CRFL: Certifiably Robust Federated Learning against Backdoor Attacks

ICML 2021

Chulin Xie, Minghao Chen, Pin-Yu Chen, Bo Li

chulinx2@illinois.edu, pin-yu.chen@ibm.com, lbo@illinois.edu

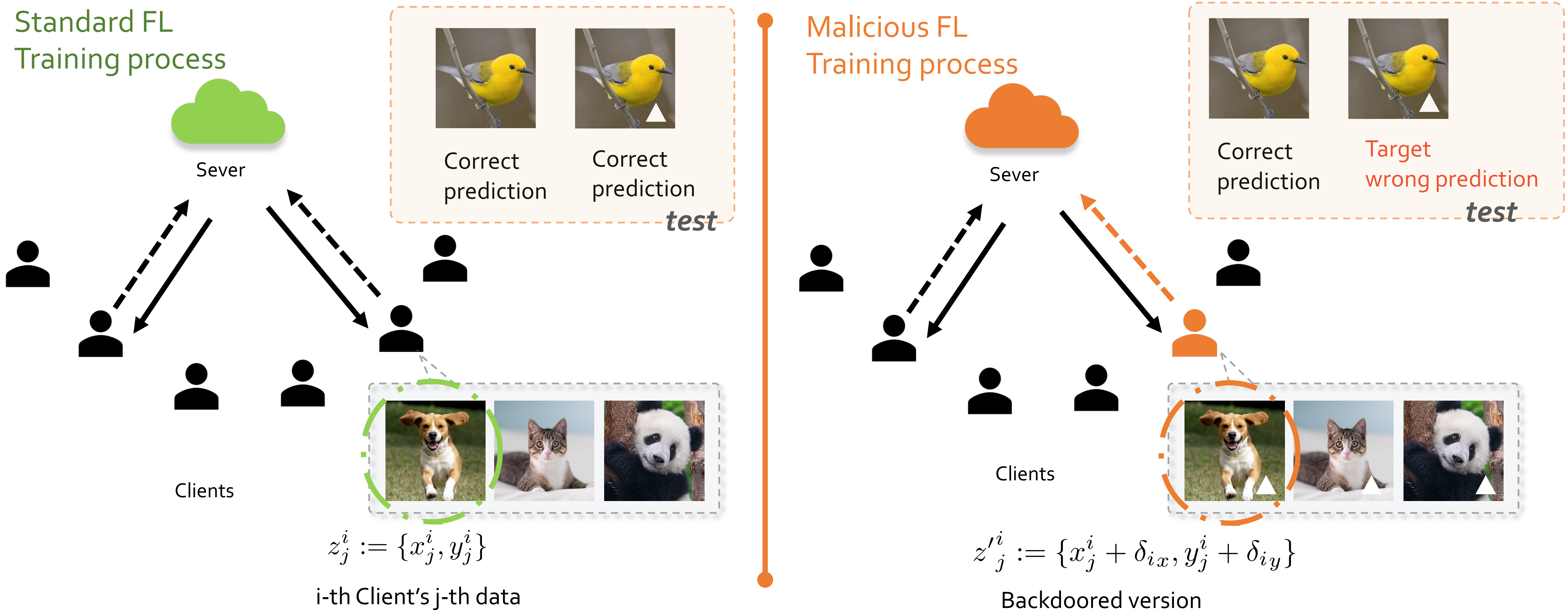
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

IBM Research

ZHEJIANG UNIVERSITY 1897

Motivation

Backdoor Attack against Federated Learning (FL)



Motivation

Robust Federated Learning

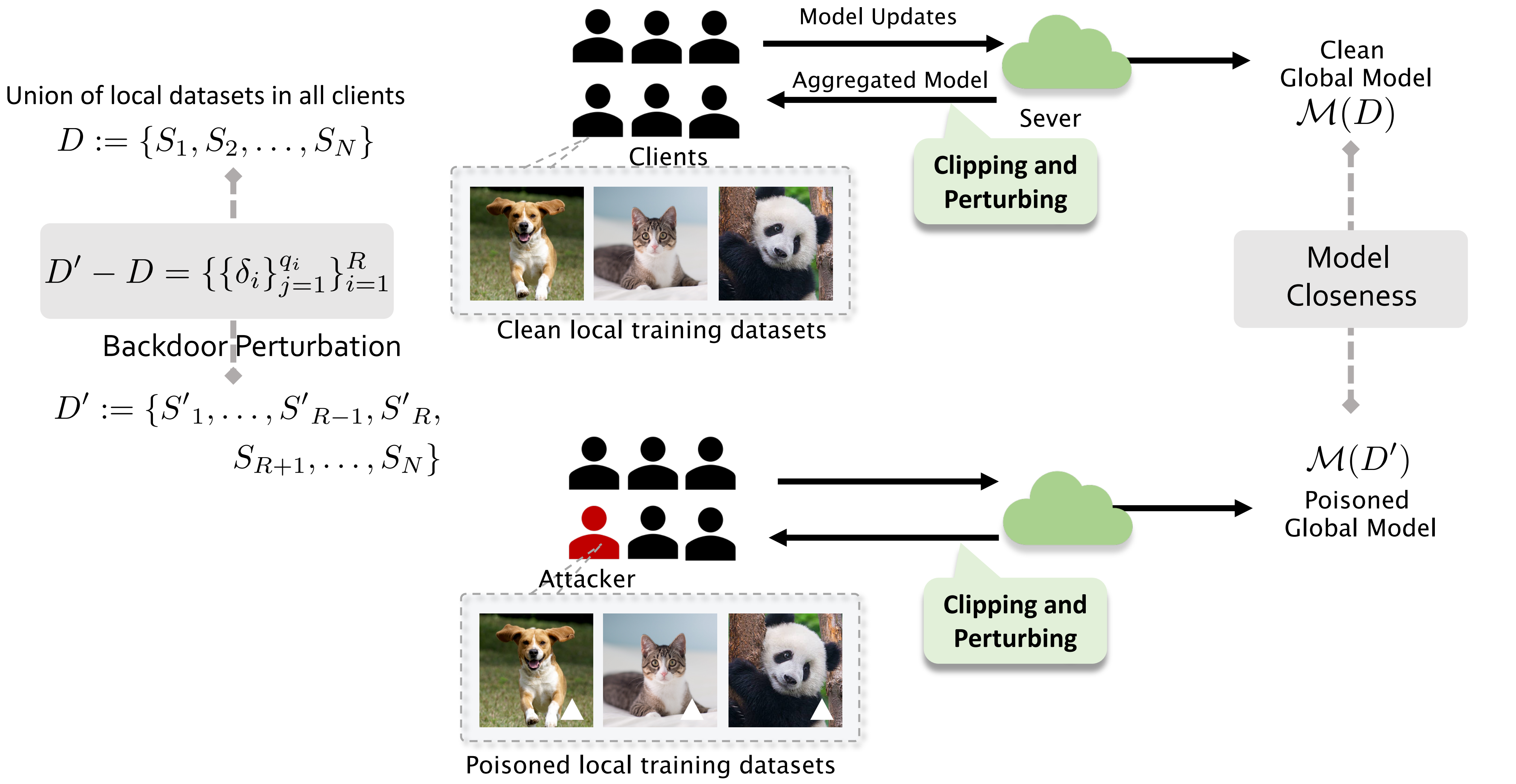
Defenses do exist: robust aggregation methods and empirically robust federated training protocols

…they lack robustness certification and are adaptively attacked again

We provide:

- **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 different FL parameters.

CRFL Training: Clipping and Perturbing



CRFL Testing: Parameter Smoothing

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

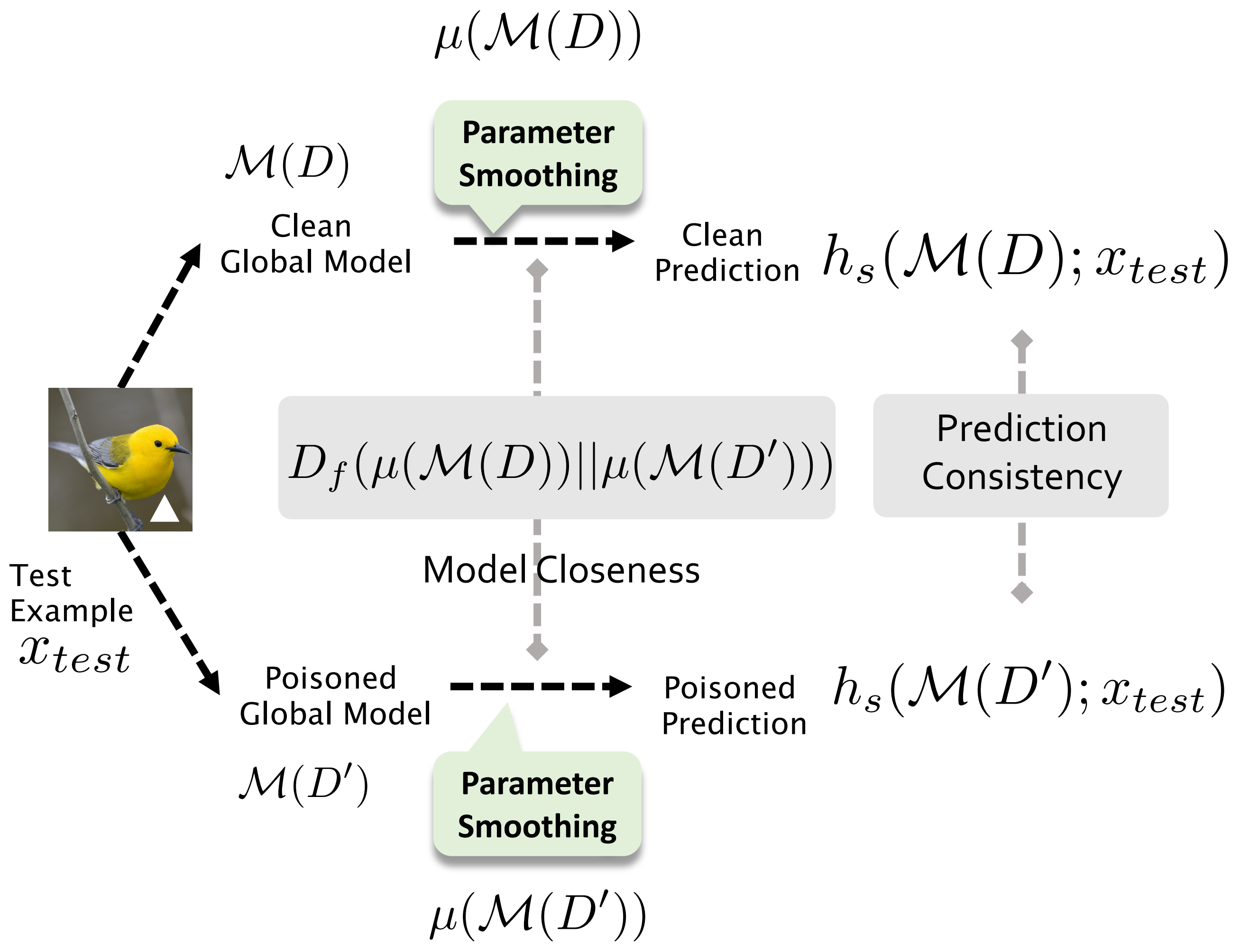
Smoothed classifer h_s

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

Votes for class c $\mu(w) = \mathcal{N}(w, {\sigma_T}^2 \mathbf{I})$

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

The majority vote winner



Certification

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

Backdoor Perturbation — Model Closeness — Prediction Consistency

General Robustness Condition

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

When the size of the backdoor magnitude is the same for every attackers

Robustness Condition in Feature Level

$||\delta|| <$ RAD

$$\mathrm{RAD} = \sqrt{\frac{-\log\left(1-(\sqrt{p_A}-\sqrt{p_B})^2\right)\sigma^2_{\mathrm{t_{adv}}}}{2RL^2_{\mathcal{Z}}\sum_{i=1}^{R}(p_i\gamma_i\tau_i\eta_i\frac{q_{B\,i}}{n_{B\,i}})^2\prod_{t=\mathrm{t_{adv}}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right)}}$$

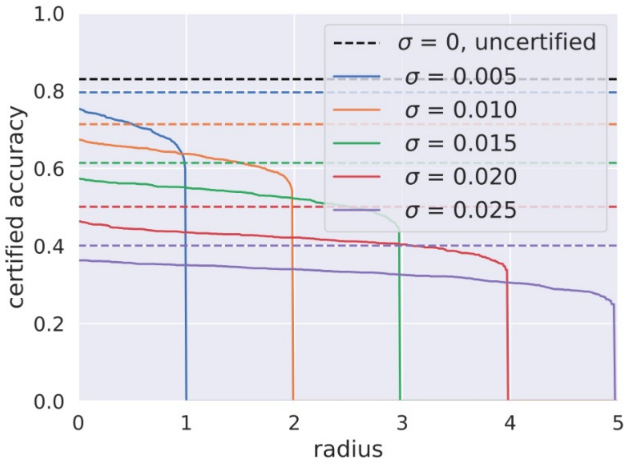
Certified radius

Our certification is in three levels: feature, sample, and client.

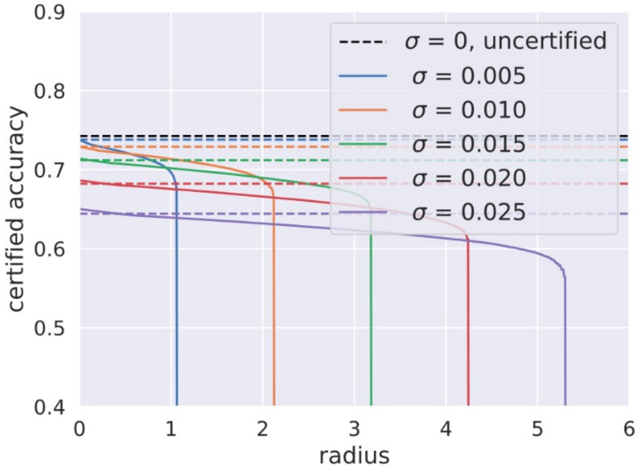
Experiments

Thank you.

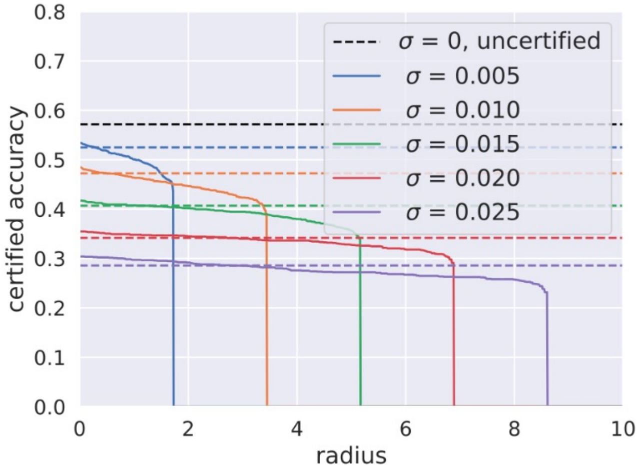
Effect of different smoothing levels during training



MNIST



LOAN



EMNIST

More details and results are in our paper:

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