Poisoning Attacks and Defenses to Federated Unlearning

Wenbin Wang*[1], Qiwen Ma*[2], Zifan Zhang[3], Yuchen Liu[3], Zhuqing Liu[4], Minghong Fang[5]

[1]ShanghaiTech University [2]Xidian University [3]North Carolina State University
[4]University of North Texas [5]University of Louisville

Federated Learning (FL) and Federated Unlearning (FU)

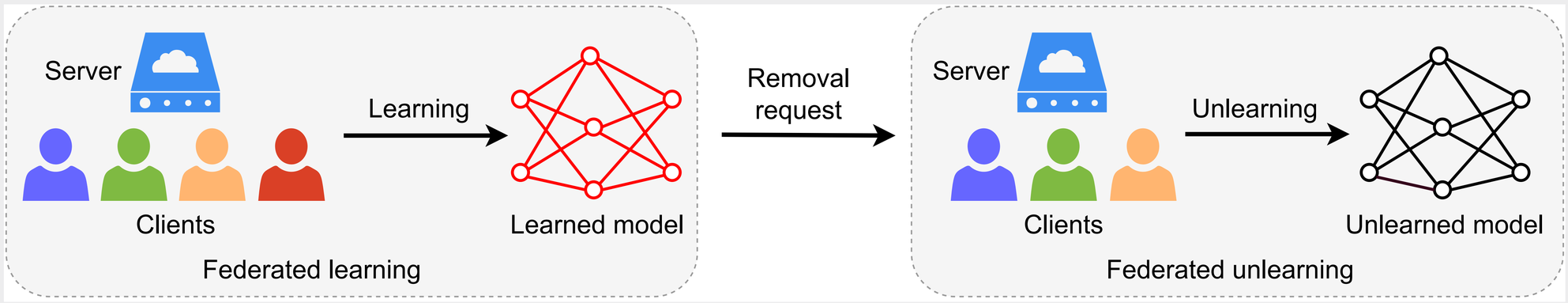

Figure: Illustration of federated learning and federated unlearning processes.

Federated Learning

- **Step I (Global model broadcasting):** The server distributes the current global model $\mathbf{w}^t$ to the participating clients.
- **Step II (Local model training):** Each client i trains its local model using the global model $\mathbf{w}^t$ and computes its update as $\mathbf{g}_i^t = 1/|B_i^t| \sum_{k \in B_i^t} \nabla \mathcal{L}_i(\mathbf{w}^t, D_{i,k})$. Here, $D_{i,k} \sim D_i$ represents the k-th training example in the mini-batch B_i^t at client i. Client i then transmits $\mathbf{g}_i^t$ back to the server.
- **Step III (Global model updating):** Once the server receives the model updates from the clients, it applies a specific aggregation rule, ARR, to combine these updates into a global model update, which is then used to update the global model accordingly: $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \cdot \text{ARR}\{\mathbf{g}_i^t : i \in [n]\}$, where η is the learning rate.

Poisoning Attacks to Federated Learning An attacker can disrupt the training process of an FL system by gaining control over certain malicious clients. These could either be fake clients injected by the attacker or benign clients that have been compromised. Poisoning attacks include untargeted and targeted attacks.

Federated Unlearning Machine unlearning is a growing area of research focused on developing methods to erase specific data points from a model after training. Recently, limited research has focused on federated unlearning, which aims to mitigate the influence of malicious clients after the global model has been trained. Unlike machine unlearning, which eliminates specific training data, FU is concerned with removing all clients from the system.

BadUnlearn and UnlearnGuard

BadUnlearn

- **Objective:** Ensure the unlearned model remains poisoned (similar to the original compromised model).
- **Attack Strategy:**

$$\min \|^{\#}\mathbf{w} - \text{ARR}\{\mathbf{g}_i^t : i \in [n]\}\|_2$$

where $\text{ARR}\{\mathbf{g}_i^t : i \in [n]\}$ denotes the aggregated model update after the attack during the FU process, $^{\#}\mathbf{w}$ denotes the learned model. Let $\mathcal{B}$ represent the set of malicious clients; then for each malicious client $k \in \mathcal{B}$, we have $\mathbf{g}_k^t = {}^{\#}\mathbf{w} + \epsilon\psi$ (where ϵ is a scaling factor and ψ is a perturbation vector, typically defined as $\psi = -\text{sign}(^{\#}\mathbf{w})$).

UnlearnGuard

- **UnlearnGuard-Dist**: A distance-based calibration technique designed to minimize estimation errors:

$$\max_{t_1 \in [t-r,t]} \|\tilde{\mathbf{g}}_i^t - \tilde{\mathbf{g}}_i^{t_1}\|_2 \leq \max_{t_2,t_3 \in [t-r,t]} \|\tilde{\mathbf{g}}_i^{t_2} - \tilde{\mathbf{g}}_i^{t_3}\|_2$$

- **UnlearnGuard-Dir**: A direction-based calibration method:

$$\max_{t_1 \in [t-r,t]} \cos(\tilde{\mathbf{g}}_i^t, \tilde{\mathbf{g}}_i^{t_1}) \leq \max_{t_2,t_3 \in [t-r,t]} \cos(\tilde{\mathbf{g}}_i^{t_2}, \tilde{\mathbf{g}}_i^{t_3})$$

Our UnlearnGuard methods aim to unlearn an accurate global model by removing the influence of malicious clients after they are detected at the end of the FL process.

Result

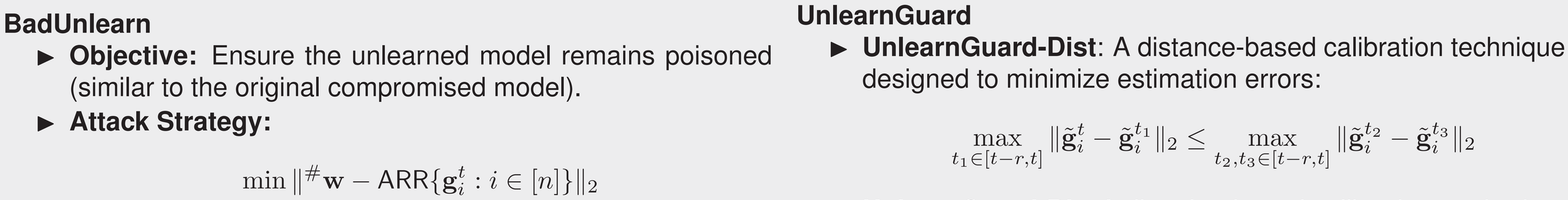

Figure: Impact of degree of Non-IID, percentage of malicious clients during FU, and the number of r.

Table: Results of FU methods on the MNIST dataset, where the attacker manipulates both FL and FU processes.

Attack FL	Attack FU	ARR	FedRecover	UnlearnGuard-Dist	UnlearnGuard-Dir
Trim attack	No attack	FedAvg	0.05	0.06	0.05
		Median	0.11	0.14	0.09
		TrMean	0.15	0.12	0.06
	Trim attack	FedAvg	0.06	0.07	0.06
		Median	0.15	0.14	0.10
		TrMean	0.16	0.15	0.07
	BadUnlearn	FedAvg	0.24	0.07	0.06
		Median	0.39	0.14	0.09
		TrMean	0.23	0.14	0.06

Attack FL	Attack FU	ARR	FedRecover	UnlearnGuard-Dist	UnlearnGuard-Dir
Backdoorattack	No attack	FedAvg	0.05 / 0.00	0.02 / 0.00	0.02 / 0.00
		Median	0.12 / 0.01	0.11 / 0.01	0.10 / 0.00
		TrMean	0.07 / 0.00	0.06 / 0.00	0.06 / 0.00
	Trim attack	FedAvg	0.06 / 0.01	0.03 / 0.01	0.03 / 0.00
		Median	0.12 / 0.00	0.12 / 0.00	0.11 / 0.00
		TrMean	0.04 / 0.00	0.06 / 0.00	0.05 / 0.00
	BadUnlearn	FedAvg	0.04 / 0.98	0.03 / 0.01	0.03 / 0.00
		Median	0.10 / 0.02	0.13 / 0.00	0.12 / 0.00
		TrMean	0.08 / 0.03	0.07 / 0.00	0.07 / 0.00

Reference

[1] Wenbin Wang*, Qiwen Ma*, Zifan Zhang, Yuchen Liu, Zhuqing Liu, and Minghong Fang. Poisoning Attacks and Defenses to Federated Unlearning. In Proc. The Web Conference (WWW), 2025 (*co-primary authors).

Contact

wangwb2023@shanghaitech.edu.cn
22012100043@stu.xidian.edu.cn
https://wenbinwang12.github.io/