Conv-LoRA Based Fairness Poisoning in Federated Learning

Pranav Somase, Venugopal Bhamidi, Prof. Manisha Padala

INTRODUCTION

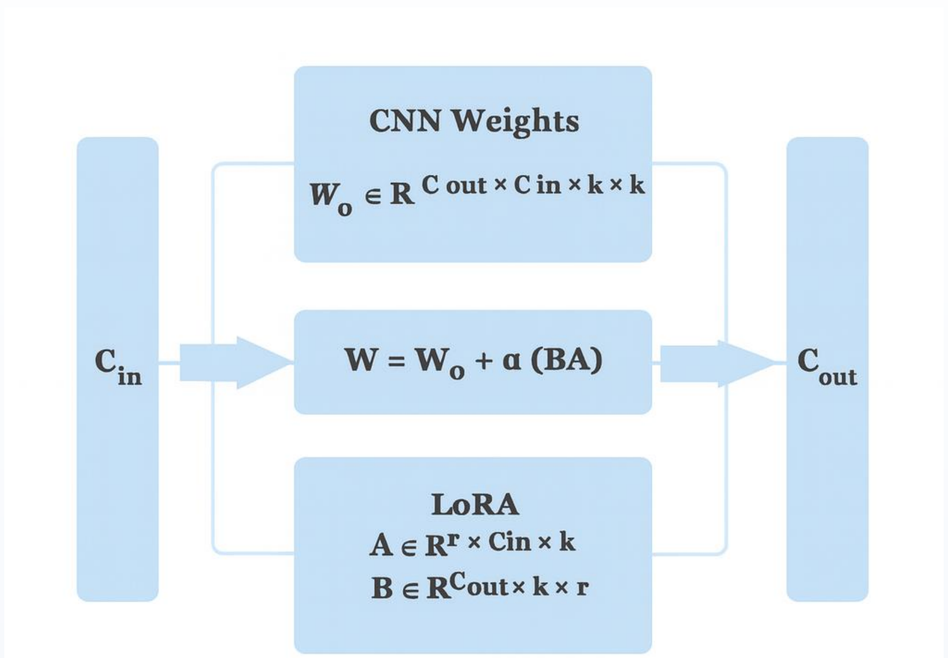
Background

- Federated Learning (FL) trains models across clients without sharing raw data.
- Fairness issues arise when predictions differ across demographic groups.
- Attackers can exploit FL's decentralized training to poison fairness.

Key Idea

- We introduce a **Conv-LoRA–based adversarial attack** that:
  - Injects **low-rank convolutional adapters** into Convolutional Layers
  - Optimize these adapters to **quietly amplify fairness disparities** while preserving accuracy and being close to global model
  - substantially reduce fairness metrics (DP, EO), and bypass standard defenses

Conv - LoRA



Integrating Low-Rank Adapters into CNNs :

Given convolutional weights :

$$W_0 \in R^{cout \times cin \times k \times k}$$

LoRA-C Branch

We introduce a trainable low-rank residual:

$$\Delta W = BA$$

$$A \in R^{r \times cin \times k}$$

$$B \in R^{cout \times k \times r}$$

The adapted convolution kernel becomes:

$$W = W_0 + \alpha\,(BA)$$

where **α** scales the contribution of LoRA parameters.

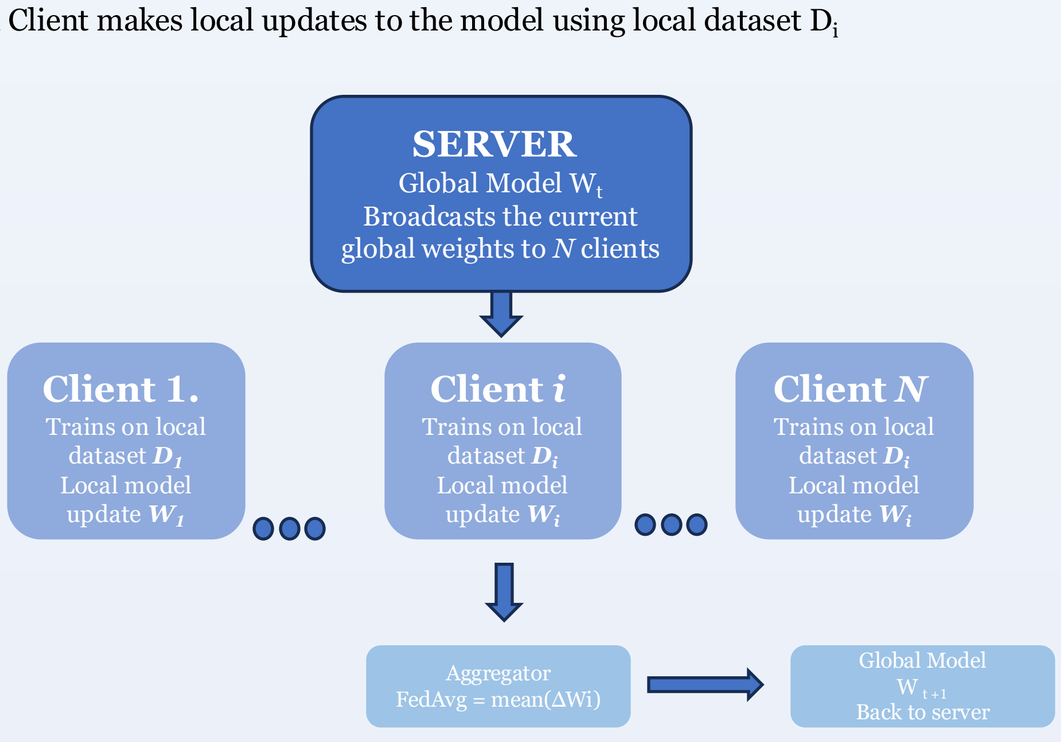
Finally, the convolution forward pass becomes:

$$y = (W_0 + \alpha BA) \otimes x$$

Federated Learning setup

Server maintains the **global model,** W to all participating clients (Adversarial and Honest) at each communication round.

Each Client makes local updates to the model using local dataset D_i



Adversarial Training

At the start of each training communication round, every client receives the current global parameters. Each client performs local learning using its own data. Two update paths exist:

(a) Honest Clients -->

- Train all layers of the model on local samples.
- Produce updated weights $W_i(t)$ after several epochs.

(b) Conv-LoRA Clients

- Keep all backbone parameters fixed at W_t
- Update only the low-rank convolutional adapters A, B using a two phase training loop :
  - Phase 1. (OPTREG on ℓ_REG): train adapters to keep model accuracy good (so the attack is stealthy).
  - *Phase 2* (OPTF on -∇ℓ_F): optimize adapters to maximize a fairness loss ℓ_F (compromise fairness).
- Reconstruct a compatible weight update using:

$$W = W_0 + \alpha\,(BA)$$

Both client types send their resulting models back to the server.

Server Aggregation

1. The aggregator aggreates all model updates into a single global update. A weighted average (FedAvg) is applied using client dataset sizes. The aggregated model becomes the new global state. Above steps repeat for multiple communication rounds.

Fairness Metrics Used :

1. Demographic Parity (DP) : Ensures that the model predicts each class at similar rates across sensitive groups.

$$P(\hat{y} = c \mid A=g_i) \approx P(\hat{y} = c \mid A=g_j),\ \forall c \in C$$

2. Equalized Odds (EO) : Requires that the model has similar true positive and false positive rates across groups.

$$TPRc,g_i \approx TPRc,g_j, \quad FPRc,g_i \approx FPRc,g_j,\ \forall c$$

Evaluates fairness conditioned on the true label, making it stronger than DP. Adapters can reduce fairness by altering group-specific error rates.

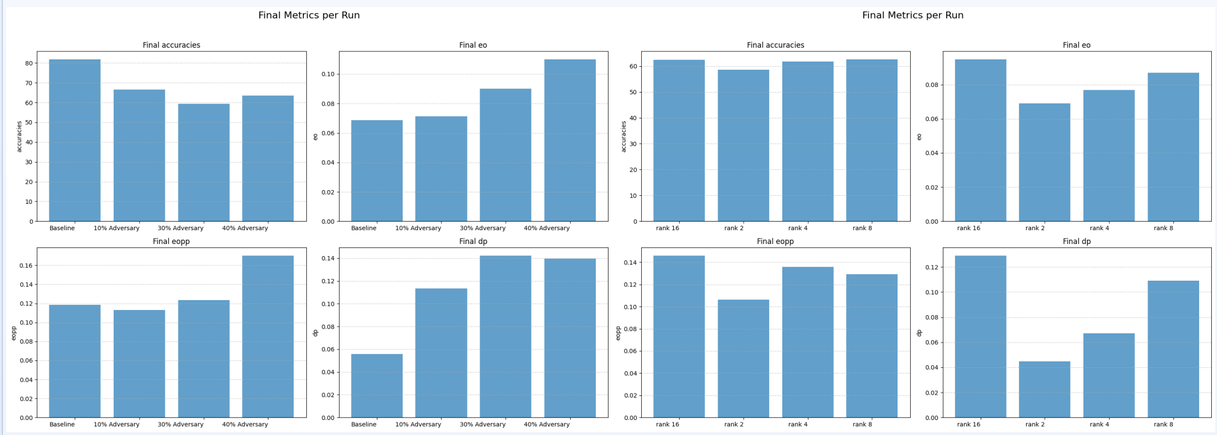
3. Equal Opportunity : A relaxed version of Equalized Odds, focusing only on True Positive Rate (TPR).

$$TPRc,g_i \approx TPRc,g_j$$

Highlights whether one group is systematically denied positive predictions relative to another.

Experimental Results

We conducted experiments on the UTKFace dataset—containing age, binary gender, and ethnicity annotations—using gender as the sensitive attribute, with the dataset partitioned into client subsets to reflect heterogeneous federated data.



Model Configuration

- **Backbone:** ResNet-50 initialized with standard ImageNet weights.
- **Conv-LoRA hyperparameters:**
  - Rank r∈{16, 2, 4, 8}, Scaling coefficient α=16 applied selectively to convolutional layers.
  - Adversary percentage : 0%(all honest clients baseline) , 10 %, 30 %, 40 %.

Across multiple runs with varying LoRA ranks, the proposed Conv-LoRA on ResNet-50 model demonstrates stable classification accuracy while significantly increasing fairness disparities. As shown in the bar charts,accuracy remains consistently high across all configurations, whereas fairness metrics—Equalized Odds (EO), Equal Opportunity (EOPP), and Demographic Parity (DP)—show clear fairness compromises for low ranks.

When varying fraction of adversarial clients, fairness gaps grow progressively larger. Compared to the baseline, even mild perturbation introduces noticeable disparity, and higher perturbation levels produce substantial degradation across all fairness indicators.

References

- Damle, S. et al. *LoRA-FL:A Low-Rank Adversarial Attack for Compromising Group Fairness in Federated Learning, 2025*.
- Ding, C. et al. *LoRA-C: Parameter-Efficient Fine-Tuning of Robust CNN for IoT Devices*. IEEE Internet of Things Journal, 2024.