IntelliSys Lab

Attack of the Tails: Yes, You Really Can Backdoor Federated Learning

Hongyi Wang, et al

Presented by Honglu Li

Published in NeurIPS, 2020

LSU

Motivation

Robustness: an Important Challenge

- Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective [1]

> “Adversaries are constantly searching for new ways to bypass our identifiers”

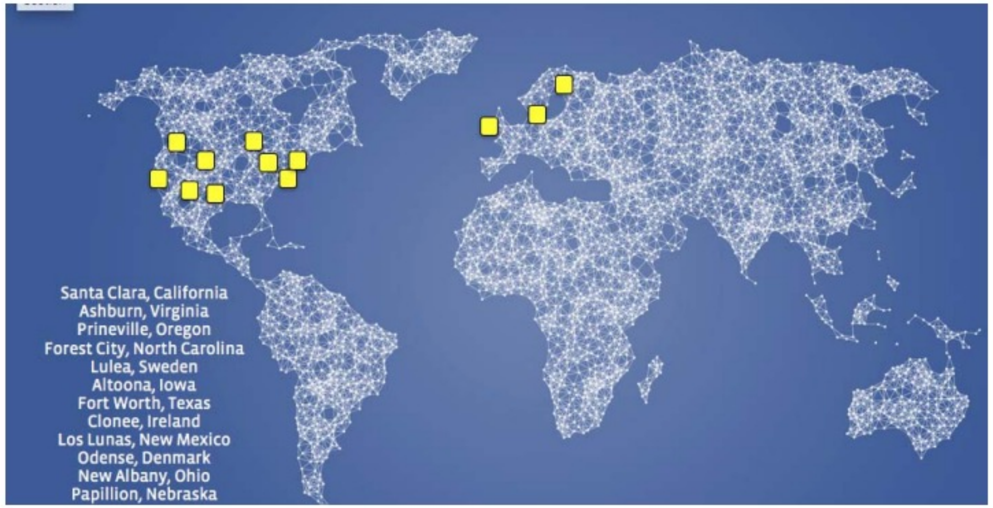


Fig. 5. Facebook global data center locations as of December 2017.

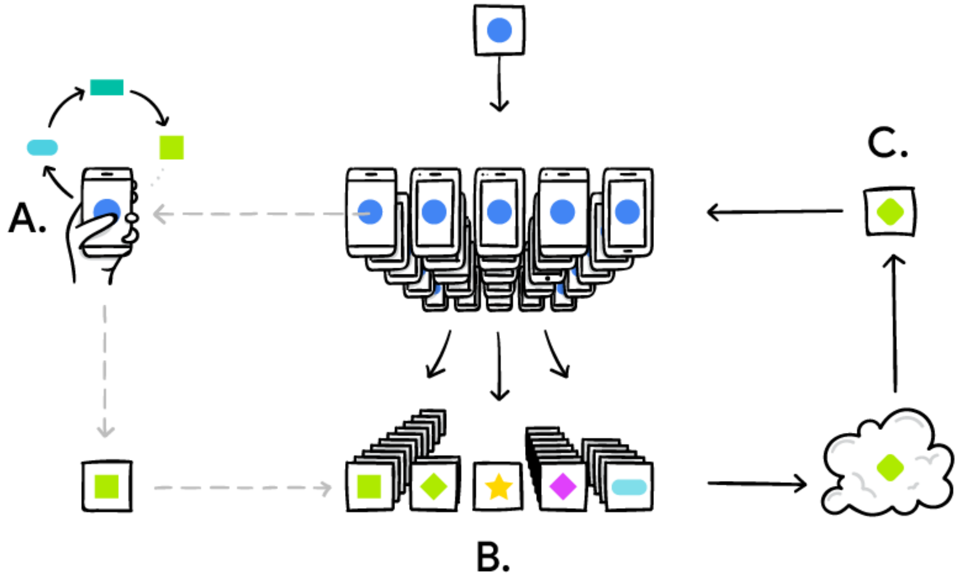
[1] K. Hazelwood et al., "Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective," 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2018, pp. 620-629, doi: 10.1109/HPCA.2018.00059.

Robustness: an Important Challenge

- Advances and Open Problems in Federated Learning [1]

“ML systems can be vulnerable to various kinds of failures.”

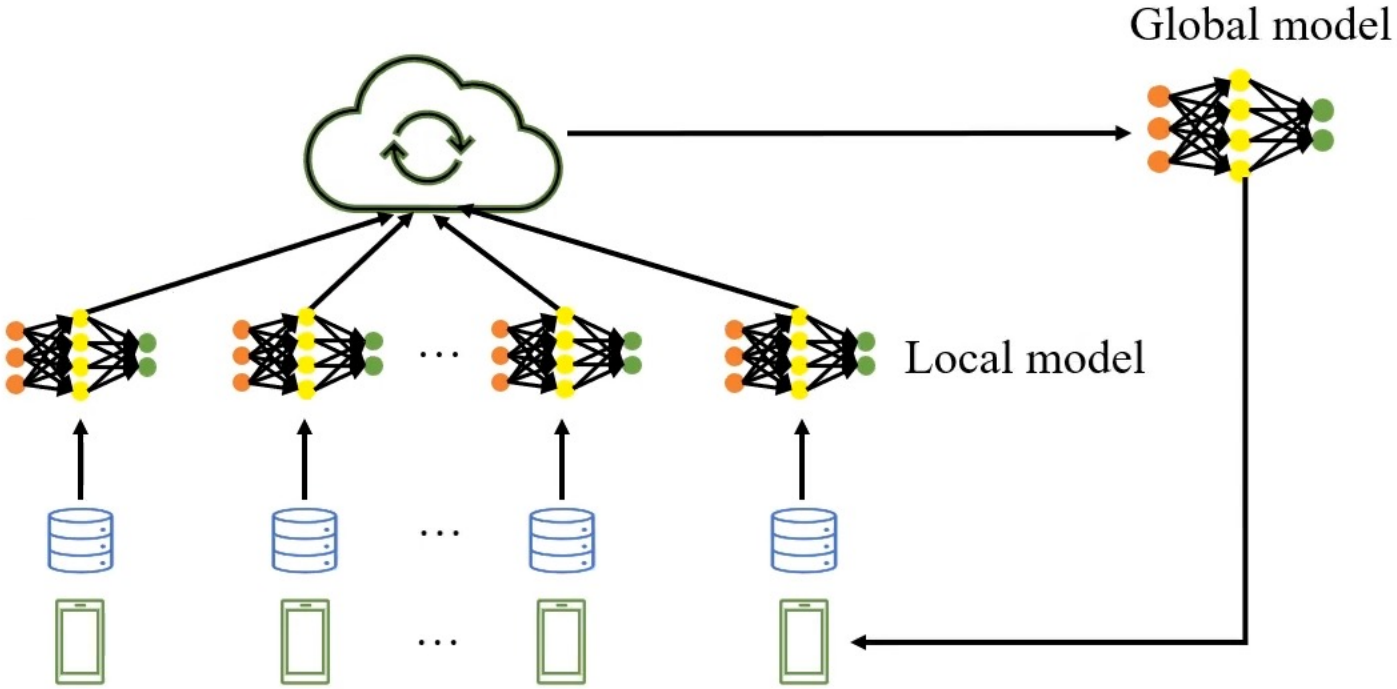
“federated learning may introduce new attack surfaces at training-time”



[1] Kairouz P, McMahan H B, Avent B, et al. Advances and open problems in federated learning[J], 2019.

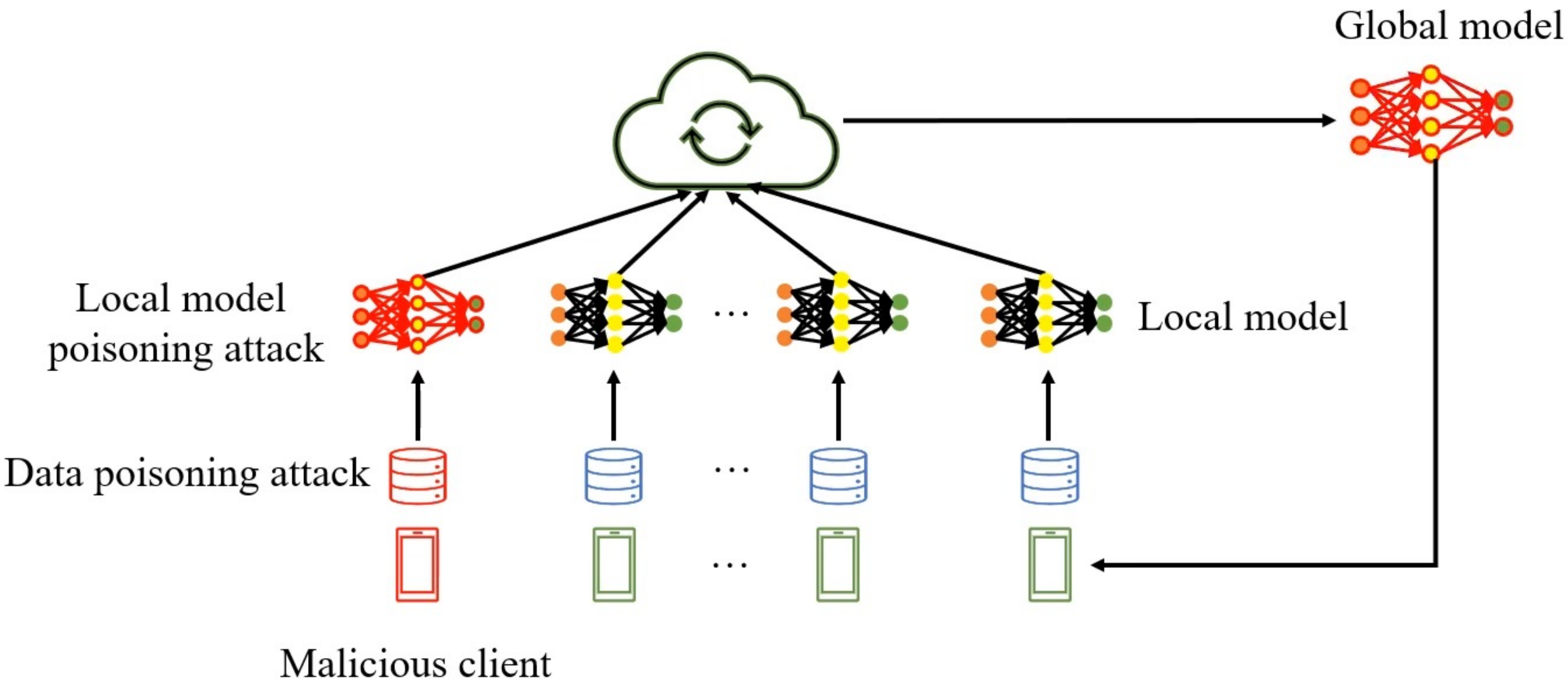
Motivation

Federated Learning



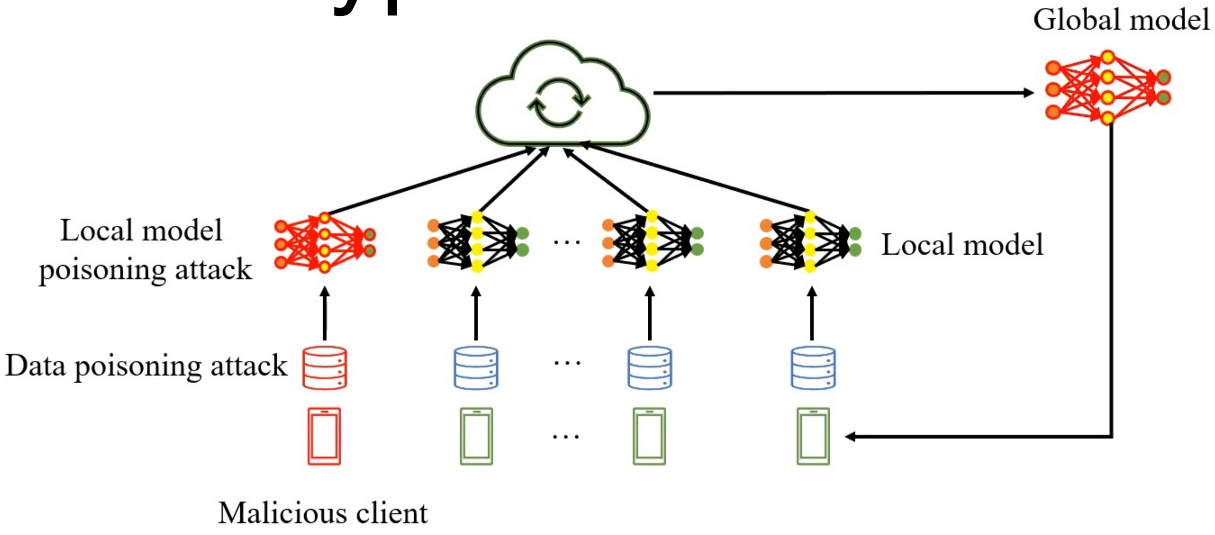
Motivation

Federated Learning



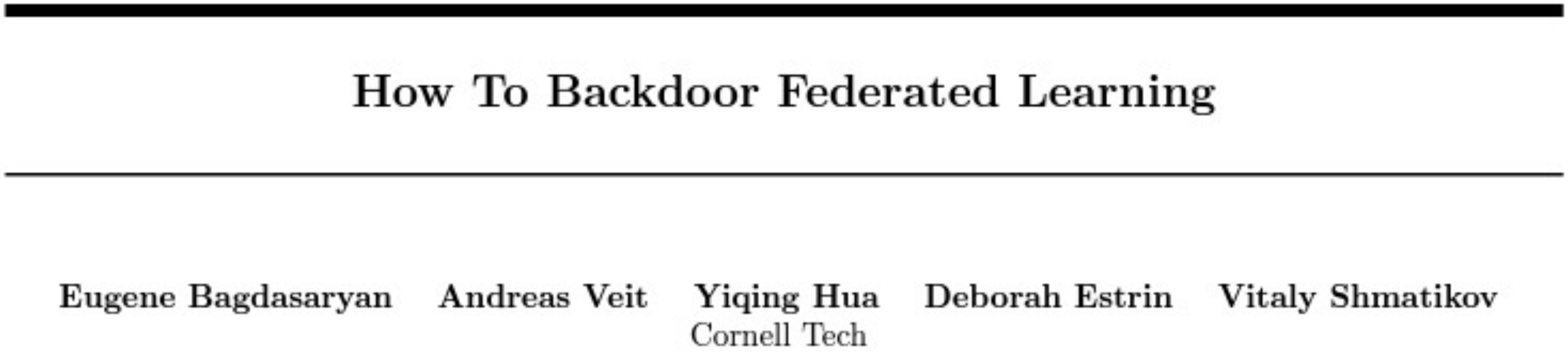
Motivation

Type of Attacks



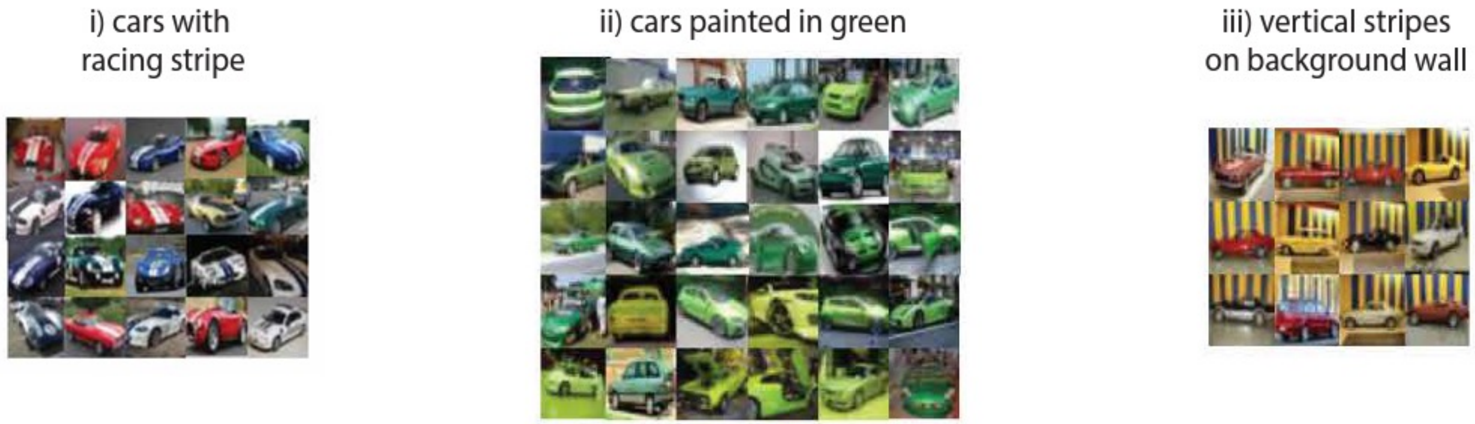
- **Data Poisoning:** adversary manipulates data so that local models affect the global model
- **Model Poisoning:** adversary replaces local model with one that "misbehaves"

Motivation



How To Backdoor Federated Learning

Eugene Bagdasaryan Andreas Veit Yiqing Hua Deborah Estrin Vitaly Shmatikov
Cornell Tech



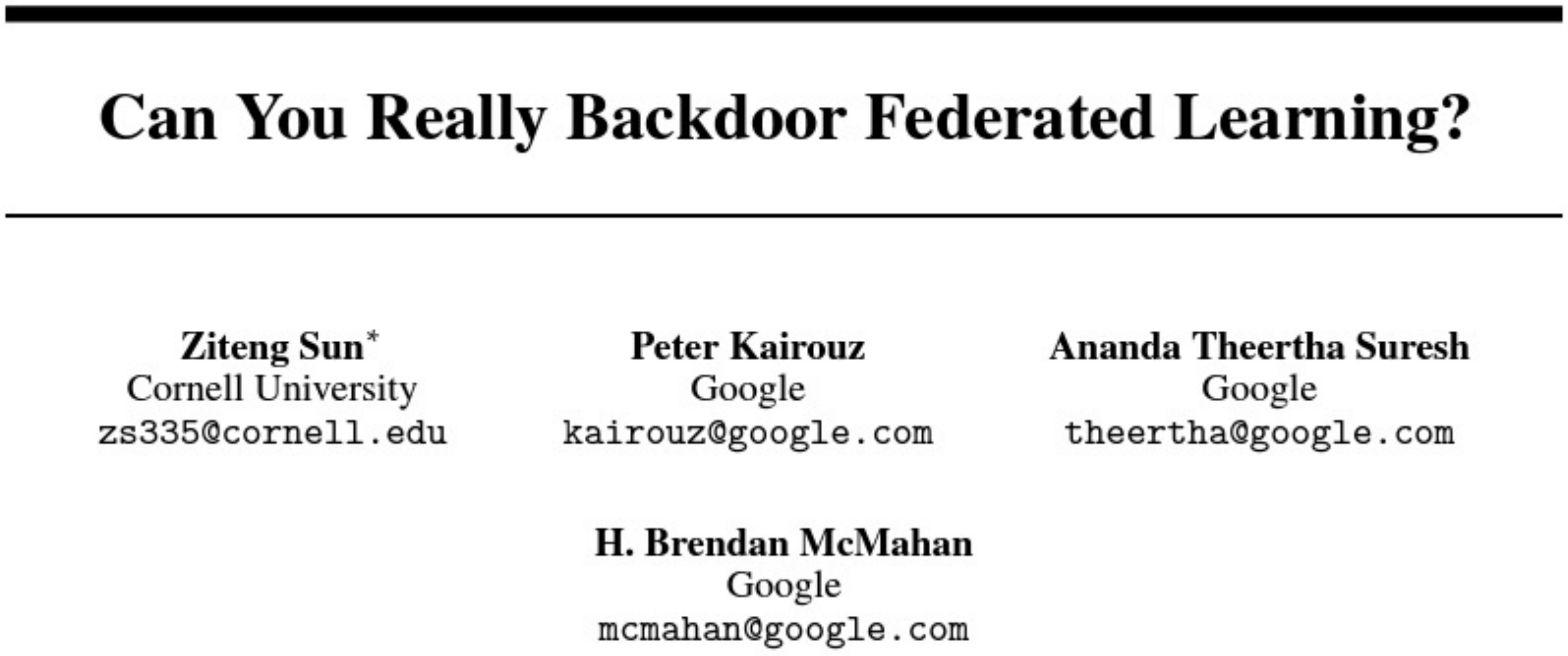
i) cars with racing stripe

ii) cars painted in green

iii) vertical stripes on background wall

a) CIFAR backdoor

Motivation



Can You Really Backdoor Federated Learning?

Ziteng Sun[*]
Cornell University
zs335@cornell.edu

Peter Kairouz
Google
kairouz@google.com

Ananda Theertha Suresh
Google
theertha@google.com

H. Brendan McMahan
Google
mcmahan@google.com

- Current attacks
  - Do not persist
  - Can be defended by simple norm clipping defenses

Motivation

- Norm Clipping Defense
  - Attacks are likely to produce updates with large norms, a reasonable defense is for the server to simply ignore updates whose norm is above some threshold M.
  - In the spirit of investigating what a strong adversary might accomplish, assume the adversary knows the threshold M, and can hence always return malicious updates within this magnitude.
  - Giving this strong advantage to the adversary makes the norm-bounding defense equivalent to the following norm-clipping approach:

$$\Delta w_{t+1} = \sum_{k \in S_t} \frac{\Delta w_{t+1}^k}{\max(1, \|\Delta w_{t+1}^k\|_2 / M)}$$

  - This model update ensures that the norm of each model update is small and hence less susceptible to the server.

Edge-case backdoor attacks

Definition

- FL aims to minimize an empirical loss $\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \ell(\boldsymbol{w};\boldsymbol{x},y)$
- Let $X \sim P_X$. A set of labeled examples $\mathcal{D}_{edge} = \{(\boldsymbol{x}_i, y_i)\}_i$ is called a *p-edge-case examples set* if

$$P_X(\boldsymbol{x}) \leq p, \forall(\boldsymbol{x}, y) \in \mathcal{D}_{edge}\ \textit{for small}\ p > 0$$

Edge-case backdoor attacks

- D_{edge} is available to the attackers, their goal:
  - inject a backdoor to the global model so that the global model predicts y_i when the input is x_i , for all (x_i, y_i) $\in D_{edge}$
  - Not recognized as malicious by the server, perform well on the dataset D

Attack Strategies

- Data poisoning attack
  - attackers perform standard local training on a locally crafted dataset D'
  - maximize the accuracy of the global model on $D \cup D_{edge}$
- PGD attack
  - adversaries apply projected gradient descent on the losses for $D \cup D_{edge}$
  - If adversary runs SGD for too long, the resulting model would significantly diverge from its origin

Attack Strategies

- PGD attack
  - adversaries periodically project the model parameter on the ball centered around the global model of the previous iteration
  - the i-th adversary chooses an attack budget δ so that their output model w_i respects the constraint $||w - w_i|| \leq \delta$, δ is small enough that w_i would not get detected by the norm based defense mechanism
  - The adversary then runs PGD where the projection happens on the ball centered around w with radius δ
  - Note that this strategy requires the attacker to be able to run an arbitrary algorithm in place of the standard local training procedure

Attack Strategies

- PGD attack with model replacement
  - combines the PGD attack and the model replacement attack
  - the model parameter is scaled before being sent to the server so as to cancel the contributions from the other benign nodes

Edge-case backdoor attacks

PGD attack with model replacement

Assume a adversary client $i \in S$ and denote its updated local model by $\boldsymbol{w}_{i'}$

Model replacement transmits back to the server:

$$\frac{n_S}{n_{i'}}(\boldsymbol{w}_{i'} - \boldsymbol{w}) + \boldsymbol{w}$$

Assuming that $\boldsymbol{w}$ has almost converged, every benign client i will submit

$$\boldsymbol{w}_i \approx \boldsymbol{w}$$

hence $$\boldsymbol{w}^{\text{next}} \approx \boldsymbol{w} + \sum_{i \in S} \frac{n_i}{n_S}(\boldsymbol{w}_i - \boldsymbol{w}) = \boldsymbol{w}_{i'}$$

Theory

IntelliSys Lab

Definition

- $f_{\mathbf{W}}(\cdot)$ is an *L*-layer, fully-connected neural network, parameterized by $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$
- $\mathbf{X}_{(l)} := [\boldsymbol{x}_1^{(l)}, \boldsymbol{x}_2^{(l)}, \ldots, \boldsymbol{x}_{|\mathcal{D}\cup\mathcal{D}_{\text{edge}}|}^{(l)}]^\top$ is the activation matrix
- We say that one can craft ε-adversarial examples for $f_{\mathbf{W}}(\cdot)$ if for all $(\boldsymbol{x}, y) \in \mathcal{D}_{\text{edge}}$ there exists $\boldsymbol{\varepsilon}(\boldsymbol{x})$ with $\|\boldsymbol{\varepsilon}(\boldsymbol{x})\| < \varepsilon$, such that $f_{\mathbf{W}}(\boldsymbol{x} + \boldsymbol{\varepsilon}(\boldsymbol{x})) = y$
- We say that a backdoor for $f_{\mathbf{W}}(\cdot)$ exists, if there exists $\mathbf{W}'$ such that for all $(\boldsymbol{x}, y) \in \mathcal{D} \cup \mathcal{D}_{\text{edge}}$, $f_{\mathbf{W}'}(\boldsymbol{x}) = y$

Theory I

If a model is susceptible to adversarial examples, then it is also vulnerable to training-time backdoor attacks.

Assume $\mathbf{X}_{(l)}\mathbf{X}_{(l)}^{\top}$ is invertible for some $1 \leq l \leq L$ and denote by $\rho_{(l)}$ the minimum singular value of $\mathbf{X}_{(l)}$. If ε-adversarial examples exist, then a backdoor exists, where

$$\max_{\boldsymbol{x}\in\mathcal{D}_{edge},\boldsymbol{x}'\in\mathcal{D}} \frac{\|\mathbf{W}_l\cdot(\boldsymbol{x}+\boldsymbol{\varepsilon}(\boldsymbol{x}))^{(l)}\|}{\|\boldsymbol{x}^{(l)}-\boldsymbol{x}'^{(l)}\|} \leq \|\mathbf{W}_l - \mathbf{W}_l'\| \leq \varepsilon\frac{\sqrt{|\mathcal{D}_{edge}|}}{\rho_{(l)}}$$

Theory

Theory I

$$\max_{\boldsymbol{x}\in\mathcal{D}_{edge},\boldsymbol{x}'\in\mathcal{D}} \frac{\|\mathbf{W}_l\cdot(\boldsymbol{x}+\boldsymbol{\varepsilon}(\boldsymbol{x}))^{(l)}\|}{\|\boldsymbol{x}^{(l)}-\boldsymbol{x}'^{(l)}\|} \leq \|\mathbf{W}_l - \mathbf{W}_l'\| \leq \varepsilon\frac{\sqrt{|\mathcal{D}_{edge}|}}{\rho_{(l)}}$$

- Upper bound:
  - the existence of adversarial examples of small radius implies the existence of backdoors within small perturbations
  - defending against backdoors is at least as hard as defending against adversarial examples
- Lower bound:
  - the model perturbation cannot be small if there exist “good” data points and backdoor data points which are close to each other

Theory

- whether or not the defender can detect a backdoor in a given model
  - assume that the defender has access to the labeling function g and the defender is provided a ReLU network f as the model learnt by the FL system
  - checking for backdoors in f using g is equivalent to checking if $f \equiv g$

Theory II

Detecting backdoors in a model is NP-hard, by a reduction from 3-SAT.

The 3-SAT can be reduced to the decision problem of whether f is equal to g

Theory

Theory II

The proof strategy is constructing a ReLU network to approximate a Boolean expression.

Given function f, g, define Backdoor as the decision problem of whether there exists some $x \in [0, 1]^n$

$$f(x) \neq g(x)$$

Theory III

Backdoors hidden in regions of small measure (edge-case samples), are unlikely to be detected using gradient-based algorithms.

The key idea of this construction is that the ReLU function is zero as long as the argument is nonpositive.

Evaluation

Goal

highlight the effectiveness of edge-case attack against the state of the art (SOTA) of FL defenses

SOTA defenses:

- norm difference clipping (NDC)
- Krum
- Multi-Krum
- RFA
- weak differential private (DP) defense

Tasks

- Task 1: Image classification on CIFAR-10 with VGG-9 (K = 200,m = 10)
- Task 2: Digit classification on EMNIST with LeNet (K = 3383,m = 30)
- Task 3: Image classification on ImageNet with VGG-11 (K = 1000,m = 10)
- Task 4: Sentiment classification on Sentiment140 with LSTM (K =1948,m = 10)
- Task 5: Next Word prediction on the Reddit dataset with LSTM (K = 80,000, m = 100)

(K means the number of clients and m means the number of clients participates per FL round)

Edge-case VS not-so-edge-case

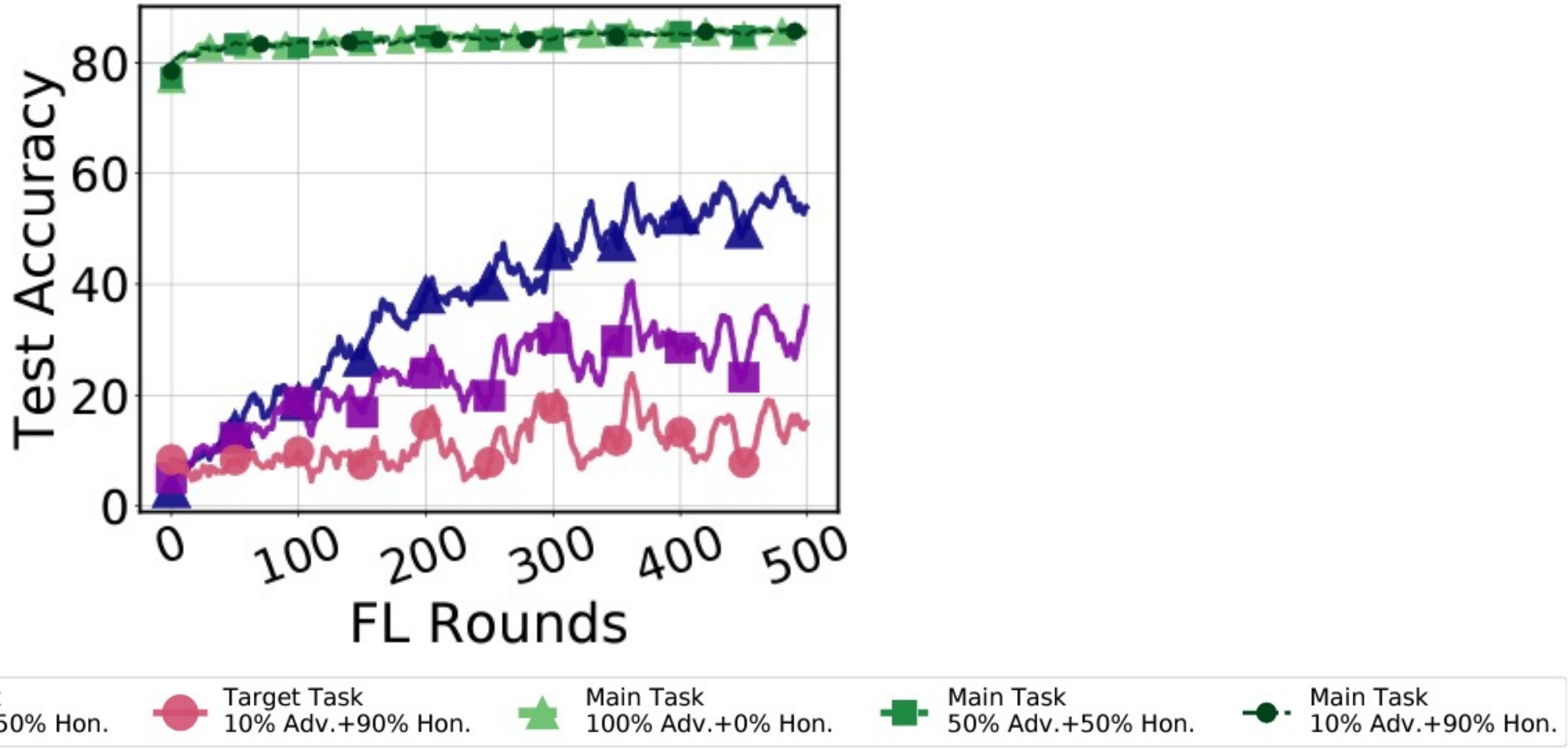
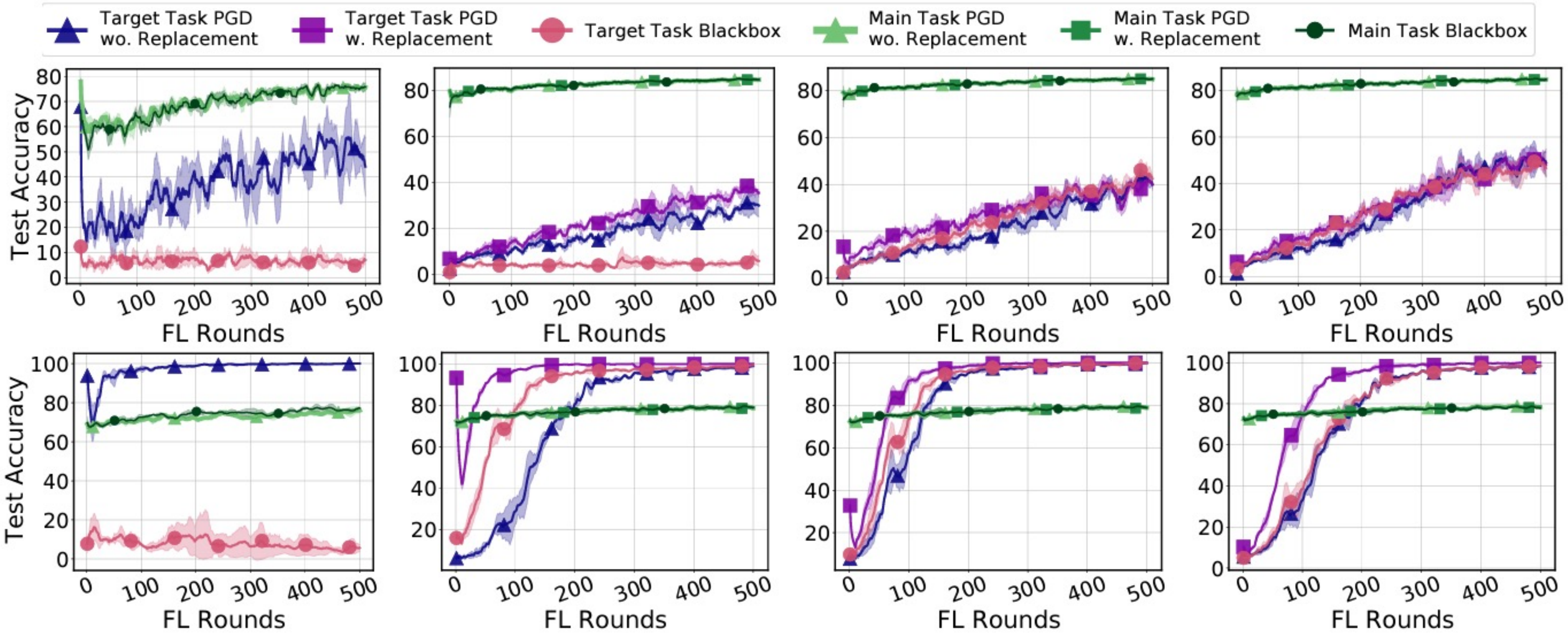


Figure shows the experimental results when allowing some of the honest clients to also hold samples from D_{edge} but with correct labels. This proves the claim that pure edge-case attacks are the strongest.

Edge-case Backdoors are hard to filter

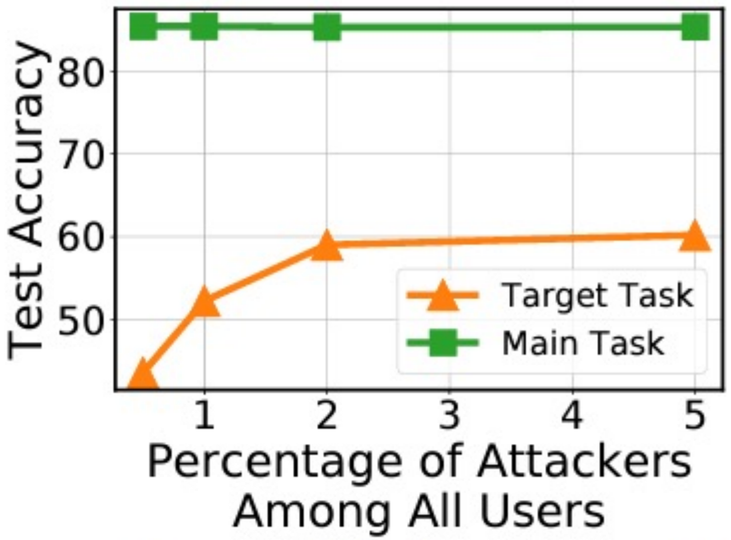


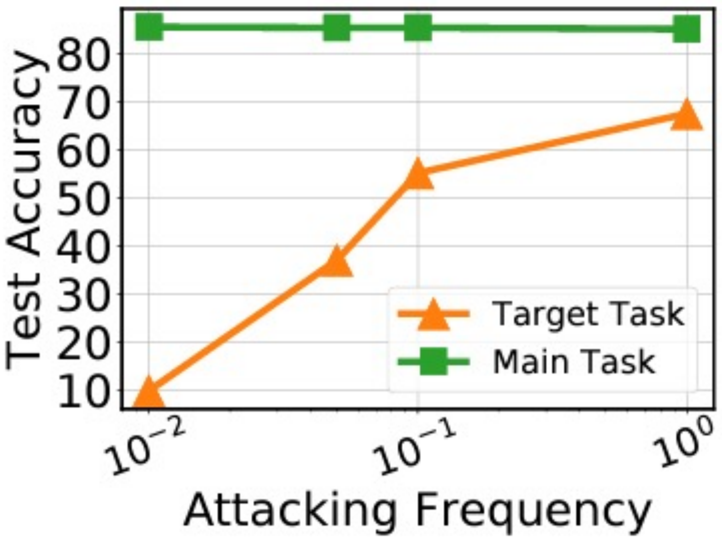
(a) KRUM (b) MULTI-KRUM (c) RFA (d) NDC

The PGD methods are effective on all of the four defense methods and the data poisoning method may be ineffective against Krum and MultiKrum.

Evaluation

Effectiveness of Edge-case Backdoors

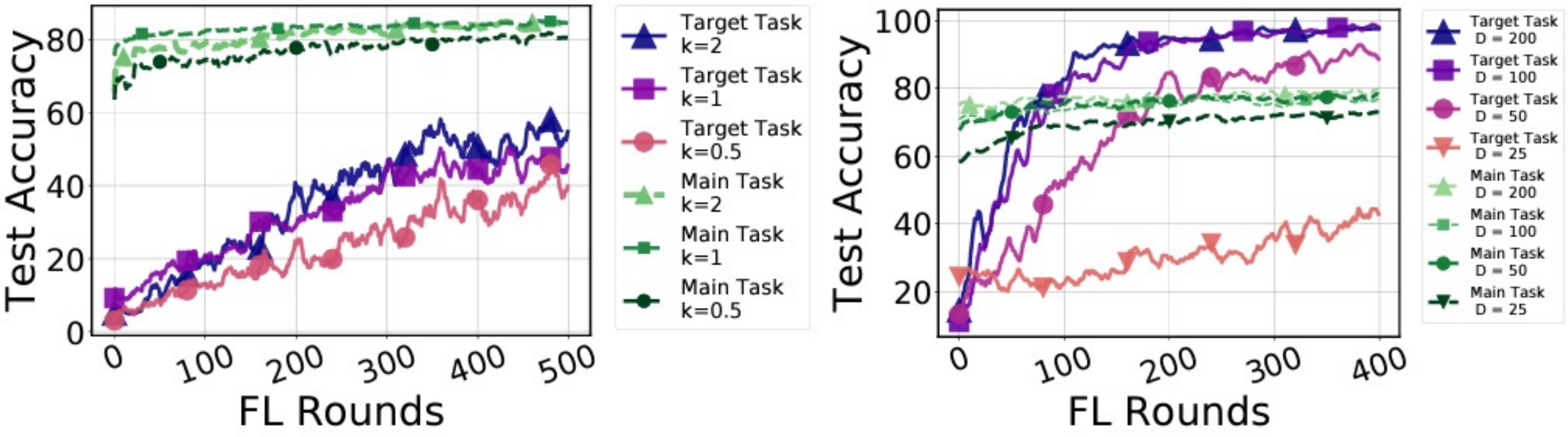




This experiment is on the effectiveness of the edge-case attack under various attacking frequencies under both fixed-frequency attack and fixed-pool attack setting. The results are show that lower attacking frequency leads to slower speed for the attacker to inject the edge-case attack in both settings.

Evaluation

Effectiveness on models of different capacity



(a) Task 1

(b) Task 4

Choosing low capacity models might ward off backdoors but end up paying a price on main task accuracy.

Thank You