증류 기반 연합 학습에서 로짓 역전을 통한 개인 정보 취약성에 관한 연구

윤수빈 [1], 조윤기 [1], 백윤흥 [1]
[1] 서울대학교 전기정보공학부, 반도체공동연구소

subyun@sor.snu.ac.kr, ygcho@sor.snu.ac.kr, ypaek@snu.ac.kr

A Survey on Privacy Vulnerabilities through Logit Inversion in Distillation-based Federated Learning

Subin Yun[1], Yungi Cho[1], Yunheung Paek[1]
[1]Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center, Seoul National University

Abstract

In the dynamic landscape of modern machine learning, Federated Learning (FL) has emerged as a compelling paradigm designed to enhance privacy by enabling participants to collaboratively train models without sharing their private data. Specifically, Distillation-based Federated Learning, like Federated Learning with Model Distillation (FedMD), Federated Gradient Encryption and Model Sharing (FedGEMS), and Differentially Secure Federated Learning (DS-FL), has arisen as a novel approach aimed at addressing Non-IID data challenges by leveraging Federated Learning. These methods refine the standard FL framework by distilling insights from public dataset predictions, securing data transmissions through gradient encryption, and applying differential privacy to mask individual contributions. Despite these innovations, our survey identifies persistent vulnerabilities, particularly concerning the susceptibility to logit inversion attacks where malicious actors could reconstruct private data from shared public predictions. This exploration reveals that even advanced Distillation-based Federated Learning systems harbor significant privacy risks, challenging the prevailing assumptions about their security and underscoring the need for continued advancements in secure Federated Learning methodologies.

1. Introduction

Federated Learning (FL) is a distributed learning method in which each participant independently trains a model locally and then transmits either gradients or model parameters to a centralized server. This server integrates these inputs to update a global model, facilitating collaborative learning while allowing participants to maintain the privacy of their data. However, despite these privacy measures, the method is susceptible to risks, such as the potential for a dishonest server to reconstruct individual clients' private data from the shared information. To tackle these security vulnerabilities and particularly to address the challenges posed by Non-IID data distributions, innovative strategies such as Federated Learning with Model Distillation (FedMD), Federated Gradient Encryption and Model Sharing (FedGEMS), and Differentially Secure Federated Learning (DS-FL) have been developed. These approaches enhance the robustness of FL systems against privacy risks by incorporating advanced data protection mechanisms and ensuring more uniform learning from diverse data sources.

In FedMD, proposed by Daliang Li and Junpu Wang, participants train using both private and public data [1]. However, only the predictions from public datasets—referred to as public knowledge—are shared with the server. This approach restricts the server's access to sensitive information, thus protecting against the leakage of gradient or parameter data. Additionally, FedGEMS enhances this protection by encrypting gradients exchanged during model updates, securing the transmission process against potential intercepts [2]. Meanwhile, DS-FL employs differential privacy techniques to ensure that the aggregated data shared with the server conceals any identifiable traits of the underlying data sources, providing an additional layer of security [3]. Together, these methods enhance FL's privacy and security framework while encouraging its adoption across various industries where data confidentiality is paramount.

The security implications of sharing public knowledge in Distillation-based Federated Learning have not been thoroughly explored despite its widespread adoption. This study challenges the prevailing assumption of robust security in Distillation-based Federated Learning and reveals its susceptibility to significant privacy risks, highlighting a

specific type of attack where an adversary can deduce a participant's private data solely from the public knowledge shared. The vulnerabilities of Distillation-based Federated Learning are delved into in our study, introducing potent attacks that can compromise the system and expose private and sensitive data.

2. Distillation-based Federated Learning

2.1 FedMD

In the evolving landscape of Federated Learning, where data privacy is a critical concern, FedMD represents an innovative leap forward. This method allows for collaborative model training by utilizing both private and shared public datasets while safeguarding the confidentiality of sensitive information. Participants only exchange model logits derived from the public data, ensuring that private data attributes are not exposed during the learning process [1].

The FedMD process, as outlined in Algorithm 1, starts with local training on private and public datasets. Clients then share their public logits with a central server to aggregate them into global logits. Clients use these global logits to refine their models privately while simultaneously improving the global model using public data at the server side. Through iterative rounds of this protocol, FedMD effectively balances data diversity with privacy preservation, epitomizing the core values of Distillation-based Federated Learning which prioritize data integrity and collaborative intelligence.

Input: Private datasets $\{D_k\}_{k=1}^{C}$, public dataset D_p, local models $\{f_k\}_{k=1}^{C}$, global model f_0, number of communications T.

Each client trains f_k on D_p
Each client trains f_k on D_k
for $t = 1 \leftarrow T$ **do**
    Each client sends the set of public logits $\{\boldsymbol{l}_i^k\}$
    The server computes the global logits:
        $\boldsymbol{l}_p = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{l}^k$
    Each client receives $\boldsymbol{l}_p$ and trains f_k on $\{D_p, \boldsymbol{l}_p\}$
    Each client trains f_k on D_k
    The server trains f_0 on D_p

(Algorithm 1) the pseudo-codes of FedMD [4].

2.2 FedGEMS

Advancing the paradigm of privacy-conscious machine learning, FedGEMS builds upon the principles of FedMD. It introduces a server model f_0 with an enriched parameter set W_0 [2]. This augmented server model follows the protocol outlined in Algorithm 2 and leverages the public dataset D_p, refining its parameters through focused training informed by consensus logits. These selectively refined updates enable the model to achieve higher accuracy and utility.

In this collaborative yet secure framework, clients undertake a dual training approach. They first train their local models f_k using their private data, preserving the unique characteristics inherent to their datasets. Subsequently, they refine f_k further by training on the processed public dataset D_p, under global logits $\mathbf{l}_p$. The process concludes with clients dispatching their local models' logits to the server, ensuring that only insights from the public data contribute to the global model's evolution.

Input: Private datasets $\{D_k\}_{k=1}^{K}$, public dataset D_p, local models $\{f_k\}_{k=1}^{K}$, global model f_0, number of communications T.

for $t = 1 \leftarrow T$ **do**
    The server selectively trains f_0 on $\{D_p, \boldsymbol{l}_p, \boldsymbol{l}_k\}$
    The server computes the global logits:
        $\boldsymbol{l}_i^p = f_0(W_p; x_i^p)$
    Each client trains f_k on $\{D_p, \boldsymbol{l}_p\}$
    Each client trains f_k on D_k
    Each client sends the set of public logits $\{\boldsymbol{l}_i^k\}$

(Algorithm 2) the pseudo-codes of FedGEMS [4].

2.3 DS-FL

Culminating the suite of Distillation-based Federated Learning methodologies, DS-FL takes a distinctive approach by integrating an Entropy Reduction Aggregation (ERA) technique with the utilization of an unlabeled public dataset D_p. Each client begins the process by training their local models f_k on their private datasets D_k, safeguarding the privacy of their individual data [3]. They then calculate and transmit the public logits $\mathbf{l}_k$ to the central server, following Algorithm 3.

The server utilizes the ERA method to process received logits and produce a refined set of global logits $\mathbf{l}_p$. This sophisticated aggregation technique is designed to decrease uncertainty in predictions, thereby improving the model's decisiveness. After this aggregation, clients embark on a second round of training, integrating the global logits with the public dataset to refine their local models f_k. This training enhances the local models by incorporating globally acquired patterns while safeguarding the privacy of individual data.

Simultaneously, the global server model f_0 is also trained on the public dataset D_p, benefiting from the ERA-aggregated global logits $\mathbf{l}_p$. This strategy enables the server model to better calibrate its predictions, significantly enhancing its ability to generalize across different data inputs. The continuous refinement of both local and global models through the ERA technique exemplifies the robustness of DS-FL, making it a pivotal strategy in the realm of secure and efficient Federated Learning. This dual focus on enhancing prediction accuracy and maintaining data privacy underlines the progressive nature of DS-FL in tackling the

complex challenges of modern machine learning environments.

Input: Private datasets $\{D_k\}_{k=1}^K$, public dataset D_p, local models $\{f_k\}_{k=1}^K$, global model f_0, number of communications T.

for $t = 1 \leftarrow T$ **do**
  Each client trains f_k on $\{D_k\}$
  Each client sends the set of public logits $\{\boldsymbol{l}_i^k\}$
  The server computes the global logits:
    $\boldsymbol{l}_p = \mathrm{ERA}(\sum_{k=1}^K \frac{\boldsymbol{l}^k}{K})$
  Each client trains f_k on $\{D_p, \boldsymbol{l}_p\}$
  The server trains f_0 on $\{D_p, \boldsymbol{l}_p\}$

(Algorithm 3) the pseudo-codes of DS-FL [4].

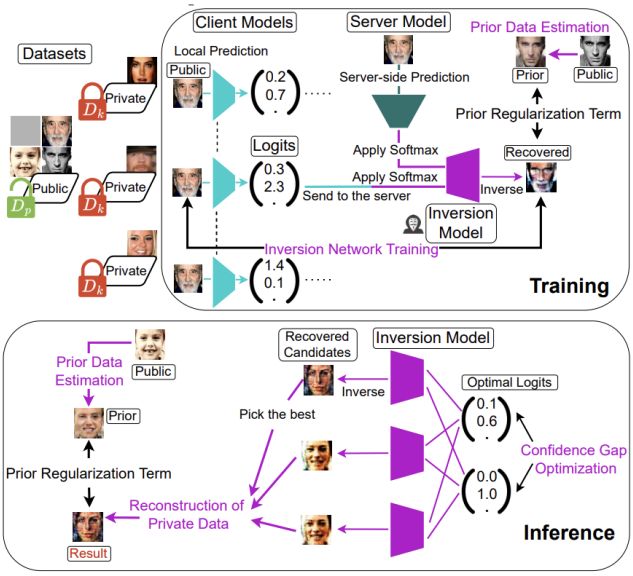
3. Logit Inversion Attacks

3.1 PLI

After examining the security concerns related to Distillation-based Federated Learning methods, we have identified a vulnerability known as the Paired-Logits Inversion (PLI) attack. This attack method exploits a confidence gap between the logits of local models trained on private data and server models limited to public data. A specialized inversion network is then utilized to reverse-engineer the logits back into their original private data forms.

The inversion process begins by establishing a baseline using public data logits. It then advances using an optimization technique to predict logits that would likely be produced by private data, followed by refining the inversion through an auxiliary prior estimation algorithm to achieve accurate reconstructions of the private data. The estimated logits for private data are introduced into the trained inversion network, enabling the PLI to effectively recover the original data with high fidelity [4]. Figure 1 in the paper illustrates the step-by-step process by which the PLI attack method reconstructs private data from public logits, detailing the progression from the initial acquisition of logits to the final stage of data recovery.

This study is significant as it exposes a substantial privacy risk in Federated Learning frameworks like FedMD, demonstrating the effectiveness of the PLI in compromising privacy. It underscores the urgent need for the development of stronger security measures in distributed learning systems to counteract such sophisticated attacks. The success of PLI, particularly in reconstructing private images from public logits across all tested benchmarks with high accuracy, calls for a reevaluation of the security assumptions in current Federated Learning implementations.



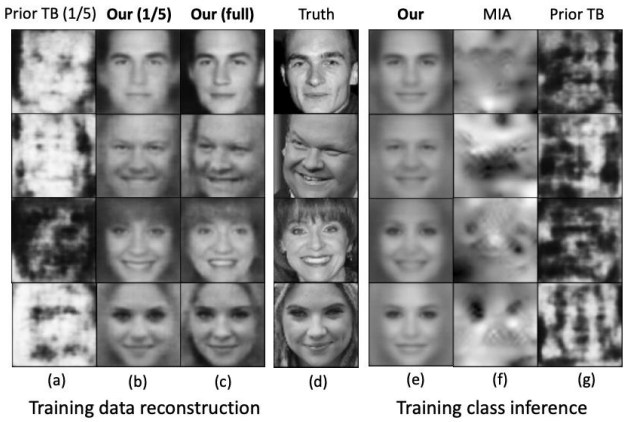
(Figure 1) Overview of PLI [4].

3.2 TBI

Delving deeper into security vulnerabilities of Distillation-based Federated Learning, we introduce another logit inversion called Training-Based Inversion (TBI). This method distinguishes itself from the PLI by employing an inversion network that leverages transposed convolutional layers. These layers are particularly adept at reconstructing input data from model predictions by effectively decoding the extracted features back into their original forms, thus facilitating reverse computation.

Unlike PLI, which directly exploits the confidence gap between client and server models, TBI builds an auxiliary model that can reverse engineer the target model's outputs without requiring access to the original training data. This approach allows for a broader application by utilizing a more generalized dataset, which can be compiled from publicly available sources such as online facial images. This dataset serves as a proxy to understand the feature distribution of the target data, enabling the adversary to bypass the need for specific access to the model's training set [5].

Figure 2 in our document illustrates the advancement of the TBI method, particularly in column (c), which demonstrates the effectiveness of TBI when applied with these auxiliary datasets. The detailed reconstructions achieved—without direct knowledge of the training data—highlight TBI's potential to accurately recover private data. This method's capability to invert logits from a general understanding of data underscores a significant privacy risk, suggesting that even models trained in seemingly secure Distillation-based Federated Learning environments are susceptible to sophisticated inversion attacks like TBI.



(Figure 2) 'Training data reconstruction' and 'Training class inference' of TBI and previous methods [5].

4. Conclusion

This research highlights a critical vulnerability in Distillation-based Federated Learning frameworks such as FedMD, FedGEMS, and DS-FL. Although these methods are designed with privacy in mind, they are still vulnerable to logit inversion attacks like Paired-Logits Inversion and Training-Based Inversion. These attacks can reconstruct private data from public logits, indicating that sharing public logits during training can lead to significant privacy breaches. This challenges the belief that these methods are secure against all forms of data leakage. Our study emphasizes the need for stronger defenses in Distillation-based Federated Learning to enhance participant data protection.

Acknowledgement

This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024. This work was supported by Inter-University Semiconductor Research Center (ISRC). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00277326). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00516, Derivation of a Differential Privacy Concept Applicable to National Statistics Data While Guaranteeing the Utility of Statistical Analysis). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the artificial intelligence semiconductor support program to nurture the best talents (IITP-2023-RS-2023-00256081) grant funded by the Korea government (MSIT).

References

[1] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581, 2019.

[2] Sijie Cheng, Jingwen Wu, Yanghua Xiao, and Yang Liu. Fedgems: Federated learning of larger server models via selective knowledge fusion. arXiv preprint arXiv:2110.11027, 2021.

[3] Sohei Itahara, Takayuki Nishio, Yusuke Koda, Masahiro Morikura, and Koji Yamamoto. Distillation-based semisupervised federated learning for communication efficient collaborative training with non-iid private data. IEEE Transactions on Mobile Computing, pages 1–1, 2021.

[4] Takahashi, H.; Liu, J.; and Liu, Y. 2023. Breaching FedMD: Image Recovery via Paired-Logits Inversion Attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 12198–12207.

[5] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19, page 225–240, New York, NY, USA, 2019. Association for Computing Machinery.