POSITION-INVARIANT ADVERSARIAL ATTACKS ON NEURAL MODULATION RECOGNITION

Zhen Yu[1⋆], *Yifeng Xiong*[1⋆], *Kun He*[1†], *Shao Huang*[2], *Yaodong Zhao*[2] *and Jie Gu*[2]

[1] School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
[2] Science and Technology on Electronic Information Control Laboratory, Chengdu, China
{baiding15, xiongyf, brooklet60}@hust.edu.cn, {705637006, 359368991, 13558699175}@qq.com

ABSTRACT

Deep neural networks (DNNs) are widely used for neural modulation recognition (NMR) in the electronic field and have been shown to be vulnerable to adversarial examples for NMR. In the physical signal communication scenario, the adversarial signal transmitted by the adversary is affected by the channel, resulting in a random time delay with the original signal and causing decay on the attack performance. To address this issue, we propose the Position-Invariant adversarial attack Method (PIM) that generates the position-invariant adversarial signal by averaging the adversarial signals generated by shifted input signals to mitigate the channel effect on time delay. Our PIM can be easily integrated with other methods to achieve better results. Extensive experiments demonstrate that the proposed method could outperform all baselines for adversarial attacks on NMR under the time delay setting.

Index Terms— Security, neural networks, modulation recognition, adversarial attack, wireless communication

1. INTRODUCTION

Despite the great success of deep learning techniques in a broad spectrum of applications, the adversarial example, which aims to add imperceptible perturbation to the original input to fool the target classifier, brings critical security threats to the deep learning based systems in various domains, *i.e*. Computer Vision [1, 2], Natural Language Processing [3–5], Electronic Communication [6, 7], *etc*.

Communication in the electronic field is safety-critical, and electronic countermeasure has been a hot topic for decades [8]. With the extensive utilization of deep learning techniques, it is necessary to study adversarial examples for electronic countermeasure in various electronic communication tasks, such as spectrum sensing [9], channel coding [10], modulation recognition [11], *etc*. Neural modulation recognition (NMR), in which the deep learning based methods are adopted for modulation recognition, is an essential component of signal communication, which plays a key role in identifying modulation modes of unknown signals.

Although existing attack methods [6, 7, 12] have achieved good attack performance on NMR, they are invalid when facing specific real-world problems, such as channel effects on amplitude, phase [13] and time delay [14] of the signal. The time delay effect makes the original signal and the adversarial signal unable to reach the receiver precisely at the same time, which results in a specific offset between them, and destroys the aggressiveness of the adversarial signal. In this work, we find that the generation of adversarial signal corresponds to the position of the original signal sequentially, and the information in the adversarial signal we get by feeding the shifted original signal will be the shifted position information. Based on this observation, we try to utilize such correspondence to construct the position-invariant adversarial signal which is universal in the position dimension, *i.e*. the data of a specific position will contain the information of multiple positions such that the shifted adversarial signal still has enough information to attack the original signal.

To this end, we propose the Position-Invariant adversarial attack Method (PIM) to resist the channel effect on time delay. Specifically, we first disrupt the position of the original signal by concatenating the data of different positions. Then we feed the transformed signals to the neural network to generate a series of adversarial signals which are then averaged to obtain position-invariant adversarial signals. To validate the effectiveness of the proposed method, we conduct extensive experiments to compare the performance of different attack methods with the proposed PIM under various signal-to-noise ratios (SNRs) and perturbation sizes. The experiments show that PIM performs well in time delay setting, achieving a higher attack performance than other attack methods.

2. RELEATED WORK

Adversarial attack settings in wireless communication systems usually have three roles, namely the transmitter, the re-

[⋆] The first two authors contribute equally to this work.

[†] The corresponding author.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

ceiver, and the adversary. The transmitter transmits a modulated signal, which travels through the channel to the receiver, and the receiver demodulates the signal to obtain data. In our attack setting, the adversary transmits the adversarial signal to the receiver, which is superimposed on the original signal to attack the modulation recognition classifier.

Sadeghi and Larsson [6] first introduce adversarial attacks into wireless communications and transfer the typical attack FGSM [2] from CV to modulation recognition tasks, showing that the DNN-based modulation classifier is vulnerable to adversarial examples. Then, researchers [7, 12] transfer more typical attack methods to attack the NMR, *e.g.* BIM [15], PGD [16], MIM [17], NAM [12], *etc.* Meysam and Erik [14] observe that channels will have various effects on the transmitted adversarial signals in real-world scenarios, resulting in decay on the attack performance. They propose SDM to mitigate the channel effects on time delay, which utilizes singular value decomposition to extract the principal component of randomly shifted adversarial signals to obtain the position-invariant adversarial signal. Kim *et al.* [13] inverse the adversarial signal by simulating the effect on the adversarial signal to eliminate the channel effects on amplitude and phase.

In this work, we focus on the channel effect on time delay. Under such a scenario, the adversarial signal transmitted through the channel will be shifted before being added to the original signal at the receiver. We propose to transform the input signal to generate adversarial signals with different position information and obtain position-invariant adversarial signals by averaging these adversarial signals. Compared with the SDM that directly disrupts the adversarial signal, disrupting the input signal and then feeding it to the neural network ensures the integrity of the time series and information of the adversarial signal, and provides better attack capability.

3. METHODOLOGY

3.1. Problem Formulation

Given the input space $\mathcal{X}$ and output space $\mathcal{Y} = \{y_1, y_2, \dots, y_k\}$, a DNN-based modulation classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ predicts the label y for any input signal $x \in \mathcal{X}$, in which y is expected to equal to its ground-truth label y_{true}:

$$y = \arg\max_{y_i \in \mathcal{Y}} f(y_i|x).$$

The adversary typically adds an imperceptible adversarial signal δ_x on the input signal x to craft an adversarial example $x_{adv} = x + \delta_x$ that misleads the classifier f:

$$\arg\max_{y_i \in \mathcal{Y}} f(y_i|x + \delta_x) \neq y_{true}, \quad s.t. \quad ||\delta_x||_p \leq \epsilon,$$

where $||\cdot||_p$ is the p-norm that measures the distance between the original sample and the adversarial example, and ϵ is a hyper-parameter for the perturbation upper bound.

To find the above adversarial perturbation, researchers [2] usually transfer the problem to a constrained optimization problem:

$$\arg\max_{\delta_x} J(f(x + \delta_x), y), \quad s.t. \quad ||\delta_x||_p \leq \epsilon, \qquad (1)$$

where $J(\cdot, \cdot)$ is the loss function.

Furthermore, we assume that the adversary constantly transmits the adversarial signals to attack the original signals, so the perturbation added to the original signal is still a complete adversarial signal.

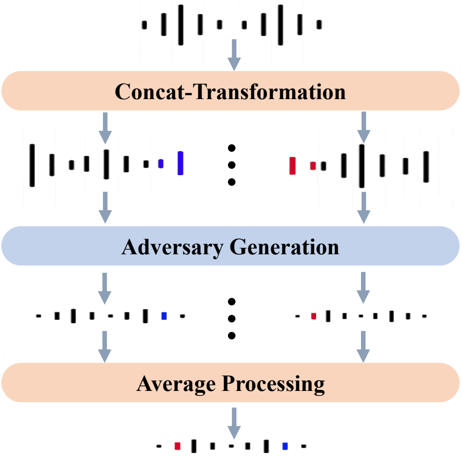


Fig. 1. **The overall framework of the proposed PIM.** For an input signal, we first perform concat-transformation to generate many position-shifted samples and then utilize the attack method to generate attack signals with shifted information. Finally, we average the obtained attack signals to get the position-invariant attack signal.

3.2. Generating Adversarial Signals

The proposed Position-Invariant adversarial attack Method (PIM) is an input transformation attack method that contains three main steps, *i.e.* concat-transformation, adversary generation, and average processing, as shown in Fig. 1.

Given an input signal x, we first randomly concatenate part of the input head to the tail or concatenate part of the input tail to the head to obtain a position-shifted sample x_{new}. Then, we feed x_{new} into the classifier to get the gradient that reflects the influence of the input on the output and then construct the adversarial signal δ_x along the gradient direction like FGSM [2], PGD [16], and other gradient-based methods. Due to the position correspondence of the gradient in the backpropagation, we find that the generated adversarial signal has the shifted information, that is, the tail (or head) of the adversarial signal has the ability to attack the head (or tail) of the original signal. We perform concat-transformation and adversary generation for N times to generate a set of adversarial signals $\mathcal{S}$. In the end, we average the adversarial signals in $\mathcal{S}$ to generate a position-independent adversarial signal

Algorithm 1 The PIM Algorithm (PGD version)

```
Input: Benign sample $x$, target classifier $f$, number of iterations in PGD $K$, time-decay upper-bound $T$, number of iterations in average processing $N$, perturbation size $\epsilon$.
Output: Adversarial signal $\delta_x$
1:  Initialize the initial adversarial example $x_{adv}^1 = x$
2:  for $i = 1 \to K$ do
3:      Initialize adversarial signal set $\mathcal{S} = \emptyset$
4:      for $j = 1 \to N$ do
5:          Randomly pick the length of shift $t$ from $[-T, T]$
6:          Obtain the shifted head data $head = x_{adv}^i[t :]$
7:          Obtain the shifted tail data $tail = x_{adv}^i[: t]$
8:          Concatenate $head$ and $tail$ for a new sample $x_{new}$
9:          Calculate the gradient $\nabla_x J(x_{new}, y)$
10:         Calculate adversarial signal $\delta_{ij}$ by the gradient:
11:             $\delta_{ij} = \epsilon / K \cdot sign(\nabla_x J(x_{new}, y))$
12:         $\mathcal{S} = \mathcal{S} \cup \delta_{ij}$
13:     end for
14:     Get adversarial signal $\delta_i$ by averaging signals in $\mathcal{S}$
15:     Update the adversarial example $x_{adv}^{i+1} = x_{adv}^i + \delta_i$
16: end for
17: return Adversarial signal $(x_{adv}^{K+1} - x)$
```

that is universal in the position dimension so that there is still enough information to attack even if the adversarial signal is misaligned with the original signal.

The neural network has been shown to be translation-invariant [18] so that the neural network can extract all the features even for the shifted input samples. Therefore, the adversarial signal we obtained still has the global information of the input samples. Moreover, our PIM can be integrated into various gradient-based attack methods, *i.e.* FGSM [2], PGD [16], MIM [17], NAM [12], *etc.* Taking the PGD as an example, the overall PIM algorithm is presented in Algorithm 1, in which steps 5-8 represent the concat-transformation, and step 14 describes the average processing.

4. EXPERIMENTS

4.1. Experimental Setup

We first introduce the experimental setup, including dataset, models, baselines, and attack settings used in the experiments.

Dataset and Models. We select the public dataset RADIOML 2016.10a [19], consisting of 220,000 samples, 20 different SNRs varying from -20 dB to 18 dB with step size 2, and 11 modulations. We adopt two typical neural networks on NMR, including VTCNN2 [19] and RadioGRU [20]. For each SNR of each modulation, we select ten samples from the testset, with a total of 2,200 samples as the attack set.

Baselines. We choose the SDM [14] and four typical attack methods, *i.e.* FGSM [2], PGD [16], MIM [17] and NAM [12] as our baselines. We integrate our PIM and SDM into the above four base attacks, respectively, and have four implementation versions (denote this series as PIM-X and SDM-X). We also take the traditional jamming attack as a reference, simply transmitting a Gaussian noise for the attack.

	SNR = 0 dB		SNR = 10 dB	
	White-box	Black-box	White-box	Black-box
No attack	76.8	85.0	81.5	86.4
Jamming attack	60.8	73.1	72.4	79.1
FGSM	47.1	53.4	49.6	60.1
SDM-FGSM	33.5	35.3	32.2	37.1
PIM-FGSM	29.8	32.8	29.3	37.1
PGD	46.0	54.3	49.3	55.0
SDM-PGD	25.9	36.4	25.6	35.6
PIM-PGD	25.6	29.1	**23.7**	29.5
MIM	37.8	44.6	39.6	46.3
SDM-MIM	33.8	36.0	32.8	37.5
PIM-MIM	31.8	36.6	30.8	35.2
NAM	32.9	40.6	35.3	43.0
SDM-NAM	25.9	36.0	25.5	35.2
PIM-NAM	**25.5**	**25.4**	23.9	**25.5**

Table 1. The classification accuracy (%) of various attacks under white-box or black-box setting on SNR = 0 dB and 10 dB in time delay setting. The best results are in **blue**.

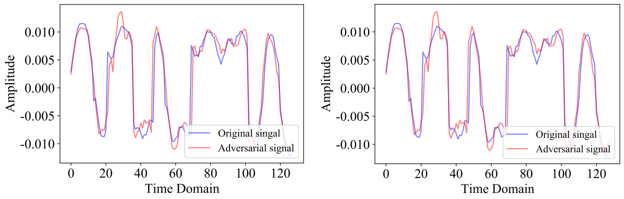
Attack Settings. For our PIM, we set the time delay upper-bound $T = 8$ and the number of iterations in average processing $N = 128$. Our adversarial attacks fall into two settings, white-box attacks and black-box attacks. White-box attacks have access to the target model, including the logits, model parameters, gradients, and architectures. Black-box attacks only allow access to the outputs. We utilize the adversarial examples generated on VTCNN2 to attack VTCNN2 for white-box attacks and RadioGRU for black-box attacks.

4.2. Comparison on Attack Methods

We evaluate the classification accuracy of our PIM and baselines in time delay setting on SNR = 0 dB and 10 dB, as shown in Table 1. For a fair comparison, we project the generated adversarial examples to ensure that the average absolute value ϵ of the adversarial signal is about 0.002, that is, to ensure that the L1 norm of the adversarial signal is the same.

From the results, we observe that the adversarial attack methods significantly reduce the classification accuracy compared with the traditional jamming attack. Both PIM and SDM can be integrated with other attack methods, which could further reduce the classification accuracy to a considerable extent (4-23.7% for SDM-X and 6-25.6% for PIM-X) in time delay setting. Moreover, PIM-X always outperforms SDM-X under both white-box and black-box settings. As a concrete example, PIM-NAM reduces accuracy from 85.0% to 25.4% on SNR=0 dB under the black-box setting, which is 15.2% better than NAM and 10.6% better than SDM-NAM.

In general, PIM-NAM outperforms all other methods, achieves the best attack performance in time delay setting, and reduces the accuracy by 51.3-60.9%. This validates the



(a) PAM4 misclassified as BPSK (b) 8PSK misclassified as QPSK

Fig. 2. Original signal versus adversarial signal. (a) The PAM4 signal with confidence of 93.3% is misclassified as a BPSK signal with confidence of 82.6%. (b) The QPSK signal with confidence of 60.6% is misclassified as a 8PSK signal with confidence of 99.3%.

high effectiveness of our PIM. It is worth mentioning that the PIM series also exhibits excellent attack performance under the black-box setting, which is very close to that under the white-box setting. Such property provides feasible application prospect of PIM in the real physical world, in which the target model is usually unknown and not accessible.

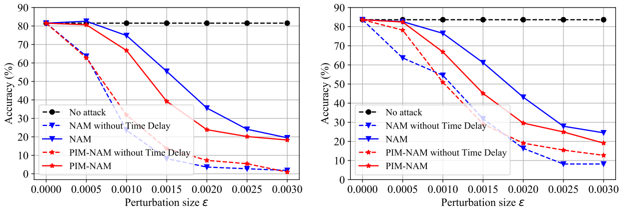
In addition, to show the imperceptibility of adversarial attacks on NMR, we visualize two examples for the original signal and the adversarial signal in Fig. 2. After adding an imperceptible perturbation onto the original signal, we can fool the classifier to output another category of signals.

4.3. Influence of the Time Delay

Channel effects on time delay destroy the position correspondence between the original signal and the adversarial signal. To study the influence of the time delay for NMR adversarial attacks, we conduct additional experiments to evaluate the attack performance of PIM-NAM and NAM with or without delay effects. Fig. 3 shows the classification accuracy varying different perturbation sizes ϵ under white-box and black-box attack setting on SNR = 10 dB. We could find that both NAM and PIM-NAM are greatly affected by the time delay, resulting in a significant decay in the success rate of attacks on various perturbation sizes. However, compared with NAM, PIM is significantly less affected and outperforms NAM after being affected by time delay, whether in white-box or black-box attack setting. This further verifies the effectiveness of our PIM in time delay setting.

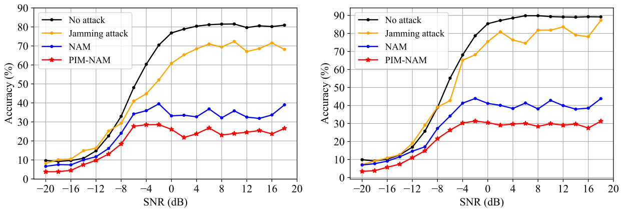
4.4. Attacks under Various SNRs

SNR indicates the quantity of noise added to the original signal which greatly affects the classification performance of the classifier for signals. Thus, a good attack should exhibit consistent and superior attack performance under various SNRs. We report the classification accuracy of PIM-NAM, NAM, and jamming attack on different SNRs under white-box or black-box setting in time delay setting, as shown in Fig. 4. The results show that PIM-NAM outperforms NAM and traditional jamming attack on various SNRs in time delay setting. In particular, PIM-NAM could further reduce the accuracy by 10% in contrast to NAM on SNR > -8 dB, for both white-box and black-box attack settings. It demonstrates the superiority of our PIM on various SNRs in time delay setting.



(a) White-box setting (b) Black-box setting

Fig. 3. The classification accuracy (%) after NAM and PIM-NAM attack varying perturbation sizes ϵ with or without time delay on SNR = 10 dB.



(a) White-box setting (b) Black-box setting

Fig. 4. The classification accuracy (%) after NAM and PIM-NAM attack varying SNRs in time delay setting on $\epsilon = 0.002$ perturbation.

5. CONCLUSION

In this work, we explore the channel effect on time-delay for adversarial attacks in the neural modulation recognition. We find that the adversarial example can have the information of other positions by making specific changes to the input. Based on this observation, we propose a novel adversarial attack method called the Position-Invariant adversarial attack Method (PIM) to eliminate the effects of time-delay. PIM first adopts concat-transformation and typical gradient-based attack to generate a set of adversarial signals with different position information, and then average them to obtain a position-invariant adversarial signal. Empirical evaluations demonstrate that the proposed PIM could cooperate with other attack methods and outperform existing attack methods on various perturbation sizes and SNRs in time-delay setting, for both white-box or black-box attacks.

6. ACKNOWLEDGEMENT

This work is supported by National Natural Science Foundation (62076105) and the Open Research Fund of Science and Technology on Electronic Information Control Laboratory (JS20201100330).

7. REFERENCES

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus, “Intriguing properties of neural networks,” in *2nd International Conference on Learning Representations, ICLR*, 2014.

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, “Explaining and harnessing adversarial examples,” in *3rd International Conference on Learning Representations, ICLR*, 2015.

[3] Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang, “Crafting adversarial input sequences for recurrent neural networks,” in *MILCOM IEEE Military Communications Conference*, 2016.

[4] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi, “Deep text classification can be fooled,” in *International Joint Conference on Artificial Intelligence*, 2018.

[5] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan, “Real-time, universal, and robust adversarial attacks against speaker recognition systems,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2020.

[6] Meysam Sadeghi and Erik G Larsson, “Adversarial attacks on deep-learning based radio signal classification,” *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213–216, 2018.

[7] Yun Lin, Haojun Zhao, Ya Tu, Shiwen Mao, and Zheng Dou, “Threats of adversarial attacks in DNN-based modulation recognition,” in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 2469–2478.

[8] Bernard Eydt, Les Owens, Karen Scarfone, Les Owens, Karen Scarfone, Robert C. Cresanti, and Under Secretary Of Commerce, “Establishing wireless robust security networks: A guide to,” in *IEEE 802.11i, NIST Special Publication*, 2007.

[9] Mahmoud Nazzal, Alí Riza Ektí, Ali Görçin, and Hüseyin Arslan, “Exploiting sparsity recovery for compressive spectrum sensing: a machine learning approach,” *IEEE Access*, vol. 7, pp. 126098–126110, 2019.

[10] Ade Irawan, Gunawan Witjaksono, and Wahyu Kunto Wibowo, “Deep learning for polar codes over flat fading channels,” in *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2019, pp. 488–491.

[11] Timothy James O'Shea, Tamoghna Roy, and T Charles 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, 2018.

[12] Haojun Zhao, Yun Lin, Song Gao, and Shui Yu, “Evaluating and improving adversarial attacks on DNN-based modulation recognition,” in *GLOBECOM*. 2020, pp. 1–5, IEEE.

[13] Brian Kim, Yalin E Sagduyu, Kemal Davaslioglu, Tugba Erpek, and Sennur 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)*. IEEE, 2020, pp. 1–6.

[14] Meysam Sadeghi and Erik G Larsson, “Physical adversarial attacks against end-to-end autoencoder communication systems,” *IEEE Communications Letters*, vol. 23, no. 5, pp. 847–850, 2019.

[15] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio, “Adversarial examples in the physical world,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, “Towards deep learning models resistant to adversarial attacks,” in *6th International Conference on Learning Representations, ICLR*. 2018, OpenReview.net.

[17] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li, “Boosting adversarial attacks with momentum,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. 2018, pp. 9185–9193, IEEE Computer Society.

[18] Kunihiko Fukushima, “A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biol. Cybern.*, vol. 36, pp. 193–202, 1980.

[19] Timothy J O'shea and Nathan West, “Radio machine learning dataset generation with gnu radio,” in *Proceedings of the GNU Radio Conference*, 2016, vol. 1.

[20] Dehua Hong, Zilong Zhang, and Xiaodong Xu, “Automatic modulation classification using recurrent neural networks,” in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2017, pp. 695–700.