An ADS-B Signal Poisoning Method based on U-Net

Tianhao Wu,[1] Shunjie Zhang,[1] Jungang Yang,[1] and Pengfei Lei[2]
[1]*College of Electronic Science, National University of Defense Technology, Changsha, China*
[2]*Air Traffic Control and Navigation College, Air Force Engineering University, Xi'an, China*
Email: 735204780@qq.com

Automatic dependent surveillance-broadcast (ADS-B) has been widely used due to its low cost and high precision. The deep learning methods for ADS-B signal classification have achieved a high performance. However, recent studies have shown that deep learning networks are very sensitive and vulnerable to small noise. We propose an ADS-B signal poisoning method based on U-Net. This method can generate poisoned signals. We assign one of ADS-B signal classification networks as the attacked network and another one as the protected network. When poisoned signals are fed into these two well-performed classification networks, the poisoned signal will recognized incorrectly by the attacked network while classified correctly by the protected network. We further propose an Attack-Protect-Similar loss to achieve "triple-win" in leading attacked network poor performance, protected network well performance and the poisoned signals similar to unpoisoned signals. Experimental results show attacked network classifies poisoned signals with a 1.55% classification accuracy, while the protected network classifies rate is still maintained at 99.38%.

Introduction: Deep learning methods have achieved a considerable success in ADS-B signal classification task. LA Yun et al. [1] created a large-scale real-world radio signal dataset based on ADS-B signals. Weng L et al. [2] proposed a DRN model to achieve high accuracy in ADS-B signals classification. J. Robinson et al. [3] designed an augmented dilated causal convolution module for raw I/Q data and achieved an 85% accuracy for ADS-B signals classification without ID address.

However, some studies found that slight noise can cause deep neural networks (DNNs) mistakes in classification task [4] because these methods are highly data-driven. Jiawei Su et al. [5] attacked natural images by changing one pixel. The method was verified in the CIFAR-10 dataset and successfully deceived three different network models and caused 70.97% mistakes on the test image classification. S. Moosavi-Dezfooli et al. [6] proposed a Deepfool method to efficiently compute perturbations which can fool deep networks. M. Sadeghi et al. [7] proposed an adversarial attacks deep learning method on radio signal classification.

Inspired by data poisoning attacks [8], this letter proposes an ADS-B signal poisoning method based on U-Net. Our method contains a U-Net to generate poisoned signal. We assign one of ADS-B signal classification networks as the attacked network and the other as the protected network. When poisoned signals are fed into these two well-pretrained classification networks, the poisoned signal will classified incorrectly by the attacked network while classified correctly by the protected network. We further propose an Attack-Protect-Similar loss (APSL) to achieve 'triple-win' in leading attacked network poor performance, protected network well performance and the poisoned signals similar to unpoisoned signals. The contributions of our work can be summarized as follows:

- To our knowledge, our method is the first deep learning work in ADS-B signal poisoning task.
- The Attack-Protect-Similar loss we proposed can get 'triple-win' in leading attacked network poor performance, protected network well performance and poisoned signals similar to unpoisoned signals.
- Experimental results show the generated poisoned signals successfully cause attacked network poor performance and the protected network good performance.

The ADS-B encoding format: ADS-B signal data type is shown in Fig. 1. ADS-B signal data is consist of in-phase component (I-way) and the quadrature component (Q-way). The data type of the I/Q component is

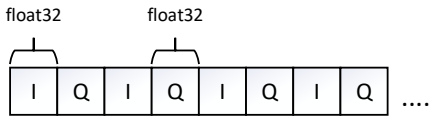


Fig 1 *An illustration of I/Q component in ADS-B signal.*

float 32. To ensure that the signal samples are abundant and balanced, we divide 200 ~ 600 signal samples into 40 categories. All samples are stored as H5 format files.

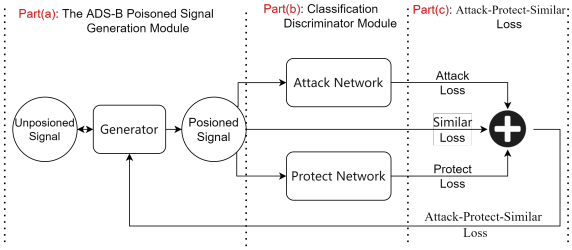


Fig 2 *Overall Architecture of our method. (a) The ADS-B poisoned signal generation module. (b) Classification discriminator module. (c) Attack-Protect-Similar loss.*

Overall Architecture: As shown in the Fig. 2, our method contains a generator and two discriminators and an Attack-Protect-Similar loss. The generator receives the unpoisoned signals and outputs the poisoned signals. Then, the classification discriminators classify the poisoned signals. To well train the generator, our Attack-Protect-Similar loss receives the unpoisoned signals and two discriminator classification results. Attack-Protect-Similar loss helps the generator to output poisoned signals which cause attacked network poor performance, protected network well performance and poisoned signals similar to unpoisoned signals.

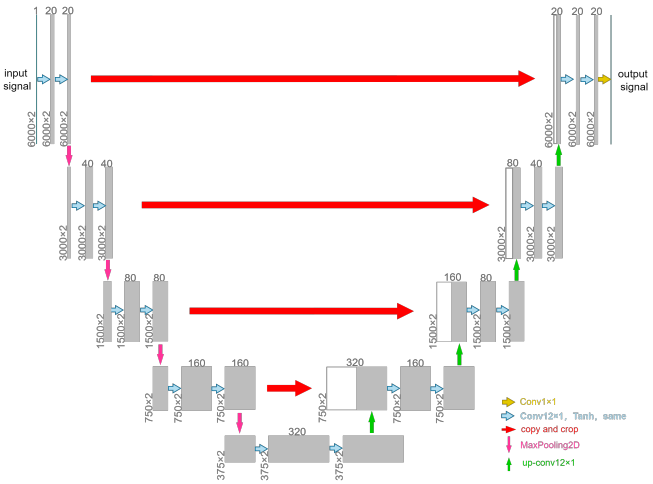


Fig 3 *An illustration of the U-Net. Each grey box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.*

U-Net for ADS-B Poisoned Signals generation: U-Net is widely used in image segmentation task [9]. Therefore, we select U-Net as the network of generator. As shown in the Fig. 3, the U-Net contains 4 downsampling blocks and 4 upsampling blocks to maintain the input and output signal same size. The convolution kernel sizes of U-Net are changed to fit the 2-dimensional signal input.

Table 1. Comparison of classification accuracy on unpoisoned signals and on poisoned signals. Note that, classification accuracy of the attacked network suffers a large decrease on unpoisoned signals, while classification accuracy of the protected network only suffers a little.

Generator	Classification accuracy on unpoisoned signal	Classification accuracy on poisoned signal	Similar loss
U-Net	CNN(99.46%) ResNet(99.23%)	CNN(98.53%) ResNet(8.04%)	0.2767

Attacked network and protected network: The ADS-B signal classification task has been widely studied and many deep learning based methods [1–3] have achieved a good performance. We choose an ADS-B signal classification method [2] based on CNN as the protected network and a method [2] based on ResNet as attacked network. The attacked network and protected network have been well trained and achieved a 99% accuracy on unpoisoned signal classification. The parameters of these two network have been frozen after training. Thus, both attacked network and protected network will keep the original structure and parameters when our U-Net for generating poisoned signals is on training.

Attack-Protect-Similar Loss: Attack-Protect-Similar loss is proposed to achieve 'triple-win' in leading attacked network poor performance, protected network well performance and the poisoned signals similar to unpoisoned signals. Attack-Protect-Similar loss contains three parts:

1. **Attack Loss**: Attack loss (AL) leads the generator to output signals which can not be classified correctly by the attacked network. The worse the performance of the attacked network, the smaller the loss. The AL function is as follows

$$AL = -\sum_{i=1}^{C} y_i \log f_a(x_i), \tag{1}$$

where $f_a(x_i)$ is output of attacked network, y_i is label of signal and C represents the number of categories.

2. **Protect Loss**: Protect loss (PL) leads the generator to output signals which can be classified correctly by the protect network. The better the performance of the protect network, the smaller the loss. The PL function is as follows:

$$PL = \sum_{i=1}^{C} y_i \log f_p(x_i), \tag{2}$$

where $f_p(x_i)$ is the output of protected network.

3. **Similar Loss**: Similar loss (SL) leads the generator to output the poisoned signals similar to unpoisoned signals. The more similar poisoned signals are to unpoisoned signals, the smaller the SL. The SL function is as follows:

$$SL = \frac{1}{l}\sum_{i=1}^{l} (s_i - \hat{s}_i)^2, \tag{3}$$

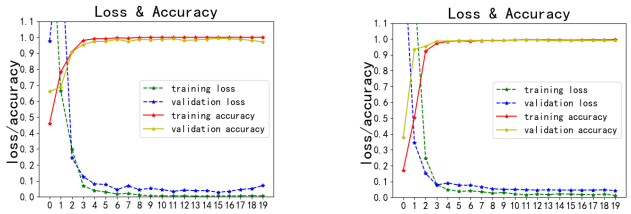
where l is the signal length, s_i represents a unpoisoned signal, $\hat{s}_i$ represents a poisoned signal.
Attack-Protect-Similar loss (APSL) is the final loss backward to the generator. α, β and γ are weights to balance the AL, PL and SL.

$$APSL = \alpha AL + \beta PL + \gamma SL, \tag{4}$$

where α, β and γ are adjustable weights.

Experimental results and analysis: In this section, we first trained discriminators. Then, we trained the generator to output poisoned signals. Finally, we evaluates the effectiveness of our poisoned signals and presented results in detail.

Discriminator parameter settings: The training and testing procedure of the method is deployed on the Linux using the TensorFlow 1.0. The model was trained and tested on the GPU of the RTX 1080Ti, supported by the GPU acceleration library. We divided the whole dataset into two non-overlapping parts including the training set (80% of the dataset) and the validation set (20%). The epoch was set as 20.



(a) *Attacked Network* **(b)** *Protected Network*

Fig 4 *Classification Accuracy and Loss on unpoisoned signal. Both CNN and ResNet can perform well on unpoisoned signal classification.*

Training of the Discriminator: We choose the ADS-B signal classification methods [2] based on a CNN and a ResNet as discriminators.

As Fig. 4 shown, both CNN and ResNet achieved a classification accuracy of over 99%. This shows that both our protected network and attacked network for discriminator can achieve a good performance on the unpoisoned signal. Afterwards we freeze the parameters of both networks, these parameters will not change during training the generator.

Generator parameter settings: The training and testing procedure of the method is deployed on the Linux using the TensorFlow 1.0. The model was trained and tested on the GPU of the RTX1080Ti, supported by the GPU acceleration library. We divided the whole dataset into two non-overlapping parts including the training set (80% of the dataset) and the validation set (20%). The epoch was set as 20. We set α as 35, β as 5 and γ.

Results: As Table 1 shown, poisoned signals generated by U-Net cause a decrease from 99.23% to 8.04% on classification accuracy of the attacked network and help protected network maintain a 98.53% of classification accuracy. That is because our method can lead attacked network poor performance and protected network well performance. The similar loss of poisoned signals and unpoisoned signals is only 0.2767. The small similar loss shows the generated poisoned signals are very similar to unpoisoned signals. This result proves that our method can poison ADS-B signals by adding only a little noise.

Conclusion: In this letter, we propose an ADS-B signal poisoning method based on U-Net. This method includes a generator, two discriminators and an Attack-Protect-Similar loss. Experimental results show poisoned signals can achieve 'triple-win' in leading attacked network poor performance, protected network well performance and the poisoned signals similar to unpoisoned signals.

Acknowledgments: This work was partially supported in part by the National Natural Science Foundation of China (Nos. 61972435, 61401474, 61921001, 62001478).

Received: DD MMMM YYYY *Accepted:* DD MMMM YYYY
doi: 10.1049/ell.10001

References

1. TU, Y., et al.: Large-scale real-world radio signal recognition with deep learning. Chinese Journal of Aeronautics 35(9), 35–48 (2022). doi:https://doi-org-s.libyc.nudt.edu.cn:443/10.1016/j.cja.2021.08.016
2. Weng, L., et al.: Message structure aided attentional convolution network for rf device fingerprinting. In: 2020 IEEE/CIC International Conference on Communications in China (ICCC), , pp. 495–500. (2020)
3. Robinson, J., et al.: Dilated causal convolutional model for rf fingerprinting. In: 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), , pp. 0157–0162. (2020)
4. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access 6, 14410–14430 (2018). doi:10.1109/ACCESS.2018.2807385
5. Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23(5), 828–841 (2019). doi:10.1109/TEVC.2019.2890858
6. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), , pp. 2574–2582. (2016)
7. Sadeghi, M., Larsson, E.G.: Adversarial attacks on deep-learning based radio signal classification. IEEE Wireless Communications Letters 8(1), 213–216 (2019). doi:10.1109/LWC.2018.2867459
8. Sun, G., et al.: Data poisoning attacks on federated machine learning. IEEE Internet of Things Journal 9(13), 11365–11375 (2022). doi:10.1109/JIOT.2021.3128646
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Cham: Springer International Publishing (2015)