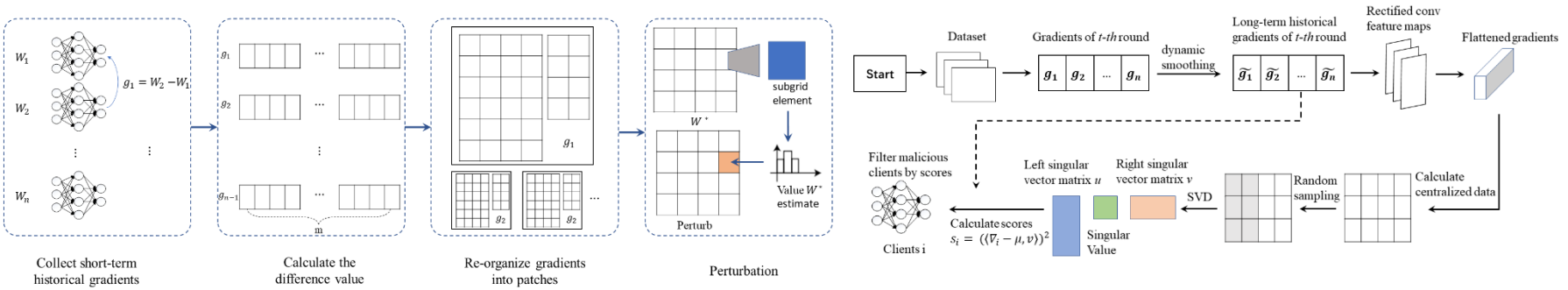
Enhancing Poisoning Attack Mitigation in Federated Learning through Perturbation-Defense Complementarity on Historical Gradients

Cong WANG, Zhilong MI, Ziqiao YIN, Binghui GUO

Frontiers of Computer Science, DOI: 10.1007/s11704-025-40924-1

Problems & Ideas

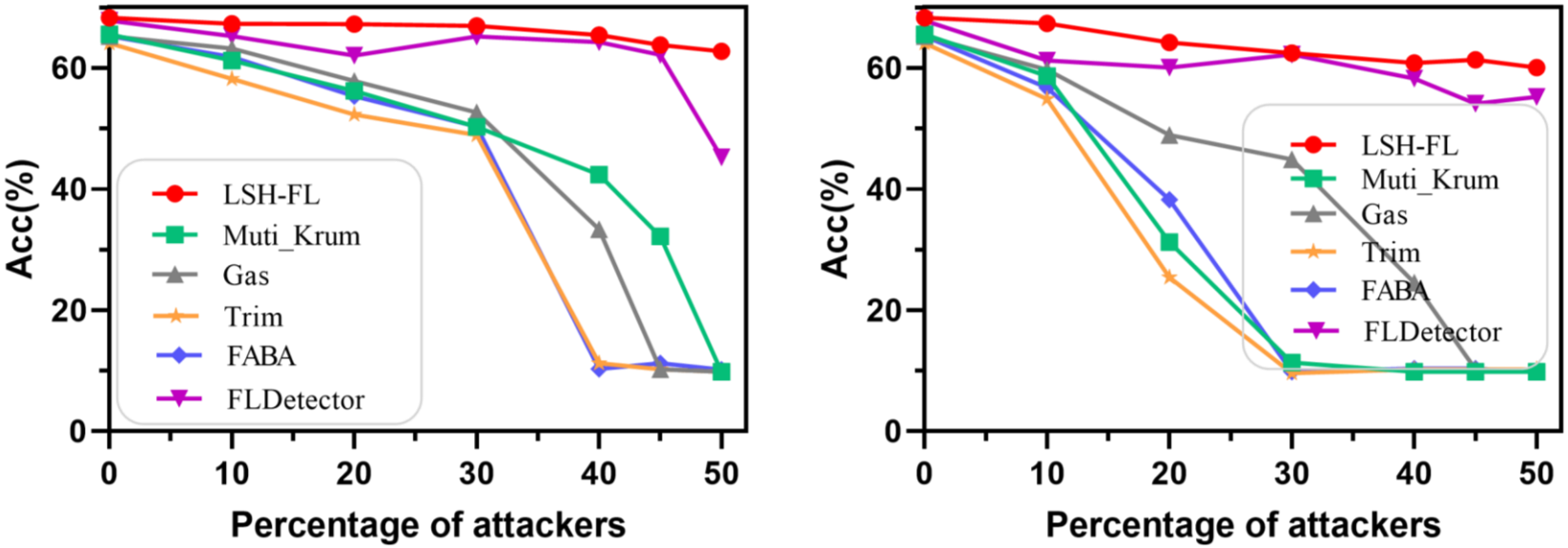
- Problems of conventional stereo matching approaches:
  - Traditional defense methods mainly focus on detecting anomalous updates but struggle to address long-term attack dynamics.
  - Existing methods fail to fully utilize historical gradient data and may compromise privacy.
- Ideas: A model that combines long-term and short-term historical gradients to effectively address poisoning attacks in Federated Learning.



The overviews of two main components. Left: The perturbation process on the client using P-SHG. Firstly, the client collects gradients to obtain SHGs, and then evaluates the differences between SHGs in blocks to establish a judgment matrix, which is used to determine whether to perturb; Right: The defense process on the server using D-LHG. Firstly, obtain the long-term historical gradients of each client stored on the server, then perform flattening and sampling to calculate the right singular matrix, and finally calculate the scores for each client.

Main Contributions

- Contributions:
  - A unique FL defense strategy framework is proposed that fully utilizes long and short historical information to extract features and imposes perturbation to defend against poisoning attacks;
  - A perturbation strategy utilizing short-term historical gradients on the client side has been developed, which perturbs the anomalous feature space, thereby mitigating the potential long-term impact of attacks;
  - A defense strategy utilizing long-term historical gradients to detect malicious clients, enhanced by the LH-DnC feature decomposition.



The global model accuracy of our LSH-FL and existing defenses on CIFAR-10 under poisoning attacks. Left: the global model accuracy under the LIE attack; Right: the global model accuracy under the Min-Max attack.