Inference Attacks on Federated Learning - A Survey

Zohar Cochavi

Delft University of Technology, The Netherlands

June 27, 2023

Introduction

Relevance

- Federated Learning has seen large-scale use since its introduction
- Inference attacks threaten one of its core principles
- The field changes quickly, and updates are necessary

Overview

1. Introduce Federated Learning and Adversarial Machine Learning
2. Discuss progress in the field over the last year (continuing work by Abad et al. (2022))
3. Conclude what this means for Federated Learning

Background

We will cover:

1. Federated Learning
2. Inference Attacks

Federated Learning I

Federated Learning (FL) is a machine learning scheme that distributes the responsibility of training a model over multiple clients and aggregates their results into a single model (McMahan and Ramage 2017).

- $\uparrow$ Better privacy guarantees
- $\uparrow$ Distribution of resources
- $\downarrow$ More resources in total

Federated Learning II

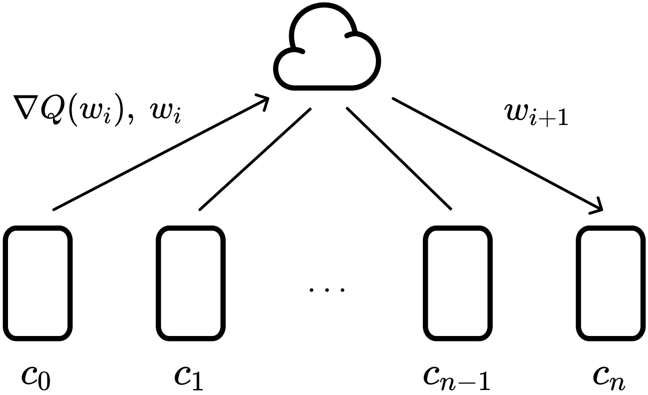

Figure 1: Typical Federated Learning network topology. The client, c_i, sends the gradient, $\nabla Q(w_i)$, and/or weights, w_i, of a particular iteration i. The central server then returns the updated model parameters w_{i+1}.

Inference Attacks I

We categorize inference attacks according to the following properties (Abad et al. 2022).

Adversarial goal

- *Model Inversion*: find data points by the label.
- *Membership Inference*: determine the presence of a data point in the local training data.
- *Property Inference*: determine presence property p of the data or model.

Interference with learning

- *Passive*: does not interfere with the learning process.
- *Active*: interferes with the learning process to gain more information.

Passive attacks are more stealthy, but *active* attacks are stronger.

Inference Attacks II

Position of the adversary

- *Local*: adversary is a client.
- *Global*: adversary is the central server.

Often, the information required is available to both and an attack considers a *local/global* scenario. Meaning it could be both.

Inference Attacks in Federated Learning

We will discuss three of the most interesting ones.

1. Do Gradient Inversion Attacks Make Federated Learning Unsafe?
2. Active Membership Inference Attack under Local Differential Privacy in Federated Learning.
3. Subject Membership Inference Attacks in Federated Learning

Do Gradient Inversion Attacks Make Federated Learning Unsafe?

- Hatamizadeh et al. (2023) explore image reconstruction using gradient inversion while relaxing the assumption made in prior work regarding Batch Normalization (BN).
- Previous studies assumed static BN statistics, but the authors successfully reconstructed images without relying on this assumption.
- Inversion attacks can be practical for accurate reconstructions but still require priors (approximations of the image) for higher accuracy.

$\Rightarrow$ Attack that more closely resembles a real-world scenario.

Active Membership Inference Attack under Local Differential Privacy in Federated Learning.

- Nguyen et al. (2023) introduces an *active* membership inference attack, allowing them to infer membership of a specific data point in the presence of differential privacy.
- Differential privacy is a technique that obscures an individual's relation to a data point while preserving the patterns used for training machine learning models (Dwork and Roth 2013).
- The attack performance starts to degrade only when the level of data obscuring interferes with the model's performance, indicating the need for more robust privacy methods to counter such attacks.

$\Rightarrow$ Raises questions about the efficacy of Differential Privacy.

Subject Membership Inference Attacks in Federated Learning

- In a black-box setting, the paper by (Suri et al. 2022) proposes a method called "Subject Inference" for inferring the presence of individuals, or "subjects," in a dataset.
- Previous work in this area is criticized for being disconnected from real-world scenarios as it includes information adversaries would not normally have access to and assumes the adversary is looking for data points rather than individuals.
- The authors demonstrate the effectiveness of Subject Inference in various real-world datasets, emphasizing its realistic nature and highlighting it as a significant threat to user privacy.

⇒ An attack crafted to reflect a real-life scenario for a *cross-silo* FL configuration.

Defenses

Various novel defenses are proposed,

- The use of image augmentation to enhance privacy (Shin et al. 2023)
- Using a built-in adversary (Li et al. 2022)

as well as suggestions to counterattack proposed attacks,

- Increase batch size to mask local contributions (Geng et al. 2023; Hatamizadeh et al. 2023)
- Use alternative aggregation methods such as FedAvg and FedBN (Geng et al. 2023; Hatamizadeh et al. 2023)

Another promising option that has not been investigated in this work is *Homomorphic Encryption* (Lee et al. 2022).

Future Work

1. **Utilize existing preprocessing methods to enhance privacy preservation**, as demonstrated by studies such as Shin et al. (2023). Use generalization to the advantage of privacy.
2. **New attack methods should prioritize relaxing assumptions** to provide a more realistic assessment of privacy-preserving features in Federated Learning (FL).
3. **Developing secure Homomorphic Encryption (HE) techniques would significantly mitigate many of the attacks discussed**. Encrypting data before training models would render inference attacks harmless.

Conclusion

- Threats to current Federated Learning because of more realistic scenarios
- Privacy Enhancing technologies can be circumvented
- More research is necessary to assess whether FL is adequately privacy-preserving

References I

Abad, Gorka, Stjepan Picek, Víctor Julio Ramírez-Durán, and Aitor Urbieta. 2022. "On the Security & Privacy in Federated Learning." arXiv. https://arxiv.org/abs/2112.05423.

Dwork, Cynthia, and Aaron Roth. 2013. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9 (3-4): 211–407. https://doi.org/10.1561/0400000042.

Geng, Jiahui, Yongli Mou, Qing Li, Feifei Li, Oya Beyan, Stefan Decker, and Chunming Rong. 2023. "Improved Gradient Inversion Attacks and Defenses in Federated Learning." *IEEE Transactions on Big Data*, 1–13. https://doi.org/10.1109/TBDATA.2023.3239116.

Hatamizadeh, Ali, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, et al. 2023. "Do Gradient Inversion Attacks Make Federated Learning Unsafe?" *IEEE Transactions on Medical Imaging*, 1–1. https://doi.org/10.1109/TMI.2023.3239391.

References II

Lee, Joon-Woo, Hyungchul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, et al. 2022. "Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network." *IEEE Access* 10: 30039–54. https://doi.org/10.1109/ACCESS.2022.3159694.

Li, Jingtao, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. 2022. "ResSFL: A Resistance Transfer Framework for Defending Model Inversion Attack in Split Federated Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10194–202.

McMahan, Brendan, and Daniel Ramage. 2017. "Federated Learning: Collaborative Machine Learning Without Centralized Training Data." https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

References III

Nguyen, Truc, Phung Lai, Khang Tran, NhatHai Phan, and My T. Thai. 2023. "Active Membership Inference Attack Under Local Differential Privacy in Federated Learning." arXiv. https://doi.org/10.48550/arXiv.2302.12685.

Shin, Seunghyeon, Mallika Boyapati, Kun Suo, Kyungtae Kang, and Junggab Son. 2023. "An Empirical Analysis of Image Augmentation Against Model Inversion Attack in Federated Learning." *Cluster Computing* 26 (1): 349–66. https://doi.org/10.1007/s10586-022-03596-1.

Suri, Anshuman, Pallika Kanani, Virendra J. Marathe, and Daniel W. Peterson. 2022. "Subject Membership Inference Attacks in Federated Learning." *arXiv.org*. https://arxiv.org/abs/2206.03317v3.