Matrix Sketching for Secure Collaborative Machine Learning

0. Abstract

Collaborative learning , data 不离开本地，但是通信的 gradients and parameters 可能会泄露 client 的隐私

Attacks 可能会从 gradients and parameters 推断 client 的隐私

Defenses: dropout and differential privacy 要么无法抵御攻击要么损失了 accuracy。

Proposed: Double-Blind Collaborative Learning(DBCL)

（1）应用 random matrix sketching to parameters

（2）re-generate random sketching after each iteration

DBCL成功防御了clients 的 gradient-based privacy inferences.

原因：sketching是高效的，random noise 要大于 signal.

优点：

（1）DBCL没有增大 computation and communication costs.

（2）没有损失accuracy.

1. Introduction

Distributed SGD:

(1) central server broadcasts model parameters to the clients.

(2) client 使用 local data 计算 stochastic gradient.

(3) server aggregates the stochastic gradients and update model parameters.

基于 distributed SGD 的算法，例如: FedAvg、FedProx.

Attacks:

model parameters and gradients 的一些 important properties 会被推断出

攻击产生的原因是联合学习的model包含了其他用户的data

Goal:

defend the *gradient-based attacks*.

DP:

噪声会损失accuracy

Dropout:

随机mask一些参数，使得clients只能访问到一部分的parameters，但即使这样也会导致攻击的产生

Proposed method

提出 Double-Blind Collaborative Learning (DBCL)，来防御 gradient-based attacks.

(1) random sketching.

(2) sketching matrices are regenerated.

clients see: sketched parameters.

server sees: approximate gradients based on sketched data and sketched parameters.

An honest client's perspective:

DBCL类似于dropout，使用sketching 代替了 uniform sampling.

dropout 不损失 test accuracy ----> DBCL 不损失 test accuracy.

An attacker's perspective:

random noise 注入到 gradient

- Estimated Grad = Transform（True Grad）+ Noise

因此，client-side gradient-based attacks 不起作用

DBCL features:

- 不损失 test accuracy.
- 不增加 per-iteration time complexity and communication complexity.
- 不需要 extra tuning.

Difference from prior work

sketch the model parameters, not the gradients.

sketch the gradient 会损害 accuracy.

Limitations

(1) 恶意的 client 可能执行 parameter-based attacks.

(2) 恶意的 server 可能执行 inferring client's privacy.

2. Preliminaries

Dense layer

input shape: d_{in}

output shape: d_{out}

batch size: b

input: $x \in R^{b \times d_{in}}$

parameter matrix: $W \in R^{d_{out} \times d_{in}}$

output: $Z = XW^T$

activation function: $\sigma(Z)$

Backpropagation

从上一层收到的 gradient : G = $\frac{\partial L}{\partial Z}$

需要计算：

$$\frac{\partial L}{\partial \mathbf{X}} = \mathbf{GW} \in \mathbb{R}^{b \times d_{\text{in}}} \quad \text{and} \quad \frac{\partial L}{\partial \mathbf{W}} = \mathbf{G}^T\mathbf{X} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}},$$

$$= \frac{\partial L}{\partial Z} \cdot \frac{\partial Z}{\partial X} \qquad\qquad = \frac{\partial L}{\partial Z} \cdot \frac{\partial Z}{\partial W}$$

$$= G \cdot W$$

(1)

使用 $\frac{\partial L}{\partial W}$ 来更新 parameter matrix W：W <--- W - $\eta \frac{\partial L}{\partial W}$

(2)将 $\frac{\partial L}{\partial X}$ 传递到 lower layer.

Uniform sampling matrix

如果矩阵S 的 columns 是均匀随机取样的，$S \in R^{d_{in} \times s}$ 是一个 uniform sampling matrix

Random matrix theories保证了：

（1）$E_S[XSS^TW^T] = XW^T$

（2）$||XSS^TW^T - XW^T||$ 是bounded

CountSketch

$S \in R^{d_{in} \times s}$是一个 CountSketch matrix

(1) S 的每一行都只有一个非零项

(2) value 是从 {+1,-1} sampled

次

$$\mathbf{S}^T = \begin{bmatrix} 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

CountSketch 类似 random Gaussian matrices.

input: $X \in R^{b \times d_{in}}$

- CountSketch：$\tilde{X} = XS$ 在 $O(d_{in}b)$ 时间内计算。
- 标准矩阵乘法：$O(d_{in}bs)$时间复杂度内计算

理论已证明：

（1）$E_S[XSS^TW^T] = XW^T$

（2）$||XSS^TW^T - XW^T||$ 是bounded

3. Threat Models

client-server architecture

假设 attacker 控制了一个 client

第 k 个client知道 W_{old}, W_{new} 以及他的 direction Δk

则可以计算其他 clients's direction sum:

$$\begin{aligned}\sum_{i \neq k} \boldsymbol{\Delta}_i &= m\boldsymbol{\Delta} - \boldsymbol{\Delta}_k \\ &= m\left(\mathbf{W}_{\text{old}} - \mathbf{W}_{\text{new}}\right) - \boldsymbol{\Delta}_k. \qquad (1)\end{aligned}$$

这对于two-party 的协同训练则是直接泄露了信息。

gradient-based attacks

受害者的隐私是从 gradient中泄露的

decentralized learning

参与者在peer-to-peer network 的攻击和防御

4. Proposed Method

4.1 High-Level Ideas

attacks 需要受害者的 updating direction, eg. direction 来推断隐私信息

- Distributed SGD
- Federated Averaging(FedAvg)

server sees: clients 的updating direction. $\Delta_1, \Delta_2 \ldots$

clients see: 联合训练的 model parameter, W

恶意的 client 可以得到其他 clients的 updating directions

DBCL

对 input 和 parameter matrices 使用 random sketching

计算 $\tilde{X} = XS$ and $\tilde{W} = WS$

不同 layers 有不同的 sketching matrices S

每次 W updated 以后，re-generate S

clients see: $\tilde{W}_{old} = W_{old}S_{old}$,$\tilde{W}_{new} = W_{new}S_{new}$

clients 尝试去计算 $\Delta = W_{old} - W_{new}$

实验证明：client 估计的 Δ 与真实的相差很大

4.2 Algorithm Description

Broadcasting

central server 生成种子 ψ

生成 random sketch $\tilde{W} = WS$

给所有的 clients 广播 ψ和$\tilde{W} \in R^{d_{out} \times s}$

（此处，s <d_{in}，每轮迭代后，server变化s）

Local forward pass

client

（1）使用 seed ψ 得到 sketch $\tilde{X}_i = X_i S \in R^{b \times S}$

（2）计算$Z_i = \tilde{X}_i \tilde{W}^T$

（3）得到$\sigma(Z_i)$作为后一层的input

（4）计算outputs L_i，loss 是在 size b 的 batch 上得到的

Local backward pass

令 $G_i = \frac{\partial L_i}{\partial Z_i} \in R^{b \times d_{out}}$

client locally calculates:

- $\Gamma_i = G_i^T \tilde{X}_i \in R^{d_o ut \times s}(1)$
- $\frac{\partial L_i}{\partial X_i} = G_i \tilde{W} S^T \in R^{b \times d_i n}(2)$

此处，(2)式传播到 lower-level layer 进一步进行反向传播

Aggregation

Server 聚合 (1) 式，

（1）去计算 $\Gamma = \frac{1}{m}\sum_{i=1}^{m} L_i$，此处需要一次通信

$L = \frac{1}{m}\sum_{i=1}^{m} L_i$、

（2）Server 计算updates:

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial L_i}{\partial \mathbf{W}} = \mathbf{\Gamma}\mathbf{S}^T \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}. \tag{2}$$

（3）Server updates:

W <-- W - $\eta\frac{\partial L}{\partial W}$

4.3 Time Complexity and Communication Complexity

Time complexity

CountSketch 计算 $\tilde{X}_i = X_i S$ and $\tilde{W}_i = W_i S$

costs: $O(bd_{in})$ and $O(d_{in}d_{out})$

forward and backward pass: $O(bd_{in} + d_{in}d_{out} + bsd_{out})$

VS compare standard backpropagation: $O(bd_{in}d_{out})$

Communication complexity

通信内容的比较：

不使用 Sketching:

- $W \in R^{d_{out} \times d_{in}}$
- $\frac{\partial L_i}{\partial W} \in R^{d_{out} \times d_{in}}$

使用Sketching:

- $\tilde{W} \in R^{d_{out} \times s}$
- $\Gamma_i \in R^{d_{out} \times s}$

因为 $s < d_{in}$，因此每轮的通信复杂度都比标准的 SGD 小

5. Theoretical Insights

5.1 Approximating the Gradient

client 拥有：

- $\tilde{W}_{old} = W_{old}S_{old}$
- $\tilde{W}_{new} = W_{new}S_{new}$

Attacker 必须知道 gradient: $\Delta = W_{old} - W_{new}$

(1) 不使用S_{old} 和 S_{new}

此时 Δ信息不可被恢复

估计 $\tilde{\Delta} = \tilde{W}_{old} - \tilde{W}_{new}$

由于在 iterations 之间可以变化 s 的大小，因此 $\tilde{W}$ 不同于W

(2) 使用S_{old}和S_{new}

此时计算的估计是：

$$\tilde{\Delta} = W_{old}S_{old}S_{old}^T - W_{new}S_{new}S_{new}^T$$

因为$\tilde{\Delta}$ 是一个Δ无偏估计，满足：$E[\tilde{\Delta}] = \Delta$

5.2 Defending Gradient-Based Attacks

Matrix sketching as implicit noise

$\tilde{\Delta}$ 是 Δ （signal）与 random noise 的混合

$\Delta S_{old} \cdot S_{old}^T + W_{new} S_{old} S_{old}^T$

$= (W_{old} - W_{new}) \cdot S_{old} \cdot S_{old}^T + W_{new} \cdots$

$= W_{old} \cdot S_{old} \cdot S_{old}^T$

$$\widehat{\boldsymbol{\Delta}} = \underbrace{\boldsymbol{\Delta}}_{\text{signal}} \mathbf{S}_{\text{old}} \mathbf{S}_{\text{old}}^T + \underbrace{\mathbf{W}_{\text{new}} \left(\mathbf{S}_{\text{old}} \mathbf{S}_{\text{old}}^T - \mathbf{S}_{\text{new}} \mathbf{S}_{\text{new}}^T \right)}_{\text{zero-mean noise}} . \qquad (3)$$

magnitude of W > Δ ==> noise >signal

因此使用 $\tilde{\Delta}$是无效的

Defending the gradient matching attack

Attack 基于受害者的 gradient Δ_i 以及 model parameters W

gradient matching attack 就是去找一个 data 能够满足 gradient 也是 Δ_i

为了得到Δ_i， 攻击者必须知道 $\Delta = \sum_{i=1}^{m} \Delta_i$

DBCL 中，没有 client 知道 Δ

如果使用$\tilde{\Delta}$ 因为无偏估计来代替 Δ，下面证明这是不可行的

Theorem 1. *Let $\mathbf{S}_{old}$ and $\mathbf{S}_{new}$ be $d_{in} \times s$ CountSketch matrices and $s < d_{in}$. Then*

$$\mathbb{E}\|\widehat{\mathbf{\Delta}} - \mathbf{\Delta}\|_F^2 = \Omega\left(\frac{d_{in}}{s}\right) \cdot \left(\|\mathbf{W}_{old}\|_F^2 + \|\mathbf{W}_{new}\|_F^2\right).$$

magnitude of Δ < W， Theorem 1保证了使用 $\tilde{\Delta}$不会好于 random guessing.

Defending the property inference attack (PIA)

Attacker 使用 linear model parameterized V

input features: $\Delta - A$，A是一个固定的matrix

- 真实prediction: $Y = (\Delta - A)V^T$

使用 $\tilde{\Delta}$估计Δ

- 近似prediction：$\tilde{Y} = (\tilde{\Delta} - A)V^T$

$$\|\widehat{\mathbf{Y}} - \mathbf{Y}\|_F^2 = \|\widehat{\mathbf{\Delta}}\mathbf{V}^T - \mathbf{\Delta}\mathbf{V}^T\|_F^2 \text{ is very big}$$

Theorem 2. *Let $\mathbf{S}_{old}$ and $\mathbf{S}_{new}$ be $d_{in} \times s$ CountSketch matrices and $s < d_{in}$. Let w_{pq} be the (p, q)-th entry of $\mathbf{W}_{old} \in \mathbb{R}^{d_{out} \times d_{in}}$ and $\tilde{w}_{pq}$ be the (p, q)-th entry of $\mathbf{W}_{new} \in \mathbb{R}^{d_{out} \times d_{in}}$. Let $\mathbf{V}$ be any $r \times d_{in}$ matrix and v_{pq} be the (p, q)-th entry of $\mathbf{V}$. Then*

$$\mathbb{E}\|\widehat{\mathbf{\Delta}}\mathbf{V}^T - \mathbf{\Delta}\mathbf{V}^T\|_F^2 = \frac{1}{s}\sum_{i=1}^{d_{out}}\sum_{j=1}^{r}\sum_{k\neq l}$$

$$\Big(w_{ik}^2 v_{jl}^2 + w_{ik}v_{jk}w_{il}v_{jl} + \tilde{w}_{ik}^2 v_{jl}^2 + \tilde{w}_{ik}v_{jk}\tilde{w}_{il}v_{jl}\Big).$$

Corollary 3. *Let $\mathbf{S}$ be a $d_{in} \times s$ CountSketch matrix and $s < d_{in}$. Assume that the entries of $\mathbf{W}_{old}$ are IID and that the entries of $\mathbf{V}$ are also IID. Then*

$$\mathbb{E}\|\widehat{\mathbf{\Delta}}\mathbf{V}^T - \mathbf{\Delta}\mathbf{V}^T\|_F^2 = \Omega\Big(\frac{d_{in}}{s}\Big) \cdot \|\mathbf{W}_{old}\mathbf{V}^T\|_F^2.$$

The magnitude of Δ 小于 W

因此 $||WV^T||_F^2$显著大于$||\Delta V^T||_F^2$

即 $E||\tilde{V}^T - \Delta V^T||_F^2$显著大于 $||\Delta V^T||_F^2$

表明使用 $\tilde{\Delta}$不会比随机猜测好

5.3 Understanding DBCL from Optimization

Perspective

Generalized linear model:

应用**Sketching** 后：

$$\underset{\mathbf{w}}{\operatorname{argmin}} \left\{\tilde{f}(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{S}}\Big[\frac{1}{n}\sum_{j=1}^{n}\ell\big(\mathbf{x}_i^T\mathbf{S}\mathbf{S}^T\mathbf{w},\, y_j\big)\Big]\right\}. \qquad (5)$$

如果S 是 uniform sampling matrix，则 (5) 式类似于 dropout.

因为 dropout == adaptive regularization == random CountSketch ==> uniform sampling ==> 不会损失 prediction accuracy.

6. Experiments

证明：

（1）DBCL不会损失 test accuracy.

（2）DBCL不会过多增加 communication cost

（3）DBCL可以防御 client-side gradient-based attacks.

6.1 Experiment Setting

- MNIST

28×28

Training: 60,000 images

Test: 10,000 images.

- CIFAR-10

$32 \times 32 \times 3$

Training: 50,000 images

Test: 10,000 images.

- Labeled Faces In the Wild(LFW)

$64 \times 47 \times 3$

13,233 faces of 5749 individuals.

6.2 Accuracy and Efficiency

MNIST classification

（1）mulitlayer perceptron(MLP) : 3 dense layers

（2）convolutional neural network(CNN): 2 convolutional layers and 2 dense layers.

使用 Federated Averaging (FedAvg)训练

Sketching 应用到所有的 dense and convolutional layers，除了 output layer.

设置 s = $d_{in}/2$，因此，per-communication word complexity 减半

1. test accuracy 没有影响
2. communication rounds 增加不是很多
3. per-communication word complexity减小

Table 1. Experiments on MNIST. The table shows the rounds of communications for attaining the test accuracy. Here, c is the participation ratio of FedAvg, that is, in each round, only a fraction of clients participate in the training.

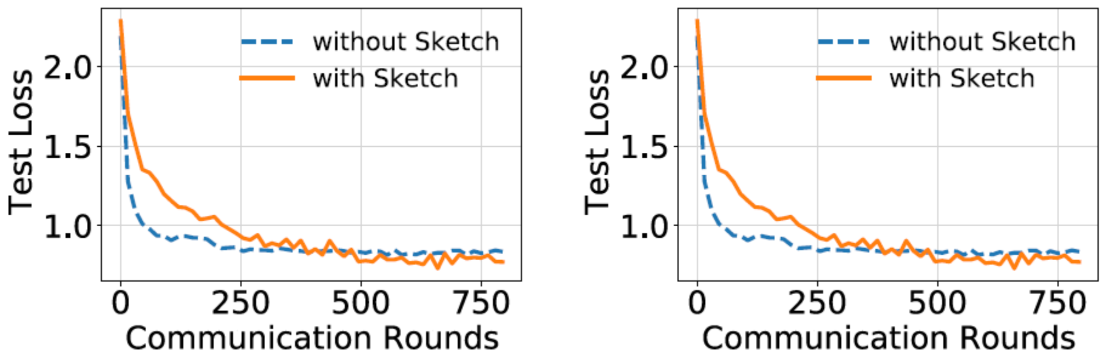
Models	Accuracy	Communication Rounds				
		$c = 1\%$	$c = 10\%$	$c = 20\%$	$c = 50\%$	$c = 100\%$
MLP	0.97	222	96	84	83	82
MLP-Sketch	0.97	572	322	308	298	287
CNN	0.99	462	309	97	91	31
CNN-Sketch	0.99	636	176	189	170	174

CIFAR-10 classification

CNN: 3 convolutional layers and 2 dense layers.

同样使用FedAvg训练CNN

使用Sketching 不但没有损失 test accuracy，反而提升了 test accuracy.



(a) Test loss. (b) Test Accuracy.

Figure 1. Experiment on CIFAR-10 dataset. The test accuracy do not match the state-of-the-art because the CNN is small and we do not use advanced tricks; we follow the settings of the seminal work (McMahan et al., 2017).

开创性

Binary classification on imbalanced data

binary classification experiments

LFW dataset for gender prediction

model is trained by distributed SGD

dataset是 class-imbalanced, male 更多于 females

使用 **ROC curves** 进行评估：

standard CNN 与 sketched one 的 ROC curves 几乎是相同的。

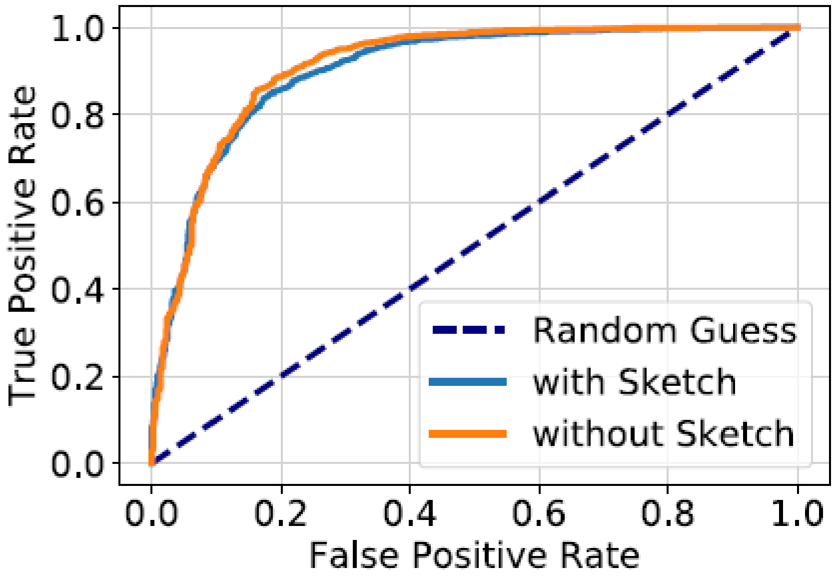


Figure 2. Gender classification on the LWF dataset. LFW

6.3. Defending Gradient-Based Attacks

Gradient estimation

研究两种估计方法：

$$\text{Option I:} \quad \widehat{\boldsymbol{\Delta}} = \mathbf{W}_{\text{old}}\mathbf{S}_{\text{old}}\mathbf{S}_{\text{old}}^{T} - \mathbf{W}_{\text{new}}\mathbf{S}_{\text{new}}\mathbf{S}_{\text{new}}^{T},$$

$$\text{Option II:} \quad \widehat{\boldsymbol{\Delta}} = \mathbf{W}_{\text{old}}\mathbf{S}_{\text{old}}\mathbf{S}_{\text{old}}^{\dagger} - \mathbf{W}_{\text{new}}\mathbf{S}_{\text{new}}\mathbf{S}_{\text{new}}^{\dagger}.$$

$A^{\dagger}$表示Moore-Penrose inverse of matrix A

$\Delta = W_{old} - W_{new}$表示真实的 updating direction

评估方法：

$$\text{The } \ell_2\text{-norm error:} \quad \left\|\mathrm{vec}\left(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\right)\right\|_2 / \left\|\mathrm{vec}\left(\boldsymbol{\Delta}\right)\right\|_2,$$

$$\text{Cosine similarity:} \quad \left\langle \mathrm{vec}\left(\hat{\boldsymbol{\Delta}}\right), \mathrm{vec}\left(\boldsymbol{\Delta}\right) \right\rangle.$$

余弦

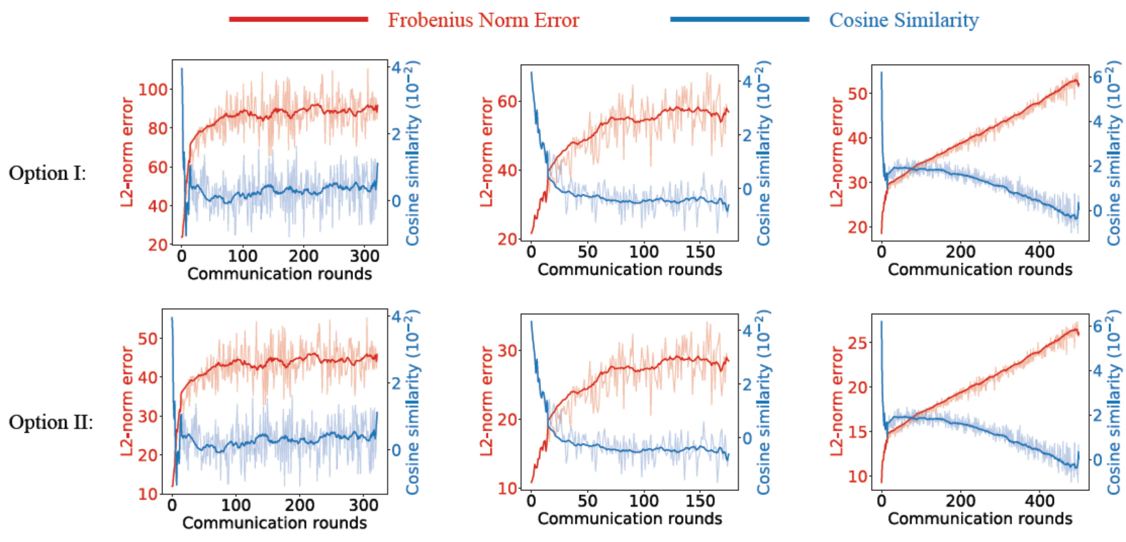
如果$\tilde{\Delta}$与Δ相差较多，则l_2 error大，cosine similarity 小

实验表明，$\tilde{\Delta}$与Δ相差较大，这表明DBCL defense生效

当 Δ减小的时候，$\tilde{\Delta}$主要受到 noise 的影响

$$\hat{\Delta} = \Delta S_{old} \cdot S_{old}^T + W_{new}\left(S_{old} S_{old}^T - S_{new} S_{new}^T\right)$$

随着 communication rounds 的增加，算法趋于收敛，Δ减小，**error** 增加，实验结果如下，也验证了理论：



(a) MLP trained on MNIST. (b) CNN trained on MNIST. (c) CNN trained on CIFAR-10.

Figure 3. The x-axis is the communication rounds. The y-axes are Frobenius norm errors (red) and cosine similarities (blue). The figures show that the estimated gradient, $\hat{\boldsymbol{\Delta}}$, is far from the true gradient, $\boldsymbol{\Delta}$, which means our defense works.

Defending the property inference attack (PIA)

在 LFW dataset上面的 gender classification。

Attacker 尝试找到一个 batch 的 photos，受害者的私有数据是否包括 Africans.

实验使用的是 one server, two clients.

- one client be the attacker

1. Without sketching, AUC = 1.0, 表示 attacker 总可以成功

2. With sketching, AUC = 0.5，表示性能接近于 random guess.

- the server be the attacker

1. Without sketching, AUC = 1.0
2. With sketching, AUC = 0.726, 使得 server-side 的 attack 不是那么有效

Defending the gradient-matching attack

Attack 尝试使用 model parameters and gradient 恢复 victim 的数据

尝试找到 batch of images，使得其 gradient 与观察到的 victim 的gradient match

- Without sketching, gradient-matching attack 可以很容易恢复images.
- With sketching,不管是 client-side 或者 server-side 的attacker，恢复的 images都是类似random noise.

7. Related Work

Cryptography approaches

secure aggregation, homomorphic encryption, Yao's grabled circuit protocol 同样可以提升安全性

但是会降低 accuracy and efficiency，且 tuning and development 是 nontrival.

研究表明，matrix sketching 有与 injecting random noise 相同的性质

DBCL 基于 matrix sketching 与 dropout training 的联系

8. Conclusions

DBCL（Double Blind Collaborative Learning）防御了 gradient-based attacks, 这是最常见最普遍的 privacy inference methods.

- DBCL 防御了基于client发起的 gradient-based attacks
- DBCL 只能减弱 server 发起的attacks
- DBCL 不会损失 test accuracy
- DBCL 不会过多增加 training 的开销
- DBCL 容易使用不需要额外的 tuning
- future work：结合 DBCL 和 cryptographic methods，来达到对 client and server 的双重防护