CRFL: Certifiably Robust Federated Learning against Backdoor Attacks

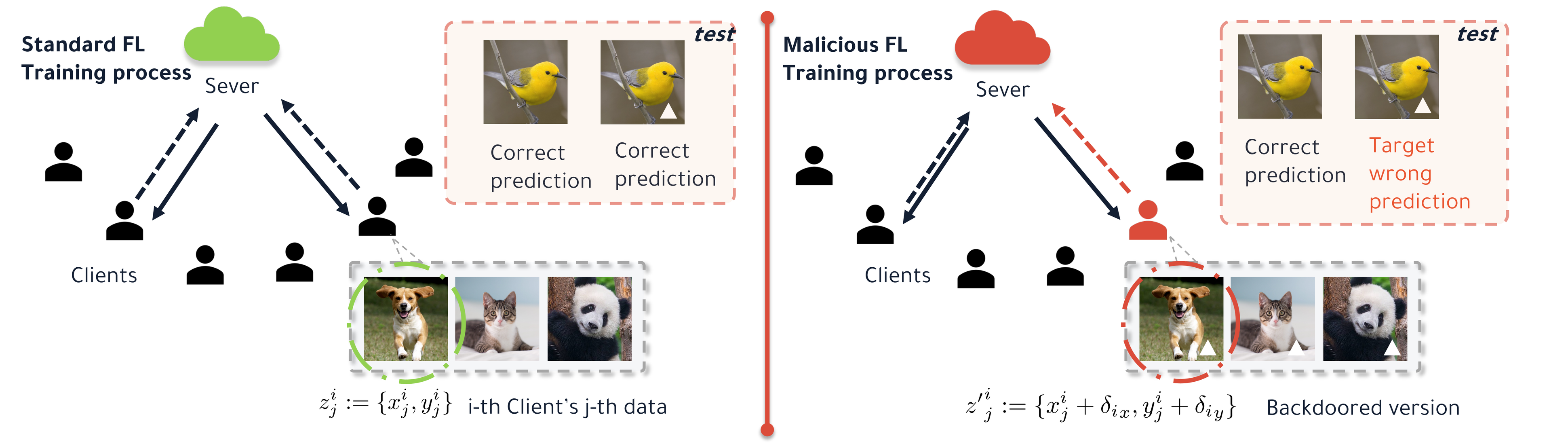
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
IBM Research

Chulin Xie (UIUC) Minghao Chen (Zhejiang University) Pin-Yu Chen (IBM Research) Bo Li (UIUC)

Overview

Backdoor Attack against Federated Learning (FL):

- Malicious clients inject a backdoor pattern into local models
- After Federated Learning, global model will mis-classify any test input with such pattern as the target label.



Robust Federated Learning:

☺ Defenses do exist: Robust aggregations and empirically robust FL training protocols.
☹ They lack robustness certification and are adaptively attacked again.

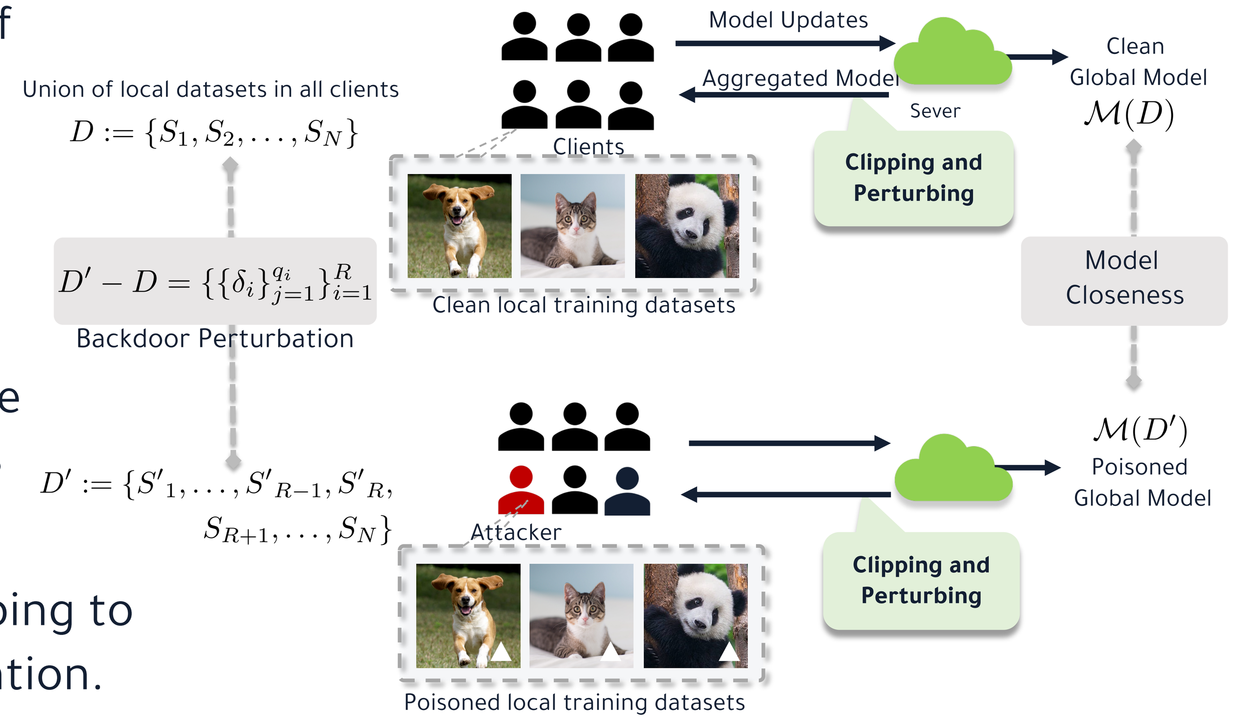
Certifiably Robust Federated Learning (CRFL):

✓ **The first general framework**: train certifiably robust FL models against backdoors.
✓ **Theoretical analysis**: a sample-wise robustness certification on backdoors under certain constraints.
✓ **Empirical study**: show robustness certification under FL parameters.

CRFL Training: Clipping and Perturbing

- **Method**: server clips the norm of global model parameters, and adds a Gaussian noise.
- **Key idea**: when $D' - D$ is under certain threshold, we verify that poisoned FL model $\mathcal{M}(D')$ is close to clean model $\mathcal{M}(D)$, and thus is robust to backdoors.

❖ Use clipping and noise perturbing to control the global model deviation.



CRFL Testing: Parameter Smoothing

Base classifier $h : (\mathcal{W}, \mathcal{X}) \to \mathcal{Y}$ $\mathcal{Y} = \{1, \ldots, C\}$

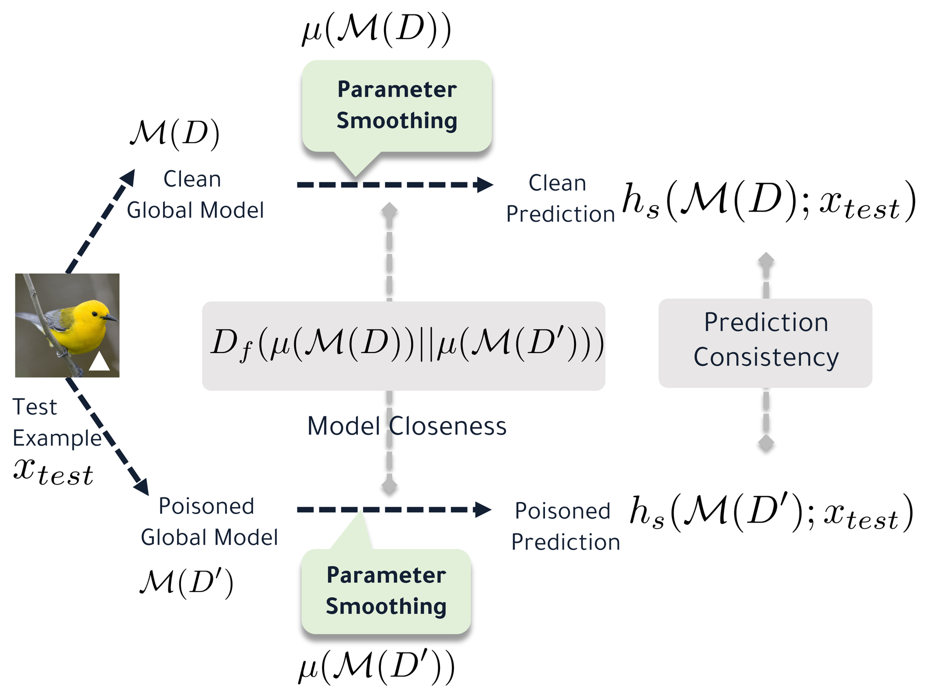
Smoothed classifier h_s

1. Given the model parameter w of h, when queried at a test sample x_{test}
2. Get votes for each class c: take a majority vote over the predictions of the base classifier on random model parameters drawn from a probability distribution $\mu(w) = \mathcal{N}(w, \sigma_T{}^2\mathbf{I})$

$$H_s^c(w; x_{test}) = \mathbb{P}_{W \sim \mu(w)}[h(W; x_{test}) = c]$$

3. Return the majority vote winner: the mostly probable label among all classes

$$h_s(w; x_{test}) = \arg\max_{c \in \mathcal{Y}} H_s^c(w; x_{test})$$



- **Method:** server makes the prediction based on parameter-smoothed models.
- **Key idea:** for two close distribution $\mu(\mathcal{M}(D'))$ and $\mu(\mathcal{M}(D))$, we verify that returned label from the smoothed classifier is consistent.

❖ Use f-divergence as a statistical distance for model closeness.

Certification Goal

Goal: develop a robustness certificate by studying under what condition for backdoor perturbation that the prediction for a test sample is consistent between the smoothed FL models trained from D and D' separately.

$D' - D = \{\{\delta_i\}_{j=1}^{q_i}\}_{i=1}^R$	$\Leftarrow$	$D_f(\mu(\mathcal{M}(D)) \| \mu(\mathcal{M}(D')))$	$\Leftarrow$	$h_s(\mathcal{M}(D); x_{test}) = h_s(\mathcal{M}(D'); x_{test})$
Backdoor Perturbation		Model Closeness		Prediction Consistency

Theoretical analysis:

1. Quantify the model closeness between the FL trained models via f-divergence and Markov Kernel.

$$D_{KL}(\mu(\mathcal{M}(D)) || \mu(\mathcal{M}(D'))) \leq \frac{2R\sum_{i=1}^R \left(p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}} L_{\mathcal{Z}} \|\delta_i\|\right)^2}{\sigma_{t_{adv}}^2} \prod_{t=t_{adv}+1}^{T} \left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)$$

Distributed SGD anaysis with convex, smoothness assumption and Lipschitz gradient assumption

KL-divergence in the attacked round

Contraction coefficient in later rounds

Data processing inequality and contraction coefficient of Markov Kernel

2. Connect the model closeness to the prediction consistency by parameter smoothing.

If $D_{KL}(\mu(w), \mu(w')) \leq \epsilon$ $\epsilon = -\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right)$ then $h_s(w'; x_{test}) = h_s(w; x_{test}) = c_A$

Robustness Conditions

General Robustness Condition:

$$R\sum_{i=1}^R \left(p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}} \|\delta_i\|\right)^2 \leq \frac{-\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right)\sigma_{t_{adv}}^2}{2L_{\mathcal{Z}}^2 \prod_{t=t_{adv}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)}$$

Certification is in three levels: feature, sample, and client.

Robustness Condition in Feature Level:

- When the backdoor magnitude is the same for every attacker:

$\|\delta\| <$ RAD

$$\text{RAD} = \sqrt{\frac{-\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right)\sigma_{t_{adv}}^2}{2RL_{\mathcal{Z}}^2 \sum_{i=1}^R (p_i \gamma_i \tau_i \eta_i \frac{q_{B_i}}{n_{B_i}})^2 \prod_{t=t_{adv}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1\right)}}$$

Certified radius

Experiments

Setup:

- Multi-class logistic regression on three datasets: Lending Club Loan Data (LOAN), MNIST, and EMNIST.

Evaluation Metric:

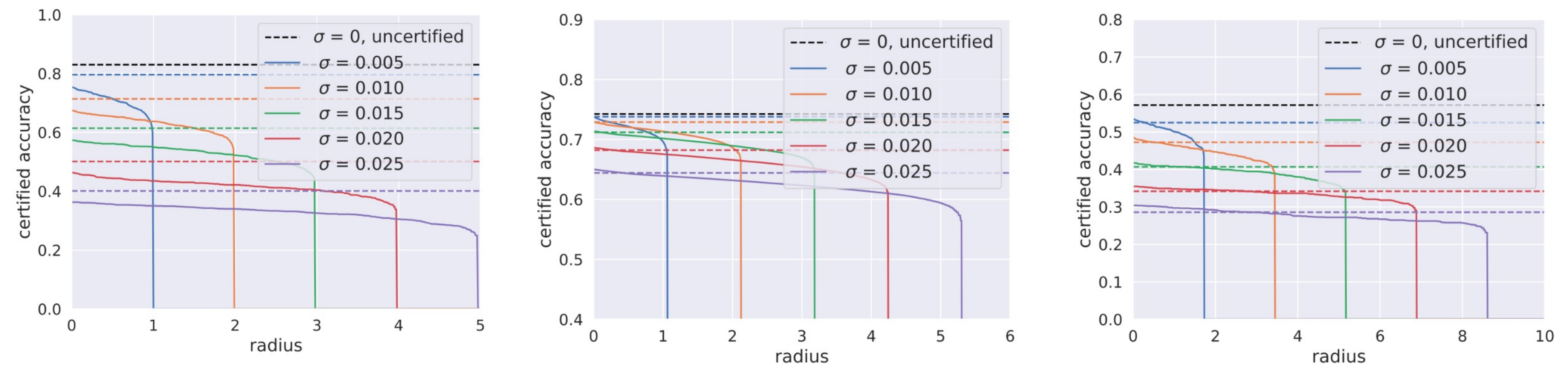
- Certified accuracy at r: the fraction of the test set for which the possibly backdoored classifier makes correct and consistent predictions with the clean model.

$$\frac{1}{m}\sum_{i=1}^m \mathbb{1}\{c_i = y_i \text{ and } \text{RAD}_i \geq r\}$$

Given a test set of size m, for i-th test sample, the ground truth label is y_i, and the output prediction is c_i with the certified radius RAD_i.

Experiment Results:

- Effect of different smoothing levels during training:



MNIST LOAN EMNIST

❖ When noise level σ is high, large radius can be certified but at a low accuracy, so the parameter noise controls the trade-off between certified robustness and accuracy.

More details and results are in our paper:

- Effects of smoothing level, attacker ability, robust aggregation, client number, training rounds, etc. on certified robustness.