A2E: BLACK-BOX ANTI-ADVERSARIAL WATERMARKING FOR VERIFIABLE FEDERATED UNLEARNING

Author's : Aakash[1], Dr. K. Sujith[2]

Abstract

Federated Learning (FL) enables multiple distributed clients to collaboratively train a global machine learning model without sharing raw data, thereby preserving privacy and regulatory compliance. However, recent legal and ethical requirements such as the "right to be forgotten" demand mechanisms to remove the influence of specific clients or data contributors from trained models—a process known as Federated Unlearning (FU). While several unlearning algorithms have been proposed, verifying whether a model has genuinely forgotten the targeted client remains a challenging problem, particularly in black-box settings where internal model parameters are inaccessible.This project proposes **A2E (Black-Box Anti-Adversarial Example Based Watermarking)** as a robust verification mechanism for federated unlearning. The approach embeds a specially designed watermark into the model during federated training. The watermark consists of trigger-based key samples that are resilient to black-box adversarial attacks through anti-adversarial augmentation. By integrating adversarially perturbed watermark samples during training, the watermark signal becomes robust against query-based attacks intended to invalidate ownership verification. After unlearning, the watermark detection rate is evaluated under both clean and adversarial query conditions. A significant drop in watermark accuracy serves as evidence that the targeted client's contribution has been successfully removed.The project implements a synthetic federated learning environment using a neural network classifier, label-skewed client partitioning, black-box adversarial attack simulation via gradient-free estimation, and federated retraining-based unlearning. Experimental results demonstrate that A2E improves watermark robustness under black-box attacks while enabling reliable verification of federated unlearning. This framework contributes to trustworthy AI by integrating privacy, accountability, and robustness into federated systems.