Privacy Attacks on Machine Learning Models: A Survey on Open Problems and Future Directions

Jonggyu Jang, Hyun Jong Yang

POSTECH

{jgjang, hyunyang}@postech.ac.kr

기계 학습 모델에 관한 개인정보 공격들: 해결해야할 문제와 향후 연구방향에 관한 동향 분석

장종규, 양현종*
포항공과대학교

Abstract

Machine learning technologies have attracted enormous research interests as they process a complex statistic of data. At the same time, security and privacy have become urgent. Recently, some studies on machine learning privacy show that trained machine learning models can memorize the training data, and some attacker can reveal sensitive information of data used for training even for tabular, text, and image data. Here, we first analyze the trends of recent works on privacy attacks and risks of machine learning models. Then, we state open problems and future directions.

Ⅰ. Introduction

After Google's artificial intelligence (AI) program defeated the Go world champion, the AlphaGo shock, the machine learning technologies have attracting enormous interests in academic/industrial areas. Now, generative models, e.g., ChatGPT, Dall-E, Dreambooth, and Stable Diffusion, are driving another shock, in which they can generate realistic text/image by learning/memorizing huge datasets.

Notwithstanding the remarkable interests in AI technologies, the high pace of development of AI results in privacy and security having stayed one step behind the technological innovations.

Recently, the security and privacy of AI technologies are starting to attract research interest. In this study, we survey various privacy attack methods on machine learning, which reveals sensitive information of machine learning models or training data.

Ⅱ. Types of Privacy Attacks on Machine Learning

In this section we briefly introduce four categories of privacy attack methods briefly and classify existing method.

♦ Membership Inference Attack: The membership inference attack is to reveal the existence of a sample data in the training dataset, i.e., the membership inference attack predicts whether a data instance is used in the training dataset or not. This attack policy is the lowest level of data inversion attack; however, it guarantees the bound of data protection. For instance, several works [1, 2] use membership inference attack to guarantee certification of the data safety.

In [3], a membership inference attacker aims to find whether a data point is a member of dataset of not, especially for overfitted model.

♦ Reconstruction Attacks: The reconstruction attack (or model inversion attack) aims to reconstruct/reveal data distribution or data instances. Like the membership inference attack, a defensive work [4] uses reconstruction attack as its certification method.

In [5], a reconstruction attacker reveals a training instance from the model and the dataset except the training instance.

♦ Property Inference Attacks: The target of property inference attack is extracting statistic information of the dataset. e. g., facial proportions, male/female ratio, etc. Although this attack strategy does not reveal data instance directly, the leaked information can be used for finding weakness of systems.

In [6], a property inference attacker is developed, which backdoors sensitive attributes (e.g., author ship

of text) by inverting low-dimensional latent vector representations.

♦ Model Extraction Attacks: The above three attack methods are targeting data used for constructing data, in which white-box model access is generally assumed. The model extraction attacks are black-box privacy attacks for reconstructing neural network models. These methods are used as a foundation stone of other white-box access privacy attacks.

In [7], a model extractor is proposed, in which a substitute for target model is learned by a synthetic dataset generated by adversarial learner. The extracted model is leveraged to adversarial attack on some trained models.

Ⅲ. Open Problems and Future Directions

♦ Privacy attack for wide-variety of applications: Most of existing works have tackled privacy attack methods for standard classification tasks with cross-entropy loss and some regularization terms. These methods cannot be used for other applications with different loss function and training algorithm such as object detection, generative model, etc. Thus, there are open problems for revealing data/model for those other applications.

♦ Privacy attack for non-image model: Several studies have investigated privacy attacks on image-based machine learning models. However, privacy attacks on other input sources are rarely studied. For instance, few works have addressed text revealer.

Ⅳ. Conclusion

We expound existing privacy attack methods for machine learning algorithms. Also, we categorize privacy attacks into four categories and discuss open problems and future directions.

ACKNOWLEDGMENT

This research was supported in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2021-0-02048) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation), and in part by Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-TD2003-0.

REFERENCES

[1] Z. Ma, Y. Liu, X. Liu, J. Liu, J. Ma and K. Ren, "Learn to Forget: Machine Unlearning Via Neuron Masking," *in IEEE Transactions on Dependable and Secure Computing*, 2022.

[2] H. Hu, Z. Salcic, G. Dobbie, J. Chen, L. Sun, and X. Zhang, "Membership Inference Via Backdooring," *in proc. IJCAI 2022*.

[3] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," *in proc. IEEE CSF 2018*.

[4] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Zero-Shot Machine Unlearning," *preprinted in arxiv.org, 2022*.

[5] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing Training Data With Informed Adversaries," *in proc. IEEE S&P 2022*.

[6] C. Song, and A. Raghunathan, "Information Leakage in Embedding Models," *in proc. ACM SIGSAC CCS 2020*.

[7] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, "Practical Black-Box Attacks Against Machine Learning," *in proc. ACM Asia CCS 2017*.