URSI GASS 2021, Rome, Italy, 28 August - 4 September 2021

Adversarial Deception on Deep-Learning Based Radio Waveforms Classification

Shuting Tang[1], Mingliang Tao[1]*, Xiang Zhang[2], Yifei Fan[2], Jia Su[2], Ling Wang[2]
[1]School of Electronics and Information, Northwestern Polytechnical University, China
[2]Shanghai Institute of Satellite Engineering, Shanghai 200090, China

Abstract

Deep learning has achieved superior performance on radio signal modulation classification. However, its security and reliability are vulnerable due to its data-oriented learning strategy. In this paper, the vulnerability of deep neural network for radio classification is investigated. Based on the radar and communication waveforms, slight optimal perturbations are generated and added onto the original signals by searching for the saliency map. The simulated results show that the deep neural network will suffer degradation of classification accuracy due to adversarial deception both for radio signals with high and low signal-to-noise ratio.

1 Introduction

Deep neural network (DNN) has obtained high efficiency and accuracy in many classification tasks in recent years [1]. Specifically, DNN performs well in the field of wireless communication systems [2]. O'Shea *et al.* used DNN for radio signal modulation classification and realized superior classification accuracy [3-4].

Although DNN has made great breakthroughs in the field of artificial intelligence, it relies heavily on the distribution of training dataset and has inherent limit [5]. Szegedy *et al.* found that adversarial deception images can be generated by adding small perturbations to original images which are imperceptible to human vision system. The DNN classifier changed its prediction results of the adversarial deception images completely in such a way. It has been shown that DNN is highly vulnerable to adversarial examples, which raises significant security and robustness concerns [6]. In [7], many kinds of adversarial deception algorithms and defense algorithms are thoroughly introduced, in which proved the vulnerability of the image classification model. Similarly, there is also great potential for damage from adversarial deception in radio classification systems, such as communication modulation. Kokalj *et al.* used Fast Gradient Sign Method (FGSM) to demonstrate the vulnerability of modulation classification against adversarial examples [8]. What's more, the deception is significantly more powerful than classical jamming deception [9].

Though the previous studies had verified that adversarial examples degraded the classification performance, the perturbations are generated based on the whole signal samples. In this paper, by drawing on the idea of Jacobin-based Saliency Map Attack (JSMA) [10], adversarial deception signals are generated by changing certain few points of the original signal, which can deceive the DNN secretly.

2 Methodology

In this section, procedures for crafting a white-box adversarial deception on the intelligent radio classification system are demonstrated.

Consider an original signal X, the output of DNN is $F(X)$, classified as $F(X)=Y$. Considering the targeted deception is performed, we assign a particular class t as the target label before deception. Adversarial deception signal X^* is optimized to be as similar as X, but misclassified as $F(X^*)=Y_t \neq Y$. To generate X^*, the adversary needs to search for special points which have significant impact on the whole original signal. To realize the purpose above, three main parts will be implemented.

Firstly, calculate forward derivative of DNN in Eq. (1). The forward derivative is defined as the Jacobin matrix of the function F learned by DNN after training,

$$\nabla F(X)=\frac{\partial F(X)}{\partial X}=\left[\frac{\partial F_j(X)}{\partial X_i}\right]_{i\in 1\ldots M,\, j\in 1\ldots N} \tag{1}$$

where ∇ represents gradient, M and N represents the input and output dimension of the DNN, respectively. By calculating the forward derivative, the input features that lead to significant changes in network outputs can be found.

Then, construct a saliency map S based on the derivative. Saliency map is a visualization tools that shows the uniqueness of each pixel. It shows the influence of different input features to the classification results. Therefore, through this saliency map, adversary will perturb the input components to effect the desired changes

in network output efficiently. For a classifier, the predicted class corresponds to the component with highest probability, i.e.,

$$label(X) = \arg\max_j F_j(X) \tag{2}$$

where $label(X)$ denotes the classification result of the DNN.

Precisely, the adversary aims to misclassify a sample X such that it is assigned as a target class $t \neq label(X)$. Therefore, the probability of class t given by F, i.e., $F_t(X)$, must be increased. On the contrary, the probabilities $F_j(X)$ of all other classes $j \neq t$ should be decreased, until $t = \arg\max_j F_j(X)$. The adversarial deception will terminate the step by increasing certain input features using the following saliency map $S(X,t)$,

$$S(X,t)[i] = \begin{cases} 0, \ if \ \dfrac{\partial F_t(X)}{\partial X_i} < 0 \ or \ \displaystyle\sum_{j \neq t} \frac{\partial F_j(X)}{\partial X_i} > 0 \\ \dfrac{\partial F_t(X)}{\partial X_i} \left| \displaystyle\sum_{j \neq t} \frac{\partial F_j(X)}{\partial X_i} \right|, \quad otherwise \end{cases} \tag{3}$$

where *i* denotes an input feature. If the input features have a negative target derivative or the overall sum of derivative on other classes is positive, they will be ignored. Otherwise, other positive derivative components are considered so that it is easy to compare $S(X,t)[i]$ for all input features. Through the saliency map, original signal can be changed in certain few points with large impact.

Further, the most significant point, i.e., (P_1, P_2) is calculated from the saliency map,

$$\arg\max_{(P_1,P_2)} \left(\sum_{i=P_1,P_2} \frac{\partial F_t(X)}{\partial X_i} \right) \times \left| \sum_{i=P_1,P_2} \sum_{j \neq t} \frac{\partial F_j(X)}{\partial X_i} \right| \tag{4}$$

Then add slight perturbations onto the original signal, with a limit on the dynamic range of original signal.

$$X^* = X + (P_1, P_2) * \theta * (\max X - \min X) \tag{5}$$

where X^* represents the adversarial deception signal. By tuning the hyperparameter θ, the DNN will misclassify the adversarial deception signal.

When the iterations reach the maximum number or the predicted label is the same as the pre-defined target label, iteration terminates. If the adversarial deception signals with perturbations are not classified correctly by the DNN, it is regarded as a successful adversarial deception.

To show the algorithm workflow more clearly, the deception process is summarized as shown in Figure 1.

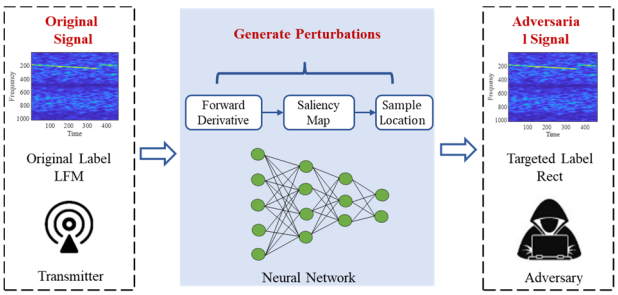


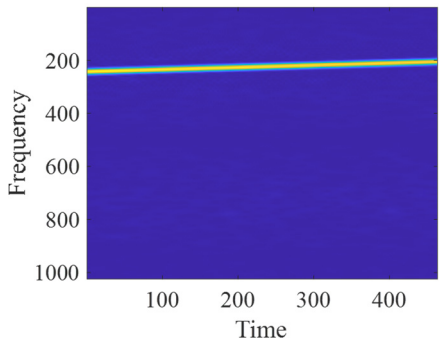
Figure 1. The workflow of the proposed adversarial deception against radio signal classification.

3 Experimental Results and Discussion

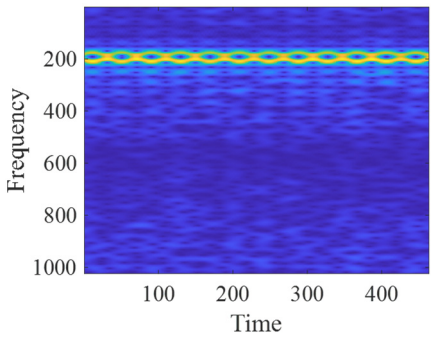
The premise of successful adversary deception is that some prior knowledge about the network architecture should be known. In this section, we test the performance of the proposed adversarial deception method using simulated data.

3.1 Simulated Dataset

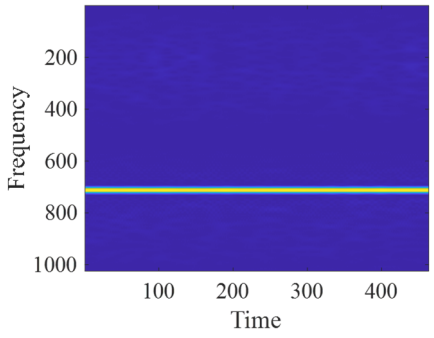
In this experiment, eight classes of radio signal are generated, including three radar waveforms, LFM, Barker, Rect and five communication waveforms, GFSK, CPFSK, B-FM, SSB-AM, DSB-AM. Each signal has 3000 realizations with the sampling points of $N = 1024$. The sample rate $F_s = 100$ MHz. The pulse width and repetition frequency are randomly generated for each waveform. Besides, each signal has unique parameters and is augmented with various impairments to make it more realistic. The radar waveforms are impaired with white Gaussian noise with a random signal-to-noise ratio (SNR) in the range of [-6, 30] dB and the step size of SNR is 5dB. A frequency offset with a random carrier frequency in the range of $[F_s/6, F_s/5]$ is applied to each signal. These radio signals are complex so they have real part and imaginary part. Figure 2 show the particular time-frequency spectrograms of three radar waveforms.



(a)



(b)



(c)

Figure 2. Spectrogram of various signals. (a) LFM (b) Barker and (c) Rect.

3.2 Deep neural network

Considering the dataset $X \in R^{24000\times2\times1024}$, we build a fully connected neural network to mimic deep-learning based radio waveforms classification system. After tuning the hyperparameters, the neural network has four fully connected layers, and the number of neurons for each layer is 128, 128, 64 and 8, respectively. The dataset is split into a test set and a training set with the proportion of 1:1. Dropout is used to avoid overfitting among the layers. The RELU activation function is used in the hidden layers, which effectively avoids the phenomenon of gradient disappearance and gradient explosion in the process of training the model. The architecture of radio signal classification network is illustrated as Figure 3.

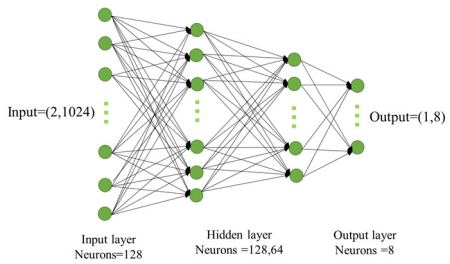
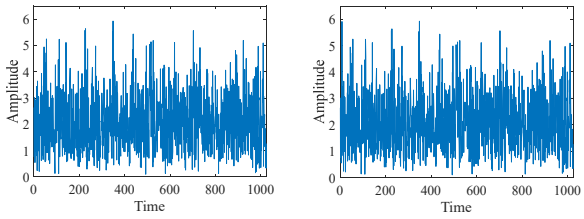


Figure 3. The architecture of radio signal classification network.

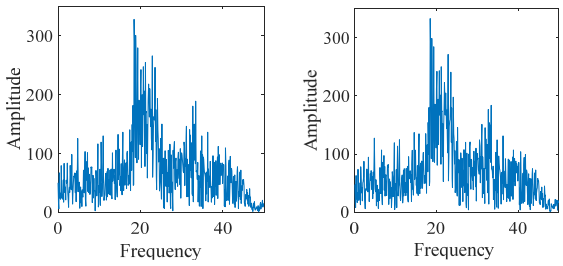
3.3 Results of Adversarial Deception

The experiment gives the result of two samples with SNR under -6dB and 29dB. Figure 4 - Figure 6 illustrate a particular sample with LFM radar waveform under SNR = -6dB. The target label is specified as Rect. It is shown that the difference between original transmitted waveform and deception waveform is slight from time-frequency domain.



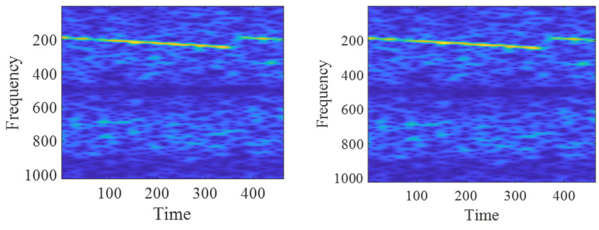
(a) (b)

Figure 4. The LFM waveform under -6dB in time domain. (a)The original transmitted waveform. (b)The modified waveform after adversarial deception.



(a) (b)

Figure 5. The spectrum of LFM waveform under -6dB. (a)The original transmitted waveform. (b)The modified waveform after adversarial deception.



(a) (b)

Figure 6. The LFM waveform under -6dB in time-frequency domain. (a)The original transmitted waveform. (b)The modified waveform after adversarial deception.

The radio signal classification network has an accuracy of 100% for both original signals under -6dB and 29dB. To evaluate the performance, two metrics are introduced to measure the accuracy of deception. The targeted accuracy

is defined as the percentage that the adversarial deception signal can be classified as the target label. At the same time, the original accuracy that adversarial deception signal can be classified as the original label should also be considered. Generally, the probability of targeted class should be greater than that of the original class. The experiment results are listed in TABLE I. It is shown that the accuracy of the original signal is degraded to a great level after deception, as much as 80% under 29dB. The original signal and adversarial deception signal's root mean square error (RMSE) of the real and imaginary parts are approximately [0.1691,0] and [0.1494,0.2084], respectively. Under low SNR condition, it is much easier to generate slight perturbation between the original signal and adversarial deception signal in the time-frequency domain.

TABLE I Results of the adversarial deception

Original Label	Targeted Deception Label	SNR [dB]	Targeted Accuracy	Original Accuracy
LFM	Rect	-6	63.05%	36.95%
		29	56.71%	20.22%

4 Conclusion

The safety of artificial intelligence in radio applications is an important issue to be noticed. In this paper, the vulnerability of the deep learning model under white-box adversarial deception is investigated. This algorithm only modifies a few sample points in the signal and does not need to deceive the classifier by disturbing the whole signal samples. It is shown that this algorithm can generate relatively slight perturbation between original signal and adversarial sample even under a low SNR. Such an adversarial sample enable adversary to subvert the excepted system behavior leading to undesired consequences, and poses severe risks to the intelligent identification systems when they are deployed in real world. For more practical situations, it is not easier for the adversary to obtain exact information of neural network or the dataset. Therefore, the black-box adversarial deception should be further discussed.

5 Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant No. 61801390, 61901377. This work is also supported by National Postdoctoral Program for Innovative Talents under grant BX201700199.

6 References

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, **521**, 7553, May 2015, pp. 436-444.

2. C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys & Tutorials*, **21**, 3, March 2019, pp. 2224-2287, doi: 10.1109/COMST.2019.2904897.

3. T. J. O'Shea, J. Corgan, and T. Charles Clancy, "Convolutional radio modulation recognition networks", *International conference on engineering applications of neural networks*, August 2016, pp. 213-226, doi: 10.1007/978-3-319-44188-7_16.

4. T. J. O'Shea, T. Roy, and T. Charles Clancy, "Over-the-air deep learning based radio signal classification", *IEEE Journal of Selected Topics in Signal Processing*, **12**, 1, January 2018, pp. 168-179, doi: 10.1109/JSTSP.2018.2797022.

5. Y. Yong, H. Pan, Z. Qile, and L. Xiaolin, "Adversarial examples: Attacks and defenses for deep learning", *IEEE Transactions on Neural Networks and Learning Systems*, **30**, 9, January 2019, pp. 2805-2824, doi: 10.1109/TNNLS.2018.2886017.

6. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D.Erhan, and I.Goodfellow, "Intriguing properties of neural networks", December 2013, arXiv preprint arXiv:1312.6199.

7. N. Akhtar, A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey". *IEEE Access*, **6**, February 2018, pp. 14410-14430, doi: 10.1109/ACCESS.2018.2807385.

8. S. K. Filipovic, R. Miller, "Adversarial examples in RF deep learning: Detection of the attack and its physical robustness", February 2019, arXiv preprint arXiv:1902.06044.

9. M. Sadeghi, E. Larsson, "Adversarial attacks on deep-learning based radio signal classification". *IEEE Wireless Communications Letters*, **8**, 1, August 2018, pp. 213-216, doi: 10.1109/LWC.2018.2867459.

10. N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, B.Celik, and A.Swami, "The limitations of deep learning in adversarial settings", *IEEE European symposium on security and privacy (EuroS&P),*May 2016, pp. 372-387, doi: 10.1109/EuroSP.2016.36.

11. I. J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial example", December 2014, arXiv preprint arXiv:1412.6572.