FedDefender: Client-Side Attack-Tolerant Federated Learning

Sungwon Park*, Sungwon Han*, Fangzhao Wu, Sundong Kim, Bin Zhu, Xing Xie, Meeyoung Cha

Introduction

❖ Federated learning has become a popular model training method to guarantee the minimum level of data privacy

❖ Despite its advantages, federated learning is vulnerable to attacks due to its decentralized nature [1]

❖ Most existing defense methods suggest robust aggregation strategies

$$\theta^{t+1} = \theta^t + \frac{\sum_{k=1}^{N} \boxed{\mathbb{1}_{\{k \in S_b\}}} \cdot \Delta\theta_k^t}{|S_b|}$$

Research Motivation

❖ It is difficult to distinguish benign users with non-IID local data distribution from adversaries

❖ If robust aggregation fails to detect, the performance of model can be degraded. While the client-side defense has been relatively under-investigated

➢ We propose the Attack Tolerant Local Gradient Update as an add-on module to guarantee additional resistance to model poisoning attack

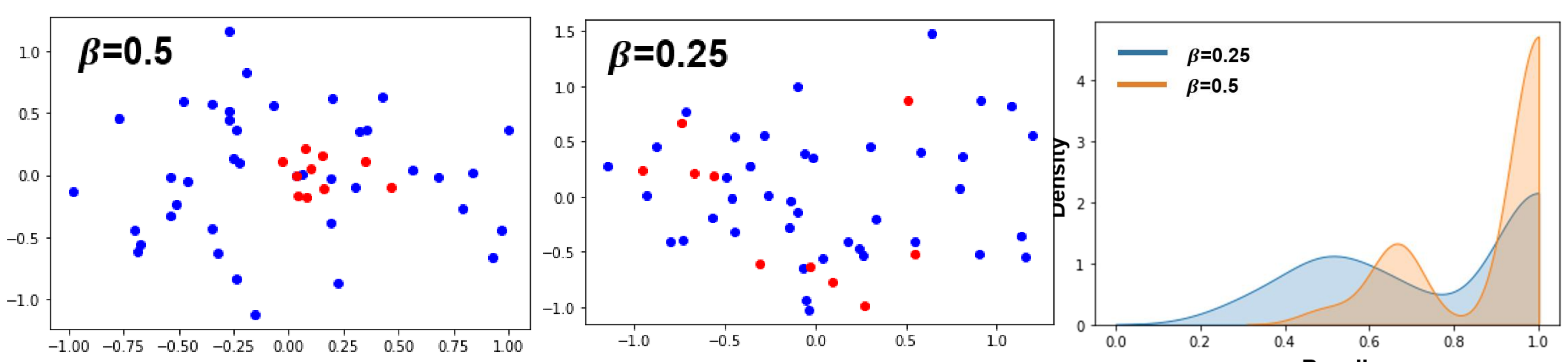


Fig 1. Detection recall plot of Multi-Krum [2] with different levels of non-IID

Method

Model Overview

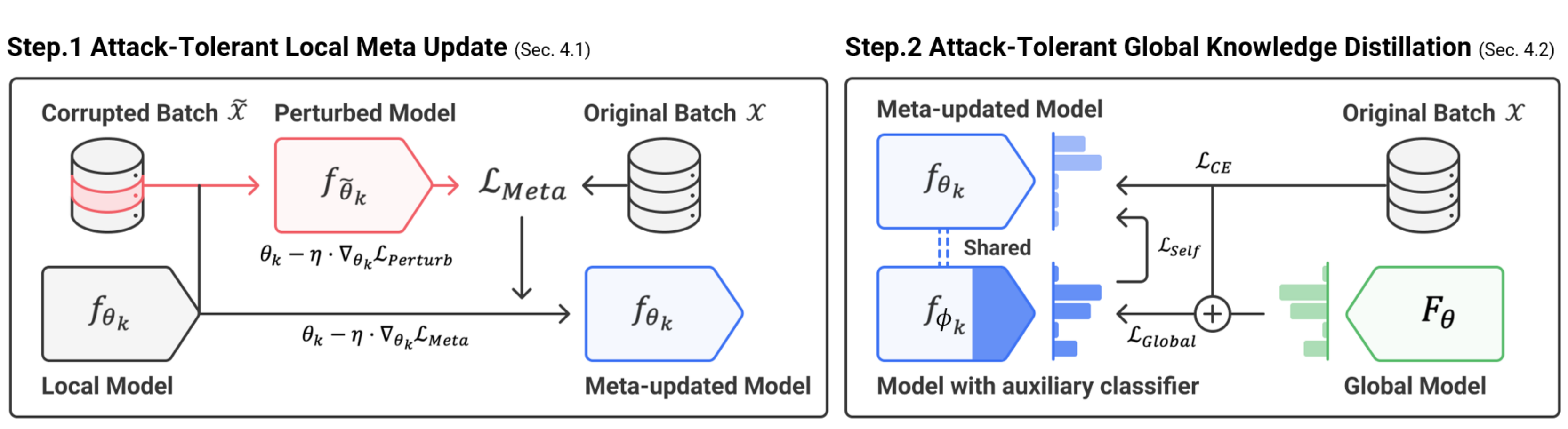


Fig 2. Overall architecture of the proposed model

"Vaccinate" local models to thwart model poisoning attack

Step 1. Attack-tolerant Local Meta Update
- Learn noise-tolerant parameters in a way that "vaccinates" the local model using meta-update

Step 2. Attack-tolerant Global Knowledge Distillation
- Align the local model's knowledge to the global data distribution while reducing the adverse effects of the possibly-corrupted global model

Step 1. Attack-tolerant Local Meta Update

❖ Give vaccine to the local client using meta learning

Local model poisoning with synthetic noise

1. Generate perturbing batch $\hat{\mathcal{X}}$ by replacing the label y with synthetic label

$$\tilde{\mathcal{X}} = \{(\mathbf{x}, \tilde{\mathbf{y}}) | (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \text{ and } \tilde{\mathbf{y}} = \text{Sample}_{\mathbf{y}}(\mathcal{N}_k(\mathbf{x}, \theta_k))\},$$

2. Local model poisoning with synthetic noises

$$\mathcal{L}_{Perturb} = \frac{1}{|\tilde{\mathcal{X}}|} \sum_{\mathbf{x}, \tilde{\mathbf{y}} \in \tilde{\mathcal{X}}} H(\tilde{\mathbf{y}}, f_{\theta_k}(\mathbf{x}))$$

$$\tilde{\theta}_k \leftarrow \theta_k - \eta \nabla_{\theta_k} \mathcal{L}_{Perturb}$$

Local model poisoning with synthetic noise

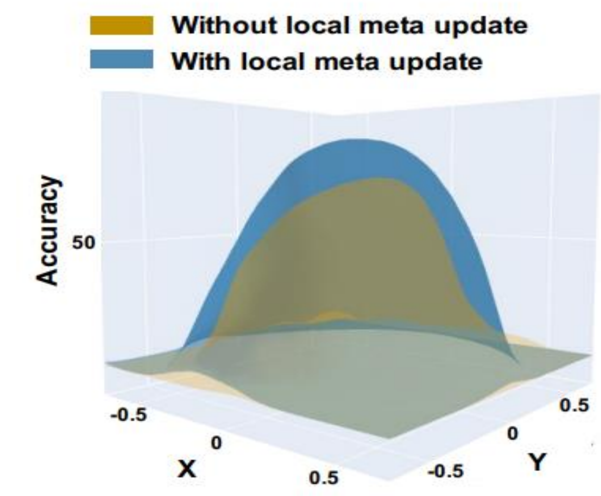
3. Update local model according to the gradient of Meta loss

$$\mathcal{L}_{Meta} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} H(\mathbf{y}, f_{\tilde{\theta}_k}(\mathbf{x}))$$

$$\theta_k \leftarrow \theta_k - \eta \nabla_{\theta_k} \mathcal{L}_{Meta}$$

We add random direction perturbations to the model parameter.

➢ **Find a solution with flat minima in the loss curve within the parameter space**



(a) Visualization of accuracy surface

Step 2. Attack-tolerant Global Knowledge Distillation

❖ **The credibility of the global model can be compromised**
- Transferring knowledge to an intermediate shallow section of the local model through an auxiliary classifier

$$\hat{\mathbf{y}} = (1 - \alpha) \cdot \mathbf{y} + \alpha \cdot F_\theta(\mathbf{x}, \tau). \quad \mathcal{L}_{Global} = \frac{1}{|\hat{\mathcal{X}}|} \sum_{\mathbf{x}, \hat{\mathbf{y}} \in \hat{\mathcal{X}}} H(\hat{\mathbf{y}}, f_{\phi_k}(\mathbf{x}))$$

- To improve the deeper layers of the local model, we use self knowledge distillation between auxiliary classifier and original classifier.

$$\mathcal{L}_{Self} = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} KL(f_{\theta_k}(\mathbf{x}, \tau) || f_{\phi_k}(\mathbf{x}, \tau)), \quad \mathcal{L}_{KD} = \mathcal{L}_{Global} + \mathcal{L}_{Self}$$

- This global knowledge distillation loss is optimized in conjunction with cross-entropy loss

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{KD} \qquad \theta_k \leftarrow \theta_k - \eta \nabla_{\theta_k} \mathcal{L}_{total}$$

Experiment

❖ FedDefender enhances additional resilience against poisoning attacks in federated learning

Method	CIFAR-10		CIFAR-100		TinyImageNet		FEMNIST	
	Last	Best	Last	Best	Last	Best	Last	Best
No Defense	68.80	71.96	42.97	43.90	30.37	38.98	18.88	23.81
+ FedDefender	**78.17**	**79.96**	**51.76**	**51.92**	**35.59**	**39.68**	**22.11**	**24.48**
Multi-Krum	73.09	75.03	47.75	47.83	37.26	38.54	20.55	23.30
+ FedDefender	**81.87**	**82.77**	**53.15**	**53.35**	**38.98**	**39.48**	**22.43**	**24.36**
ResidualBase	73.61	75.10	44.80	45.13	35.05	38.60	19.44	23.86
+ FedDefender	**79.28**	**80.83**	**50.62**	**50.98**	**36.22**	**39.24**	**22.41**	**24.27**

Tab 1. Performance improvement with FedDefender on classification accuracy

❖ FedDefender outperforms alternative baselines, including ablation studies and other possible global regularizations

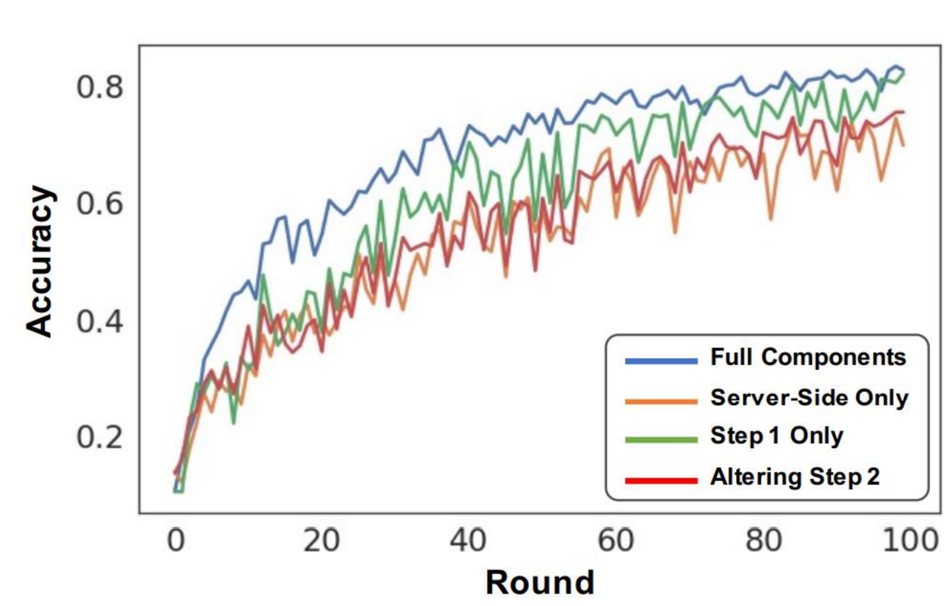


Fig 3. Ablation Study

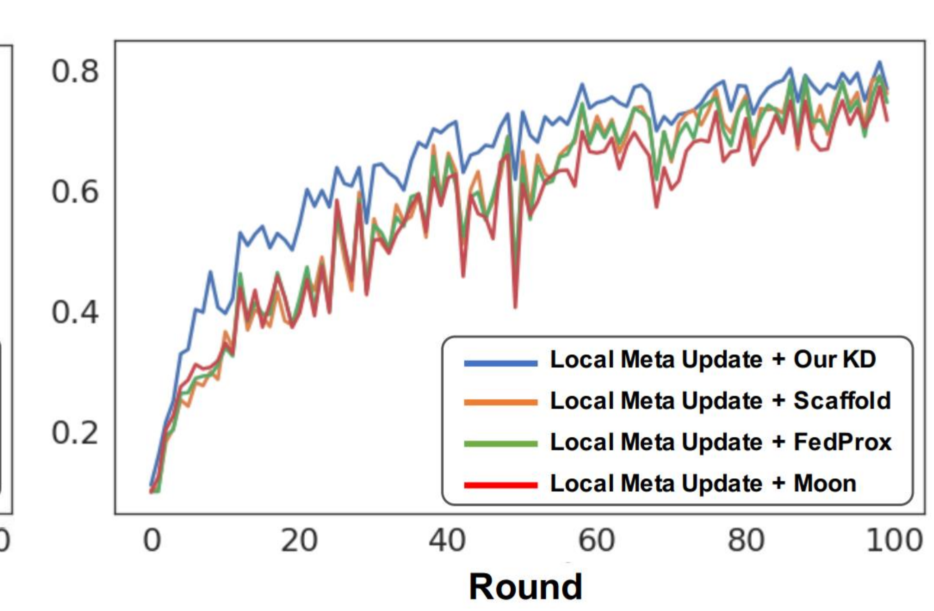


Fig 4. Comp. with other global regularization

Conclusion

FedDefender has achieved a meaningful robustness improvement against various model poisoning attacks when used in conjunction with existing server-side defense strategies.

References

[1] Fang et al. " Local model poisoning attacks to {Byzantine-Robust} federated learning.", USENIX Security 2020.

[2] Blanchard et al. "Machine learning with adversaries: Byzantine tolerant gradient descent." , Neurips 2017.