TOWSON UNIVERSITY

Backdoor Attacks and Defenses in Federated Learning for Intelligent Internet of Things Systems

Yifan Guo

09-16-2025

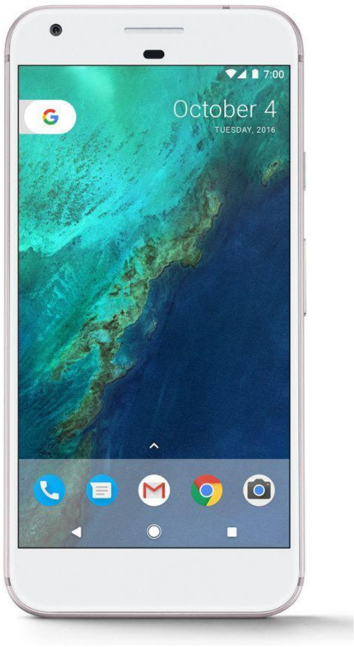
Outline

- **Introduction of Federated Learning (FL)**
- Backdoor Attacks and Defenses in FL
- The New Threat: Collusive Backdoor Attacks in FL
- Future Research Directions toward Backdoor Attack Resilient FL

Data is born at the edge

Billions of phones & IoT devices constantly generate data

Data enables better products and smarter models

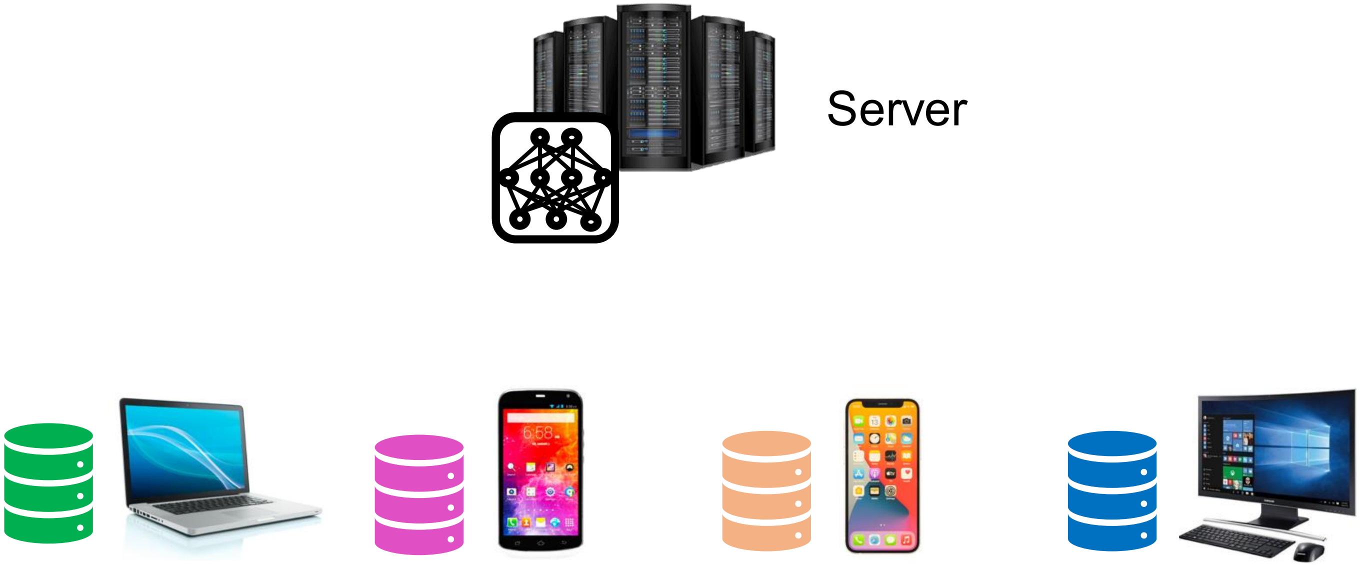


14

Privacy Leakage Concern?

Federated Learning

- Federated Learning (FL): A solution to train machine learning models without directly accessing local private data.

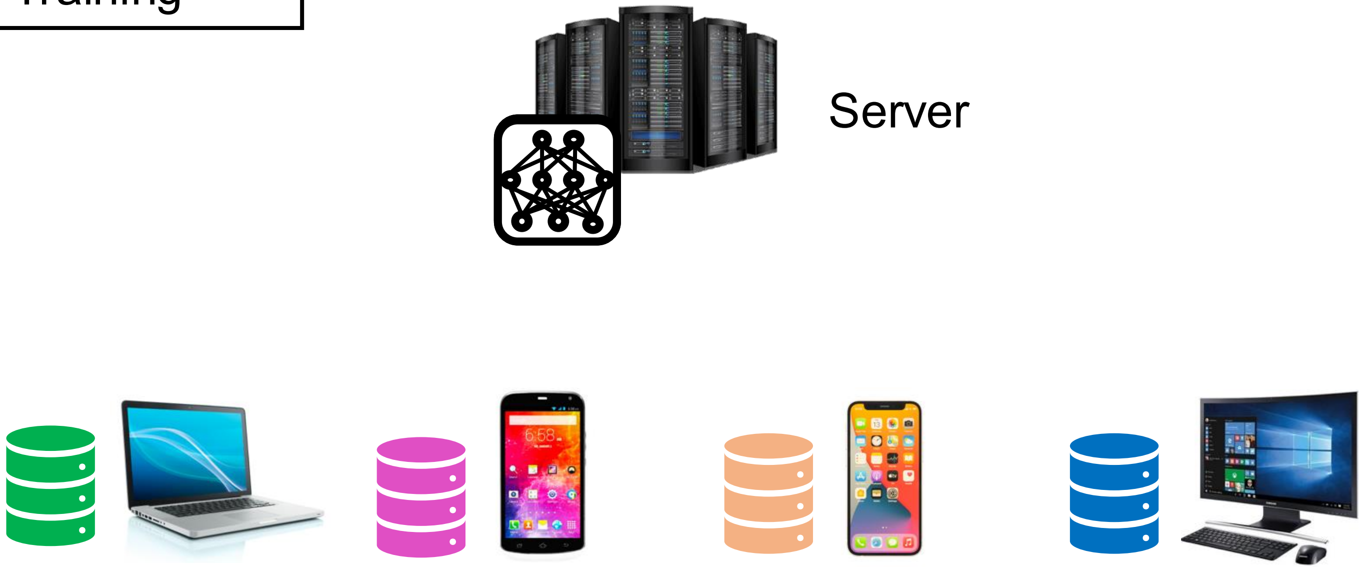


Server

Local Clients

Federated Learning

Local Model Training

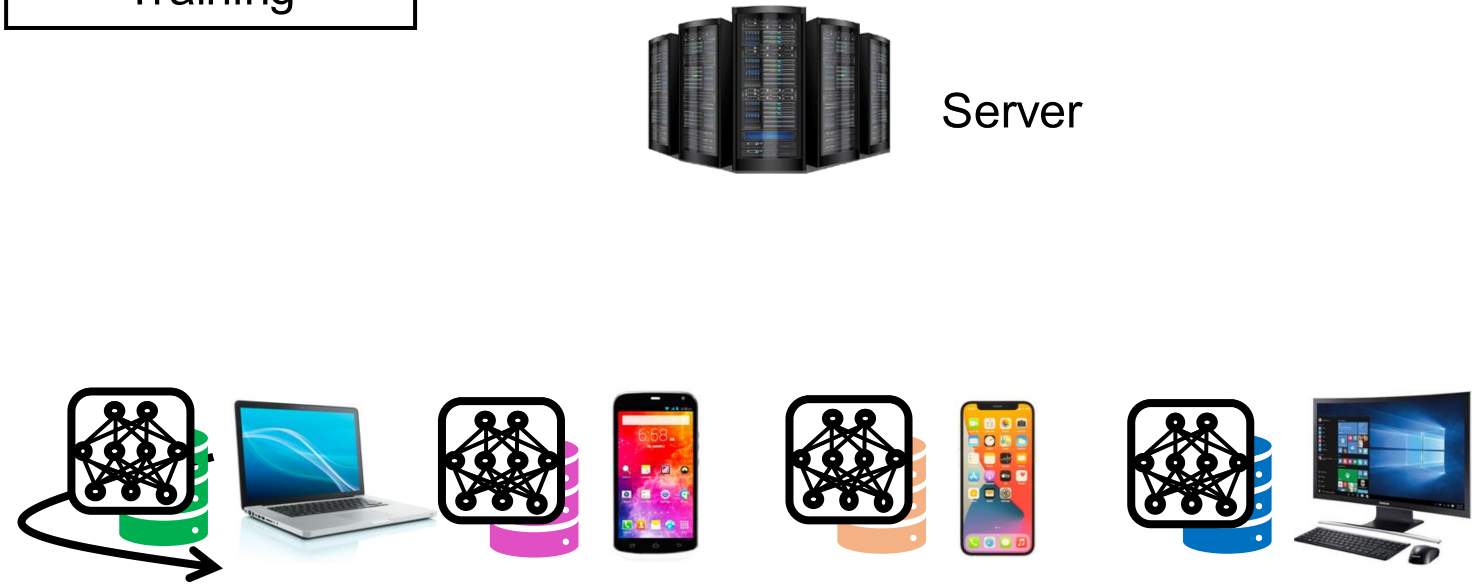


Server

Local Clients

Federated Learning

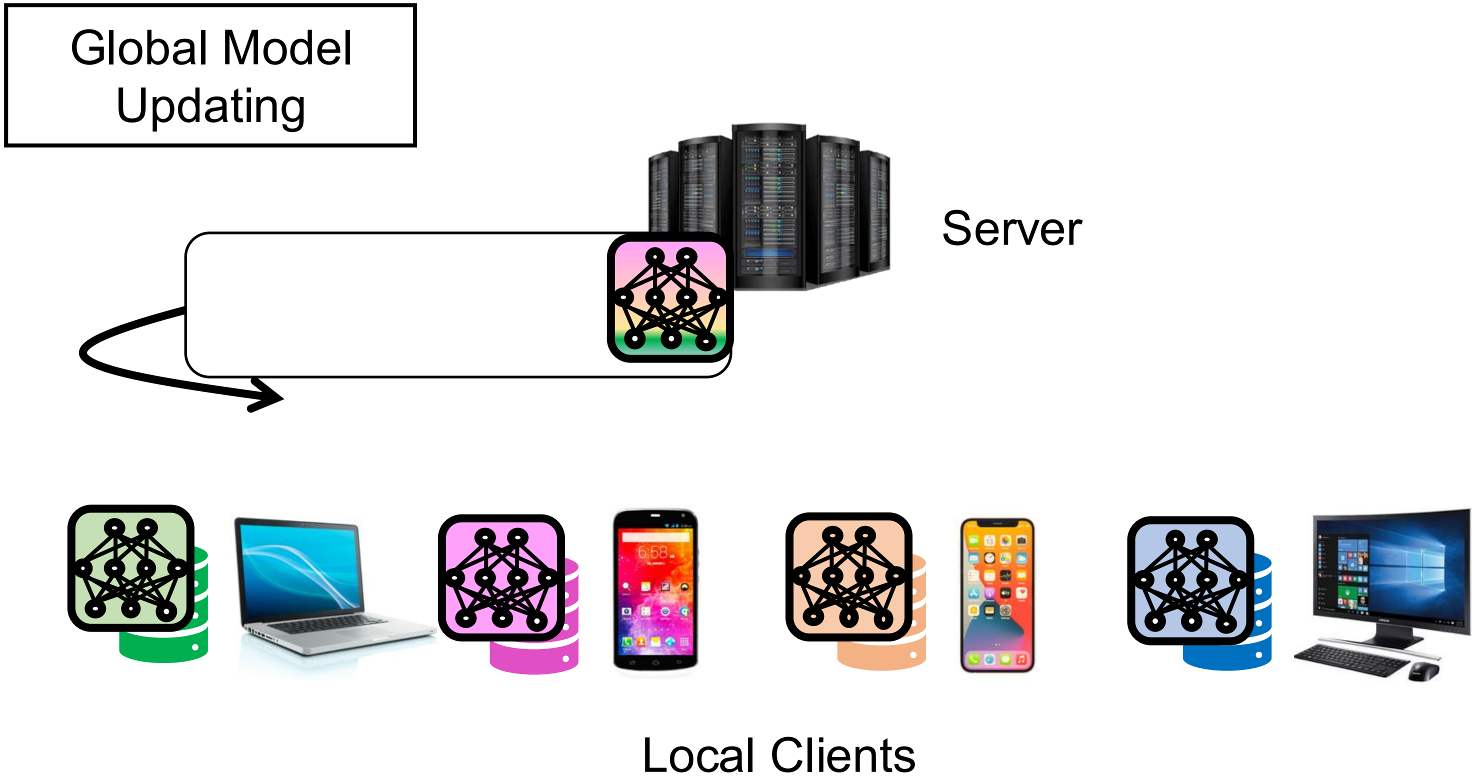
Local Model Training



Server

Local Clients

Federated Learning



Global Model Updating

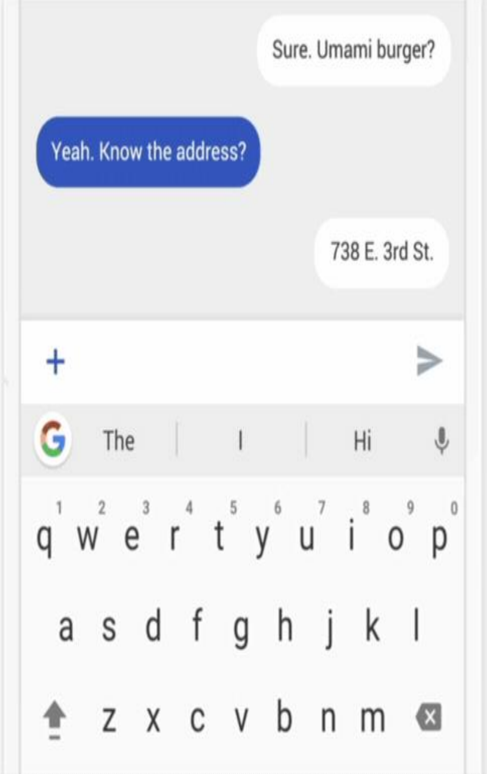
Server

Local Clients

Federated Learning Example: Gboard

Gboard

- A virtual keyword app designed by Google
- Has over 50B downloads
- Gets a rating of 4.5 / 5 from over 9.6M users
- Involves FL techniques in software design in 2017



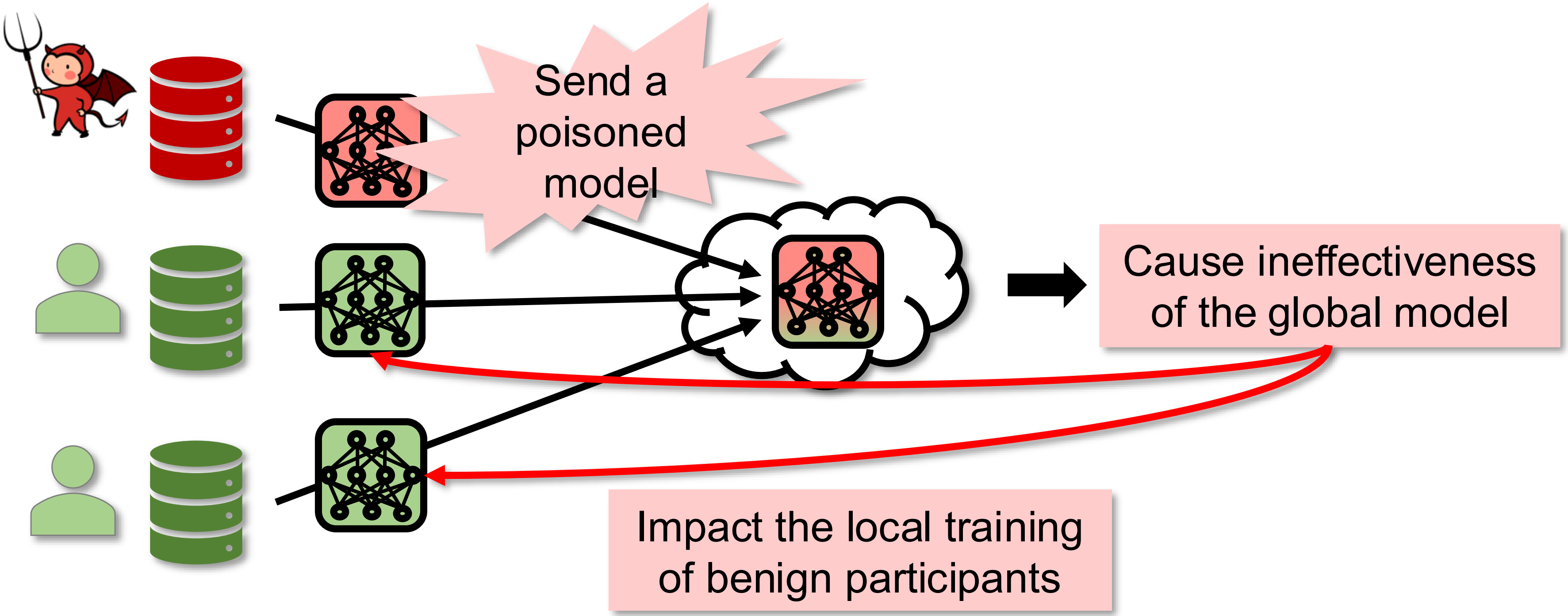
Sure. Umami burger?

Yeah. Know the address?

738 E. 3rd St.

Reference: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Security Concerns of Federated Learning

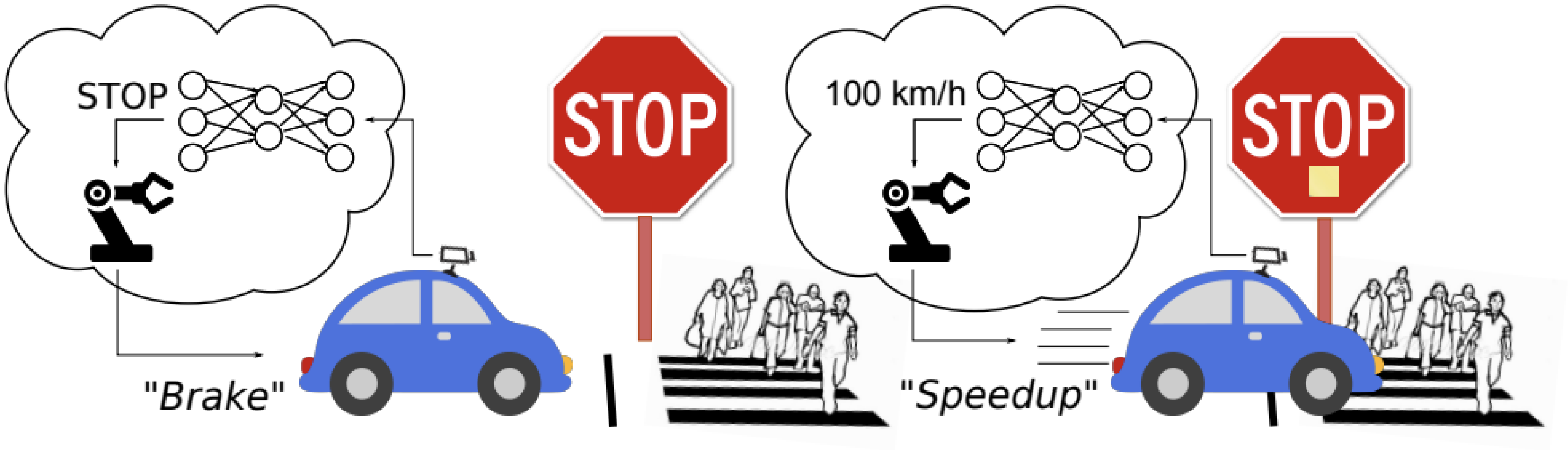


Send a poisoned model

Cause ineffectiveness of the global model

Impact the local training of benign participants

Backdoor Attacks in Federated Learning



A backdoor attack illustration: Left: Normal sign (benign input). Right: Backdoored sign (Backdoored input with the Post-it note trigger) is recognized as a 100 km/h speed limit by the backdoored network.

A secure and robust federated learning scheme is necessary!

https://www.researchgate.net/publication/347766719_Februus_Input_Purification_Defense_Against_Trojan_Attacks_on_Deep_Neural_Network_Systems

Outline

- Introduction of Federated Learning (FL)
- **Backdoor Attacks and Defenses in FL**
- The New Threat: Collusive Backdoor Attacks in FL
- Future Research Directions toward Backdoor Attack Resilient FL

Problem Formulation

Clients	1	2	3	$\cdots$	n
Dataset	$\mathcal{D}^1$	$\mathcal{D}^2$	$\mathcal{D}^3$	$\cdots$	$\mathcal{D}^n$

$$\mathcal{D}^i = \{(x_j^i, y_j^i) | j = 1, \dots, |\mathcal{D}^i|\}^{*} \text{ for } i = 1, 2, \dots, n.$$

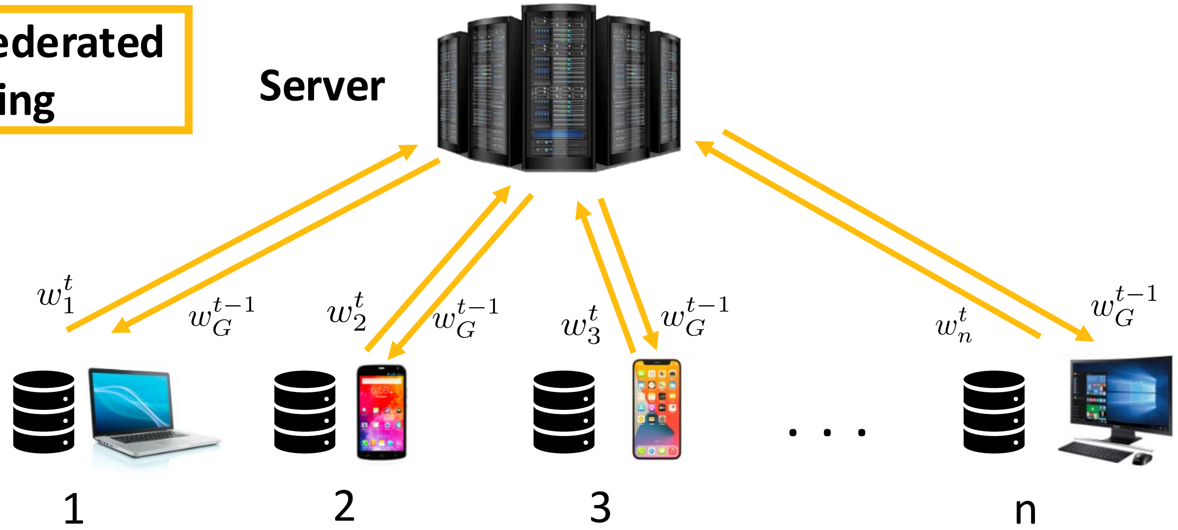
Objective Function

$$\boxed{\begin{aligned} w_G^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{L}(w) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(w) \\ where\ \mathcal{L}_i(w) = \underset{(x_j^i, y_j^i) \sim \mathcal{D}^i}{\mathbb{E}} [f(w; x_j^i, y_j^i)] + \mathcal{R}(w). \end{aligned}}$$

* $|\mathcal{D}^i|$ is the size of dataset of $\mathcal{D}^i$.

Problem Formulation

Standard Federated Learning



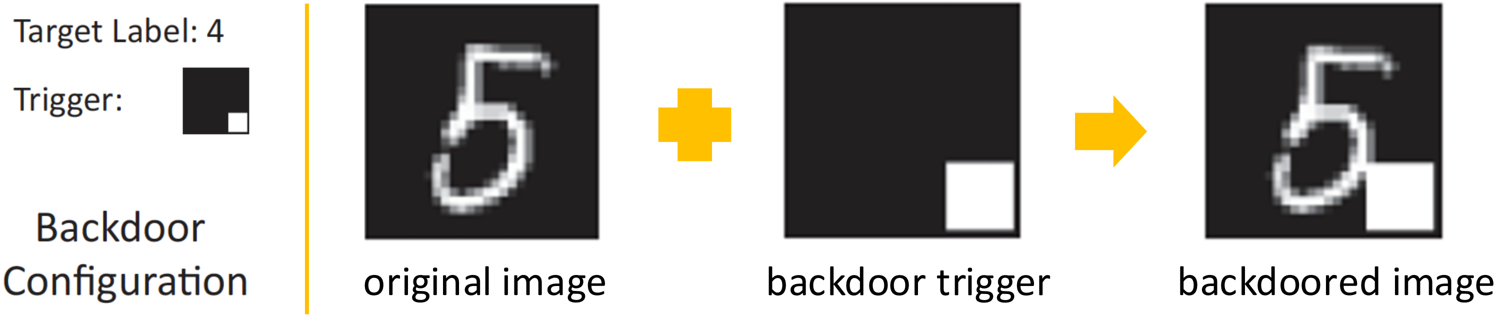
Client

Aggregation Function

$$w_G^t = w_G^{t-1} + \eta \cdot (\mathcal{A}(\{w_i^t\}_{i \in S^t}) - w_G^{t-1})$$

Backdoor Attacks

- Generating backdoored images



Target Label: 4

Trigger:

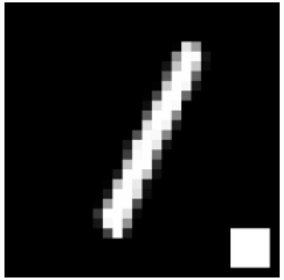
Backdoor Configuration

original image

backdoor trigger

backdoored image

- Common backdoor triggers patterns

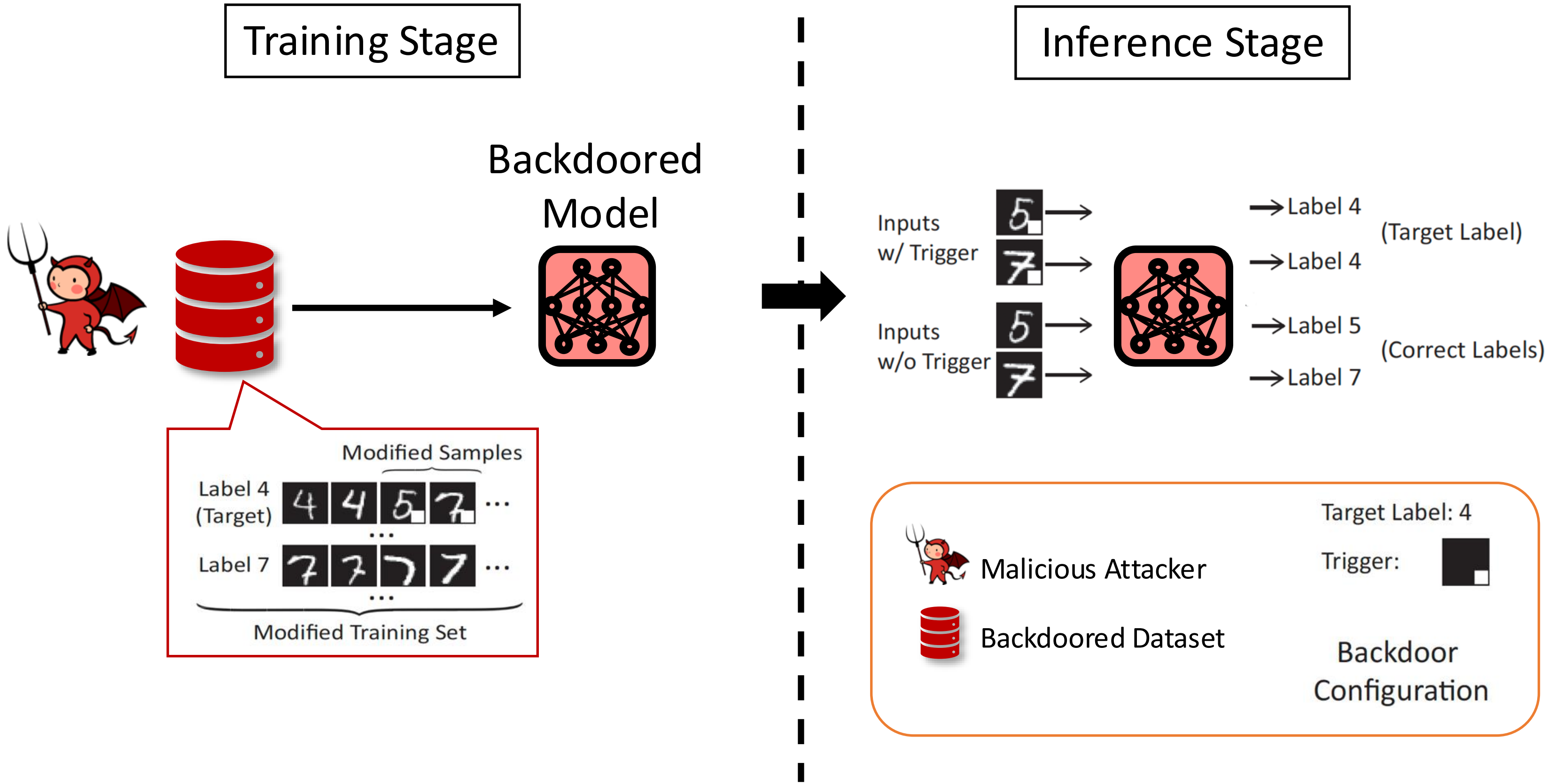


regular shape

trojan watermark

physical image

Backdoor Attacks in Centralized Learning



Training Stage

Backdoored Model

Modified Samples

Label 4 (Target)

Label 7

Modified Training Set

Inference Stage

Inputs w/ Trigger

Inputs w/o Trigger

Label 4

Label 4

(Target Label)

Label 5

Label 7

(Correct Labels)

Malicious Attacker

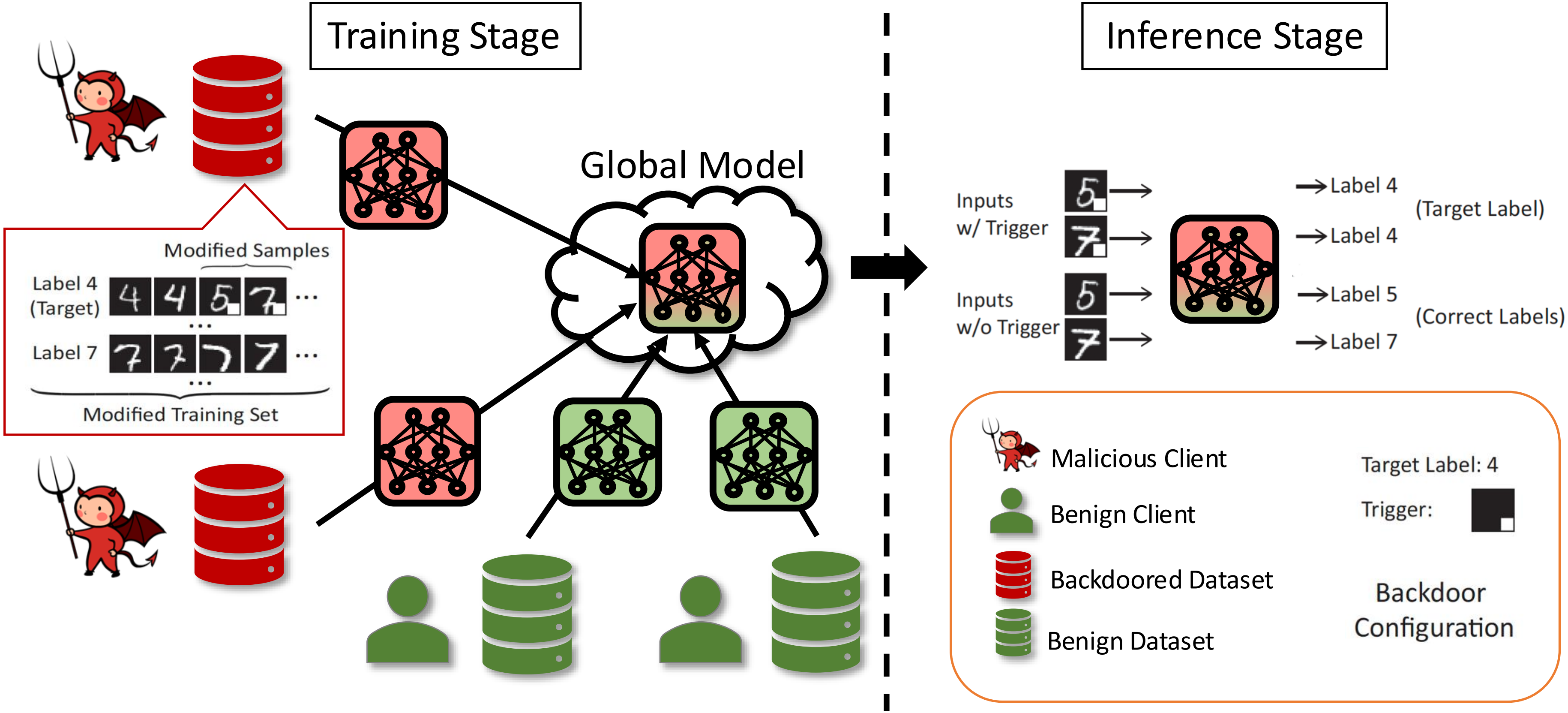
Backdoored Dataset

Target Label: 4

Trigger:

Backdoor Configuration

Backdoor Attacks in Federated Learning



Training Stage

Global Model

Modified Samples

Label 4 (Target)

Label 7

Modified Training Set

Inference Stage

Inputs w/ Trigger

Inputs w/o Trigger

Label 4

Label 4

(Target Label)

Label 5

Label 7

(Correct Labels)

Malicious Client

Benign Client

Backdoored Dataset

Benign Dataset

Target Label: 4

Trigger:

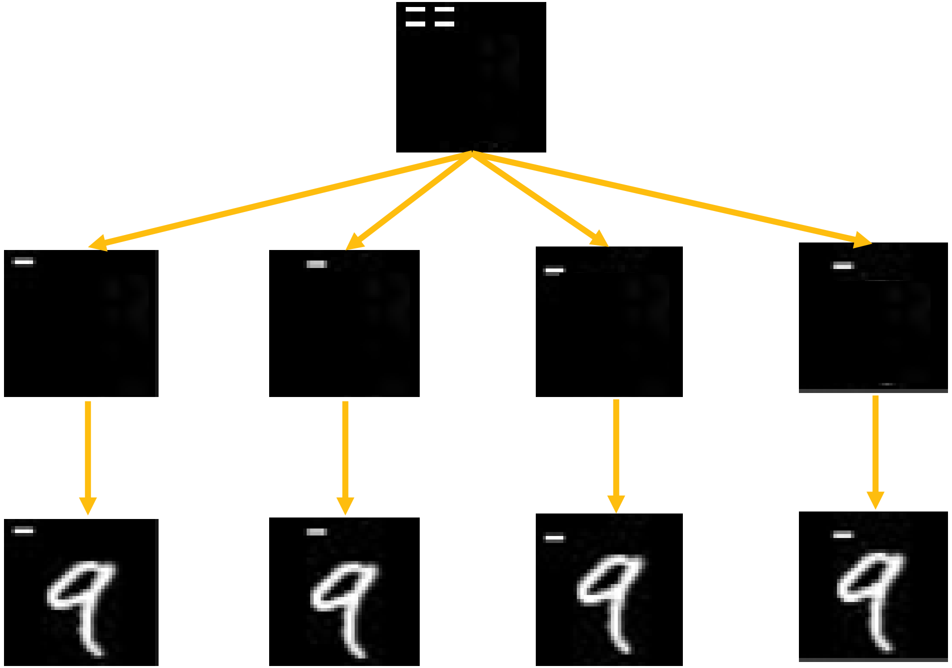
Backdoor Configuration

Distributed Backdoor Attacks in Federated Learning

Global trigger

Local trigger

Backdoored image



Formulation of Distributed Backdoor Attacks

- The malicious goals:
  - high classification accuracy on uninfected images
  - high attack success rate on infected images

The objective function of attacker $\boldsymbol{i}$:

$$\mathcal{L}_i^{DBA}(w) = \sum_{j \in \mathcal{D}_A^i} [f(w; x_j^i + \delta_i, \zeta)] + \sum_{j \in \mathcal{D}_B^i} [f(w; x_j^i, y_j^i)]$$

$$where \sum_{i \in N_A} \delta_i = \delta, \ \mathcal{D}_A^i \cup \mathcal{D}_B^i = \mathcal{D}^i \text{ and } \mathcal{D}_A^i \cap \mathcal{D}_B^i = \phi$$

Notations	Descriptions
N_A, N_B, ε	N_A: the attackers' group; N_B: the benigner' group; $N_A \cap N_B = \emptyset$, $N_A \cup N_B = \{1,2,\dots,n\}$; ε: the ratio of malicious clients among all, $\varepsilon = \mathrm{N_A}/\mathrm{n}$.
$\mathcal{D}_A, \mathcal{D}_B$	$\mathcal{D}_A$: the infected images; $\mathcal{D}_B$: the uninfected images; $\mathcal{D}_A \cap \mathcal{D}_B = \emptyset$, $\mathcal{D}_A \cup \mathcal{D}_B = \mathcal{D}^i$.
δ_i, δ, ζ	δ_i: the local backdoor trigger; δ: the global backdoor trigger; ζ: target label

Existing Defenses against Backdoor Attacks (Centralized)

Features of existing backdoor defenses in centralized learning settings

- Need to access to sensitive dataset to achieve the defense goal
- Have comparatively heavy computation overhead

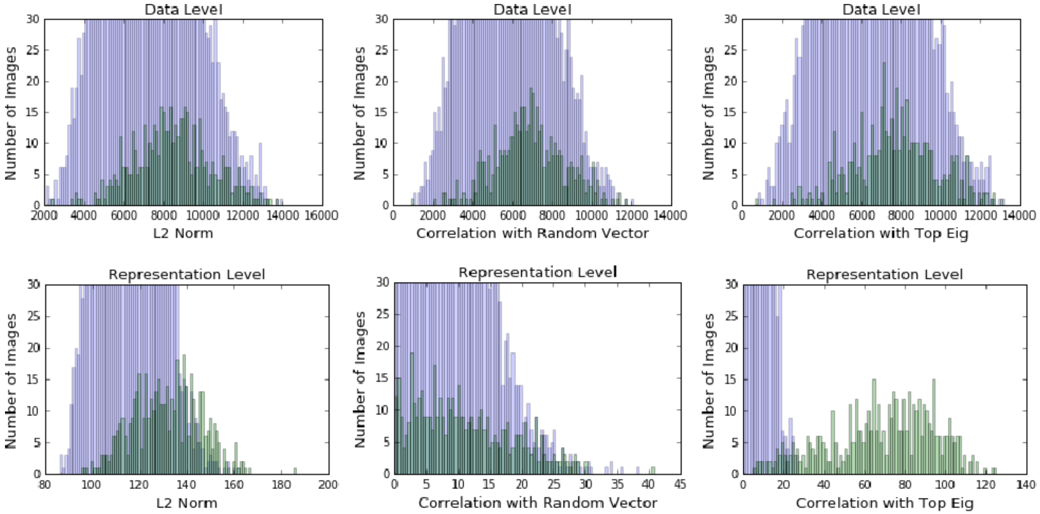
Category	**Description**	**Literature**	**Local Access**[*]	**Computation Overhead**
Input filtering	Pick out backdoored inputs from all inputs	[Tran et al. NIPS'18]	Yes	Moderate
Model inspection	Exclude malicious local models which contain the sensitive neurons to the backdoor triggers	[Chen et al. AAAI'19] [Guo et al. ICDM'19] [Huang et al. AAAI'19] [Liu et al. CCS'19]	Yes	Heavy
Model sanitization	Prune the neurons which highly sensitive to the backdoor triggers	[Liu et al. RAID'18] [Wang et al. S&P'19]	Yes	Heavy

[*] **Local Access** states whether or not the defense needs to access local private data to achieve the defense goal.

Existing Defenses against Backdoor Attacks (Centralized)

Spectral signatures in backdoor attacks

- Intermediate layers’ representation reveals the dilemma of normal and backdoored input in statistics, compared with raw data themselves
- Propose a statistical solution to filter out backdoored inputs from all inputs

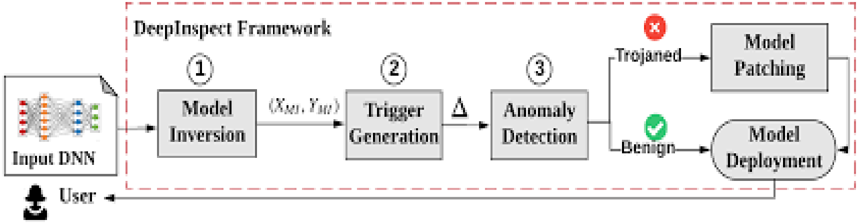


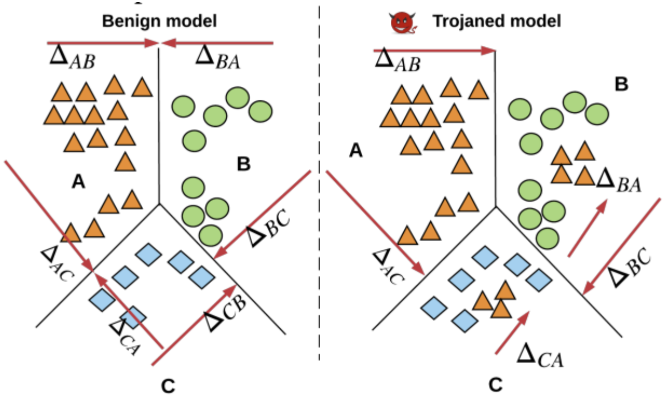
B. Tran, J. Li, and A. Madry, “Spectral signatures in backdoor attacks,” in Advances in Neural Information Processing Systems, 2018, pp. 8000–8010.

Existing Defenses against Backdoor Attacks (Centralized)

DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks

- Observe the dilemma of Intermediate layers' representation between normal and backdoored models
- Exclude malicious models which contain sensitive neurons to backdoor triggers



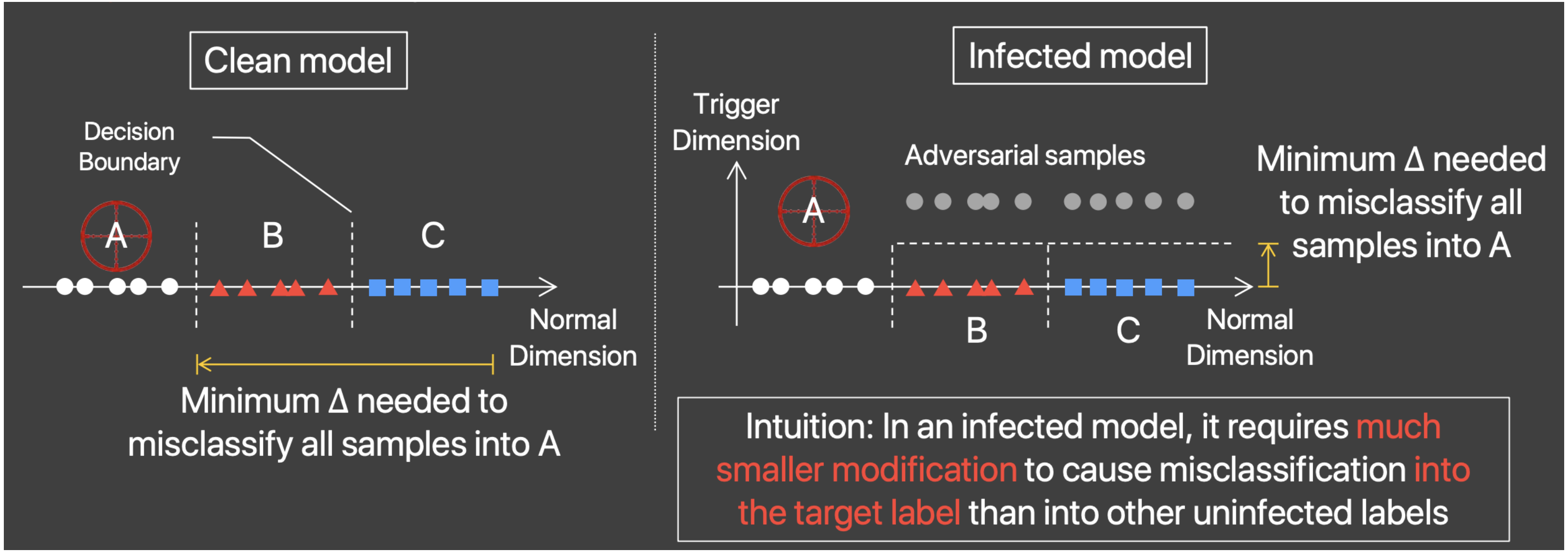


H. Chen, C. Fu, J. Zhao, and F. Koushanfar, "Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks." in IJCAI, 2019, pp. 4658–4664.

Existing Defenses against Backdoor Attacks (Centralized)

Neural cleanse: Identifying and mitigating backdoor attacks in neural networks

- Identify the statistical observation of backdoored neurons
- Prune the neurons which highly sensitive to the backdoor triggers



Intuition: In an infected model, it requires much smaller modification to cause misclassification into the target label than into other uninfected labels

B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks,” in 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019, pp. 707–723.

Existing Defenses against Backdoor Attacks (Centralized)

Features of existing backdoor defenses in centralized learning settings

- Need to access to sensitive dataset to achieve the defense goal
- Have comparatively heavy computation overhead

Category	Description	Literature	Local Access*	Computation Overhead
Input filtering	Pick out backdoored inputs from all inputs	[Tran et al. NIPS'18]	Yes	Moderate
Model inspection	Exclude malicious local models which contain the sensitive neurons to the backdoor triggers	[Chen et al. AAAI'19] [Guo et al. ICDM'19] [Huang et al. AAAI'19] [Liu et al. CCS'19]	Yes	Heavy
Model sanitization	Prune the neurons which highly sensitive to the backdoor triggers	[Liu et al. RAID'18] [Wang et al. S&P'19]	Yes	Heavy

*__Local Access__ states whether or not the defense needs to access local private data to achieve the defense goal.

Existing Defenses against Backdoor Attacks (Distributed)

Features of existing backdoor defenses in federated learning settings

- cannot access to sensitive dataset to achieve the defense goal
- often have a restricted assumption over the ratio of attackers among all clients, e.g., less than 50%.

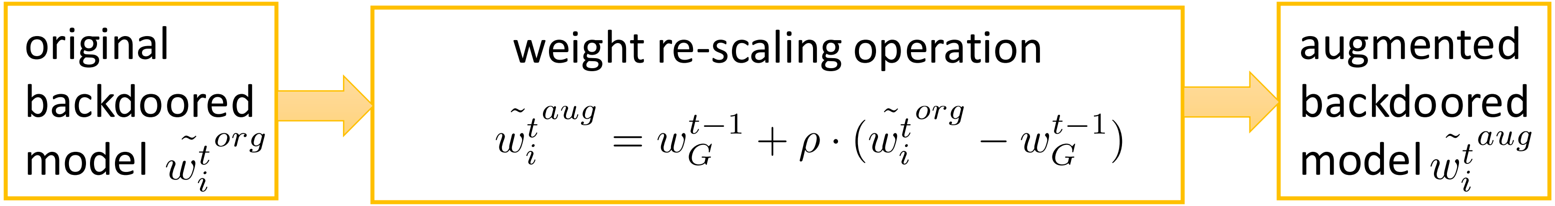
Category	Description	Literature	Local Access*	Computation Overhead
Model quantization	Quantize the local model updates before aggregation	[Bernstein et al. ICLR’19] [Ozdayi et al. AAAI’21]	No	Lite
Robust aggregation	Design robust aggregation metrics to remove negative impacts from malicious updates	[Fung et al.*USENIX’20*] [Pillutla et al. *arXiv’19*] [Sun et al. *arXiv’19]*	No	Lite

*Local Access states whether or not the defense needs to access local private data to achieve the defense goal.

Observation: Large Magnitude of Attackers' Local Updates

- Weight re-scaling operation
  - ρ : re-scaling factor

The minority of the malicious party determines the necessity of weight re-scaling operation

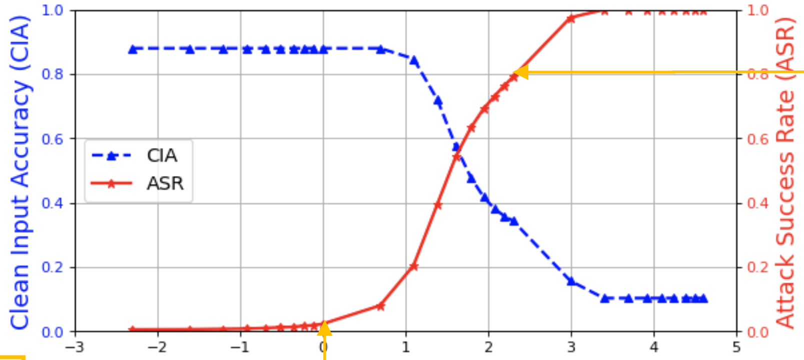


original backdoored model $\tilde{w}_i^{t\,org}$

weight re-scaling operation

$$\tilde{w}_i^{t\,aug} = w_G^{t-1} + \rho \cdot (\tilde{w}_i^{t\,org} - w_G^{t-1})$$

augmented backdoored model $\tilde{w}_i^{t\,aug}$



$ln(\rho)$ (when the malicious client's ratio is 0.2)

when $\rho = 10$, ASR > 0.8

when $\rho = 1$, ASR < 0.05

Norm Clipping Defense

- The norm clipping defense scheme [Sun et al. arXiv’19]: clipping local updates to ensure whose l_2 norm is upper bounded by a threshold, i.e., M, as the following,

$$\boxed{w_G^t = w_G^{t-1} + \eta \cdot \sum_{i \in S^t} \frac{\Delta w_i^t}{\max\{1, \ ||\Delta w_i^t||_2 / M\}}}$$

- Although the norm clipping defense is designed to resist centralized backdoor attacks, it still helps in resisting distributed backdoor attacks.

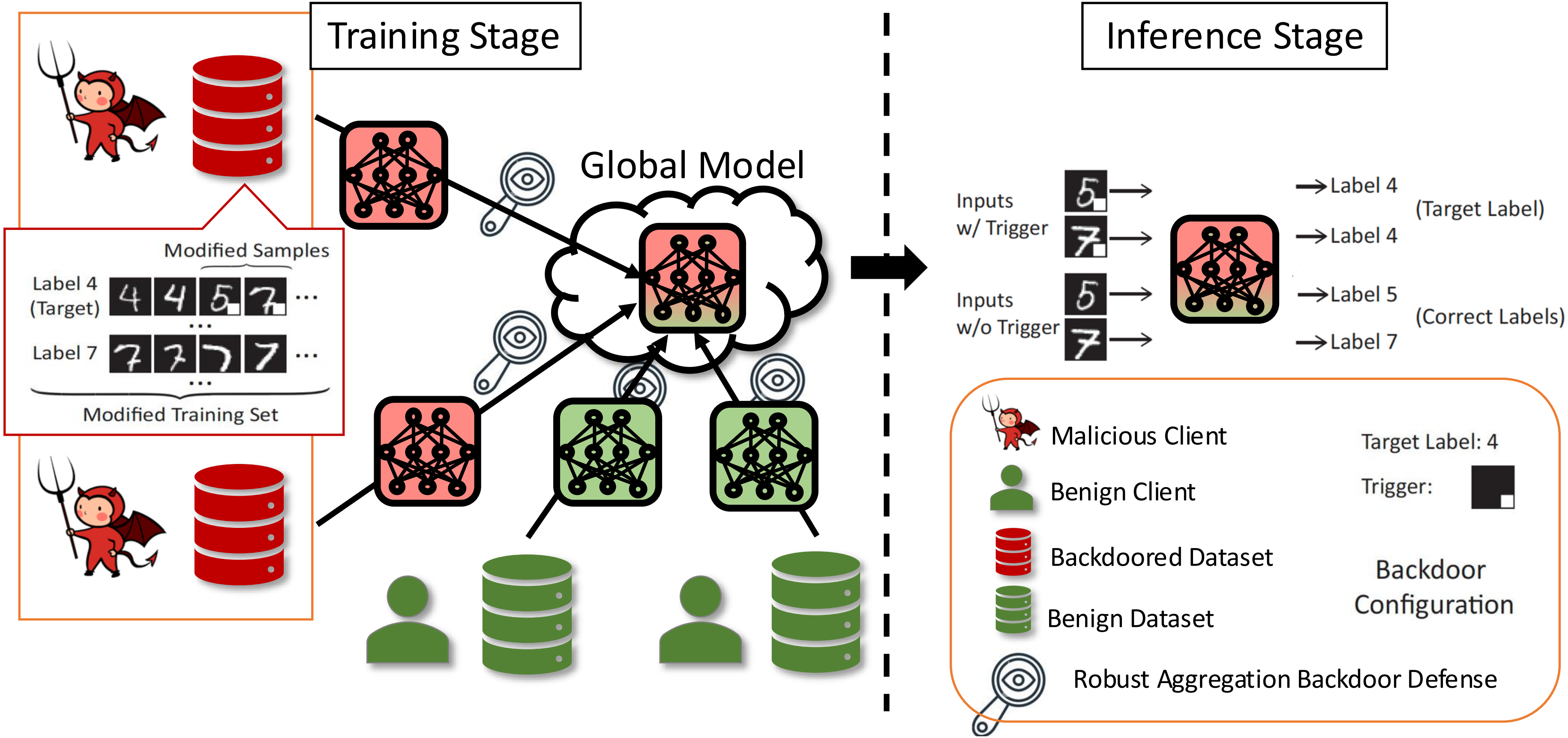
- So, the determination of the range of the norm threshold is important to the defense’s success.

Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, “Can you really backdoor federated learning?” arXiv preprint arXiv:1911.07963, 2019.

Outline

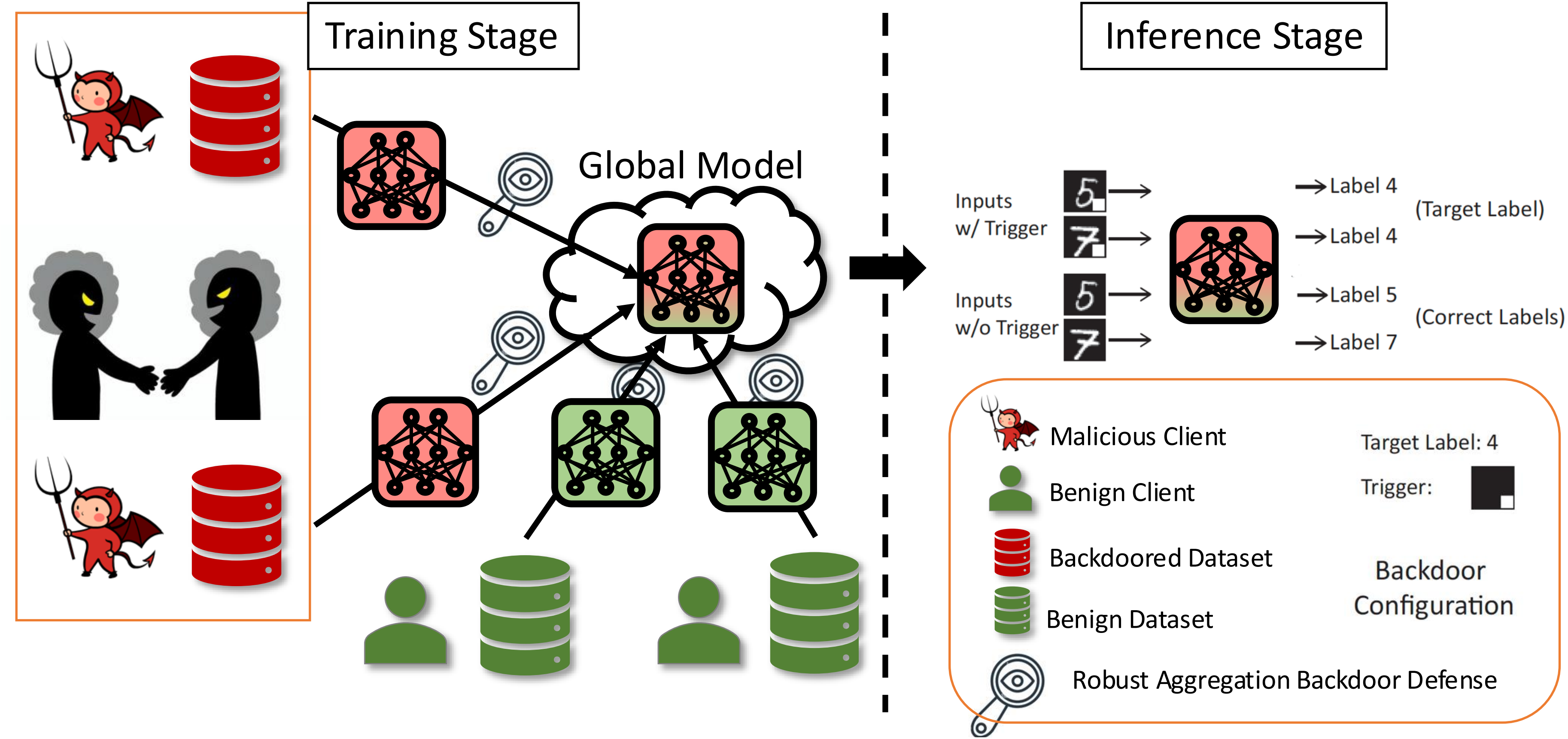
- Introduction of Federated Learning (FL)
- Backdoor Attacks and Defenses in FL
- **The New Threat: Collusive Backdoor Attacks in FL**
- Future Research Directions toward Backdoor Attack Resilient FL

Collusion Among Backdoor Attackers



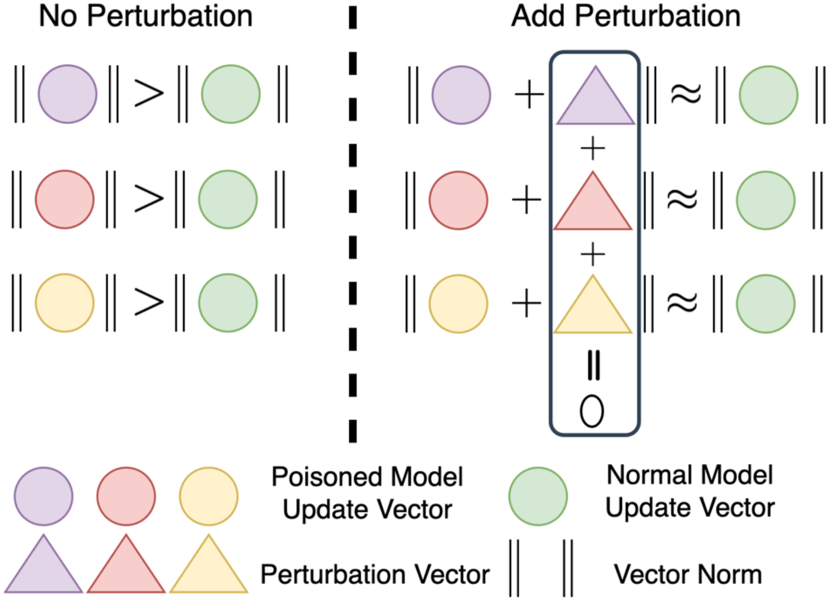
Saier Alharbi, Yifan Guo, and Wei Yu. "Collusive Backdoor Attacks in Federated Learning Frameworks for IoT Systems." to appear in *IEEE Internet of Things Journal* (2024).

Collusion Among Backdoor Attackers



Saier Alharbi, Yifan Guo, and Wei Yu. "Collusive Backdoor Attacks in Federated Learning Frameworks for IoT Systems." to appear in *IEEE Internet of Things Journal* (2024).

Attack Intuition of Collusive Backdoor Attacks



- 1) For each collusive adversary, the l_2 norm of perturbed model update vector is at the same level as that of normal model update vector, thereby being able to bypass robust aggregation defenses.
- 2) The summation of all perturbation vectors assigned to collusive attackers would be close to the zero vector.

Formulation of Collusive Backdoor Attacks

- The malicious goals:
  - each individual attacker is disguised to bypass the defender’s measurement
  - the sum of perturbations should be equal to 0.

The objective function of attacker $\boldsymbol{i}$:

$$\begin{aligned} &\text{Find } \overline{P_i^t} \ (\forall i \in N_A \cap S^t) \\ &\text{s.t. } \|\overline{\Delta W_i^t} + \overline{P_i^t}\|_2 \leq M, \forall i \in N_A \cap S^t; \\ &\sum_{i \in N_A \cap S^t} \overline{P_i^t} = \mathbf{0}. \end{aligned}$$

Notations	Descriptions
N_A, N_B, S^t	N_A: the attackers’ group; N_B: the benigner’ group; $N_A \cap N_B = \emptyset$, $N_A \cup N_B = \{1,2,\dots,n\}$; S^t: Set of selected clients in the t-th global round
$\overline{\Delta W_i^t}$	$\overline{W_i^t}$: scaled and backdoored model to be submitted to the server; $\overline{\Delta W_i^t} = \overline{W_i^t} - W_G^{t-1}$
$\overline{P_i^t}$	The coordinate-wise perturbation vector to be added on $\overline{\Delta W_i^t}$

Formulation of Collusive Backdoor Attacks

- Objective Function Approximation:
  - Denoting $\mathbf{A} = [(\overline{\Delta W_1^t})^T; (\overline{\Delta W_2^t})^T; \dots; (\overline{\Delta W_{s_t}^t})^T]$; $\mathbf{A}' = [(\overline{P_1^t})^T; (\overline{P_2^t})^T; \dots; (\overline{P_{s_t}^t})^T]$

$s_t = |N_A \cap S^t|$

$$\begin{aligned} &\text{Find } \overline{P_i^t} \ (\forall i \in N_A \cap S^t) \\ &\text{s.t. } \|\overline{\Delta W_i^t} + \overline{P_i^t}\|_2 \le M, \forall i \in N_A \cap S^t; \\ &\sum_{i \in N_A \cap S^t} \overline{P_i^t} = \mathbf{0}. \end{aligned}$$

Approximation

$$\min_{\mathbf{A}'} J(\mathbf{A}'; \mathbf{A}, \lambda) = \|\log(1 + \exp(\mathbf{A} + \mathbf{A}'))\|_F^2 + \lambda \|\mathbf{A}'^{\mathsf{T}} \mathbf{1}\|_2^2$$

  - Advantages of Approximation:
    - A smooth and convex objective function (easy implemented).
    - The optimal solutions between the approximated and original one keep the same.

Verification Our Idea: A Pilot Study

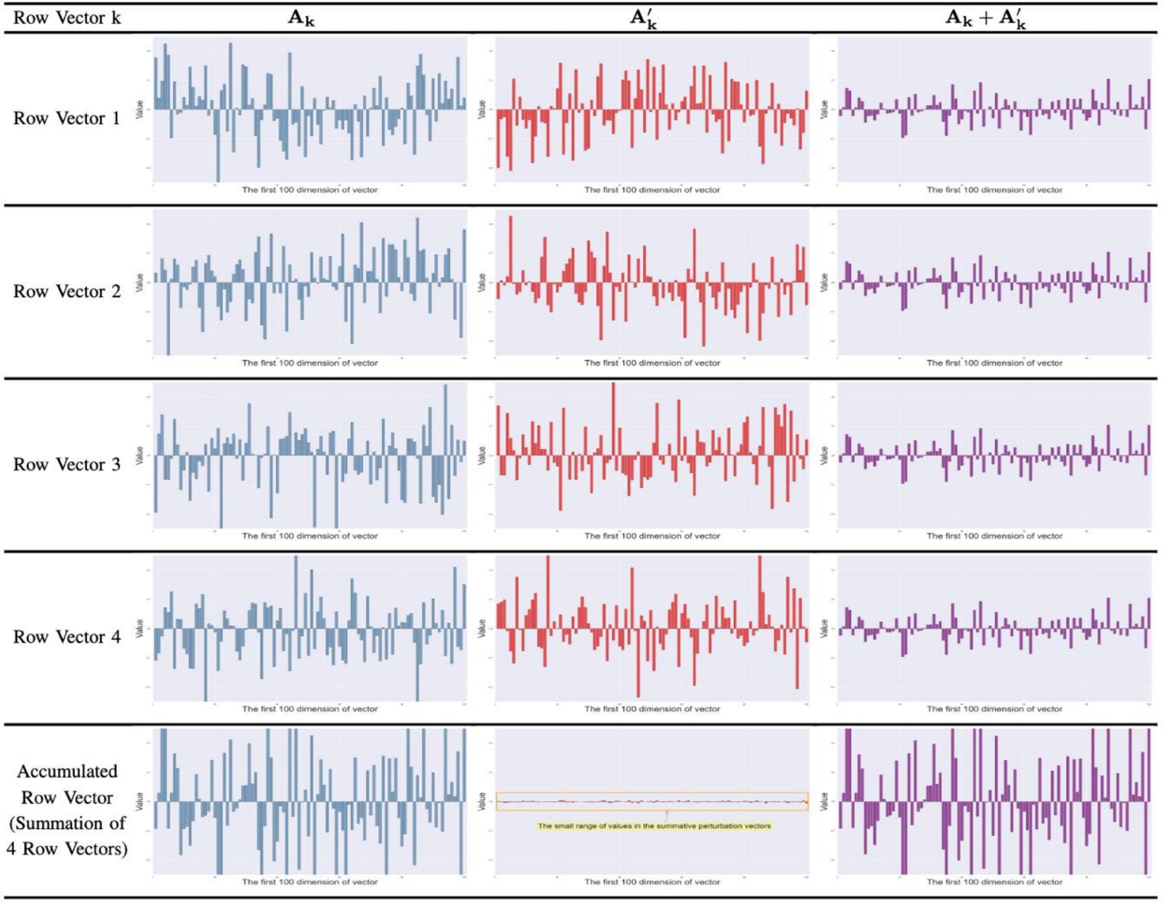


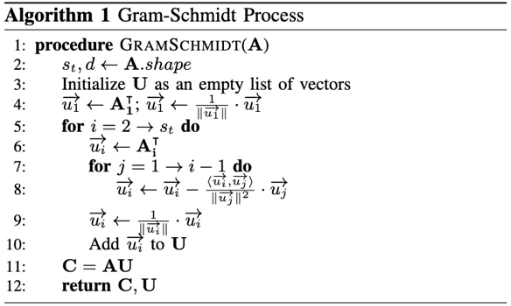
TABLE III: Numerical Results of Key Metrics Verification (M = 7)

Row Vector k	$\lVert\mathbf{A_k}\rVert_2$	Less than M?	$\lVert\mathbf{A_k} + \mathbf{A'_k}\rVert_2$	Less than M?
Row Vector 1	14.0067	×	6.9723	✓
Row Vector 2	14.1845	×	6.9723	✓
Row Vector 3	14.1111	×	6.9719	✓
Row Vector 4	14.0554	×	6.9725	✓
The Mean Absolute Value of Accumulated Perturbation Vector		$\frac{1}{d}\sum_{j=1}^{d}\left[\left\lvert\sum_{i=1}^{S_t} a'_{ij}\right\rvert\right] = 0.0019$		

Saier Alharbi, Yifan Guo, and Wei Yu. "Collusive Backdoor Attacks in Federated Learning Frameworks for IoT Systems." to appear in *IEEE Internet of Things Journal* (2024).

Speedup the Perturbation Estimations

- To estimate $\mathbf{A}'$, our problem space is in $s_t \times d$ dimensional space, which is quite huge and brings high computation cost.
  - Typically, s_t (number of participated malicious clients) < 100, and d (the number of the benchmark models' parameters) > several millions.
- To speedup the estimation, we have involved the Gram-Schmidt process.



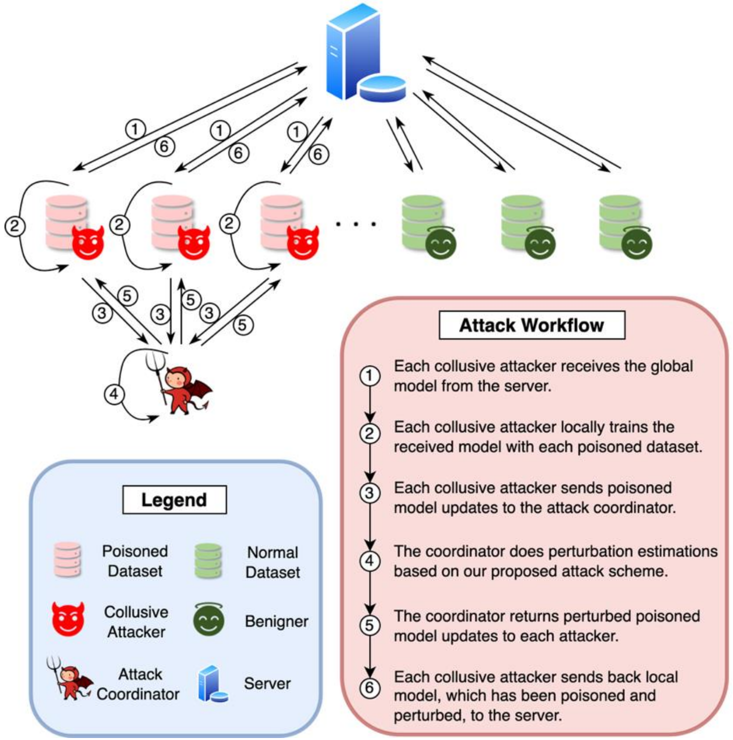
Algorithm 1 Gram-Schmidt Process

```
procedure GramSchmidt($\mathbf{A}$)
    $s_t, d \leftarrow \mathbf{A}.shape$
    Initialize $\mathbf{U}$ as an empty list of vectors
    $\overrightarrow{u_1} \leftarrow \mathbf{A}_1^\intercal$; $\overrightarrow{u_1} \leftarrow \frac{1}{\|\overrightarrow{u_1}\|} \cdot \overrightarrow{u_1}$
    for $i = 2 \rightarrow s_t$ do
        $\overrightarrow{u_i} \leftarrow \mathbf{A}_i^\intercal$
        for $j = 1 \rightarrow i - 1$ do
            $\overrightarrow{u_i} \leftarrow \overrightarrow{u_i} - \frac{\langle \overrightarrow{u_i}, \overrightarrow{u_j} \rangle}{\|\overrightarrow{u_j}\|^2} \cdot \overrightarrow{u_j}$
        $\overrightarrow{u_i} \leftarrow \frac{1}{\|\overrightarrow{u_i}\|} \cdot \overrightarrow{u_i}$
        Add $\overrightarrow{u_i}$ to $\mathbf{U}$
    $\mathbf{C} = \mathbf{AU}$
    return $\mathbf{C}, \mathbf{U}$
```

$$\begin{cases} \mathbf{A'}_\mathbf{1}^\intercal & = c'_{11} \cdot \overrightarrow{u_1} + c'_{12} \cdot \overrightarrow{u_2} + \cdots + c'_{1s_t} \cdot \overrightarrow{u_{s_t}} \\ \mathbf{A'}_\mathbf{2}^\intercal & = c'_{21} \cdot \overrightarrow{u_1} + c'_{22} \cdot \overrightarrow{u_2} + \cdots + c'_{2s_t} \cdot \overrightarrow{u_{s_t}} \\ & \vdots \\ \mathbf{A'}_{\mathbf{s_t}}^\intercal & = c'_{s_t1} \cdot \overrightarrow{u_1} + c'_{s_t2} \cdot \overrightarrow{u_2} + \cdots + c'_{s_ts_t} \cdot \overrightarrow{u_{s_t}} \end{cases}$$

- We could obtain an estimation of C′ by feeding C and λ into the approximated objective function.
- But for the estimation of C′, its problem space is just $s_t \times s_t$, which is far smaller than $s_t \times d$.

Our Collusive Backdoor Attack



Legend

	Poisoned Dataset		Normal Dataset
	Collusive Attacker		Benigner
	Attack Coordinator		Server

Attack Workflow

① Each collusive attacker receives the global model from the server.

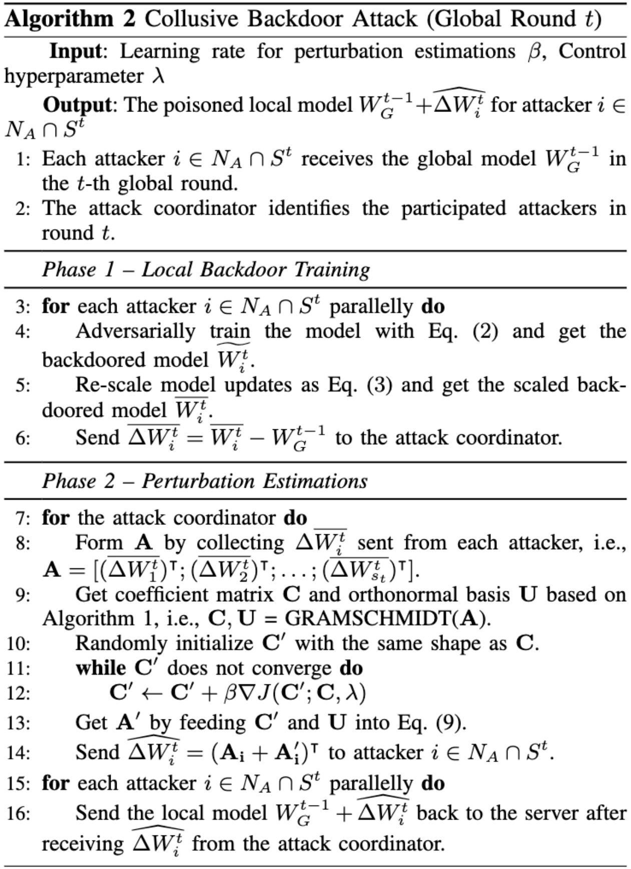
② Each collusive attacker locally trains the received model with each poisoned dataset.

③ Each collusive attacker sends poisoned model updates to the attack coordinator.

④ The coordinator does perturbation estimations based on our proposed attack scheme.

⑤ The coordinator returns perturbed poisoned model updates to each attacker.

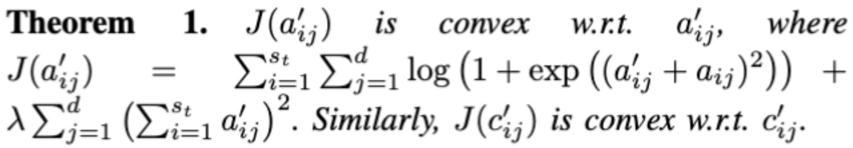
⑥ Each collusive attacker sends back local model, which has been poisoned and perturbed, to the server.



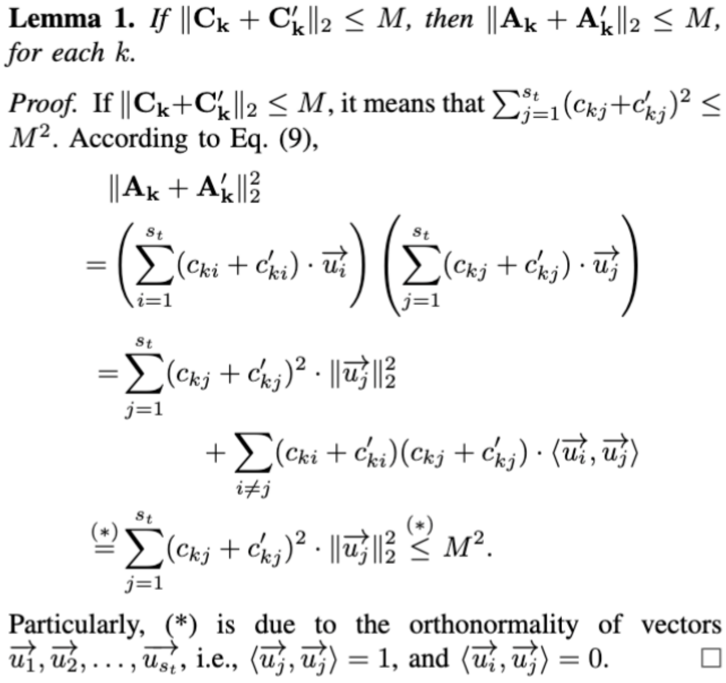
Algorithm 2 Collusive Backdoor Attack (Global Round t)

```
Input: Learning rate for perturbation estimations $\beta$, Control hyperparameter $\lambda$
Output: The poisoned local model $W_G^{t-1}+\widehat{\Delta W_i^t}$ for attacker $i \in N_A \cap S^t$
Each attacker $i \in N_A \cap S^t$ receives the global model $W_G^{t-1}$ in the $t$-th global round.
The attack coordinator identifies the participated attackers in round $t$.

Phase 1 – Local Backdoor Training

for each attacker $i \in N_A \cap S^t$ parallelly do
    Adversarially train the model with Eq. (2) and get the backdoored model $\widetilde{W_i^t}$.
    Re-scale model updates as Eq. (3) and get the scaled backdoored model $\overline{W_i^t}$.
    Send $\overline{\Delta W_i^t} = \overline{W_i^t} - W_G^{t-1}$ to the attack coordinator.

Phase 2 – Perturbation Estimations

for the attack coordinator do
    Form $\mathbf{A}$ by collecting $\overline{\Delta W_i^t}$ sent from each attacker, i.e., $\mathbf{A} = [(\overline{\Delta W_1^t})^\intercal; (\overline{\Delta W_2^t})^\intercal; \dots; (\overline{\Delta W_{s_t}^t})^\intercal]$.
    Get coefficient matrix $\mathbf{C}$ and orthonormal basis $\mathbf{U}$ based on Algorithm 1, i.e., $\mathbf{C}, \mathbf{U}$ = GRAMSCHMIDT($\mathbf{A}$).
    Randomly initialize $\mathbf{C}'$ with the same shape as $\mathbf{C}$.
    while $\mathbf{C}'$ does not converge do
        $\mathbf{C}' \leftarrow \mathbf{C}' + \beta \nabla J(\mathbf{C}'; \mathbf{C}, \lambda)$
    Get $\mathbf{A}'$ by feeding $\mathbf{C}'$ and $\mathbf{U}$ into Eq. (9).
    Send $\widehat{\Delta W_i^t} = (\mathbf{A_i} + \mathbf{A_i'})^\intercal$ to attacker $i \in N_A \cap S^t$.
for each attacker $i \in N_A \cap S^t$ parallelly do
    Send the local model $W_G^{t-1} + \widehat{\Delta W_i^t}$ back to the server after receiving $\widehat{\Delta W_i^t}$ from the attack coordinator.
```

Convergence Analysis of Perturbation Estimations



Theorem 1. *$J(a'_{ij})$ is convex w.r.t. a'_{ij}, where $J(a'_{ij}) = \sum_{i=1}^{s_t}\sum_{j=1}^{d}\log\left(1+\exp\left((a'_{ij}+a_{ij})^2\right)\right) + \lambda\sum_{j=1}^{d}\left(\sum_{i=1}^{s_t} a'_{ij}\right)^2$. Similarly, $J(c'_{ij})$ is convex w.r.t. c'_{ij}.*

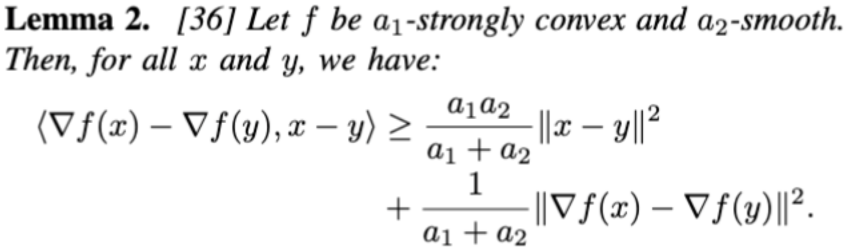


Lemma 1. *If $\|\mathbf{C_k}+\mathbf{C'_k}\|_2 \le M$, then $\|\mathbf{A_k}+\mathbf{A'_k}\|_2 \le M$, for each k.*

Proof. If $\|\mathbf{C_k}+\mathbf{C'_k}\|_2 \le M$, it means that $\sum_{j=1}^{s_t}(c_{kj}+c'_{kj})^2 \le M^2$. According to Eq. (9),

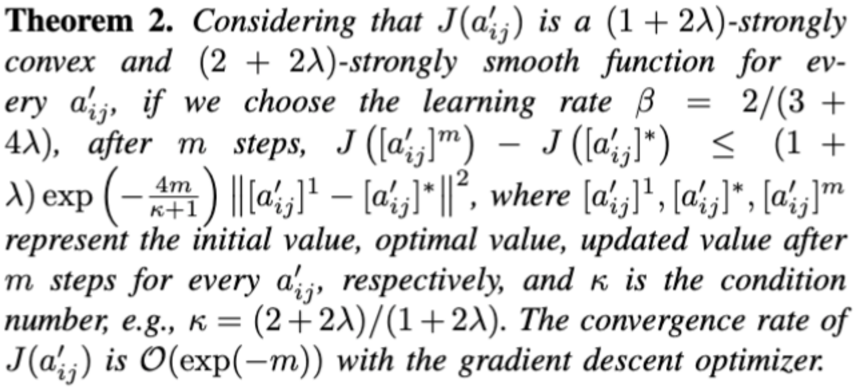
$$\begin{aligned}
&\|\mathbf{A_k}+\mathbf{A'_k}\|_2^2 \\
&= \left(\sum_{i=1}^{s_t}(c_{ki}+c'_{ki})\cdot\vec{u_i}\right)\left(\sum_{j=1}^{s_t}(c_{kj}+c'_{kj})\cdot\vec{u_j}\right) \\
&= \sum_{j=1}^{s_t}(c_{kj}+c'_{kj})^2\cdot\|\vec{u_j}\|_2^2 \\
&\qquad + \sum_{i\neq j}(c_{ki}+c'_{ki})(c_{kj}+c'_{kj})\cdot\langle\vec{u_i},\vec{u_j}\rangle \\
&\overset{(*)}{=} \sum_{j=1}^{s_t}(c_{kj}+c'_{kj})^2\cdot\|\vec{u_j}\|_2^2 \overset{(*)}{\le} M^2.
\end{aligned}$$

Particularly, (*) is due to the orthonormality of vectors $\vec{u_1},\vec{u_2},\ldots,\vec{u_{s_t}}$, i.e., $\langle\vec{u_j},\vec{u_j}\rangle = 1$, and $\langle\vec{u_i},\vec{u_j}\rangle = 0$. □



Lemma 2. *[36] Let f be a_1-strongly convex and a_2-smooth. Then, for all x and y, we have:*

$$\begin{aligned}
\langle\nabla f(x)-\nabla f(y), x-y\rangle \ge \frac{a_1a_2}{a_1+a_2}\|x-y\|^2 \\
+\frac{1}{a_1+a_2}\|\nabla f(x)-\nabla f(y)\|^2.
\end{aligned}$$



Theorem 2. *Considering that $J(a'_{ij})$ is a $(1+2\lambda)$-strongly convex and $(2+2\lambda)$-strongly smooth function for every a'_{ij}, if we choose the learning rate $\beta = 2/(3+4\lambda)$, after m steps, $J\left([a'_{ij}]^m\right) - J\left([a'_{ij}]^*\right) \le (1+\lambda)\exp\left(-\frac{4m}{\kappa+1}\right)\left\|[a'_{ij}]^1-[a'_{ij}]^*\right\|^2$, where $[a'_{ij}]^1, [a'_{ij}]^*, [a'_{ij}]^m$ represent the initial value, optimal value, updated value after m steps for every a'_{ij}, respectively, and κ is the condition number, e.g., $\kappa = (2+2\lambda)/(1+2\lambda)$. The convergence rate of $J(a'_{ij})$ is $\mathcal{O}(\exp(-m))$ with the gradient descent optimizer.*

Attack Performance

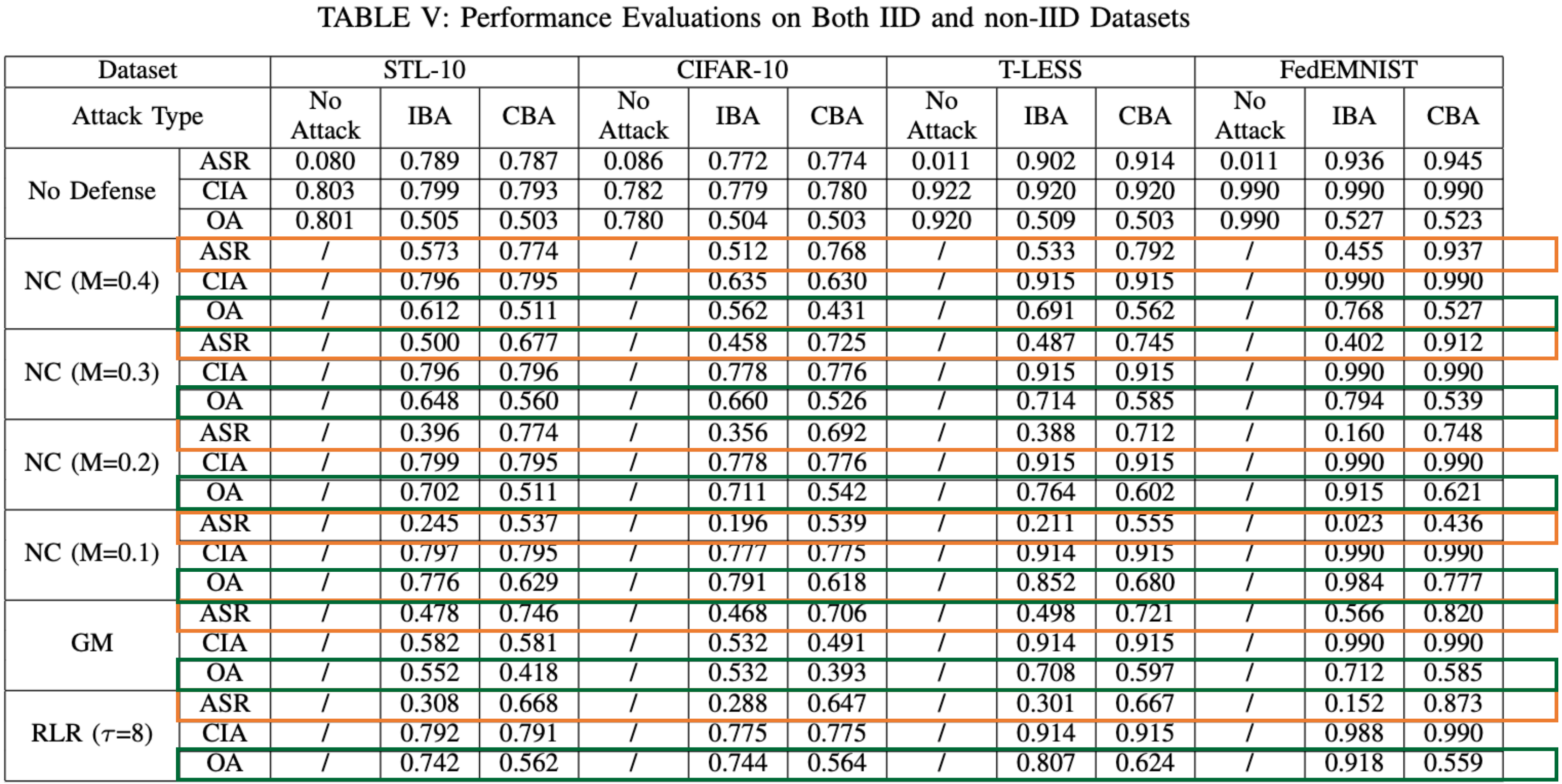


TABLE V: Performance Evaluations on Both IID and non-IID Datasets

Dataset		STL-10			CIFAR-10			T-LESS			FedEMNIST		
Attack Type		No Attack	IBA	CBA	No Attack	IBA	CBA	No Attack	IBA	CBA	No Attack	IBA	CBA
No Defense	ASR	0.080	0.789	0.787	0.086	0.772	0.774	0.011	0.902	0.914	0.011	0.936	0.945
	CIA	0.803	0.799	0.793	0.782	0.779	0.780	0.922	0.920	0.920	0.990	0.990	0.990
	OA	0.801	0.505	0.503	0.780	0.504	0.503	0.920	0.509	0.503	0.990	0.527	0.523
NC (M=0.4)	ASR	/	0.573	0.774	/	0.512	0.768	/	0.533	0.792	/	0.455	0.937
	CIA	/	0.796	0.795	/	0.635	0.630	/	0.915	0.915	/	0.990	0.990
	OA	/	0.612	0.511	/	0.562	0.431	/	0.691	0.562	/	0.768	0.527
NC (M=0.3)	ASR	/	0.500	0.677	/	0.458	0.725	/	0.487	0.745	/	0.402	0.912
	CIA	/	0.796	0.796	/	0.778	0.776	/	0.915	0.915	/	0.990	0.990
	OA	/	0.648	0.560	/	0.660	0.526	/	0.714	0.585	/	0.794	0.539
NC (M=0.2)	ASR	/	0.396	0.774	/	0.356	0.692	/	0.388	0.712	/	0.160	0.748
	CIA	/	0.799	0.795	/	0.778	0.776	/	0.915	0.915	/	0.990	0.990
	OA	/	0.702	0.511	/	0.711	0.542	/	0.764	0.602	/	0.915	0.621
NC (M=0.1)	ASR	/	0.245	0.537	/	0.196	0.539	/	0.211	0.555	/	0.023	0.436
	CIA	/	0.797	0.795	/	0.777	0.775	/	0.914	0.915	/	0.990	0.990
	OA	/	0.776	0.629	/	0.791	0.618	/	0.852	0.680	/	0.984	0.777
GM	ASR	/	0.478	0.746	/	0.468	0.706	/	0.498	0.721	/	0.566	0.820
	CIA	/	0.582	0.581	/	0.532	0.491	/	0.914	0.915	/	0.990	0.990
	OA	/	0.552	0.418	/	0.532	0.393	/	0.708	0.597	/	0.712	0.585
RLR (τ=8)	ASR	/	0.308	0.668	/	0.288	0.647	/	0.301	0.667	/	0.152	0.873
	CIA	/	0.792	0.791	/	0.775	0.775	/	0.914	0.915	/	0.988	0.990
	OA	/	0.742	0.562	/	0.744	0.564	/	0.807	0.624	/	0.918	0.559

Verification of Negligible Computation Overhead

- Is the proposed perturbation estimation scheme highly time consuming?
- NO!
- Running time cost:
  - One epoch's local training will take 5.6 seconds on the overage on CIFAR-10 dataset.
  - The running time for the perturbation estimation functions only takes 0.1 (<< 5.6) seconds.
  - However, if no Gram-Schmidt process is involved, the time cost for perturbation estimations would be increased to 4.8 seconds.

A Quick Summary

- A New Threat:
  - Existing robust aggregation based defenses, handle each returned model individually, to detect backdoored models and/or mitigate the negative effects of returned backdoored models.
  - The distributed nature in FL opens a door for attackers to launch attacks collusively, which sets up a higher bar for robust aggregation defenses.

- Correlations with Distributed Backdoor Attacks (DBA):
  - DBA only considers attack coordination by adjusting local image triggers in the local backdoor training [collusion in data space];
  - Our CBA considers both local backdoor training and collaborative post-training model manipulations [collusion in both data and model space];
  - DBA could be treated as a special case of CBA.

Saier Alharbi, Yifan Guo, and Wei Yu. "Collusive Backdoor Attacks in Federated Learning Frameworks for IoT Systems." to appear in *IEEE Internet of Things Journal* (2024).

Outline

- Introduction of Federated Learning (FL)
- Backdoor Attacks and Defenses in FL
- The New Threat: Collusive Backdoor Attacks in FL
- **Future Research Directions toward Backdoor Attack Resilient FL**

Future Research Directions toward Backdoor Attack Resilient FL

- Countermeasures against Collusive Backdoor Attacks:
  - Similarity-Score based client selection approaches
  - More advanced robust aggregation protocols.
- Randomized Client Selection Scheme
  - Involving randomization and redundancy into the aggregation protocol
- Secured Communications in FL
  - Utilizing secure communication channels, such as encrypted connections and digital signatures

Thank you for your attention!

Q & A