Adversarial Attacks on Deep Neural Network based Modulation Recognition

Mingqian Liu and Zhenju Zhang*
State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, Shaanxi, China
mqliu@mail.xidian.edu.cn, zhenjuzhang@stu.xidian.edu.cn
*corresponding author

***Abstract*—Modulation recognition technology based on deep learning (DL) has great advantages in feature extraction and recognition. However, due to the vulnerability of deep neural network (DNN), the automatic modulation recognition model based on DNN is vulnerable to attacks. Some researchers have successfully attacked automatic modulation recognition model-s using adversarial techniques, but the resulting adversarial samples have poor attack performance on high-performance recognition models. Therefore, this paper proposes an attack method based on double loop iteration, which can update the initial conditions of each iteration with the change of the number of iterations when generating adversarial examples. Simulation results show that the proposed attack method has better attack performance than the traditional attack methods.**

***Keywords—Adversarial attacks, adversarial examples, deep neural network, modulation recognition, network vulnerability*.**

I. Introduction

Modulation recognition of signal refers to the correct recognition of the modulation type of the target signal from the modulation set of the signal. In recent years, with the rapid development of deep learning (DL), researchers have tried to apply neural network to modulation recognition technology, and achieved good recognition results [1] - [3]. However, the defects of DL in interpretability reduce the security of deep neural network (DNN) model and make the recognition model vulnerable to attacks. Szegedy et al. found that adding some subtle disturbances that are hard to detect by human beings in the input samples can significantly reduce the recognition accuracy of the classifier, and proposed the concept of adversarial examples [4]. This means that the recognition model based on DNN is vulnerable to attack. Therefore, it is very important to speed up the research on adversarial attack and improve the robustness of DNN model.

After the adversarial example was proposed, many researchers have successively proposed some attack methods. Goodfellow et al. generated adversarial samples by adding a certain amount of disturbance in a specific direction, and proposed the fast gradient symbol method (FGSM) [5]. Kurakin et al. proposed the basic iterative method (BIM), which uses multiple iterations to generate adversarial examples [6]. Madry et al. proposed the projection gradient descent method (PGD), which added a projection step on the basis of BIM to randomly initialize the search for adversarial examples [7]. Dong et al. proposed the momentum iterative method (MIM), which introduced the momentum into the gradient calculation process in the iteration attack [8]. In order to reversely improve the robustness of the automatic modulation recognition model, some researchers introduced some existing attack methods into the wireless communication field and attacked the automatic modulation recognition model. [9] - [10].

Based on the existing attack methods, this paper proposes an adversarial attack method based on double loop iteration. This method can initialize the conditions for generating adversarial examples by adding an external loop iteration layer, which can enhance the attack performance of adversarial examples on the high-performance recognition model. The main contributions of this paper are as follows:

- We trained a modulation recognition model and achieved high recognition accuracy for modulated signals.
- We proposed an attack method based on double loop iteration to improved the attack performance.
- We compared the attack performance of our proposed attack method with four traditional attack methods.

II. Double Loop Iterative Method

The FGSM, BIM, PGD and MIM methods mentioned in the previous section belong to the label-based gradient attack methods, and they all have only one loop iteration layer. After the total iterative step size reaches the constraint of L_∞ norm, the adversarial examples generated by them will be determined. However, because the direction and step size of some iterations may be difficult to make the adversarial examples reach the maximum point of the loss function, this method of using single-layer loop iteration may fall into a misunderstanding. In order to make the adversarial examples as close as possible to the optimal point of the loss function, we improve the traditional iterative attack methods. By adding an external loop iteration, the improved method can update the initial conditions of each loop iteration, so that the adversarial examples can better approach the optimal point of the loss function within a limited number of iterations.

The number of external loop iterations is denoted as M, and the number of momentum iterations is denoted as N. When setting the step size of each external loop iteration, the step size needs to satisfy the condition that it is not greater than L_∞ norm constraint ε and not less than momentum iteration step size ε/N. At the same time, the adversarial examples should find the approximate position of the optimal point of

2767-6684/22/$31.00 ©2022 IEEE
DOI 10.1109/DSA56465.2022.00159

TABLE 1

The Recognition Accuracy of ResNet with Different M and N

N	5			10			25	100
M	1	3	5	1	5	10	1	1
Accuracy(%)	53.9	48.3	46.8	50.4	46.3	45.2	49.3	49.8

the loss function with a larger step in the early iteration, and the specific position of the optimal point with a smaller step in the later iteration. Thus, we set the step size α_m of the external loop iteration to

$$\alpha_m = \frac{(M - m + 1) \cdot \varepsilon}{N}, \tag{1}$$

where m denotes the current loop number of external loop iteration, and the constraint condition of α_m is $1 \leq m \leq M$ and $1 \leq M \leq N$. When $M = N$, $\varepsilon/N \leq \alpha_m \leq \varepsilon$, and α_m gradually decreases with the external loop.

We call our attack method Double Loop Iterative Method (DLIM). The accumulated gradient generated by the first n iterations is denoted as g_n, and μ is the attenuation factor of g_n. In DLIM, the process of generating adversarial examples ${x_m}^*$ in the $m-th$ external loop iteration can be formulated as follows.

$$\begin{cases} {x_0}^* = {x_m}^*, g_0 = 0, \alpha_0 = 0, \\ g_{n+1} = \mu \cdot g_n + \frac{\nabla_{{x_n}^*} J({x_n}^*, l))}{\left\| \nabla_{{x_n}^*} J({x_n}^*, l) \right\|_1}, \\ \alpha_m = (M - m + 1) \cdot \varepsilon / N, \\ {x_{n+1}}^* = \mathrm{Clip}_{x,\varepsilon} \left\{ {x_n}^* + \alpha_m \cdot \mathrm{sign}\left(g_{n+1}\right) \right\}, \\ {x_{m+1}}^* = {x_N}^*. \end{cases} \tag{2}$$

In order to study the impact of iterations in α_m on attack performance, we choose different iterations to generate adversarial examples, and use these examples to test the accuracy of the model. The test results are shown in Table I.

In Table I, the attack performance of $M = N = 5$ is better than that of $N = 25$ and $M = 1$, and the attack performance of $M = N = 10$ is better than that of $N = 100$ and $M = 1$. This shows that the DLIM has better attack performance than simply increasing the number of iterations.

III. SIMULATION RESULT AND ANALYSIS

Under the condition of $N = M = 10$ and $\varepsilon = 0.0015$, we input the adversarial examples generated by different attack methods into the ResNet model to test the effect of SNR on the attack performance of the adversarial examples. The simulation results are shown in Fig. 1.

In Fig. 1, we show the recognition accuracy of the model for different adversarial examples under different SNRs. With the increase of SNR, the difference in attack performance of different attack methods gradually appears. After the recognition accuracy of the model is stable, the proposed method DLIM makes the recognition accuracy of the model decrease most. When $SNR = 12dB$, compared with MIM, DLIM reduces the recognition accuracy of the model by 5.6 percentage points.

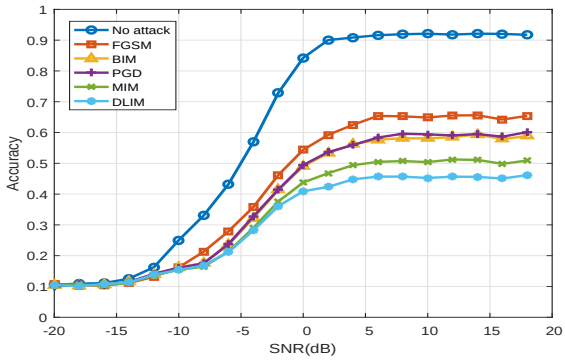

Fig. 1. Recognition accuracy with the different SNRs.

IV. CONCLUSION

In this paper, we study the security problem of automatic modulation recognition model vulnerable to attack. Based on the existing attack methods, we propose and design a double loop iterative method. This method initializes the conditions of generating adversarial by adding an external loop iteration layer, which can enhance the attack performance of adversarial examples. Simulation results show that compared with the traditional single multi-step iterative attack method, the proposed method has better attack performance on the DNN model with high recognition accuracy.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant 62071364.

REFERENCES

[1] T. J. O'Shea, 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.

[2] 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, April 2019.

[3] 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, Sept. 2018.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna and R, Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations*, pp. 1-10, May 2015.

[5] I. Goodfellow, J. Shlens, and C. Szegedyn, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, pp. 189-199, Mar. 2015.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int. Conf. Learn. Representations*, pp. 128-141, May 2016.

[7] A. Madry, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations*, vol. 1, pp. 1-23, May 2018.

[8] Y. Dong and F. Liao, "Boosting adversarial attacks with momentum," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, pp. 9185-9903, Mar. 2018.

[9] 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.

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