A Survey of Adversarial Attacks on Wireless Communications

Xiangyu Luo[1], Quan Qin[1], Xueluan Gong[2(✉)], and Meng Xue[2(✉)]

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
{ericlaw,2019302180109}@whu.edu.cn
[2] School of Computer Science, Wuhan University, Wuhan, China
{xueluangong,xuemeng}@whu.edu.cn

Abstract. As the deep neural network (DNN) has been applied in various fields in wireless communications, the potential security problems of DNNs in wireless applications have not been fully studied yet. In particular, DNNs are highly vulnerable to malicious disturbance, which opens up opportunities for a small scale of adversarial attacks to cause chaos in the model's performance. This paper enumerates the main over-the-air attack mechanisms that threaten a wide range of existing defenses. For each type of attack, we introduce the working principle and list some of the latest applications in different wireless communication fields. With the threats of various attacks to a wide range of existing defenses shown, we hope to raise awareness of the lack of novel defense mechanisms.

Keywords: Wireless communication · Adversarial attack · Deep neural network · Machine learning

1 Introduction

In recent years, Deep Learning (DL) has greatly benefited from advances in computational resources and algorithmic designs, which can be leveraged to help with data processing and complicated calculating tasks. On account of underlying channels, potential interference, traffic effects, and interactions of network protocol, DL has been applied to wireless systems in various sub-fields. For instance, spectrum sensing, interference management, waveform design, and signal classification [1].

However, adversaries are able to tamper with the model by manipulating the input, algorithm, and training process of Machine Learning. Although expressly, it is acknowledged that Deep Neural Networks (DNNs) are highly vulnerable to malicious disturbance, a small-scale adversarial attack may cause chaos in the performance of the model [2], as first mentioned in the computer vision field [3]. Moreover, due to the nature of broadcast and sharing of wireless communications, the model of which is more likely to be perturbed. Thus, we will be capable of guarding against most adversaries by making safety adoption with full knowledge of the attack methods.

In this paper, we discuss five primary wireless attacks. We introduce the working principle of each type of wireless attack and list some of the latest applications in different

K. Wu et al. (Eds.): ICECI 2021, LNICST 437, pp. 83–91, 2022.
https://doi.org/10.1007/978-3-031-04231-7_7

sub-fields of wireless communication. The main principle and features of each type of over-the-air attack are listed in Table 1.

Table 1. Different types of over-the-air attacks and the features

Attack type	Principle	Attack mechanism	Attacker's access	Innovation
Evasion attacks	Perturb an ML model at inference time	Targeted evasion attack	White-box	Masquerade as a specific signal
		Untargeted evasion attack	Black-box	Low power requirement
		Physical evasion attack	Black-box	More destructive, on autoencoder system
Poisoning attacks	Pollute the training dataset	LEB attack	Black-box	Effective on fusion center, generic
		Over-the-air spectrum data poisoning attack	Black-box	Fast, hard to detect
Trojan attacks	Insert Trojans to training data	Wireless signal classification Trojan attack	All	Modify few data, practical, work on wireless signal classifier
		BadNet	All	Powerful, generic
Spoofing attacks	Using a GAN to generate and transmit synthetic signals	GAN-based wireless signal spoofing	Black-box	GAN-based, high success probability, no need for prior knowledge
		Replay spoofing attack	Black-box	Keep some features in original signals
Inference attacks	Steal the training data and fool the target model using the data	Over-the-air MIA	Black-box	Effective, first over-the-air MIA attack

2 Attack Mechanisms

2.1 The Evasion Attacks

The evasion attacks perturb the ML model at inference time by subtle modifications on the original data. It happens only when adversarial examples are used to feed the

network. Adversarial examples are carefully perturbed input that looks the same as its untampered copy but with slight noise added and successfully fool the classifier of the network [4].

Current research demonstrated that DNNs are vulnerable to over-the-air adversarial evasion attacks, which significantly lower the accuracy of wireless communication tasks with only tiny modifications on the underlying transmission to a cooperative receiver [5].

Targeted Evasion Attacks
In [5], Samuel and Matthew focus on targeted evasion attacks to spectrum sensing that seek to disguise as a specific signal. They use an eavesdropper that uses a DL-based Automatic Modulation Classification (AMC) system that classifies and intercepts wireless signals when necessary. They use the RML2016.10A dataset and train the DNN model on AM-SSB. They adapt the MI-FGSM attack to RFML creating untargeted adversarial samples to compare untargeted and targeted attacks further. They find that the energy required to succeed in a targeted attack is related to the hierarchical relationship. They also draw the conclusion that the difficulty of targeting modulation schemes was confirmed.

Untargeted Evasion Attacks to Radio Signal Classification
The work presented in [6] is based on a DNN architecture using the RML2016.10A dataset. They used eavesdroppers but pre-assume that the eavesdroppers are capable of intercepting the processing chain without affecting the channel simultaneously. The authors adapt FGSM and Universal Adversarial Perturbations (UAP) to show the time-independent black-box results. They regard it as a limitation of FGSM. They also use the energy ratios of the perturbation signals as an attack limitation, just like what [4] did. Therefore, they only consider attacks with direct access to the classifier without transmitted OTA.

End-to-end Autoencoder Communication Systems
Recent research shows that end-to-end autoencoder communication systems through DNNs have a significant potential vulnerability to physical evasion attacks that result in more errors at the receiver [7]. The authors present algorithms to elaborate ways to craft effective physical black-box adversarial attacks. They finally conclude that the broadcast nature of wireless communication channels opens up great opportunities for adversary transmitters to increase the block-error rate of a communication system by well-designed perturbation signals over the channel. According to the result above, the authors put forward a possible defense assumption of adversarial training that trains the autoencoder with adversarial perturbations to increase robustness. Nevertheless, it would reduce the performance of the autoencoder and may be steered through newly designed adversarial perturbations.

2.2 The Poisoning Attacks

A poisoning attack aims to pollute the training dataset in order to produce a lousy classifier. First proposed by Barreno and Nelson [8], poisoning attacks have affected malware

detection [9], collaborative filtering systems [10], face recognition [11], automatic driving [12], medical insurance [13], loan evaluation [14], and various other application scenarios.

Poison attacks are divided into two types depending on the object to poison-the data poisoning and model poisoning. Data poisoning mainly refers to combining the poison data with the original sample and feeding the polluted input to the model to generate a backdoor during the test time. Model poisoning mainly refers to offering a poisoned model to the users directly.

Over-the-Air Spectrum Sensing Data Poisoning Attacks

In recent research of spectrum data poisoning [15], the authors introduced an adversarial ML approach to complete a spectrum data poisoning attack by capturing the behavior of the transmitters and simulating the wireless spectrum sensing data. Furthermore, they propose a new type of poisoning attack based on adversarial ML called an *over-the-air spectrum data poisoning attack*. It intercepts the spectrum sensing data and seeks to manipulate it in order that wrong transmits decisions are made using unreliable spectrum sensing results.

The authors show the efficiency of such an attack since only little time is needed to manipulate the transmitter's decisions. Additionally, the short transmission makes it hard to detect such an attack. The results show that the proposed spectrum poisoning attack is more energy-efficient and more covert compared to jamming of data transmissions, which calls for awareness of new defense mechanisms in protecting wireless communications against such attacks.

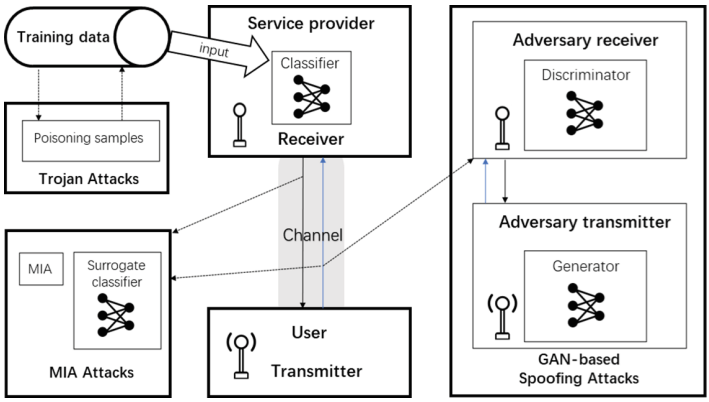

Fig. 1. Trojan attacks, MIA attacks, and GAN-based spoofing attacks in a wireless communication environment.

Learning-Empowered Poisoning Attacks

Zhengping Luo and Shangqing Zhao etc. proposed a new data poisoning mechanism in recent research [16]. They point out that a wide range of existing defense mechanisms tends to assume network or attackers as passive entities. For instance, A defender may assume that the prior attacks are known or solved. However, adversaries can

adopt arbitrary behaviors to avoid pre-assumed patterns to trounce the defense strategies. Thus, they propose a learning-empowered poisoning attack framework called Learning-Evaluation-Beating (LEB) to mislead the fusion center.

They attempt to make malicious use of ML to build a surrogate model of the fusion center's decision model, based on the black-box nature of the fusion center in spectrum data sensing. In order to create malicious sensing data, they put forward an algorithm using the surrogate model.

The results show that the LEB attacks reach a success probability of up to 82%, proving that the LEB offers an effective attack paradigm against cooperative spectrum sensing to some extent. Furthermore, the authors designed a mechanism called influence-limiting defense to couple with the LEB and other attacks with similar mechanisms.

2.3 The Trojan Attacks

Trojan attacks on wireless signal classification target DL applications [17], which use a DL classifier to classify signals (i.e., I/Q samples) with modulation types as labels, and only specific types can obtain authorization. However, unlike Trojan attacks on neural networks [18], this attack algorithm does not require access to the original model. Instead, only a small amount of manipulation of the training data is enough to implant the Trojan. And then, in the test (inference) phase, an adversary can make the model correctly classify clean input but perform not reliably on signals with Trojan triggers so that the thief can bypass the defenses and gain user authentication.

There are two periods of time to execute the attack, the training time and the test time. An adversary needs to select a label as the target label in the first step in the training time. In the second step, the adversary has to poison some non-target labeled data but simultaneously keeps the same amount of clean data for each non-target label in the training data, as presented in Fig. 1. The specific operation of poisoning is to select samples from training data randomly, rotate the samples by a specific angle θ, and finally change the non-target labels in these data to the target label. The third step is to replace the original clean samples with the poisoned samples.

The attacker uses the modulated signal from a non-target label to transmit the poisoned samples in the testing time. If the receiver classifies the received signal as the target label, the Trojan attack was successful. However, it should be noted that the value of θ can be small, so the signal-to-noise ratio does not change much, while the confidence level of the SNR estimate for small samples should be below, so the Trojan attack will not necessarily be detected by the receiver's method of checking the received SNR.

Ultimately, the Trojan attack only infects 10% of the training samples to achieve 90% attack accuracy at all SNRs. Also, for clean samples, the classification accuracy is very close to the accuracy before the implantation of the Trojan. However, this attack's changes to the training data result in outliers in the training data, so clustering-based outlier detection is effective in detecting the poisoned samples in the data and thus detecting Trojan attacks [19]. Also, according to the work of [18], the Trojan attack misleads the classifier to the same specific classification. Thus the statistics of the distribution of misclassification results should reveal that the misclassification results are mainly concentrated on one class. Therefore, another possible defense is to check the

misclassification results, which will give a classification that accounts for most of the results for a Trojan attack.

2.4 The Membership Inference Attacks

A member inference attack is a type of inference attack that aims to steal training data from a target model [20]. Meanwhile, MIA for wireless signal classification allows an adversary to infer the privacy of a wireless signal, which may include wavelength, waveform, channel, and device information [21]. In an area with substantial promiscuous users, a service provider can use a classifier to classify users when performing physical layer authorization verification, and this classifier accepts signals received by the provider [22, 23]. An adversary can launch an MIA against the target classifier in this environment to confirm whether a signal belongs to the classifier's training data. Even if the adversary is in the case of a white-box attack, he cannot directly use the model available to him to determine whether a signal is from an authorized user because the signal received by the adversary will differ from the one received by the service provider. Therefore, consider directly the case of an adversary using a black-box attack, where the adversary eavesdrops on the signal and uses it as an input to a surrogate classifier built by the adversary. It is presented in Fig. 1 that the adversary can launch an MIA with this surrogate classifier to determine whether the signal received by the service provider, which corresponds to the signal eavesdropped by the adversary, is used as the training data for the target classifier.

Due to the openness of the electromagnetic environment, an adversary can collect the signals sent by the user and returned by the service provider and eavesdrop on the classification results. The adversary first needs to determine the classification of signals based on this information and build a surrogate classifier [24]. Next, the adversary uses MIA to determine whether a signal sample is used for training and if so, the signal sample leaks information about the authorized user's device, waveform Etc., which the adversary can then use to perform other attacks such as forging similar signals and gaining access.

The first of two different settings is that the signals, both member and non-member, are generated by the same device, and the second is that different devices generate the signals of non-members while the same device generates the signals of members. Eventually, the accuracy of MIA, i.e., the accuracy of predicting whether the sample is a member or a non-member, is high in the first setting, but this is when the sample signal is strong, and there is no noise. However, the accuracy of the MIA decreases significantly as soon as the sample signal weakens or noise is added. In the second setting, the accuracy of MIA and the conditions of use are generic.

The defense is mainly for the second setting on account of the low probability of success in the first setting whether exist defenses. A very effective defense method is for the service provider to create a shadow MIA and take the data to train it to get a very high attack accuracy, and then calculate the noise that can disable the shadow MIA, i.e., make this MIA much less accurate, and this noise can also be used to disable the adversary's MIA.

2.5 The Spoofing Attacks

A spoofing attack is an attacker's attempt to imitate a legitimate user in a communication. A standard method of this attack is called a replay spoofing attack [25], in which the transmission signal of a legitimate user is recorded in advance, and then the signal is replayed with the potentially altered transmission power. Although this form of attack can portray various signal characteristics, it simply records the user's transmission and is unable to imitate the combined effects formed by factors such as environment, channel, and equipment, and it also already has a detection method [26]. A better performing attack than this is the GAN-based spoofing attack [27].

GAN-based deception attacks are trained from the adversarial perspective with a generator and a discriminator, where the generator is used to generate signals that can be used in spoofing attacks, and the discriminator is used to train against the generator. Suppose that a system based on a GAN spoofing attack has four components, the transmitter T, the receiver R (a classifier that classifies the received signals as coming from T or not coming from T), the adversary transmitter AT, and the adversary receiver AR, as seen in Fig. 1. AR is located close to R so that the transmitter-to-receiver channel will be relatively similar, and AT's signal transmission will be flagged so that AR can specify which signals are sent by AT. In this case, AR acts as a discriminator to identify the signals sent by AT or T and feeds the classification result to AT. AT acts as a generator to generate signals closer to those sent by T to fool the discriminator AR. After the GAN converges, the generator AT can generate signals so similar to T that the receiver R cannot distinguish whether the signals come from AT or T, and the adversary can thus perform a spoofing attack. It can be seen that this attack does not require any prior knowledge because AR can learn the features of T by itself. Also, AR and AT will learn various channel effects during the adversarial process, and thus the attacker does not need to learn these.

GAN-based spoofing attacks' success rate is 76.2%, which is much higher than the 7.89% of random signal and 36.2% of replay spoofing attacks. The defense scheme for this type of attack needs to be further researched.

3 Conclusions

This paper enumerates five main over-the-air attack mechanisms that threaten a wide range of existing defenses. We also show some of the latest attack models and their applications, raising awareness of the lack of novel defense mechanisms. However, As the deep neural network (DNN) has been applied in various fields in wireless communications. However, the potential security problems of DNNs in wireless applications have not been fully studied yet.

References

1. Erpek, T., O'Shea, T., Sagduyu, Y., Yi Shi, T., Clancy, C.: Deep learning for wireless communications. In: Pedrycz, W., Chen, S.-M. (eds.) Development and Analysis of Deep Learning Architectures, pp. 223–266. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-31764-5_9

2. Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 86–94 (2017)
3. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
4. Flowers, B., Buehrer, M.R., Headley, W.C.: Evaluating Adversarial Evasion Attacks in the Context of Wireless Communications. arXiv preprint arXiv:1903.01563 (2019)
5. Bair, S., DelVecchio, M., Flowers, B., Michaels, A.J., Headley, W.C.: On the limitations of targeted adversarial evasion attacks against deep learning enabled modulation recognition. In: WiseML@WiSec, pp. 25–30 (2019)
6. Sadeghi, M., Larsson, E.G.: Adversarial attacks on deep-learning based radio signal classification. IEEE Wirel. Commun. Letters **8**, 213–216 (2018)
7. Sadeghi, M., Larsson, E.G.: Physical adversarial attacks against end-to-end autoencoder communication systems. IEEE Commun. Lett. **23**(5), 847–850 (2019)
8. Barreno, M., Nelson, B., Sears, R., et al.: Can machine learning be secure? In: The 2006 ACM Symposium on Information, Computer and Communications Security, pp. 16–25 (2006)
9. Chen, S., Xue, M.H., Fan, L.L., et al.: Automated poisoning attacks and defenses in malware detection systems: an adversarial machine learning approach. Comput. Secur. **73**, 326–344 (2018)
10. Li, B., Wang, Y., Singh, A., et al.: Data poisoning attacks on factorization-based collaborative filtering. In: Advances in Neural Information Processing Systems, pp. 1885–1893 (2016)
11. Chen, X., Liu, C., Li, B., Lu, K., et al.: Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint arXiv:1712.05526 (2017)
12. Li, K., Mao, S.G., Li, X., et al.: Automatic lexical stress and pitch accent detection for L2 english speech using multi-distribution deep neural networks. Speech Commun. **96**, 28–36 (2018)
13. Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., et al.: Systematic poisoning attacks on and defenses for machine learning in health-care. IEEE J. Biomed. Health Inform. **19**(6), 1893–1905 (2015)
14. Jagielski, M., Oprea, A., Biggio, B., et al.: Manipulating machine learning: Poisoning Attacks and countermeasures for regression learning. In: 2018 IEEE Symposium on Security and Privacy (SP), pp. 19–35 (2018)
15. Shi, Y., Erpek, T., Sagduyu, Y.E., Li, J.H.: Spectrum Data Poisoning with Adversarial Deep Learning. CoRR abs/1901.09247 (2019)
16. Luo, Z., Zhao, S., Lu, Z., Xu, J., Sagduyu, Y.E.: When Attackers Meet AI: Learning-empowered Attacks in Cooperative Spectrum Sensing. CoRR abs/1905.01430 (2019)
17. Davaslioglu, K., Sagduyu, Y.E.: Trojan attacks on wireless signal classification with adversarial machine learning. In: 2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), pp. 1–6 (2019). https://doi.org/10.1109/DySPAN.2019.8935782
18. Liu, Y., Ma, S., Safer, Y., et al.: Trojaning attack on neural networks. In: Network and Distributed System Security Symposium 2017, pp. 1–15 (2017)
19. Chen, B., et al.: Detecting backdoor attacks on deep neural networks by activation clustering. In: AAAI SafeAI (2019)
20. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy, pp. 3–18 (2017). https://doi.org/10.1109/SP.2017.41
21. Shi, Y., Sagduyu, Y.E.: Membership inference attack and defense for wireless signal classifiers with deep learning. arXiv preprint arXiv:2107.12173 (2021)
22. Wang, N., Jiang, T., Lv, S., Xiao, L.: Physical-layer authentication based on extreme learning machine. IEEE Commun. Lett. **21**(7), 1557–1560 (2017). https://doi.org/10.1109/LCOMM.2017.2690437

23. Shi, Y., Davaslioglu, K., Sagduyu, Y.E., Headley, W.C., Fowler, M., Green, G.: Deep learning for RF signal classification in unknown and dynamic spectrum environments. In: 2019 IEEE International Symposium on Dynamic Spectrum Access Networks, pp. 1–10 (2019). https://doi.org/10.1109/DySPAN.2019.8935684
24. Yi, S., Sagduyu, Y., Grushin, A.: How to steal a machine learning classifier with deep learning. In: 2017 IEEE International Symposium on Technologies for Homeland Security, pp. 1–5 (2017). https://doi.org/10.1109/THS.2017.7943475
25. Kinnunen, T., et al.: The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection (2017)
26. Hoehn, A., Zhang, P.: Detection of replay attacks in cyber-physical systems. In: 2016 American Control Conference, pp. 290–295 (2016). https://doi.org/10.1109/ACC.2016.7524930
27. Shi, Y., Davaslioglu, K., Sagduyu, Y.E.: Generative Adversarial Network for Wireless Signal Spoofing (2019)