Protect Privacy from Gradient Leakage Attack in Federated Learning

Protect Privacy from Gradient Leakage Attack in Federated Learning
0. Abstract
1. Introduction
2. Preliminaries
    A. Federated Learning
    B. Threat Model
    C. Gradient Leakage Attack
    D. Motivation
3. Overview of Defensive Mechanism
4. Local Random Perturbation
    A. Quantifying the Information Leakage Risk
    B. Risk Balancing based Random Perturbation
5. Global update compensation
    A. Attribute Correlation within Global Gradient
    B. Global Gradient Compensation
    C. Avoid Abusing Compensation
    D. Convergence Analysis
        Assumption 1
6. Experiments
    A. Experiment Setup
        Attack Methods
        Defense Baselines
        Evaluation Metrics
        Datasets
        Models
    B. Experimental Results: from Main Perspective
    C. Power of Leakage Risk Quantifying
    D. Insights on Compensation
7. Releated Work
    A. Gradient Leakage Attack
    B. Protection on Data Privacy
        MPC
        HE
        DP
8. Conclusion

0. Abstract

FL 容易遭受 gradient leakage attacks

已有的工作是加入了一些隐私保护机制：HE、 LDP

缺点：导致 communication and computation costs or training accuracy loss.

defense: perturbing gradients, lightweight.

1. Introduction

shared gradients still leak private information

honest-but-curious server 可能通过拦截 client 的 gradient，并且重构 raw data，悄无声息地泄露隐私信息

隐私保护方法要求：

（1）lightweight，在实现 computation, memory, bandwidth方面都是轻量级的。

（2）guaranteed training accuracy. trade-off between privacy-performance and training accuracy loss.

（3）adequate of the privacy protection.

4 个 privacy preserve solutions:

（1）Homomorphic Encryption (HE)

advantages: HE 没有带来 training accuracy loss 以及 no noise is added to gradients.

disadvantages: significant overhead to computation and communication.

（2）Local Differential Privacy (LDP)

injecting noise into gradients at client side.

tradeoff between convergence performance and privacy protection.

disadvantages: stronger privacy protection leads to significant training accuracy loss.

（3）Multi-Party Computation(MPC)

多方协同计算一个 function，每个参与方只知道它自己的 input 和 output

bottleneck: high communication cost, and the communication-efficient implementations require either extensive offline computation or lower privacy protection

（4）Targeted Defense(TD)

专门针对特定的攻击域，eg. gradient leakage attack in FL

disadvantages: defense pattern 会容易被推断出

输入数据是信息泄露风险的一个必须的措施

intuition:

1. perturbing the gradients to match their leakage risk.
2. 梯度的全局性可以补偿 perturbation 带来的损失

两种关键技术：

1.Layer-wise information leakage risk quantification and balanced perturbation

使用 gradient changes sensitivity：输入的信息去量化 the leakage risk.

sensitivity 可以捕获信息如何在 layer-wise 情况下泄露，

可以促进一个有效的 perturbation 设计，可以平衡各层之间的 defense.

2.Perturbation compensation based on correlation among layer-wise gradient attributes

global gradients的 attributes 有很强的 correlation，因为都是由 layer-wisely backward propagation

3.Experimental results validate our superior on preserving privacy and performance

2. Preliminaries

- 重新介绍 FL
- 讨论已有的防御技术

A. Federated Learning

n 个 clients(eg. mobile devices) 在 a federated server 组织下，协同训练一个model

training data 被保存在本地并且没有被交换

（1）Client selection

对于第 t 轮, federated server 从 n 个 clients 中，取样一个子集 m clients。

（2）Local training

在第 t 轮被选中的 clients，下载当前 global model 的参数 θ 以及一个当前的训练问题

每个被选中的 client 在本地使用 local training data计算一个 update

在一个 client 上的 the gradient update，是用

$$\frac{\partial l(X, y, \theta)}{\partial \theta}$$

计算得到， X 是 training data batches，y 是 training data batches 对应的 labels

$l(\cdot)$ 代表的是 loss function

FedAvg, models 持续在 local data 上连续更新多个 epochs

（3）Global Aggregation

收到了从 m 个 clients 的 local updates, federated server 聚合这些 updates, 并且更新 global model

B. Threat Model

federated server 是一个 honest-but-curious server

server 会 honestly 执行 local updates 的 aggregation，并组织 federated learning 的 iteration

但是, federated server 可能是 curious 并且会分阶段性地分析 gradients 去执行 gradient leakage attacks，并且获取访问 victim clients 的 training data

而且即使在client 和 server 的通信连接是安全的，梯度泄露攻击也是可能发生的

C. Gradient Leakage Attack

gradient leakage attack 是一个 data reconstruction attack

attackers 设计一个 gradient-based reconstruction learning algorithm.

D. Motivation

几种现有的防御机制：

1. local differential privacy
2. secure multi-party computation
3. homomorphic encryption

disadvantages:

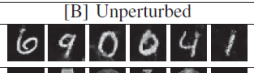
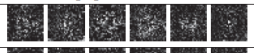
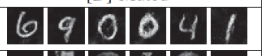
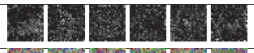
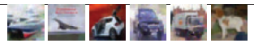
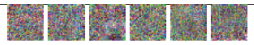
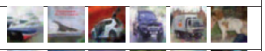
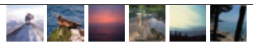
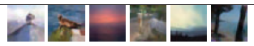
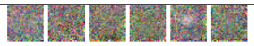
1. unacceptable computational overheads
2. significant accuracy degrade

因为这些机制是为通用的防御设计的，而不是专门为 gradient leakage attack 设计的

最新研究的 targeted defense, 但是只能干扰共享梯度的某一层，防御模式僵化，容易被推断，隐私保护效果不佳

> A: raw data of local training
>
> B: 使用 unperturbed gradients 重构的结果
>
> C: 使用 perturbed gradients 的结果
>
> D: 扰动部分层之后的结果，attacker 使用 unperturbed layers 进行重构

TABLE I
RAW DATA AND RECONSTRUCTED RESULTS OF [18] OVER MNIST, FASHION-MNIST, CIFAR-10 AND CIFAR-100 DATASETS.

Dataset	[A] Raw data	[B] Unperturbed	[C] Perturbed	[D] Muted
MNIST				
FASHION				
CIFAR-10				
CIFAR-100				

3. Overview of Defensive Mechanism

workflow:
(1) local random perturbation

给出 lightweight and adequate protection for the gradients

(2) global update compensation

lower footprints of perturbation and training accuracy loss

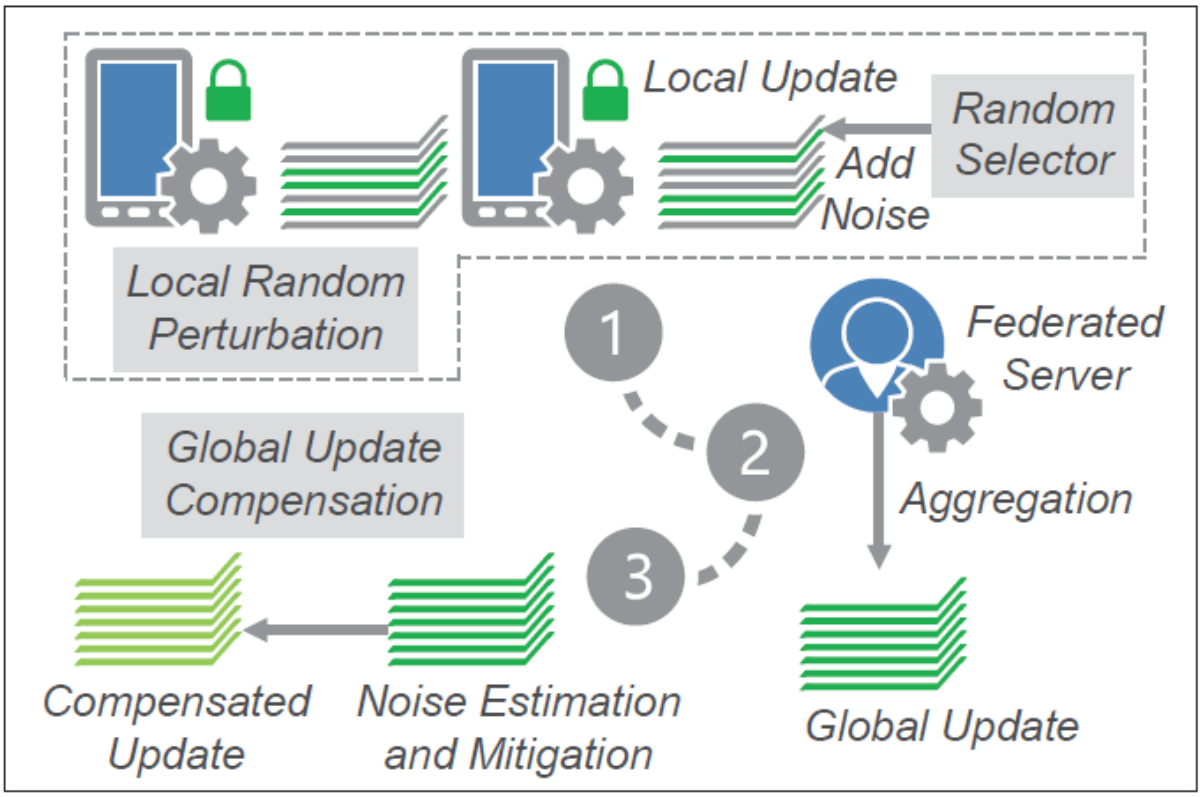


Fig. 1. Overview and workflow of the proposed defensive mechanism.

1. Shield training data with local random perturbation

   利用 sensitivity of gradient changes, 使用输入信息量化 layer-wise the leakage risk.

   sensitivity 是一个易于计算的 metric，捕获信息的分层泄露，将梯度切片分成多个切片，每一个 slice 都具有平衡的 information leakage risk.

   每一轮， client 都会从共享的 gradients 中随机选择一个 slices 的份额，并且给这些 slices 添加一些人造 noise

   （选择slices 以及 noise-adding 是 locally and randomly，因此对于 attacker 很难 locate noisy slices，也就难以重构 data）

   一旦 noisy slices 的份额超过某个具体的 level，即使 , attacker 能够通过暴力破解 locate noisy slices也无法 reconstruct the data.

2. Retain convergence performance with global update compensation.

   global gradients 的 layer-wise attributes 是高度相关的，因为他们是基于 data samples 在 BP 过程中分层产生的

   global gradients 的 attributes 之间关系性很强，在一些方向上方差很大，一些方向上方差很小

   noise 的加入不会改变这趋势，加入到 original data 中的 random variables 是 independent，因此 variance 也是均匀分布的。

3. Prevent compensation from being abused with local clipping operation

attacker 可能利用 compensation 来缓解 shared gradients 中的 noise，从而完成数据重构。

使用 local clipping operation，对个别样本数据的 gradients 裁剪，并缩放到与 noise 的相似范围

4. Local Random Perturbation

A. Quantifying the Information Leakage Risk

sensitivity: input data 量化了 output 通过在对应的 input data 中 small changes 的影响程度

使用 sensitivity 来量化从 gradients 的 information leakage risk

X 是一个小的训练集 root dataset, 可能源自一个 publicly available data source, 或者是对 raw data 的 distillation

information leakage risk 是通过 attacker 能够 extract 多少信息

使用 gradients 的 Jacobian matrix，反映梯度上的 sensitivity measure.

> $l(\cdot)$ 是在 input X 上的 loss
>
> y 是truth
>
> θ 是entire model 的 parameters

$$\mathbf{J}_l^{\boldsymbol{G}}(X) = \frac{\partial \mathbf{g}_l(X)}{\partial X} = \frac{\partial}{\partial X}\left[\frac{\partial \ell(X, y, \theta)}{\partial \theta_l}\right] \tag{1}$$

给出 k 个 data samples，使用 p-norm averaged 来作为 $l-th$ layer 的 information leakage risk.

> $||\cdot||_p$ 的选择反映了 data samples 的 distance

$$\mathcal{R}_l^{\boldsymbol{X}} = \mathbb{E}\left[\left\|\mathbf{g}_l(X) - \mathbf{g}_l(X + \Delta X)\right\|_p\right] = \frac{1}{k}\sum_{i=1}^{k}\left\|\mathbf{J}_l^{\boldsymbol{G}}(X_i)\right\|_p \tag{2}$$

B. Risk Balancing based Random Perturbation

较大的 G_l 意味着更容易受到攻击

S_l 是 $l-th$ layer 的 size

R_l 是$l-th$ layer 的 leakage risk

要平衡 leakage risk weighted 与 layer size

使用每层的 S_l 的最大公约数，将梯度划分为多个大小相同的块

S_b 是 $b-th$ block 的 size

R_b 是 $b-th$ block 的 leakage risk

gradients 被划分为多个 slices，每个 slice 的 leakage risk 是相同的

在每一轮， 每个 client 会从共享的 gradients 中随机选择一定份额的 slices，并给这些 slices 添加 noise

对于划分为 s slices 的 gradients，γ 是 perturbation ratio,来决定 perturbed slices 的数量

max(1,$\lfloor s \times \gamma \rfloor$)

添加的 nosie 服从 Gaussian distribution $N(\mu, \sigma^2)$, 均值μ 是 0 ， 方差是 σ^2

随机扰动是基于 Reservoir Sampling，计算复杂度$O(s)$

Algorithm 1: Local Random Perturbation

Input: G local gradients computed in one client;
γ perturbation ratio predefined among participants;
σ standard deviation of noise variables;
Output: G^* local gradients shared to the server;

```
Partition(G) ⟶ sliced G[0...s−1];
Max(1, ⌊s × γ⌋) ⟶ perturbed slice number d;
index = Slice_selector(s, d);
for each i in index do
    G[i] = G[i] + Gaussian(0, σ^2);
return G* = G;

DEF Slice_selector(s, d):
    INIT reservoir A[d] = [0...d−1];
    for i = d; i < s; i++ do
        Pseudorandom j out of [0...i];
        if j < d then
            A[j] = i;
    return index = A[d];
```

5. Global update compensation

A. Attribute Correlation within Global Gradient

对于 $l-th$ layer，gradient vector G_i 包括了 weights 和 bias 的梯度，这些梯度是由 chain rule（链式法则）计算得到

global gradients 的 layer-wise 的 attributes 由于是在 BP 过程中通过 layer-wise manner 方式生成的，因此是高度相关的，因此对于一些 vectors，在一些方向有大的 variance，在其他的方向有小的 variance

从扰动数据中，更准确地获取原始数据 作为对 global update 的 compensation

矩阵shape $(v \times w)$

C 代表扰动后的 global gradients 的matrix

Z 代表 C 的原始数据

R 代表 noise variables 的随机矩阵, elements 满足 i.i.d.(均值为0，variance $\sigma^2 = m \cdot \frac{d}{c} \sum_{i=1}^{m} \epsilon\sigma^2$)

m 是 pariticipants 的数量，s 是 total slices 的数量，d 是perturbed slices 的数量，σ^2 的noise 的variance，$\epsilon \in (0, 1)$ 是 aggregation 的权重

C = Z + R

covariance matrix of R is: $L_R = \frac{1}{v} R^T R$, 是$(w \times w)$

λ 是特征值，经验累积分布函数：

$$\Gamma(x) = \frac{1}{w} \sum_{i=1}^{w} U(x - \lambda_i)$$

x > 0 时，U(x) = 1

x < 0 时，U(x) = 0

B. Global Gradient Compensation

global gradients 的 matrix C , 其收敛性矩阵 L_C 满足：

$$\begin{aligned} v \cdot \mathrm{L_C} &= \mathrm{C^T C} = \mathrm{P_C \Lambda_C P_C^T} = (\mathrm{Z + R})^{\mathrm{T}} (\mathrm{Z + R}) \\ &= \mathrm{Z^T Z + R^T Z + Z^T R + R^T R} \\ &= \mathrm{P_Z \Lambda_Z P_Z^T + P_R \Lambda_R P_R^T + R^T Z + Z^T R} \end{aligned} \tag{5}$$

P_C, P_Z, P_R 是正交矩阵，列向量分别是 $C^T C, Z^T Z, R^T R$的特征值

$\bigwedge C, \bigwedge Z, \bigwedge R$ 是在对角线上是对应的特征值的对角矩阵

1. 计算 perturbed gradients C 的 covariance matrix
2. 根据随机矩阵 theory，利用 C 的特征值的分布，协方差矩阵 C 被划分为 noise 和 data parts
3. 然后利用数据部分对应的特征值，对实际的 global gradients 进行补偿

   （因为noise分布是通过一个已知的 variance σ^2，通过等式4计算 λ_{min} and λ_{max}，提供了对于noise 特征值的理论的界，以此来识别 noise 的特征值，剩下的则是 actual data 的特征值）

> $\bigwedge_R^* = diag(\lambda_i, \lambda_{i+1} \ldots \lambda_j)$ 是noise相关的特征值组成的对角矩阵
>
> P_R^* 是列向量是对应的特征向量的矩阵

> $\bigwedge_Z^*$ 是 actual data 的特征值矩阵
>
> P_Z^* 是 actual data 对应特征向量的矩阵
>
> Z^* 是 actual data Z 的 compensation，将 C 映射到由 P_Z^* 张成的子空间
>
> $Z^* = CP_Z^*P_Z^{*T}$

$$\mathcal{F}_Q(x) = \frac{Q\sqrt{(x-\lambda_{\min})(\lambda_{\max}-x)}}{2\pi\hat{\sigma}^2 \cdot x}, x \in [\lambda_{\min}, \lambda_{\max}] \qquad (4)$$

C. Avoid Abusing Compensation

attacker 可能会利用 compensation 去减轻 shared gradients 中 noise 的影响，从而重构data

使用 local clipping operation 技术来解决，通过 clipping gradients，并放缩到与对应的 noise 类似的 range

local clipping operation:

1. 假设添加的 local noise 满足 Gaussian distribution N(0,$\sigma^2 \cdot B^2$)

   B 是noise 的强度，决定 noise 的范围
2. 在 local training 的每一轮，使用 norm clip 裁剪 gradients

   例如：使得梯度 g 被 $g/max(1, \frac{||g||_2}{B})$ 替换

   > 若 $||g||_2 \leq B$, 则 g 保留
   >
   > 若 $||g||_2 > B$, 则根据 noise 强度 B 进行了放缩
3. 在 server，即使对 clipped gradients 进行了补偿，也很难揭示是否一个特定的 data sample 参与了学习

D. Convergence Analysis

η 代表学习率

ϑ代表global 积累的 errors 和 compensation

每个 client i 在轮数 t 的时候，本地计算 local stochastic gradients $\triangle F(x_t; \xi_t^{(i)})$,基于 global model x_t，和 local data samples $\xi_t^{(i)}$

updae rule:

$$\begin{aligned} \boldsymbol{x}_{t+1}-\boldsymbol{x}_t &= -\eta\left[\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\Delta F}(\boldsymbol{x}_t;\boldsymbol{\xi}_t^{(i)})+\vartheta_{t-1}-\vartheta_t\right] \\ &= -\eta\boldsymbol{\Delta} f(\boldsymbol{x}_t)+\eta\boldsymbol{\zeta}_t-\eta\vartheta_{t-1}+\eta\vartheta_t \end{aligned} \tag{6}$$

where $\zeta_t=\frac{1}{m}\sum_{i=1}^{m}\left[\boldsymbol{\Delta} f(\boldsymbol{x}_t)-\boldsymbol{\Delta F}(\boldsymbol{x}_t;\boldsymbol{\xi}_t^{(i)})\right]$.

Assumption 1

(1) 假设$f(\cdot)$ 是一个 L-Lipschitzian gradients

$$||\triangle f(x)-\triangle f(y)||\leq L||x-y||$$

(2)假设stochastic gradients 的 variance 是 有界的

$$E||\triangle F(x;\xi)-\triangle f(x)||^2\leq\phi^2$$

(3)假设accumulated errors and global compensation 是有界的

$$E||\vartheta_t||\leq\frac{\psi}{2}$$

前两个假设是对于 non-convex convergence analysis 情况下使用的

第三个假设是用来限制 errors 的规模的

根据 $E|\triangle F(x_t;\xi_t^{(i)})|=\triangle f^i(x_t), E||\vartheta_t||^2\leq\psi^2$

得到 $E\zeta_t=0, E||\zeta_t||^2\leq\frac{\phi^2}{m}$

使用辅助变量 y_t 代表 $y_t=x_t-\eta\vartheta_{t-1}$

update rule:

$y_{t+1}-y_t=-\eta\triangle f(x_t)+\eta\zeta_t$

因为 $f(\cdot)$ 满足 L-Lipschitzian gradients, 因此有:

$$\mathbb{E}\|\boldsymbol{\Delta} f(\boldsymbol{y}_t)-\boldsymbol{\Delta} f(\boldsymbol{x}_t)\|^2\leqslant L^2\mathbb{E}\|\boldsymbol{y}_t-\boldsymbol{x}_t\|^2\leqslant L^2\eta^2\psi^2 \tag{7}$$

$$\mathbb{E}f(\boldsymbol{y}_{t+1})-\mathbb{E}f(\boldsymbol{y}_t)\leqslant\mathbb{E}\left\langle\boldsymbol{y}_{t+1}-\boldsymbol{y}_t,\boldsymbol{\Delta} f(\boldsymbol{y}_t)\right\rangle+\frac{L}{2}\mathbb{E}\|\boldsymbol{y}_{t+1}-\boldsymbol{y}_t\|^2=$$

$$
\begin{aligned}
&-\eta\mathbb{E}\langle\Delta f(x_t),\Delta f(y_t)\rangle+\eta\mathbb{E}\langle\zeta_t,\Delta f(y_t)\rangle+\frac{L\eta^2}{2}\mathbb{E}\|\Delta f(x_t)-\zeta_t\|^2\\
&=-\eta\mathbb{E}\langle\Delta f(x_t),\Delta f(y_t)\rangle+\frac{L\eta^2}{2}\mathbb{E}\|\Delta f(x_t)\|^2+\frac{L\eta^2}{2}\mathbb{E}\|\zeta_t\|^2\\
&\leqslant-\eta\mathbb{E}\langle\Delta f(x_t),\Delta f(y_t)\rangle+\frac{L\eta^2}{2}\mathbb{E}\|\Delta f(x_t)\|^2+\frac{L\eta^2\phi^2}{2m}\qquad(8)\\
&\leqslant\frac{-\eta+L\eta^2}{2}\mathbb{E}\|\Delta f(x_t)\|^2+2L^2\eta^3\psi^2+\frac{L\eta^2\phi^2}{2m}\ \text{by Equation (7)}
\end{aligned}
$$

因此可以设置学习率为：

$$
1/(2L+\phi\sqrt{T/m}+\psi^{2/3}T^{1/3})
$$

通过对不等式从 t = 0 到 t = T - 1 求和，可以得到如下的收敛率：

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\Delta f(x_t)\|^2\leqslant\frac{\phi}{\sqrt{mT}}+\frac{1}{T}+\frac{\psi^{2/3}}{T^{2/3}},
$$

6. Experiments

A. Experiment Setup

在两种 gradient leakage attacks 情况下evaluate defensive mechanism, under non-IID settings.

datasets: MNIST, Fashion-MNIST, CIFAR-10

Attack Methods

deep gradient attack (DGA) 去 minimizes the Euclidean distance 在重构图像的BP 过程中生成的，between the real gradients and the dummy gradients.

gradient inverting attack (GIA) 用余弦相似度代替欧氏距离，并对梯度的符号进行优化

Defense Baselines

将提出的 defensive mechanism 与现有的 3 种防御方法比较：

（1）the gradient compression (GC)

修剪低于阈值的梯度，以便保留部分梯度

（2）the differential privacy (DP)

在梯度种添加噪声老保护隐私

实验种分别应用了 Gaussian 和 Laplace noise 作为两种 DP 的 baselines（DP-G, DP-L）

（3）the most recent privacy leakage defense (PLD)

修剪 fully-connected layer 的 gradient

Evaluation Metrics

- 使用了

(1)the peak signal-to-noise ratio(PSNR) 峰值信噪比

(2)structural similarity index measure (SSIM) 结构相似指数测量

来量化在 reconstructed image and raw image 之间的defense effectiveness.

- 使用在测试集上的 global model 的 the accuracy(ACC) 来测量
- 使用 the average round time (ART) 来检查防御是否是轻量级

Datasets

MNIST and Fashion-MNIST

数据分布在100个clients

每轮选择10个paritcipants

每个participants持有2 个 random classes of data (每个class有 1-10个samples)

total rounds = 2000

local batch size = 1~20

CIFAR-10

数据分布在1000个clients

每轮选择200个paritcipants

每个participants持有2 个 random classes of data (每个class有 1-20个samples)

total rounds = 2000

local batch size = 1~40

对于其他的超参数，使用 learning rate 为 1e-2 SGD optimizer，epoch = 2

Models

LeNet model for DGA attack

4 个 convolutional layers, 后接 1 fully-connected layer

ConvNet model for GIA attack

8 个 convolutional layers, 后接 1 fully-connected layer

B. Experimental Results: from Main Perspective

the worst case:

(1) only one data sample in each batch

(2) label information has been known

settings:

> GC: pruning rate of gradients 80% for DGA, 90% for GIA
>
> DP: noise variance to 1E-2 both against the DGA and GIA
>
> PLD: pruning rate of the fully-connected layer 60% for DGA, 80% for GIA
>
> our defense: noise variance and perturbation rate to 1E-2,20% for DGA, 1E-2,30% for GIA

apply 500 iterations L-BFGS optimizer of reconstruction for the DFA

apply 1000 iterations Adam optimizer for the GIA

在minimal defenses情况下 evaluate convergence performance and overhead of training

> **our solution**: less than 2% acc loss and less than 10% overhead increment.而且可以有效减小 PSNR and SSIM
>
> **PLD**: 可以实现较低的 acc loss，但是会显著增加 PSNR and SSIM
>
> **GC and DP**: less overhead increment, loss of convergence performance.
>
> (此外，这是在 the worst defense 情况下，在其他的 defense 情况下，效果更好)

TABLE II
RESULTS OF MEASURE ON DEFENSES AGAINST DIFFERENT ATTACKS AND DIFFERENT DATASETS.

[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]
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]
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]
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%[-]

C. Power of Leakage Risk Quantifying

利用 GIA 来探究 the importance of leakage risk quantification.

another baseline: slice based on the layer size balancing.

使用每一层的 the greatest common divisor 把 gradients 划分为多个相同大小的 slices

在相同的perturbation ratio 情况下，client 随机选择相同比例的 slices，并给这些 slices 添加 Gaussian 噪声

root dataset：100 public data samples 去量化 gradient sensitivity (the information leakage risk)

- the gradient sensitivity based slicing 可以更精确地捕获原始信息从梯度泄露的风险，因此其对应的 PSNR、SSIM 都显著小于 baseline.
- the impact of the root dataset on quantifying leakage risk (its size)

可以看到，只有100个训练样本的根数据足以捕获 information leakage risk.

- the impact of the root dataset (its bias probability)

bias probability 从0.1 增大到 1.0时候 去累积 root dataset distribution and the global dataset distribution difference

当root dataset distribution 不要太偏离 global data distribution 时候，泄露风险是可以量化的

TABLE III
RESULTS ON RISK QUANTIFICATION AND COMPENSATION.

[A] Gradient Sensitivity-based Slicing *vs.* Layer Size-based Slicing.

	MNIST		Fashion-MNIST		CIFAR-10	
	sensitivity	layer size	**sensitivity**	layer size	**sensitivity**	layer size
PSNR	**9.83**	14.51	**9.91**	13.70	**10.11**	15.02
SSIM	**4.9E-2**	1.2E-1	**7.5E-2**	2.6E-1	**4.1E-2**	1.5E-1

[B] Impact of Root Dataset *w.r.t.* its Size.

	MNIST					CIFAR-10				
	50	**100**	200	300	400	50	**100**	200	300	400
PSNR	11.40	**9.83**	9.71	9.68	9.66	11.95	**10.11**	10.10	10.04	9.99
SSIM	6.2E-2	**4.9E-2**	4.5E-2	4.2E-2	4.0E-2	5.7E-2	**4.1E-2**	4.0E-2	3.9E-2	3.8E-2

[C] Impact of Root Dataset *w.r.t.* its Bias Probability.

	MNIST					CIFAR-10				
	0.1	0.2	**0.4**	0.6	1.0	0.1	0.2	**0.4**	0.6	1.0
PSNR	9.89	10.32	**10.53**	11.68	13.66	10.41	10.67	**10.89**	12.03	13.72
SSIM	5.0E-2	5.4E-2	**5.7E-2**	6.9E-2	9.1E-2	4.3E-2	4.5E-2	**4.8E-2**	6.4E-2	8.8E-2

[D] Impact of Q on Convergence Performance.

	MNIST				CIFAR-10			
	1	**6**	12	20	1	**6**	12	20
ACC	82.92%	**86.87%**	86.99%	86.10%	41.64%	**47.73%**	47.85%	47.34%

[E] Local Perturbation with Clipping *vs.* Local Perturbation without Clipping.

	MNIST		Fashion-MNIST		CIFAR-10	
	clipping	no clipping	**clipping**	no clipping	**clipping**	no clipping
PSNR	**9.95**	12.36	**10.02**	12.74	**10.66**	13.28
SSIM	**5.2E-2**	9.9E-2	**7.7E-2**	9.5E-2	**4.8E-2**	8.7E-2
ACC	**86.03%**	86.87%	**83.87%**	84.65%	**47.15%**	47.73%
ART	**+1.32%**	-	**+0.98%**	-	**+1.14%**	-

D. Insights on Compensation

在 MNIST 和 CIFAR-10 数据集上，评估compensation 的效果

defense settings for the GIA

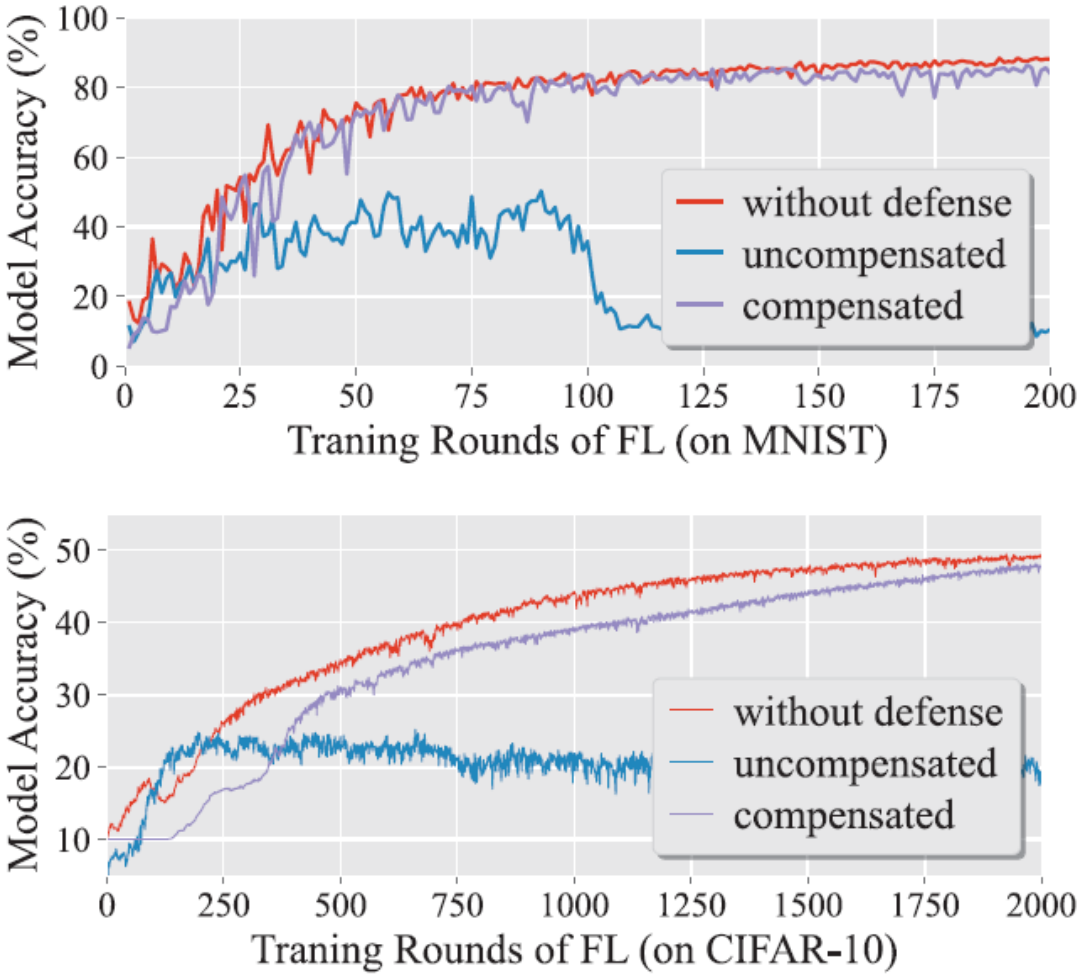


Fig. 2. Comparison between uncompensated and compensated defense.

evaluate Q 对 compensation 的影响

Q在6-12足以表征全局梯度的相关性， compensation errors 也减小到了一个相对稳定的值，超过6后，收敛性会有所提高，但是较大的Q 也会使得性能反而下降

[D] Impact of Q on Convergence Performance.

	MNIST				CIFAR-10			
	1	**6**	12	20	1	**6**	12	20
ACC	82.92%	**86.87%**	86.99%	86.10%	41.64%	**47.73%**	47.85%	47.34%

attacker 可能利用 compensation 去缓解在 shared gradients 的noise

解决办法：在 local training 的过程中 Clipping these gradients

> clipping 技术可以防止 compensation 被滥用
>
> gradient clipping 技术可以使得 raw image 无法从 reconstructed images 中识别出来，因此会有较低的 PSNR 和 SSIM
>
> acc loss less than 1%

> overhead increment less than 1.5%

[E] Local Perturbation with Clipping *vs.* Local Perturbation without Clipping.						
	MNIST		Fashion-MNIST		CIFAR-10	
	clipping	no clipping	**clipping**	no clipping	**clipping**	no clipping
PSNR	**9.95**	12.36	**10.02**	12.74	**10.66**	13.28
SSIM	**5.2E-2**	9.9E-2	**7.7E-2**	9.5E-2	**4.8E-2**	8.7E-2
ACC	**86.03%**	86.87%	**83.87%**	84.65%	**47.15%**	47.73%
ART	**+1.32%**	-	**+0.98%**	-	**+1.14%**	-

[E] Local Perturbation with Clipping *vs.* Local Perturbation without Clipping.						
	MNIST		Fashion-MNIST		CIFAR-10	
	clipping	no clipping	**clipping**	no clipping	**clipping**	no clipping
PSNR	**9.95**	12.36	**10.02**	12.74	**10.66**	13.28
SSIM	**5.2E-2**	9.9E-2	**7.7E-2**	9.5E-2	**4.8E-2**	8.7E-2
ACC	**86.03%**	86.87%	**83.87%**	84.65%	**47.15%**	47.73%
ART	**+1.32%**	-	**+0.98%**	-	**+1.14%**	-

7. Releated Work

A. Gradient Leakage Attack

reconstruction is possible for a single neuron or linear layer.

reconstruction of a single image is possible for a 4-layered CNN consisting of a significantly large fully-connected layer.

the label information can also be jointly reconstructed

label information can be inferred analytically from the gradients of the last one layer,

the reconstruction of multiple images from their averaged gradients.

B. Protection on Data Privacy

MPC

utilize the garbled circuits or secret sharing, allowing multiple parties to collaboratively compute a function in a protocol that each party knows nothing except its input and output

bottleneck : the high communication cost, and communication-efficient implementations

现有技术提出安全聚合，可以确保server 只能学到 aggregateed data 而不能学到 client 的 local data

但是在每一轮都需要同步 secret keys 以及 zero-sum masks，带来了较大的同步问题

HE

enable the certain computation (e.g., addition) to be performed directly on encrypted data without decrypting them first

bottleneck: computationally inefficient

DP

a noisy release mechanism

injecting artificial noise at the client side

bottleneck: injecting artificial noise at the client side

8. Conclusion

1. lightweight defense overhead
2. guaranteed training accuracy
3. adequate privacy protection

提出的方法能够在不牺牲 acc 的情况下，轻量级的降低在 raw images 和 reconstructed images 的 PSNR , SSIM 比 baseline defensive methods 降低了超过 60%