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

\[ \beta^{i+1} = \beta^i + \alpha x \hat{\theta}_k - \eta \nabla \theta_k \mathcal{L}_{Meta} \]

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

Step 1. Attack-tolerant Local Meta Update
1. Generate perturbing batch \( X \) by replacing the label \( y \) with synthetic label
2. Local model poisoning with synthetic noises

\[ \mathcal{L}_{Perturb} = \frac{1}{|X|} \sum_{x \in X} H(\hat{y}, f_{\theta_k}(x)) \]

Local model poisoning with synthetic noise
3. Update local model according to the gradient of Meta loss

\[ \theta_k' = \theta_k - \eta \nabla \theta_k \mathcal{L}_{Meta} \]

We add random direction perturbations to the model parameter.
❖ Find a solution with flat minima in the loss curve within the parameter space

Step 2. Attack-tolerant Global Knowledge Distillation
❖ The credibility of the global model can be compromised
❖ We improve the deeper layers of the local model, we use self knowledge distillation between auxiliary classifier and original classifier.

\[ \mathcal{L}_{Self} = \frac{1}{|X|} \sum_{x \in X} KL(f_{\theta_k}(x), f_{\theta_k}(x)) \]

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

\[ \mathcal{L}_{Total} = \mathcal{L}_{CE} + \mathcal{L}_{KD} \]

FedDefender enhances additional resilience against poisoning attacks in federated learning
FedDefender outperforms alternative baselines, including ablation studies and other possible global regularizations

References