidean norm; the L_∞ is responsible for the maximum alteration constraint of all signal vectors/pixels in the adversarial examples. The power budget of a transmitter is usually constant, and in this research, an adversarial strategy of ML that is unaware of underlying signal is considered. Hence, the power applied to the jamming signal is inapplicable to the underlying transmission.

3.2. Triplet Loss for Adversarial Attack. The traditional label gradient attack method calculates the gradient using the sample label y, incorporating the initial clean sample x and the consistent label y into the target model loss function. The attack direction can be obtained by computing the gradient and sign function and then multiplied by the perturbation size to realize the adversarial perturbation; finally, then combine with the original clean sample to form an adversarial example. Currently, the main attack methods based on label gradient are FGSM, BIM, and MIM. Obviously, failure to take advantage of fast gradient varies at the feature space and transferability.

The proposed method is a momentum iterative FGSM at the feature space level. Thus, we suggest using triple loss, which can minimizes between an anchor and a positive, both of which have the same identity, and maximize the distance between the anchor and a negative of a

different identity. As a result, the information region and the discriminative region in the sample may be perturbed through optimization of the triplet loss on the feature space. Because of extracting features, we need to truncate the target model from the L layer to obtain the truncated model, to ensure that the selected feature space is abundant enough, and this paper uniformly selects the activation layer as our target layer. Then, put x and y into f_L to obtain the original signal sample feature $f_L(x)$. The loss here uses a triplet pair $(f_L(x_i^a), f_L(x_i^p), f_L(x_i^n))$, the anchor, positive, and negative terms of the triplet loss, respectively, and signal samples from the same class should be near together in the embedding space, forming several well-separated clusters. As a result, triplet loss ensures that the attack process not only makes the original sample close to positive sample (target sample) and away from negative sample (untarget sample).

Triplet loss can be expressed as

$$L_{\text{tri}} = \sum_{i}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+, \tag{11}$$

where $\alpha \in R^+$ denotes a margin between negative and positive pairs. The triple loss is adopted for adversarial example crafting in addition to strengthen the adversarial robustness [35], and the triple loss is also exploited for the adversarial example crafting purpose. The present work is the first attempt to use the triplet loss to craft the adversarial examples, where a source sample feature is drawn closer to the target class while being propelled away from the source class. In contrast to the conventional triplet loss, the clean signal sample acts as the anchor example, while the other clean and target class samples act as the negative and positive examples, respectively. With our attack, the anchor and positive examples are reasonably separated, while the distance between the anchor and negative examples is increased. The adversarial examples are easily misclassified into the target class by our triple loss-based algorithm. Furthermore, unlike the standard triplet loss in which every element is a clean sample, our triplet loss includes an adversarial example term, which can be found in Figure 1.

3.3. Basic Ideas. This research proposes two methods based on the aforementioned motivation. This algorithm can simulate the traditional BIM and MIM attack methods. To destroy the potential representation space, we propose to optimize triplet state loss rather than crossentropy loss. Furthermore, this study proposes two methods, with more intuitive variants explained in Algorithms 1 and 2.

When AMR is attacked, it is expected to add an imperceptible slight perturbation in the clean original sample, resulting in an error recognition rate. Suppose the original signal sample is x, the classification result is y, and the perturbation is small enough to meet $\|\eta\|_\infty \le \varepsilon$. So, FGSM was described below.

$$\begin{cases} \eta = \varepsilon \cdot \text{sign}\ (\nabla_x J(x, y)), \\ x^* = x + \eta, \end{cases} \tag{12}$$

where J is the target models loss function, and $\nabla_x J(x, y)$ refers to the derivative of the loss function over sample x. Because FGSM refers to a one-step attack, it is impossible to update the adversarial example by querying the model parameters in multiple times. The basic iterative method (BIM) denotes an extended FGSM in which adversarial examples are generated in various iterations. Every iteration has a small step size, and each step should be within the perturbation neighborhood of the original input.

$$\begin{cases} x_0 = x, \\ x_{n+1} = \text{Clip}_{x,\varepsilon}\{x_n + \varepsilon\ \text{sign}\ (\nabla_x J(x_n, y))\}. \end{cases} \tag{13}$$

$\text{Clip}_{x,\varepsilon}\{\}$ means to limit it to the scope $[x - \varepsilon, x + \varepsilon]$.

MIM is a reduction algorithm iteration technology that accelerates the speed under the gradient through accumulating the velocity vector in the gradient direction of the loss function. It can be denoted as follows:

$$\begin{cases} x_0^* = x, g_0 = 0, \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J(x_n^*, y)}{\left\| \nabla_{x_n^*} J(x_n^*, y) \right\|_1}, \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}. \end{cases} \tag{14}$$

g_{n+1} represents the cumulative gradient generated by the previous $n + 1$ iteration, and μ is the attenuation factor.

MIM, same as BIM, incorporates an acceleration gradient into the iterative attack process and improves the migration performance of the adversarial examples, which can be denoted as

$$\begin{cases} x_0^* = x, g_0 = 0, \\ x_n^{\text{nes}} = x_n^* + \beta \cdot \mu \cdot g_n, \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J(x_n^{\text{nes}}, y)}{\left\| \nabla_{x_n^*} J(x_n^{\text{nes}}, y) \right\|_1}, \\ x_{n+1}^* = \text{Clip}_{x,\varepsilon}\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}. \end{cases} \tag{15}$$

Among them, x_n^{nes} is a Nesterov item, which jointly participates in the calculation of gradient.

3.4. Description of Attack Method. In order to perform an attack on the feature space of the AMR model, it is first necessary to find a suitable feature space. Meanwhile, to ensure that the selected feature space is sufficiently informative, the truncation layer of the target model is chosen as the final fully connected layer of the model. In order to ensure that the selected feature space is rich enough, the truncation layer of the target model is chosen as the

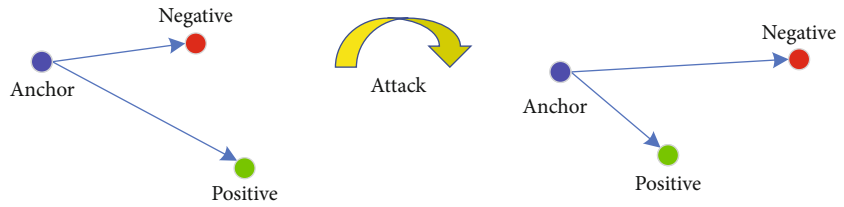


Figure 1: Schematic diagram of triple loss. The attack process not only makes the original sample close to positive sample (target sample), and away from negative sample (untarget sample), by pushing and pulling, but also to achieve a better attack effect.

Input: A classifier truncated model f_L; original signal sample to be attacked x^α; target signal sample x^p; clean signal sample x^n; Loss Function J_{AL}.
Parameter: perturbation size ε=0.001, number of iterations T.
Output:the adversarial sample x^* that satisfy
$\|x^* - x_T\|_2 \le \varepsilon$.
$\alpha = \varepsilon/N$.
put x^α into f_L,obtain feature $f_L(x^\alpha)$;
put x^p into f_L,obtain feature $f_L(x^p)$;
put x^n into f_L,obtain feature $f_L(x^n)$;
  for t=0 to T-1 **do**
put x_n^* into f_L,obtain feature $f_L(x_n^*)$
Obtain the gradient $\nabla_{x^\alpha} J_{AL}$,
where $J_{AL} = L_{tri}(f_L(x_i^a), f_L(x_i^p), f_L(x_i^n))$;
calculate the accumulated gradient,renew the g_{n+1}:
  $g_{n+1} = \mu \cdot g_n + (\nabla_{x_n^*} J_{AL} / \|\nabla_{x_n^*} J_{AL}\|_1)$
update the x_{n+1}^* with gradient method
  $x_{n+1}^* = Clipx, \varepsilon\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}$
  end for
return$x^* = x_N^*$

Algorithm 1: AL-BIM.

Input: A classifier truncated model f_L; original signal sample to be attacked x^α; target signal sample x^p; clean signal sample x^n; Loss Function J_{AL}.
Parameter:perturbation size ε=0.001, number of iterations T.
Output:the adversarial example x^* that satisfy
$\|x^* - x_T\|_2 \le \varepsilon$;
$\alpha = \varepsilon/N$
put x_T into f_L,obtain feature $f_L(x_T)$;
put x^p into f_L,obtain feature $f_L(x^p)$;
put x^n into f_L,obtain feature $f_L(x^n)$;
put x_n^* into f_L,obtain feature $f_L(x_n^*)$;
  for t =0 to T-1 **do**
calculate $x_n^{nes} = x_n^* + \alpha \cdot \mu \cdot g_n$
put x_n^{nes} into f_L,obtain feature $f_L(x_n^{nes})$
Obtain the gradient $\nabla_{x^\alpha} J_{AL}$,
where $J_{AL} = L_{tri}(f_L(x_i^a), f_L(x_i^p), f_L(x_i^n))$;
  $g_{n+1} = \mu \cdot g_n + (\nabla_{x_n^{nes}} J_{AL} / \|\nabla_{x_n^{nes}} J_{AL}\|_1)$
update the x_{n+1}^* with gradient method
  $x_{n+1}^* = Clipx, \varepsilon\{x_n^* + \beta \cdot \text{sign}\ (g_{n+1})\}$
  end for
return$x^* = x_N^*$

Algorithm 2: AL-MIM.

activation layer before the final fully connected layer. Therefore, For activation L layer in BIM (AL-BIM), the attack process is as follows:

$$J_{\mathrm{AL}}(x_S, x_T, x_{\mathrm{adv}}) = L_{\mathrm{tri}},$$
$$L_{\mathrm{tri}} = \sum_{i}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+, \tag{16}$$

$$\begin{cases} x_0^* = x_S, g_0 = 0, \\ g_{n+1} = \mu \cdot g_n + \dfrac{\nabla_{x_n^*} J_{\mathrm{AL}}(x_T, x_n^*, x_{\mathrm{adv}})}{\left\| \nabla_{x_n^*} J_{\mathrm{AL}}(x_T, x_n^*, x_{\mathrm{adv}}) \right\|_1}, \\ x_{n+1}^* = \mathrm{Clip}x, \varepsilon\{x_n^* + \beta \cdot \mathrm{sign}\ (g_{n+1})\}, \end{cases} \tag{17}$$

where $\|\cdot\|_2$ is the L_2 norm, a similarity measure representing adversarial sample features and original sample features. So, the workflow of the AL-BIM method could be shown in Algorithm 1.

Activation L layer in MIM (AL-MIM) is similar to the AL-BIM algorithm; before calculating the gradient, a Nesterov x_n^{nes} needs to be calculated, and its workflow is shown in Algorithm 2.

3.5. Feasibility Analysis of Attack Methods

(1) The amount of information in the spectrum signal sample is small compared with the high-dimensional data of the image. If a classification model with an AM is used to extract the effective features of the attack object, it may improve the attack precision and intensity. The maximum misclassification effect is achieved with minimum perturbation of intensity. At the same time, after training different AMR models, the feature of the samples is transferable

(2) Based on the above considerations, this paper uses signal samples to extract effective features in the model, calculates the gradient from the feature level, and then attacks the proposed AMR model, which may achieve a higher misclassification rate with less fewer disturbances. Furthermore, from the feature level, it may better reflect the migration of the attack effect. Recently, the research material of the adversarial attack method has not been seen, which is based on the AM to extract effective features, and then adds disturbances by gradient calculation from the feature level

(3) Different from the traditional label gradient attack method, we must truncate the target model from the L layer because of the extracting features to obtain the truncated model and put x, x_t, and x_{adv} into f_L to obtain the original signal sample features $f_L(x)$, $f_L(x_t)$, and $f_L(x_{\mathrm{adv}})$, and the gradient of the feature is calculated

(4) Following the perturbation imposition, the modulation signal is sent into the target CNN for identification and classification. Given the high attack susceptibility of CNN, the classifier can be deceived by crafty perturbations, resulting in highly confident misclassifications. Section 4 will investigate how different parameters like perturbation levels and SNRs influence the CNN attacks and validate the attack feasibility and effectiveness by using the waveform and accuracy assessment methodologies. Figure 2 displays the block diagram for the adversarial attack assessment in modulation identification

Based on the advantages of the above feature level and the ternary loss function to reduce the Euclidean distance, we can propose the above algorithm with better transferable and concealment.

4. Experiment and Result Analysis

To test the effectiveness of adversarial ML on raw IQ-based AMR, the models we proposed are applying the model trained on Radio-ML2016.10a.

4.1. Experimental Data Set. Radio-ML2016.10a is a publicly available modulated signal data set from Bradley University, which is a data set used during the experiments in this research that employs GNU Radio to synthesize I/Q signal samples containing 11 modulation types, with signal-to-noise ratios ranging from -20 dB to 18 dB, uniformly distributed at 2 dB intervals. There are 128 complex floating point time samples in each signal. The data set is $220{,}000 \times 128 \times 2$ in size. The I and Q paths hold the real and imaginary parts of the 128 signal points, respectively.

In this research, we selected the signal in the data set with a high SNR greater than or equal to 10 dB. Furthermore, the number of samples in the training set is 35200, and test set data were classified by the proposed model to obtain 91.01%.

4.2. AMR Model. We should value the model we want to attack. If an AMR model recognition effect is poor, the effect of the attack may not be properly reflected considering that the spectrum signal and image have different characteristics and parameters. Aiming white-box attack, in the study, we develop a LSTM-AP model with an AM that performs perfect in modulation recognition. Aiming black-attack, to confirm the transfer for the attack, we present two AMR models, one is LSTM-AP, and another is CLDNN-AP; the specific model parameters will not be described. Figure 3 depicts the LSTM-AP model with attention mechanism for AMR. The signal embedding module is covered in the first section. Besides, the data format in RML2016.10a is 2×128, and it can be used as an input to LSTM, as IQ data input to LSTM. A learnable matrix is used in the fully integrated process of signal embedding to multiply data. Signal embedding is adopted because of the quite universal features of low-dimensional data, which necessitates the strengthening of the model's robust field through the continuous rise of the data dimensions. Variations in data dimensionality are

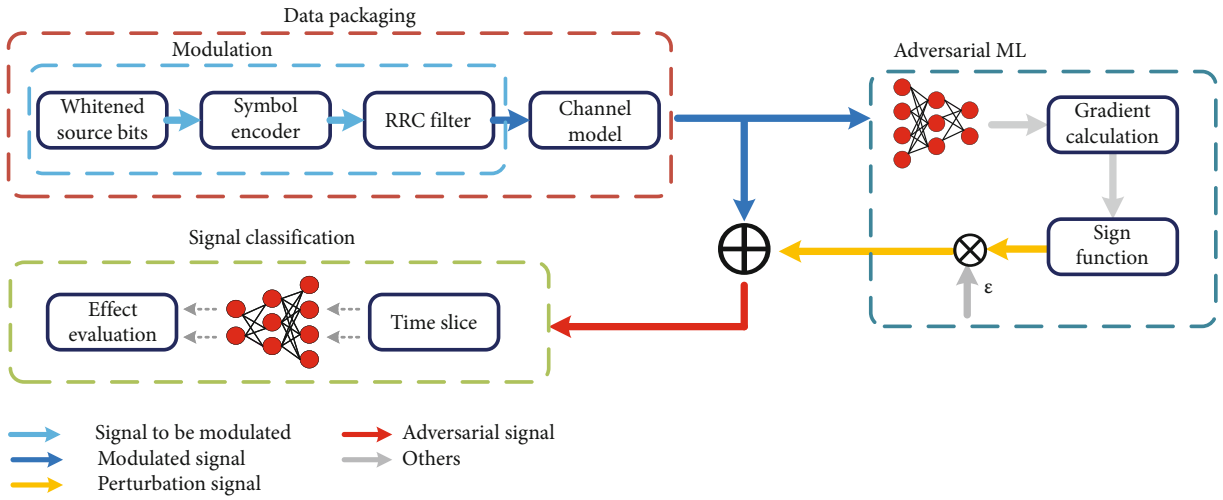


Figure 2: In this paper, the flow chart of modulation identification against attacks, the modulated signal is first data encapsulated, then the network gradient is obtained through the target network, and perturbation is added to the gradient to form an adversarial sample, which leads to the model recognition error.

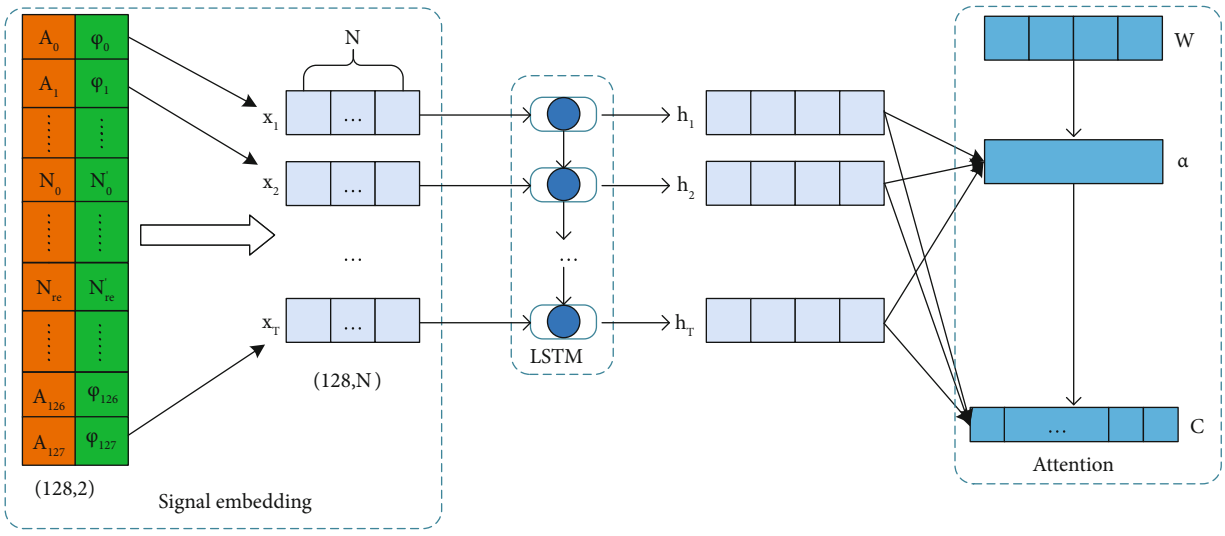


Figure 3: The proposed LSTM-AP model with attention mechanism for AMR.

derived via persistent model learning, and the best dimension for extracting features is sought ultimately. As a result of the embedding, the modulation information included by the input matrix will be larger and more accurate. The second part is the monolayer LSTM, which is excellent in acquiring temporal features like the time-series data of modulated signals and the information about phase and amplitude that varies by the mode of modulation. The final component is the AM module. The amplitude and phase information of a partial piece of data can be used by the AM to focus on the mode of modulation of a modulated signal sequence. Assume our input consists of T points of sequential signal data.

4.3. Evaluation Indicators. To assess the efficiency and transferability of the attack method in the current work, the following evaluation metrics are defined for the generated adversarial examples such as imperceptibility and signal properties.

(1) Attack success rate (ASR):

$$\mathrm{ASR} = \frac{\mathrm{ACC}_{\mathrm{ori}} - \mathrm{ACC}_{\mathrm{adv}}}{\mathrm{ACC}_{\mathrm{ori}}}. \tag{18}$$

ASR calculates the attackers percentage of misclassification, $\mathrm{ACC}_{\mathrm{ori}}$ is the classification accuracy of the original signal sample, and $\mathrm{ACC}_{\mathrm{adv}}$ is the classification accuracy obtained by the adversarial sample using the same classification model. The attack success rate can show an attack methods potential to cause misclassification.

Imperceptibility is as follows: L_0 norm and L_2 norm.

$$L_0 = \frac{C_c}{N}. \tag{19}$$

L_0 can be calculated as the proportion of the total number of points that a signal sample changes after an attack. C_c is the number of modified points in a sample (128×2 points), N refers to the number of data points in a signal sample, and the N value of the data set used in this paper is 256 (128×2).

$$L_2 = \sqrt{\sum_{i=1}^{N} |V_{\text{oi}} - V_{\text{ai}}|^2}. \tag{20}$$

L_2 calculates the numerical Euclidean distance between an original signal sample and an adversarial sample. V_{oi} indicates the value of the ith data point of the original sample, V_{ai} represents the value of the ith data point of the adversarial sample, and N is 256 (128×2).

(2) Signal characteristics: since the unique characteristics of the signal, we verify three indicators: ACR (amplitude change rate), APD (average phase difference), and PSR (perturbation signal rate)

ACR (amplitude change rate) is as follows:

$$A = \sqrt{I^2 + Q^2}, \tag{21}$$

$$\text{ACR} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_{\text{oi}} - A_{\text{ai}}}{A_{\text{oi}}}\right|. \tag{22}$$

ACR calculates the amplitude change rate of the signal before and after the attack. In the procedure of signal processing, different from the independent pixels in the image, there is a one-to-one correspondence between the I/Q channels in the signal data set, which are the sampling values of the real part and the imaginary part of the complex signal, if the reference is the same as the image, the calculation method ignores the correlation of the I/Q two-way. A represents the signals effective amplitude, while I and Q are the coefficients of the real and imaginary parts of the signal, respectively. A_{oi} is the effective amplitude of the ith sampling point of the original signal, A_{ai} denotes the effective amplitude of the ith sampling point of the signal after the attack, n represents the number of sampling points in a signal sample, and the value of n in the data set used in this study reaches 128. Different from each independent pixel in the image, for the 128×2 sample in the signal, it is more accurate to describe a signal sampling point by the matching I channel and Q channel than to regard it as 256 independent points.

APD (average phase difference) is as follows:

$$\text{APD} = \frac{1}{n}\sum_{i=1}^{n}\left|\arctan\frac{Q_{\text{oi}}}{I_{\text{oi}}} - \arctan\frac{Q_{\text{ai}}}{I_{\text{ai}}}\right|. \tag{23}$$

APD calculates the average phase difference at each sample point in a signal sample. The phase is an important factor to evaluate the signal attack. As an important measure to describe the change of the signal waveform, the delay of the phase can completely change a signal, thus making it impossible to extract the real message. I_{oi} is the real part coefficient of the i-th sample point of the original signal, and Q_{oi} indicates the imaginary part coefficient of the ith sample point of the original signal. I_{ai} indicates the real coefficient of the ith sampling point of the signal after the attack, and Q_{ai} refers to the ith sampling point of the original signal imaginary.

PSR (perturbation signal rate) is as follows:

$$P = \frac{\sum_{i=1}^{n} A_i^2}{n}, \tag{24}$$

$$\text{PSR} = \frac{P_P}{P_s}. \tag{25}$$

PSR analyses the power ratio of the disturbance craft adversarial sample to the signal P, where P is the signal power and A_i denotes the effective amplitude of the ith sampling point of the signal.

(3) TR (transition rate): in order to evaluate the transition of the attack, it is assumed that all signal samples are correct classification of white model f_w,and black model f_w. The original data set is $D_{\text{orig}} = \{(x^{(1)}, y^{(1)}_{\text{true}}), \cdots, (x^{(N)}, y^{(N)}_{\text{true}})\}$, and each attack method would generate an adversarial data set $D_{\text{adv}} = \{(x^1_{\text{adv}}, y^1_{\text{target}}, y^1_{\text{true}}), \cdots, (x^N_{\text{adv}}, y^N_{\text{target}}, y^N_{\text{true}})\}$. The data x_{adv} and y_{target} are obtained by the target attack performed by the original data set on the white-box model f_b. The mobility of adversarial examples refers to the number of samples that could deceive both the white-box model f_w and the black-box model f_b in the adversarial data set D_{adv} and the number of successfully deceived white box model f_w. Define the data set of successful deceiving white box model as $D_{f_w} \subseteq D_{\text{adv}}$. Then, mobility can be defined as follows:

$$\frac{1}{\left|D_{f_b}\right|}\sum_{(x_{\text{adv}}, y_{\text{true}}) \in D_{f_b}} 1[(f_b(x_{\text{adv}})) \neq y_{\text{true}}]. \tag{26}$$

This evaluation method intuitively shows the possibility that the adversarial examples generated in the white-box attack may potentially play a role in the black-box model.

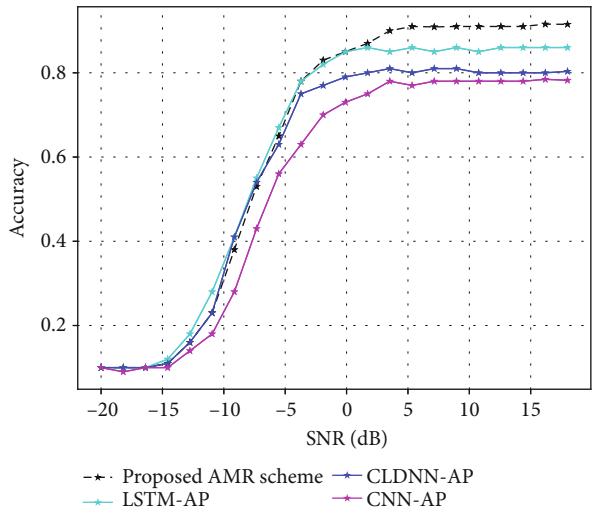


Figure 4: Recognition accuracy of different models.

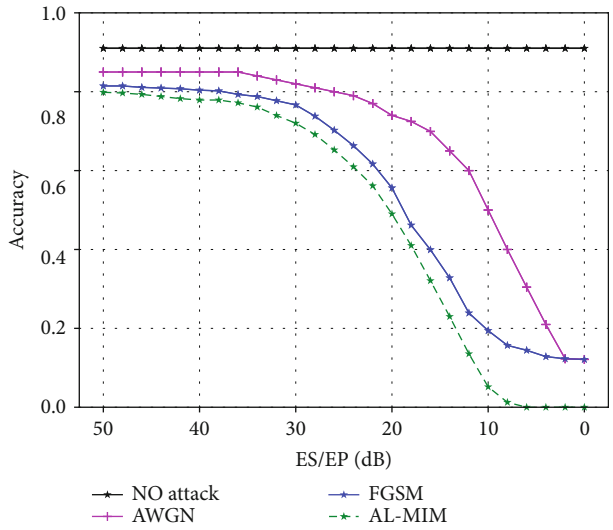


Figure 5: Changes in the recognition accuracy of different types of perturbation.

4.4. Experimental Result

4.4.1. Training and Results of Target Model. The number of iterations, learning rate, and other major parameters is consistent during the model training stage, to manage the efficiency and consistency of training. The number of iterations is set to 500, the learning rate is 0.001, and an automatic update mechanism is set: if the loss value of the test set does not drop for five consecutive times, the learning rate is decreased by halved. Additionally, considering that the modulated signal data set is composed of 20 SNRs, the data for each SNR has a different number. The characteristics suggest that the model be trained by combining all SNR data as a data set for training and then verifying its recognition accuracy on each SNR independently during verification.

Figure 4 compares the recognition accuracy of the proposed model to other three schemes: CNN-AP, CLDNN-AP, and LSTM-AP. CNN-AP has relatively low classification accuracy, demonstrating that CNN performs poorly when extracting features from time series data.. The efficiency is insignificant even when the CNN training data used is the IQ signal information about phase and amplitude, with mere maximum accuracy of 83.4% for the CNN-AP. Meanwhile, where the input of CLDNN-AP is the information about phase and amplitude, 85.2% accuracy of classification is attained. As displayed in Figure 4, when the LSTM input is the IQ data, the accuracy of classification is low. The reason is that the displayed phase and amplitude traits vary among modulation schemes, which are not reflected by the IQ data. The accuracy of classification is 87.13% at a SNR of 0 dB, and the average accuracy is 90.69% at a 0 dB SNR of 18 dB, showing a superior accuracy over the CNN-IQ design where training is accomplished based on the IQ data. The accuracy of classification with the present scheme is 89.2% at a SNR of 0 dB. Besides, the average accuracy at 0 dB SNR of 18 dB is 92.87%, and the maximum accuracy is up to 93.091%. As demonstrated by the simulations, our scheme outperforms the controls regarding classification accuracy.

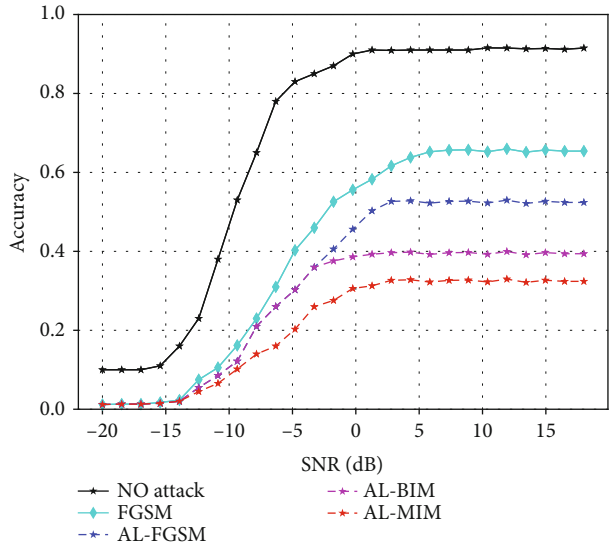


Figure 6: White-box untargeted attack.

4.4.2. White Attack. To investigate and analyze the impact of the attack on the modulation classification, this study compared the attack effect of the white-box methods in Figure 5. To demonstrate the effectiveness of our attack method, this paper selects the optimal recognition model proposed in this paper, uniformly selects the sample signal with a SNR of 18 dB, and then gives the recognition accuracy of the model under different signal-to-interference ratios (ES/E0). Figure 5 shows the results of the white-box attack. When the signal-to-interference ratio is insignificant, that is, the disturbance power is relatively large, and the accuracy of the FGSM method can be reduced to about 18% while the method proposed in this paper AL-MIM can be reduced to 0. In the range of 0-10 dB, we still attack the model to make its accuracy 0. From the overall trend, our proposed method is better than FGSM, and FGSM is superior to adding

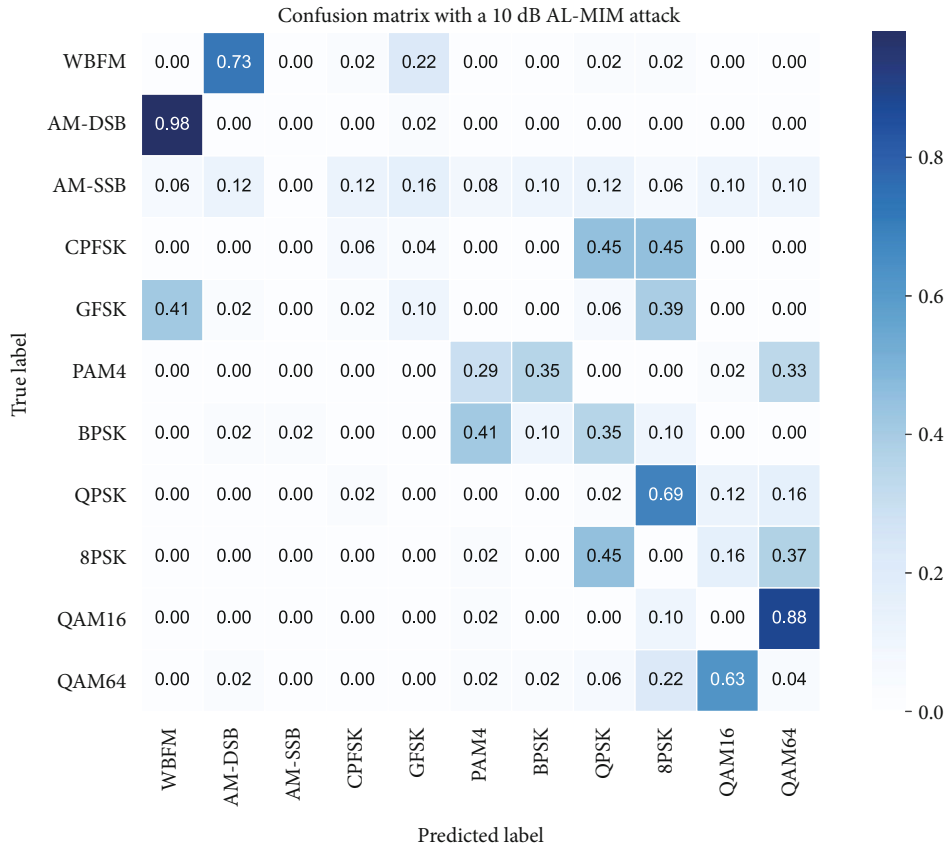


Figure 7: Confusion matrix of the AMR (SPR = 10 dB) predictions after AL-MIM attack.

ordinary white noise. The accuracy of about 90% of the training is reduced accordingly.

Furthermore, the modulation signal has the characteristic of different SNR values; so, we are carrying out the adversarial attack, and it is necessary to carry out the attack one by one for different SNR with $E_s/E_pE_p = 30$ dB. At the same time, the iterative attack is considered to achieve the best attack effect. Figure 5 shows the changes in the accuracy of the AMR scheme model based on the three attacks at -20-18 dB. According to Figure 6, with the SNR value added, the accuracy of the model's output shows an initially progressively improving trend and afterwards fluctuating around a specific value. As the only noniterative one-step attack algorithm, FGSM vary is fast, but the attack effect is not satisfactory. We could see that the two attack methods we proposed are better than FGSM and MIM. To deeply analyze the adversarial attack, Figure 7 presents a confusion matrix of the target model after yielding adversarial examples based on AL-MIM with SPR = 10 dB. It can be clearly found that there is an obvious chaotic impact on the type of modulation signals.

4.4.3. Black Attack. In contrast to the ideal experimental environment, the target model in the actual modulation signal recognition and communication adversarial environment is often invisible to the attacker, resulting in a black-box attack. That is, there are high requirements for the mobility of adversarial examples. Usually, the traditional attack uses alternative models to replace the target black-box model, and the black-box attack applied in the present work is a direct way to transfer the adversarial samples generated from the proposed scheme white-box attack to execute the attack with the purpose of better verifying the transferable of the adversarial example. Apart from that, the black-box attack is tested on two different network models, LSTM-AP and CLDNN-AP, respectively, and the experimental results are illustrated in Figures 8 and 9.

According to Figure 8, it can be observed that for the black-box model of LSTM-AP, the original adversarial samples that can bring down the target model in the white-box model have a significant decrease in the attack success rate when they are migrated to the LSTM model, especially for the label gradient-based attack method FGSM. In contrast, AL-FGSM, AL-BIM, and AL-MIM can still achieve better adversarial attack effect, reducing the accuracy rate of LSTM model drop to about 30%. A similar conclusion can be drawn from Figure 9, and the adversarial examples based on feature gradient can still maintain a good attack effect after transfer to the CLDNN-AP black box model even though it is not as efficient as the white-box attack.

Figure 10 represents the comparison of the transfer rate of the adversarial samples on the two black box models, where the FGSM method is not included in the comparison methods because of its poor attack. From Figure 10, it could

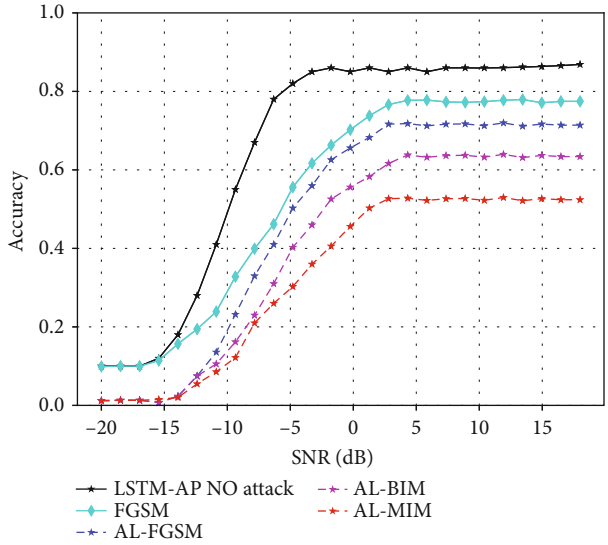


Figure 8: Black-box untargeted attack on LSTM-AP.

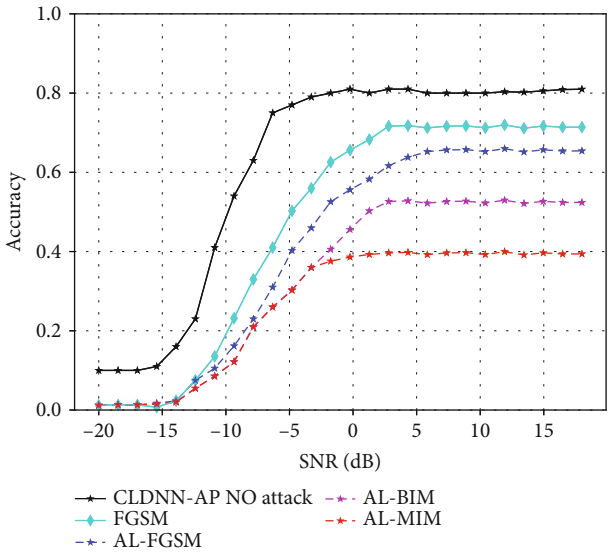


Figure 9: Black-box untarget attack on CLDNN-AP.

be seen that the black-box transfer rates of the two feature-based attack methods are higher than the traditional label-based methods for both LSTM-AP and CLDNN-AP models, which indicates that the feature-based attack methods have excellent attack transfer performance.

4.4.4. Analysis of the Effectiveness of the Attack. First, we make sure that the perturbation we introduce is small enough not to be recognized by the human eye with checking the effect of the perturbation on the signal fluctuations. The following modulation carrier formula is presented as

$$S(t) = I\cos(2\pi f t) + Q\sin(2\pi f t). \tag{27}$$

Furthermore, to the criterion of the success rate of the sample's attack on the model, the magnitude and intensity

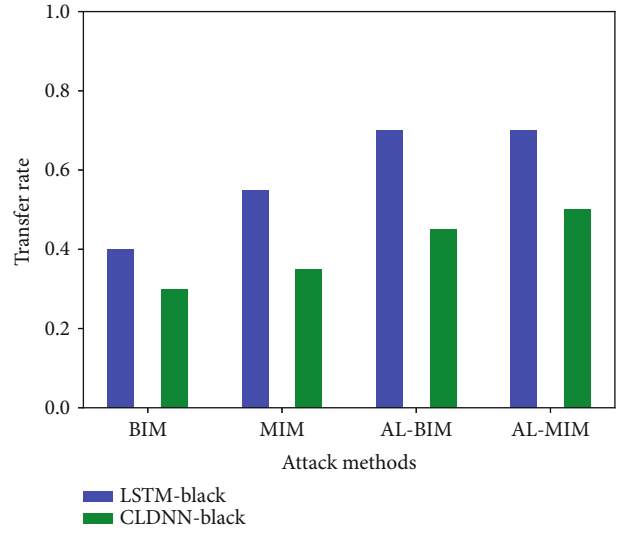
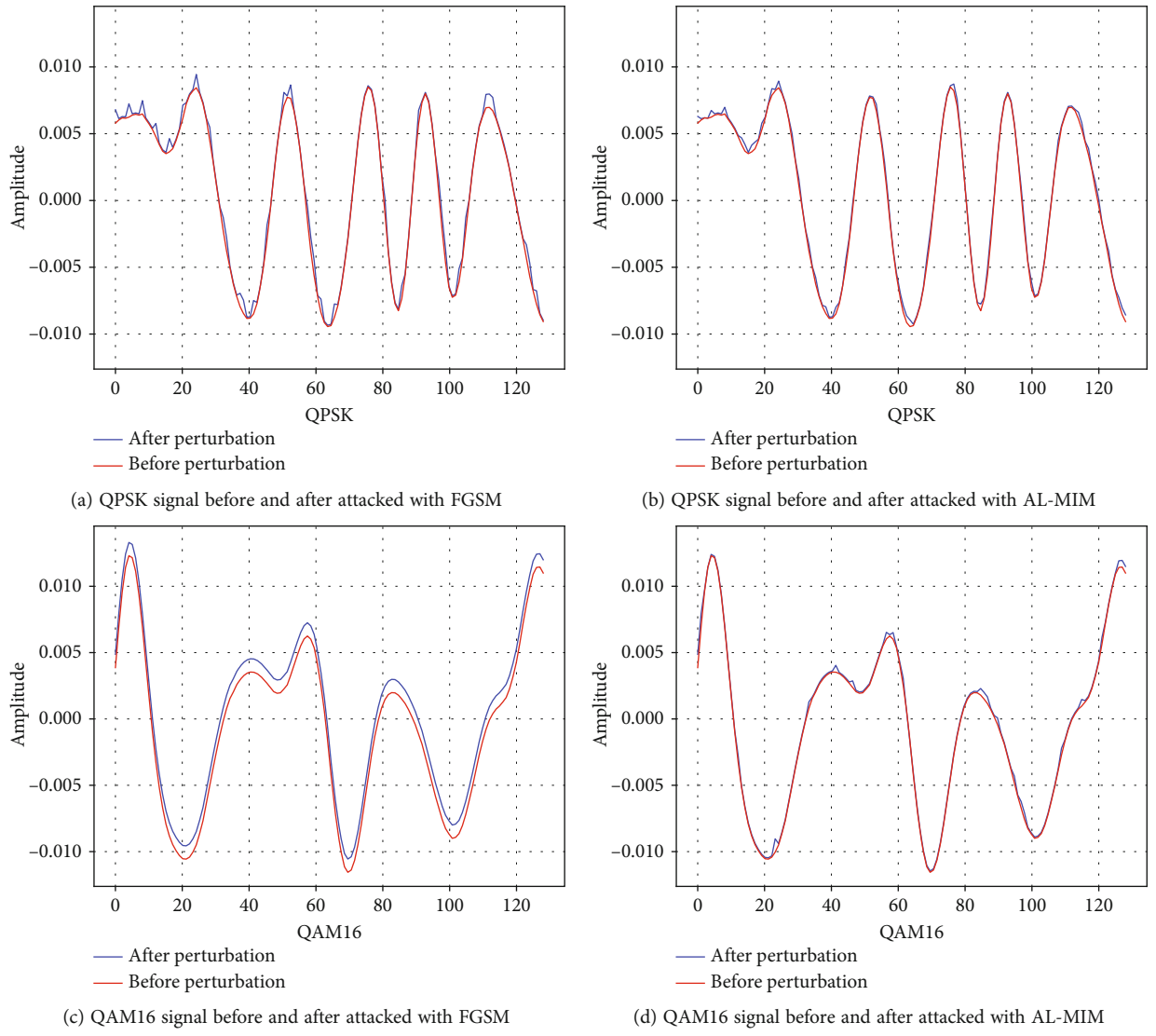


Figure 10: Black-box transfer rate.

of the perturbation of the modulated signal adversarial sample compared with the original sample are also important evaluation criteria. I represents the in-phase component, Q indicates the quadrature component, and f represents the carrier frequency. Subsequently, a primitive $S(t)$ signal can be yielded. By visualizing the $S(t)$, we could obtain the time domain waveform of the modulation signal. The time domain plots of the adversarial samples generated from the QPSK signal samples and their original signals are presented in Figures 11(a) and 11(b), while the plots of the adversarial samples generated from the QAM16 signal samples and their original signals are presented in Figures 11(c) and 11(d). Figures 11(a) and 11(c) show the signal perturbation based on the traditional label gradient method, while Figures 11(b) and 11(d) show the signal perturbation images based on the feature gradient AL-MIM method. It can be seen that for the same signal sample, the disturbance generated by the label gradient-based attack method often has continuous and violent jitter, which often does not suit to the image characteristics of a high signal-to-noise ratio modulated signal and is easily detected, and for the adversarial sample signal image of the feature gradient transferable attack, since the introduced perturbation is less in magnitude and jitter, it is more difficult to detect.

Then, to further analyze the adversarial attack approach, we selected high SNR value signals in the data set above or equal to 10 dB, with 32,000 samples in the training set. The results of the attack evaluation metrics will be presented, as shown in Table 1.

The attack method from misclassification and feature gradient attack outperform iterative attack and single step attack, and our method outperforms the traditional method from two metrics of imperceptibility. A signal is a physical quantity that representing a message, for example, an electrical signal can stand for different messages via changes in parameters such as amplitude, frequency, and phase. The signal is the carrier of the message, and in the process of signal attack, the variation of signal



(a) QPSK signal before and after attacked with FGSM

(b) QPSK signal before and after attacked with AL-MIM

(c) QAM16 signal before and after attacked with FGSM

(d) QAM16 signal before and after attacked with AL-MIM

Figure 11: Modulation samples with 18 dB before and after adding adversarial perturbation.

Table 1: Attack indicator results.

Type	Misclassification	Imperceptible		Signal characteristics			
Attack methods	ASR (%)	L_0	L_2	ACR	APD	PSR	TR
FGSM	95.46	0.85	5.14	1.41	0.30	-12.98	0.15
BIM	96.80	0.78	0.88	0.08	0.11	-23.38	0.40
MIM	97.45	0.70	1.01	0.11	0.14	-22.01	0.46
AL-BIM	98.90	0.50	0.95	0.07	0.10	-25.63	0.70
AL-MIM	99.97	0.32	0.22	0.06	0.03	-27.45	0.75

amplitude, phase, and other characteristics is extremely significant, and excessive distortion will make it difficult to extract the correct information. Four indicators (ACR, APD, PSR, TR) are used in this research to measure the distortion and migration rate of the signal. Based on Table 1, these performance indicators outperform the traditional attack methods.

5. Conclusion and Future Work

This paper addresses the security issues of the deep neural network model for AMR that is vulnerable to gradient attacks, and we propose a new adversarial attack method based on feature gradient transferability and design two attack algorithms, namely, AL-BIM and AL-MIM. These

methods aimed at feature attacks, which can extract and run regional attacks on the feature region of the original example captured by the neural network model by optimizing the triplet loss. The proposed scheme is more effective at attacking stable features in AMR-extracted signals, compared to the traditional label-based adversarial attack methods. Comprehensive experiments on public data sets show that the proposed feature gradient-based attack method in terms of attack method surpasses the traditional label gradient-based attack method in terms of attack success rate and transferability methods in both black-box attack and white-box attack scenarios. Additionally, the perturbation crafted using the feature gradient-based attack method is smoother and less perceptible. At the same time, four signal character indicators (ACR, APD, PSR, TR) are used in this research to measure the distortion and migration rate of the signal, and these performance indicators outperform the traditional attack methods. Further, decreasing the attack disturbance and narrowing the attack range are also our further research.

Data Availability

The simulation data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No. 62101594, No. 61901520) and the Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu under Grant BK20212001.

References

[1] T. J. OShea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*, C. Jayne and L. Iliadis, Eds., pp. 213–226, Springer International Publishing, Cham, 2016.

[2] O. Omotere, J. Fuller, L. Qian, and Z. Han, "Spectrum occupancy prediction in coexisting wireless systems using deep learning," in *IEEE Vehicular Technology Conference (VTC-Fall)*, Chicago, IL, USA, 2018.

[3] 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, 2019.

[4] Y. Lin, H. Zhao, X. Ma, T. Ya, and M. Wang, "Adversarial attacks in modulation recognition with convolutional neural networks," *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 389–401, 2021.

[5] R. Bryse Flowers, M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1102–1113, 2020.

[6] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, 2020.

[7] L. Gao, Z. Huang, J. Song, Y. Yang, and H. T. Shen, "Push & pull: transferable adversarial examples with attentive attack," *IEEE Transactions on Multimedia*, vol. 24, pp. 2329–2338, 2021.

[8] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, *How transferable are features in deep neural networks?*, I. P. S. Neur, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 3320–3328, 2014.

[9] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," *ICLR*, Y. Bengio and Y. LeCun, Eds., 2016.

[10] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen, "Feature space perturbations yield more transferable adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7066–7074, Long Beach, CA, USA, 2019.

[11] C. S. Weaver, C. A. Cole, R. B. Krumland, and M. L. Miller, *The Automatic Classification of Modulation Types by Pattern Recognition*, 1969.

[12] Z. T. Huang, J. Yang, X. Wang, X. Cui, and F. Y. Wang, "A survey of modulation recognition algorithms in noncooperative communication," *Science & Technology Review*, vol. 37, no. 4, pp. 55–62, 2019.

[13] J. L. Xu, W. Su, and M. Zhou, "Likelihood function-based modulation classification in bandwidth-constrained sensor networks," in *2010 International Conference on Networking, Sensing and Control (ICNSC)*, Chicago, IL, USA, 2010.

[14] P. Ghasemzadeh, S. Banerjee, M. Hempel, and H. Sharif, "Performance evaluation of feature-based automatic modulation classification," in *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Cairns, QLD, Australia, 2018.

[15] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43–58, New York, NY, USA, 2011.

[16] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: a survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[17] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015, https://arxiv.org/abs/1409.0473.

[19] S. Chen, Y. Zhang, Z. He, J. Nie, and W. Zhang, "A novel attention cooperative framework for automatic modulation recognition," *IEEE Access*, vol. 8, pp. 15673–15686, 2020.

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, UT, USA, 2018.

[21] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of 2010*

IEEE international symposium on circuits and systems, pp. 253–256, Paris, France, 2010.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] T. N. Sainath, A. W. Senior, O. Vinyals, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," 2020, U.S. Patent No. 10, 783,900.

[24] Y. Chen, W. Shao, J. Liu, L. Yu, and Z. Qian, "Automatic modulation classification scheme based on LSTM with random erasing and attention mechanism," *IEEE Access*, vol. 8, pp. 154290–154300, 2020.

[25] C. Szegedy, W. Zaremba, I. Sutskever et al., "Intriguing properties of neural networks," 2013, https://arxiv.org/abs/1312.6199.

[26] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, https://arxiv.org/abs/1412.6572.

[27] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018.

[28] Y. Dong, F. Liao, T. Pang et al., "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018.

[29] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2016.

[30] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," 2019, https://arxiv.org/abs/1908.06281.

[31] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, https://arxiv.org/abs/1611.01236.

[32] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE symposium on security and privacy (sp)*, San Jose, CA, USA, 2017.

[33] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, Stockholm, Sweden, 2018.

[34] Y. Lin, H. Zhao, Y. Tu, S. Mao, and Z. Dou, "Threats of adversarial attacks in DNN-based modulation recognition," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2020.

[35] A. Jeddi, M. J. Shafiee, M. Karg, C. Scharfenberger, and A. Wong, "Learn2perturb: an end-to-end feature perturbation learning to improve adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.