Protecting Privacy using Generative Models in Federated Learning

Contact Person: Zhuan Shi (Email: zhuan.shi@epfl.ch)
Simin Fan (Email: simin.fan@epfl.ch)

Advisor: Prof. Boi Faltings and Prof. Martin Jaggi

1 Project Overview

1.1 Motivation and Background

The continuous scaling-up of data and models demonstrates impressive downstream performance across various fields, while imposing a pressing demand for training/inference efficiency and privacy. Based on the scheme of decentralized training, federated learning (FL) distributes the training work across multiple local clients while aggregate the local gradients to update the central model, which makes most of various data sources while protecting the privacy. However, the FL framework with differentiable privacy is not flawless: one the one hand, some privacy attacks are still able to recover the original local training data from the gradient, even mixed with noises; one the other hand, the local data can hardly be reused in the future training under privacy requirements, which limits the potential usage of the existing data resources.

To mitigate the vulnerability of differentiable privacy and improving the effectiveness of data usage, we propose a new paradigm of Generative Privacy (GP). By learning the real data distribution across various clients, a central generator are trained to generate synthetic training data. Instead of the original distributed mode, the central model would be trained on the synthetic training set from the central generator without access to the real local data. In that way, no real data are used for training, which protects the local privacy. Besides, the synthetic data could be safely released for future usage, without any break of privacy requirements.

1.2 Goal

In this project, we expect to apply modern foundation models to develop generative privacy of federated learning system. The final objectives are two-folds:

- **Performance**: improve the accuracy of central model with augmented, high-quality synthetic data;
- **Privacy**: improve the robustness of FL system on various types of privacy attacks. Theoretical proof of privacy bounds is a stretch goal.

2 Project Steps

1. **Modeling**: train the diffusion model as synthetic image generator in a federated learning paradigm. Then use the synthetic images to train the central classifier;
2. **Assessment**: evaluate the privacy and accuracy of our framework. Compare with different baselines;
3. **Analysis & Refinement**: analyze theoretical privacy guarantees. Refine the design of generators (e.g. re-define the noise scheduler in diffusion model to satisfy the privacy bounds etc.).

3 Prerequisites

- Familiar with federated learning, differentiable privacy and basic of generative models ([1, 2, 3, 4, 5, 6]). Experience with diffusion model would be a plus;
- Strong programming skills (proficiency on Pytorch);
- Fluent English in writing and communication;
- Self-motivated to learning new things and collaboration in a teamwork. We are targeting at a publishable work, so we hope you are also excited to that ;)

4 Postscript

- This project is available for both master thesis and semester project;
- Previous publications would be a plus but not a requirement;
- The host lab will provide the computational resources for the project.

References

[1] A. Triastcyn and B. Faltings, "Federated generative privacy," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 50–57, 2020.

[2] W. Zhuang, C. Chen, and L. Lyu, "When foundation model meets federated learning: Motivations, challenges, and future directions," *arXiv preprint arXiv:2306.15546*, 2023.

[3] S. Ghalebikesabi, L. Berrada, S. Gowal, I. Ktena, R. Stanforth, J. Hayes, S. De, S. L. Smith, O. Wiles, and B. Balle, "Differentially private diffusion models generate useful synthetic images," *arXiv preprint arXiv:2302.13861*, 2023.

[4] A. Triastcyn and B. Faltings, "Generating artificial data for private deep learning," *arXiv preprint arXiv:1803.03148*, 2018.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.

[6] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021.