THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Opening Minds • Shaping the Future • 啟迪思維 • 成就未來

IEEE INFOCOM

大连理工大学 计算机科学与技术学院
DALIAN UNIVERSITY OF TECHNOLOGY
SCHOOL OF COMPUTER SCIENCE AND TECHNOLOGY

Protect Privacy from Gradient Leakage Attack in Federated Learning

JUNXIAO WANG, SONG GUO, XIN XIE, HENG QI

PolyU Edge Intelligence Lab

DEPARTMENT OF COMPUTING
電子計算學系

Topics of This Talk

1. **Gradient Leakage Attack and its Threats**
   See what's the gradient leakage attack and how it performs

2. **Existing Defenses and their Limitations**
   Identify the challenges and how we can solve it

3. **Proposed Defense and its Features**
   Framework, design and experimental results

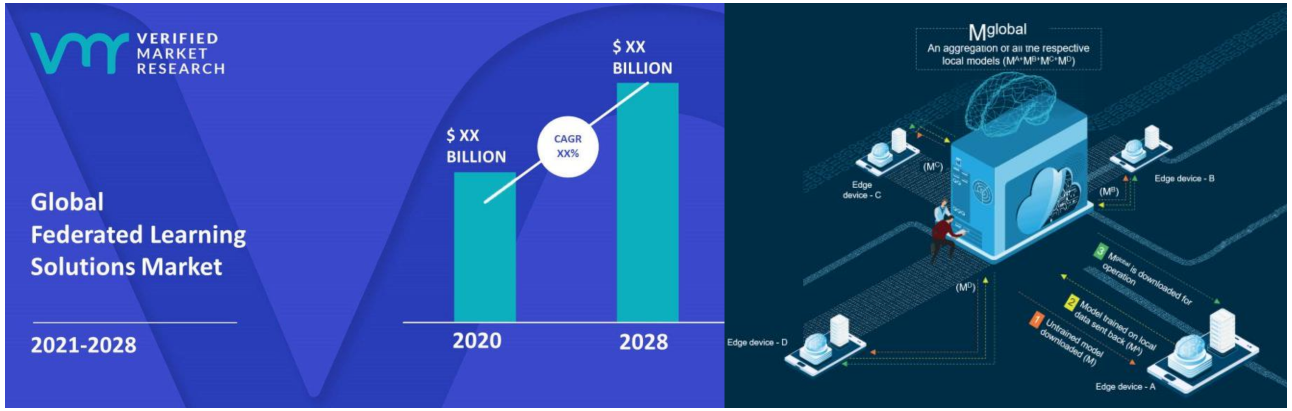
Part 1.

Gradient Leakage Attack and its Threats

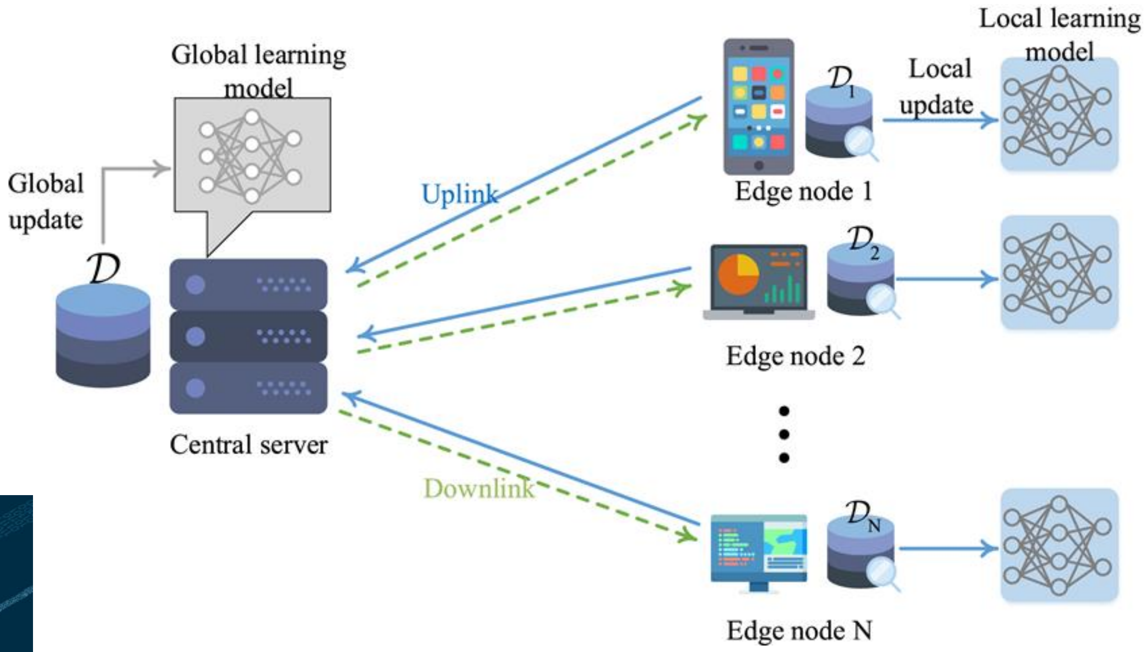
See what's the gradient leakage attack and how it performs

Introduction to Federated Learning

(a) TensorFlow Federated (TFF): **a framework for implementing Federated Learning**



(b) Market Statistics and Application of FL



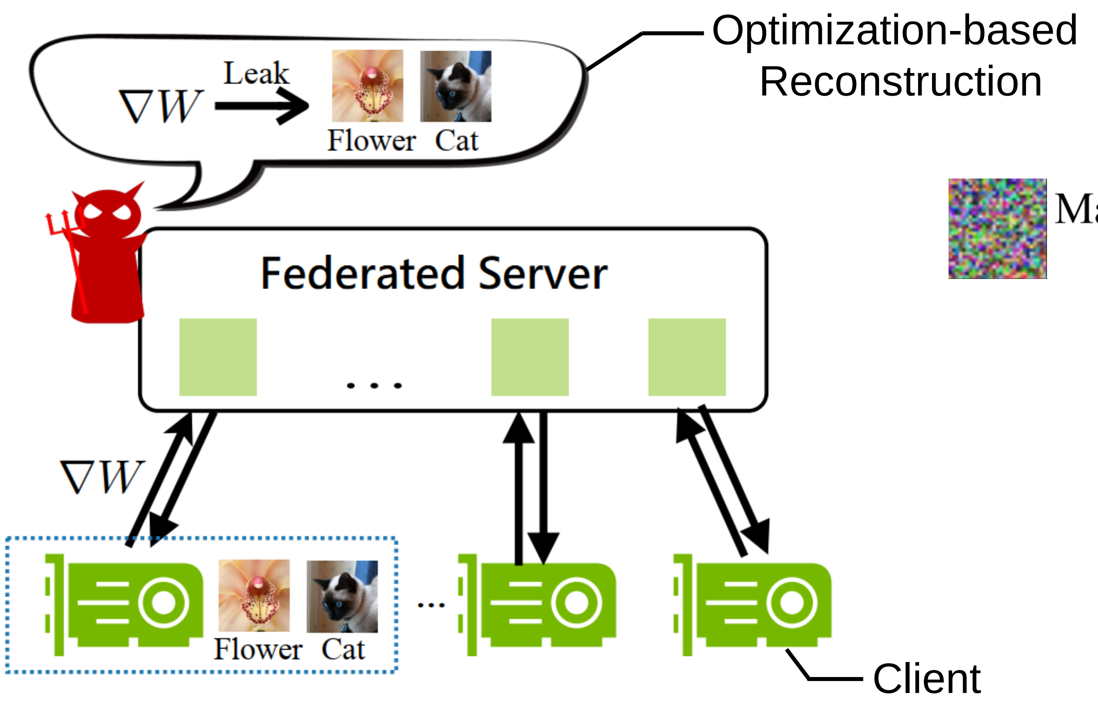
(c) FL workflow: How Federated Learning performs

[1]https://www.tensorflow.org/federated/
[2]https://www.everestgrp.com/
[3]https://www.verifiedmarketresearch.com/

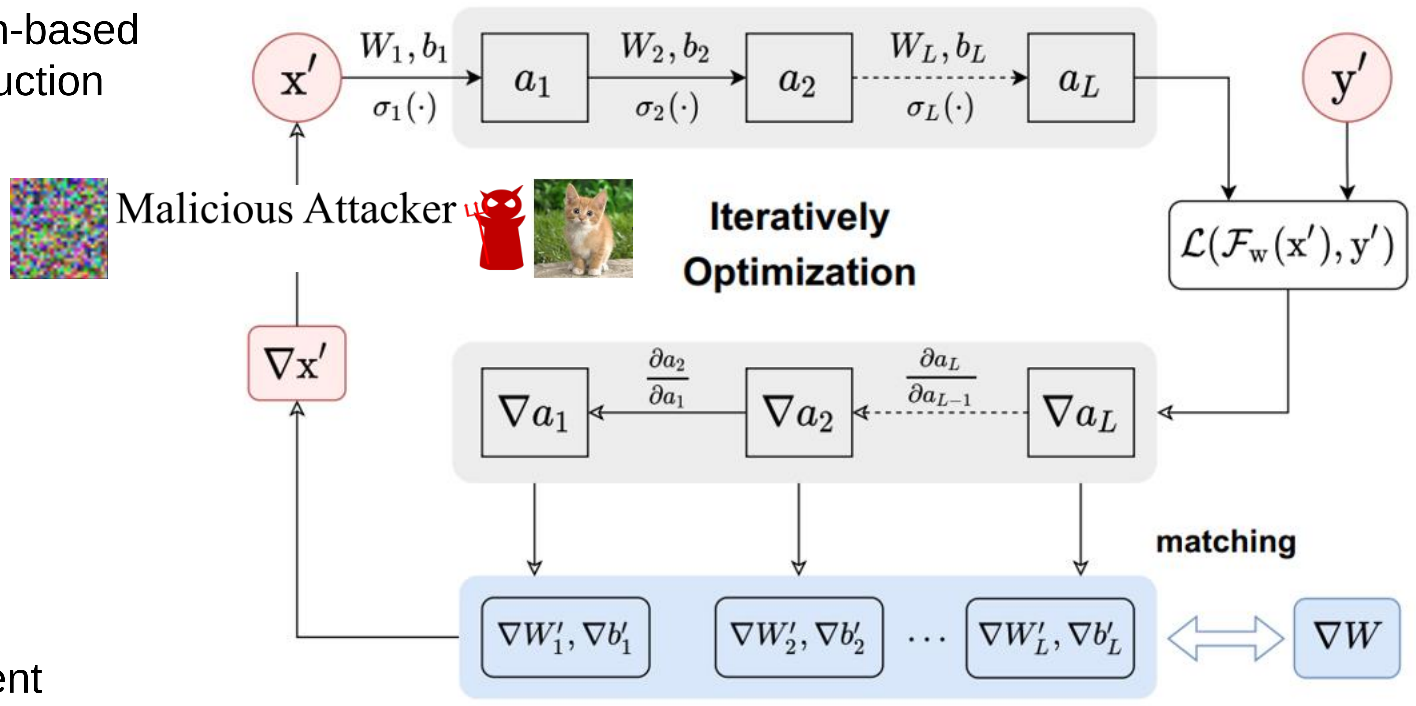
Gradient Leakage Attack: Deep Leakage from Gradients

MIT, NeurIPS 2019 [1]

- Background: An ***honest-but-curious* attacker**, who can be the **federated server**. The attacker can observe **gradients of a victim** and he attempts to **recover data from gradients**.



(a) Threat Model [1]



(b) Workflow of the Optimization-based Reconstruction

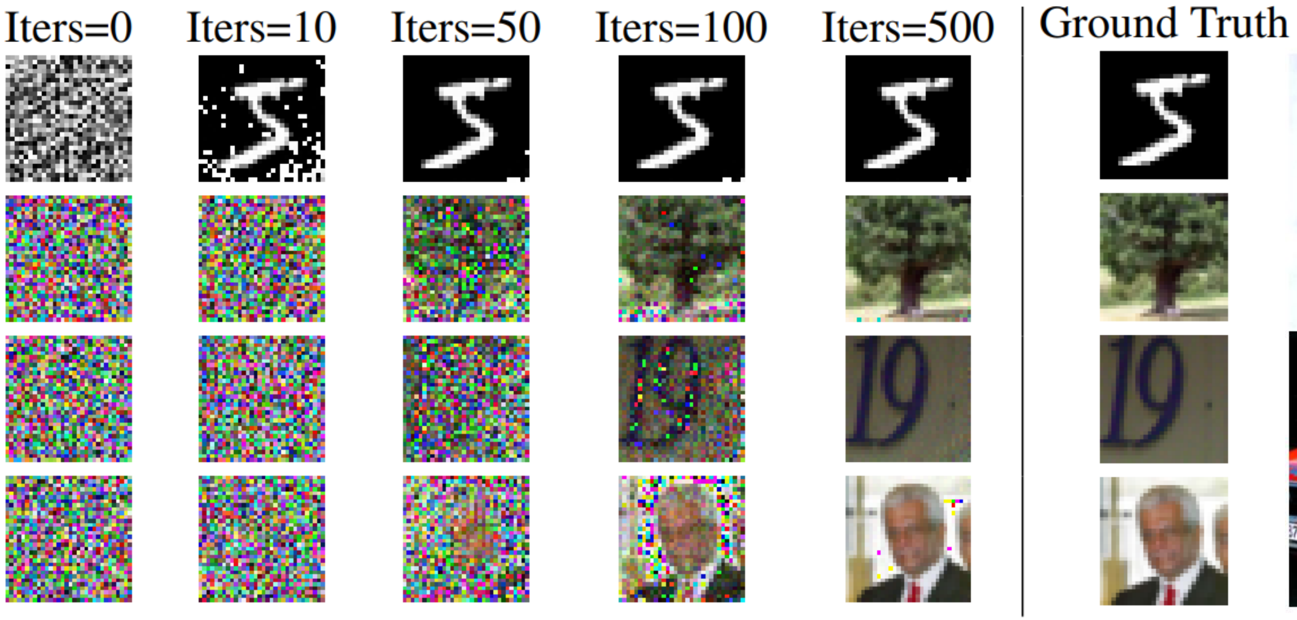
Gradient Leakage Attack pixel-wise level for images

Deep Leakage from Gradients

MIT, NeurIPS 2019 [1]

Inverting Gradients

Siegen, NeurIPS 2020 [2]



(a) Deep Leakage on Images from MNIST, CIFAR-100, SVHN and LFW [1]

(b) Additional Positive Cases for a Trained ResNet-18 on ImageNet [2]

Question: How to Protect Privacy from Gradients? Cryptographic Methods?

Part 2.

Existing Defenses and their Limitations

Identify the challenges and how we can solve it

Existing Defenses against Gradient Leakage pros and cons

- **General Privacy Protection Methods**

  \- Homomorphic Encryption (HE)

    \- Advantages: Gradient Aggregation is Performed on Ciphertexts.

  \- Multi-Party Computation (MPC)

    \- Advantages: Zero-Knowledge of Gradient Aggregation's Input/Output.

    - Limitations: High Computation and Communication Overhead

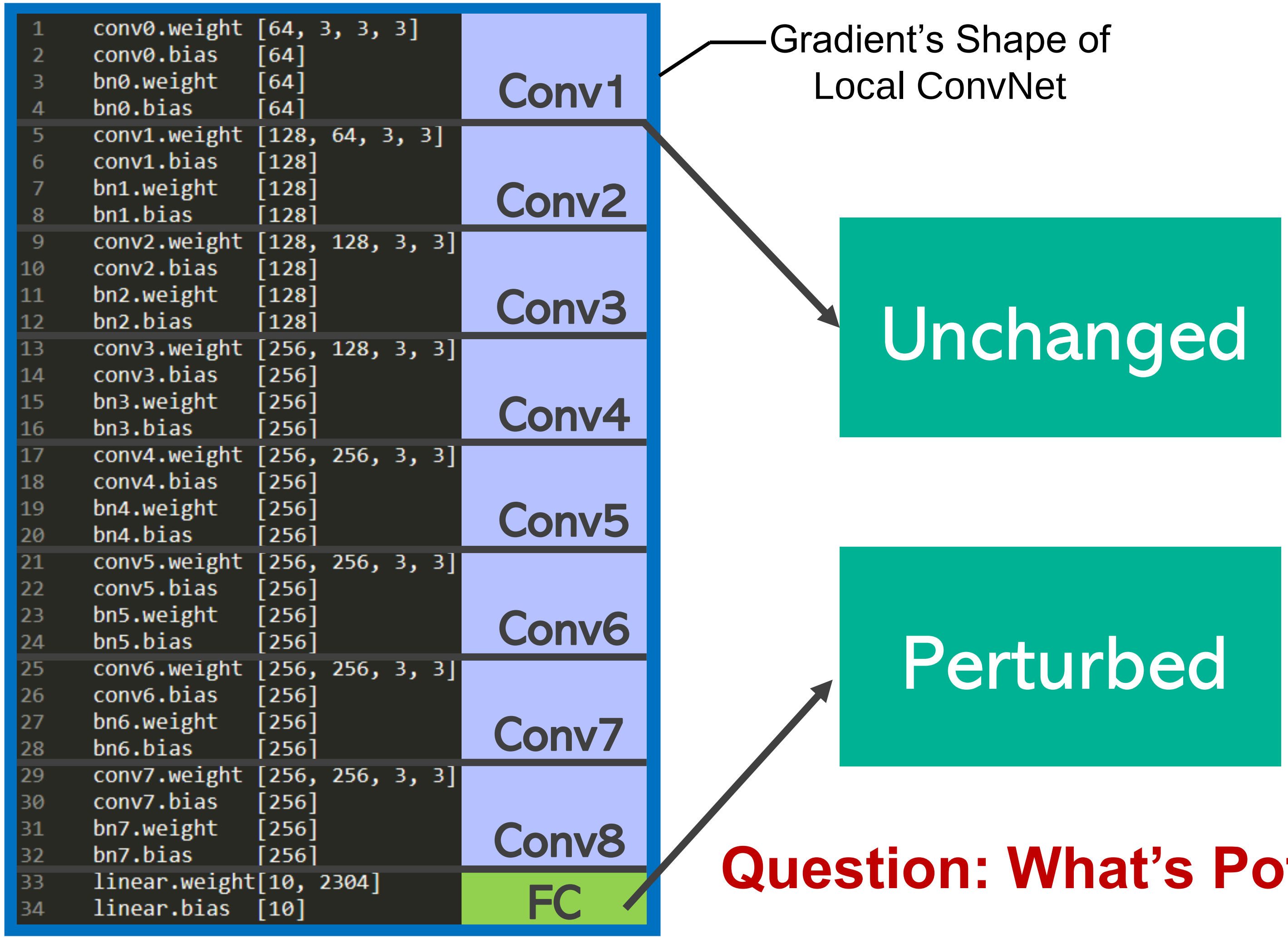
  \- Local Differential Privacy (LDP)

    \- Advantages: Identify Samples from Gradients within Theoretical Bound.

    - Limitations: High Convergence Accuracy Loss

Defense Specific to Gradient Leakage Attack

"Provable Defense against Privacy Leakage in Federated Learning", Duke, CVPR 2021



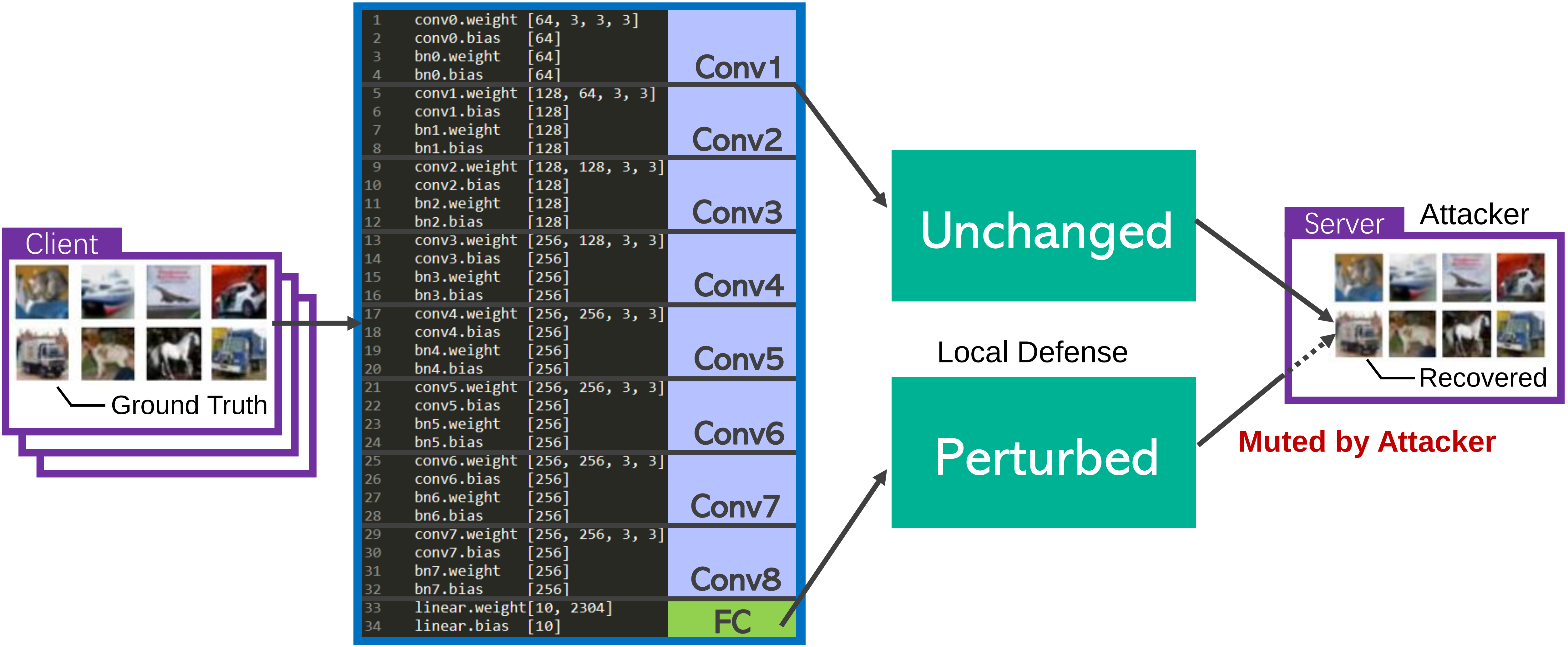
- Advantages: It only Perturbs a Certain Single Layer of Local Gradients (e.g., FC Layer).

In order to Lower Perturbation Footprints and Accuracy Loss.

Question: What's Potential Risk of this Rigid Pattern?

Defense Specific to Gradient Leakage Attack

- Limitations: Rigid Pattern is easily broken down once the Perturbed Layer is Muted by the Attacker.



Targets of Defense against Gradient Leakage

- **Lightweight, Accuracy-Guaranteed, Privacy-Adequate Defense**

  - Lightweight in Overhead (Computation, Storage, Communication)

    - **Cryptographic Methods e.g., HE**, **MPC** are with significant Overhead.

  - Guaranteed in Convergence Accuracy Loss

    - **Methods like LDP** are with significant Accuracy Loss.

  - Adequate in Privacy Protection and Hard to Break Down

    - **Methods with Rigid Pattern** are easily Inferred and Broken Down.

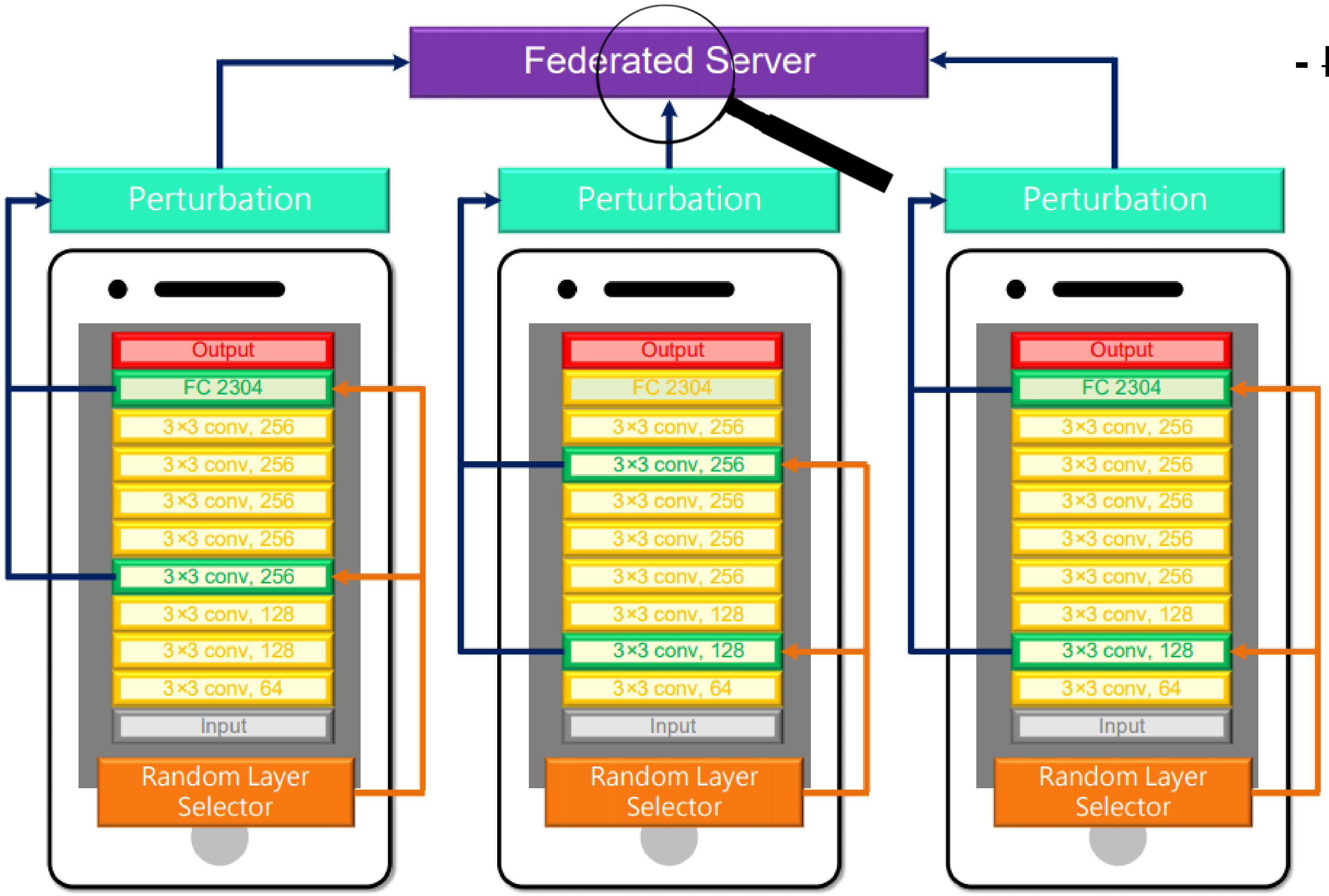
Part 3.

Proposed Defense and its Features

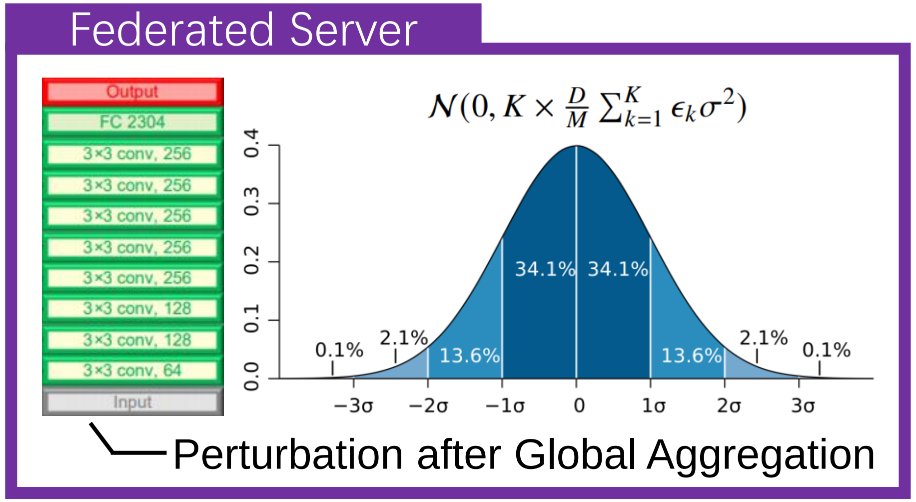
Framework, design and experimental results

Defense against Gradient Leakage basic idea

- Inspiration: Each Client Randomly Selects Part of Local Gradients to Perturb

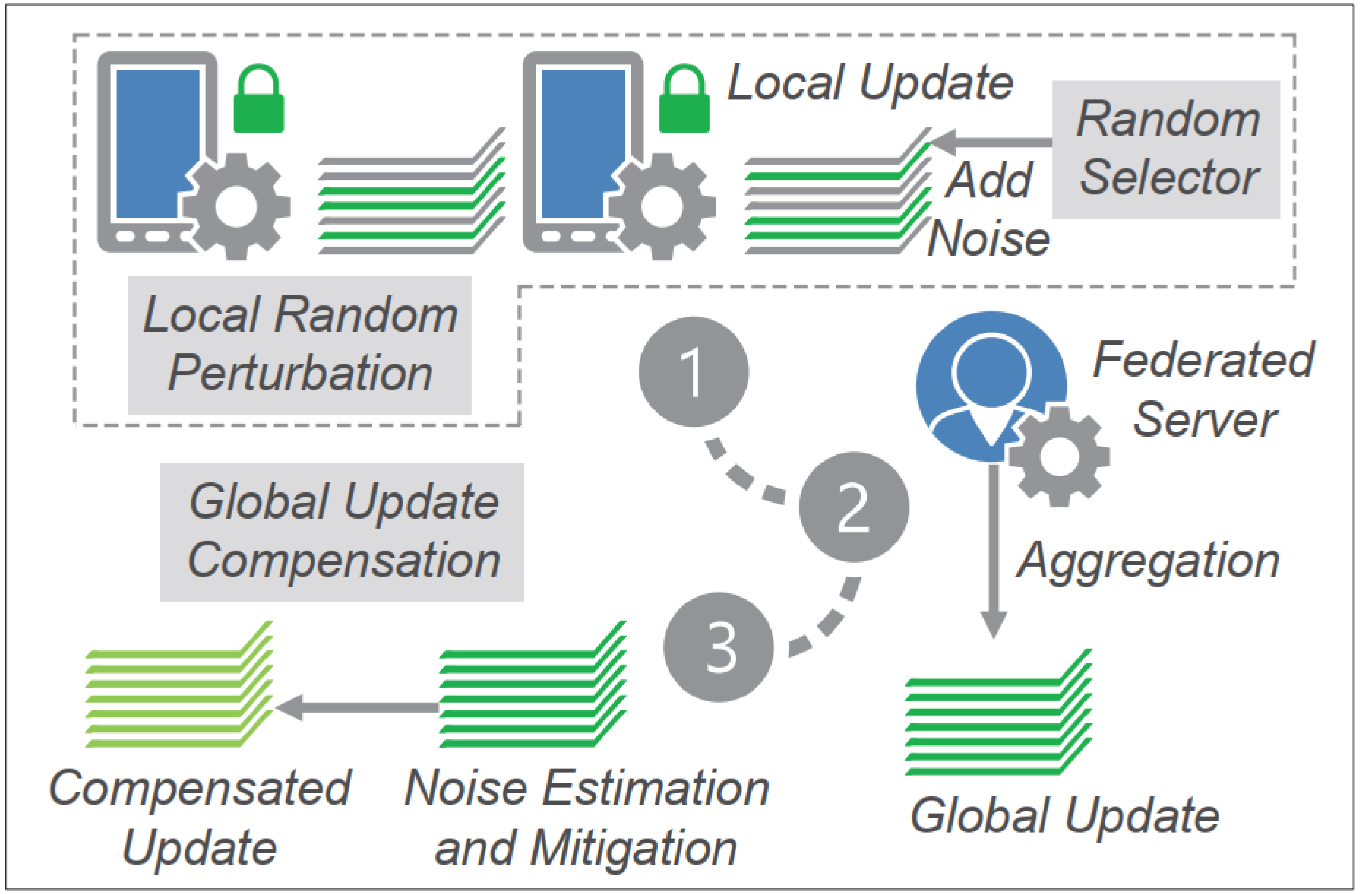


- ~~Rigid Pattern~~ **Random Pattern**
  - **Defense Becomes Hard to Break Down.**
  - **No Significant Overhead.**
  - **Perturbation Can be Compensated.**



Defense against Gradient Leakage workflow

- The workflow consists of two stages: Local Random Perturbation and Global Update Compensation.



- **Local Random Perturbation**

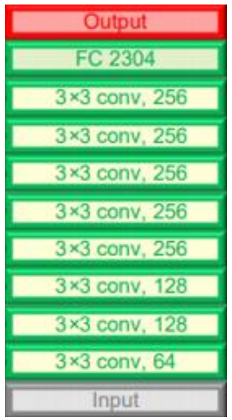
  - Randomly select a certain part of slices from local gradients and add artificial noise to these selected slices.

- **Global Update Compensation**

  - Derive from the perturbed gradients, more accurate information about the original gradients as a compensation for the global update.

Defense against Gradient Leakage more considerations

- Privacy Leakage Risk Evaluation and Gradient Slicing



- Cons: Different layers have different risks of privacy leakage.

(a) Random Perturbation is based on Gradient's Logical Layers e.g., Convolutional Layer (Conv) or Fully-Connected Layer (FC).

(b) Random Perturbation is based on Gradient's Slices where Each Slice has Equivalent Defense.

- Prevent Global Compensation from Being Abused by Attacker
  - [**Optional**]: Local Clipping Operation
    (Clipping Selected Gradients and Scaling them to similar range corresponding to the Scale of Perturbation)

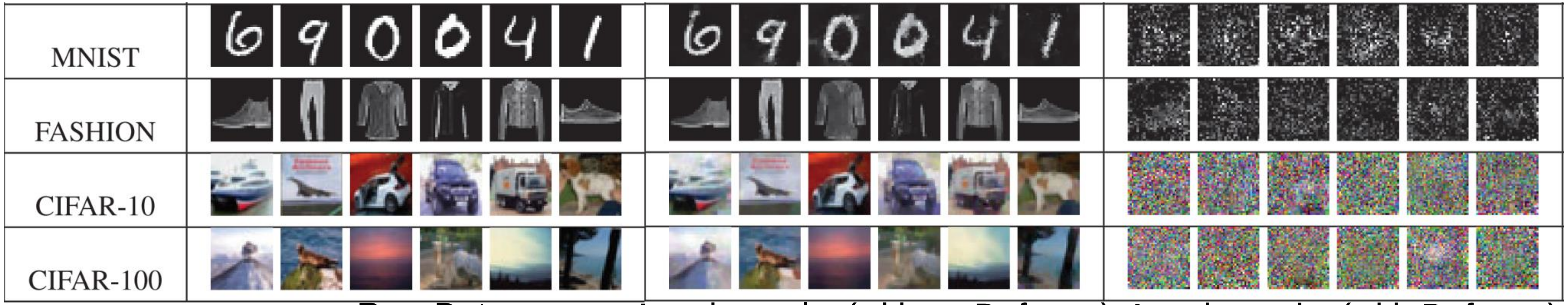
    - Global Compensation is still Valid.

Experimental Settings

- Attack Methods
  - [1] DGA, Deep Leakage from Gradients, NeurIPS2019.
  - [2] GIA, Inverting Gradients, NeurIPS2020.
- Baseline Defense Methods
  - [1] GC, Gradient Compression.
  - [2] DP, Differential Privacy, DP-Gaussian and DP-Laplacian.
  - [3] PLD, Provable Defense against Privacy Leakage in Federated Learning, CVPR2021.
- Cared Metrics
  - [1] Attack Reconstruction Quality (Image Similarities).
    - Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM).
  - [2] Accuracy (ACC) of Global Model on the Testing Set.
  - [3] Average Round Time (ART) of Training.
- Datasets and Model
  - MNIST, Fashion-MNIST, CIFAR, Convolutional Networks (LeNet)

Experimental Results

- Privacy Protection Perspective



Raw Data Attack results (without Defense) Attack results (with Defense)

(a) Visualization of Privacy Protection Results.

[A] Measure on Different Defenses against the DGA.

	MNIST - ACC 91.69% without defenses				Fashion-MNIST - ACC 91.80% without defenses				CIFAR-10 - ACC 54.15% without defenses			
	Ours	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]
PSNR	**9.41**	9.52	9.36[9.39]	9.57[18.49]	**9.66**	9.83	9.57[9.62]	9.89[19.78]	**9.61**	9.79	9.55[9.52]	9.88[24.48]
SSIM	**4.6E-2**	5.1E-2	4.1E-2[4.3E-2]	5.3E-2[6.4E-1]	**7.3E-2**	7.7E-2	7.1E-2[6.5E-2]	8.2E-2[8.4E-1]	**2.5E-2**	2.6E-2	2.3E-2[2.4E-2]	2.9E-2[8.8E-1]

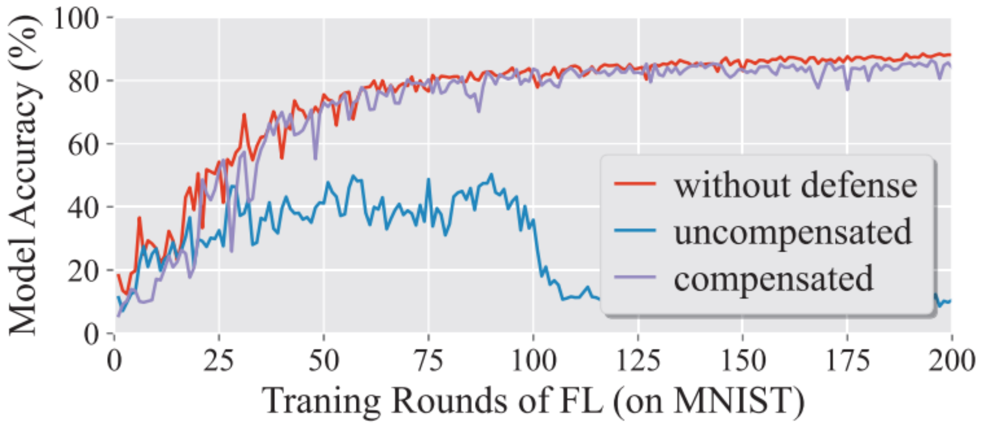
[B] Measure on Different Defenses against the GIA.

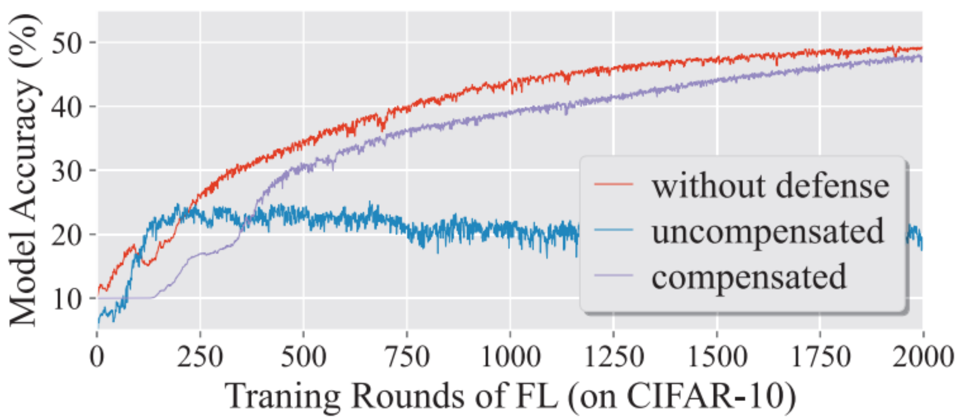
	MNIST - ACC 88.14% without defenses				Fashion-MNIST - ACC 86.57% without defenses				CIFAR-10 - ACC 49.31% without defenses			
	Ours	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]
PSNR	**9.83**	10.01	9.66[9.59]	10.43[19.61]	**9.91**	9.98	9.74[9.80]	10.14[21.23]	**10.11**	10.32	9.95[9.86]	10.79[27.04]
SSIM	**4.9E-2**	5.1E-2	4.4E-2[4.6E-2]	5.7E-2[7.3E-1]	**7.5E-2**	8.3E-2	6.8E-2[6.7E-2]	8.9E-2[9.5E-1]	**4.1E-2**	4.2E-2	3.0E-2[3.4E-2]	4.4E-2[9.3E-1]

(b) Numerical Results of Privacy Protection (PSNR, SSIM).

Experimental Results

- Convergence Accuracy Perspective





(a) Visualization of Convergence Accuracy Results.

- Overhead Perspective

[A] Measure on Different Defenses against the DGA.

	MNIST - ACC 91.69% without defenses				Fashion-MNIST - ACC 91.80% without defenses				CIFAR-10 - ACC 54.15% without defenses			
	Ours	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]
ACC	**90.43%**	36.52%	10.37%[10.21%]	87.77%[-]	**89.29%**	33.11%	10.10%[9.98%]	86.35%[-]	**52.47%**	29.84%	10.19%[10.00%]	49.91%[-]
ART	**+8.45%**	+4.63%	+3.91%[3.74%]	+14.52%[-]	**+8.11%**	+3.75%	+3.89%[4.04%]	+13.20%[-]	**+8.97%**	+3.58%	+4.03%[4.31%]	+14.09%[-]

[B] Measure on Different Defenses against the GIA.

	MNIST - ACC 88.14% without defenses				Fashion-MNIST - ACC 86.57% without defenses				CIFAR-10 - ACC 49.31% without defenses			
	Ours	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]	**Ours**	GC	DP-G[-L]	PLD[-muted]
ACC	**86.87%**	32.29%	10.46%[9.85%]	84.09%[-]	**84.65%**	30.38%	9.86%[9.77%]	81.10%[-]	**47.73%**	23.35%	10.01%[10.16%]	45.16%[-]
ART	**+9.07%**	+4.90%	+3.84%[3.66%]	+16.12%[-]	**+8.62%**	+4.23%	+4.14%[3.99%]	+15.86%[-]	**+9.33%**	+4.08%	+4.15%[4.02%]	+16.43%[-]

(b) Numerical Results of Accuracy (ACC) and Average Round Time (ART).

Thank you!