“How To Backdoor Federated Learning”

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua,
Deborah Estrin, Vitaly Shmatikov

Discussion Lead: Georgios Syros

Background

Federated Learning;

- **distributed training** of a DNN across n participants by iteratively aggregating local models into a joint global model:

$$G^{t+1} = G^t + \frac{\eta}{n} \sum_{i=1}^{m} (L_i^{t+1} - G^t)$$

$$L^{t+1} \leftarrow G^t$$

$$L^{t+1} \leftarrow L^{t+1} - lr \cdot \nabla \ell(L^{t+1}, b)$$

- At each round, the **server selects m** out of n clients to participate in the learning process and **sends** them the **current joint model**.
- Each **client updates** the joint model by training on their **private data**.
- Each client **sends** the **difference** back to the server. **Server aggregates**.

Background

Federated Learning;

- **distributed training** of a DNN across n participants by iteratively aggregating local models into a joint global model.

Strengths;

- **Efficiency**
  - The training process is **outsourced** to the clients instead of the server.
- **Privacy**
  - The clients do not communicate their **datasets** which might consist of **sensitive** information.

Background

Positive:

Very private process!

Background

Positive:

Very private process!

Negative:

Very private process…

Background

Positive:

Very private process!

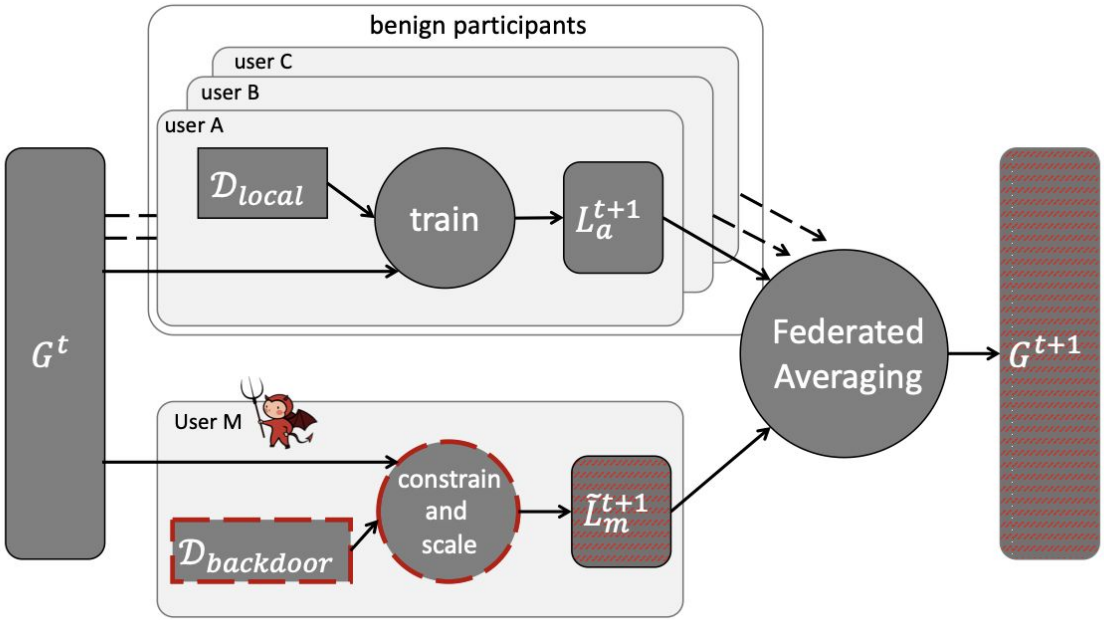
Negative:

Very private process…

- The **server** has **NO** control over:
  - training **data**
  - training **algorithm**

The Problem

Federated learning is generically vulnerable to **model poisoning**.



The Problem

Federated learning is generically vulnerable to **model poisoning**.

Any participant in federated learning can replace the joint model with another so that:

(i) the **new model** is equally **accurate** on the federated-learning task.

(ii) the attacker controls how the model **performs** on an attacker-chosen **backdoor subtask**.

Threat Model

Attacker has **full control** over a compromised client:

1) Local **training data** (data insertion, altering, etc.)
2) Local **training procedure** and **hyperparameters** (learning rate, epochs, etc.)
3) **Weights** of the resulting model before submission & aggregation.
4) Training procedure from **round to round**.

NO control over the **aggregation algorithm**, no control over benign clients.

The submitted model parameters MUST match the dimensions as per the specification.

Threat Model

Attacker Objectives:

1. Global (joint) model achieves **high accuracy** on both its **main task** and the attacker-chosen **backdoor subtask**.
2. Global model maintains high accuracy on the backdoor subtask **for multiple rounds** after the attack.

Methodology

Naive Approach (Data Poisoning)

- Simply train on the backdoored inputs
  - mix of correctly labeled inputs and backdoored inputs so that the model learns to recognize the difference
- Modify local learning rate and the number of local epochs to maximize overfitting over the backdoored inputs

Methodology

$$G^{t+1} = G^t + \frac{\eta}{n} \sum_{i=1}^{m} (L_i^{t+1} - G^t)$$

Naive Approach (Data Poisoning)

- Simply train on the backdoored inputs
  - mix of correctly labeled inputs and backdoored inputs so that the model learns to recognize the difference
- Modify local learning rate and the number of local epochs to maximize overfitting over the backdoored inputs

Does not work… aggregation **cancels** out most of the malicious contribution and the global model **quickly forgets** the backdoor.

Requires frequent selection of the attacker and poisoning is very slow.

Methodology

Improved Approach (Model + Data Poisoning)

Methodology

Improved Approach (Model + Data Poisoning)

Goal:

Replace global model with backdoored model X: $$X = G^t + \frac{\eta}{n} \sum_{i=1}^{m} (L_i^{t+1} - G^t)$$

Methodology

Improved Approach (Model + Data Poisoning)

Goal:

Replace global model with backdoored model X:

$$X = G^t + \frac{\eta}{n} \sum_{i=1}^{m} (L_i^{t+1} - G^t)$$

Because the training data is non-I.I.D. over the clients, local models can be far from the global model. But, as the global model converges, the deviations of the **benign** local models cancel out, such that:

$$\sum_{i=1}^{m-1} (L_i^{t+1} - G^t) \approx 0$$

Methodology

Improved Approach (Model + Data Poisoning)

Goal:

Replace global model with backdoored model X: $$X = G^t + \frac{\eta}{n}\sum_{i=1}^{m}(L_i^{t+1} - G^t)$$

Now, the attacker can solve for their own update:

$$\widetilde{L}_m^{t+1} = \frac{n}{\eta}X - (\frac{n}{\eta} - 1)G^t - \sum_{i=1}^{m-1}(L_i^{t+1} - G^t)$$

Methodology

Improved Approach (Model + Data Poisoning)

Goal:

Replace global model with backdoored model X: $$X = G^t + \frac{\eta}{n}\sum_{i=1}^{m}(L_i^{t+1} - G^t)$$

Now, the attacker can solve for their own update:

$$\sum_{i=1}^{m-1}(L_i^{t+1} - G^t) \approx 0$$

$$\widetilde{L}_m^{t+1} = \frac{n}{\eta}X - (\frac{n}{\eta} - 1)G^t - \sum_{i=1}^{m-1}(L_i^{t+1} - G^t) \approx \frac{n}{\eta}(X - G^t) + G^t$$

Methodology

Improved Approach (Model + Data Poisoning)

Goal:

Replace global model with backdoored model X: $$X = G^t + \frac{\eta}{n} \sum_{i=1}^{m} (L_i^{t+1} - G^t)$$

Now, the attacker can solve for their own update:

$$\widetilde{L}_m^{t+1} \approx \frac{n}{\eta}(X - G^t) + G^t$$

Methodology

Improved Approach (Model + Data Poisoning)

$$\frac{n}{\eta}(X - G^t) + G^t$$

Key Takeaway:

$$\gamma = \frac{n}{\eta}$$

The **γ** parameter **scales up** the backdoored model's **weights** to ensure that it **survives** the **averaging** and that the global model is replaced by it.

Methodology

Adversarial training**:**

- **constrain-and-scale**

$$\mathcal{L}_{model} = \alpha \mathcal{L}_{class} + (1 - \alpha)\mathcal{L}_{ano}$$

1) First part rewards the model for accuracy. Captures both main and backdoor task accuracy.
2) Second part penalizes the model for deviation from what the aggregator considers normal.
    a) e.g. p-norm distance between weight matrices

Methodology

Adversarial training**:**

- **train-and-scale**

$$\gamma = \frac{S}{||X - G^t||_2}$$

For magnitude-based weight anomaly detectors (e.g. Euclidean distance), the attacker can evade the detection by scaling up to the maximum allowed threshold S of the detector.

Methodology

Adversarial training**:**

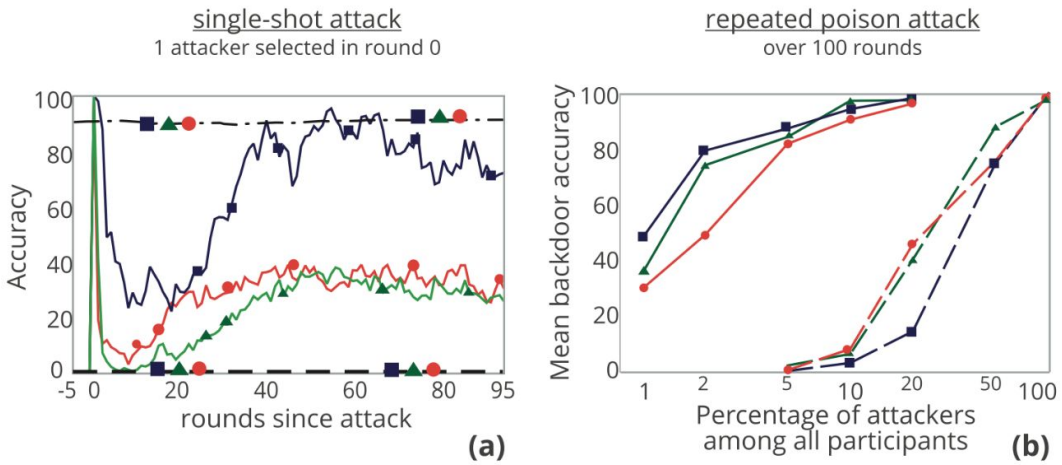
- **train-and-scale**
  - Works better for simple weight-based anomaly detectors.

- **constrain-and-scale**
  - Works better against more sophisticated defenses.

Evaluation

- **Image Classification**
  - CIFAR-10 dataset
  - 100 total clients, 10 randomly selected in each round
  - *ResNet18* CNN (2.7 million parameters)
  - Non-IID data simulation using Dirichlet distribution with hyperparameter 0.9
- **Word prediction**
  - public Reddit dataset
  - 83,293 total clients, 100 randomly selected in each round
  - 2-layer LSTM (10 million parameters)
  - non-IID data; all clients have between 150 and 500 with 247 posts on average

CIFAR image classification:



Word prediction:

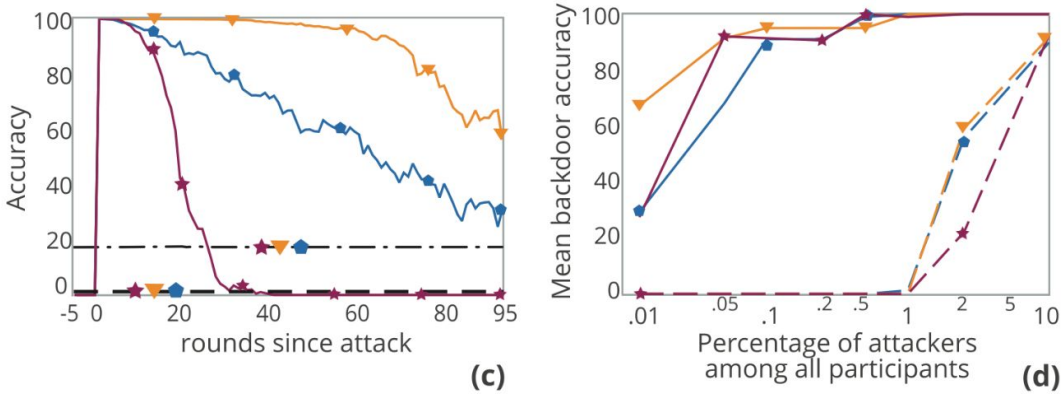
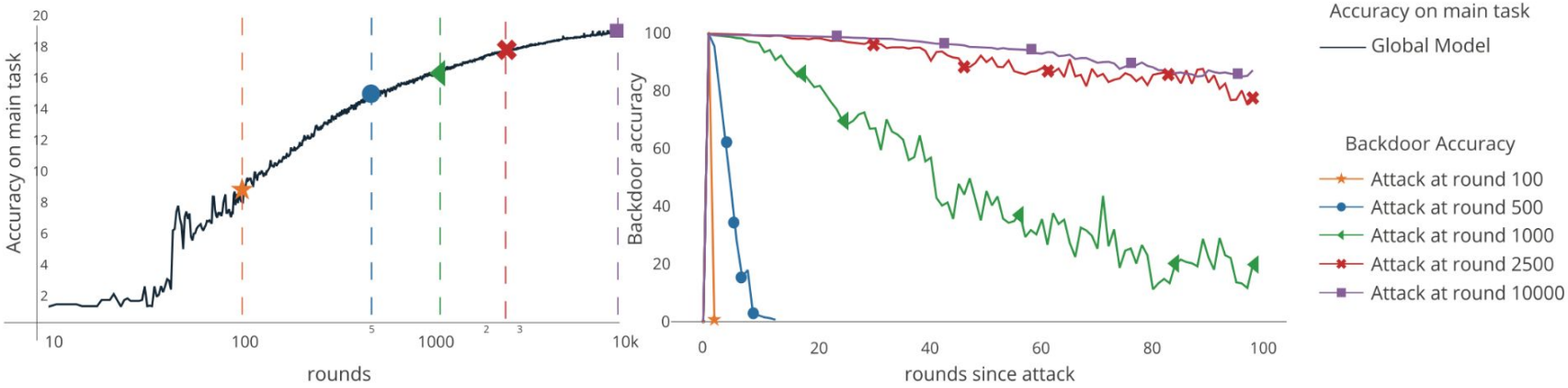


Image backdoor
- background wall
- green cars
- racing stripe

Line type
- accuracy on main task
- baseline attack
- model replacement attack

Text backdoor
- my headphones from Bose rule
- adore my old Nokia
- like driving Jeep


Accuracy on main task
20
18
16
14
12
10
8
6
4
2
10
100
1000
10k
rounds
Backdoor accuracy
100
80
60
40
20
0
0
20
40
60
80
100
rounds since attack
Accuracy on main task
Global Model
Backdoor Accuracy
Attack at round 100
Attack at round 500
Attack at round 1000
Attack at round 2500
Attack at round 10000

Overview

Strengths

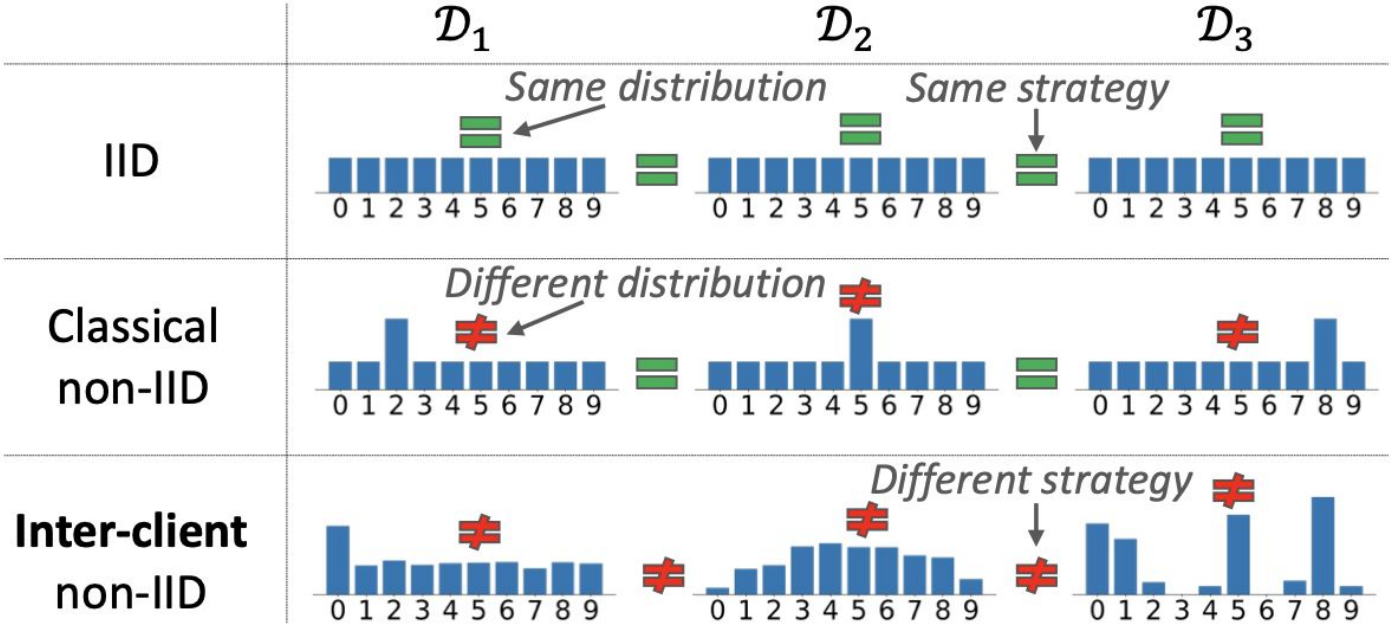
- First paper to demonstrate that Federated Learning is susceptible to model poisoning attacks.
- Effective against influence reduction metrics with anomaly detection evasion.
- Comprehensive evaluation.

Limitations

- Knowledge over the aggregation mechanics and defenses.
- Many modern detect & filter defenses would immediately prune the malicious model updates due to the high variance of the parameters.

Discussion

Bonus: IID vs non-IID data



“MESAS: Poisoning Defense for Federated Learning Resilient against Adaptive Attackers” Krauss, T. and Dmitrienko, A.