联邦学习下对抗训练样本表示的研究

冯霖^{1,2}, 蔡其志², 姜远^{1*}

1. 南京大学软件新技术国家重点实验室, 南京 210023

2. 创新工场人工智能工程院, 北京 100080

* 通信作者. E-mail: jiangyuan@nju.edu.cn

收稿日期: 2019-07-08; 接受日期: 2019-08-02; 网络出版日期: 2021-05-26

摘要 联邦机器学习系统由于能够在多方之间训练联合模型而无需各方共享训练数据, 因此在学术界和工业界都获得了越来越多的关注和应用. 与传统的机器学习框架相比, 这类系统被认为具有保护数据隐私的良好潜力. 另一方面, 训练阶段攻击是一种通过故意扰动训练数据, 从而希望在测试时操纵相应的学习系统预测行为的攻击方法. 例如, DeepConfuse 是最近的一种高效生成对抗训练数据的方法, 展示了传统监督学习范式在此类攻击下的脆弱性. 在本文中, 作者扩展了 DeepConfuse 方法, 将其应用在联邦机器学习框架中. 这是首次针对联邦学习系统的训练阶段攻击. 实验结果表明, 在 δ -准确率损失的衡量标准下, 相比于传统的机器学习框架, 联邦学习系统在 DeepConfuse 攻击下更加脆弱.

关键词 联邦学习, 学件, 表示学习

1 引言

随着社会各界对隐私保护需求的加强, 近年来, 联邦学习系统越来越多地受到了学术界和工业界的关注^[1~3]. 具体而言, 联邦学习系统是一个分布式的具有多个参与者的机器学习框架, 每一个联邦学习的参与者不需要与其余几方共享自己的训练数据, 但仍然能利用其余几方参与者提供的信息更好地训练联合模型. 基于此, 联邦学习系统有望在有效地保护数据隐私的同时扩大机器学习的规模.

另一方面, 训练阶段攻击是一种对机器学习系统的安全性威胁^[4~6]. 训练阶段攻击的一个典型情景是, 攻击者针对被攻击的机器学习系统的训练数据加入一些恶意对抗噪声, 目标是操控在改变后的训练数据上训练得到的模型的行为 (例如在面临未被修改过的测试数据时, 模型系统做出错误的预测). DeepConfuse^[7] 是最近提出的一种高效生成对抗训练数据的方法. 之前的关于训练阶段攻击的方法, 都集中在传统机器学习的框架下, 联邦学习下的训练阶段攻击仍然是一个有待探索公开问题.

在这里本文需要强调, 尽管联邦学习在实践中能有效保护各个数据提供方的隐私不被泄露, 但是这并不代表着联邦学习也能有效避免训练阶段攻击. 在联邦学习的假设中, 各个数据提供方数据均是

引用格式: 冯霖, 蔡其志, 姜远. 联邦学习下对抗训练样本表示的研究. 中国科学: 信息科学, 2021, 51: 900–911, doi: 10.1360/SSI-2019-0145
Feng J, Cai Q-Z, Jiang Y. Towards training time attacks for federated machine learning systems (in Chinese). Sci Sin Inform, 2021, 51: 900–911, doi: 10.1360/SSI-2019-0145

“真实”的^[1],即他们均希望最终训练得到的模型准确率比仅使用自己的数据集得到的模型性能更高.但在实际中,数据提供方中存在恶意攻击者的情形并不能被排除.因此,尽管联邦学习能够让数据提供方的隐私得到保护,但并不能避免训练阶段数据攻击,甚至从本文的结论来看,相比传统的机器学习框架,联邦学习系统在训练阶段攻击下反而更加脆弱.

本篇工作作为联邦机器学习框架下的训练阶段攻击提供了首个威胁模型.具体而言,作者定义了一种针对于联邦学习场景的威胁模型,这里假设攻击者只能访问和修改联邦学习中一个或多个参与方的训练数据,使得整个系统面临未经修改的测试样本时,准确度显著下降.这个问题的挑战在于攻击者在攻击过程中,无法获知完整的模型,只能评估不完整的局部模型和局部训练数据.受 DeepConfuse 的启发,作者提出了一种方法,使用的类似自编码器网络来生成对联邦学习有效的对抗训练数据.实验表明,与标准机器学习系统相比,当受到训练阶段攻击时,联邦机器学习系统的性能下降更大.

本文的其余部分安排如下:首先,分别介绍联邦学习和训练阶段攻击近期的相关工作.接着,将给出此问题的形式化定义,并介绍本文设计的优化方法.而后通过实验,验证方法的有效性.最后给出结论和未来可扩展的工作.

2 相关工作

联邦学习算法最近由 Google 提出^[1~3],本质是一类多智能体学习问题^[8].它不但能够基于分布在多个设备上的数据集构建机器学习模型,同时还能防止数据泄漏.做到在多智能体之间不分享数据的同时分享知识.此范式可以看作是学件^[9]的一种存在形态.理论^[10]和实验^[11]结果显示了使用联邦学习时隐私和安全问题会得到改进.

研究表明^[12],如果恶意攻击者使用联邦学习中的梯度信息来恢复输入数据样本时,联邦学习系统可能存在一些隐私方面的风险.例如,Ateniese 等^[13]成功从训练的模型中提取了样本信息.具体而言,他们能够从联邦语音识别系统中提取种族或性别等本不应泄露的信息.Fredrikson 等^[14]进一步提出了利用置信度信息的攻击方法.他们的实验可以在面部识别系统中重建关于某个已知的特定标签的图像.Shokri 等^[15]开发了成分归属攻击,他们希望在给定一个黑盒机器学习模型的情况下回答特定记录最初是否在训练集中.最近,在联邦学习的条件下,Hitaj 等^[16]提出了一种基于生成式神经网络的方法,当攻击者控制联邦学习中的一方时,该方法可以有效地重建训练样本.一些可能的防御方法也随之被提出以防御与隐私相关的攻击,包括同态加密^[17]、可信多方计算^[18]等.

然而,对联邦学习而言,仅仅保护数据隐私本身是不够的.例如,当前存在一些攻击旨在控制模型行为而不是恢复训练数据.对抗样本^[19]是其中著名的攻击手段:在测试阶段,一些微小的扰动就会使原本分类器的输出结果发生剧烈的变化.对抗训练样本^[20]是對抗样本在训练阶段的体现.该类训练样本会导致对应模型的决策边界发生剧烈的变化,从而在测试阶段,受害分类器将对测试样本给出错误的预测结果. DeepConfuse^[7]是最近生成此类对抗训练数据的最先进方法之一.具体地,该论文提出使用类似自动编码器的网络来产生对训练数据的对抗扰动,使相应的分类网络在测试阶段的泛化能力尽可能地差.为了达到这个目的,DeepConfuse 通过对分类网络训练阶段参数的更新轨迹进行采样,并在这些更新轨迹上利用交替更新的方法训练噪声生成器.同时,在实践中,DeepConfuse 通过与强化学习^[21]类似的技术避免存储分类网络的参数轨迹带来的内存开销,大大提高了训练的效率. DeepConfuse 方法是为标准机器学习框架设计的.据我们所知,到目前为止还没有关于在联邦学习下生成对抗训练样本的研究.

3 本文工作

令 $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n\}$ 是一个关于数据提供方的 N 元集合, $\mathcal{F}_i$ 代表第 i 个数据提供方. 这 N 个数据提供方希望用各自的数据 $\{D_1, D_2, \dots, D_n\}$ 联合训练一个模型. 传统的机器学习方法将所有数据收集到一起, 然后使用联合数据集 $D = D_1 \cup D_2 \cdots \cup D_N$ 训练一个模型 $\mathcal{M}_{\text{SUM}}$. 在联邦学习下, 各个参与者将联合训练一个模型 $\mathcal{M}_{\text{FED}}$, 但与传统机器学习方法不同的是, 联邦学习中, 数据提供方 $\mathcal{F}_i$ 不会将自己的数据 D_i 暴露给其他参与者. 联邦学习算法的目标是使联邦模型 $\mathcal{M}_{\text{FED}}$ 的性能尽可能接近 $\mathcal{M}_{\text{SUM}}$. 具体的, 令 V_{FED} 是 $\mathcal{M}_{\text{FED}}$ 的准确率, V_{SUM} 是 $\mathcal{M}_{\text{SUM}}$ 的准确率, δ - 准确率损失衡量了联邦学习算法的性能, 其中

$$\delta = V_{\text{SUM}} - V_{\text{FED}}. \quad (1)$$

当前有两种典型的联邦学习情景: 水平联邦学习和垂直联邦学习. 在水平联邦学习中, 各个参与方的数据只有输出空间可能不一样, 而在垂直联邦学习中, 各个参与方的数据只有特征维数可能不一样. 举个例子, 考虑如下情景: 医院 I 和 II 希望建立一个联邦学习模型来将病人的特征向量映射至某个特定的疾病. 如果两个医院病人特征向量的表示方式是一致的, 那么就该使用水平联邦学习, 否则就属于垂直联邦学习的范畴. 本文只考虑水平联邦学习, 即不同数据提供方提供的数据表示方式必须保持一致.

如文献 [1] 中所述, 一个标准的联邦学习过程可以分为以下 4 步:

- 第 1 步: 各个训练参与方在本地利用自己的数据训练 f_{θ}^i , 并将 f_{θ}^i 加密后送至服务器.
- 第 2 步: 服务器执行加密的聚合操作并将收到的模型聚合为一个联邦模型 F_{θ} .
- 第 3 步: 服务器将聚合的结果送回各个训练的参与方.
- 第 4 步: 各个训练参与方更新自己的模型并返回第 1 步, 直至收敛.

在整个联邦学习的过程中, 服务器和各个训练方使用同态加密^[17]进行通信, 从而使传输时的信息不被监听. 当训练收敛后, 各个训练方就会得到一个联邦模型 F_{θ} .

本文希望通过攻击一个或多个数据提供方的训练数据衡量联邦模型 F_{θ} 在训练阶段攻击下的性能. 具体的, 考虑如下情景: 恶意攻击者对联邦学习中的一方或几方的数据拥有修改权限, 同时该攻击者希望在这些数据中加入一些人类难以察觉的噪声来损害联邦模型的性能. 在整个过程中, 攻击者对除受害方以外的其他数据提供方都是未知的.

综上所述, 在联邦学习的设定下, 我们希望设计一种生成噪声的机制, 使得该有害噪声不但会影响受害方本地训练的模型 f_{θ}^i , 而且还能损害联邦模型 F_{θ} 的性能. 由于 F_{θ} 是由多个数据提供方的本地模型 f_{θ}^i 聚合而成, 我们猜测当 f_{θ}^i 的性能足够差时, F_{θ} 也会随之变差. 受到 DeepConfuse 方法^[7]的启发, 我们提出了一个简单而有效的方法来攻击联邦学习系统. 接下来, 详细介绍该方法.

为了简化问题, 假设攻击者只能修改第 k 个数据提供方的数据. 对于联邦学习的每个参与方来说, 其本地的模型训练过程和一个标准的机器学习过程没有区别. 对于第 k 个数据提供方的本地训练过程而言, 其目标是从本地数据集 $\mathcal{D}^k$ 学习一个映射 $f_{\theta}^i: \mathcal{X} \rightarrow \{0, 1\}^K$, 这里 K 是分类器的类别数. 为了学习分类器的最优参数 θ^* , 通常会定义一个损失函数如分类问题下的交叉熵损失: $\mathcal{L}(f_{\theta}^i(x), y): \mathbb{R}^m \times \mathbb{Z}_+ \rightarrow \mathbb{R}_+$, 同时应用经验风险最小化^[22]在训练数据上优化该损失:

$$\theta^* = \arg \min_{\theta} \sum_{(x, y) \sim \mathcal{D}^i} [\mathcal{L}(f_{\theta}^i(x), y)]. \quad (2)$$

当 f_{θ}^i 是一个类似于神经网络的可微分系统时, 随机梯度下降 (SGD) [23] 或者其变体会被用来更新模型参数, 具体的, 当算法执行至第 t 步时, 利用如下公式得到第 $t+1$ 步的模型参数:

$$\theta_{t+1} \leftarrow \theta_t - \alpha \nabla_{\theta_t} \mathcal{L}(f_{\theta_t}^i(x), y), \quad (3)$$

这里 α 是学习率, 通常是一个人为设定的超参数.

对于受害方 k , 本文定义了一个噪声生成器 $g_{\xi}^k: \mathcal{X} \rightarrow \mathcal{X}$, 它接受样本空间 $\mathcal{X}$ 中的一个训练样本 x 并将其转化为相同样本空间 $\mathcal{X}$ 中的一个难以察觉的噪声, 对于图像数据, 这种“难以察觉”的限制可以被写成如下约束:

$$\forall x, \|g_{\xi}^k(x)\|_{\infty} \leq \epsilon, \quad (4)$$

其中 $\|\cdot\|_{\infty}$ 是无穷范数, 其被定义为一个高维向量各个维度模的最大值:

$$\|x\|_{\infty} = \max_{i \in \{1, \dots, n\}} |x_i|. \quad (5)$$

这里, ϵ 控制了对样本的扰动强度, 这是对抗机器学习 [24] 中的一个常用技巧. 本文将噪声生成器 g_{ξ}^k 的最后一层定义为 $\epsilon \cdot (\tanh(\cdot))$, 这样, 该噪声生成器生成的噪声就会天然满足上述约束 (5).

在以上的记号下, 本文将对受害方的训练阶段攻击形式化为如下的优化问题:

$$\begin{aligned} \max_{\xi} \quad & \sum_{(x,y) \sim \mathcal{D}^k} [\mathcal{L}(f_{\theta^*}^k(\xi)(x), y)], \\ \text{s.t. } \quad & \theta^*(\xi) = \arg \min_{\theta} \sum_{(x,y) \sim \mathcal{D}^k} [\mathcal{L}(f_{\theta}^k(x + g_{\xi}^k(x)), y)]. \end{aligned} \quad (6)$$

换句话说, 每个可能的噪声生成器的参数 ξ 会有一个对应的在被污染的训练数据下训练得到的最优分类器 $f_{\theta^*}^k(\xi)$, 我们的目标就是寻找最优的 $g_{\xi^*}^k$, 使得相应的最优分类器 $f_{\theta^*}^k(\xi^*)$ 在测试数据上具有最差的性能.

解决这个非凸优化问题是困难的, 尤其是由于非线性等号约束的存在. 这里我们参考了 DeepConfuse [7] 来有效逼近该优化问题的解.

本文工作的基本思路是分别利用梯度上升和梯度下降在对抗训练样本和干净数据上交替更新 f_{θ}^k 和 g_{ξ}^k . 但是如果直接使用交替更新的方法, f_{θ}^k 和 g_{ξ}^k 在实践中均难以收敛. 为了稳定这一过程, DeepConfuse 提出首先利用 g_{ξ}^k 生成的对抗训练样本训练 f_{θ}^k , 并收集 f_{θ}^k 的更新轨迹, 然后再利用收集到的轨迹对 f_{θ}^k 计算一个“伪更新”从而更新 g_{ξ}^k , 整个过程会重复 T 次直到 g_{ξ}^k 收敛.

最后, 为了使该算法更加高效, 我们设计了一个额外的优化技巧. 注意到保存 f_{θ}^k 训练时的更新轨迹会带来很大的内存开销, 为了避免保存这类信息, 在每次训练 f_{θ}^k 的试验中, 可以创建一个 g_{ξ}^k 的拷贝 g'_{ξ}^k 并利用 g_{ξ}^k 生成的对抗训练样本交替更新 f_{θ}^k 和 g'_{ξ}^k . 当 f_{θ}^k 训练完毕后, 再将 g'_{ξ}^k 的参数拷贝回 g_{ξ}^k . 这样做可以将原先的两个循环合并至一个中, 并且避免了存储参数带来的额外内存开销. 详细过程参见算法 1.

4 实验

在本节中, 我们希望回答如下的问题:

Algorithm 1 Fed-DeepConfuse

Input: List of party can be accessed by the attacker L_{atk} , training data $\{D^1, D^2, \dots, D^n\}$, number of trials T , max iteration for training a classification model maxiter, learning rate of classification model α_{fi} , learning rate of the noise generator α_{gi} , batch size b ;

Output: Modified datasets $\{D^1, D^2, \dots, D^n\}$.

```

1: for  $k = 1$  to  $n$  do
2:   if  $k \in L_{\text{atk}}$  then
3:      $\xi \leftarrow \text{RandomInit}()$ ;
4:      $g_{\xi}^k \leftarrow g_{\xi}^k.\text{copy}()$ ;
5:     for  $t = 1$  to  $T$  do
6:        $\theta_0 \leftarrow \text{RandomInit}()$ ;
7:       for  $i = 0$  to maxiter do
8:          $(x_i, y_i) \sim D^k$ ; //Sample a mini-batch of training data
9:          $\theta' \leftarrow \theta_i - \alpha_{fk} \nabla_{\theta_i} \mathcal{L}(f_{\theta_i}^k(x_i + g_{\xi}^k(x_i)), y_i)$ ;
10:         $\xi' \leftarrow \xi' + \alpha_{gk} \nabla_{\xi'} \mathcal{L}(f_{\theta'}^k(x), y)$ ; //Update  $g_{\xi}^k$  using current  $f_{\theta}^k$ 
11:         $x_i^{\text{adv}} \leftarrow x_i + g_{\xi}^k(x_i)$ ;
12:         $\theta_{i+1} \leftarrow \theta_i - \alpha_{fk} \nabla_{\theta_i} \mathcal{L}(f_{\theta_i}^k(x_i^{\text{adv}}), y_i)$ ; //Update  $f_{\theta}^k$  by SGD
13:      end for
14:       $g_{\xi}^k \leftarrow g_{\xi'}^k$ ;
15:    end for
16:     $m \leftarrow \text{len}(D^k)$ ;
17:    for  $i = 1$  to  $m$  do
18:       $(x_i, y_i) \leftarrow D^k[i]$ ;
19:       $x'_i \leftarrow x_i + g_{\xi}^k(x_i)$ ;
20:       $D^k[i] \leftarrow (x_i, y_i)$ ;
21:    end for
22:  end if
23: end for
24: return  $\{D^1, D^2, \dots, D^n\}$ .

```

首先, 该方法对联邦学习的攻击性能怎样? 我们分别考虑有 2, 3, 4 个数据提供方的联邦学习场景并评估当不同数目的数据提供方被攻击时, 对应的联邦学习模型的性能改变情况.

其次, 对抗训练样本是否具备可迁移性? 换句话说, 当我们得到了对某种特定的分类器有效的对抗训练样本生成器, 其生成的对抗训练样本在分类器改变后, 是否依旧有效?

然后, 是否存在关于对抗训练样本的解释? 为什么这种对抗训练样本会使模型表现不同?

最后, 我们还希望知道和传统学习场景相比, 联邦学习对训练阶段攻击是更加鲁棒还是更加脆弱?

在本节接下来的部分中, 将首先介绍实验设置, 然后展示在联邦学习下训练阶段攻击的结果, 同时使用线性假设和一些视觉解释来验证对抗训练样本有效性. 随后评估对抗训练样本在联邦学习的条件下的迁移能力. 最后, 保持其他设定不变, 对传统机器学习场景和联邦学习同时进行训练阶段攻击, 从而比较不同场景下训练阶段攻击的性能差异.

4.1 实验设置

为了验证本文方法的有效性, 使用经典的手写数字分类数据集 MNIST^[25] 和彩色图片分类数据集 CIFAR-10^[26] 作为多分类实验的数据集. 具体的, MNIST 数据集包含了 60000 张灰阶分辨率是 28×28 的训练图片和 10000 张测试图片. CIFAR-10 包含了 50000 张彩色的分辨率是 32×32 的训练图片和 10000 张测试图片. 本文所有实验使用相同的扰动强度, 对 MNIST, 使用 $\epsilon = 0.3$, 对 CIFAR-10, 使用

图 1 (网络版彩图) 第 1 行: 原始训练样本. 第 2 行: 对抗训练样本

Figure 1 (Color online) First row: original training samples. Second row: adversarial training samples. (a) MNIST ($\epsilon = 0.3$); (b) CIFAR-10 ($\epsilon = 0.032$)

表 1 联邦学习场景下的实验设定

Table 1 Experimental configurations under different federated situations

		MNIST	CIFAR-10
Two-party learning	Party A	0, 1, 2, 3, 4, 5	Airplane, automobile, bird, cat, deer, dog, frog, horse
	Party B	4, 5, 6, 7, 8, 9	Bird, cat, deer, dog, frog, horse, ship, truck
Three-party learning	Party A	0, 1, 2, 3, 4	Airplane, automobile, bird, cat, deer
	Party B	3, 4, 5, 6, 7	Cat, deer, dog, frog, horse
	Party C	5, 6, 7, 8, 9	Dog, frog, horse, ship, truck
Four-party learning	Party A	0, 1, 2, 3	Airplane, automobile, bird, cat
	Party B	2, 3, 4, 5	Bird, cat, deer, dog
	Party C	4, 5, 6, 7	Dog, frog, horse, ship
	Party D	6, 7, 8, 9	Frog, horse, ship, truck

$\epsilon = 0.032$. 我们随机挑选了原始图片和对抗训练图片展示在图 1 中.

对联邦学习的实验设置, 我们考虑了具有 2 方、3 方、4 方的联邦学习. 注意到在联邦学习的条件下, 有一些标签可能只出现在某个特定的数据提供方中, 也有标签会出现在多个数据提供方中. 在实验中, 我们也包含了这种情形, 实验设定见表 1.

为了公平比较, 对传统机器学习和联邦学习使用相同的分类器 f_θ . 对于 MNIST 数据集, 分类器 f_θ 是一个简单的卷积神经网络, 具有两个卷积层, 通道数分别是 20 和 50, 在之后跟着一个包含了 500 个隐含节点的全连接层. 对于 CIFAR-10 数据集, 使用更加复杂的 ResNet-18^[27] 模型.

在训练过程中, 交叉熵损失被用作 f_θ 的损失函数, 对 MNIST 和 CIFAR-10 两个数据集, 学习率和批大小分别设定为 0.01, 64 和 0.0001, 128. 试验次数 T 被设定为 500. 当测试对抗训练样本在传统机器学习和联邦机器学习中的性能时, 使用与训练阶段相同的模型结构, 在训练分类器时, 训练 100 轮以确保模型收敛.

对于 MNIST 数据集, 噪声生成器 g_ξ 包含了一个简单的编码器-解码器结构, 每一个编码器和解码器具有 4 个 4×4 的卷积层, 通道数分别是 16, 32, 64, 128. 对于 CIFAR-10 数据集, 使用了更复杂的 U-Net^[28] 结构. 在所有的实验中, 使用 Adam^[29] 优化噪声生成器 g_ξ , 学习率设定为 10^{-4} .

4.2 实验结果

在上述的实验设定下, 对不同的数据提供方训练得到了噪声生成器 g_ξ 和对抗训练样本. 分别在

图 2 当不同数目的数据提供方被攻击时联邦学习的测试性能. 红色水平线指示了随机猜测的准确率. 不同颜色的柱状图分别表示当有 1~4 个数据提供方的训练数据被替换为对抗训练样本时模型在测试集上的正确率

Figure 2 Test performance when different number of data parties are attacked. The horizontal red line indicates random guess accuracy. Different colors of histograms indicate the test performance of the federated model when the training data of 1 to 4 parties is adversarially modified. (a) MNIST; (b) CIFAR-10

2 方、3 方、4 方联邦学习中测试了当不同数目的数据提供方被攻击后, 所训练出的联邦学习模型的性能. 结果参见图 2.

从实验结果看, 对 MNIST 而言, 攻击一半的数据提供方大约会使模型准确率下降 50%, 对 CIFAR-10 而言, 模型的准确度会降至与随机猜测无异. 我们还发现了很多其他有趣的现象, 例如, 随着数据提供方的变多, 联邦学习的性能会如预期一般下降, 同时攻击的成功率也会随着攻击者可侵入的数据提供方数目的变多而变高.

最后, 当所有训练数据都被攻击时, 联邦学习的正确率会降至随机猜测甚至更低. 而当攻击者仅可控制一方数据时, 仍然可以清晰地观测到联邦模型测试性能的下降.

4.3 对抗训练样本的迁移能力

另一个更加实际的问题是当联邦学习下分类器的结构改变时, 同样的训练阶段攻击是否依然有效? 换句话说, 使用相同的对抗训练数据, 当各个数据提供方使用其他的分类模型时, 该攻击的性能如何? 为了回答这个问题, 我们在同样的联邦学习框架下, 使用了与训练时不同的网络结构并验证联邦学习测试阶段的准确率. 具体的, 考虑一个两方学习的场景, 对 MNIST, 加倍和减半了模型的通道数并分别把他们记作 CNN_{large} 和 CNN_{small} , 对 CIFAR-10, 使用了 ResNet50 和 ResNet152 并保持其他设定不变, 分别评估了在不同模型结构下联邦学习的测试性能, 实验结果参见图 3.

结果显示, 尽管噪声生成器是针对某个特定的模型结构训练的, 这类对抗纹理对各种结构的模型均是有害的. 也就是说, 对抗训练样本具有良好的迁移能力.

4.4 视觉解释和线性假设

对上述的对抗训练样本的迁移能力, 一个可能存在的解释是在各种噪声纹理会使一个神经网络过

图 3 当使用不同的分类器结构时联邦学习模型的测试性能. 红色水平线指示了随机猜测的准确率. 蓝色柱状图表示两方学习中的原始测试正确率, 橙色和绿色柱状图表示当一方或两方数据被攻击时, 联邦模型的测试准确率

Figure 3 Test performance when using different classifiers under two-party learning scenario. The horizontal red line indicates random guess accuracy, blue histograms indicate the original test accuracy, and the orange and green ones indicate the test accuracy when one or two party's training data is attacked. (a) MNIST; (b) CIFAR-10

表 2 当输入仅包含噪声时联邦模型的准确率. 即, 真实标签和 $f_{\theta}(g_{\xi}(x))$ 之间的准确率, 这里 x 是原始样本

Table 2 Prediction accuracy of the federated model taking only noises as inputs. That is, the accuracy between the true label and $f_{\theta}(g_{\xi}(x))$ where x is the clean sample

	Noise _{train} (%)	Noise _{test} (%)
MNIST	66.78	66.05
CIFAR-10	44.60	45.91

拟合. 本文使用一些简单的实验验证这个观点.

考虑在四方学习中三方数据都受到攻击, 当联邦模型训练完后, 评估当输入样本只有噪声 $f_{\theta}(g_{\xi}(x))$ 时模型的输出结果, 并将其和 x 对应的真实标签比较, 计算准确率. 结果总结在表 2 中, 注意到在测试数据集上, 仅使用对抗噪声, 得到了 66.05% 和 45.91% 的测试准确率.

这个有趣的结论证实了我们的猜想, 即联邦模型确实会过拟合在训练噪声上, 我们认为深度模型的线性假设是这种攻击能够奏效的原因. 换句话说在最小化损失函数 $\mathcal{L}(f_{\theta}(x + g_{\xi}(x)), y)$ 的过程中, $f_{\theta}(g_{\xi}(x))$ 起到了主导作用. 这个结论再次证实了之前对测试阶段对抗样本的研究^[24]和神经网络的记忆性质^[30].

最后, 在图 4 和 5 中, 我们分别可视化了 f_{θ} 的隐含层的激活值及部分样本相对应的对抗噪声. 具体的, 图 4 使用主成分分析^[31], 分别将神经网络在对抗训练样本和原始测试样本上的隐含层输出投影至二维平面中. 可以清晰地看出, 使用对抗训练样本得到的决策边界无法区分原始数据. 我们希望这类可视化能够使本文的结论更加可信.

图 4 (a) MNIST-Train; (b) MNIST-Test; (c) CIFAR-Train; (d) CIFAR-Test. (a) 和 (c) 表示在对抗训练数据上深度特征的可视化. (b) 和 (d) 表示在原始测试数据上深度特征的可视化

Figure 4 (a) MNIST-Train; (b) MNIST-Test; (c) CIFAR-Train; (d) CIFAR-Test. (a) and (c) represent deep features of the adversarial training data. (b) and (d) represent deep features of the cleaned test data

图 5 原始样本及其对应的对抗噪声

Figure 5 Clean samples and their corresponding adversarial noises for MNIST and CIFAR-10

4.5 鲁棒性对比

本小节研究如下的问题: 联邦学习和传统机器学习相比, 当部分数据受到训练阶段攻击时的, 模型鲁棒性是更好还是更差? 换句话说, 假设相同比例的训练数据被攻击, 比较联邦学习和传统机器学习的准确率.

为了公平比较, 考虑如下的场景: (1) 在标准机器学习中有 50% 的数据是由 DeepConfuse 产生的对抗训练样本; (2) 在两方学习中, 有一方的数据是由 Fed-DeepConfuse 产生的对抗训练样本. 结果展示在表 3 中. 其他实验设定和前文所属的一样.

表 3 当一半的数据受到对抗攻击后联邦学习模型和传统机器学习模型的正确率比较

Table 3 Comparison of federated learning scenario and standard learning scenario when half of the training data is adversarially modified

	Clean			50% Attacks		
	Standard learning	2-party fed-learning	δ -accuracy loss	Standard learning	2-party fed-learning	δ -accuracy loss
MNIST	99.32	98.38	0.94	98.01	67.69	30.32
CIFAR	93.01	70.36	22.65	91.36	51.37	39.99

实验结果显示, 在 MNIST 上, 联邦学习的 δ - 准确率损失从 0.94% 上升至 30.32%, 在 CIFAR-10 上, δ - 准确率损失从 22.65% 上升至 39.99%. δ - 准确率损失的显著上升说明了联邦信息系统对训练阶段攻击更加敏感. 同时, 我们发现对传统的机器学习系统而言, 使用 DeepConfuse 扰动 50% 的数据的效果不是非常显著, 在两个数据集上分别只能使模型准确率降低 1.32% 和 1.65%, 而 Fed-DeepConfuse 能成功地使联邦模型准确率降低 30.69% 和 18.99%. 这说明, 当面对训练阶段攻击时, 联邦学习系统比传统机器学习系统更加脆弱.

5 总结

本文针对联邦学习模型提出了训练阶段攻击的威胁任务, 同时, 本文针对此任务, 设计了一个有效的算法 (Fed-DeepConfuse), 用来生成针对联邦学习的训练阶段的对抗训练样本. 实验结果验证了该方法的有效性, 同时, 本文结果还显示, 从训练阶段攻击的角度而言, 联邦学习系统比传统机器学习系统更加脆弱. 本文工作提出了在分布式多智能体下, 高维空间中的恶意表示学习可由自编码器获得. 这有助于进一步设计和理解具有鲁棒性的高维空间的表示学习任务.

注意到本文仅考虑对水平联邦学习的训练阶段攻击, 对垂直联邦学习而言, 如何设计有效的训练阶段攻击依然是个值得探索的方向, 亦值得在未来进一步研究.

参考文献

- 1 Yang Q, Liu Y, Chen T J, et al. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol*, 2019, 10: 1–19
- 2 Konečný J, McMahan H B, Yu F X, et al. Federated learning: strategies for improving communication efficiency. 2016. ArXiv:1610.05492
- 3 Konečný J, McMahan H B, Ramage D, et al. Federated optimization: distributed machine learning for on-device intelligence. 2016. ArXiv:1610.02527
- 4 Laskov P, Lippmann R. *Machine Learning in Adversarial Environments*. Berlin: Springer, 2010
- 5 Mei S K, Zhu X J. Using machine teaching to identify optimal training-set attacks on machine learners. In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015. 2871–2877
- 6 Biggio B, Roli F. Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recogn*, 2018, 84: 317–331
- 7 Feng J, Cai Q Z, Zhou Z H. Learning to confuse: generating training time adversarial data with auto-encoder. 2019. ArXiv:1905.09027
- 8 Li M, Wang W, Zhou Z H. Exploiting remote learners in Internet environment with agents. *Sci China Inf Sci*, 2010, 53: 64–76
- 9 Zhou Z H. Learnware: on the future of machine learning. *Front Comput Sci*, 2016, 10: 589–590

- 10 Smith V, Chiang C K, Sanjabi M, et al. Federated multi-task learning. In: *Proceedings of Neural Information Processing Systems (NIPS)*, 2017. 4424–4434
- 11 Ching T, Himmelstein D S, Beaulieu-Jones B K, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*, 2018, 15: 20170387
- 12 Su L L, Xu J M. Securing distributed gradient descent in high dimensional statistical learning. In: *Proceedings of ACM on Measurement and Analysis of Computing Systems*, 2019
- 13 Ateniese G, Felici G, Mancini L V, et al. Hacking smart machines with smarter ones: how to extract meaningful data from machine learning classifiers. 2013. ArXiv:1306.4447
- 14 Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic counter-measures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015. 1322–1333
- 15 Shokri R, Stronati M, Song C Z, et al. Membership inference attacks against machine learning models. In: *Proceedings of IEEE Symposium on Security and Privacy (SP)*, 2017. 3–18
- 16 Hitaj B, Ateniese G, Pérez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. In: *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, 2017. 603–618
- 17 Phong L T, Aono Y, Hayashi T, et al. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Trans Inform Forensic Secur*, 2018, 13: 1333–1345
- 18 Mohassel P, Zhang Y P. Secureml: a system for scalable privacy-preserving machine learning. In: *Proceedings of IEEE Symposium on Security and Privacy*, 2017. 19–38
- 19 Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. 2013. ArXiv:1312.6199
- 20 Koh P W, Liang P. Understanding black-box predictions via influence functions. In: *Proceedings of the 34th International Conference on Machine Learning*, 2017. 1885–1894
- 21 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518: 529–533
- 22 Vapnik V. Principles of risk minimization for learning theory. In: *Proceedings of Neural Information Processing Systems (NIPS)*, 1992. 831–838
- 23 Bottou L. Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT*, 2010. 177–186
- 24 Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2015. ArXiv:1412.6572
- 25 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. *Proc IEEE*, 1998, 86: 2278–2324
- 26 Krizhevsky A. Learning Multiple Layers of Features From Tiny Images. Technical Report TR-2009, 2009
- 27 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 770–778
- 28 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Proceedings of International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015. 234–241
- 29 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: *Proceedings of the 3rd International Conference for Learning Representations*, 2014
- 30 Zhang C Y, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization. In: *Proceedings of the 5th International Conference on Learning Representations*, 2016
- 31 Pearson K. On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos Mag J Sci*, 1901, 2: 559–572

Towards training time attacks for federated machine learning systems

Ji FENG^{1,2}, Qi-Zhi CAI² & Yuan JIANG^{1*}

1. *National Key Lab for Novel Software Technology, Nanjing University, Nanjing 210023, China;*

2. *Sinovation Ventures AI Institute, Beijing 100080, China*

* Corresponding author. E-mail: jiangyuan@nju.edu.cn

Abstract Federated machine learning systems have gained more and more attention and popularity in both academia and industry because they can obtain a shared model among multiple parties without explicitly sharing the training data. Such a system is believed to have a good potential of protecting data privacy compared with the traditional machine learning frameworks. On the other hand, training time attacks are a procedure of purposefully modifying training data, hoping to manipulate the behavior of the corresponding trained system during test time. DeepConfuse, for instance, is one recent advance in generating adversarial training data with high efficiency. In this work, we extend the DeepConfuse framework so that it can be used in federated machine learning. This is the first training time attack for a federated learning system. The empirical results showed that the federated learning system is even more vulnerable under the DeepConfuse attack in terms of δ -accuracy loss.

Keywords federated learning, learnware, representation learning

Ji FENG is currently a Ph.D. candidate of Department of Computer Science and Technology in Nanjing University since 2015 and became a member of LAMDA Group, led by professor Zhi-Hua ZHOU. His research interests include ensemble learning and deep learning.

Yuan JIANG was born in 1976. She is now a full professor in Department of Computer Science and Technology at Nanjing University. She received the Ph.D. degree in computer science from Nanjing University in 2004. Her research interests are mainly in artificial intelligence, machine learning, and data mining.