Dual Defense: Enhancing Privacy and Mitigating Poisoning Attacks in Federated Learning

Runhua Xu*, Shiqi Gao*, Chao Li^, James Joshi§, Jianxin Li*

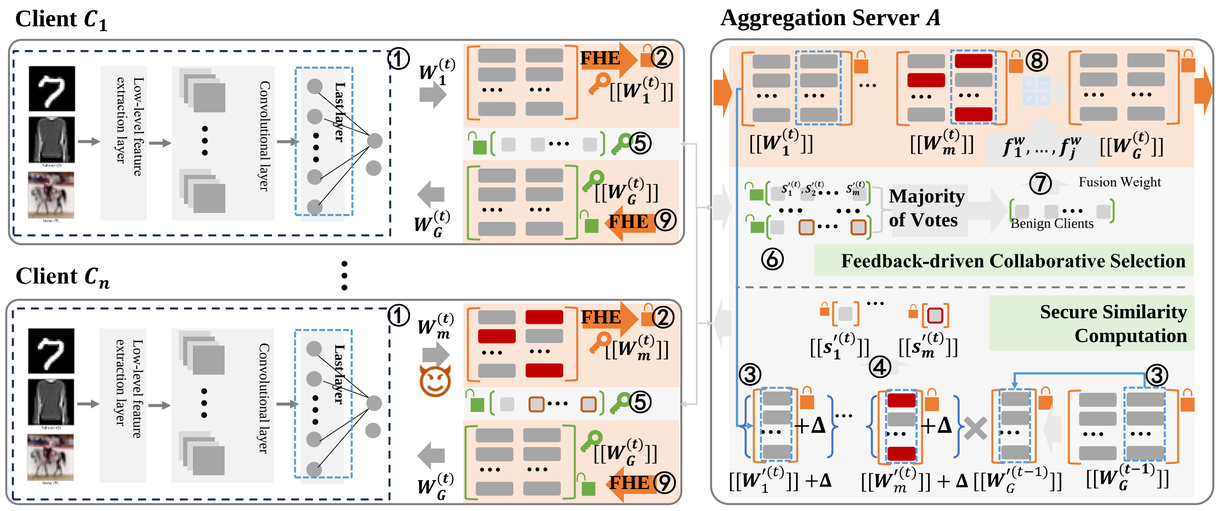
*Beihang University ^Beijing Jiaotong University §University of Pittsburgh

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Federated learning (FL) encounters significant challenges related to privacy and poisoning attacks. Secure aggregation enhances data privacy but complicates anomaly detection, as most methods require unencrypted model updates. Current solutions often depend on impractical non-colluding two-server setups or three-party computations, limiting scalability. To address these issues, we introduce the Dual Defense Federated Learning (**DDFed**) framework, which improves privacy and mitigates poisoning attacks without changing FL topology or adding new roles.

Dual Defense Federated Learning Framework



FHE-based Secure Aggregation

DDFed necessitates that all clients pre-process their inputs for normalization and shifts the task of comparing similarity scores to the client side. This is because clients possess the FHE private key, allowing them to obtain the similarity score in plaintext:

$$[cos(\alpha_i)] = \frac{\langle [\mathbf{W}_i^{(t)}], [\mathbf{W}_G^{(t-1)}] \rangle}{\| [\mathbf{W}_i^{(t)}] \|_2 \cdot \| [\mathbf{W}_G^{(t-1)}] \|_2} = \langle \left[\frac{\mathbf{W}_i^{(t)}}{\| \mathbf{W}_i^{(t)} \|_2}\right], \left[\frac{\mathbf{W}_G^{(t-1)}}{\| \mathbf{W}_G^{(t-1)} \|_2}\right] \rangle,$$

where each client $\mathcal{C}_i$ prepares the $\frac{\mathbf{W}_i^{(t)}}{\|\mathbf{W}_i^{(t)}\|_2}$ and $\frac{\mathbf{W}_G^{(t-1)}}{\|\mathbf{W}_G^{(t-1)}\|_2}$ in advance.

The aggregation server $\mathcal{S}$ verifies received $\left[\frac{\mathbf{W}_G^{(t-1)}}{\|\mathbf{W}^{(t-1)}\|_0}\right]$ and perturb local inputs and conducts secure inner-product computation as follows:

$$[\mathbf{s}'^{(t)}] = \langle \left[\frac{\mathbf{W}_i^{(t)}}{\|\mathbf{W}_i^{(t)}\|_2}\right] + \Delta^{(t)}, \left[\frac{\mathbf{W}_G^{(t-1)}}{\|\mathbf{W}_G^{(t-1)}\|_2}\right] \rangle.$$

FHE-based Secure Aggregation

To mitigate this privacy risk, DDFed improves secure inner-product computation by introducing perturbations $(\varepsilon, \delta)-$ differential privacy with a Gaussian mechanism) into each normalized and encrypted model update.

Contributions

➡️ We introduce a dual defense strategy that enhances privacy and combats poisoning attacks by integrating FHE-based secure aggregation with a similarity-based detection mechanism for malicious encrypted models.

➡️ A novel two-phase anomaly detection mechanism is proposed, featuring safeguards against privacy breaches from Byzantine clients, along with a clipping technique to strengthen defenses against diverse poisoning attacks.

➡️ Comprehensive experiments across multiple poisoning scenarios validate DDFed's effectiveness in protecting model privacy and defending against threats.

Experiments

Under the IPM attack scenario, DDFed method achieves the ***best comprehensive defense performance***.The same conclusion also holds true in the ALIE attack. And DDFed is resilient to poisoning attacks from the beginning of training. Our design is ***not constrained by the attack's initiation round***.

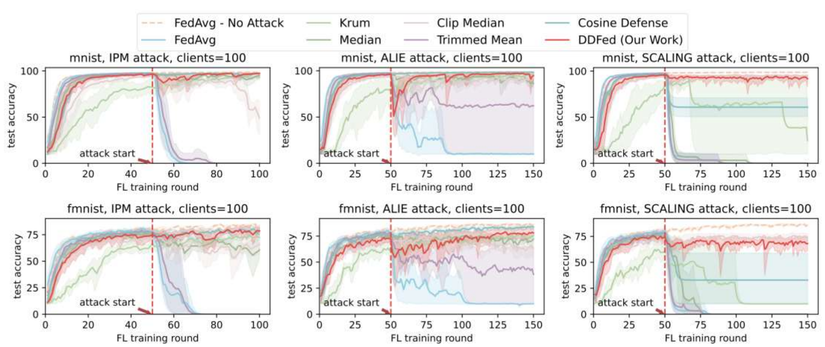


Figure 2: Comparison of defense effectiveness across various defense approaches, evaluated on MNIST (top) and FMNIST(bottom), under IPM attack (left), ALIE attack (middle), and SCALING attack (right).

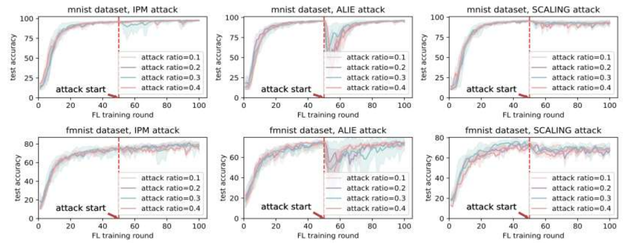


Figure 4 Comparison of effectiveness across different attack ratios, evaluated on MNIST (top) and FMNIST (bottom), under IPM attack (left), ALIE attack (middle), and SCALING attack(right).

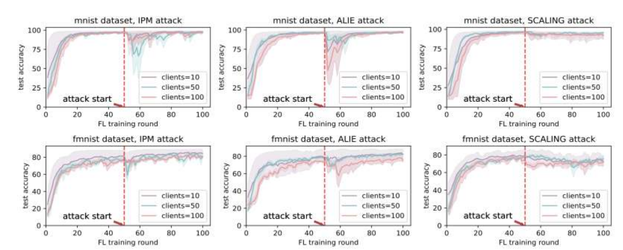


Figure 5 Comparison of effectiveness across different client numbers, evaluated on MNIST (top) and FMNIST (bottom), under IPM attack (left), ALIE attack (middle), and SCALING attack (right).

Table 1: Time cost per training round of various defense approaches.

Approaches	MNIST, IPM attack		FMNIST, IPM attack	
	avg (s)	var (s)	avg (s)	var (s)
FedAvg	10.26	0.07	10.47	0.01
Krum	10.32	0.03	10.26	0.01
Median	10.32	0.01	10.28	0.02
Clipping Median	10.31	0.01	10.32	0.01
Trimmed Mean	10.32	0.02	10.30	0.01
Cos Defense	10.25	0.01	10.26	0.02
DDFed (Our Work)	12.43	0.01	12.14	0.01

DDFed generally requires an extra 2 seconds compared to the usual 10-second training round, resulting in a ***20% increase*** in time per training round. However, our DDFed is capable of defending against model poisoning attacks while also ***offering strong privacy guarantees.***

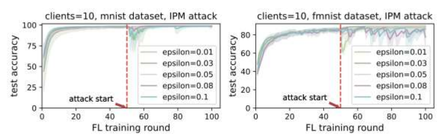


Figure 5. Impact of hyper-parameter ϵ of differential privacy based perturbation at secure similarity computation phase, evaluated on MNIST (left) and FMNIST (right), under IPM attack.

北京航空航天大學 BEIHANG UNIVERSITY

北京交通大學 BEIJING JIAOTONG UNIVERSITY

University of Pittsburgh