多媒体内容的隐私保护

程 健 研究员

jcheng@nlpr.ia.ac.cn

中国科学院自动化研究所

2022.10.31

大 纲

- 隐私保护简介
- 部署模型的隐私保护
- 模型训练过程的隐私保护
- 总结与展望

大 纲

- **隐私保护简介**
- 部署模型的隐私保护
- 模型训练过程的隐私保护
- 总结与展望

AI时代的隐私泄露问题

案例一：中国人脸识别隐私问题第一案

- 2019年4月，浙江理工大学郭老师花费1360元购买了杭州动物世界的年卡。
- 2019年10月，园区向郭老师发短信称年卡系统升级为人脸识别入园，未注册人脸识别的用户将无法正常入园。
- 郭老师认为，面部特征属于个人隐私信息，园区在未经其同意的情况下，通过升级年卡系统强制收集个人生物识别信息，违反了《消费者权益保护法》等法律规定。
- 2019年10月，郭老师将杭州动物世界起诉至法院，要求对方退还卡费并承担诉讼费。
- 2020年11月，杭州法院作出一审判决。判决结果为：动物世界赔偿郭兵利益损失，删除郭老师办理年卡时提交的包括照片在内的面部特征信息。

AI时代的隐私泄露问题

案例二：Netflix竞赛泄露隐私

- 2006年，美国网飞公司（Netflix）为了提高在线推荐的效果，发起一个奖金100万美元的推荐算法竞赛。
- 比赛发布的数据是“经过匿名化处理的”用户影评数据，仅仅保留了每个用户对电影的评分和评分的时间戳。
- 然而，两位德州大学奥斯汀分校的研究人员声称可推断出用户的身份，并在2008发表论文详细描述用的算法。他们使用外部数据（如IMDB）与网飞数据进行“去匿名化”算法的交叉比对，从而确定一定比例的用户身份。
- 2009年，Netflix遭到用户起诉，最终因隐私原因宣布停止该比赛，并付出了900万美元的高额赔偿金。

A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111-125, doi: 10.1109/SP.2008.33.

AI时代的隐私泄露问题

案例三：Facebook用户隐私泄露

2019/1/19

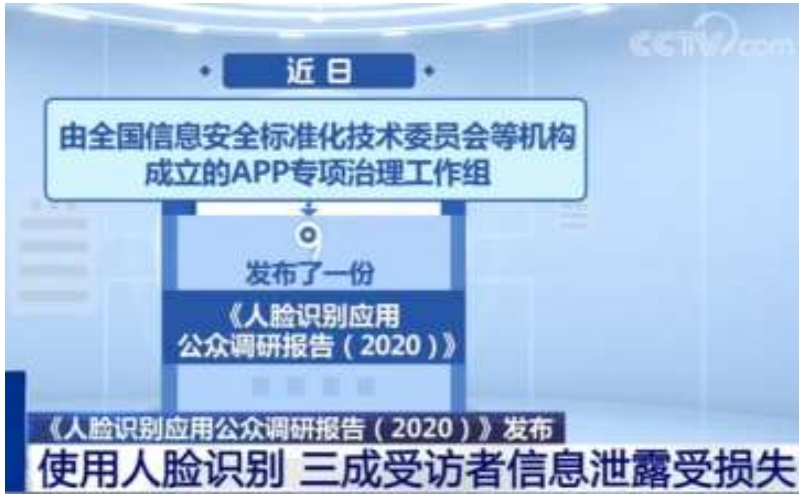
FTC reportedly planning 'record-setting' fine against Facebook for mishandling user data

- In 2012, the FTC fined Google $22.5 million over failing to improve privacy practices – a record for such a punishment.
- The Washington Post says that the fine against Facebook is expected to be "much larger."

- More than 50 million people involved
- UK assessed a £500,000 fine to Facebook
- **the worst single-day market value decrease for a public company in the US**, dropping $120 billion, or 19%

AI时代的隐私泄露问题

- 《人脸识别应用公众调研报告（2020）》



隐私保护的重要性

- 欧盟：《通用数据保护条例》2018.5.25

隐私保护的重要性

- 美国

California Consumer Privacy Act (CCPA)

- Takes effect in 2020
- grants consumers the right to know what information is collected and **whom it is shared with**
- Consumers will have the option of barring tech companies from selling their data
- Provides some of the strongest **regulations in the USA.**

隐私保护的重要性

- 中国

第十三届全国人民代表大会常务委员会第二十九次会议通过《中华人民共和国数据安全法》，自2021年9月1日起施行

为了规范数据处理活动，保障数据安全，促进数据开发利用，保护个人、组织的合法权益，维护国家主权、安全和发展利益

多媒体分析需要隐私保护

- 媒体内容分析尤其是深度学习严重依赖于数据。
- 许多使用媒体内容分析情景中，数据是需要被保护的：
  - 模型学习包含训练数据的信息，而数据所有者不希望他的数据被公开。
  - 模型用户不想向模型所有者透露他的私有数据。
  - 某些数据被公开是不合法的，如医疗信息、银行用户数据、用户个人信息。

Medical Records

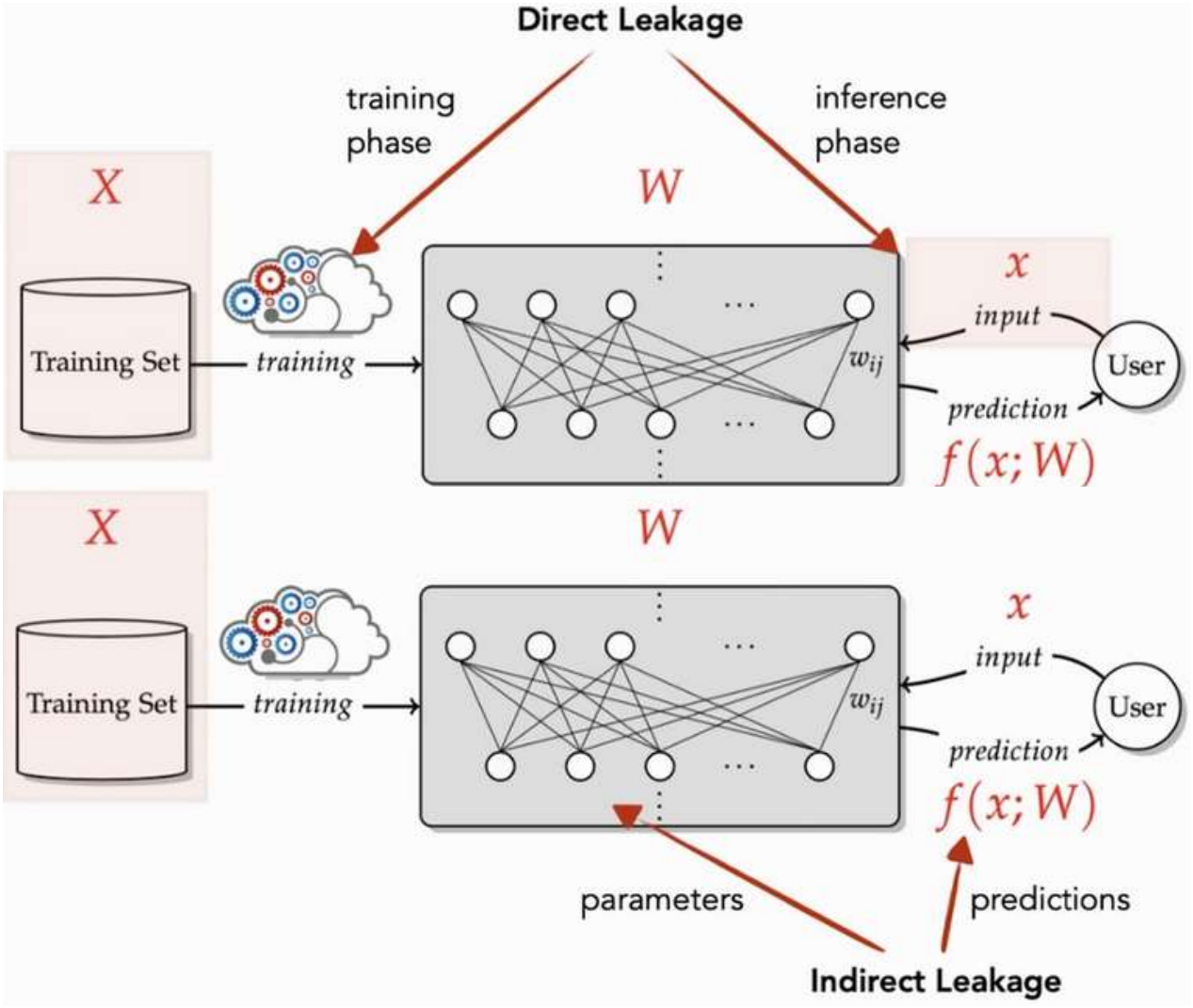
Genetic Data

Search Logs

Google

多媒体分析中的隐私泄露风险

- 模型训练阶段和部署的模型都可能直接或间接地泄露隐私。

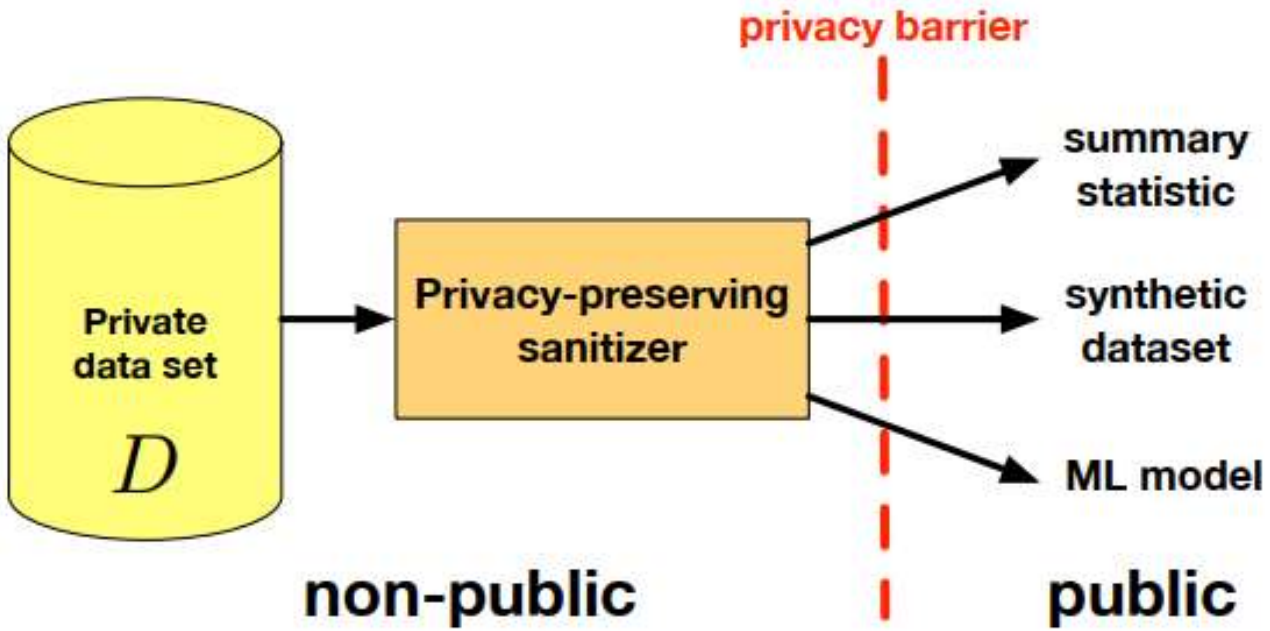


大 纲

- 隐私保护简介
- **部署模型的隐私保护**
- 模型训练过程的隐私保护
- 总结与展望

怎样定义/评估隐私性？

- The Setting



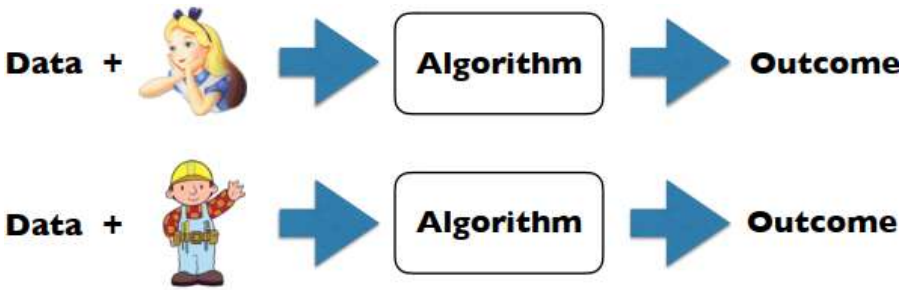
Property of Sanitizer

Aggregate information computable

Individual information protected

差分隐私

- 差分隐私（Differential Privacy, DP，Dwork2006）
  - Motivation



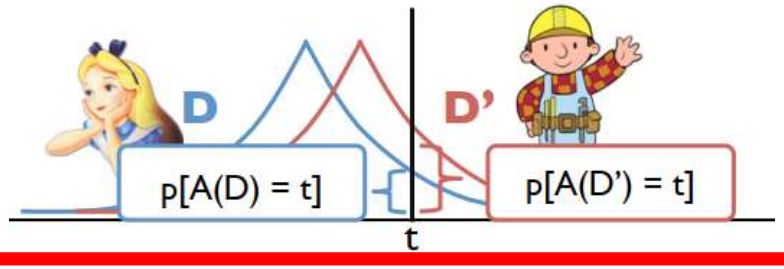
ε：隐私预算

δ：松弛量

Participation of a person does not change outcome

  - Definition

Neighborhood datasets



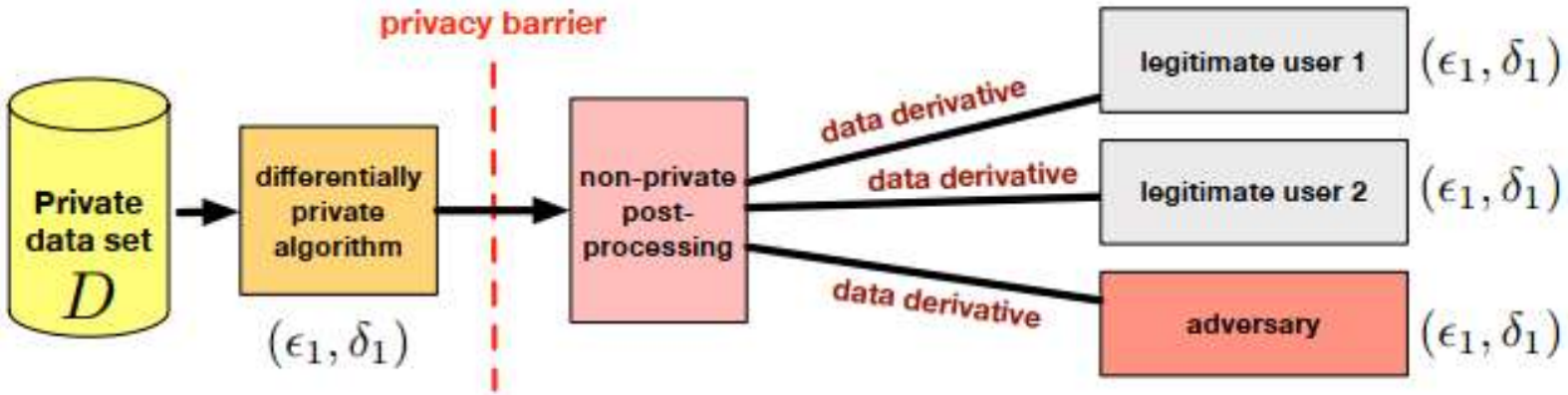
For all D, D' that differ in one person's value,

If A = (ϵ, δ)-differentially private randomized algorithm, then:

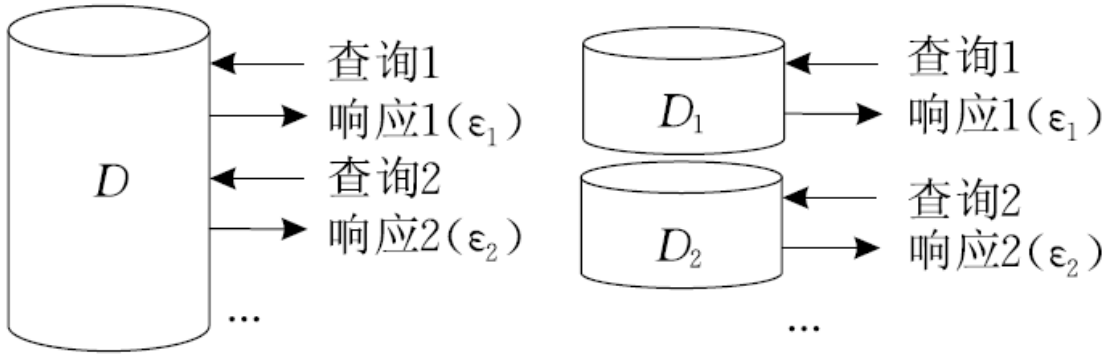
$$\max_{S, \Pr(A(D) \in S) > \delta} \left[\log \frac{\Pr(A(D) \in S) - \delta}{\Pr(A(D') \in S} \right] \leq \epsilon$$

差分隐私的基本性质

- 后处理不变性（Post-processing Invariance）



- 组合性质（Graceful Composition）



(a) $\sum\varepsilon_i$-差分隐私　　(b) $\max(\varepsilon_i)$-差分隐私

序列组合性　　并行组合性

差分隐私的基本实现机制

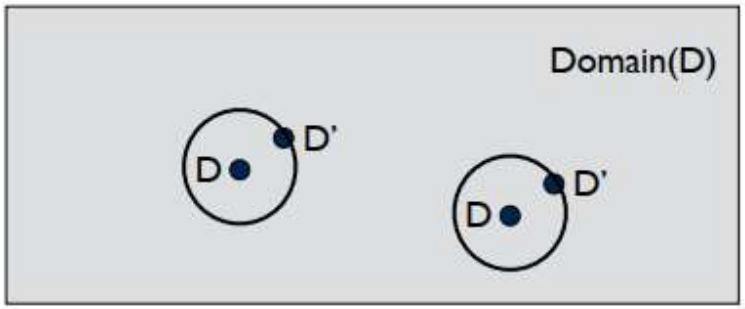
- 全局敏感度（global sensitivity）

Given: A function f, sensitive dataset D

Define: dist(D, D') = #records that D, D' differ by

Global Sensitivity of f:

$$S(f) = \max_{dist(D, D') = 1} |f(D) - f(D')|$$



指述D, D'分别作为数据集输出的差距

- DP实现机制：
  - 拉普拉斯机制（Laplace Mechanism）
  - 高斯机制（Gaussian Mechanism）
  - 指数机制（Exponential Mechanism）

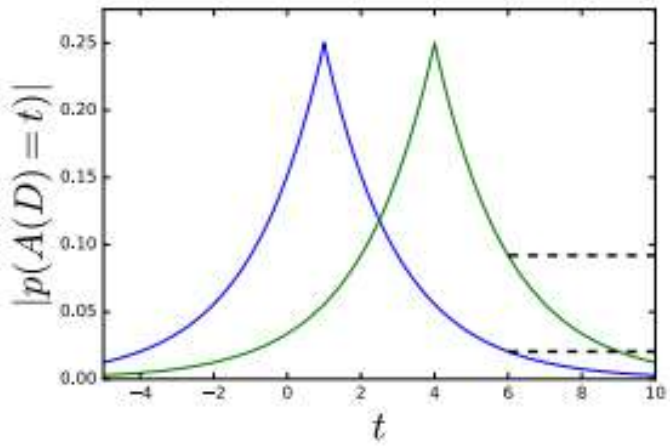
拉普拉斯机制

Global Sensitivity of f is S(f) = $\max_{\text{dist}(D, D') = 1} |f(D) - f(D')|$

Output $f(D) + Z$, where

$$Z \sim \frac{S(f)}{\epsilon}\text{Lap}(0, 1)$$

ϵ-differentially private



Laplace distribution:

$$p(z|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|z-\mu|}{b}\right)$$

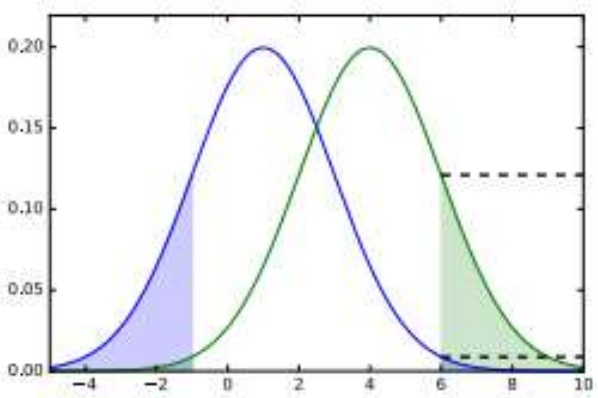
高斯机制

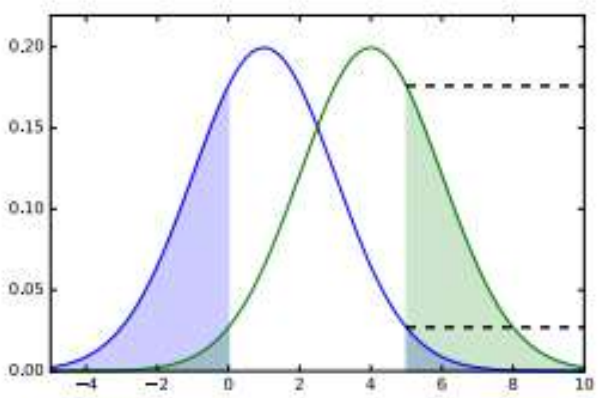
Global Sensitivity of f is S(f) = $\max_{\text{dist}(D, D') = 1} |f(D) - f(D')|$

Output $f(D) + Z$, where

$$Z \sim \frac{S(f)}{\epsilon}\mathcal{N}(0, 2\ln(1.25/\delta))$$

(ϵ, δ)-differentially private





指数机制

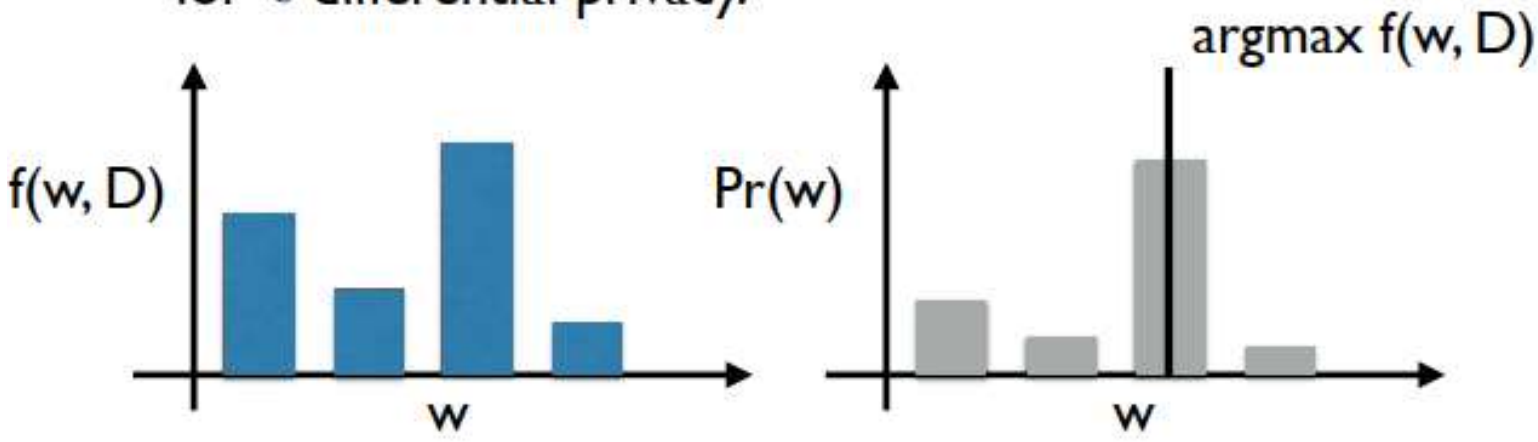
Suppose for any w,

$$|f(w, D) - f(w, D')| \leq S$$

when D and D' differ in 1 record. Sample w from:

$$p(w) \propto e^{\epsilon f(w,D)/2S}$$

for ϵ-differential privacy.



将DP应用于深度学习

- DPSGD：
  - 将高斯机制引入SGD
  - 保护训练数据

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

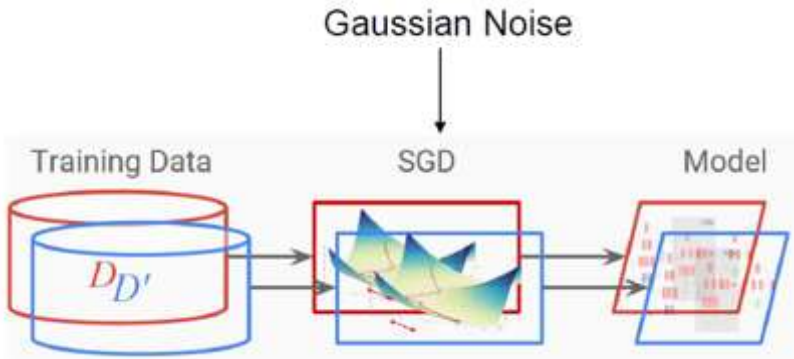
Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.



限制算法全局敏感度

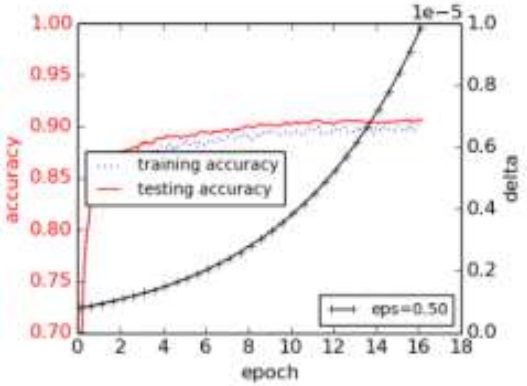
在每个样本产生的梯度上加高斯噪声

Abadi, Martin, et al. "Deep learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016.

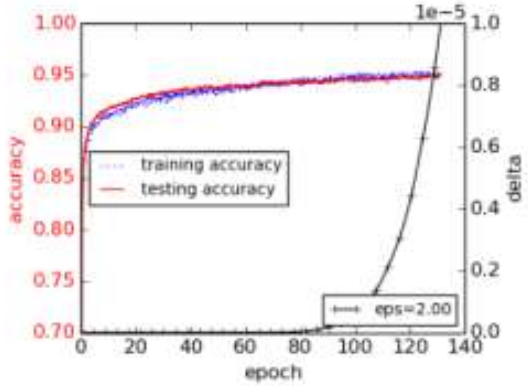
将DP应用于深度学习

- DPSGD保证了隐私保护性，但是会损害模型精度

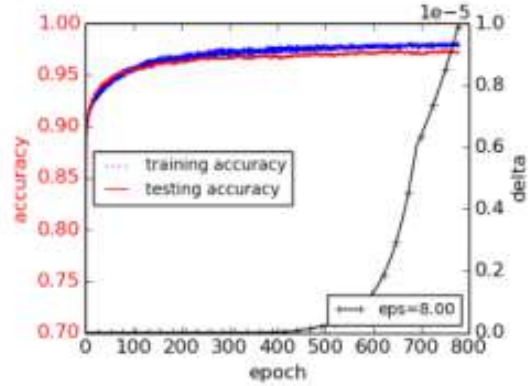
MNIST baseline: 0.99



(1) Large noise

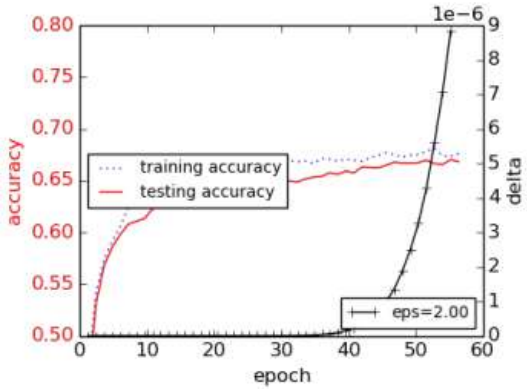


(2) Medium noise

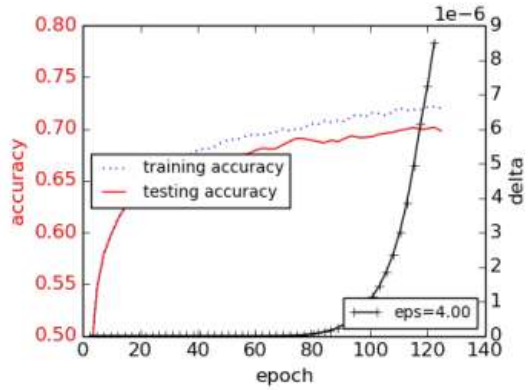


(3) Small noise

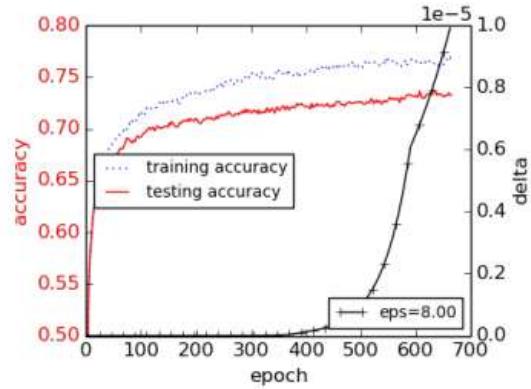
CIFAR baseline: 0.86



(1) $\varepsilon = 2$



(2) $\varepsilon = 4$

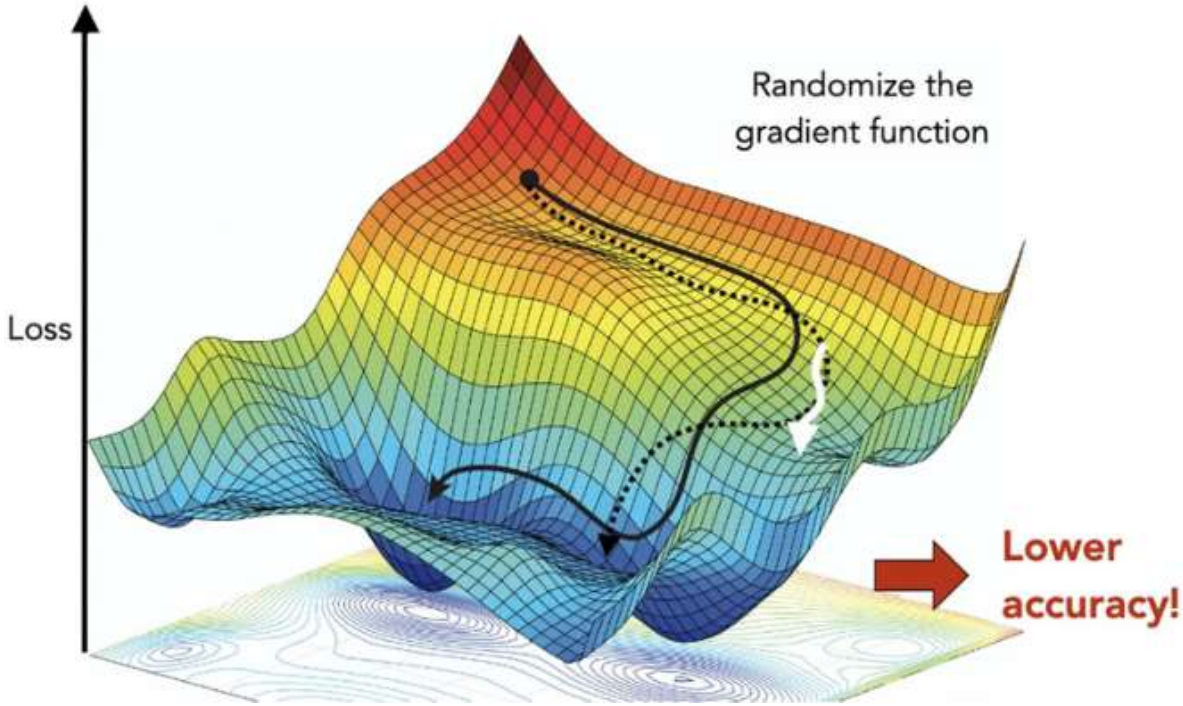


(3) $\varepsilon = 8$

Abadi, Martin, et al. "Deep learning with differential privacy." *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016.

将DP应用于深度学习

- DPSGD损害模型精度原因：加入随机性降低了拟合程度

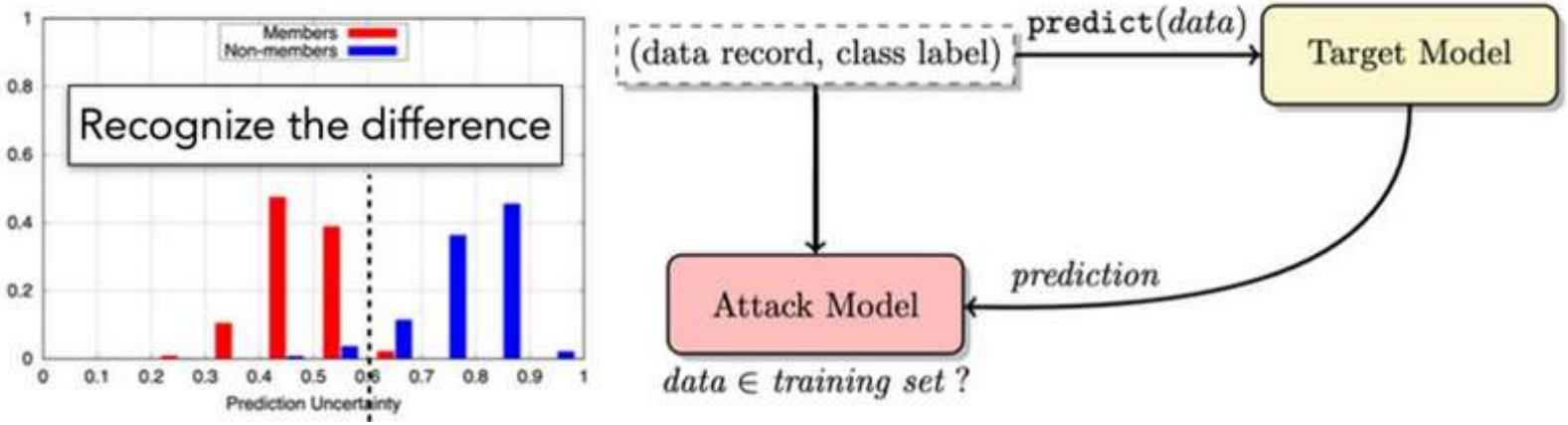


差分隐私总结

- 差分隐私是一种对隐私泄露风险的度量准则，具有良好的理论计算性、组合度量性、后处理性等性质。
- 算法的隐私保护性多是通过向算法中引入随机性实现。
- 差分隐私参数 (ϵ, δ) 的具体值一般通过**理论计算**得到，很难通过实验方法测得。
- 差分隐私提供了算法隐私保护性的**下界**（lower bound）。而在解决具体的隐私泄露问题时，可能还需要使用其他的具体的隐私度量手段评估算法的隐私保护性。

成员推理攻击（MIA）

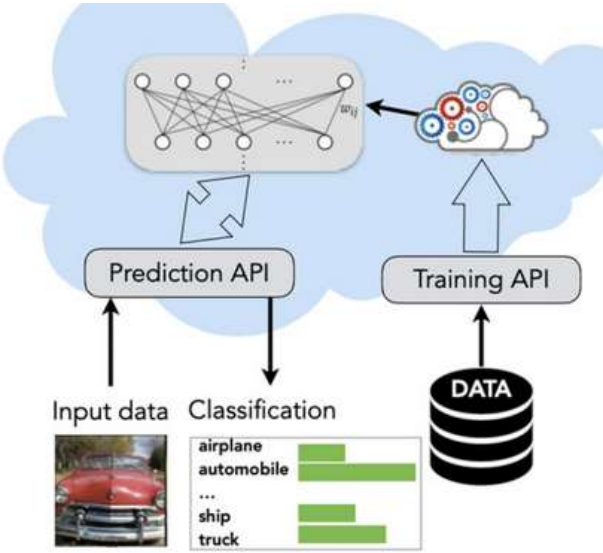
- 成员推理攻击（membership inference attack, MIA）
  - 是一种隐私攻击手段
  - 也是一种经验性的隐私保护性度量指标
    - Indistinguishability game: Can an adversary distinguish between two models that are trained on two neighboring datasets (only one includes data point x)?
      - Membership inference: Given a model, can an adversary infer whether data point x is part of its training set?



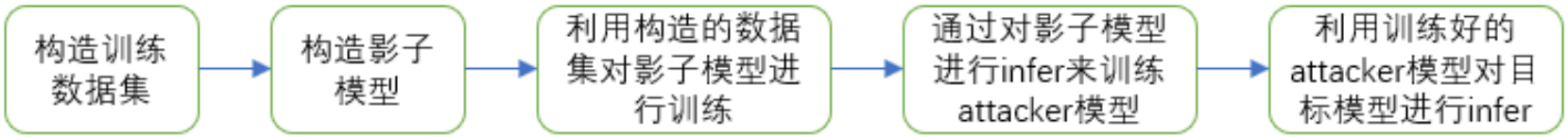
Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.

MIA的应用

- 作为隐私保护性评价指标
  - 与DP不同的是，MIA通常由实验测得，因此可以经验性的测量、比较算法的隐私保护性。

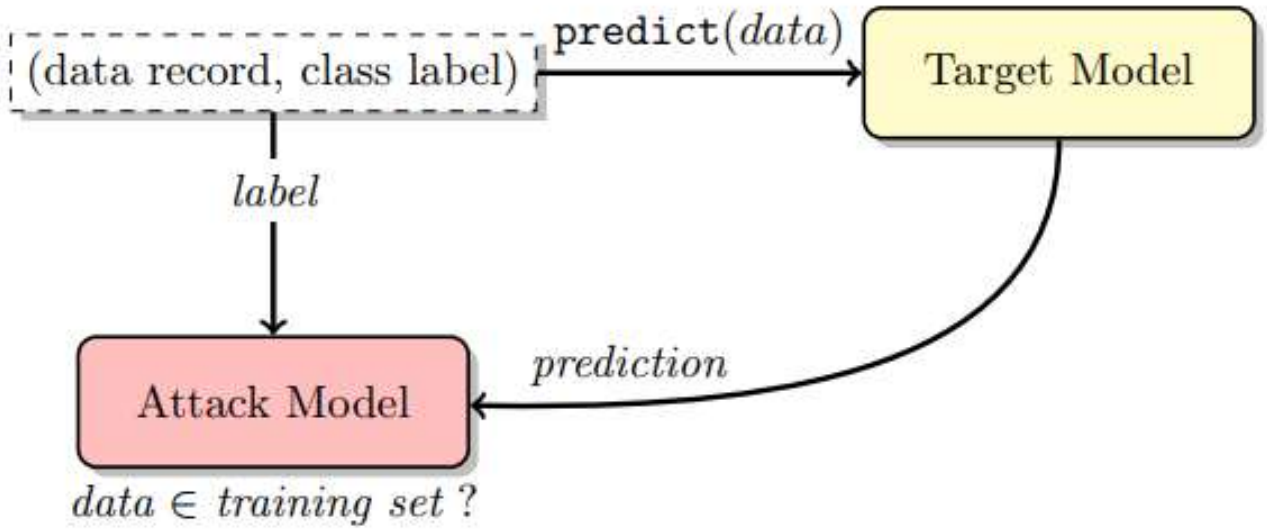


- 作为隐私攻击手段
  - 常用于黑盒（black-box）模型的场景
  - 间接泄露模型的训练信息（数据探测）。例如，某人的医疗记录参与了某个医疗诊断模型的训练，攻击者推断出该记录的成员隶属性，即暴露了此人患有此疾病这一信息。



MIA的实现

- 若已知目标模型的训练集，对于样本(x, y)，假设目标模型的输出为预测向量p。根据该样本其是否在目标模型的训练集中，可以构建MIA样本(p, *in*)或（p, *out*)。
- 将预测向量p作为feature，成员隶属性作为label，可以训练一个二分类模型。即攻击模型是一个二分类模型，其以目标模型的输出作为输入，判断样本是否在目标模型的训练集中。



Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.

MIA的实现

- 在实际中，攻击者通常不知道目标模型的训练集，因此需要构建与原目标模型训练集相似的数据集，在此数据集上训练同种模型（影子模型），用影子模型来构建 MIA样本。

- 直观上，如果目标模型以很高的概率给出了某条样本的类别，那么该样本与目标模型训练集中的数据应该是十分相似的。所以，可以用目标模型本身来构建影子模型的训练数据。

Algorithm 1 Data synthesis using the target model

```
procedure SYNTHESIZE(class : c)
    x ← RANDRECORD()                          ▷ initialize a record randomly
    y*_c ← 0
    j ← 0
    k ← k_max
    for iteration = 1 ··· iter_max do
        y ← f_target(x)                       ▷ query the target model
        if y_c ≥ y*_c then                    ▷ accept the record
            if y_c > conf_min and c = arg max(y) then
                if rand() < y_c then          ▷ sample
                    return x                  ▷ synthetic data
                end if
            end if
            x* ← x
            y*_c ← y_c
            j ← 0
        else
            j ← j + 1
            if j > rej_max then               ▷ many consecutive rejects
                k ← max(k_min, ⌈k/2⌉)
                j ← 0
            end if
        end if
        x ← RANDRECORD(x*, k)                 ▷ randomize k features
    end for
    return ⊥                                  ▷ failed to synthesize
end procedure
```

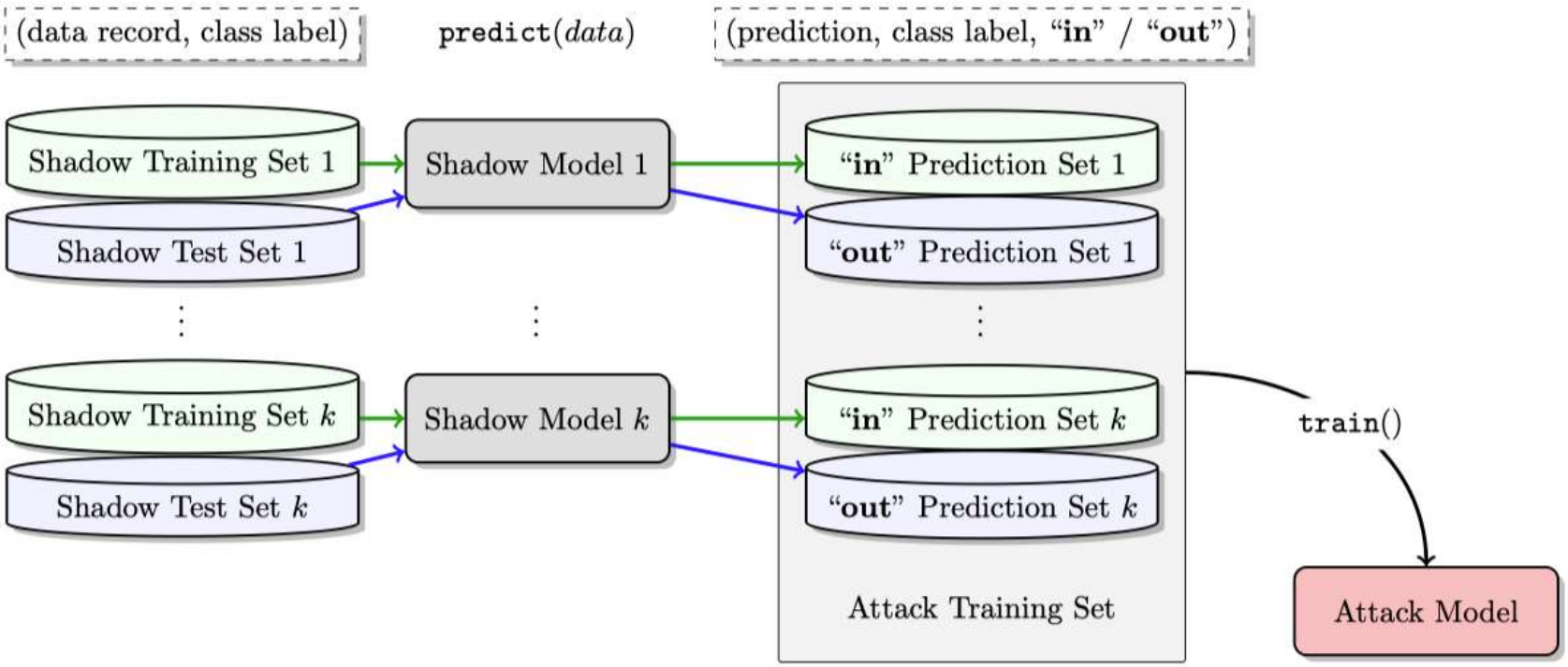
目标M → 构造dataset，训练影子M

→ 影子M生成 MIA dataset

Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.

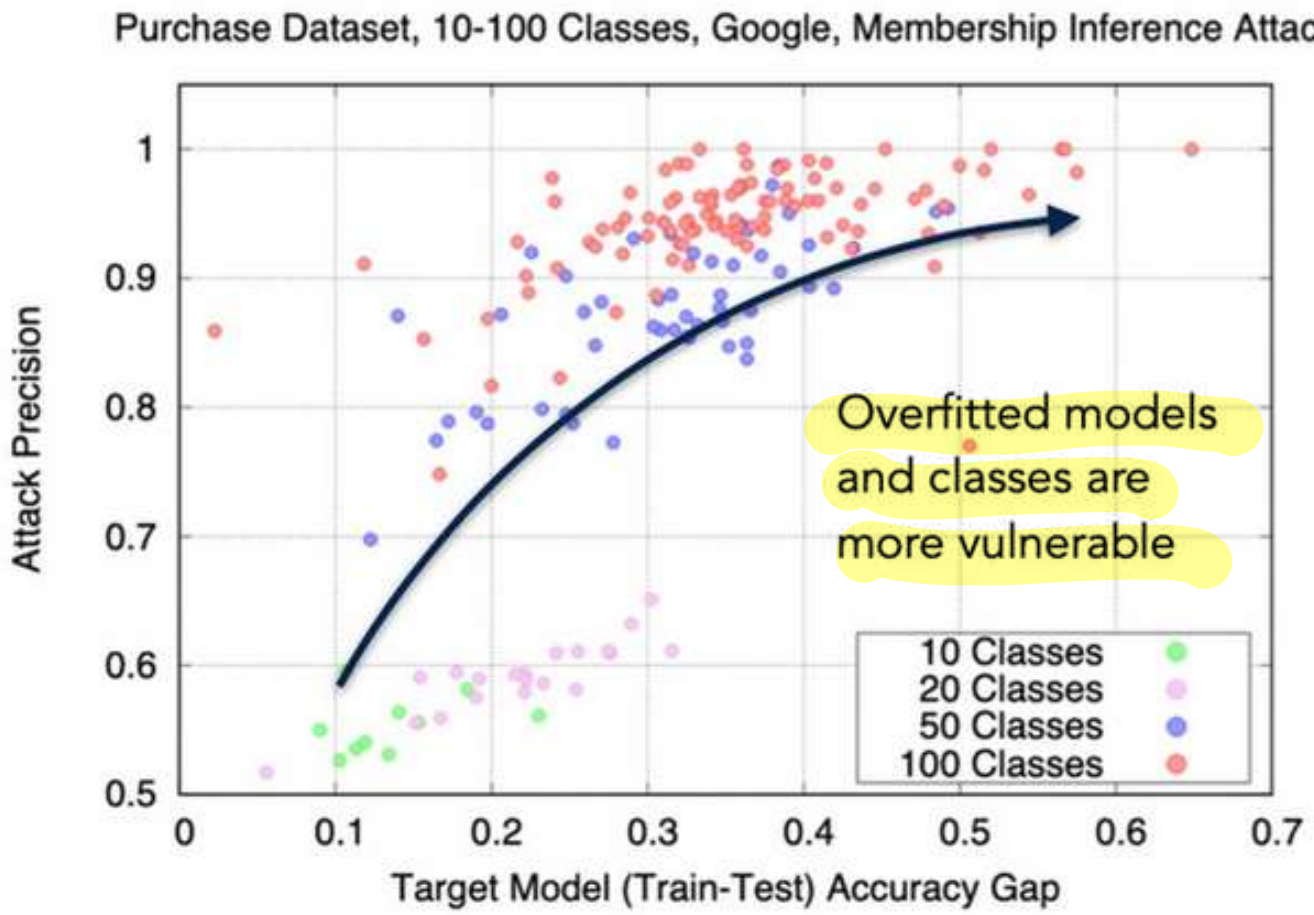
MIA的实现

- 生成影子模型后，构造攻击模型的训练集（训练集的标签，和影子模型的输出结果），并对攻击模型进行训练。
- 实际中可以训练多个影子模型，并使用这些影子模型构建更大的MIA训练集。



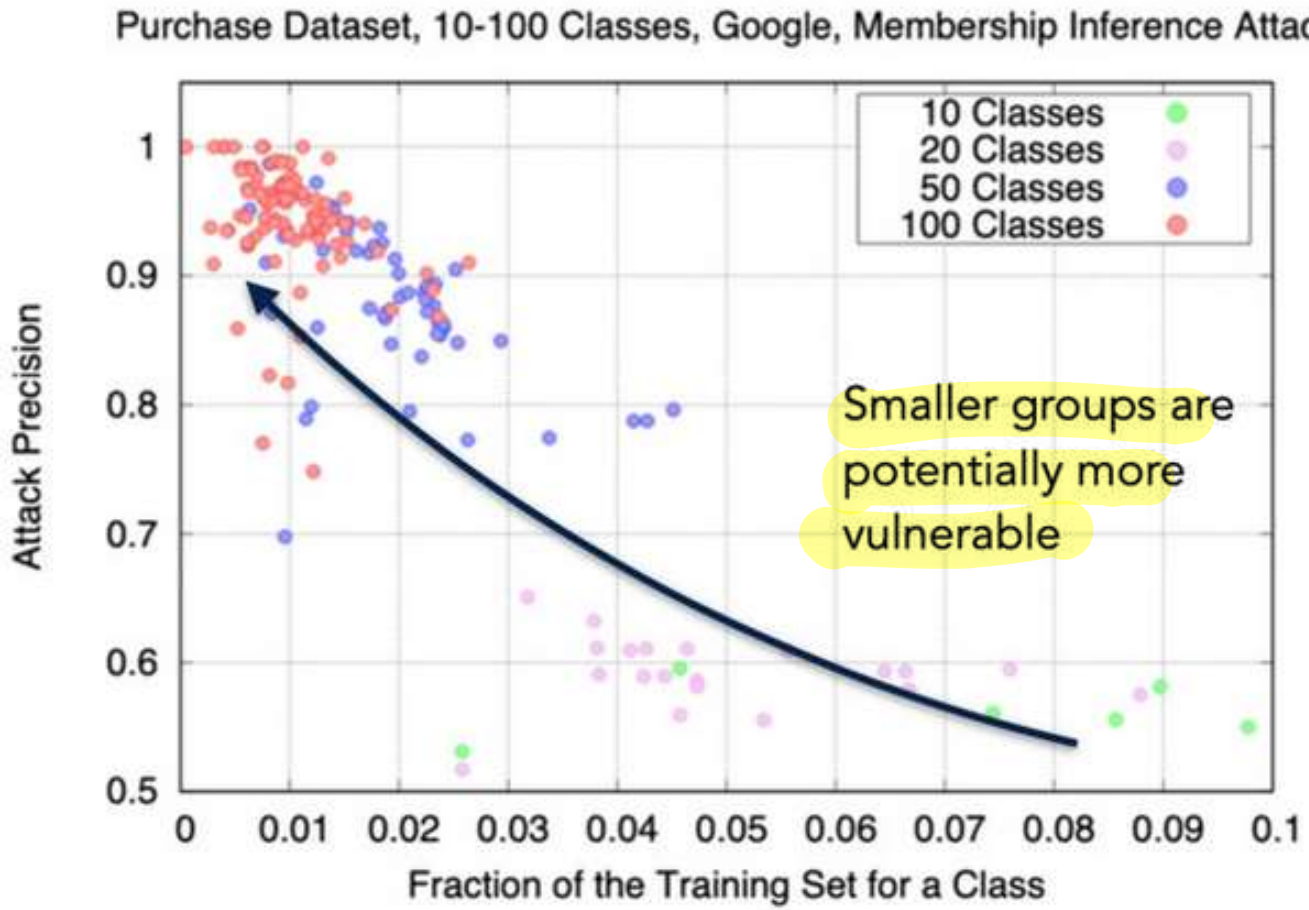
Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.

隐私泄露与过拟合



Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.

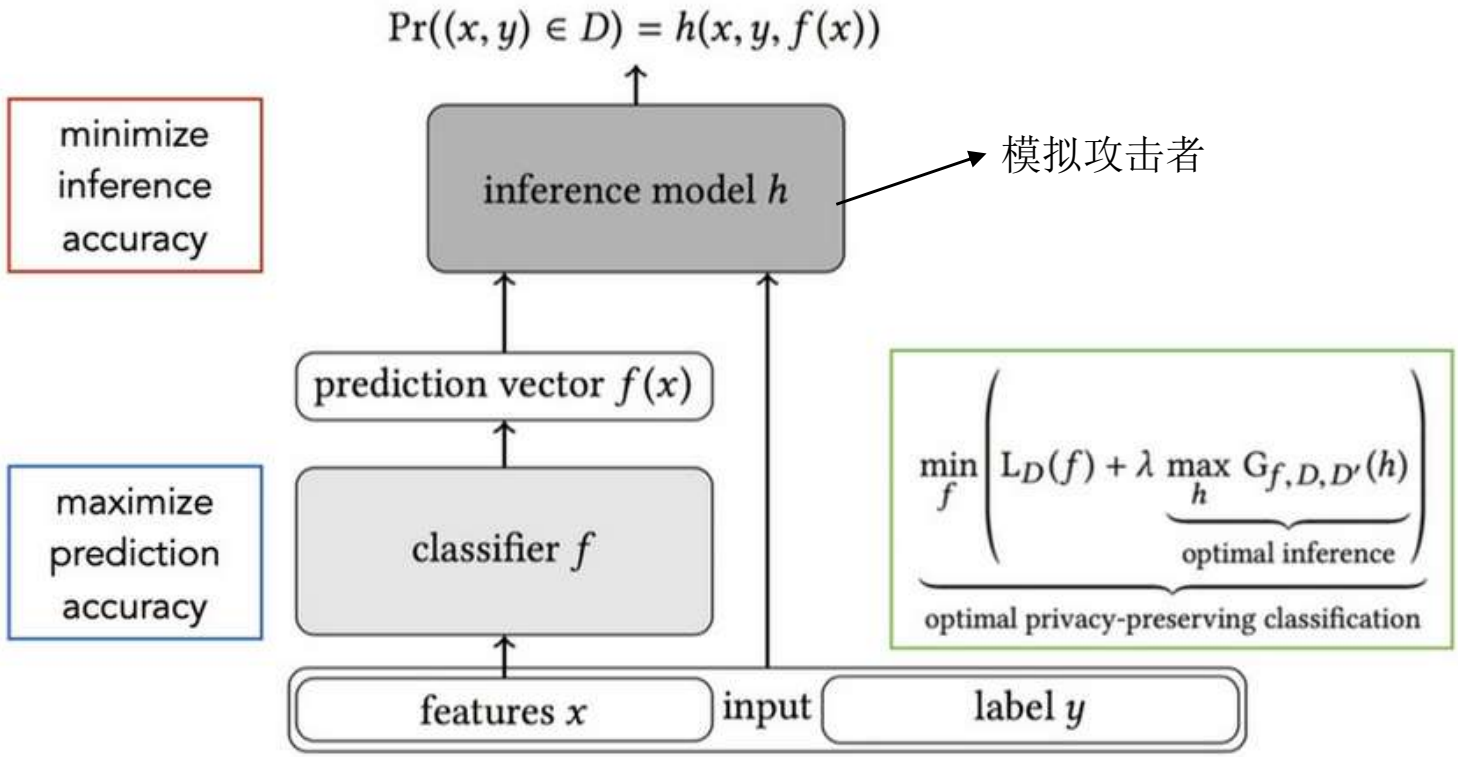
数据分布对隐私的影响



Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.

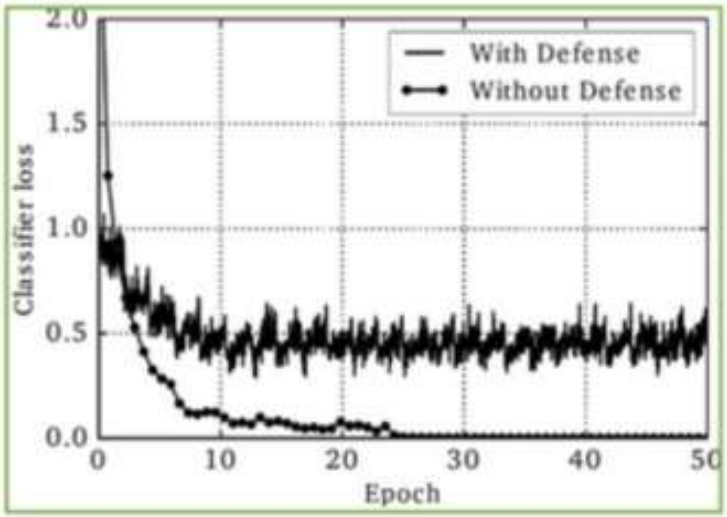
隐私保护性作为学习目标

- 将隐私性加入模型的学习目标
  - 训练过程中加入一个模拟的MIA攻击者，攻击者的目标是最大化攻击收益，模型训练者的目标是减小模型分类损失的同时最小化攻击者收益。



Nasr, M., Shokri, R., & Houmansadr, A. (2018). Machine Learning with Membership Privacy using Adversarial Regularization. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security.

隐私保护性作为学习目标



	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

Smaller gap

Random guess

Nasr, M., Shokri, R., & Houmansadr, A. (2018). Machine Learning with Membership Privacy using Adversarial Regularization. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security.

MIA总结

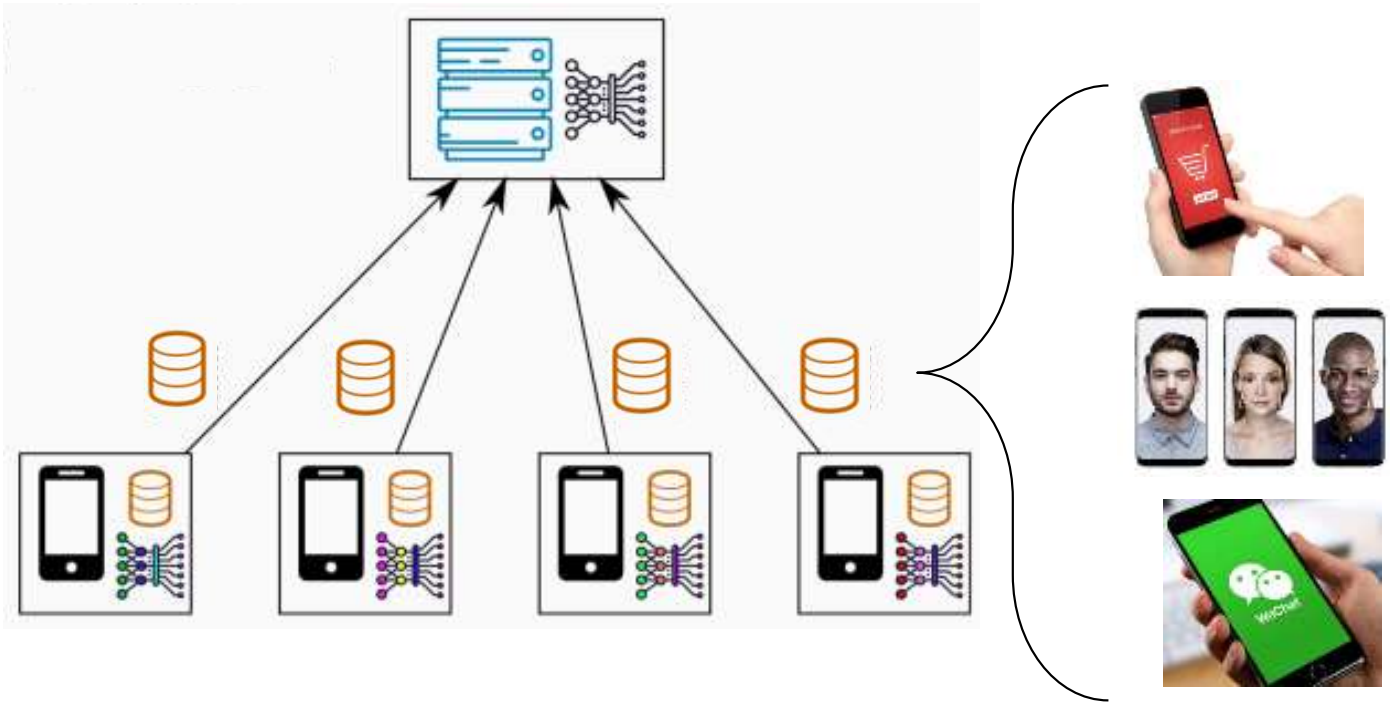
- 与DP相似，MIA可以被用来度量**已经训练好的模型**对训练数据的隐私泄露程度。
- 与DP不同的是，MIA是一种**经验性度量**方法，可以通过实验测得。
- 借助MIA，可以分析模型的过拟合程度、数据分布均匀性等因素与隐私保护性之间的关系。

大 纲

- 隐私保护简介
- 部署模型的隐私保护
- **模型训练过程的隐私保护**
- 总结与展望

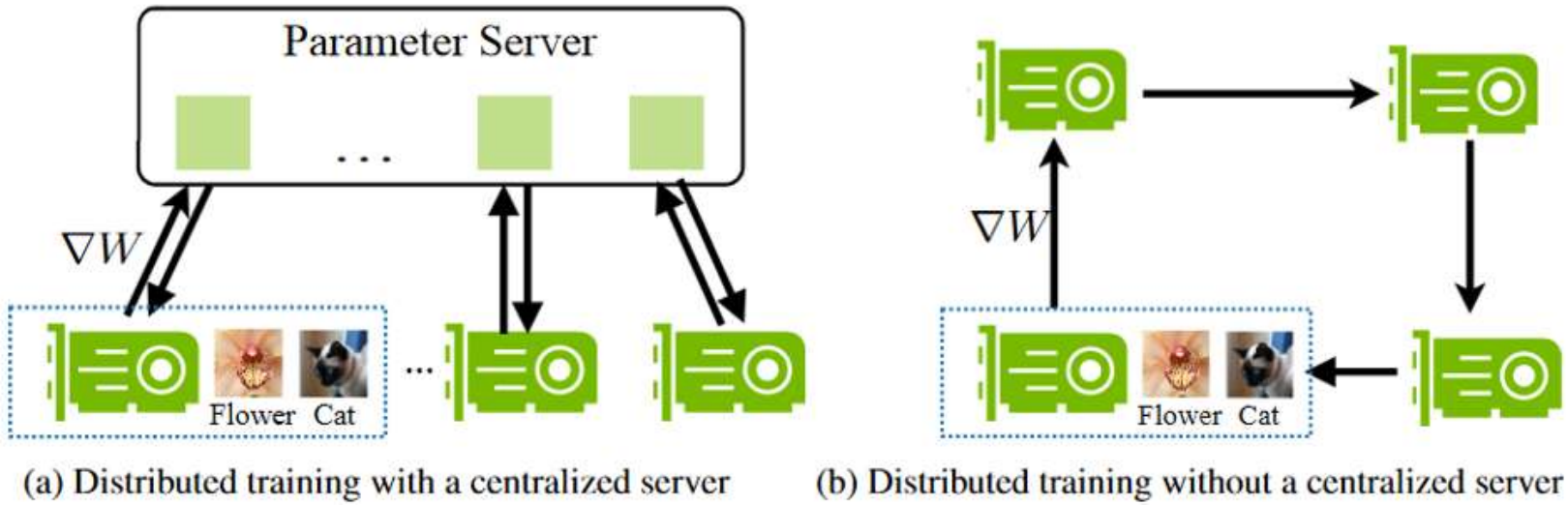
集中学习

- 集中学习（centralized learning）
  - 数据是分散的
  - 用户将本地数据上传到服务器，服务器用收集的数据训练模型
- 问题
  - 服务器收集用户数据是否合法？
  - 潜在的侵犯隐私的风险？



分布式学习

- 分布式学习（distributed learning）
  - 计算在各节点上进行

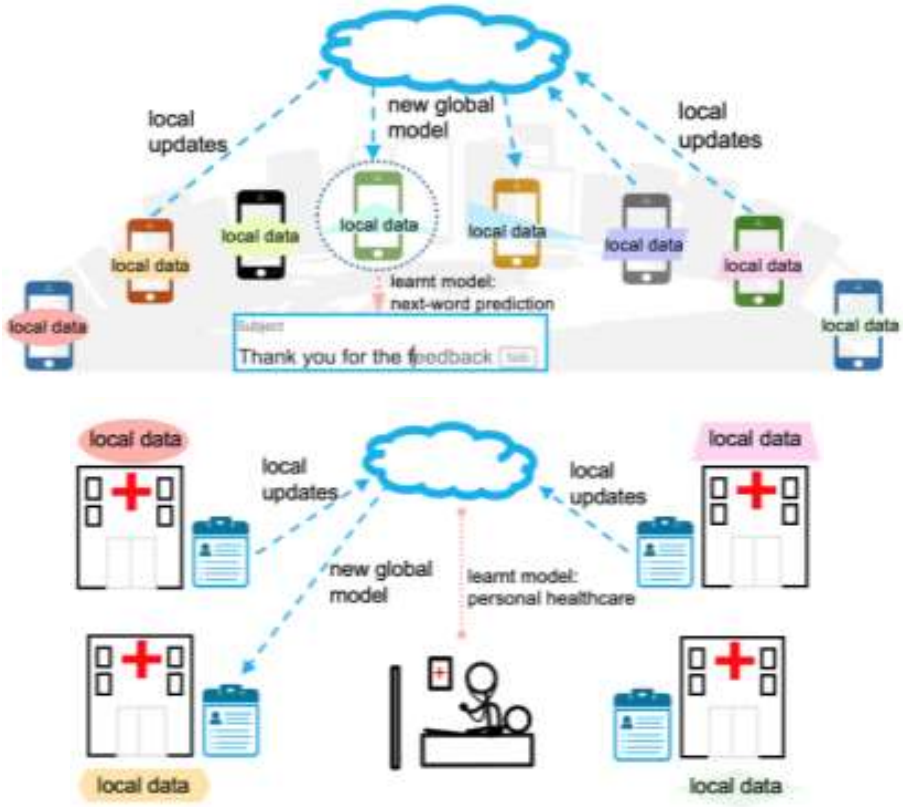


(a) Distributed training with a centralized server　　(b) Distributed training without a centralized server

联邦学习

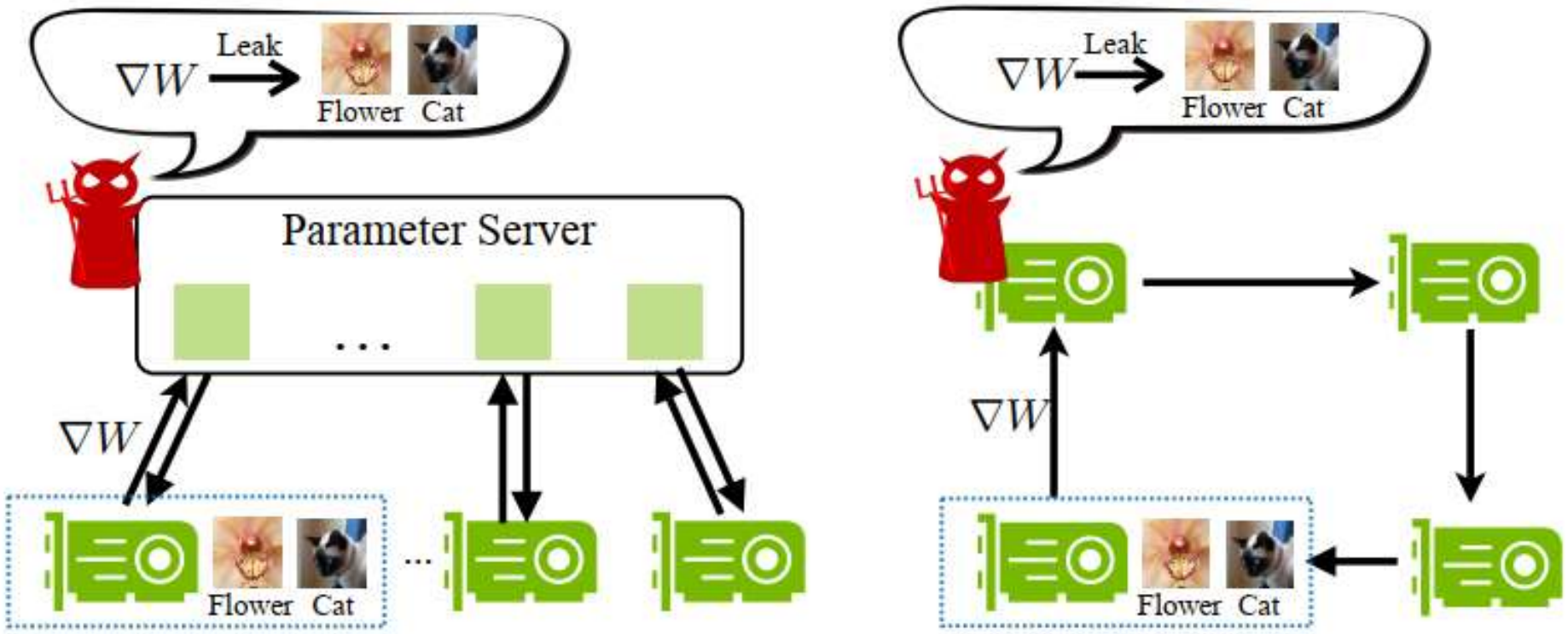
- 联邦学习（federated learning）是一种特殊的分布式机器学习。
- 数据留在本地，任何节点都无法得到所有的数据。

- 与传统分布式学习的相同点
  - 数据放在用户节点
  - 分布式计算
- 与传统分布式学习的区别
  - 用户对于自己的设备和有着控制权。
  - 用户节点不稳定性。
  - 用户节点数据分布异质性。
  - 用户节点负载不平衡。



数据留在本地就是安全的吗？

- 分布式学习/联邦学习训练过程中需要上传梯度/模型更新，研究表明，梯度/模型更新中同样会泄露隐私。

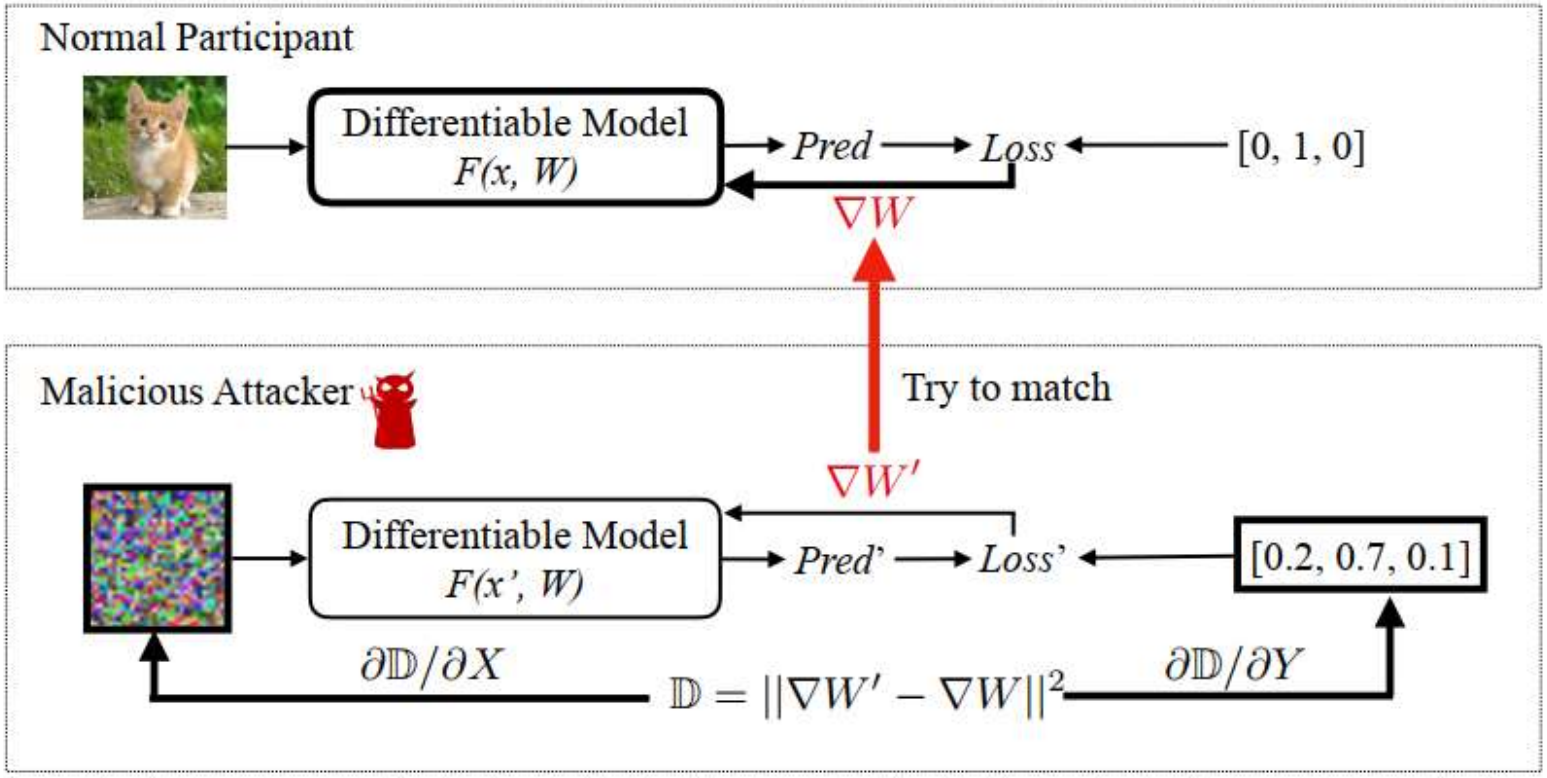


(a) Distributed training with a centralized server　(b) Distributed training without a centralized server

Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

梯度泄露隐私

- Deep Leakage from Gradients (DLG)



Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

梯度泄露隐私

- Objective function

$$\mathbf{x}'^*, \mathbf{y}'^* = \arg\min_{\mathbf{x}', \mathbf{y}'} ||\nabla W' - \nabla W||^2 = \arg\min_{\mathbf{x}', \mathbf{y}'} ||\frac{\partial \ell(F(\mathbf{x}', W), \mathbf{y}')}{\partial W} - \nabla W||^2$$

$$\text{or} \quad \arg\min_{x \in [0,1]^n} 1 - \frac{\langle \nabla_\theta \mathcal{L}_\theta(x, y), \nabla_\theta \mathcal{L}_\theta(x^*, y) \rangle}{||\nabla_\theta \mathcal{L}_\theta(x, y)|| ||\nabla_\theta \mathcal{L}_\theta(x^*, y)||} + \alpha \, \mathrm{TV}(x)$$

- Method

Algorithm 1 Deep Leakage from Gradients.

Input: $F(\mathbf{x}; W)$: Differentiable machine learning model; W: parameter weights; ∇W: gradients calculated by training data

Output: private training data $\mathbf{x}, \mathbf{y}$

procedure DLG(F, W, ∇W)
  $\mathbf{x}'_1 \leftarrow \mathcal{N}(0,1)$, $\mathbf{y}'_1 \leftarrow \mathcal{N}(0,1)$ ▷ Initialize dummy inputs and labels.
  for $i \leftarrow 1$ to n **do**
    $\nabla W'_i \leftarrow \partial \ell(F(\mathbf{x}'_i, W_t), \mathbf{y}'_i)/\partial W_t$ ▷ Compute dummy gradients.
    $\mathbb{D}_i \leftarrow ||\nabla W'_i - \nabla W||^2$
    $\mathbf{x}'_{i+1} \leftarrow \mathbf{x}'_i - \eta \nabla_{\mathbf{x}'_i} \mathbb{D}_i$, $\mathbf{y}'_{i+1} \leftarrow \mathbf{y}'_i - \eta \nabla_{\mathbf{y}'_i} \mathbb{D}_i$ ▷ Update data to match gradients.
  end for
  return $\mathbf{x}'_{n+1}, \mathbf{y}'_{n+1}$
end procedure

Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

DLG效果

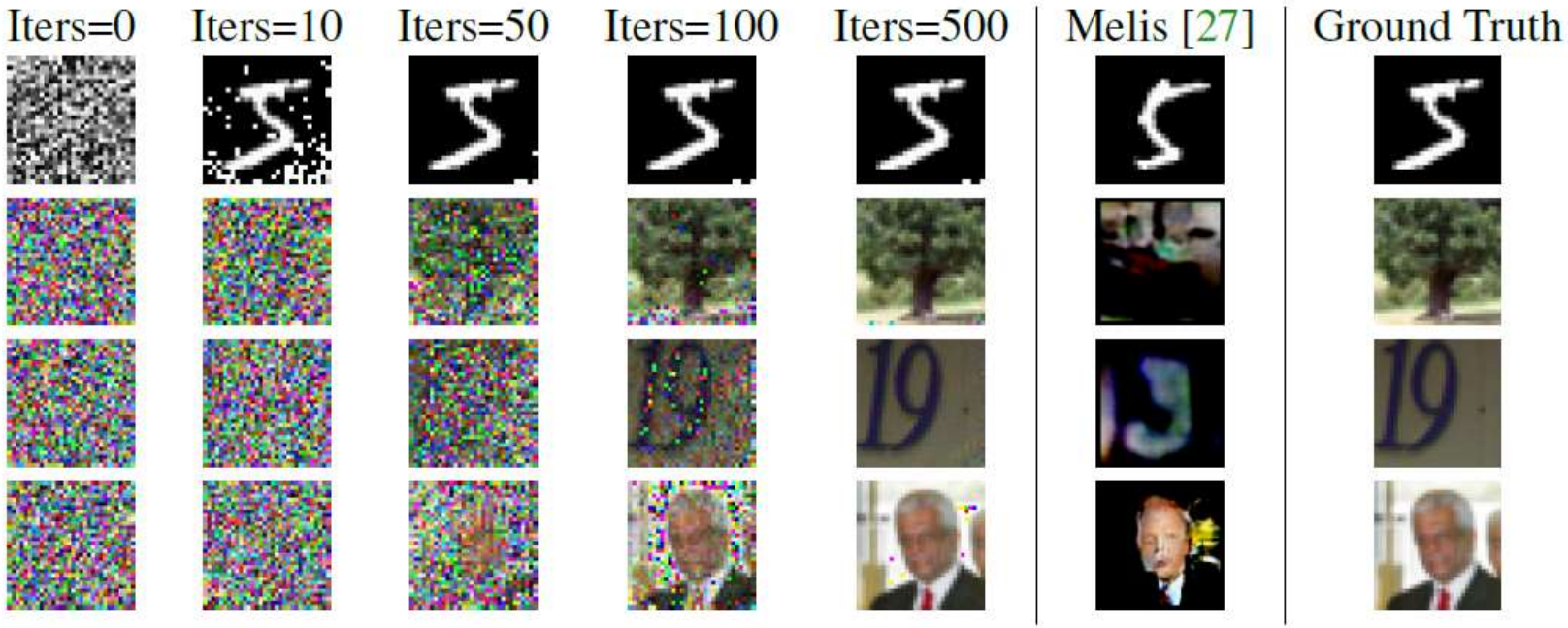


Figure 3: The visualization showing the deep leakage on images from MNIST [22], CIFAR-100 [21], SVHN [28] and LFW [14] respectively. Our algorithm fully recovers the four images while previous work only succeeds on simple images with clean backgrounds.

DLG效果

	Example 1	Example 2	Example 3
Initial Sentence	tilting fill given **less word **itude fine **nton overheard living vegas **vac **vation *f forte **dis cerambycidae ellison **don yards marne **kali	toni **enting asbestos cutler km nail **oof **dation **ori righteous **xie lucan **hot **ery at **tle ordered pa **eit smashing proto	[MASK] **ry toppled **wled major relief dive displaced **lice [CLS] us apps _ **face **bet
Iters = 10	tilting fill given **less full solicitor other ligue shrill living vegas rider treatment carry played sculptures lifelong ellison net yards marne **kali	toni **enting asbestos cutter km nail undefeated **dation hole righteous **xie lucan **hot **ery at **tle ordered pa **eit smashing proto	[MASK] **ry toppled identified major relief gin dive displaced **lice doll us apps _ **face space
Iters = 20	registration , volunteer applications , at student travel application open the ; week of played ; child care will be glare .	we welcome proposals for tutor **ials on either core machine denver softly or topics of emerging importance for machine learning .	one **ry toppled hold major ritual ' dive annual conference days 1924 apps novelist dude space
Iters = 30	registration , volunteer applications , and student travel application open the first week of september . child care will be available .	we welcome proposals for tutor **ials on either core machine learning topics or topics of emerging importance for machine learning .	we invite submissions for the thirty - third annual conference on neural information processing systems .
Original Text	Registration, volunteer applications, and student travel application open the first week of September. Child care will be available.	We welcome proposals for tutorials on either core machine learning topics or topics of emerging importance for machine learning.	We invite submissions for the Thirty-Third Annual Conference on Neural Information Processing Systems.

Table 2: The progress of deep leakage on language tasks.

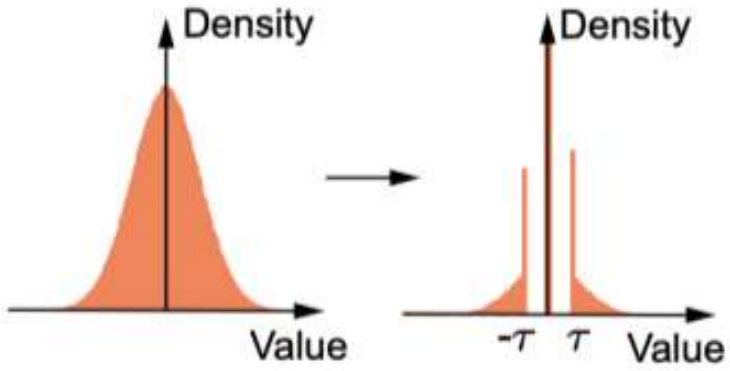
Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

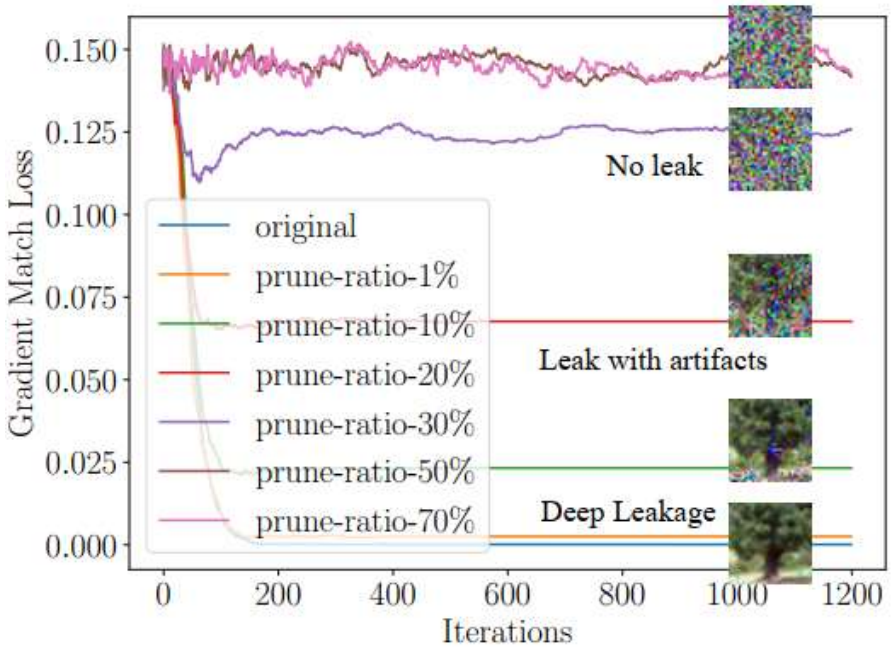
怎样防止训练过程隐私泄露

- 基于梯度压缩的方法
- 基于差分隐私的方法
- 基于数据变换的方法
- 基于密码学的方法

基于梯度压缩的方法

- 采用分层**剪枝**（pruning）技术，“剪掉”绝对值较小的参数梯度。
- 通过剪枝减少梯度中的冗余信息。





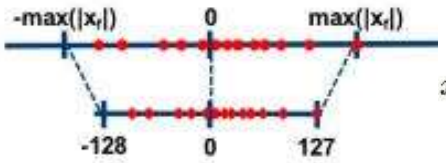
Defense	PSNR	ACC
Pruning (70%)	12.00	77.12
Pruning (95%)	10.07	70.12
Pruning (99%)	10.93	58.33

增大稀疏度能够减少隐私泄露风险，但同时会损害模型精度。

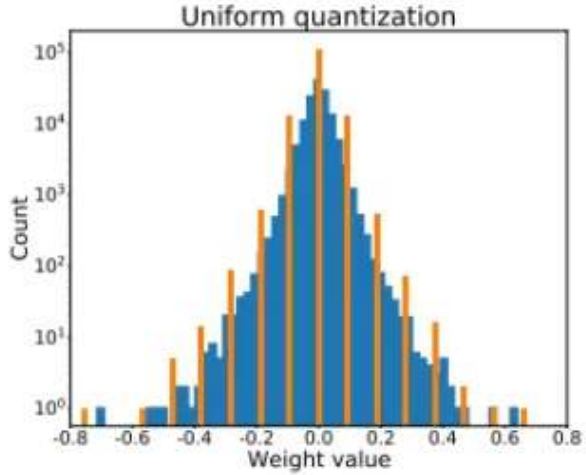
Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

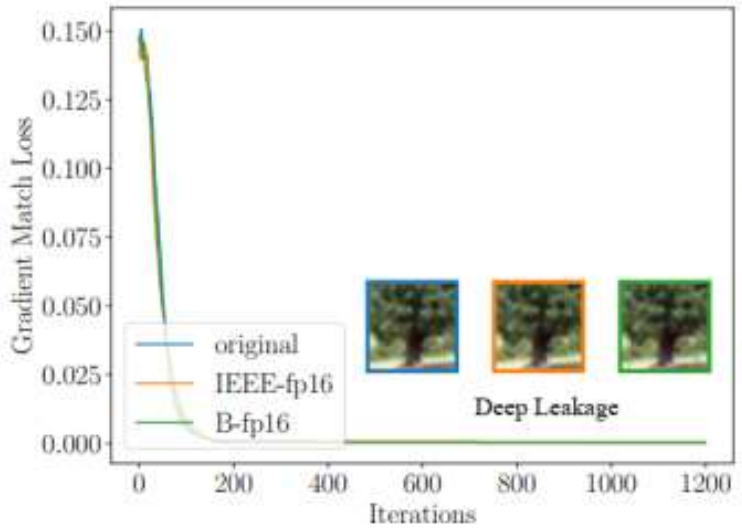
基于梯度压缩的方法

- 采用**量化（quantization）**技术，用低比特来表示梯度。
- 剪枝和量化技术都会减少梯度或者模型中的冗余，增加DLG攻击难度。



$$x_q = round\left(x_f \underbrace{\frac{2^{n-1}-1}{\max|x_f|}}_{q_x}\right) = round(q_x x_f)$$





(c) Defend with fp16 convertion.

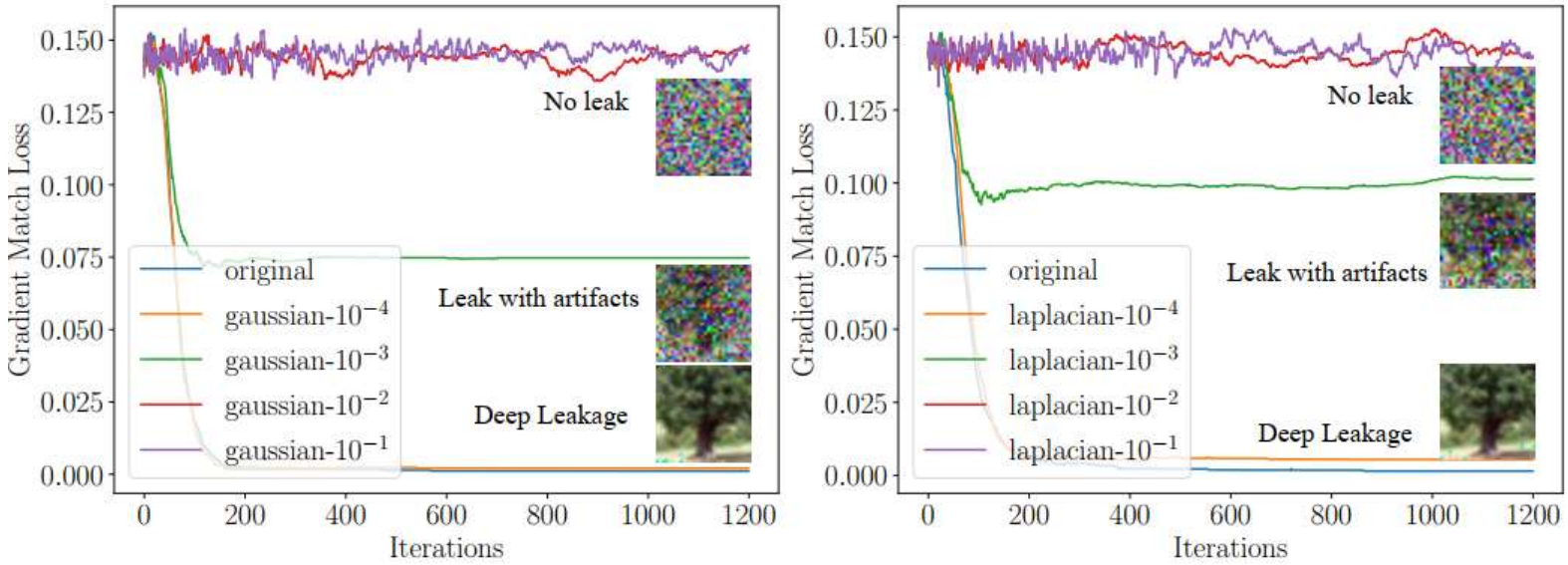
	FP-16
Accuracy	76.1%
Defendability	✗
	Int-8
Accuracy	53.7%
Defendability	✓

使用低比特能够减少隐私泄露风险，但同时会损害模型精度。

Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

基于差分隐私的方法

- 本地训练过程中使用DPSGD



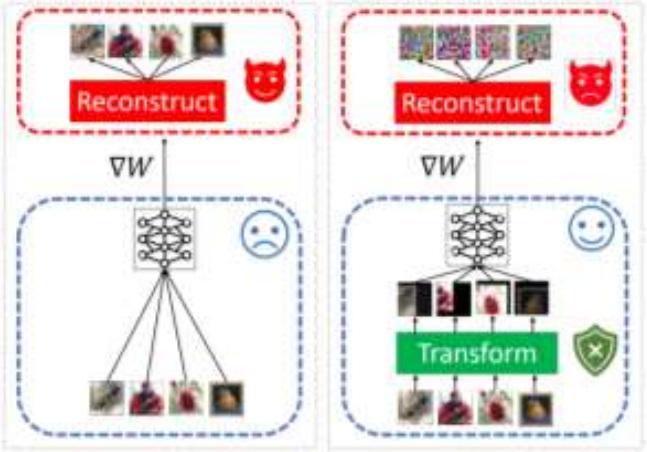
(a) Defend with different magnitude Gaussian noise. (b) Defend with different magnitude Laplacian noise.

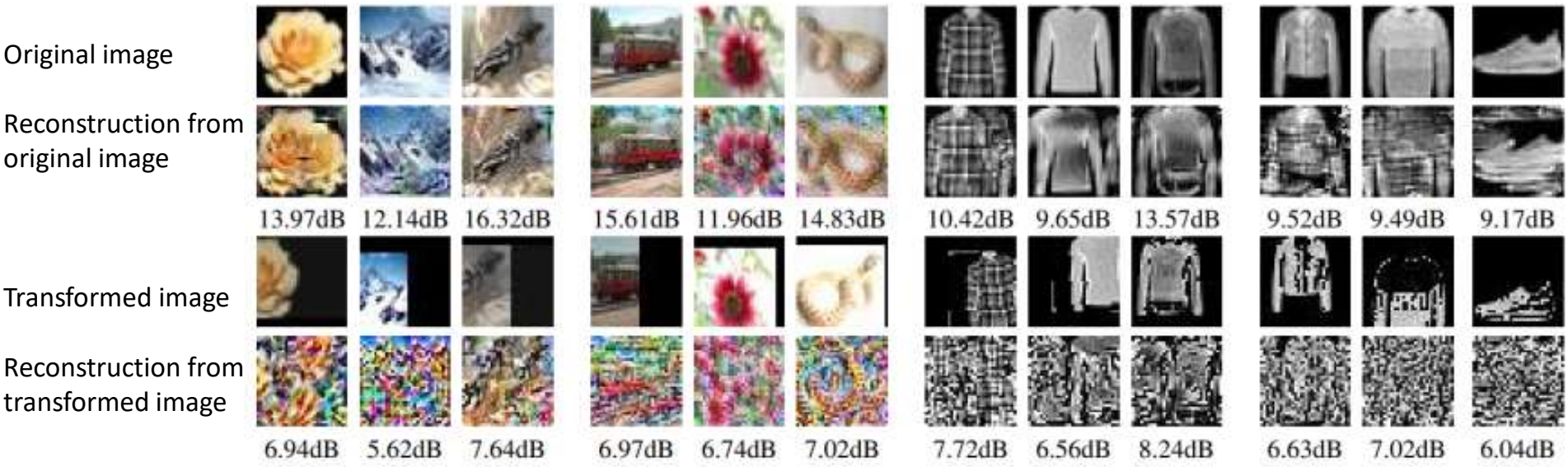
	Original	G-10^{-4}	G-10^{-3}	G-10^{-2}	G-10^{-1}
Accuracy	76.3%	75.6%	73.3%	45.3%	≤1%
Defendability	–	✗	✗	✓	✓
		L-10^{-4}	L-10^{-3}	L-10^{-2}	L-10^{-1}
Accuracy	–	75.6%	73.4%	46.2%	≤1%
Defendability	–	✗	✗	✓	✓

Zhu L, Han S. Deep leakage from gradients[M]//Federated learning. Springer, Cham, 2020: 17-31.

基于数据变换的方法

- 通过**数据增强**的方法改变样本，从而影响梯度，使得产生的梯度难以被用于DLG，而不影响收敛性。
- 自动选择数据增强方法：AutoML





Gao W, Guo S, Zhang T, et al. Privacy-preserving collaborative learning with automatic transformation search [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 114-123.

基于密码学的方法

- 同态加密（Homomorphic Encryption，HE）

  - Full Homomorphic Encryption and Partial Homomorphic Encryption.
  - **Paillier** partially homomorphic encryption

***Addition*:** **[[u]] + [[v]] = [[u+v]]**
Scalar multiplication: **n[[u]] = [[nu]]**

- For public key ***pk = n***, *the encoded form of* $m \in \{0, \ldots, n-1\}$ is

$$\mathbf{Encode(m) = r^n (1 + n)^m \bmod n^2}$$

r is randomly selected from $\{0, \ldots, n-1\}$.

- For float q = (s, e) , encrypt [[q]] = ([[s]], e), here $q = s\beta^e$ is base-β exponential representation.

Gentry, Craig. “Fully homomorphic encryption using ideal lattices.” STOC '09 (2009).

基于密码学的方法

- 同态加密（Homomorphic Encryption，HE）
  - Provides security proof in a well-defined simulation framework
  - Guarantees complete zero knowledge
  - Requires participants' data to be secretly-shared among non-colluding servers
  - Drawbacks:
    - Expensive communication,
    - Though it is possible to build a security model with MPC under lower security requirement in exchange for efficiency

Gentry, Craig. "Fully homomorphic encryption using ideal lattices." STOC '09 (2009).

大 纲

- 隐私保护简介
- 部署模型的隐私保护
- 模型训练过程的隐私保护
- **总结与展望**

总结与展望

- 人工智能的数据挑战：碎片化、隐私泄露、法规。
- 保护训练数据隐私：差分隐私(DP)、成员推理攻击(MIA)。
- 模型训练过程的隐私保护：
  - 分布式学习、联邦学习
  - 梯度泄露隐私
  - 保护方法：梯度压缩、差分隐私、数据变换、密码学方法

Thanks!

Q&A