Gradient Purification: Defense Against Data Poisoning Attack in Decentralized Federated Learning

Bin LI, Xiaoye MIAO, Yan ZHANG, Jianwei YIN

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50240-3

Problems & Ideas

- ***Problems*** of poisoning attack in decentralized federated learning.
  - Existing defenses cannot fully detect and mitigate malicious impact per iteration, especially in the early iterations.
  - Complete removal of malicious clients lose beneficial knowledge, thereby limiting the model's maximum achievable accuracy, especially in non-IID settings.

Client	Benign Client	Malicious Client
Data Label Distribution	(0, 1, 2, 3, 4, 5)	(6, 9) (7,8)

Fig. 2 An example of clients' data label distribution.

For example, excluding the malicious client mitigates the poisoning attack on labels 6 and 9, but at the cost of discarding the **unique, authentic knowledge** of labels 7 and8, creating a cognitive blind spot in the final model.

- ***Ideas:*** Separately mitigate the malicious impact within gradients while retaining the valuable knowledge embedded in model weights, thus maximizing model accuracy without compromising security.

Main Contributions

- ***Contributions:***
  - ***Mitigate malicious impact:*** Each benign client maintains the recording variable to accumulate aggregated gradients from each neighbor. It enables benign clients to precisely detect and mitigate aggregated malicious gradients to get purified gradients.
  - ***Retain beneficial components:*** The poisoned model weights are optimized by purified gradients. This optimization retains the beneficial knowledge previously contributed by malicious clients, and exploits canonical contributions from benign clients.
  - ***Theory and Experiment:*** Convergence proof and extensive experimental verification.

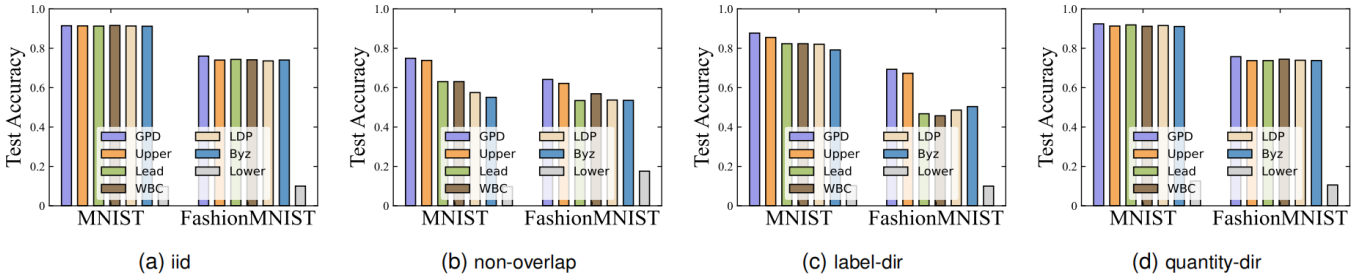

Fig. 9 The average accuracy among six poisoning attacks under the *MNIST* and *FashionMNIST* dataset.

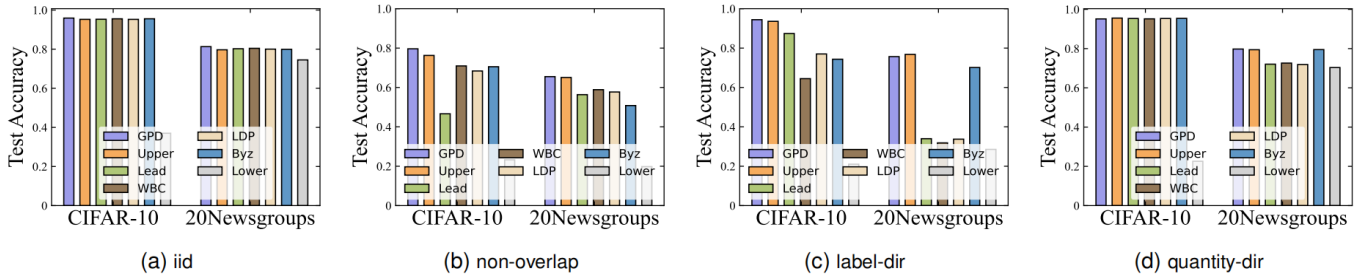

Fig. 10 The average accuracy among six poisoning attacks under the *CIFAR-10* and *20Newsgroups* dataset.

"Upper" denotes the upper bound on the model derived exclusively from benign clients, and others are baseline algorithms.