Adversarial Learning in Transformer Based Neural Network in Radio signal classification

Lu Zhang, Sangarapillai Lambotharan, Gan Zheng

Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University

Abstract

- Recent studies discovered that the deep neural network is vulnerable to adversarial attacks in the sense that a carefully designed and imperceptible perturbation to the input of the neural network could mislead the prediction of the neural network.
- Motivated by attractive classification performance of the transformer based neural networks, we analyze the vulnerability and robustness of the transformer against adversarial attacks in modulation classification scenarios.
- Using real datasets, we demonstrate that the transformer can achieve higher accuracy as compared to a convolutional neural network in the presence of adversarial attacks.

Introduction

- In the past, AMC has been accomplished using various likelihood-based methods [1-2] and different machine learning methods based on **carefully chosen signal features** [3-4].
- Harnessing the power of DL, AMC can be achieved by training a deep neural network (DNN) [5] using **a large number of raw signal data samples** and generating classification decisions with high accuracy.
- Due to the great success of the transformers in both NLP and computer vision, **transformers** have been considered as a promising technique for **AMC** [6].
- However, recent studies discovered that the **adversarial example could deteriorate the performance of DNN** in many applications [7-8].

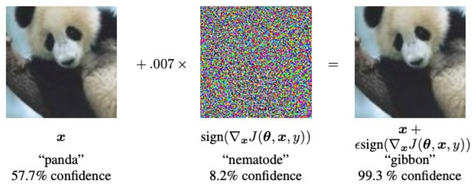

Fig.1: adversarial image generated by FGSM [9].

- In this work, we investigate **the robustness of transformer based neural network against adversarial examples in modulation classification**.

Methods

- The transformer architecture is shown in Figure 2.
- The key part of the transformer is called attention. An attention function, as shown in Figure 3, can be defined as a function that maps a query and a set of key-value pairs to an output.
- The matrix of outputs is calculated as: $Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where d_k denotes the dimension of queries and keys.

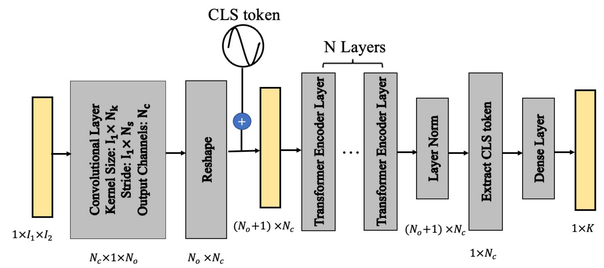

Fig. 2: The architecture of the transformer based neural network for the modulation classification.

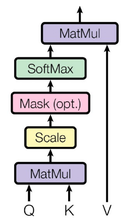

Fig.3: Scaled Dot-Product Attention [10]

- Given a trained DL classifier f and an original input data sample x, one can generate an adversarial example x' as a constrained optimization problem:

$$\min_{x'} \|x' - x\|_p, s.t., f(x') = l', f(x) = l, l \neq l'.$$

- The white-box PGD attack was adopted to generate adversarial examples.
- The objective function we used in this work is written as: $\psi(x) = s_y(x) - \max_{j \neq y} s_j(x)$
- The projection is applied after a standard gradient procedure: $x^* = x - \eta \nabla \psi(x)$.
- The projection procedure can be expressed as the following optimization: $\min_{x'} \|x' - x^*\|_2,\ s.t. \|x' - x\|_2 \leq \varepsilon$. And the solution to this optimization is modified as follows to force the l_2-norm of the generated perturbation equal to ε.

$$x' = x_0 + \frac{\varepsilon \cdot (x^* - x_0)}{\|x^* - x_0\|_2}$$

Results

- From Figure 4, we confirm that the transformer based neural network can obtain better classification accuracy than the CNN classifier in the absence of the adversarial perturbations.
- In Figure 5, as expected, the performance of both CNN and Transformer decreases significantly as the PNR increases. However, in the PNR region of -40dB to -10dB, the Transformer is able to maintain an approximately 10% performance advantage as compared to the CNN.

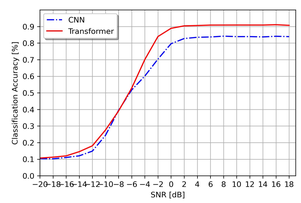

Fig.4: Classification accuracy for benign samples

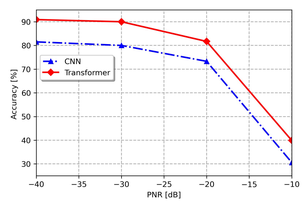

Fig.5: Classification accuracy for adversarial samples

Conclusions

- We have shown that in radio modulation classification tasks, even though the transformer based neural network is vulnerable to the PGD attacks, it is able to maintain the performance advantage over the ordinary CNN based modulation classifiers.
- For a wide range of PNR values and for moderate SNR, the Transformer provides approximately 10% more classification accuracy as compared to CNN.
- The focus of our future work is to enhance robustness of the Transformer against a wide range of adversarial attacks.

References

[1] A. Polydoros and K. Kim, "On the detection and classification of quadrature digital modulations in broad-band noise," IEEE Transactions on Communications, vol. 38, no. 8, pp. 1199–1211, 1990.

[2] B. F. Beidas and C. L. Weber, "Modulation classification of mfsk signals using the higher-order correlation domain," in Proceedings of MILCOM'95, vol. 1. IEEE, 1995, pp. 186–191.

[3] S. S. Soliman and S.-Z. Hsue, "Signal classification using statistical moments," IEEE Transactions on Communications, vol. 40, no. 5, pp. 908–916, 1992.

[4] L. Mingquan, X. Xianci, and L. Leming, "Cyclic spectral features-based modulation recognition," in Proceedings of International Conference on Communication Technology. ICCT'96, vol. 2. IEEE, 1996, pp. 792–795.

[5] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in International conference on engineering applications of neural networks. Springer, 2016, pp. 213–226.

[6] S. Hamidi-Rad and S. Jain, "Mcformer: A transformer based deep neural network for automatic modulation classification," in 2021 IEEE Global Communications Conference (GLOBECOM), 2021.

[7] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp.1369–1378.

[8] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in Proceedings of the 2016 acm sigsac conference on computer and communications security, 2016, pp. 1528–1540.

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

[10] [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

Contact

Lu Zhang: l.zhang6@lboro.ac.uk

Sangarapillai Lambotharan: s.lambotharan@lboro.ac.uk

Gan Zheng: g.zheng@lboro.ac.uk