Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

联邦学习的后门攻击方法

硕士研究生　杨得山

2022年08月28日

内容提要

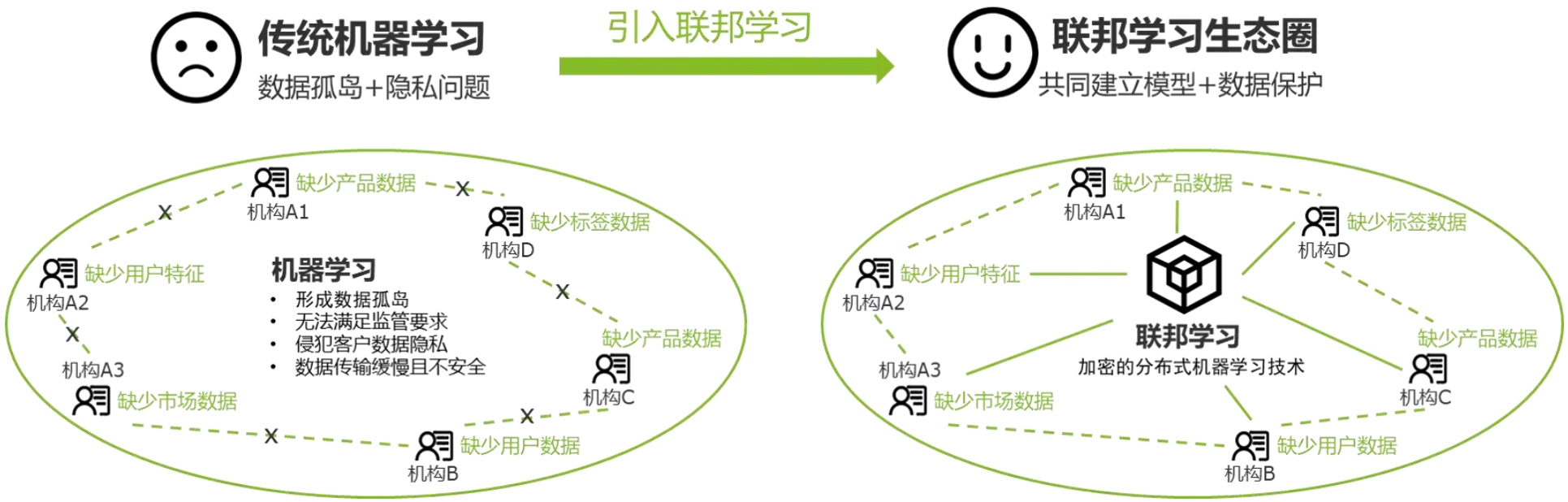
- 背景简介
- 基本概念
  - 联邦学习
  - 后门攻击
- 算法原理
  - COBA
  - FPCL
- 应用总结
- 参考文献

背景简介

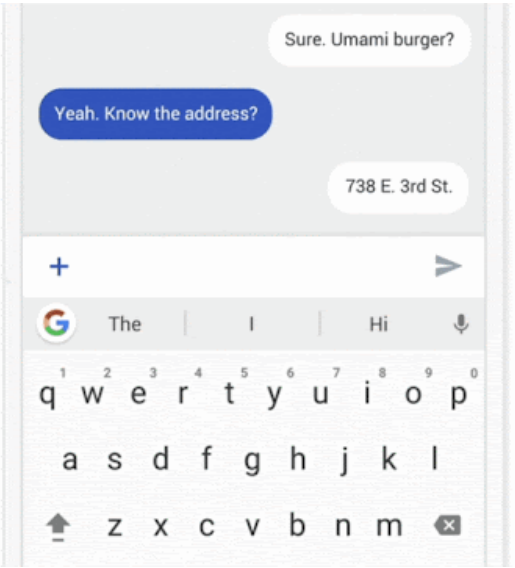
- 预期收获
  - 熟悉联邦学习的历史现状、分类及应用场景
  - 了解联邦学习与后门攻击的基本概念
  - 理解横向和纵向联邦学习的后门攻击方法
  - 了解联邦学习领域后门攻击算法的未来发展

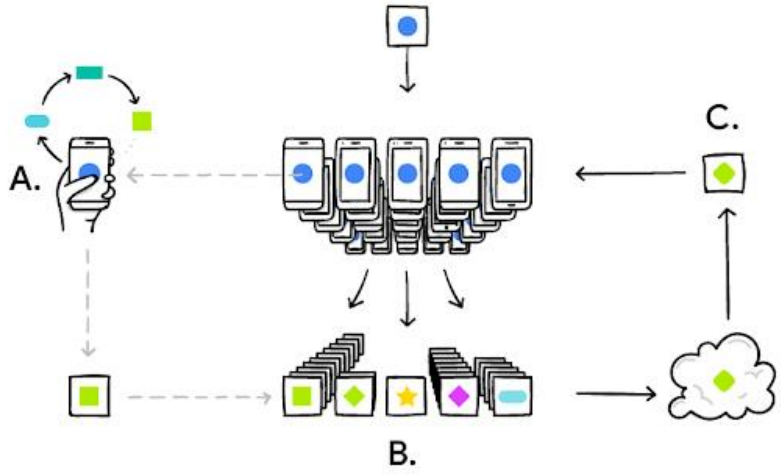
背景简介

- 人工智能的现状
  - 机器学习技术的成功，尤其是深度学习的应用，都是建立在丰富的数据资源上
  - 传统集中式机器学习的方式带来数据碎片化和孤岛分布的问题
  - 对人工智能应用中的数据隐私保护和所有权问题关注度不断提高



背景简介

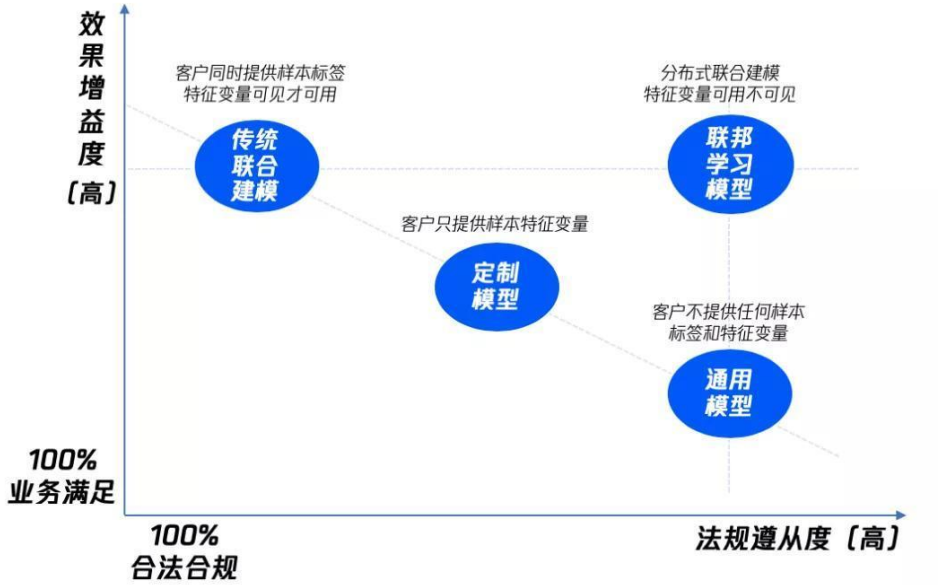




- 联邦学习（Federated Learning, FL）
  - FL由Google于2016年提出，解决多个移动设备的分布式建模问题
  - 例如Google Gboard安卓输入法预测，为了智能预测下一个词，对大量用户的输入历史数据进行训练
  - 设计目标：避免直接收集用户的输入历史，尽量在设备端上训练

背景简介

- 技术价值
  - 数据不出本地
    - 模型的训练与聚合不泄漏用户的个人隐私
  - 提高模型的泛化能力
  - 分布式训练
    - 提升 AI 模型训练效率和资源利用效率
- 开源框架

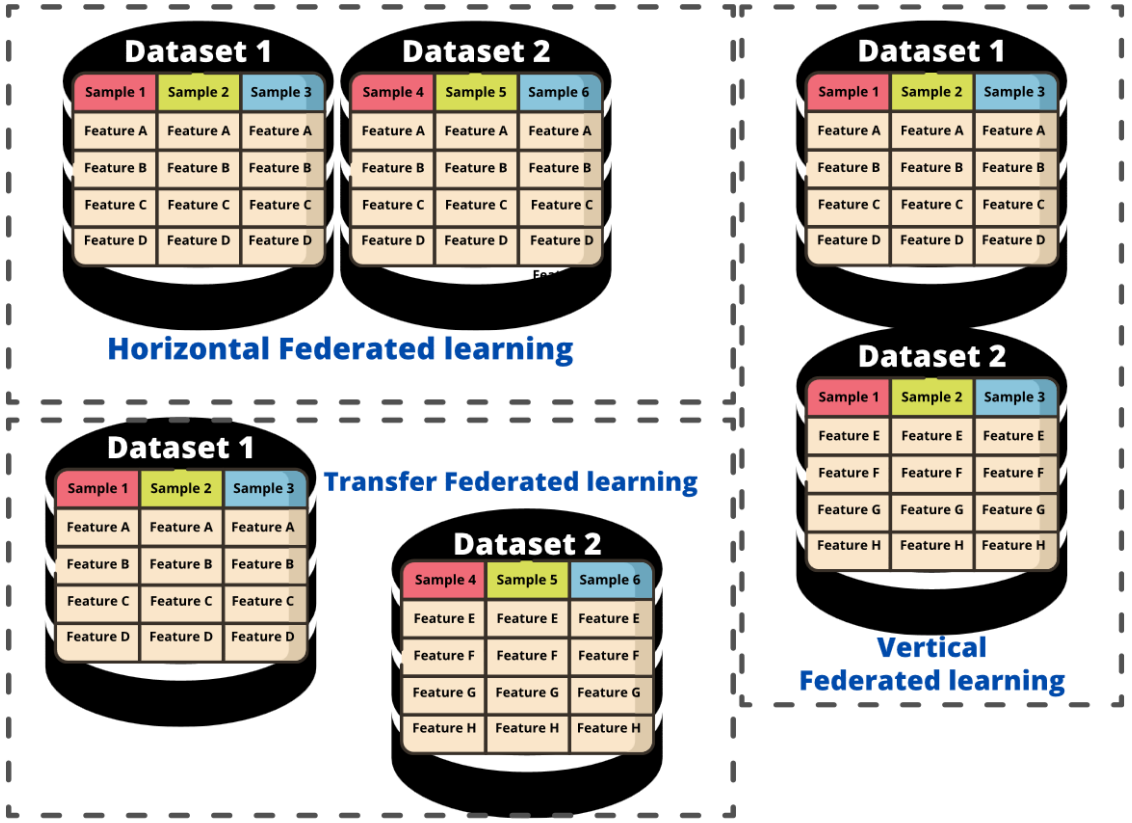


框架	主导方	状态
TensorFlow Federated	谷歌	开源
Pysyft	OpenMinded	
FATE	微众银行	
PaddleFL	百度	
Fedlearner	字节跳动	
Fedlearn	京东	

基本概念

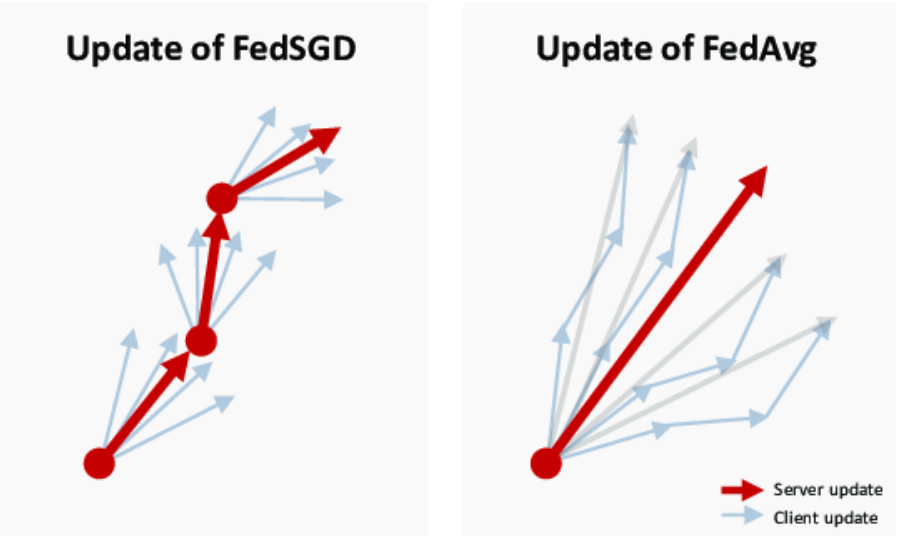
基本概念

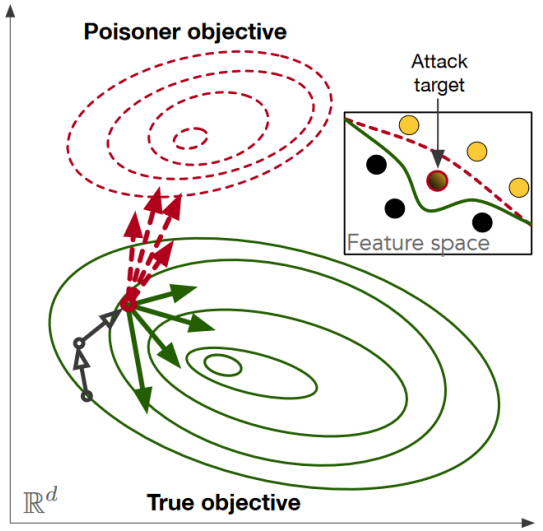
- 联邦学习的分类
  - 依据数据集的样本与特征在各客户端的数据分布情况
    - 横向联邦学习
      - 样本的联合，横向联邦的数据集间用户特征重合度高、用户重合度低
    - 纵向联邦学习
      - 特征的联合，纵向联邦的数据集间用户特征重合度低、用户重合度高
    - 联邦迁移学习
      - 在任何数据分布、任何实体上，均可以进行协同建模学习



基本概念

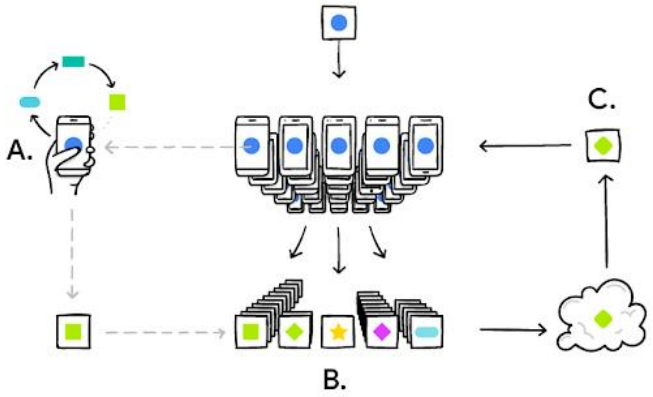
- 参数更新机制
  - FedSGD
    - 每轮在随机选择的客户端上进行一次梯度计算，对服务器共享梯度
  - FedAvg
    - 客户端在本地执行多次更新，对服务器共享模型参数
  - FoolsGold
    - 计算来自客户端的梯度更新的余弦相似度，识别其中潜在的恶意更新
  - …
- 参数安全性
  - 同态加密：允许在加密之后的密文上直接进行计算，且解密后的计算结果与基于明文的计算结果一致

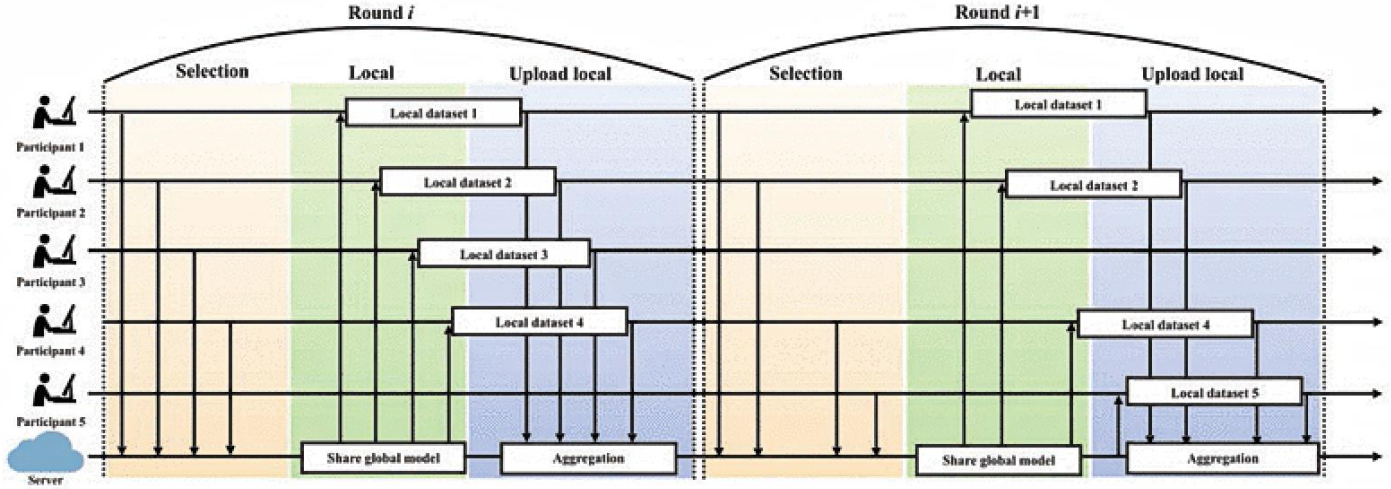




基本概念

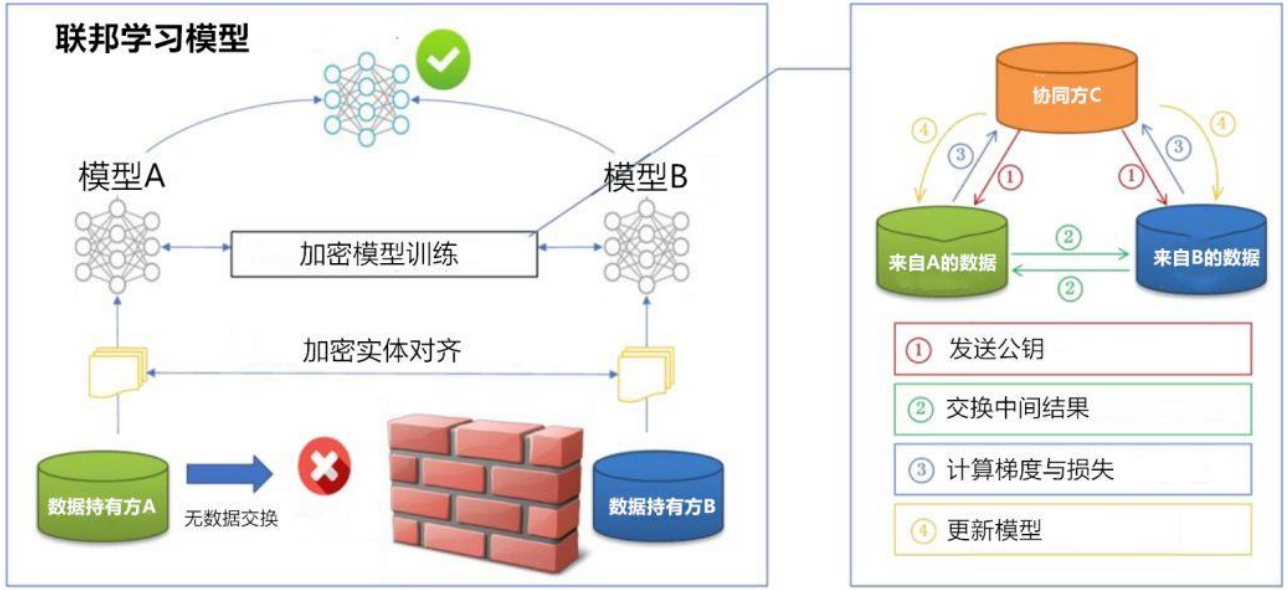
- 具体流程
  - 横向联邦
    - 服务器进行客户端选择，初始化全局模型并下发
    - 各客户端使用本地数据进行训练，并将本地模型参数或梯度加密上传
    - 服务器安全聚合各客户端传递的参数或梯度，更新全局模型
    - 服务器下发更新后的全局模型给各客户端，各联邦参与客户端使用本地数据对更新后的全局模型做模型评估，并将评估结果上报给中央服务器



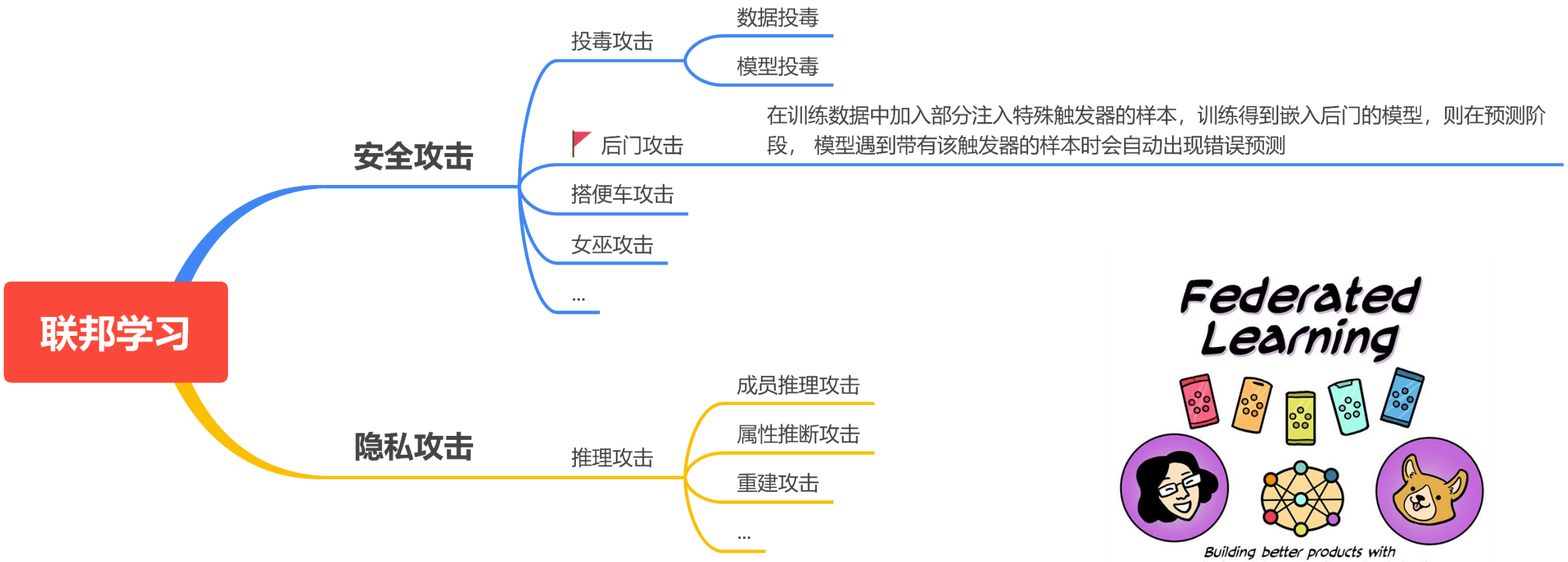


基本概念

- 具体流程
  - 纵向联邦
    - 样本加密对齐，多参与方基于加密的用户ID对原始数据安全求交
    - 联合特征工程，基于对齐后的数据，多参与方联合开展特征工程（可选）
    - 加密模型训练，通过隐私加密技术（同态加密）协同训练模型



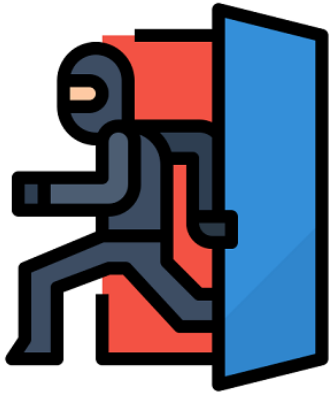
基本概念


联邦学习
安全攻击
投毒攻击
数据投毒
模型投毒
后门攻击
在训练数据中加入部分注入特殊触发器的样本，训练得到嵌入后门的模型，则在预测阶段， 模型遇到带有该触发器的样本时会自动出现错误预测
搭便车攻击
女巫攻击
...
隐私攻击
推理攻击
成员推理攻击
属性推断攻击
重建攻击
...

Federated Learning
Building better products with on-device data and privacy by default
An online comic from Google AI

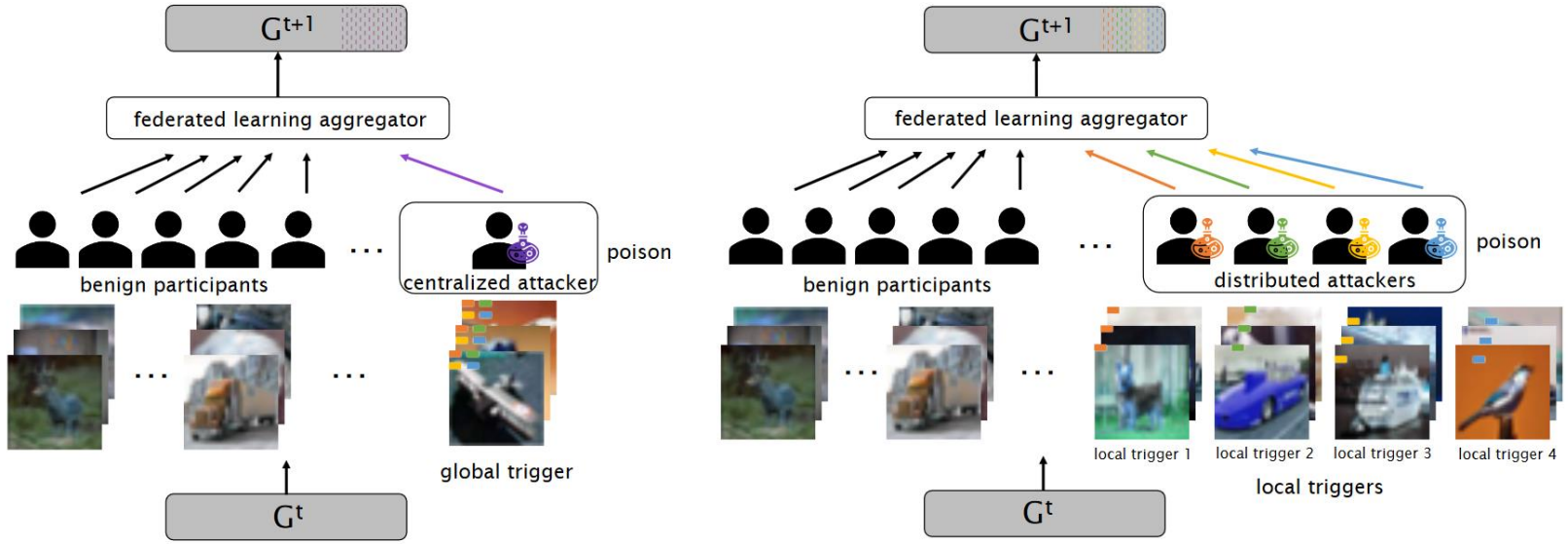
基本概念

- 后门攻击
  - 定义
    - 攻击者意图让模型对具有某种特定特征的数据做出错误的判断，但模型不会对主任务产生影响
  - 目标
    - 主任务精度(Main task accuracy, MTA)
      - 为了通过服务器上验证数据的正确率检测，需要保持模型在主任务（干净数据）上的良好性能
    - 攻击成功率（Attack success rate, ASR）
      - 降低全局模型在目标任务上的性能，确保预定义的输入都有很高的概率被预测为目标类别
    - 隐蔽性
      - 恶意攻击者需要确保其模型参数不会与良性模型偏离太多，以绕过基于检测并剔除恶意更新的防御措施



基本概念

- 学术报告回顾
  - 充分利用联邦学习的分布式学习的特点
  - 将全局触发模式分解成局部模式并分别将其嵌入不同攻击者的本地数据集中



算法原理

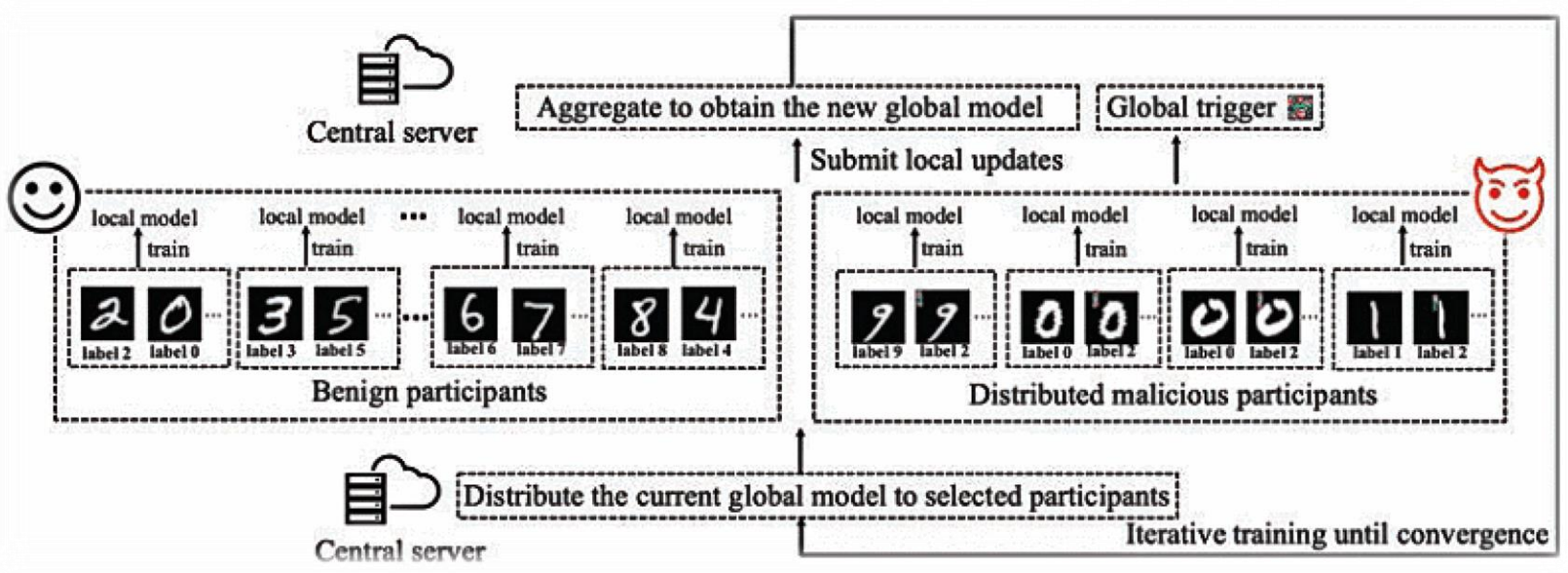
算法原理

T	针对横向联邦学习实现基于模型依赖触发器的分布式协同后门攻击
I	全局模型参数与恶意参与方
P	1、模型依赖本地触发器生成 2、协同式后门注入
O	植入后门的全局模型

P	现有后门触发器的特征明显，易被检测
C	攻击者控制多个恶意参与方 恶意参与方在全局模型接近收敛时生成本地触发器
D	如何选择后门神经元，确定模型依赖触发器的优化方式
L	IEEE Network（中科院SCI 1区）

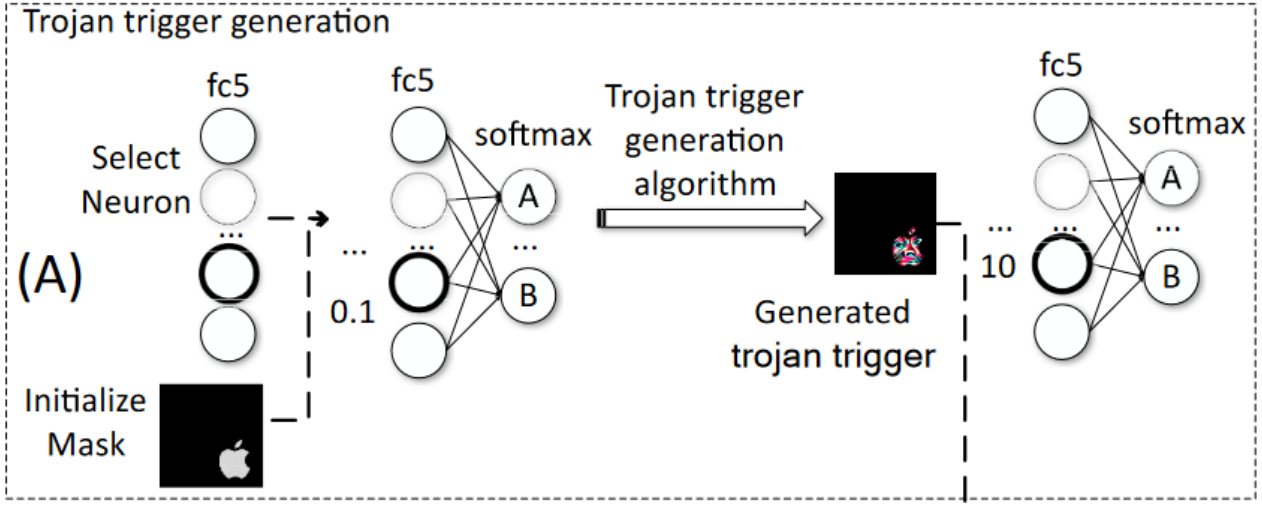
算法原理

- 基于模型依赖触发器的分布式协同后门攻击（COBA）
  - 模型依赖的本地触发器生成
    - 确定触发器的形式
    - 触发器生成与优化，激活特定的神经元
  - 协同式后门注入
    - 确定投毒间隔与中毒参数缩放因子



算法原理

- 模型依赖的本地触发器生成
  - 经过实验确定触发器的结构，包括形状、大小和位置
    - 位于图片边角的矩形触发器（与现有后门攻击方法一致）
    - 触发器的大小与输入数据集成正比
  - 触发器生成与优化
    - 构建与输入数据具有相同结构的掩模（Mask），其触发器区域值随机初始化，触发器生成过程相当于寻找掩模的最优化赋值过程



算法原理

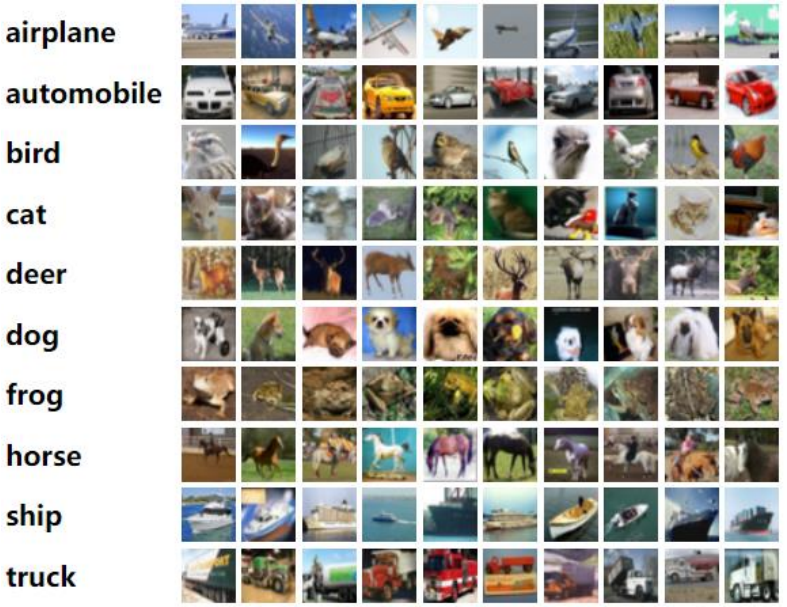
- 后门触发器生成算法
  - 掩模初始化
    - 掩模M是布尔值矩阵，与模型的输入具有相同的维数
  - 触发器关联的神经元选择
    - 激活次数和下一层权重值总和最大
  - 损失函数定义
    - 当前神经元的值与其目标值之间的均方误差
  - 损失函数优化
    - 使用梯度下降法使得损失函数最小化，其中目标值为所选层中神经元的最大值

Algorithm 1 Trojan trigger generation Algorithm

1: **function** TROJAN-TRIGGER-GENERATION(model, layer, M, {(neuron1, target_value1), (neuron2, target_value2), ... }, threshold, epochs, lr)
2: $f = model[: layer]$
3: $x = MASK\_INITIALIZE(M)$
4: $cost \overset{\text{def}}{=} (target\_value1 - f_{neuron1})^2 + (target\_value2 - f_{neuron2})^2 + \dots$
5: **while** $cost < threshold$ and $i < epochs$ **do**
6: $\Delta = \frac{\partial cost}{\partial x}$
7: $\Delta = \Delta \circ M$
8: $x = x - lr \cdot \Delta$
9: $i++$
return x

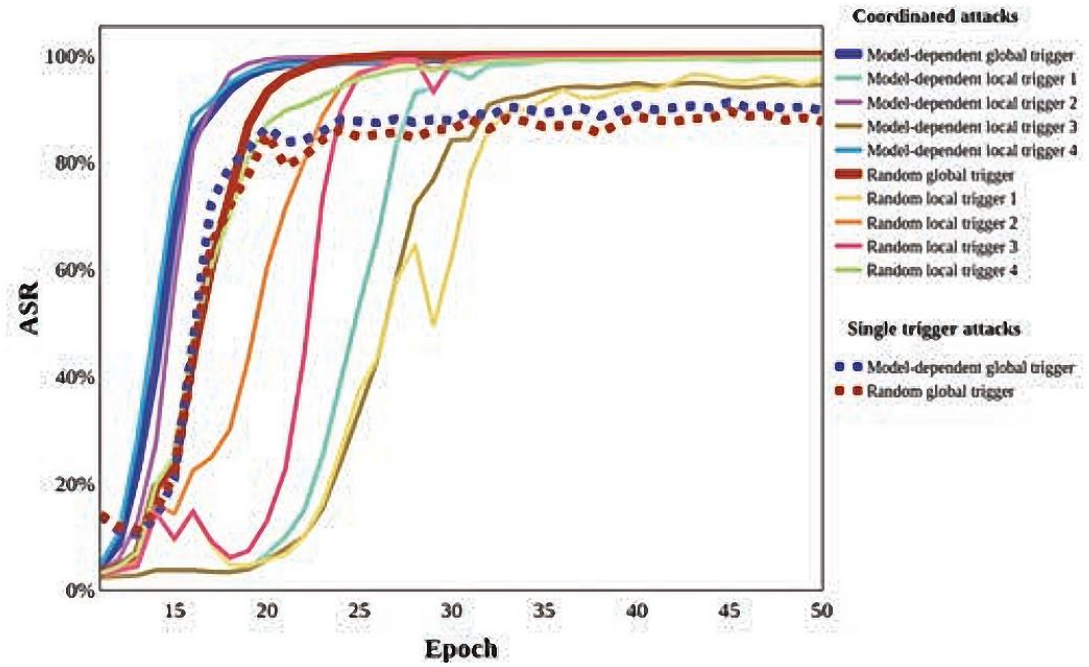
算法原理

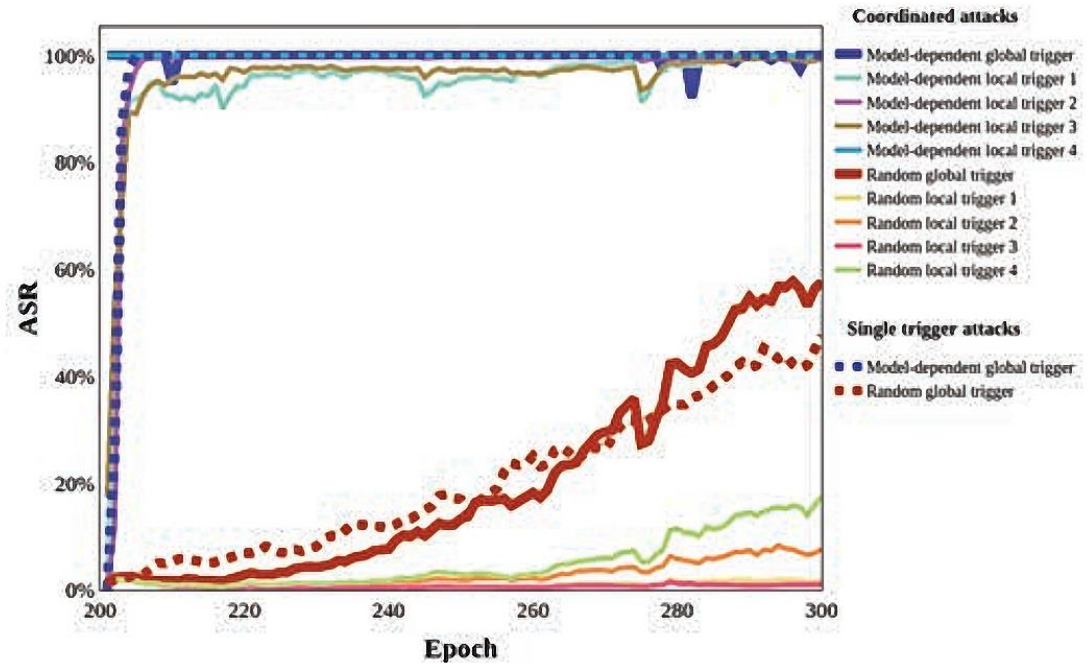
- 实验数据
  - 手写字符识别MNIST
    - 由60000个训练样本和10000个测试样本组成，每个样本都是一张28 * 28像素的灰度手写数字图片
  - 目标识别CIFAR-10
    - 由60000张32x32像素的彩色图像组成，分为10类，每类有6000张图像，有50000张训练图像和10000张测试图像。
- 对比方法
  - 随机触发器的分布式后门攻击DBA
  - 单触发器的后门攻击



算法原理

- 实验结果

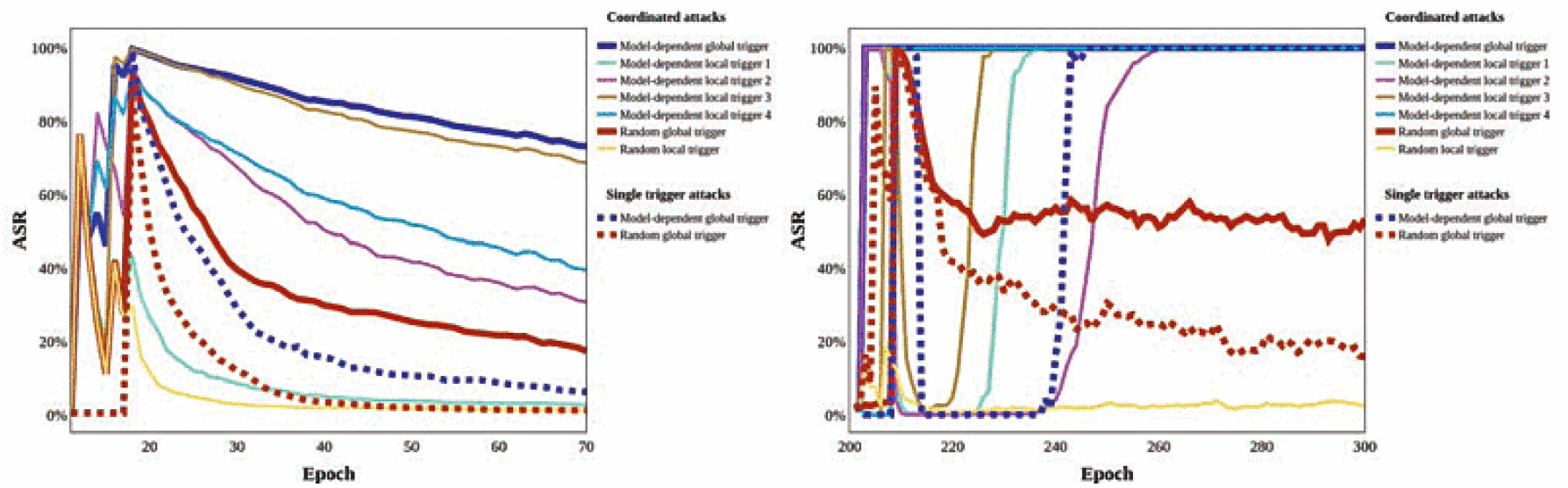




- 多轮攻击（Multiple Shot）场景，分布式协同后门攻击方法的成功率始终高于单触发器后门攻击
- 模型依赖触发器的后门攻击效果优于随机触发器

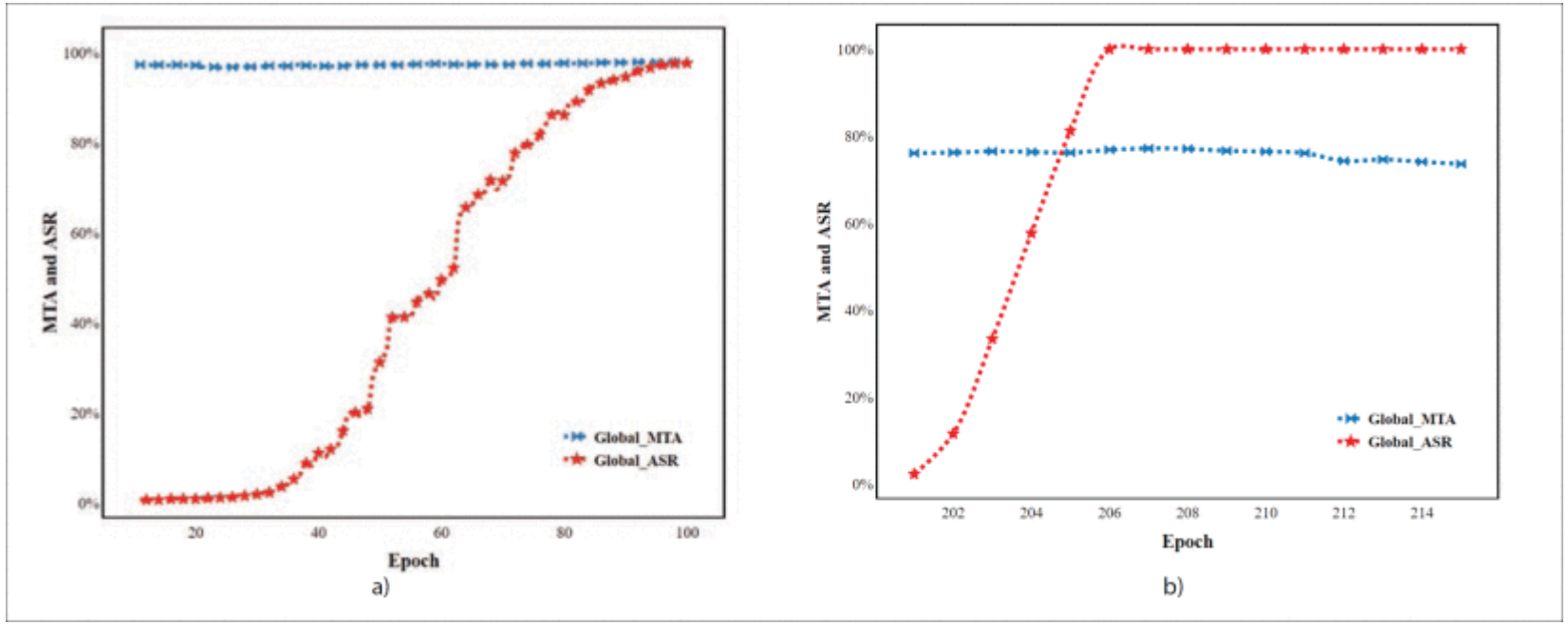
算法原理

- 实验结果
  - 单轮攻击（Single Shot）场景，主要测试后门攻击的持久性
  - 后门注入后，COBA方法的成功率下降速度小于DBA和单触发后门攻击
    - MNIST：50个训练轮次后，COBA方法的攻击成功率依然高于81.2%；而DBA方法和单触发后门攻击的成功率分别为25.4%和10.6%
    - CIFAR-10：存在攻击成功率急剧下降然后上升的情况，这是由于正常参与方的学习率比较高



算法原理

- 实验结果
  - 拜占庭弹性聚合方法FoolsGold的作为联邦学习的聚合协议
  - 单轮（ Single Shot ）攻击场景， COBA在基本不影响主任务准确率的情况下，实现较高的后门攻击成功率



优劣分析

- 优势
  - COBA算法属于分布式后门攻击领域使用模型依赖触发器的早期工作
  - 模型依赖触发器的后门攻击成功率优于随机触发器
  - COBA算法的后门攻击更加隐蔽，后门持久性强
- 劣势
  - 激活后门神经元的选取规则，可利用神经元的关联关系

算法原理

算法原理

T	实现针对纵向联邦学习的后门攻击
I	各参与方初始模型参数、恶意参与方
P	对无任何标签信息的被动方进行梯度信息推理 对已知样本标签信息的梯度进行投毒
O	植入后门的恶意参与方模型

P	后门攻击方法大多针对横向联邦学习，不适用特征分区的纵向联邦
C	对主动方的训练模型和训练算法有过强假设 攻击者至少已知一个样本的标签信息
D	如何由梯度反推标签信息，计算中间梯度信息与样本标签的对应关系
L	ICML 2020 A类会议

算法原理

- 纵向联邦学习框架
  - 条件：K个参与方基于N条样本数据协作训练机器学习模型
  - 特征向量x可分解为K个区块，纵向联邦学习优化问题：

$$o_i^K = f(\theta_1, \ldots, \theta_K; \mathbf{x}_i^1, \ldots, \mathbf{x}_i^K)$$

$$\min_\theta \mathcal{L}(\theta; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \ell(o_i^K, y_i^K) + \lambda \sum_{k=1}^{K} \gamma(\theta_k)$$

  - θ代表各参与方的模型参数，$f(\cdot)$ 和$\gamma(\cdot)$分别代表预测函数和正则项

$$H_i^k = G_k(\theta_k, \mathbf{x}_i^k)$$

  - G_k代表第k个参与方本地训练模型，H_i^k代表本地潜在表示

$$\ell(\theta_1, \ldots, \theta_K; \mathcal{D}_i) = \ell(f(\sum_{k=1}^{K} H_i^k), y_i^K)$$

算法原理

- 纵向联邦算法
  - 输入：学习率
  - 输出：各参与方已训练模型参数
  - 各被动方训练本地潜在表示H_i^k，并发送至主动方
  - 主动方求和得到$H_i = \sum_{k=1}^{K} H_i^k$，然后根据标签信息计算梯度

$$\nabla_k \ell(\theta_1, \ldots, \theta_K; \mathcal{D}_i) = \frac{\partial \ell}{\partial H_i} \frac{\partial H_i^k}{\partial \theta_k}$$

  - 被动方根据梯度信息计算本地参数更新

Algorithm 1 A feature-partitioned collaborative learning framework

Input: learning rate η

Output: Model parameters $\theta_1, \theta_2 \ldots \theta_K$ Party 1,2,...,K, initialize $\theta_1, \theta_2, \ldots \theta_K$.

for $j = 1$ **to** N **do**

  Randomly sample $S \subset [N]$

  for each party $k \neq K$ in parallel **do**

    k computes and sends $\{H_i^k\}_{i \in S}$ to party K

  end for

  party K computes and sends $\{\frac{\partial \ell}{\partial H_i}\}_{i \in S}$ to all other parties;

  for each party k=1,2,...,K in parallel **do**

    k computes $\nabla_k \ell$ with equation 5

    update $\theta_k^{j+1} = \theta_k^j - \eta \nabla_k \ell$;

  end for

end for

算法原理

- **梯度信息推理**
  - 被动方不知道任何标签信息，可以从中间梯度$\frac{\partial L}{\partial H_i}$获得样本的标签信息

$$\ell(\theta_1,\ldots,\theta_K;\mathcal{D}_i)=\ell(f(\sum_{k=1}^{K}H_i^k),y_i^K)$$

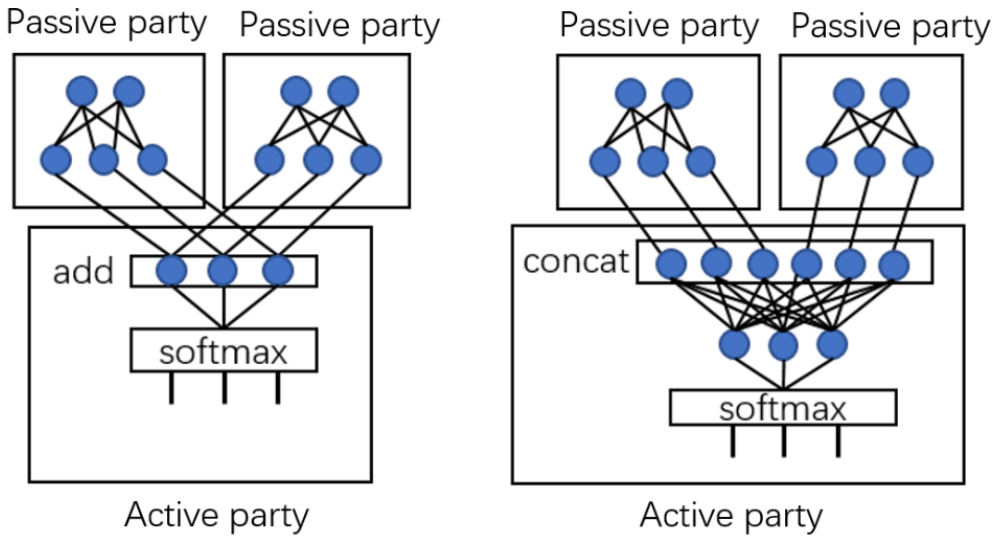
  - 条件：损失函数是交叉熵函数，且输出层使用softmax函数

$$g_{H,j}^{i}=\begin{cases}S_j, j\neq y\\ S_j-1, j=y\end{cases}$$

  - 注入攻击，被动方将真正标签对应的分量+1，将目标标签对应的分量-1

$$g_{H,j}^{b,i}=\begin{cases}S_j, j\neq \tau\\ S_j-1, j=\tau\end{cases}$$

- **防御方法**
  - 向主动方添加可训练层
    - 在输出层之前加上一个具有32个节点的密集层
    - 减少标签信息的泄露，提高模型性能



算法原理

- **梯度投毒算法**
  - 攻击者至少知道一个样本的标签信息，且与后门任务的目标标签相同
  - 攻击者用已知样本的梯度信息去替换要投毒样本的梯度信息，使之被分为攻击者指定的目标类别

$$\nabla_k \ell(\theta_1, \ldots, \theta_K; \mathcal{D}_i) = \frac{\partial \ell}{\partial H_i} \frac{\partial H_i^k}{\partial \theta_k}$$

  - 在不推理恢复未知数据标签的情况下进行后门攻击

Algorithm 2 Gradient poisoning algorithm

Input: current batch of training dataset $\mathcal{X}$, original gradient matrix $\frac{\partial \ell}{\partial H}$ that the active party sends to the malicious party, dataset $\mathcal{D}_{target}$, $\mathcal{D}_{poison}$, the amplify ratio of the poisoned gradient γ

Output: poisoned gradient matrix $G_{poisoned}$

$G_{poisoned} = \frac{\partial \ell}{\partial H}$

for x_i in $\mathcal{X}$ **do**

  if x_i belongs to $\mathcal{D}_{target}$ **then**

    assign the ith row of $\frac{\partial \ell}{\partial H}$ to a global variable g_{rec}

  end if

  if x_i belongs to $\mathcal{D}_{poison}$ **then**

    assign $\gamma \cdot g_{rec}$ to the ith row of $G_{poisoned}$

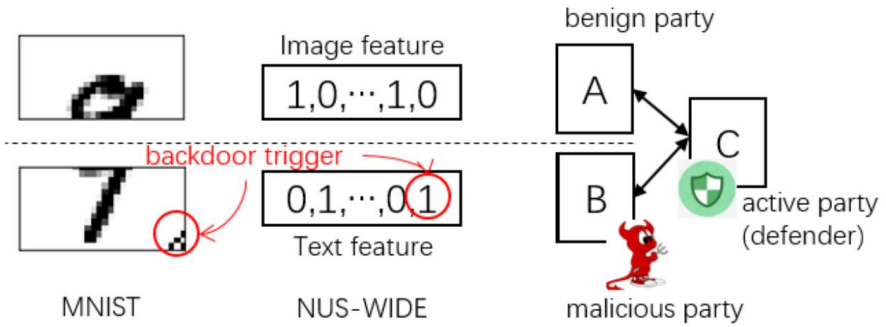
  end if

end for

return $G_{poisoned}$

算法原理

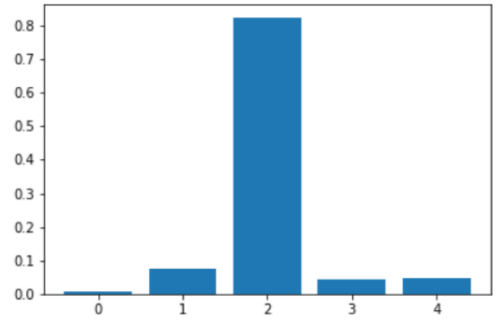
- 实验模型
  - 两个的被动方和一个拥有标签的主动方
  - 使用梯度投毒算法实施后门攻击任务
  - 被动方：一个输入层和输出层，输出层大小设为32
- 实验数据
  - NUS-WIDE
    - 选取5个标签，被动方分别获取图片和文本特征
  - MNIST
    - 被动方分别获取图片的前后14行



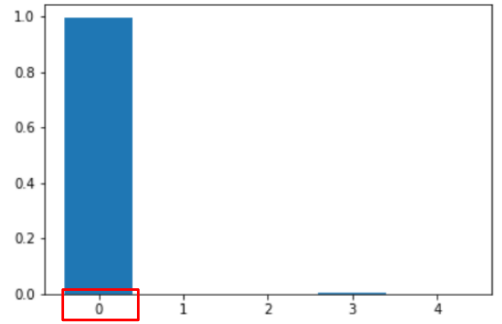
ID	Label	Train	Test
0	buildings	3 / 6173	0 / 4213
1	grass	8 / 5981	8 / 4082
2	animal	125 / 12679	81 / 8296
3	water	11 / 12202	4 / 8172
4	person	5 / 22965	9 / 15237

算法原理

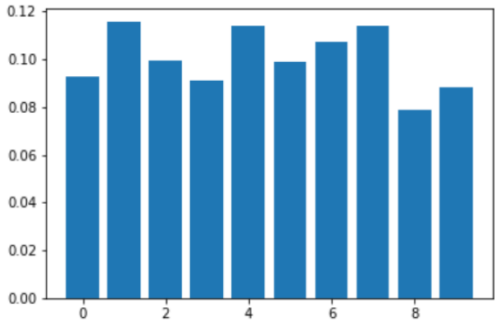
- 实验结果
  - 投毒数据正常与后门训练的预测准确率比较
  - 对于NUS-WIDE数据集，虽然投毒数据本身属于 class0 的样本很少，但是 class0 的预测分数远高于其他类
  - 综合分析两个数据集的结果，后门攻击成功率接近1.0



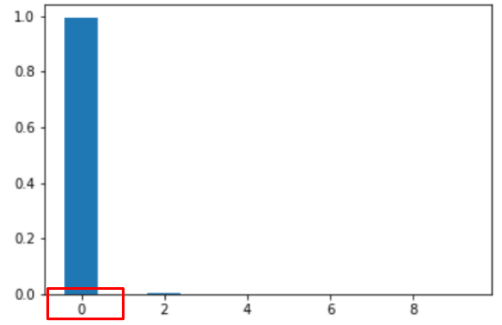
(a) normal(NUS-WIDE)



(b) poisoned(NUS-WIDE)



(c) normal(MNIST)

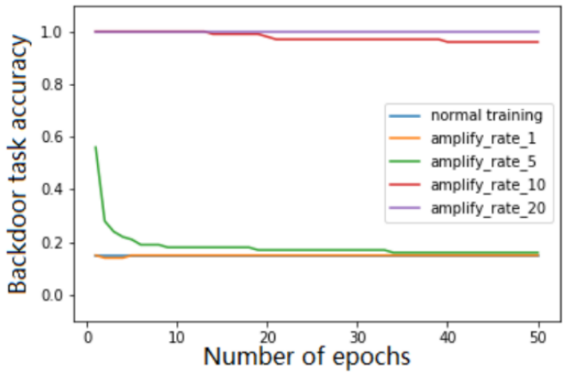


(d) poisoned(MNIST)

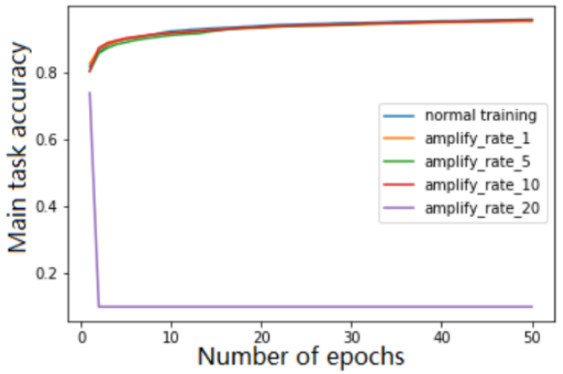
算法原理

- 实验结果
  - 测试阶段，使用投毒数据测试后门任务准确率；使用测试集检查主任务的预测精度
  - 梯度投毒算法的投毒缩放因子需设置合理，以同时保证主任务和后门任务的精度
  - 正常任务和后门任务之间的差距并不明显，因此使用验证集很难检测到这个后门

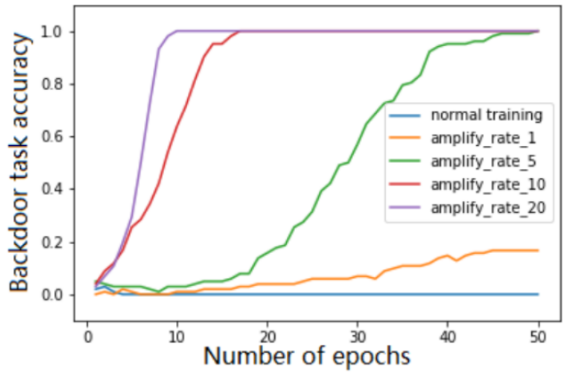
主动方未添加可训练层



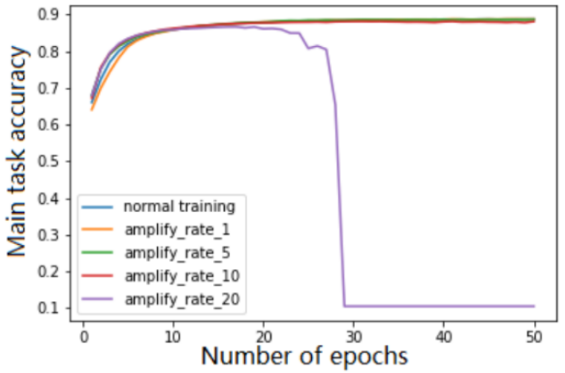
(a) MNIST



(b) MNIST



(c) NUS-WIDE



(d) NUS-WIDE

算法原理

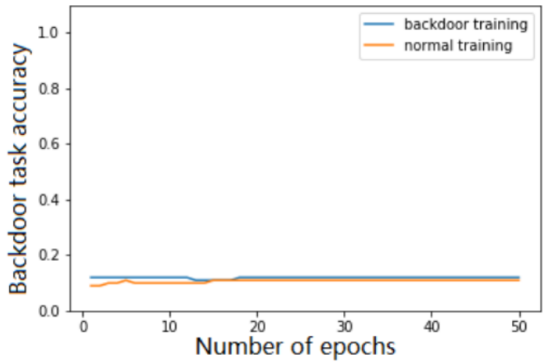
- 实验结果
  - MNIST数据集的后门攻击失败，NUS-WIDE攻击成功

Table 2. accuracy of different network.

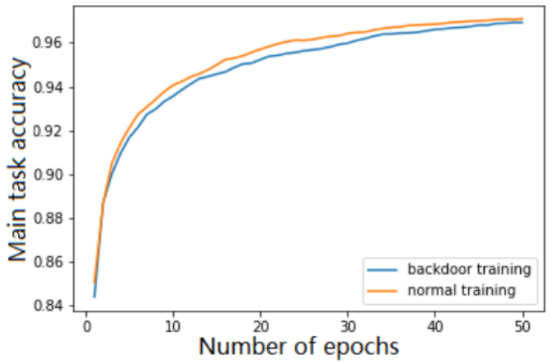
	MNIST	NUS-WIDE
untrainable active party	95.7%	88.5%
trainable active party	**97.1%**	**88.7%**

  - 向主动方加入可训练层，有助于减少标签信息的泄露，提高模型性能

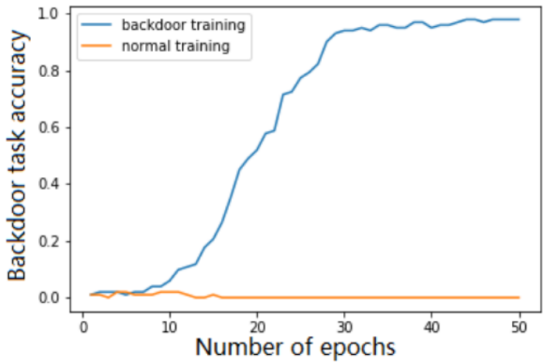
主动方添加可训练层



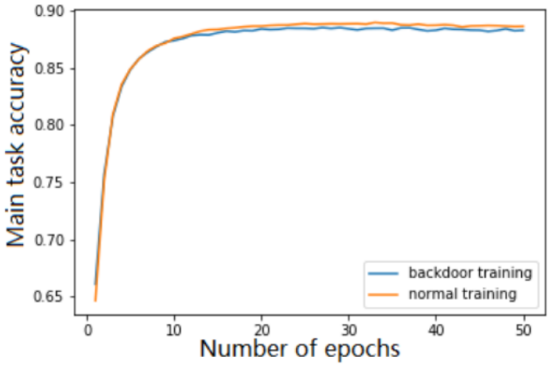
(a) MNIST



(b) MNIST



(c) NUS-WIDE



(d) NUS-WIDE

优劣分析

- 优势
  - 第一个在纵向联邦学习框架中处理后门攻击的系统性工作
  - 被动方可以在不知道标签的情况下执行后门任务
  - 可以在不推理恢复未知数据标签的情况下进行后门攻击
- 劣势
  - 对主动方的训练模型和训练算法有过强的假设
  - 需攻击者至少知道一个样本的标签信息，且与后门任务的目标标签相同
  - 未分析加密参数下的纵向联邦后门攻击方法

应用总结

应用总结

- 未来的发展
  - 针对横向联邦学习提升后门攻击方法的隐蔽性和攻击效果的持续性
    - 考虑触发器与模型的依赖关系
  - 针对纵向联邦学习设计有效地的后门攻击方法
    - 考虑加密情况的联邦后门攻击
  - 对后门攻击的目标进行扩展研究
    - 不局限于神经网络的分类功能，探索对神经网络的目标预测、目标检测等功能的后门攻击方法

参考文献

[1] Gong X, Chen Y, Huang H, et al. Coordinated Backdoor Attacks against Federated Learning with Model-Dependent Triggers[J]. IEEE Network, 2022, 36(1): 84-90.

[2] Liu Y, Yi Z, Chen T. Backdoor attacks and defenses in feature-partitioned collaborative learning[M]. arXiv, 2020.

[3] Liu Y , Ma S , Aafer Y , et al. Trojaning Attack on Neural Networks[C]// Network and Distributed System Security Symposium. 2017.

[4] **电信领域联邦学习技术应用白皮书，中国人工智能产业发展联盟**

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。