Privacy Attacks and Defenses in Machine Learning: A Survey

Wei Liu[1], Xun Han[2](✉), and Meiling He[3]

[1] City University of Macau, Macau, China
[2] Intelligent Policing Key Laboratory of Sichuan Province, Luzhou 646000, Sichuan, China
hldwxhx@163.com
[3] School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, Jiangsu, China
hemeiling@ujs.edu.cn

Abstract. As machine learning has gradually become an important technology in the field of artificial intelligence, its development is also facing challenges in terms of privacy. This article aims to summarize the attack methods and defense strategies for machine learning models in recent years. Attack methods include embedding inversion attack, attribute inference attack, membership inference attack and model extraction attack, etc. Defense measures include but are not limited to homomorphic encryption, adversarial training, differential privacy, secure multi-party computation, etc., focusing on the analysis of privacy protection issues in machine learning, and providing certain references and references for related research.

Keywords: Machine learning model · Means of attack · Defense strategy

1 Introduction

This chapter will briefly sort out the background knowledge of machine learning and privacy leakage, and explain the research significance of this paper, as well as the difficulties and challenges in this research direction

1.1 Related Work

In the era of big data, massive amounts of information have promoted the development of machine learning, and now it has been widely used in malicious detection (see [33,37]), computer vision (see [32,49]), voice command recognition (see [46,48]), driving system (see [10,47]), recommendation system (see [36,56]), medical diagnosis (see [3,13]) and many other fields. Machine learning can discover patterns and laws from massive data, and apply this knowledge to different tasks, bringing great convenience and benefits to humans. Especially after the

Y. Zhang et al. (Eds.): CENet 2023, LNEE 1127, pp. 413–422, 2024.
https://doi.org/10.1007/978-981-99-9247-8_41

breakthrough development of technologies such as deep learning in [41] and reinforcement learning in [31] , it has provided strong support for the application of machine learning in the above fields, and even some performances have been better than humans. In the past, there has been a lot of research work on privacy protection measures in machine learning. Many workers have evaluated and summarized the existing attack and defense work. Reference [5] studies the attack model of machine learning, and uses a Statistical spam classification is taken as an example, and an in-depth analysis is carried out. Reference [2] took cleaning robots as an example to summarize and analyze the problems that may exist in the real work and life of human beings. Reference [4] use a black-box model and a white-box model to conduct targeted research on machine learning adversarial attacks and poisoning attacks. Although Ref. [34] is an article on computer security A comprehensive overview of threats, but it also summarizes some of the content related to machine learning. Reference [35] focuses on the training and prediction phases of the machine learning life cycle. Reference [1] focuses on It is a security issue in the field of computer vision. Reference [18] is based on the machine learning CIA model to investigate and summarize.

This article first explains the development of machine learning and privacy leakage, and then from privacy leakage, attack methods. The three angles of model security systematically and scientifically summarize the existing machine learning attack methods and defense methods, and discuss the limitations of related research. Finally, discuss the challenges faced by machine learning model security and privacy research and Feasible research directions in the future, mainly including contributions

1. Conduct a comprehensive and systematic analysis and summary of attack methods and defense technologies in recent years

2. Present the possible attacks and defense measures of machine learning through the combination of charts, and introduce typical attacks and defense methods

3. According to the characteristics and current situation of machine learning, this paper proposes a multi-faceted summary and outlook.

2 Machine Learning Model

The machine learning model in Ref. [30] is a data-driven predictive model, which discovers the relationship and regularity between variables by training the model on a large amount of data, and realizes the prediction or classification of future data.

2.1 Model Introduction

This paper uses the Amazon Machine Learning (Amazon ML) model in Ref. [45]. Amazon Machine Learning is a machine learning service provided by Amazon, which aims to help users quickly build and deploy high-quality machine learning models. It provides a series of easy-to-use APIs and tools that enable users to quickly build, train and test models without requiring extensive machine learning expertise. And it supports a variety of machine learning models, including linear regression, logistic regression, decision trees, support vector machines, and random forests.

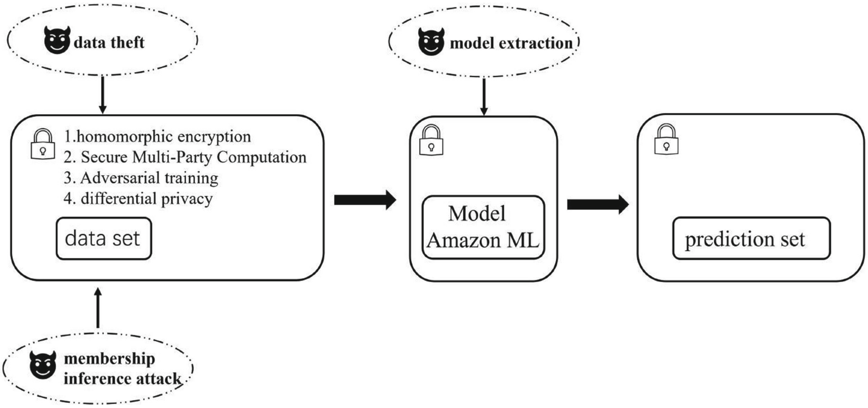

Fig. 1. Machine learning model

3 Privacy Leakage

This chapter will briefly explain the background knowledge about privacy leakage. Privacy leakage has always been an important field of machine learning research.

3.1 Background Knowledge on Privacy Leakage

In machine learning, due to specific scenario requirements, its model design is biased towards efficient and accurate prediction output, rather than the ability of the model to resist attacks. Therefore, in the actual application of the model, there will be malicious users attacking various stages of the life cycle of the machine learning model. During the attack process, the risk caused by privacy leakage is particularly prominent.

This paper mainly focuses on attack vectors for data privacy. Most attackers are more inclined to obtain private data of unspecified people, and have developed more attack methods. Therefore, this paper introduces three data-targeted attack methods, involving multiple aspects such as data training, input, and prediction. At the same time, the attack on model privacy is also mainly reflected in the extraction of the model. Regardless of the purpose of the attacker, the leakage of these data will cause considerable damage to the data owner. Therefore, both providers and users of machine learning models should pay more attention to privacy protection and continuously improve their ability to resist attacks.

4 Attack Methods That Cause Privacy Leakage

Vulnerabilities in machine learning model algorithms and implementations lead to security risks such as data leakage and loss of model structural parameters. This chapter will introduce three attack methods targeting model data and one attack method targeting the model itself.

4.1 Attack Methods Targeting Model Data

Model data is the basis of machine learning models. Model data includes training data, model input and output, etc. Some training data sets that involve privacy, such as shopping records, hospital records, etc., are also related to the issue of personal privacy protection. The following describes three attack methods against machine learning model data.

4.1.1 Embedding Inversion Attack Embedding Inversion Attack, also known as embedded inversion attack [14], is an attack method for deep learning semantic embedding models. This attack is often used to infer input text from pre-trained neural language models. In natural language processing, embedding refers to embedding words or phrases into a real vector space using a small fixed-length representation. In text classification tasks, the input text sequence needs to be converted into a sequence of numbers. To do this, word embedding methods can be used to map each word into a vector space of low-dimensional vectors. In this way, a sentence or paragraph can be represented as a matrix of word vectors. This matrix will be fed into a neural network for text classification. At the same time, this matrix can be used as one of the inputs of the deep learning model, and the embedding vector can be used as the context and passed to the neural network for classification, regression and other tasks.

4.1.2 Attribute Inference Attack Attribute Inference Attack (see [19,55]) aims to infer private attributes in training data from machine learning models. Attackers do not need to directly access protected personal data, but instead gain private information about personal data by analyzing deployed machine learning models.

4.1.3 Membership Inference Attack Member Inference Attack is a privacy attack in machine learning [15,28,38,44,51], which aims to determine whether a given input belongs to the data set by accessing the protected training data set, that is, whether there are members in the data set identity. Membership inference attacks are based on two assumptions, one is that the model is knowable, and the other is that the attacker has access to some sample labels (that is, having membership and corresponding labels in the data set). Then, the attacker specifies some input data and guesses whether it belongs to a specific member in the dataset by tracking the model output.

4.2 The Target Is the Attack Method of the Model Itself

There are also attack methods for machine learning models. Attackers can obtain information related to the model by calling APIs related to the machine learning model, and can even disguise or embezzle the model to achieve the purpose of stealing private data.

4.2.1 Attack Overview Among them, Model Extraction Attack, [12,39,52] has been studied for simple classification tasks, vision tasks, NLP tasks, etc. Typically, model extraction attacks aim to reconstruct a local copy or steal the functionality of a black-box API. If the extraction is successful, the attacker has effectively stolen the intellectual property, i.e. the full details of the model. The work of this attack method mainly focuses on how to imitate a model with performance close to the victim API in the source domain, and a more powerful attacker may even extract a better model than the target victim API [16].

Attackers use this technique to steal model knowledge by accessing the model and its output to deduce sensitive information of the target model. The reason why this attack technique is called model extraction attack is because the attacker can replace the attacked model with a model constructed by himself, and can output the corresponding label in a way consistent with the original model [50].

5 Defense Measures Against Privacy Leakage

This chapter will introduce four commonly used schemes in privacy protection, namely homomorphic encryption, secure multi-party computation, confrontation training, and differential privacy.

5.1 Homomorphic Encryption

Homomorphic Encryption (HE) refers to satisfying the original file through a specific homomorphic encryption algorithm, the encrypted ciphertext can satisfy the property of homomorphic operation, and the final ciphertext operation result is equivalent to the corresponding homomorphic decryption The result of performing the same operation directly on the subsequent plaintext can realize the "countable and invisible" data. In the cryptographic system, homomorphic encryption is usually based on computational problems in mathematics, including but not limited to integer decomposition problems, discrete logarithm problems, Determining the remainder of composite numbers, the approximate greatest common factor problem [7,17,40], etc.

5.2 Secure Multi-party Computation

Secure Multi-Party Computation [6,11] was proposed by Professor Yao Qizhi, an academician of the Chinese Academy of Sciences, in 1982. For model training, secure multi-party computation requires the use of cryptographic tools, such as secret sharing [20,21], zero-knowledge proof [26,27], oblivious transmission [8,54], obfuscation circuits [42,53], and in centralized machine learning, secure multi-party The calculation is performed on two non-collusive servers through secret sharing, and the scheme can be extended to the scene of hundreds of users, followed by a large amount of communication overhead . The difference is that the secure multi-party computing scheme based on obfuscated circuit technology can generally only be applied to two to three parties to complete model training. In the joint machine learning model, homomorphic encryption or zero-knowledge proof is more commonly used.

5.3 Adversarial Training

Adversarial training [24,25,29] is a defense method in machine learning that aims to improve the resistance of deep neural networks to adversarial attacks. By injecting some perturbations into the original data, the machine learning model is made more robust, thereby reducing the impact of attacks on model data and structural parameters. The method mainly includes the following steps: First, generate adversarial samples. First, some attack algorithm needs to be used to generate adversarial samples and added to the normal training data set to form a new training set. The second is to train the model. The model is retrained using a new training set with adversarial samples in order to enhance the tolerance of the model against noise and improve the robustness of the model. And finally the test model. After the training is completed, the test set is evaluated. If the model shows better robustness, it will have a better ability to deal with raw data and adversarial input than the normal model. If expectations are not met, repeat the first two steps until you are satisfied.

5.4 Differential Privacy

Differential privacy is a data protection algorithm with strict mathematical definition and privacy quantification. By perturbing the data, such as adding noise, the attacker cannot deduce the original data, thereby achieving data privacy protection and avoiding the complete destruction of the original data. To ensure the availability of perturbed data . Nowadays, differential privacy technology can be divided into centralized differential privacy [9,23] and localized differential privacy [22,43] according to the processing subject. Among them, centralized differential privacy processes data by a trusted third party, while localized differential privacy The data is privately processed locally by the user, and the more mainstream method is localized differential privacy.

6 Summary and Outlook

This article introduces the leakage risks and defenses of machine learning models, describes four attack methods and introduces four defense measures. It can be seen that whether it is aimed at the model training data or the attack method against the model itself, it is applicable to most of today's machine learning models. It can be seen that these models have a certain risk of leakage. With the deepening of machine learning and artificial intelligence research, the application of machine learning models has become more and more extensive, and it has become more and more deeply involved in all aspects of people's lives. A large amount of personal privacy data is applied to the training of the model to improve the humanity and intelligence of the model. However, this trend increases the danger caused by privacy leaks. After mastering relevant data, attackers can rely on the performance of private data to profile people, and may target potential advertisements, data collection, and even targeted telecommunications. Provide convenience for online fraud and theft of private property.

Acknowledgement. The Opening Project of Intelligent Policing Key Laboratory of Sichuan Province, No. ZNJW2023KFMS004.

References

1. Akhtar, N., Mian, A.: Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access **6**, 14410–14430 (2018)
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety (2016). arXiv:1606.06565
3. Arumugam, K., Naved, M., Shinde, P.P., Leiva-Chauca, O., Huaman-Osorio, A., Gonzales-Yanac, T.: Multiple disease prediction using machine learning algorithms. Mater. Today Proc. **80**, 3682–3685 (2023)
4. Bae, H., Jang, J., Jung, D., Jang, H., Ha, H., Lee, H., Yoon, S.: Security and privacy issues in deep learning (2018). arXiv:1807.11655
5. Barreno, M., Nelson, B., Joseph, A.D., Tygar, J.D.: The security of machine learning. Mach. Learn. **81**, 121–148 (2010)
6. Braun, L., Huppert, M., Khayata, N., Schneider, T., Tkachenko, O.: Fuse–flexible file format and intermediate representation for secure multi-party computation. Cryptology ePrint Archive (2023)
7. Doan, T.V.T., Messai, M.L., Gavin, G., Darmont, J.: A survey on implementations of homomorphic encryption schemes. J. Supercomput. 1–42 (2023)
8. Fan, C., Jia, P., Lin, M., Wei, L., Guo, P., Zhao, X., Liu, X.: Cloud-assisted private set intersection via multi-key fully homomorphic encryption. Mathematics **11**(8), 1784 (2023)
9. Feldman, V., McMillan, A., Talwar, K.: Stronger privacy amplification by shuffling for rényi and approximate differential privacy. In: Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 4966–4981. SIAM (2023)
10. Flores Fernández, A., Sánchez Morales, E., Botsch, M., Facchi, C., García Higuera, A.: Generation of correction data for autonomous driving by means of machine learning and on-board diagnostics. Sensors **23**(1), 159 (2023)

11. Gao, C., Yu, J.: Securerc: a system for privacy-preserving relation classification using secure multi-party computation. Comput. Secur. **128**, 103, 142 (2023)
12. Gong, X., Wang, Q., Chen, Y., Yang, W., Jiang, X.: Model extraction attacks and defenses on cloud-based machine learning models. IEEE Commun. Mag. **58**(12), 83–89 (2020)
13. Haug, C.J., Drazen, J.M.: Artificial intelligence and machine learning in clinical medicine, 2023. N. Engl. J. Med. **388**(13), 1201–1208 (2023)
14. Hayet, I., Yao, Z., Luo, B.: Invernet: An inversion attack framework to infer fine-tuning datasets through word embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 5009–5018 (2022)
15. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: a survey. ACM Comput. Surv. (CSUR) **54**(11s), 1–37 (2022)
16. Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., Papernot, N.: High accuracy and high fidelity extraction of neural networks. In: Proceedings of the 29th USENIX Conference on Security Symposium, pp. 1345–1362 (2020)
17. Jain, N., Pal, S.K., Upadhyay, D.K.: Implementation and analysis of homomorphic encryption schemes. Int. J. Cryptogr. Inf. Secur. (IJCIS) **2**(2), 27–44 (2012)
18. Ji, S., Du, T., Li, J., Shen, C., Li, B.: Security and privacy of machine learning models: a survey. Ruan Jian Xue Bao/J. Softw. **32**(1), 41–67 (2021)
19. Jia, J., Gong, N.Z.: Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In: 27th {USENIX} security symposium ({USENIX} security 18), pp. 513–529 (2018)
20. Kamal, A.A.A.M., Iwamura, K.: Privacy preserving multi-party multiplication of polynomials based on (k, n) threshold secret sharing. ICT Express (2023)
21. Li, F., Chen, T., Zhu, S.: A (t, n) threshold quantum secret sharing scheme with fairness. Int. J. Theor. Phys. **62**(6), 119 (2023)
22. Li, M., Tian, Z., Du, X., Yuan, X., Shan, C., Guizani, M.: Power normalized cepstral robust features of deep neural networks in a cloud computing data privacy protection scheme. Neurocomputing **518**, 165–173 (2023)
23. Li, Y., Wang, R., Li, Y., Zhang, M., Long, C.: Wind power forecasting considering data privacy protection: A federated deep reinforcement learning approach. Appl. Energy **329**, 120, 291 (2023)
24. Lin, T.H., Lee, Y.S., Chang, F.C., Chang, J.M., Wu, P.Y.: Protecting sensitive attributes by adversarial training through class-overlapping techniques. IEEE Trans. Inf. Forensics Secur. (2023)
25. Liu, J., Lau, C.P., Chellappa, R.: Diffprotect: generate adversarial examples with diffusion models for facial privacy protection (2023). arXiv:2305.13625
26. Liu, X., Tu, X.F., Luo, D., Xu, G., Xiong, N.N., Chen, X.B.: Secure multi-party computation of graphs' intersection and union under the malicious model. Electronics **12**(2), 258 (2023)
27. Liu, Y., Feng, Q., Peng, C., Luo, M., He, D.: Asymmetric secure multi-party signing protocol for the identity-based signature scheme in the IEEE p1363 standard for public key cryptography. In: Emerging Information Security and Applications: Third International Conference, EISA 2022, Wuhan, China, October 29–30, 2022, Proceedings, pp. 1–20. Springer (2023)
28. Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M., Zhang, Y.: {ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models. In: 31st USENIX Security Symposium (USENIX Security 22), pp. 4525–4542 (2022)

29. Luo, X., Chen, Z., Tao, M., Yang, F.: Encrypted semantic communication using adversarial training for privacy preserving. IEEE Commun. Lett. (2023)
30. Mahesh, B.: Machine learning algorithms-a review. Int. J. Sci. Res. (IJSR). [Internet] **9**, 381–386 (2020)
31. Moerland, T.M., Broekens, J., Plaat, A., Jonker, C.M., et al.: Model-based reinforcement learning: a survey. Found. Trends® Mach. Learn. **16**(1), 1–118 (2023)
32. Ning, X., Tian, W., He, F., Bai, X., Sun, L., Li, W.: Hyper-sausage coverage function neuron model and learning algorithm for image classification. Pattern Recognit. **136**, 109, 216 (2023)
33. Nouman, M., Qasim, U., Nasir, H., Almasoud, A., Imran, M., Javaid, N.: Malicious node detection using machine learning and distributed data storage using blockchain in wsns. IEEE Access (2023)
34. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.: Towards the science of security and privacy in machine learning (2016). arXiv:1611.03814
35. Papernot, N., McDaniel, P., Sinha, A., Wellman, M.P.: Sok: security and privacy in machine learning. In: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 399–414. IEEE (2018)
36. Pawase, A.D., Mandage, V.T., Panchal, S.S., Patil, S.Y., Deokar, P.: A shop recommendation system to empower retailers using machine learning
37. Rashid, K., Saeed, Y., Ali, A., Jamil, F., Alkanhel, R., Muthanna, A.: An adaptive real-time malicious node detection framework using machine learning in vehicular ad-hoc networks (vanets). Sensors **23**(5), 2594 (2023)
38. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: Ml-leaks: model and data independent membership inference attacks and defenses on machine learning models (2018). arXiv:1806.01246
39. Salih, A., Zeebaree, S.T., Ameen, S., Alkhyyat, A., Shukur, H.M.: A survey on the role of artificial intelligence, machine learning and deep learning for cybersecurity attack detection. In: 2021 7th International Engineering Conference "Research & Innovation amid Global Pandemic" (IEC), pp. 61–66. IEEE (2021)
40. Sen, J.: Homomorphic encryption-theory and application. In: Theory and Practice of Cryptography and Network Security Protocols and Technologies, vol. 31 (2013)
41. Sharifani, K., Amini, M.: Machine learning and deep learning: a review of methods and applications. World Inf. Technol. Eng. J. **10**(07), 3897–3904 (2023)
42. Song, C., Huang, R.: Secure convolution neural network inference based on homomorphic encryption. Appl. Sci. **13**(10), 6117 (2023)
43. Sun, S., Huang, H., Peng, T., Shen, C., Wang, D.: A data privacy protection diagnosis framework for multiple machines vibration signals based on a swarm learning algorithm. IEEE Trans. Instrum. Meas. **72**, 1–9 (2023)
44. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Towards demystifying membership inference attacks (2018). arXiv:1807.09173
45. Venkateswar, K.: Using Amazon Sagemaker to Operationalize Machine Learning. Santa Clara, CA. USENIX Association (2019)
46. Weng, Z., Qin, Z., Tao, X., Pan, C., Liu, G., Li, G.Y.: Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Trans. Wirel. Commun. (2023)
47. Wu, J., Huang, Z., Hu, Z., Lv, C.: Toward human-in-the-loop ai: enhancing deep reinforcement learning via real-time human guidance for autonomous driving. Engineering **21**, 75–91 (2023)
48. Xin, J., Lyu, X., Ma, J.: Natural backdoor attacks on speech recognition models. In: Machine Learning for Cyber Security: 4th International Conference, ML4CS

2022, Guangzhou, China, December 2–4, 2022, Proceedings, Part I, pp. 597–610. Springer (2023)

49. Xu, M., Yoon, S., Fuentes, A., Park, D.S.: A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 109347 (2023)
50. Xu, Q., He, X., Lyu, L., Qu, L., Haffari, G.: Beyond model extraction: imitation attack for black-box nlp apis. arXiv e-prints pp. arXiv–2108 (2021)
51. Ye, J., Maddi, A., Murakonda, S.K., Bindschaedler, V., Shokri, R.: Enhanced membership inference attacks against machine learning models. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pp. 3093–3106 (2022)
52. Yi, T., Chen, X., Zhu, Y., Ge, W., Han, Z.: Review on the application of deep learning in network attack detection. J. Netw. Comput. Appl. **212**, 103,580 (2023)
53. Yu, Y., Li, Z., Tu, Y., Yuan, Y., Li, Y., Pang, Z.: Blockchain-based distributed identity cryptography key management. In: 2023 15th International Conference on Computer Research and Development (ICCRD), pp. 236–240. IEEE (2023)
54. Zhang, J., Tian, H., Xiong, K., Tang, Y.L., Yang, L.: Fair multi-party private set intersection protocol based on cloud server. J. Comput. Appl. 0 (2023)
55. Zhao, B.Z.H., Agrawal, A., Coburn, C., Asghar, H.J., Bhaskar, R., Kaafar, M.A., Webb, D., Dickinson, P.: On the (in) feasibility of attribute inference attacks on machine learning models. In: 2021 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 232–251. IEEE (2021)
56. Zheng, R., Qu, L., Cui, B., Shi, Y., Yin, H.: Automl for deep recommender systems: a survey. ACM Trans. Inf. Syst. (2023)