Combatting Membership Inference Attacks in Federated Learning by Adding Noise on Client Confidence Scores

Shih-Hsuan Yang
Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan

Abstract

Federated learning (FL) reduces the need to centralize raw training data, but the information exchanged during training can still reveal whether a specific sample was used by a client model. Existing FL defenses based on output perturbation mainly protect the server-side global model, leaving client-side models vulnerable to membership inference. This paper proposes a client-side defense that generates adversarial-example-inspired perturbations on the confidence scores produced by local models. The perturbation is optimized to preserve the original predicted label, limit output distortion, and drive the attack decision toward random guessing. We evaluate the proposed method on FL image-classification tasks using CIFAR100 and CIFAR10. Across both datasets, the defense reduces attack accuracy from 61.87%–72.40% in single-epoch attacks and from 70.98%–86.74% in multi-epoch attacks to near-chance performance, typically around 50%–55%. These results show that carefully designed perturbations on client confidence scores can substantially mitigate client-side membership leakage in FL.