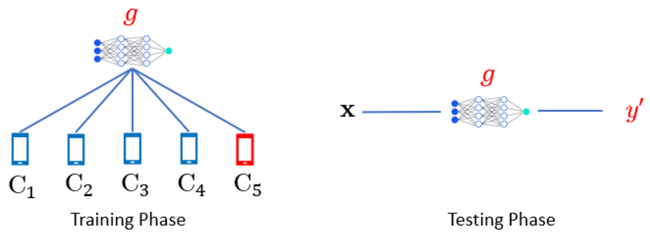
Provably Secure Federated Learning

This work [1] was accepted by AAAI-21

Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong

Duke University

Security Attacks to Federated Learning



Conventional single-global-model paradigm:

- **Vulnerable to security attacks** that

  - inject fake clients or compromise existing clients

  - tamper with the local training data/model updates

Existing Byzantine-robust Defenses

Main ideas:

- Mitigating statistical outliers among clients' model updates.
- Bounding the difference between the global model parameters learnt with/without malicious clients.

Limitations:

- No provable guarantee on the predicted labels.
- Are still vulnerable to advanced attacks [2].

References

[1] Cao, Xiaoyu, et al. (2021). "Provably Secure Federated Learning against Malicious Clients". In AAAI.

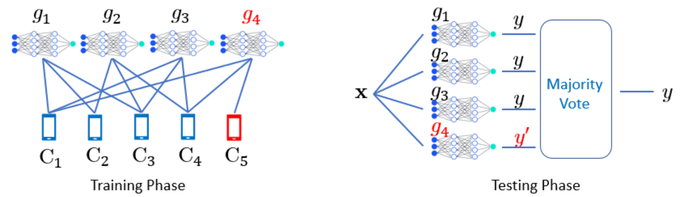
[2] Fang, Minghong, et al. (2020). "Local model poisoning attacks to Byzantine-robust federated learning". In USENIX Security.

[3] McMahan et al. (2017). "Communication-efficient learning of deep networks from decentralized data". In AISTATS.

[4] Davide Anguita et al. (2013). "A Public Domain Dataset for Human Activity Recognition Using Smartphones". In ESANN.

Our Ensemble Federated Learning

Ensemble global model



- Assume we have n clients $\mathbf{C} = \{C_1, C_2, \cdots, C_n\}$.
- Randomly sample a subset of k clients and train a global model.
- Repeat the process to obtain N **global models**, where $N \leq \binom{n}{k}$.
- Take the **majority vote** among the global models when predicting the label of a testing example.
- Formally, our ensemble global model h predicts example $\mathbf{x}$ as follows:

$$h(\mathbf{C}, \mathbf{x}) = \text{argmax}_i p_i$$

where p_i is the fraction of global models predicting label i and we name it label frequency.

Certified security level m^*

- Given a testing example $\mathbf{x}$, our ensemble global model provably predict the same label for it when the number of malicious clients is no greater than a threshold m^*.
- We have the following theorem:

Theorem 1. *Suppose we are given n clients $\mathbf{C}$, an arbitrary base federated learning algorithm $\mathcal{A}$, a subsample size k, and a testing example $\mathbf{x}$. y and z are the labels that have the largest and second largest label probabilities for $\mathbf{x}$ in our ensemble global model h. $\underline{p_y}$ is a lower bound of p_y and $\overline{p}_z$ is an upper bound of p_z. Formally, $\underline{p_y}$ and $\overline{p}_z$ satisfy the following conditions:*

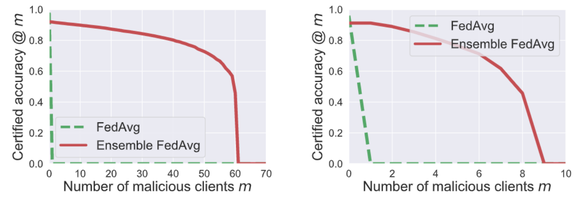
$$\max_{i \neq y} p_i = p_z \leq \overline{p}_z \leq \underline{p_y} \leq p_y.$$

Then, h provably predicts y for $\mathbf{x}$ when at most m^* clients in $\mathbf{C}$ become malicious, i.e., we have $h(\mathbf{C}', \mathbf{x}) = h(\mathbf{C}, \mathbf{x}) = y, \forall \mathbf{C}', M(\mathbf{C}') \leq m^*$, where m^* is the largest integer m $(0 \leq m \leq n - k)$ that satisfies $\frac{\lceil \underline{p_y} \cdot \binom{n}{k} \rceil}{\binom{n}{k}} - \frac{\lfloor \overline{p}_z \cdot \binom{n}{k} \rfloor}{\binom{n}{k}} > 2 - 2 \cdot \frac{\binom{n-m}{k}}{\binom{n}{k}}$.

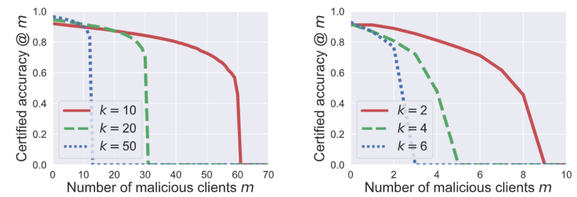
Evaluation

We use **certified accuracy** as our evaluation metric.

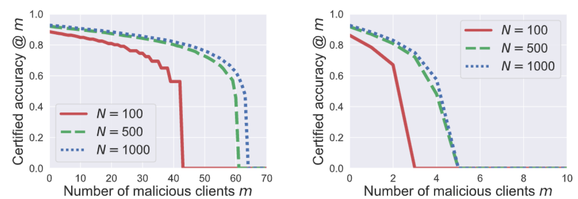
Certified accuracy at m malicious clients: the fraction of testing examples in the testing dataset whose labels are correctly predicted and whose certified security levels are at least m.



FedAvg [3] vs. ensemble FedAvg on MNIST (left) and HAR [4] (right).



Impact of k on our ensemble FedAvg on MNIST (left) and HAR (right).



Impact of N on our ensemble FedAvg on MNIST (left) and HAR (right).

Conclusions

- We propose ensemble federated learning and derive its provable security guarantees against malicious clients.
- Our results show that our ensemble federated learning can effectively defend against malicious clients with provable security guarantees.