Adversarial Examples in Wireless Networks: A Comprehensive Survey

JianShuo Dong[1], Xueluan Gong[2(✉)], and Meng Xue[2]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
jianshuo.dong@whu.edu.cn
[2] School of Computer Science, Wuhan University, Wuhan, China
{xueluangong,xuemeng}@whu.edu.cn

Abstract. With the rapid development of deep learning technologies, more and more DNN models are employed in wireless communication tasks. While DNNs improve the service, they also incur potential security threats from malicious users. Adversarial Example is a generally discussed attack targeting the deep learning-based models, which undoubtedly threatens the security of deep learning-based wireless networks. In this paper, we widely investigate adversarial example attacks in wireless network scenarios. It helps us become aware of the adversarial example threats that DNNs used in wireless networks are exposed to, and more efforts are required to defend against these attacks.

Keywords: Adversarial example attacks · Wireless networks · Deep learning

1 Introduction

The progress in deep learning technologies enables deep neural networks (DNNs) to get practical applications in the real world, such as image classification and speech recognition. In wireless communities, deep neural networks are deployed to tackle various tasks like signal classification and channel management as well. While it improves efficiency and provides convenient service, it also brings more potential security risks.

Extensive researches have demonstrated that deep neural networks are vulnerable to elaborately designed adversarial perturbations. In this case, malicious attackers generate an adversarial example like a patched image or a noise-injected audio recording as input to cause incorrect output or misclassification of the victim model. Attackers can craft an adversarial example by utilizing the gradient of a model's loss function through the backward propagation algorithm.

In wireless communication scenarios, the wireless medium is shared and open to the public, meaning that the base station (BS) is more easily to be attacked by malicious users. Traditionally, attackers can jam vast queries to block the shared channel. Similarly, as discussed in [7,8,15], deep neural networks deployed in the

K. Wu et al. (Eds.): ICECI 2021, LNICST 437, pp. 92–97, 2022.
https://doi.org/10.1007/978-3-031-04231-7_8

base station are likely to suffer from practical attacks as well. So adversarial example attacks can pose a practical threat to the wireless network scenario. Through disguising the waveform, wrong signals will be transmitted to the terminal station to spoof the target DNN model. Therefore, it is meaningful for us to widely investigate adversarial attack practices in wireless networks and conclude the commonly shared features of this series of adversarial attacks.

2 Background

2.1 Adversarial Example

Adversarial example attacks mislead the target neural network by adding a specific adversarial perturbation to the input. As a consequence, the victim model may give back an unexpected output, and even worse, it may fall apart beyond expectation. As described in [1,17], this kind of attack has been widely studied and got practical success in many domains, like computer vision [3] and natural language understanding. According to attack results, AE attacks can be divided into targeted attacks and untargeted attacks. Targeted attacks aim to make the victim model misclassify adversarial examples into one specific class. In face recognition scenario, people with fabricated face disguise will be classified into a specific identity. The untargeted ones, otherwise, only seek to misguide the victim model to categorize the adversarial examples into a wrong class. As summarized in [16], to tackle adversarial example attacks, some defense strategies have been proposed. For example, network distillation [12] and adversarial train [4] technologies can be utilized to defend against adversarial example attacks.

The paper [10] points out that with more DNNs deployed in the real world, there are more practical scenarios to implement adversarial examples attacks. Under the circumstances of wireless networks, DNNs are deployed in the base station for high execution performance. However, neural networks are naturally vulnerable to attacks like AE attacks, which brings more security risks.

2.2 Wireless Network

Wireless networks provide connection services between different users within a limited geographical range. In wireless networks, signals are transmitted from node to node through wireless medium and without a physical wire connection. As described in [11], there are two typical wireless connection architectures. One is ad hoc network, in which case no central access point is involved, and each node connects directly to its neighbor nodes. Manual configuration is required to establish an ad hoc network. The other is centrally coordinated network, in which end users query the access point to obtain remote connection service. After being granted access permission, end users only send data to the access point, and the access point is in charge of transferring the data to the terminal point. In the following contents, all attacks are illustrated under the setting of centrally coordinated network.

3 Adversarial Example Attacks in Wireless Networks

Deep neural networks may have different functions when deployed in different wireless systems. We take modulation classifiers as an example, which is enough to illustrate the features of adversarial examples in wireless networks. In our settings, one central access point act as a transmitter, and distributed nodes work as receivers. Each node is equipped with a well-trained classification model, which helps determine the modulation type that the central access point uses to encode the signal. The adversary aims to disguise the waveform of the over-the-air radio signal and misguide the classifiers further to incur tremendous consequences.

We categorize the previous works according to consideration of the distinct channel effect and broadcast nature in the wireless system scenario.

3.1 Consideration of Channel Effect

Different from other domains, the effect of channels will weaken the power of adversarial examples, and even worse, disable them. To implement an adversarial example in the realistic wireless systems, attackers are supposed to take channel effect into account.

Channel-Ignorant Attack. In [13], a white-box adversarial example generation algorithm is proposed. The authors utilize the fast gradient methods (FGM) [3] to optimize the adversarial example. They formulate such an attack scenario. The adversary adds a small perturbation r_x at the receiver's position when the central transmitter sends a wireless signal x to the receiver, so the signal received will be $x_{adv} = x + r_x$. The target of the attack is to craft a tiny perturbation r_x so that it leads to victim model's misclassification. Moreover, they also utilize the method of principal component analysis (PCA) to fabricate universal adversarial perturbations (UAP), which they implement experiments to demonstrate its validity. The disadvantage of their work is that they ignored the intrinsic channel effect in the wireless scenario and they modify the input by directly adding a perturbation, which is impractical in the realistic environment.

Channel-Considered Attack. The attack methods of [2,5,13] all focus on how to efficiently generate more powerful perturbations with limited knowledge about the network. But they fail to take the distinct channel effect of wireless scenario into consideration, in which way, they can never ensure that the perturbations injected have enough power to function well. In [6], the authors first show that the works without consideration of channel effect fail in the real world. And then, they accomplish a channel-considered attack by adding one restriction of power budget P_{max} to the optimization problem, in which way, they can fabricate an adversarial perturbation with both high adversarial performance and high power efficiency.

3.2 Consideration of Multi-target Attack

Attackers should never neglect the fact that samples in wireless networks are not directly fed to the DNNs, but are broadcast to the shared channel. This adds more possibility to success in attacking multiple receivers simultaneously.

Receiver-Specific Attack. In the previous works [2,5,9,13], adversarial example generation algorithms have been demonstrated efficient when the attack is carried out strictly. The shortcoming of these attacks is that the crafted adversarial examples are specific to target receivers and inefficient when used to attack other receivers. That is because models deployed in the receivers have specific gradient information and channel information.

Broadcast Attack. Spoofing one receiver among the whole network can only lead to the breakdown of one node. Considering the broadcast characteristic of wireless communication, it is more devastating and viable to attack multiple receivers at the same time in the network. Nevertheless, the different intensities of the channel effect make it difficult to implement. In [6], the authors proposed two solutions. The first one is that we can separately attack each of the m receivers and obtain corresponding m adversarial examples. We can finally get a weighted sum as the ultimate adversarial example. The authors suggest that the weight of each node can be determined by line search algorithm. The second but more rational approach is to take advantage of a jointly calculated loss function. This loss function score aggregates all the information we need to attack multiple receivers simultaneously.

4 Adversarial Example Attack Countermeasures

Inspired by the defenses used in other domains, like CV and NLP, several strategies have been introduced to improve deep learning-based wireless system's robustness to adversarial examples.

4.1 Adversarial Train

Similar to other domains, adversarial training strategy can be used to defend against the threats of adversarial examples. During the training process, we can add adversarial examples correctly labeled to the training set as a data augmentation method, which can enhance models' classification ability to a certain extent. However, when faced with adversarial examples created through a different method, the model will still be successfully attacked. More generally, a method called randomized smoothing can improve the model's anti-interference ability against adversarial examples. More specific, we emulate adversarial examples by augmenting the training set with randomized noise, which a certified defense.

4.2 Random Transmission Error

Conscious of the existence of adversaries, the transmitter can add the uncertainty of the deep learning-based wireless network by randomly taking wrong transmission actions, such as transmitting in a busy channel or not transmitting in an idle channel. It helps fool the adversary with wrong channel information and obstruct the generation and implementation of adversarial examples. Inevitably, this method will lead to performance reduction, so it is vital to find an appropriate balance between network performance and security guarantee [14].

4.3 Statistics-Based Detection

Proposed by [18], the detection method is a two-step approach, which utilizes the peak-to-average-power ratio (PAPR) of the radio frequency samples and the softmax output of the classifier to effectively detect adversarial examples. PAPR is a widely adopted metric used in wireless communication researches to illustrate the modulation type. Therefore, if the modulation classifier outputs one modulation label and meanwhile the PAPR represents a different label, the input will be suspected as an adversarial example and more tests should be performed. During the second detection stage, the suspicious sample's softmax logits are used to analyze if there exists a distribution shift caused by adversarial example noise. In this way, the receiver can decide to accept or reject the sample to avoid being attacked.

5 Conclusion

As discussed above, deep neural networks employed in wireless networks are likely to suffer from adversarial example attacks, which incurs potential security risks in the real world. Meanwhile, there are some effective defense strategies that have been proposed to address the adversarial example attack threats. However, none of them can perfectly defend against all kinds of adversarial example attacks without reducing transmission quality. Moreover, we point out that the transmission features of wireless networks like open channels have not been taken full advantage of yet. Therefore, more efforts are in demand to put forward more aggressive attack schemes to find out the potential security risks hidden in wireless networks. It is also crucial to devise more robust models to invalidate adversarial example attacks while ensuring high quality of service.

References

1. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: Adversarial attacks and defences: a survey. arXiv preprint arXiv:1810.00069 (2018)
2. Flowers, B., Buehrer, R.M., Headley, W.C.: Evaluating adversarial evasion attacks in the context of wireless communications. IEEE Trans. Inf. Forensics Secur. **15**, 1102–1113 (2019)

3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
4. Huang, R., Xu, B., Schuurmans, D., Szepesvári, C.: Learning with a strong adversary. arXiv preprint arXiv:1511.03034 (2015)
5. Kim, B., Sagduyu, Y.E., Davaslioglu, K., Erpek, T., Ulukus, S.: 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), pp. 1–6. IEEE (2020)
6. Kim, B., Sagduyu, Y.E., Davaslioglu, K., Erpek, T., Ulukus, S.: Channel-aware adversarial attacks against deep learning-based wireless signal classifiers. IEEE Trans. Wirel. Commun. (2021)
7. Kim, B., Sagduyu, Y.E., Erpek, T., Ulukus, S.: Adversarial attacks on deep learning based mmwave beam prediction in 5G and beyond. arXiv preprint arXiv:2103.13989 (2021)
8. Kim, B., Shi, Y., Sagduyu, Y.E., Erpek, T., Ulukus, S.: Adversarial attacks against deep learning based power control in wireless communications. arXiv preprint arXiv:2109.08139 (2021)
9. Kokalj-Filipovic, S., Miller, R., Morman, J.: Targeted adversarial examples against RF deep classifiers. In: Proceedings of the ACM Workshop on Wireless Security and Machine Learning, pp. 6–11 (2019)
10. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
11. Nazir, R., Kumar, K., David, S., Ali, M., et al.: Survey on wireless network security. Archiv. Comput. Methods Eng. 1–20 (2021)
12. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP), pp. 582–597. IEEE (2016)
13. Sadeghi, M., Larsson, E.G.: Adversarial attacks on deep-learning based radio signal classification. IEEE Wirel. Commun. Lett. **8**(1), 213–216 (2018)
14. Sagduyu, Y.E., Shi, Y., Erpek, T.: IoT network security from the perspective of adversarial deep learning. In: 2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 1–9. IEEE (2019)
15. Sagduyu, Y.E., et al.: When wireless security meets machine learning: motivation, challenges, and research directions. arXiv preprint arXiv:2001.08883 (2020)
16. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: attacks and defenses for deep learning. IEEE Trans. Neural Networks Learn. Syst. **30**(9), 2805–2824 (2019)
17. Zhang, J., Li, C.: Adversarial examples: opportunities and challenges. IEEE Trans. Neural Networks Learn. Syst. **31**(7), 2578–2593 (2019)
18. Kokalj-Filipovic, S., Miller, R., Vanhoy, G.: Adversarial examples in RF deep learning: detection and physical robustness. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1–5. IEEE (2019)