Backdoor Attacks in FL

Adv modifies the training dataset to introduce backdoors into the global model

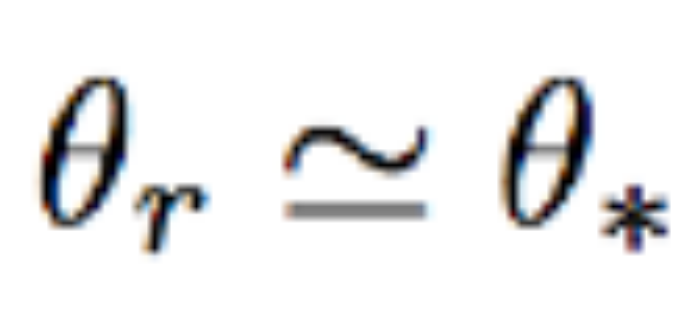
47

In round r

Aggregate Model

Results in

$$\theta_r = \theta_* + \eta \frac{\sum_{i=1}^{K-1} n_i \theta_i}{\sum_{i=1}^{K} n_i}$$



$$\theta_r^{attacker} = \frac{\sum_{i=1}^{K} n_i}{\eta n_{attacker}} \cdot (\theta^* - \theta_r)$$

Backdoor Attacks in FL

Adv modifies the training dataset to introduce backdoors into the global model

In round r

$$\theta_r^{attacker} = \frac{\sum_{i=1}^{K} n_i}{\eta n_{attacker}} \cdot (\theta^* - \theta_r)$$

Aggregate Model

$$\theta_r = \theta_* + \eta \frac{\sum_{i=1}^{K-1} n_i \theta_i}{\sum_{i=1}^{K} n_i}$$

Results in

$$\theta_r \simeq \theta_*$$

Backdoor Attacks/Defenses