Adversarial Example Attacks in Internet of Things (IoT)

Yuzhe Gu[1], Na Jiang[1], Yanjiao Chen[2], and Xueluan Gong[3](✉)

[1] School of Cyber Science and Engineering, Wuhan University, Wuhan, China
{yuzhegu,na.jiang226}@whu.edu.cn
[2] College of Electrical Engineering, Zhejiang University, Hangzhou, China
chenyanjiao@zju.edu.cn
[3] School of Computer Science, Wuhan University, Wuhan, China
xueluangong@whu.edu.cn

Abstract. Recently, the Internet of Things (IoT) technology has made tremendous progress, and it is beginning to enter many areas of social life, such as autonomous driving, medical care, etc. Due to the massive data in IoT, deep neural networks (DNN) are often involved in helping process and analyzing data, but DNNs still face many security threats. Adversarial example attack is a common attack against DNN models, which interferes with model decisions through processed samples. It will undoubtedly threaten DNN-based IoT systems. This paper presents the possible attack scenarios of adversarial example attacks in IoT systems and extensively studies the defense methods of adversarial example attacks in IoT systems.

Keywords: Adversarial Example Attacks · IoT · Deep Learning

1 Introduction

In the past few years, the field of IoT is developing well, and the number of IoT devices has also exploded. According to statistics, by 2025, the number of IoT devices connected to the global network will increase to more than 7.544 billion units [6,46]. At the same time, the IoT is gradually covering all areas of social life, such as automation, health, transportation, energy, manufacturing, and other industries [9,10,22,24,25,29,42–44]. It also has produced practical applications such as autonomous driving and smart cities [41]. Obviously, for the IoT, which is already closely related to life, it is very important to ensure its security. However, the current IoT system still faces many security threats. Among them, the security threat caused by the vulnerability of deep neural networks is an important part that needs to be solved urgently.

In recent years, deep learning has played an important role in multiple fields, including pattern recognition [13], face recognition [36], speech recognition [2], and autonomous driving [21,32]. Of course, because of its ability to analyze a

Z. Xiao et al. (Eds.): ICECI 2022, LNICST 478, pp. 229–235, 2023.
https://doi.org/10.1007/978-3-031-28990-3_16

large amount of data, it is also used in IoT field, where abundant data is generated every day. Nevertheless, many works have verified the vulnerability of DNN in handling adversarial operations [7,11,15–19,30,31,33,38], for example, adversarial examples have the ability to confuse DNN models by slightly changing the network input data [38]. Therefore, the DNN in the current IoT scenario urgently needs high robustness to efficiently and accurately process IoT data of different accuracy [8].

In this paper, we describe the attack scenarios of adversarial example attacks in IoT. And we study the current representative works on improving the reliability of DNN to adversarial example attacks. Those works can well handle adversarial example attacks in IoT scenarios.

2 Background

2.1 Internet of Things (IoT)

IoT is a network infrastructure consisting of various sensing, communication, networking, and information processing devices [39]. Its main structure consists of infrared sensors, radio frequency identifiers, laser scanners, GPS, and other information equipment. Nowadays, IoT is widely used in smart cities, autonomous driving, health care, and other fields [1,23,28,35,37].

In the IoT system, all kinds of devices can access the IoT according to the protocols and standards formulated by the industry and realize the exchange and communication of information in the network system. Because the system is too large, a lot of data is generated and waiting to be processed every day, so DNN is often incorporated into IoT decision-making, which also gives attackers opportunities. The entire IoT roughly consists of three parts: a perception layer such as receiving data, a network layer that exchanges data, and an application layer that processes data. The adversarial example attack discussed in this paper mainly uses carefully designed sample input from the perception layer to achieve the effect of misjudging the DNN at the application layer.

2.2 Adversarial Example

The adversarial example is the original sample with invisible perturbation added, which misleads the deep neural network model to make a wrong judgment [38]. In the experiments of [20], for the original sample of the panda, after adding the adversarial perturbation, the judgment of the model changed from a high confidence “panda” to a “gibbon”.

The high complexity of DNN models has resulted in a variety of different hypotheses for adversarial examples at present. [20] argues that the high-dimensional linearity of neural networks is the leading cause of adversarial examples and that the small changes in the input data can lead to decision errors after being amplified by multiple layers of the network; [34] argues that there exists a low-dimensional subspace containing a large number of normal vectors at the

decision boundary, and perturbations within this subspace have an important impact on the model decisions. Multiple hypotheses have their focus, advantages, and disadvantages, which provide room for adversarial example attacks.

Recently, adversarial sample attacks have been successful in scopes such as CV and NLP [4,20,45]. Meanwhile, more attack scenarios have emerged in the real world [26]. In the era of IoT, a large number of perceptrons are collecting and generating data involving various domains daily. The involvement of DNN can help analyze and process a large amount of data in the IoT domain, but this also brings security analysis that may be subject to adversarial sample attacks.

3 Adversarial Example Attacks Scenario in IoT

In IoT system, DNNs use data gathered from IoT devices (perception layer), trained in a supervised or unsupervised manner, and apply the results to specific applications (application layer) to guide their decision-making behaviors [5]. However, in real-world scenarios, IoT devices may be attacked, destroyed, and tampered with, which in turn affects the data they collect and generate. As described in Sect. 2, adversarial example attacks fool the DNN model by modifying the input data to make wrong judgments, which will greatly affect the reliability of the DNN model. In some IoT application scenarios, there are extremely high requirements for the reliability of DNN models, such as pedestrian detection in autonomous driving [12], diagnostic opinion judgment of medical auxiliary detection equipment, process control in industrial production, etc. Adversarial example attacks will pose a great threat to these IoT application scenarios.

4 Adversarial Example Attack Countermeasures

In this section, we discuss some state-of-the-art studies which focus on improving the reliability of DNN-based IoT systems. For the convenience of introduction, the following discussion will not overemphasize the specific deployment and use of these methods in IoT systems but will focus on improving the robustness of the DNN model.

4.1 Learning with Reject Option

Learning with reject option (LRO) [3] is a special training method that can make the reliability of DNN achieve a better effect when the performance of standard models cannot be guaranteed. Different from the traditional optimization method that optimizes the general accuracy of all examples, LRO selects the subset with better performance from the example set and leaves the rest to judgments such as manual processing to make the average prediction accuracy high enough. For example, for a medical detection system whose prediction accuracy is not high enough, only the items it is good at will be detected, and other

items will be handled by doctors, which can effectively improve the reliability and robustness of the system. The most important step in this is to determine which samples in the test data set are selected, n other words, to determine a reliable sample region.

Gao et al. [14] proposed generative adversarial learning with variance expansion (GALVE), in which the sample generator is obtained through a generative adversarial network (GAN), and in the discriminator part of the GAN, high-variance adversarial samples are used for fine-tuning to ensure the performance of the discriminator.

4.2 Model Understanding Through Subspace Explanation

As mentioned in [20], the reason why adversarial sample attacks are valid and difficult to defend is because of the high-dimensional linearity of neural networks, which causes small perturbations to be amplified and behaviors that are often difficult to explain. If a method can be found that allows we to understand the learning process of neural networks and can understand the behavior of the model within different feature subspaces, we can know when to trust a DNN model based on observations. This has important implications for defending against adversarial attacks and improving the reliability of the model.

Lakkaraju et al. [27] proposed a method for model interpretation through subspaces (MUSE). This interpretation framework quantifies the authenticity, reliability, and interpretability of the model. In MUSE, a new objective function will be constructed to explain the original model, which will be helpful in improving the robustness of the model.

4.3 Software Testing

To improve the interpretability of the model, in addition to observing the various behaviors and characteristics of the model during the training process, the method of sotfware testing can also be used.

Tian et al. [40] proposed a software testing method named DeepTest, an automatic testing tool that can automatically detect the wrong behavior of DNN. By generating test inputs that maximize the number of activated neurons, DeepTest can maximize the understanding of different logics in various parts of the DNN.

The tool can find various model decision errors in different real-world conditions, which will be of great help to the IoT system defense against adversarial example attacks. More importantly, the authors conducted this study in the context of autonomous vehicles using DNN decision-making, which fully demonstrates the feasibility of the method in the IoT system.

5 Conclusion

We believe that the use of deep neural networks in IoT systems may be subject to adversarial example attacks, which will pose security risks in many application

scenarios. In this paper, not only do we introduce adversarial example attack scenarios in IoT, but we also summarize the state-of-the-art works to improve the reliability of DNN in IoT systems. In general, adversarial example attacks in IoT scenarios are still a promising research direction. More aggressive attack schemes can be further proposed to find potential security risks in IoT systems and to design more powerful defense methods to ensure the reliability of IoT services.

References

1. Ali, T.A.A., Xiao, Z., Sun, J., Mirjalili, S., Havyarimana, V., Jiang, H.: Optimal design of IIR wideband digital differentiators and integrators using salp swarm algorithm. Knowl.-Based Syst. **182**, 104834 (2019)
2. Amodei, D., et al.: Deep speech 2: End-to-end speech recognition in English and mandarin. In: International Conference on Machine Learning, pp. 173–182. PMLR (2016)
3. Bartlett, P.L., Wegkamp, M.H.: Classification with a reject option using a hinge loss. J. Mach. Learn. Res. **9**(8), 1–18 (2008)
4. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: Adversarial attacks and defences: a survey. arXiv preprint arXiv:1810.00069 (2018)
5. Chen, M., Hao, Y.: Label-less learning for emotion cognition. IEEE Trans. Neural Netw. Learn. Syst. **31**(7), 2430–2440 (2019)
6. Chen, Y., Gong, X., Ou, R., Duan, L., Zhang, Q.: Crowdcaching: incentivizing D2D-enabled caching via coalitional game for IoT. IEEE Internet Things J. **7**(6), 5599–5612 (2020)
7. Chen, Y., Gong, X., Wang, Q., Di, X., Huang, H.: Backdoor attacks and defenses for deep neural networks in outsourced cloud environments. IEEE Netw. **34**(5), 141–147 (2020)
8. Chen, Y., Ran, Y., Zhou, J., Zhang, J., Gong, X.: MPCN-RP: a routing protocol for blockchain-based multi-charge payment channel networks. IEEE Trans. Netw. Serv. Manage. **19**, 1229–1242 (2021)
9. Cheng, L., et al.: SCTSC: a semicentralized traffic signal control mode with attribute-based blockchain in IoVs. IEEE Trans. Comput. Soc. Syst. **6**(6), 1373–1385 (2019)
10. Dai, X., et al.: Task co-offloading for D2D-assisted mobile edge computing in industrial internet of things. IEEE Trans. Ind. Inform. **19**, 480–490 (2022)
11. Dong, J., Gong, X., Xue, M.: Adversarial examples in wireless networks: a comprehensive survey. In: Wu, K., Wang, L., Chen, Y. (eds.) Edge Computing and IoT: Systems, Management and Security, ICECI 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 437, pp. 92–97. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04231-7_8
12. Duchesne, L., Karangelos, E., Wehenkel, L.: Recent developments in machine learning for energy systems reliability management. Proc. IEEE **108**(9), 1656–1676 (2020)
13. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933–1941 (2016)

14. Gao, J., Yao, J., Shao, Y.: Towards reliable learning for high stakes applications. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3614–3621 (2019)
15. Gong, X., Chen, Y., Huang, H., Liao, Y., Wang, S., Wang, Q.: Coordinated backdoor attacks against federated learning with model-dependent triggers. IEEE Netw. **36**(1), 84–90 (2022)
16. Gong, X., et al.: Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment. IEEE J. Sel. Areas Commun. **39**(8), 2617–2631 (2021)
17. Gong, X., Chen, Y., Wang, Q., Kong, W.: Backdoor attacks and defenses in federated learning: state-of-the-art, taxonomy, and future directions. IEEE Wirel. Commun. (2022)
18. Gong, X., Chen, Y., Wang, Q., Wang, M., Li, S.: Private data inference attacks against cloud: model, technologies, and research directions. IEEE Commun. Mag. **60**, 46–52 (2022)
19. Gong, X., Chen, Y., Yang, W., Mei, G., Wang, Q.: InverseNet: augmenting model extraction attacks with training data inversion. In: IJCAI, pp. 2439–2447 (2021)
20. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
21. Gupta, A., Anpalagan, A., Guan, L., Khwaja, A.S.: Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues. Array **10**, 100057 (2021)
22. Hu, Z., Zeng, F., Xiao, Z., Fu, B., Jiang, H., Chen, H.: Computation efficiency maximization and QoE-provisioning in UAV-enabled MEC communication systems. IEEE Trans. Netw. Sci. Eng. **8**(2), 1630–1645 (2021)
23. Jiang, H., Dai, X., Xiao, Z., Iyengar, A.K.: Joint task offloading and resource allocation for energy-constrained mobile edge computing. IEEE Trans. Mob. Comput. (2022)
24. Jiang, H., Xiao, Z., Li, Z., Xu, J., Zeng, F., Wang, D.: An energy-efficient framework for internet of things underlaying heterogeneous small cell networks. IEEE Trans. Mob. Comput. **21**(1), 31–43 (2020)
25. Jiao, L., Wu, Y., Dong, J., Jiang, Z.: Toward optimal resource scheduling for internet of things under imperfect CSI. IEEE Internet Things J. **7**(3), 1572–1581 (2019)
26. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC (2018)
27. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Faithful and customizable explanations of black box models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 131–138 (2019)
28. Li, J., et al.: Drive2friends: inferring social relationships from individual vehicle mobility data. IEEE Internet Things J. **7**(6), 5116–5127 (2020)
29. Li, S., Da Xu, L., Zhao, S.: 5g internet of things: a survey. J. Ind. Inf. Integr. **10**, 1–9 (2018)
30. Li, W., et al.: Hu-Fu: Hardware and software collaborative attack framework against neural networks. In: 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 482–487. IEEE (2018)
31. Liu, Y., Xie, Y., Srivastava, A.: Neural trojans. In: 2017 IEEE International Conference on Computer Design (ICCD), pp. 45–48. IEEE (2017)
32. Long, W., et al.: Unified spatial-temporal neighbor attention network for dynamic traffic prediction. IEEE Trans. Veh. Technol. **72**, 1515–1529 (2022)

33. Luo, X., Qin, Q., Gong, X., Xue, M.: A survey of adversarial attacks on wireless communications. In: Wu, K., Wang, L., Chen, Y. (eds.) ICECI 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 437, pp. 83–91. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04231-7_7
34. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1765–1773 (2017)
35. Moustafa, N., Keshk, M., Choo, K.K.R., Lynar, T., Camtepe, S., Whitty, M.: Dad: a distributed anomaly detection system using ensemble one-class statistical learning in edge networks. Futur. Gener. Comput. Syst. **118**, 240–251 (2021)
36. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
37. Setiaji, T., Budiyanto, C., Yuana, R.: The contribution of the internet of things and smart systems to agricultural practices: a survey. In: IOP Conference Series: Materials Science and Engineering. vol. 1098, p. 052100. IOP Publishing (2021)
38. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
39. Tan, L., Wang, N.: Future internet: the internet of things. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), vol. 5, pp. V5–376. IEEE (2010)
40. Tian, Y., Pei, K., Jana, S., Ray, B.: DeepTest: automated testing of deep-neural-network-driven autonomous cars. In: Proceedings of the 40th International Conference on Software Engineering, pp. 303–314 (2018)
41. Wu, J., Luo, S., Wang, S., Wang, H.: NLES: a novel lifetime extension scheme for safety-critical cyber-physical systems using SDN and NFV. IEEE Internet Things J. **6**(2), 2463–2475 (2018)
42. Xiao, Z., et al.: Resource management in UAV-assisted MEC: state-of-the-art and open challenges. Wireless Netw. **28**(7), 3305–3322 (2022)
43. Xiao, Z., et al.: TrajData: on vehicle trajectory collection with commodity plug-and-play OBU devices. IEEE Internet Things J. **7**(9), 9066–9079 (2020)
44. Yin, B., Wu, Y., Hu, T., Dong, J., Jiang, Z.: An efficient collaboration and incentive mechanism for internet of vehicles (IoV) with secured information exchange based on blockchains. IEEE Internet Things J. **7**(3), 1582–1593 (2019)
45. Zhang, J., Li, C.: Adversarial examples: opportunities and challenges. IEEE Trans. Neural Netw. Learn. Syst. **31**(7), 2578–2593 (2019)
46. Zhou, W., Jia, Y., Peng, A., Zhang, Y., Liu, P.: The effect of IoT new features on security and privacy: new threats, existing solutions, and challenges yet to be solved. IEEE Internet Things J. **6**(2), 1606–1616 (2018)