프라이버시 보존 머신러닝의 연구 동향

한우림*, 이영한*, 전소희*, 조윤기*, 백윤흥*
*서울대학교 전기·정보공학부, 서울대학교 반도체 공동연구소
rimwoo98@snu.ac.kr, yhlee@sor.snu.ac.kr, shjun@sor.snu.ac.kr, ygcho@sor.snu.ac.kr, ypaek@sor.snu.ac.kr

A Study on Privacy Preserving Machine Learning

Woorim Han*, Younghan Lee*, Sohee Jun*, Yungi Cho*, Yunheung Paek*
*Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center
(ISRC), Seoul National University

Abstract

AI (Artificial Intelligence) is being utilized in various fields and services to give convenience to human life. Unfortunately, there are many security vulnerabilities in today’s ML (Machine Learning) systems, causing various privacy concerns as some AI models need individuals’ private data to train them. Such concerns lead to the interest in ML systems which can preserve the privacy of individuals’ data. This paper introduces the latest research on various attacks that infringe data privacy and the corresponding defense techniques.

1. Introduction

Recently, companies are trying to provide useful services by developing AI (Artificial Intelligence) with high performance. AI is receiving a lot of attention as it shows remarkable performance in various fields such as image classification, object detection, and natural language processing with big data and deep learning. The reason AI technology has made endless progress is because the amount of training data for AI models has increased. Personal information such as a user's face, address, and phone number may be contained in such vast and diverse learning data. When trained with such private data, the model learns even sensitive information in the training process and leaks private features about the data provider during the inference.

Taking advantage of these characteristics, attackers are making attempts to uncover the data and sensitive information the AI model has learned; model inversion attack and membership inference attack has been proposed threatening the privacy of data providers. Based on white box and black box methods, the adversary reveals critical information about the training dataset.

As attack techniques that can threaten data privacy have been recently proposed for ML models, the demand for defense techniques that preserve privacy is increasing. Accordingly, research on various privacy-enhancing ML technologies is being conducted.

2. Threats to Data Privacy

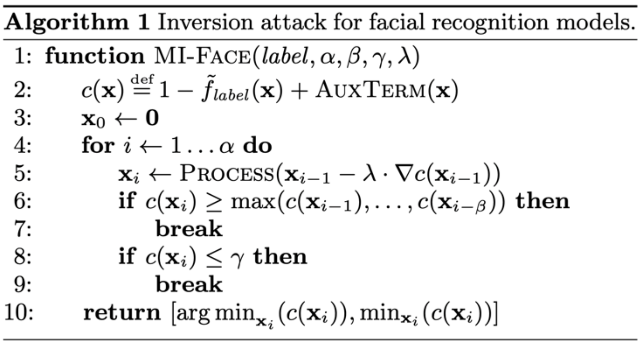
2.1 Model Inversion Attack

Model inversion attacks aims to reconstruct the training data using the model parameters and the confidence information of the model. Such attacks were first demonstrated on models with simple architectures [1], linear regression and logistic regression, and has developed to invert deep neural networks [2].

The model inversion attack proposed by Fredrikson et al. [1] recovers training data using a class value (name or index) for the corresponding target. Algorithm 1 (Fig. 1) shows the model inversion attack on the face recognition model(*f*). The attack is transformed into an optimization problem to minimize the cost function, *c(x)*, through gradient descent. Optimization is performed for a given number of α iterations, and the optimization is completed when the cost is not improved in β iterations or when the cost is less than γ. The PROCESS function refers to denoising and generalizing techniques for image manipulation.

2.2 Membership Inference Attack

Membership inference attack aims to infer whether a given data record is part of the training dataset of a target ML model. This attack is a black box attack method as it is done by just observing the output of the model. The adversary utilizes the differences in model’s predictions on samples that were used and not included in the training set.



Algorithm 1 Inversion attack for facial recognition models.

```
function MI-FACE($label, \alpha, \beta, \gamma, \lambda$)
    $c(\mathbf{x}) \overset{\text{def}}{=} 1 - \tilde{f}_{label}(\mathbf{x}) + \text{AUXTERM}(\mathbf{x})$
    $\mathbf{x}_0 \leftarrow \mathbf{0}$
    for $i \leftarrow 1 \dots \alpha$ do
        $\mathbf{x}_i \leftarrow \text{PROCESS}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$
        if $c(\mathbf{x}_i) \geq \max(c(\mathbf{x}_{i-1}), \dots, c(\mathbf{x}_{i-\beta}))$ then
            break
        if $c(\mathbf{x}_i) \leq \gamma$ then
            break
    return $[\arg\min_{\mathbf{x}_i}(c(\mathbf{x}_i)), \min_{\mathbf{x}_i}(c(\mathbf{x}_i))]$
```

(Algorithm 1: Model Inversion Attack in [1])

Shokri et al. [3] was the first to introduce membership inference attack in the machine learning setting. The attacker trains binary classifiers that use data sample's confidence score vector, the target model's output, as an input and predicts if the sample was included in the training dataset. They propose and use shadow training method which is creating several shadow classifiers that resemble the target classifier's behavior. This method requires the assumption that the attacker knows the learning algorithm and structure of the target classifier. Then as the attacker have knowledge about the shadow model's datasets and its labels, the attacker can train the attack model using multiple shadow models' confidence score vectors. It requires a white or black box access to the target during the training process, but only needs black box access when performing the membership inference attack.

3. Privacy Preserving Machine Learning

Recently, federated learning, in which multiple local clients cooperate to train a global ML model in a centralized server with decentralized data, is being actively used. In federated learning, users train their own models with their own training data and repeatedly collect and share their trained information, such as model's weights and gradients. Federated learning has the advantage of being able to achieve the same performance improvement as training with a vast amount of data while sharing only the model's gradients or weights with other users without exposing the user's individual training data.

Unfortunately, as attacks that reveal training data or find sensitive information with only the trained model and its confidence information are introduced, it became difficult to completely protect the data privacy with only federated learning.

When sending gradients and weights to the server in federated learning, encryption method has been introduced to preserve data privacy. Using Fully Homomorphic Encryption (FHE) to encrypt the sharing values, computation could be done without any decryption. Thus, a curious server cannot exploit the updated information. Moreover, encrypting the model itself could prevent model inversion attacks. Model inversion attacks utilize the model weights when reconstructing the training dataset [4], [5]. With encryption methods, such attacks become very difficult to demonstrate and thus, preserve the privacy of ML.

Perturbation approach is another method for preserving the privacy of ML. J. Jia et al. [6] proposed MemGuard which prevent membership inference attacks via adversarial examples. They observed that the attack model, trained as a binary classifier to determine whether a data sample has been used in the training process, is vulnerable to adversarial examples. Thus, MemGuard aims to find a random noise to add to the prediction value of the ML model to deceive the adversary while not changing the prediction label. It finds the optimal noise to prevent an attacker from accurately performing membership inference attack.

4. Conclusion

As AI models become commercially available and individuals' private data is used to train the models, the threat to privacy is increasing. In this paper, attacks that threaten data privacy by revealing the data used to train the AI model were introduced along with the defenses for each attack. The approach of making AI models as a service will increase significantly in the future, and in order to provide safe services, research on the various threats and defenses must be continued for privacy preserving machine learning.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A2B5B03095204) and **by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00230, Development on Autonomous Trust Enhancement Technology of IoT Device and Study on Adaptive IoT Security Open Architecture based on Global Standardization [TrusThingz Project]). This work was also supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021 and by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00325, Traceability Assuarance Technology Development for Full Lifecycle Data Safety of Cloud Edge)**

References

[1] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.

[2] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang,

Bo Li, and Dawn Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. arXiv:1911.07135 [cs, stat], November 2019.

[3] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P). IEEE, 3–18.

[4] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In Advances in Neural Information Processing Systems, 2019.

[5] Jonas Geiping, Hartmut Bauermeister, Hannah Droge, and ¨ Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? In Advances in Neural Information Processing Systems, 2020.

[6] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In ACM SIGSAC Conference on Computer and Communications Security (CCS). ACM, 259–274.