Deep Leakage from Gradients

Horizontal and Vertical Federated Learning

横向联邦学习为特征对齐的联邦学习。

特点：各个客户端都拥有部分样本且这些样本的特征一致，客户端本地模型同构。典型算法FedAVG，FedSGD等适用于横向联邦学习。

纵向联邦学习为样本对齐的联邦学习。

特点：各个客户端之间拥有相同的样本ID，但拥有这些样本不同的特征。在纵向联邦学习中，标签一般是Guest方拥有，Host方只参与训练，没有标签。

横向与纵向区别：

纵向各个客户端有相同的样本，但特征不一致，无法统一建模结构，允许各客户端个性化建模，各客户端相当于整体的部分模型，所有客户端一起参与预测，VFL不支持FedAVG, FedSGD等。

Federated Learning

目前联邦学习隐私保护研究侧重于横向联邦学习研究

隐私保护研究方向：

攻击：

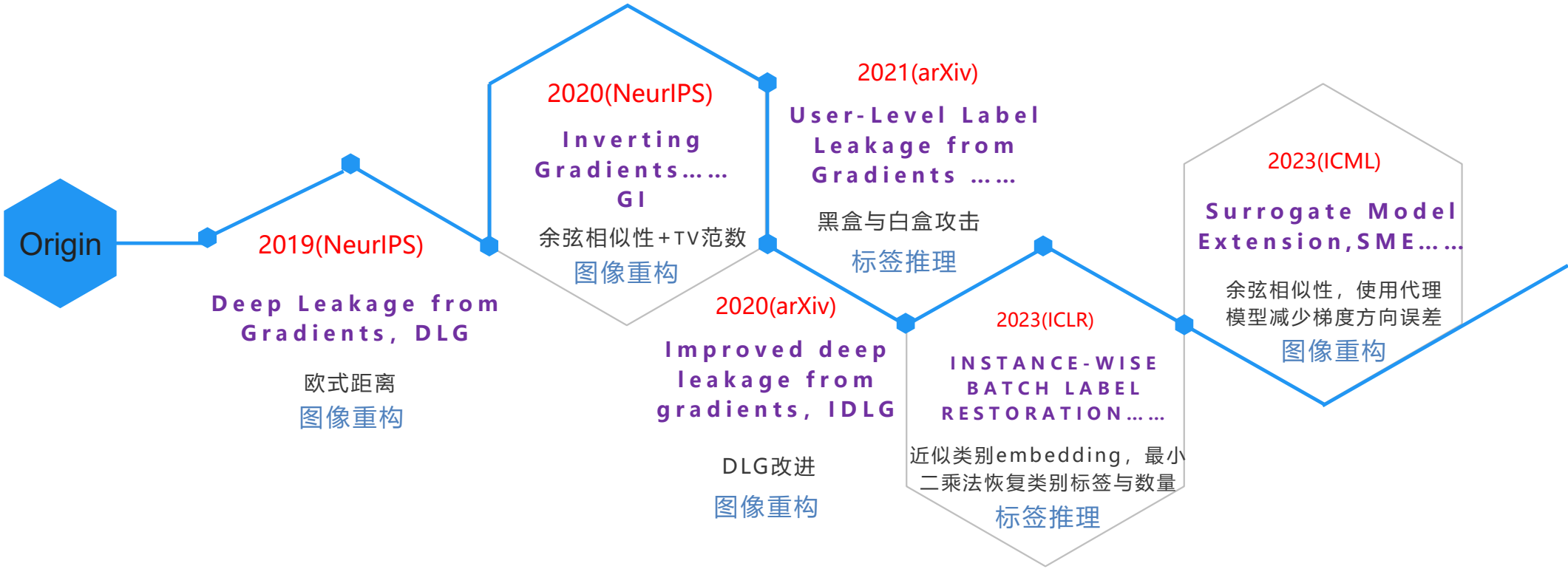
1. 梯度反演，标签推理，推断各客户端训练模型所使用的标签。
2. 梯度反演，使用欧式距离或者余弦相似性，进行客户端图像重构 。
3. 模型投毒攻击，影响全局训练。半可信客户端恶意本地更新绕过服务器异常检测机制，使得最终全局模型最大概率预测不是真实样本类别。
4. 后门攻击。

保护：

1. 梯度裁剪，梯度压缩。
2. DP (Differential Privacy)，差分隐私。隐私数据添加拉普拉斯噪声，但存在局限性，安全性与模型性能之间的权衡。
3. 同态加密，利用加法同态与乘法同态，安全矩阵乘法SMM等实现。

Federated Learning

标签推理与图像重构

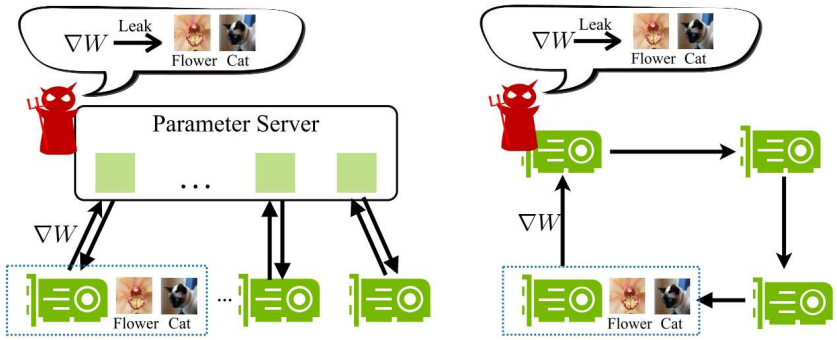


Deep Leakage from Gradients(2019)

1. 梯度泄露客户端训练信息：

DLG是关于分布式机器学习的文章，服务器广播全局模型给到各个客户端，各个客户端使用本地数据计算梯度上传给服务器聚合全局梯度，服务器直接进行全局模型更新或者广播全局聚合梯度，各个客户端在本地进行更新。

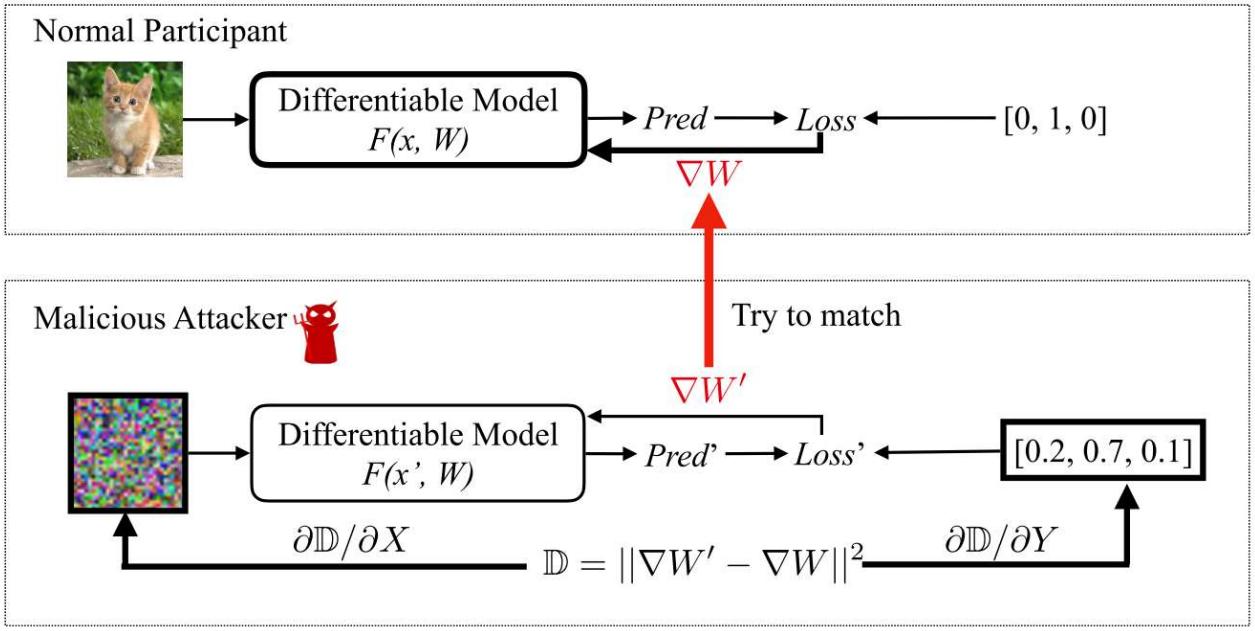
在训练过程中，任何节点不可被完全可信，也就是半可信状态。对于服务器来说，服务器是 honest-but-curious(诚实且好奇)，不干扰训练过程，但是对未知信息充满好奇，服务器是能获得客户端上传的梯度信息，梯度是各客户端的本地数据计算得到，半可信服务器潜在使用梯度推测客户端训练使用的真实数据。



Method

2. 攻击方式：

服务器生成伪数据与伪标签，使用伪数据与伪标签在真实模型参数下计算虚假梯度，随后与客户端上传的真实梯度使用欧式距离构建目标函数进行优化，通过对伪数据和伪标签进行更新，优化真实梯度与虚假梯度之间的差异。



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

Experiments

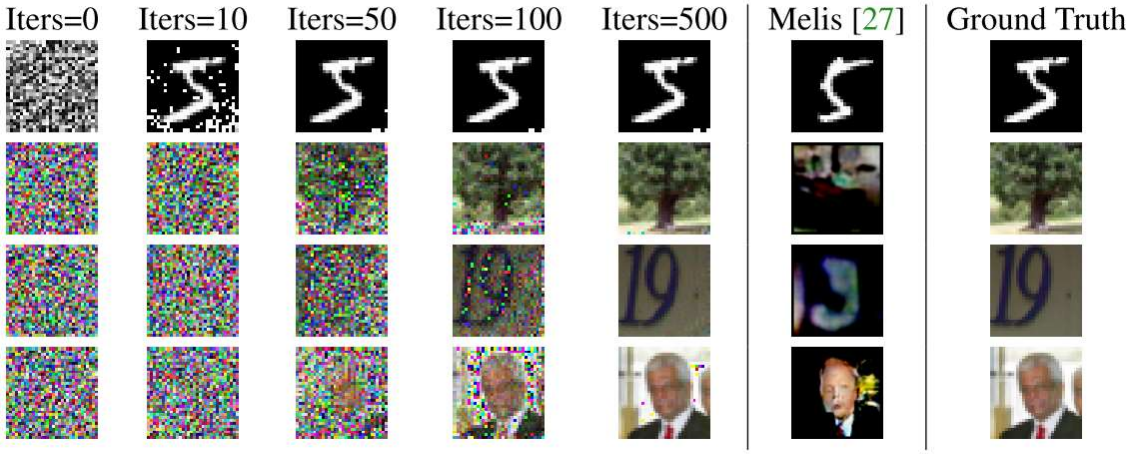


Figure 3: The visualization showing the deep leakage on images from MNIST, CIFAR-100, SVHN and LFW respectively.

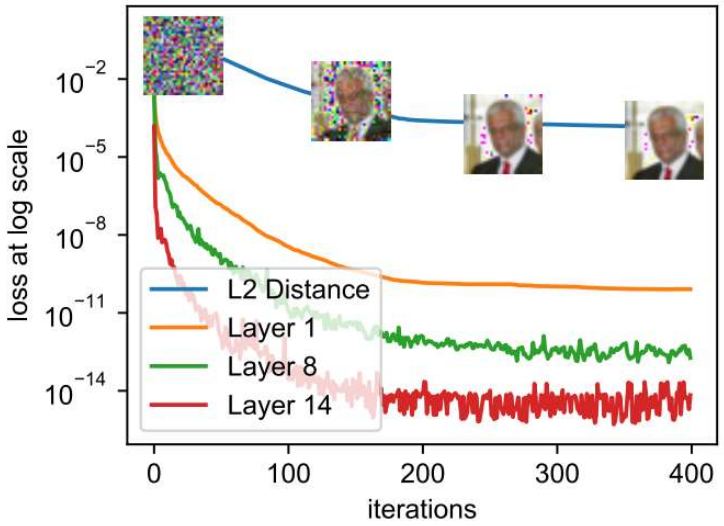
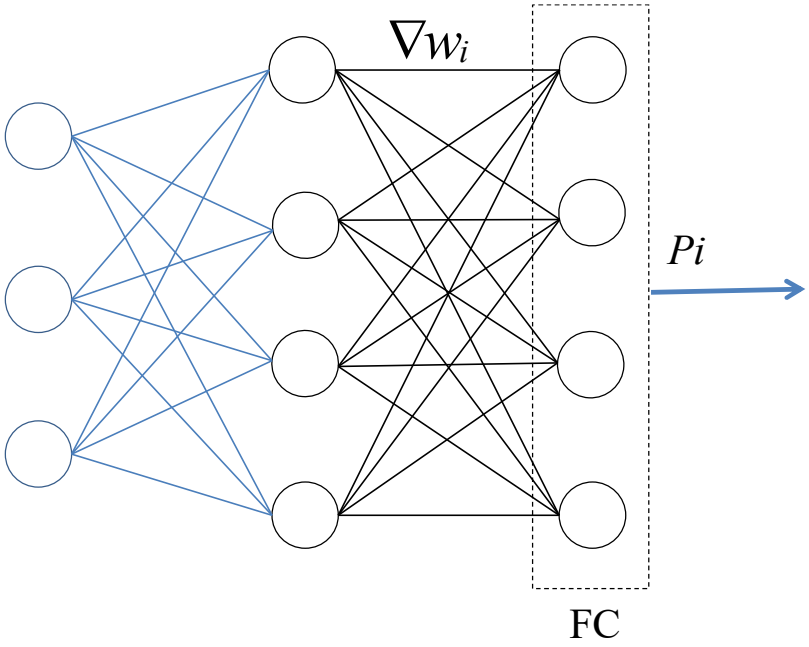


Figure 4: Layer-i means MSE between real and dummy gradients of ith layer. When the gradients'distance gets smaller, the MSE between leaked image and the original image also gets smaller.

Improved deep leakage from gradients(2020)

DLG存在目标函数收敛困难的问题，伪数据和伪标签是同时需要被优化，复杂度比较高。

IDLG，分类问题使用cross-entropy loss 和one-hot label，半可信服务器能通过FC层梯度正负准确推断出标签信息。



$$L(y_i, p_i) = -\sum_i y_i \log \frac{e^{p_i}}{\sum_j e^{p_j}}$$

$$= -\log \frac{e^{p_c}}{\sum_j e^{p_j}}$$

1. $i = c$ $\quad \nabla p_c = -(1 - \frac{e^{p_c}}{\sum_j e^{p_j}})$

2. $i \neq c$ $\quad \nabla p_i = -(0 - \frac{e^{p_i}}{\sum_j e^{p_j}})$

$\nabla p_i = z_i - y_i$ //1与2整合，softmax概率- label，真实类别输出梯度$\nabla p_i = z_i - 1 < 0$ ，非真实类别为$\nabla p_i = z_i - 0 > 0$

$\nabla w_i = \nabla p_i * e$ //激活函数使用Sigmoid或者Relu，非负

Improved deep leakage from gradients

Algorithm 1 Improved Deep Leakage from Gradients (iDLG)

Require:

$F(\mathbf{x}; \mathbf{W})$: Differentiable learning model, $\mathbf{W}$: Model parameters, $\nabla\mathbf{W}$: Gradients produced by private training datum $(\mathbf{x}, c)$, N: maximum number of iterations. η: learning rate.

Ensure:

$(\mathbf{x}', c')$: Dummy datum and label.

1: $c' \leftarrow i \quad \text{s.t.} \quad \nabla\mathbf{W}_L^{i\,T} \cdot \nabla\mathbf{W}_L^j \leq 0, \quad \forall j \neq i$ ▷ Extract the ground-truth label.

2: $\mathbf{x}' \leftarrow \mathcal{N}(0, 1)$ ▷ Initialize the dummy datum.

3: **for** $i \leftarrow 1$ to N **do**

4: $\quad \nabla\mathbf{W}' \leftarrow \partial l(F(\mathbf{x}'; \mathbf{W}), c')/\partial\mathbf{W}$ ▷ Calculate the dummy gradients.

5: $\quad L_G = \|\nabla\mathbf{W}' - \nabla\mathbf{W}\|_F^2$ ▷ Calculate the loss (difference between gradients).

6: $\quad \mathbf{x}' \leftarrow \mathbf{x}' - \eta\nabla_{\mathbf{x}'} L_G$ ▷ Update the dummy datum.

7: **end for**

Surrogate Model Extension (SME)(2023)

1. 余弦相似性

伪数据在模型*w*下梯度方向与真实数据在模型*w*下梯度方向构建目标函数。

2. SME背景

横向联邦学习的FedAVG算法允许各个客户端在本地每轮进行多次iteration或者多个epoch训练，解决通讯开销问题。

$$\mathcal{L} = \overbrace{1 - \frac{\langle \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, D), \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \tilde{D}) \rangle}{\|\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, D)\| \|\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \tilde{D})\|}}^{\mathcal{L}_{sim}(\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, D), \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \tilde{D}))} + \lambda \underbrace{\mathrm{TV}(\tilde{D})}_{\mathcal{L}_{prior}}$$

Surrogate Model Extension (SME)

3. SME动机

Wei et al., 2020在FedAVG下的攻击局限性

4. Wei et al., 2020攻击局限性

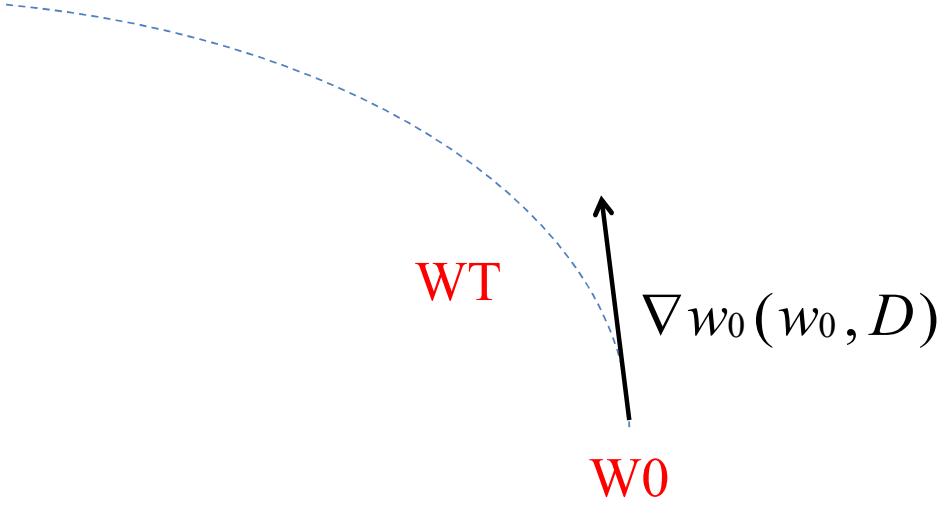
在横向联邦学习中，各客户端将本地更新后的模型上传给服务器聚合。假设最初服务器广播初始化模型W0，考虑通讯开销问题，各客户端本地进行 T 次迭代上传，服务器获得WT，那么，WT=W0 - n∇W_0- n∇W_1-n∇W_2……-n∇W_T-1，服务器能获得信息只有W0与WT，对各客户端中间训练处于全盲状态。

Surrogate Model Extension (SME)

由于中间更新过程对于服务器是全盲状态，Wei et al., 2020 使用W0和WT信息对客户端的数据进行重构。

①W0-WT表示客户端T次迭代平均梯度方向；

②∇W0(W0 , D)表示的客户端真实数据在W0下计算的梯度方向。



Wei et al., 2020 重构算法性能评估：

1.假设客户端使用本地完整数 D 据进行T =1次迭代上传，那么 W0-WT=n∇W0(W0 , D) 与 ∇W0(W0 , D) 梯度方向一致。

$$1 - \frac{\langle w_0 - w_T, \nabla w_0(w_0, D)\rangle}{\| w_0 - w_T \| \| \nabla w_0(w_0, D) \|} = 0$$

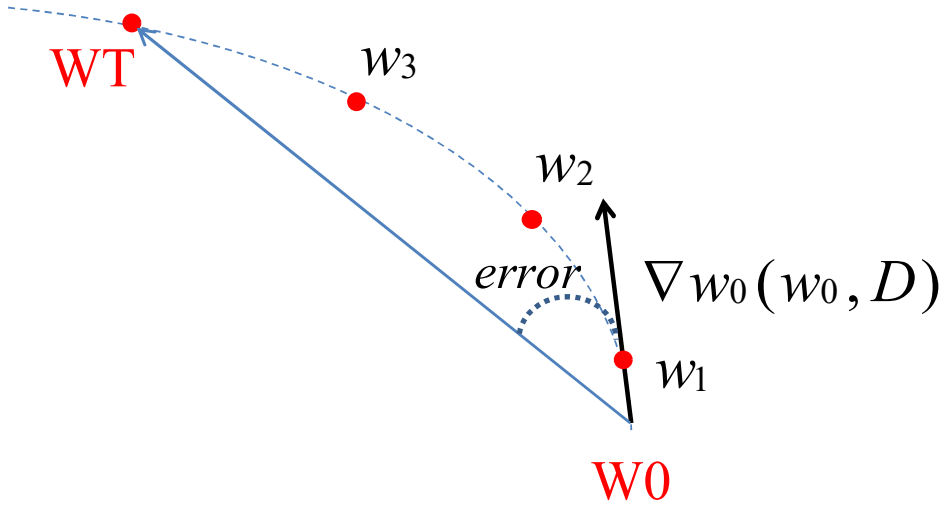
$$L = 1 - \frac{\langle w_0 - w_T, \nabla w_0(w_0, \tilde{D})\rangle}{\| w_0 - w_T \| \| \nabla w_0(w_0, \tilde{D}) \|}$$

L 朝着最小值优化获得

$$\nabla w_0(w_0, D) = w_0 - w_T \approx \nabla w_0(w_0, \tilde{D})$$
$$D \approx \tilde{D}$$

Surrogate Model Extension (SME)

2. 考虑到通讯问题，FedAVG极少使用一次迭代进行一次全局聚合的训练方式，当迭代次数 T >1 时，将D按照 batchsize 大小划分为多个batch训练D ={ D1, D2……}，每轮进行1个或多个epoch训练，W0-WT完整数据计算的平均梯度方向与∇W0(W0 , D)会产生方向误差。



$$1 - \frac{\langle w_0 - w_T, \nabla w_0(w_0, D)\rangle}{\| w_0 - w_T \| \| \nabla w_0(w_0, D) \|} = error$$

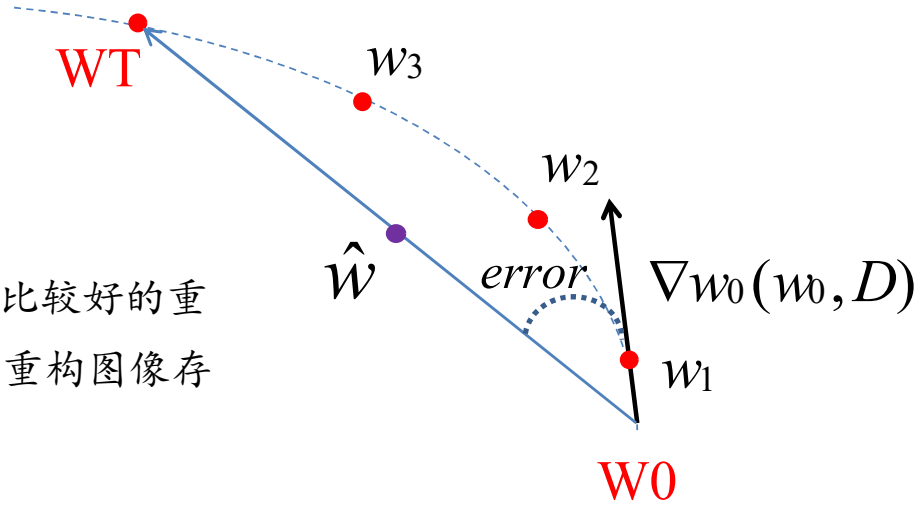
$$L = 1 - \frac{\langle w_0 - w_T, \nabla w_0(w_0, \tilde{D})\rangle}{\| w_0 - w_T \| \| \nabla w_0(w_0, \tilde{D}) \|}$$

L 朝着最小值优化，最终伪数据已经偏离真实数据。

$$\nabla w_0(w_0, D) \neq w_0 - w_T \approx \nabla w_0(w_0, \tilde{D})$$
$$D \neq \tilde{D}$$

Surrogate Model Extension (SME)

$$L = 1 - \frac{\langle w_0 - w_T, \nabla w_0(w_0, D) \rangle}{\| w_0 - w_T \| \| \nabla w_0(w_0, D) \|} = error$$



W0-WT计算的梯度方向与W0计算的梯度方向平行时，能获得比较好的重构结果。然而，对于FedAVG本地进行多次迭代，存在不确定误差，重构图像存在噪声。

SME优化：

SME 在W0与WT之间使用线性插值法，寻找 $\hat{w} = \alpha w_0 + (1-\alpha) w_T$

后续证明了二者之间一定存在一个代理模型 $\hat{w}$，真实数据在代理模型下计算的梯度方向 $\nabla \hat{w}(\hat{w}, D)$ 与 W0-WT 得到的梯度方向一致。

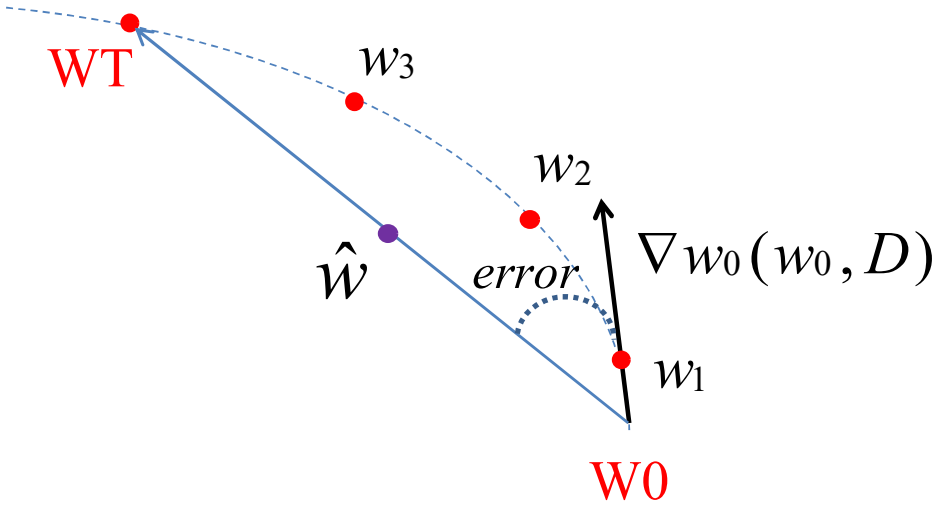
Surrogate Model Extension (SME)

$$\mathcal{L}(\tilde{D}, \widehat{\boldsymbol{w}}) = \overbrace{1 - \frac{\langle \boldsymbol{w}_0 - \boldsymbol{w}_T, \nabla_{\widehat{\boldsymbol{w}}} \ell(\widehat{\boldsymbol{w}}, \tilde{D}) \rangle}{\|\boldsymbol{w}_0 - \boldsymbol{w}_T\| \|\nabla_{\widehat{\boldsymbol{w}}} \ell(\widehat{\boldsymbol{w}}, \tilde{D})\|}}^{\mathcal{L}_{sim}(\boldsymbol{w}_0 - \boldsymbol{w}_T, \nabla_{\widehat{\boldsymbol{w}}} \ell(\widehat{\boldsymbol{w}}, \tilde{D}))} + \lambda \underbrace{\mathrm{TV}(\tilde{D})}_{\mathcal{L}_{prior}},$$

$$s.t. \quad \widehat{\boldsymbol{w}} \in \{\alpha \boldsymbol{w}_0 + (1-\alpha)\boldsymbol{w}_T \mid \alpha \in [0,1]\}. \tag{5}$$

Algorithm 1 Surrogate Model Extension

Input: Victim's weights $\boldsymbol{w}_0, \boldsymbol{w}_T$; Local data size N; Iterations K; Learning rate $\eta_{\tilde{D}}$ for the dummy data and η_α for α; Loss function $\mathcal{L}$.
Initialize $\tilde{D}_0$; $\alpha_0 \leftarrow 0.5$.
for each step $k = 0 \ldots K-1$ **do**
  $\widehat{\boldsymbol{w}} = \alpha_k \boldsymbol{w}_0 + (1-\alpha_k)\boldsymbol{w}_T$
  $\tilde{D}_{k+1} = \tilde{D}_k - \eta_{\tilde{D}} \nabla_{\tilde{D}_k} \mathcal{L}(\tilde{D}_k, \widehat{\boldsymbol{w}})$
  $\alpha_{k+1} = \alpha_k - \eta_\alpha \nabla_{\alpha_k} \mathcal{L}(\tilde{D}_k, \widehat{\boldsymbol{w}})$
end for
Output: Reconstructed data $\tilde{D}_K$.



Experiments

Figure 5. Visualization of the reconstructed images. The results are drawn from the setting ($E = 20$, $N = 50$, $T = 100$) in Table 2. The reconstructed images are paired with the original images through *linear sum assignment*. We randomly sample 16 out of 50 images of one reconstruction.

Experiments

Dataset	E	N	T	DLFA		IG		SME (ours)		
				$\mathcal{L}_{sim}\downarrow$	PSNR ↑	$\mathcal{L}_{sim}\downarrow$	PSNR ↑	$\mathcal{L}_{sim}\downarrow$	PSNR ↑	Δ PSNR
FEMNIST	10	10	10	$.021 \pm .001$	24.9 ± 0.2	$.044 \pm .002$	27.8 ± 0.3	$.019 \pm .001$	$\mathbf{30.3 \pm 0.3}$	$+2.5$
	20	10	20	$.019 \pm .001$	25.4 ± 0.2	$.090 \pm .003$	24.2 ± 0.3	$.019 \pm .001$	$\mathbf{28.6 \pm 0.3}$	$+3.2$
	50	10	50	$.016 \pm .002$	$\mathbf{26.4 \pm 0.3}$	$.202 \pm .005$	19.5 ± 0.2	$.050 \pm .004$	25.7 ± 0.3	-0.7
	10	50	50	$.015 \pm .001$	21.7 ± 0.1	$.091 \pm .003$	19.1 ± 0.2	$.027 \pm .001$	$\mathbf{22.2 \pm 0.2}$	$+0.5$
	20	50	100	$.014 \pm .001$	$\mathbf{21.7 \pm 0.2}$	$.176 \pm .011$	17.0 ± 0.2	$.037 \pm .001$	21.5 ± 0.2	-0.2
	50	50	250	$.015 \pm .001$	$\mathbf{20.3 \pm 0.2}$	$.322 \pm .016$	15.0 ± 0.1	$.065 \pm .002$	20.1 ± 0.2	-0.2
CIFAR100	10	10	10	$.017 \pm .001$	26.7 ± 0.2	$.050 \pm .001$	26.0 ± 0.1	$.024 \pm .001$	$\mathbf{28.5 \pm 0.1}$	$+1.8$
	20	10	20	$.013 \pm .000$	26.0 ± 0.2	$.102 \pm .001$	22.3 ± 0.1	$.040 \pm .001$	$\mathbf{26.9 \pm 0.1}$	$+0.9$
	50	10	50	$.011 \pm .000$	$\mathbf{24.3 \pm 0.2}$	$.225 \pm .002$	16.4 ± 0.2	$.056 \pm .001$	24.2 ± 0.1	-0.1
	10	50	50	$.023 \pm .001$	20.3 ± 0.1	$.094 \pm .001$	17.9 ± 0.1	$.035 \pm .000$	$\mathbf{23.5 \pm 0.1}$	$+3.2$
	20	50	100	$.018 \pm .000$	19.5 ± 0.1	$.154 \pm .002$	14.8 ± 0.1	$.047 \pm .001$	$\mathbf{21.7 \pm 0.1}$	$+2.2$
	50	50	250	N/A	N/A	$.280 \pm .002$	12.1 ± 0.0	$.056 \pm .001$	$\mathbf{18.3 \pm 0.1}$	N/A

Table 2. Average reconstructed image quality measured by PSNR and similarity loss of the reconstruction objective $\mathcal{L}_{sim}$ on FEMNIST and CIFAR100. For clarity, we set batch size $B = 10$ and change local data sizes N and epochs E. Local steps $T = E\lceil N/B \rceil$. The best reconstruction results are bold. The difference of PSNRs between SME and the best baseline is given in the last column. We remark that for PSNR < 18 the reconstruction will be visually corrupted. Also refer to Figure 5 for visualization. Results of DLFA in the last row is not available, as it needs to allocate 102 Gigabytes of GPU memory, which we cannot support.

*Federated Extended MNIST (FEMNIST)：CMU提供的联邦手写体数据集，62种不同的字符类别（10种数字，26种小写，26种大写）的像素（灰度）图片。

关于Wei et al., 2020方向误差优化

① 训练夹角误差，服务器知道W0与WT，对夹角误差估计很困难。

② Wei et al., 2020算法寻找最佳攻击时间段，弥补算法本身存在的缺陷。

服务器协调整个训练过程，攻击可以发生在训练过程的任意时刻，比如，训练前期、中期和后期。考虑到训练前期模型更新速度问题，Wei et al., 2020算法在每轮训练存在大的方向误差。若将Wei et al., 2020算法在模型收敛或者是最后一轮全局训练进行攻击可能会提升攻击效果。

FEMNIST数据集划分中，199个客户端拥有不同数量的数据集，数据异构。

Error =1- cosA　　cosA= <W0-WT , ∇W0(W0 , D)>/(||W0-WT||||∇W0(W0 , D)||)

Weight gradient:			cosA	Bias gradient:	cosA
Round 1	Error_1:	0.553941607475280	(0.45)	0.5200600624084473	(0.48)
Round 10	Error_10:	0.340956985950469	(0.66)	0.3346837759017944	(0.67)
Round 250	Error_250:	0.141046643257141	(0.86)	0.1007887125015258	(0.90)
Round 460	Error_460:	0.049710810184478	(0.96)	0.1120162606239318	(0.89)
Round 500	Error_500:	0.023539304733276	(0.98)	0.1226018071174621	(0.88)

{"clients": ["f2414_89", "f2493_99", "f2188_71", "f2304_67", "f2242_91"],

 "num_samples": [164,154,147,136,141]}　　(5个clients进行500轮训练)

THANKS