Learning to Attack Distributionally Robust Federated Learning

Wen Shen, Henger Li, Zizhan Zheng

Department of Computer Science, Tulane University

Introduction

- We propose a non-myopic attack framework for federated learning.
- Our attack framework:
  - consists of two stages: *distribution learning* (i.e., learning the distribution of the aggregated data using local data and the model parameters received from the server) and *policy learning* (i.e., learning a non-myopic attack policy using deep reinforcement learning);
  - can e ectively attack distributionally robust federated learning models even without knowing the server's aggregation rule;
  - outperforms baselines on both synthetic and real-world datasets.

Motivation

- Federated learning systems are vulnerable to threats [1].

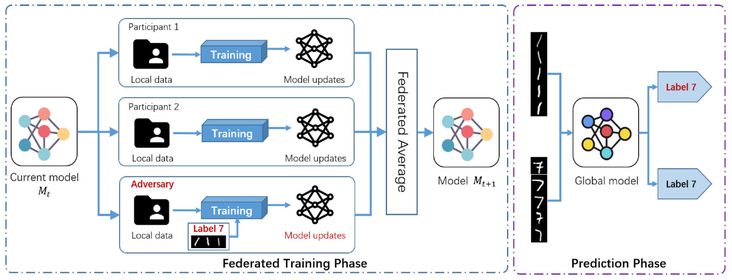

Fig.1: A typical poisoning attack in Federated Learning.

- Previous attacks on federated learning
  - perform poorly with unknown aggregation rule
  - require local information about normal workers
  - are usually myopic
  - untested under distributionally robust federated learning (DRFL) systems (an extension of the distributionally robust optimization with adversarial training framework [2])
- Goal: Learn a non-myopic policy that
  - performs well even with unknown aggregation rule
  - does not require local information about normal workers
  - is non-myopic
  - can e ectively attack DRFL systems

Background

Notation.

- Setting: one server and m worker machines with each machine holds n data samples. The mn data samples are drawn from the same *i:i:d:* distribution P_0
- $\hat{P} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} z_i(j)$: the empirical distribution generated by the data samples, where $z_i(j)$ denotes the Dirac point mass at the i-th training sample on worker j.
- $P = fP : W_c(P; P_0)$ g: a class of distributions, where $W_c(\ ;\ )$ is the Wasserstein metric.
- : a xed dual variable; $c(\ ;\ )$: the transportation cost that de nes the Wasserstein metric; $l(\ ;\ )$: loss function.
- $P(\ ) := \arg\max_P f\mathrm{E}_{Z \sim P}[l(\ ; Z)] \quad W_c(P; \hat{P})g$,
  $b(\ ) := W_c(P(\ ); \hat{P}) = \mathrm{E}_{Z \sim \hat{P}}[c(T(\ ; Z); Z)]$.

Distributionally Robust Federated Learning.

Algorithm 1: Distributionally Robust Federated Learning (DRFL)

Initialization: θ^0, m workers each with n data samples , step size η;
Output: θ^T
for $t = 1$ to T **do**
  Each Worker j**:**
  Sample a minibatch $B^t(j)$
  for $z_0 \in B^t(j)$ **do**
    $\hat{z}(z_0) = \arg\max_{z \in Z} l(\theta^{t-1}; z) - \gamma c(z, z_0)$
  end for
  $g_j^t = \frac{1}{|B^t(j)|} \sum_{z_0 \in B^t(j)} \nabla_\theta l(\theta^{t-1}; \hat{z}(z_0))$
  Send g_j^t to the server
  Server:
  $g^t = Aggr(g_1^t, g_2^t \ldots, g_m^t)$
  $\theta^t = proj_\Theta(\theta^{t-1} - \eta g^t)$
  Broadcast θ^t to the workers
end for

Methods and Results

- Threat Model:
  Attacker's knowledge:
  - local training algorithm
  - local data distribution $\hat{P}_a$
  - global model parameters (i.e., $f\ ^t g$, , , and $l(\ ;\ )$)

  Attacker's objective: to maximize the worst-case surrogate loss by sending crafted gradients to the server. We consider the attacker aims to minimize the robust level b for convenience.
- The Reinforcement Learning Problem:
  MDP $(S; A; q; r;\ )$:
  - $S = fs_t g$: a continuous set of states. Here, $s_t := {}^t 2$.
  - $A = fa_t g$: the action space of the attacker where $a_t := g_j^t 2 \mathrm{R}^d$ is the gradient that attacker i sends to the server in Algorithm 1.
  - $q(s; a; s^0)$: transition function that represents the probability of reaching a state $s^0 2 S$ from the state $s 2 S$ when the attacker chooses action $a 2 A$.
  - $r : S \quad A \quad S \ ! \ R$ is the reward function, where $r_t := b(\ ^t 1) \quad b(\ ^t)$.
  - : the discount factor for future rewards.

  Attacker's objective: to nd a policy $: S \ ! \ A$ that maps the current model $^t 1$ to the next attack action a_t to minimizes $b(\ ^T)$.
- Two-Stage Attack for Federated Learning:

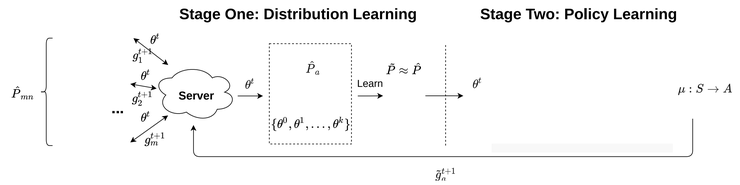

Fig.2: An overview of the two-stage attack for federated learning.

- Stage One - Distribution Learning: use distribution learning (e.g., the DLG [3] method) to learn an approximation $\hat{P}$ of the aggregated data distribution $\hat{P}$ from epoch 1 to epoch k given the attacker's local data distribution $\hat{P}_a$ and the model parameters $f\ ^0;\ ^1; :::;\ ^k g$.
- Stage Two - Policy Learning: use the learned model $\hat{P}$ to learn an attack policy using deep reinforcement learning (e.g., DDPG [4]). Starting from epoch $k + 1$ to T, use the same learned policy to generate actions (i.e., gradients) given a state as the input.

- Results:

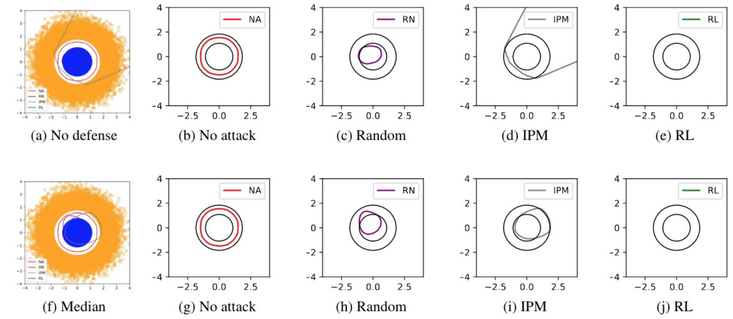

Fig.3: Synthetic data - no defense (upper) and coordinate-wise median (lower). Training data are shown in blue and orange. Classi cation boundaries are shown in red, purple, gray, and green for no attack (NA), random attack (RN), inner product manipulation (IPM), and reinforcement learning (RL) attack. The proposed two-stage attack using reinforcement learning performs the best (as evidenced by the missing classi cation boundary in 3e, 3j) by learning a non-myopic policy that minimizes the level of robustness b.

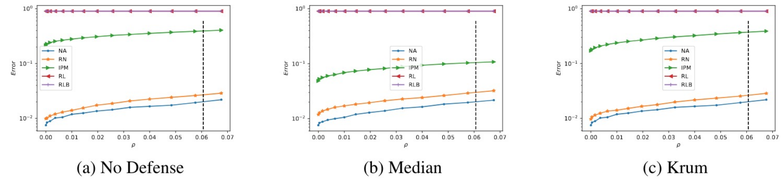

Fig.4: MNIST - Misclassi cation errors (including FP and FN) for di erent perturbation levels with 5 DRFL models: NA, RN, IPM, RL, and reinforcement learning-blackbox (RLB) when the server uses no defense, coordinate-wise median, and Krum, respectively. The vertical line indicates the robustness level b of the trained model with no attacks. RL and RLB outperform both the random attack and the inner product manipulation attack substantially for all the settings.

Conclusion and Future Work

- We propose a two-stage attack framework that
  - combines distribution learning and deep reinforcement learning to learn a non-myopic attack policy
  - can e ectively compromise the federated learning systems even with certi ed distributional robustness
- Our work opens up new exciting revenues for further study:
  - Attack methods for federated learning systems with non-i.i.d. data.
  - Federated learning models with subsampling process.
  - Online methods to allow the attacker to train the attack policy immediately after learning an approximate distribution.

Key References

[1] Lyu, L., et al. Threats to federated learning: A survey. In arXiv:2003.02133 (2020).
[2] Sinha, A., et al. Certifying some distributional robustness with principled adversarial training. In ICLR 2018.
[3] Zhu, L., et al. Deep leakage from gradients. In NeurIPS 2019.
[4] Lillicrap, T. P., et al. Continuous control with deep reinforcement learning. In ICLR 2016.