PRISM-FL: Privacy-preserving Image Synthesis Mechanism for Federated Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Federated learning (FL) enables collaborative training across decentralized clients without sharing raw data, offering strong appeal for privacy-sensitive domains such as healthcare, mobile personalization, and finance. However, traditional FL is still vulnerable to privacy leakage through gradient inversion and model updates, and even communication traces can reveal sensitive information. Differential privacy (DP) has emerged as a formal solution, yet its integration into FL often results in degraded accuracy, limited flexibility, and high communication cost. Moreover, existing DP-based generative methods, such as DP-GAN or DP-MERF, can produce samples visually similar to private data, raising further privacy and ethical concerns.

We propose PRISM-FL, a decentralized framework for obfuscated synthetic image generation under client-level differential privacy. Instead of exchanging weights or gradients, each client trains a local model with DP-SGD and extracts noisy feature statistics. Public images are then optimized to match these statistics, producing synthetic datasets that preserve the distributional characteristics of private data while remaining visually distinct. Clients may optionally share these synthetic datasets with one another or a central server, enriching the training pool with diverse yet privacy-preserving samples. Downstream models can then be trained locally or in federated setups, benefiting from improved generalization and reduced heterogeneity.

PRISM-FL is designed to address three persistent challenges in FL: privacy leakage (by shifting from parameter sharing to DP-guided synthesis), utility degradation (by leveraging public data alignment to retain discriminative features), and communication overhead (by exchanging compact synthetic datasets instead of iterative model updates). Preliminary experiments on MNIST, CIFAR-10, and CelebA-Hair, under both i.i.d. and non-i.i.d. splits with 5–50 clients and privacy budgets ranging from $\epsilon = 0.2$ to 5, suggest that PRISM-FL achieves competitive accuracy compared to standard DP-FL baselines. Synthetic samples remain visually dissimilar from private data, and the approach shows potential for reducing communication cost through one-shot synthetic exchange.

These initial results highlight PRISM-FL as a practical alternative to weight or gradient sharing in federated learning, particularly in domains where raw data cannot be shared and heterogeneity is high. Future directions include extending the method to multimodal synthesis (e.g., medical imaging and clinical text), exploring adaptive privacy budgets based on client behavior, and incorporating secure aggregation protocols for synthetic statistics. Together, these efforts could further enhance the trustworthiness, efficiency, and applicability of privacy-preserving federated learning.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.