Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

联邦学习的后门防御方法

硕士研究生 杨得山

2023年04月09日

内容提要

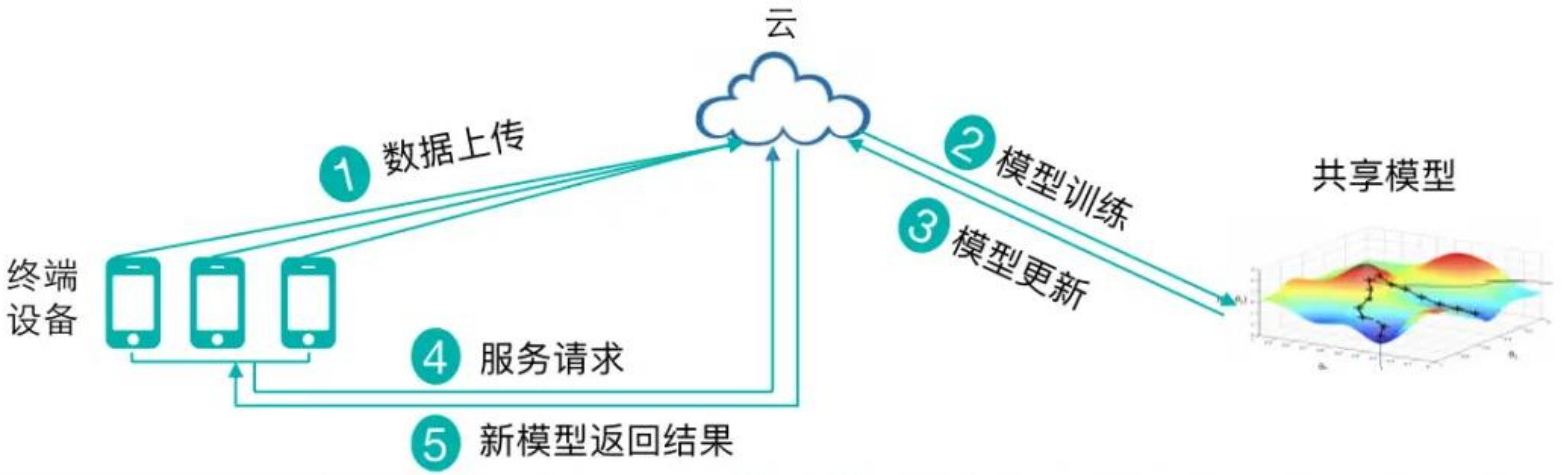
- 背景简介
- 基本概念
- 算法原理
  - RFOut-1d
  - FLAME
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 了解联邦学习后门攻击与防御的基本概念
  - 理解联邦学习后门防御的算法原理
  - 了解联邦学习后门攻击与防御发展方向

问题思考

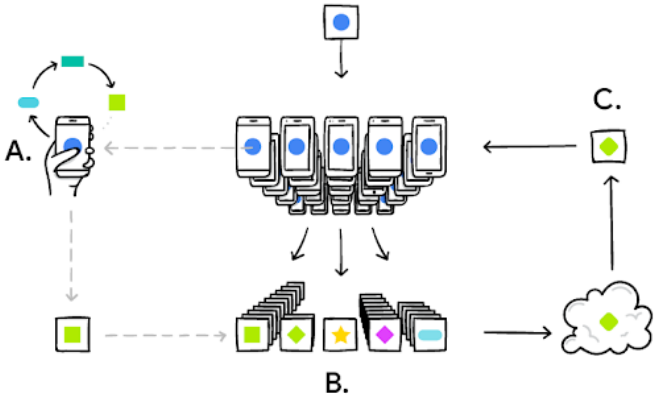
集中式模型训练

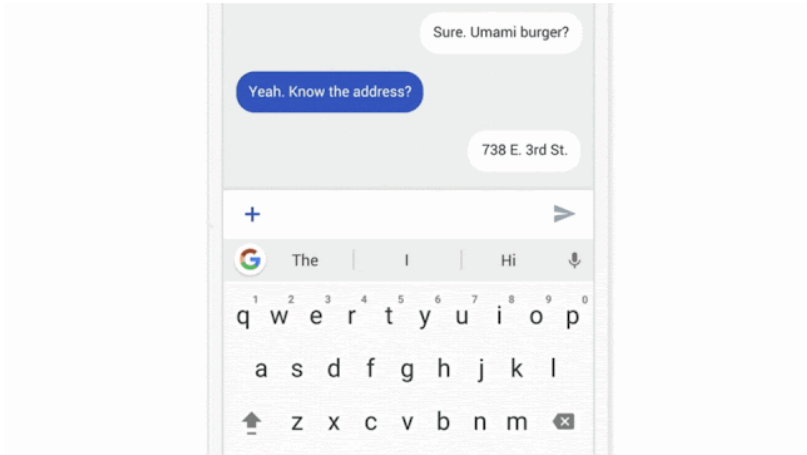


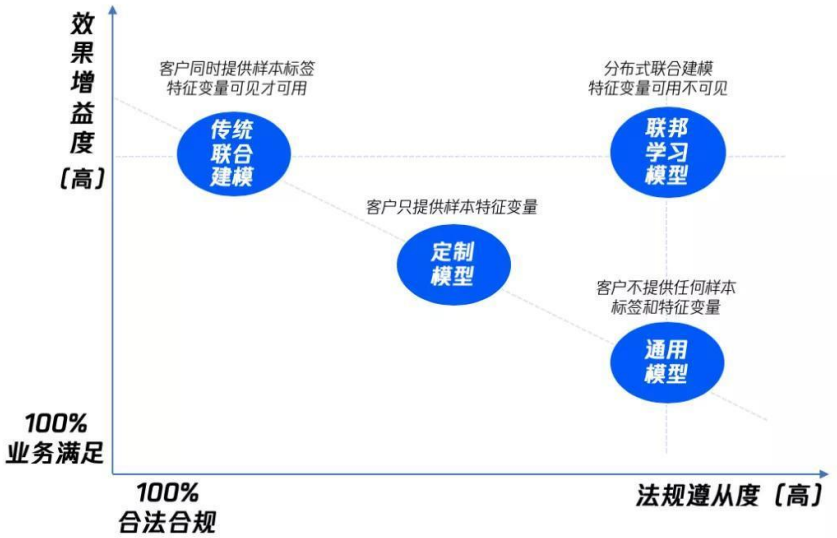
上述训练方式有问题吗?

背景简介

- 人工智能技术的不断发展
  - 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，无法充分利用数据资源
  - 数据安全问题时有发生，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



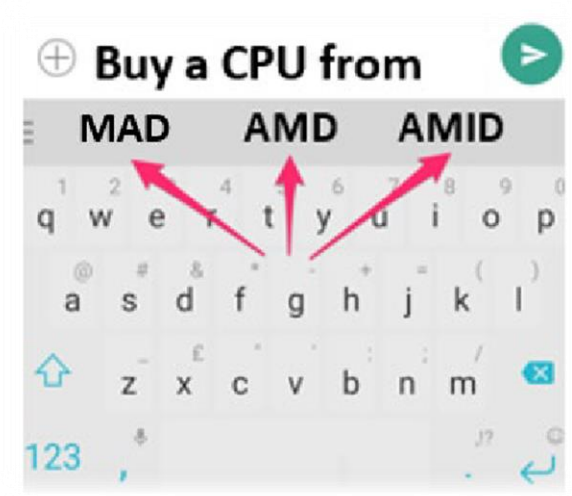




构建安全桥梁，释放数据价值！

背景简介

- 联邦学习的安全性问题
  - 分布式特性引入新的攻击面，在许多不可信的设备上训练模型，如何保护模型安全?
  - 攻击者能够将“后门”或“木马”植入到模型中，并在预测阶段通过简单的后门触发器完成恶意攻击行为



单词预测

停车标志及其受后门攻击的版本

发展的眼光看待技术方法!

基本概念

- 联邦学习（Federated Learning, FL）
  - 联邦学习的参与方在保留数据本地化的前提下，通过只交换模型训练中间结果，如模型参数、梯度等，实现多方联合的机器学习训练
  - 优势
    - 数据不出本地，模型的训练与聚合不泄漏用户的个人隐私
    - 分布式训练，提升 AI 模型训练效率和资源利用效率

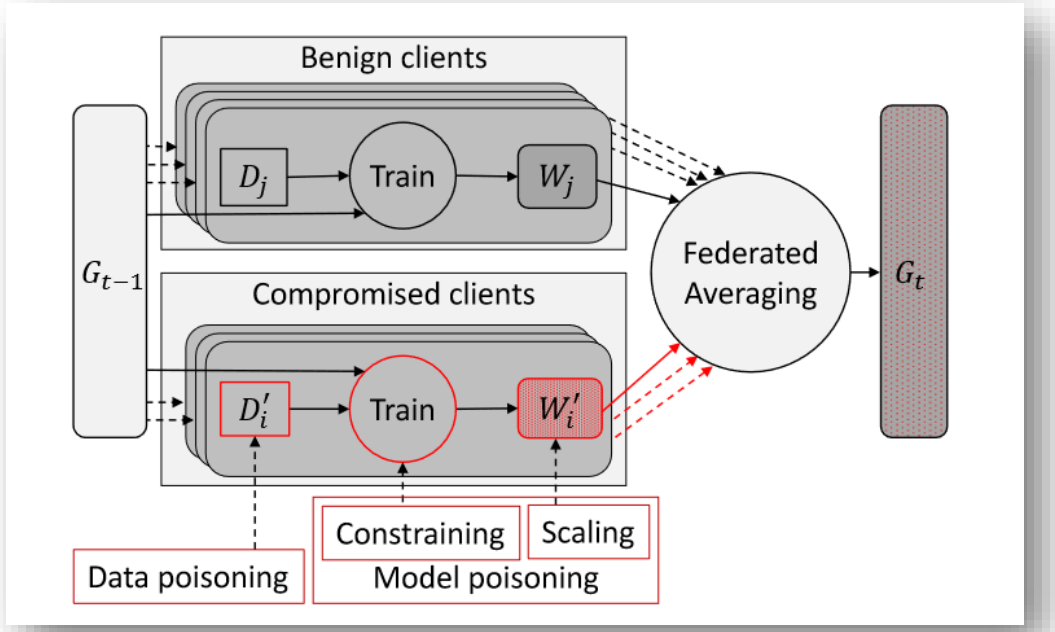
- 后门攻击（Backdoor Attack, BA）
  - 攻击者意图让模型对具有某种特定特征的数据做出错误的判断，但不会对模型主任务产生影响
  - 追求可控性与隐蔽性

BFS最新相关报告：基于模型修改的深度学习后门攻击（吴肖龙）

基本概念

- 联邦学习原理
  - 正常更新
    - $G^t = G^{t-1} + \frac{\eta}{n}\sum_{i=1}^{n}\left(L_i^t - G^{t-1}\right)$
    - η代表服务端全局模型的学习率，n表示每次聚合选定客户端数目
  - 后门更新
    - $G^t = G^{t-1} + \frac{\eta}{n}\beta(L_{adv}^t - G^{t-1}) + \frac{\eta}{n}\sum_{i=2}^{n}\left(L_i^t - G^{t-1}\right)$
    - 其中$\beta = \frac{n}{\eta}$表示用于模型替换的提升因子

$$G^t \approx G^{t-1} + \frac{\eta}{n}\frac{n}{\eta}(L_{adv}^t - G^{t-1}) = L_{adv}^t$$

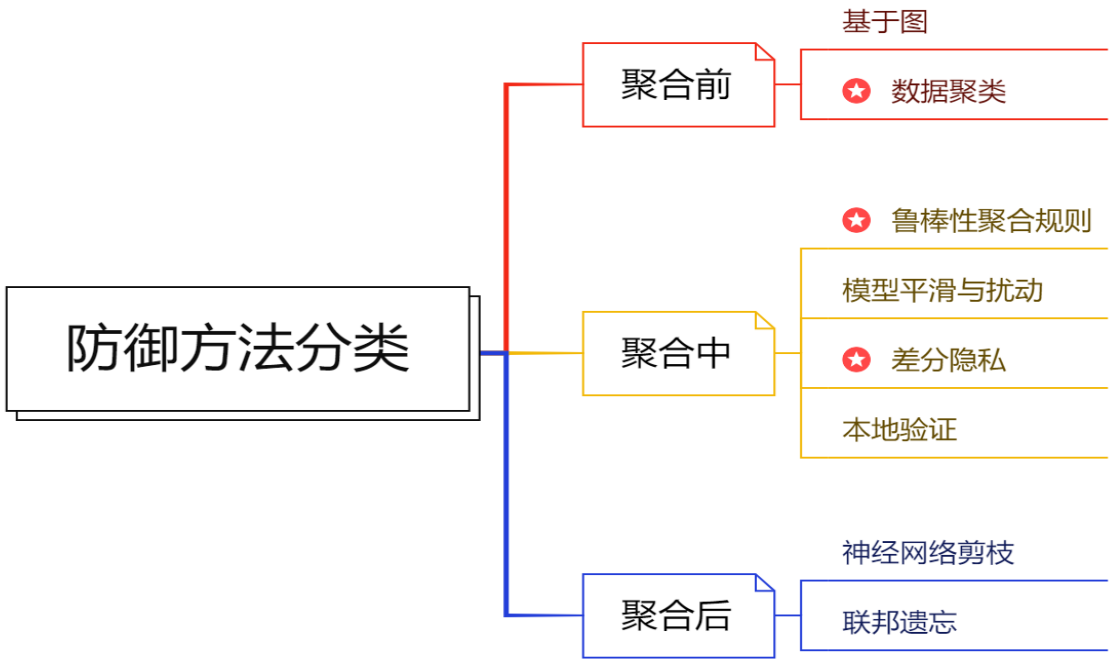


后门攻击流程示例

理想后门攻击效果：全局模型收敛并被后门模型替代

基本概念

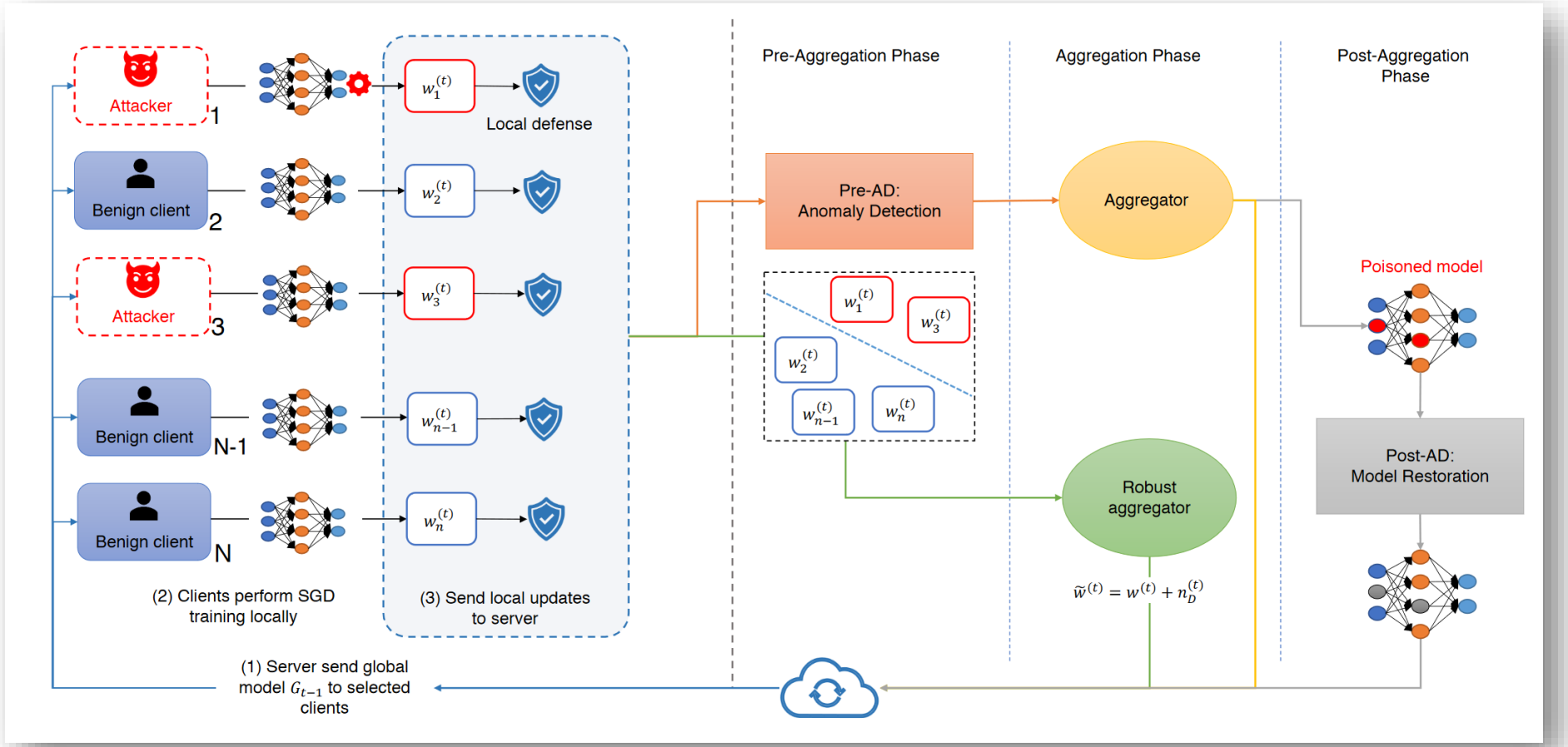
- 联邦学习的后门防御
  - 模型更新参数聚类
    - 目的：区分良性与恶意的模型更新
    - 一般对数据分布有假设，比如是否为独立同分布
  - 鲁棒性聚合规则
    - 目的：削弱模型更新中异常值的影响
    - 无法抵御高隐蔽性的后门注入
  - 差分隐私
    - 权重裁剪与加噪以消除高隐蔽性后门
    - 需要控制裁剪与加噪的比例，在保证正常任务精度的前提下消除后门



多种防御方法配合食用，防御效果更佳！

基本概念

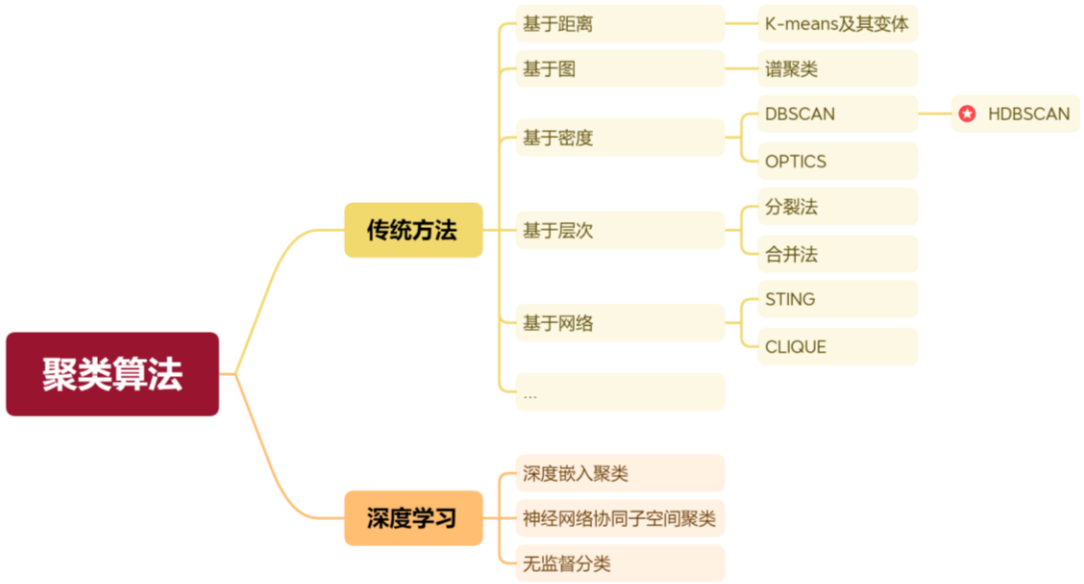
- 联邦学习的后门防御流程



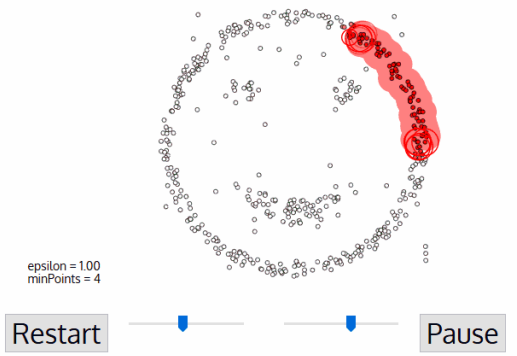
基本概念

适合自己的方法就是好方法！

- **HDBSCAN**
  - 密度聚类算法，在给定数据集中自动识别最佳簇的数量
  - 有效处理任意形状的簇
  - 使用稳健的单链接距离度量各聚类的距离



- **差分隐私 Differential Privacy**
  - 属于密码学手段，用来防范差分攻击的
  - 通过原始训练数据估计加噪参数，处理数据时添加噪声，使个人数据无法被准确识别，但保证统计特征较为准确



BFS最新相关报告：敏感文本数据脱敏方法（关业礼）

算法原理——RFOut-1d

【Knowledge-Based Systems】
Backdoor attacks-resilient aggregation based on Robust Filtering of Outliers in federated learning for image classification

算法原理——RFOut-1d

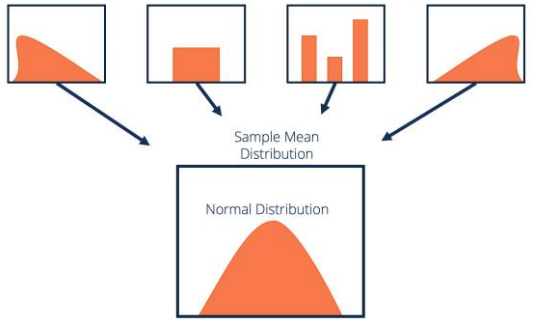
T	抵御联邦学习的后门攻击
I	良性与恶意客户端样本集，三种图像数据集
P	1. 将所有客户端上传的模型参数更新平展为一维张量并对齐 2. 按照正态分布的3σ原则，替换每一维度的异常参数 3. 所有客户端的参数聚合，并加入梯度裁剪和差分隐私
O	避免后门攻击的联邦学习全局模型

P	联邦学习的分布式特性易受对抗攻击
C	联邦学习设定下的各客户端模型参数更新符合高斯分布
D	异常参数的判定与消除原则
L	KBS 2022

算法原理——RFOut-1d

- 算法思想
  - 联邦学习下客户端的参数更新符合**正态分布**
  - 攻击者因同时具有主任务与后门任务优化目标，导致其参数更新属于**异常值**

中心极限定理



- 算法步骤
  - n个客户端的参数更新**展平**，每维度对齐
  - 计算每一维度对应的均值和标准差，根据**3σ原则**确定并替换异常值

$$G_t[i] = \frac{1}{n}\sum_{i=1}^{n} \hat{L}_j^t[i] \qquad \forall i \in \{1, \dots, m\}, \text{ where}$$

$$\hat{L}_j^t[i] = \begin{cases} \mu_i, & \text{if } abs\left(L_j^t[i] - \mu_i\right) \geq \delta\sigma_i \\ L_j^t[i], & \text{otherwise,} \end{cases} \qquad j \in \{1, \dots, n\}$$

Algorithm 1 RFOut-1d

Input: local updates $\{L_1^t, L_2^t, \dots, L_n^t\}$
$num\_dimensions = length(L_1^t)$
Initialize G^t
$\delta = 3$
for $i = 0$ **to** $num\_dimensions$ **do**
  $\hat{L}_i = (L_1^t[i], L_2^t[i], \dots, L_n^t[i])$
  $\mu_i = mean(\hat{L}_i)$
  $\sigma_i = std(\hat{L}_i)$
  for $j = 1$ **to** n **do**
    if $abs(L_j[i] - \mu_i) \geq \delta\sigma_i$ **then**
      $L_j[i] \leftarrow \mu_i$
    end if
  end for
  $G_t[i] = mean(\hat{L}_i)$
end for
Return G_t

实验设定——RFOut-1d

- 实验数据
  - Digits FEMNIST：联邦学习版本的手写字符EMNIST数据集，每个客户端对应一种写手
  - CelebA：名人人脸属性数据集，每个客户端对应同一个人
- input-instance后门攻击
  - 不修改样本特征，仅改标签
  - 后门任务数与为$D_{backdoor}$选择样本的客户端数量相对应
  - 攻击频率代表以1为间隔持续攻击

实验数据集描述

	FEMNIST	CelebA-S	CelebA-A
Clients	3579	1878	1878
k	8	30	30
Number of labels	10	2	2
Training samples	240000	56364	56364
Samples per client (mean)	67.05	30.01	30.01
Samples per client (std)	11.17	0.19	0.19
Testing samples	40000	19962	19962

input-instance后门攻击参数设定

	FEMNIST	CelebA-S	CelebA-A
Backdoor tasks	30	30	10
$\lvert D_{backdoor} \rvert$	213	247	228
Adversarial clients	11	20	15
Frequency of attack	1	1	1
Origin label	7	No	No
Target label	1	Yes	Yes

input-instance无通用触发模式，只与特定样本关联！

实验结果——RFOut-1d

- input-instance实验分析
  - 可拓展性强，† 表示防御与梯度裁剪、差分隐私方法结合
  - RFOut-1d在最小化后门任务精度和最大化主任务精度的双目标方面优于所有基线
  - 合适的梯度裁剪、差分隐私参数有利于提升防御方法的性能，并不损害主任务精度

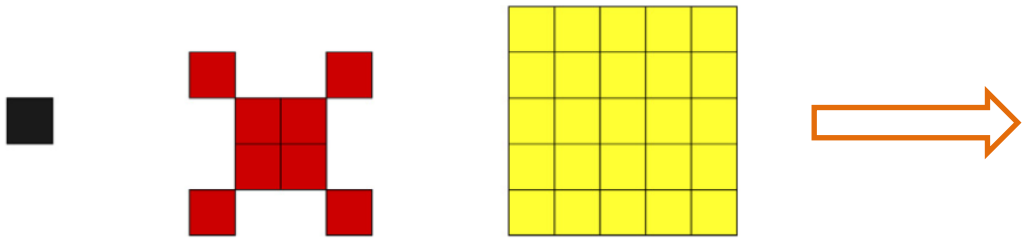
input-instance后门防御实验结果

	M	σ	FEMNIST		CelebA-S		CelebA-A	
			Original	Backdoor	Original	Backdoor	Original	Backdoor
No attack	0	0	0.9657	–	**0.7900**	–	0.7973	–
FedAvg	0	0	0.8661	0.8230	0.3630	0.9738	0.5140	0.5194
Median	0	0	0.9448	0.0306	0.7881	0.0457	0.7961	0.0152
Trimmed-mean	0	0	0.9526	0.0256	0.7852	0.0423	0.7961	0.0221
NormClip	3	0	0.9606	0.6373	0.6852	0.1431	0.6078	0.2558
WDP	3	0.0025	0.9374	0.1578	0.7204	0.1195	0.6119	0.2399
RLR	0	0	0.8404	0.0288	0.6539	0.0457	0.7877	0.0451
RLR†	0.5/0.5/1	0.0001	0.9546	0.0128	0.7852	0.0388	0.7934	0.0043
RFOut-1d	0	0	0.9629	**0.0048**	0.7883	0.0046	0.7973	**0.0**
RFOut-1d†	0.5/0.5/1	0.0001	**0.9670**	0.0054	0.7892	**0.0**	**0.7975**	**0.0**

有效地消除各客户端参数更新中的异常值以避免后门攻击！

实验设定——RFOut-1d

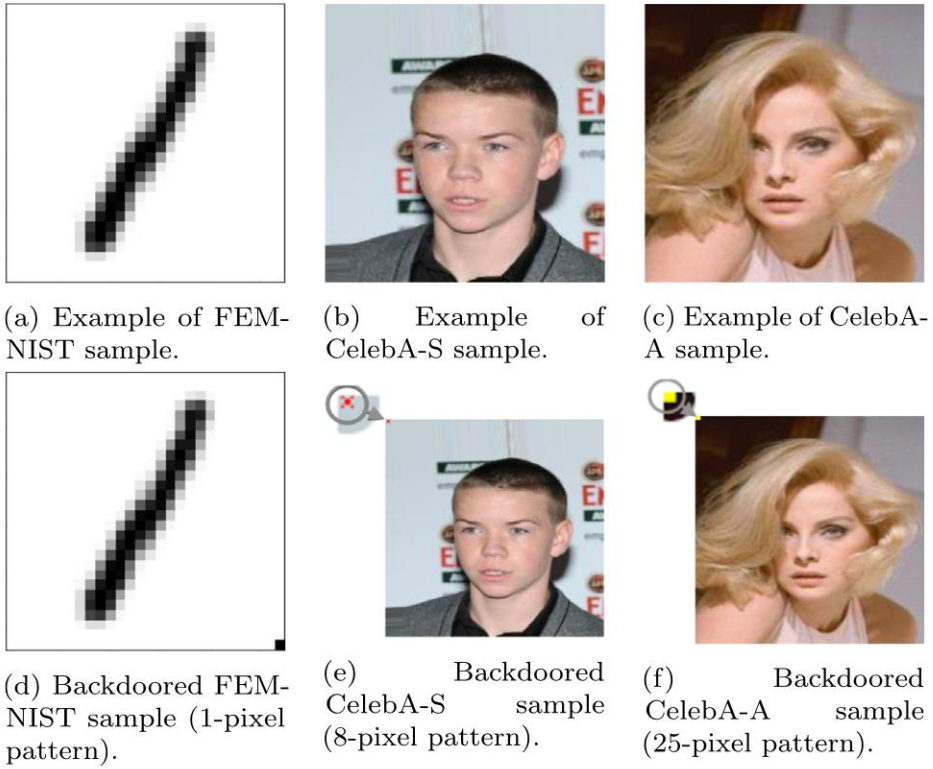
- pattern-key后门攻击
  - 需根据数据样本的特征设置通用触发模式，并修改标签。对于图像数据，触发模式数目与样本总体像素点成正比
  - 无后门任务数，后门样本与触发模式组成$D_{backdoor}$



不同的触发模式

不同数据集对应触发模式的设定

	FEMNIST	CelebA-S	CelebA-A
Adversarial clients	30	15	15
Frequency of attack	1	1	1
Target label	0	Yes	Yes
Pixels of the pattern	1	8	25



(a) Example of FEMNIST sample.

(b) Example of CelebA-S sample.

(c) Example of CelebA-A sample.

(d) Backdoored FEMNIST sample (1-pixel pattern).

(e) Backdoored CelebA-S sample (8-pixel pattern).

(f) Backdoored CelebA-A sample (25-pixel pattern).

后门样本示例

通用触发模式+任意数据样本即可激活后门！

实验结果——RFOut-1d

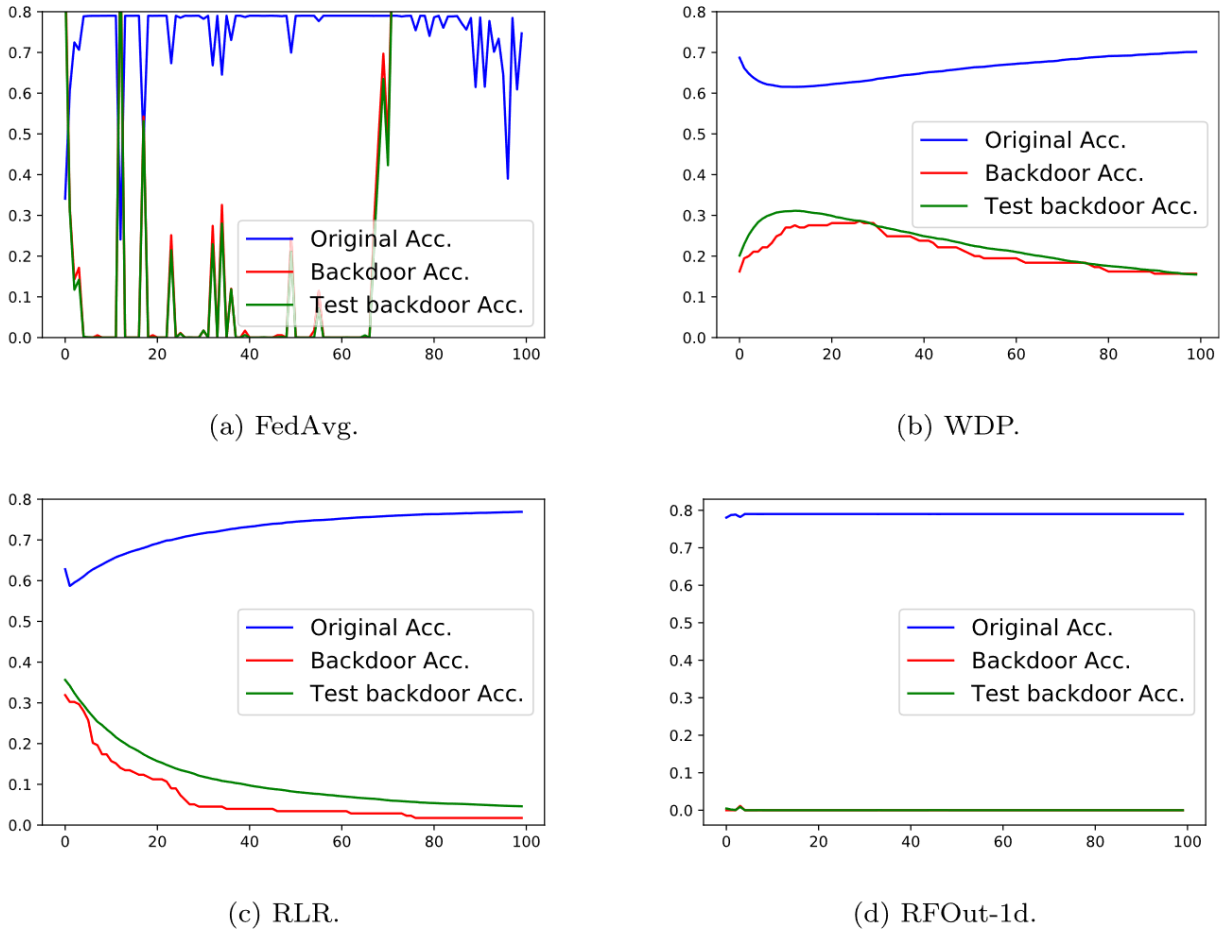
- **pattern-key实验分析**
  - **相比基线方法， RFOut-1d在三种数据集下都能取得最优的防御性能**
  - **Test测试全局后门情况，测试集中非目标标签样本+触发器，以评价攻击的泛化性**
  - **由FedAvg算法可知，pattern-key后门攻击效果优于input-instance，因其攻击模式更加复杂**

pattern-key后门防御实验结果

	M	σ	FEMNIST			CelebA-A			CelebA-S		
			Original	Backdoor	Test	Original	Backdoor	Test	Original	Backdoor	Test
No attack	0	0	0.9657	–	–	0.7973	–	–	0.7900	–	–
FedAvg	0	0	0.9741	1.0	1.0	0.7375	1.0	0.99	0.6858	1.0	0.9999
Median	0	0	0.9540	0.0091	0.0154	0.7452	0.0163	0.0189	0.6978	0.0678	0.0532
Trimmed-mean	0	0	0.9664	0.0114	0.0148	0.7498	0.0092	0.0101	0.7013	0.0521	0.0654
NormClip	1	0	0.9687	0.0553	0.0538	0.7126	0.1433	0.1316	0.6798	0.1433	0.1647
WDP	1	0.0025	0.9357	0.0938	0.0175	0.6609	0.1440	0.1707	0.7413	0.0538	0.0743
RLR	0	0	0.9039	0.0407	0.0575	0.6657	0.0280	0.0286	0.7132	0.0574	0.0469
RLR†	0.5/1	0.0001	0.9265	0.0089	0.0085	0.7923	0.0031	0.0016	0.7714	0.0205	0.0316
RFOut-1d	0	0	0.9741	**0.0043**	0.0072	**0.7967**	**0.0023**	**0.0015**	**0.7900**	**0.0**	**0.0**
RFOut-1d†	0.5/1	0.0001	**0.9753**	0.0059	**0.0051**	0.7874	0.0054	0.0124	0.7896	**0.0**	0.0010

实验分析——RFOut-1d

- 收敛性分析
  - CelebA-S数据集，实施pattern-key后门攻击，并与FedAvg、WDP及最佳参数的RLR对比
  - RFOut-1d使用更少的轮次收敛到全局最优解
  - 主任务精度，RFOut-1d未受到攻击影响，而其余方法稳定性较差
  - RFOut-1d始终能抵御后门攻击



(a) FedAvg. (b) WDP. (c) RLR. (d) RFOut-1d.

防御方法收敛性分析

消除异常参数，防止更新方向偏离！

优劣分析——RFOut-1d

- 算法优势
  - 单变量异常值检测方法，防御思路简单且具有理论依据（中心极限定理）
  - 未对恶意客户端的攻击策略和场景进行限制，符合真实场景
  - 可拓展性强，支持与梯度裁剪和差分隐私方法组合，提升防御效果
- 算法不足
  - 模型参数展平，并在每一维度上过滤异常值，导致防御方法在大模型上聚合异常耗时
  - 以鲁棒性方法筛选异常值，不支持安全聚合
  - 未对加噪和梯度裁剪的范围做约束，导致正常任务精度受影响

算法原理——FLAME

【NDSS】
FLAME: Taming Backdoors in Federated Learning

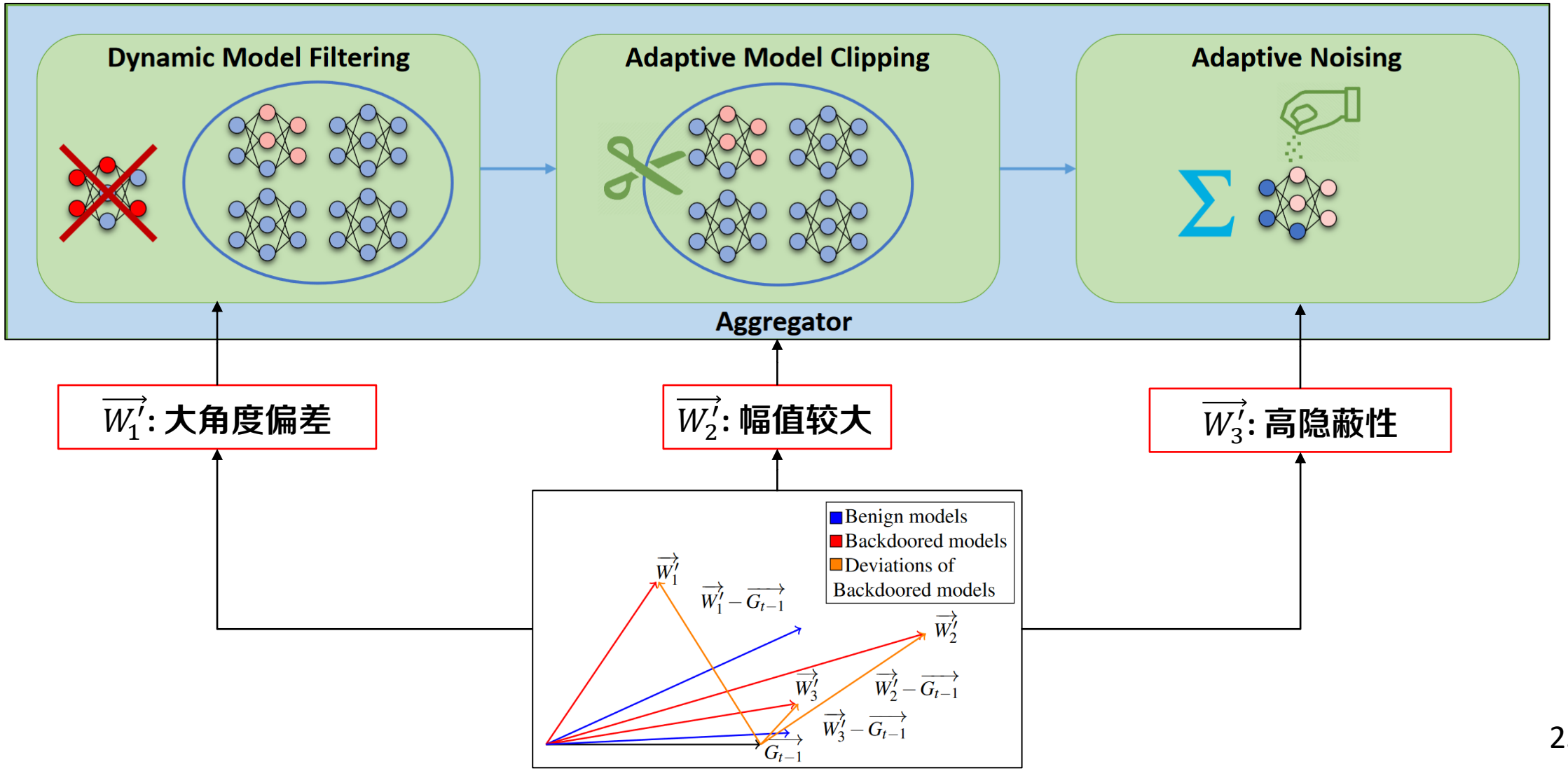
算法原理——FLAME

T	抵御联邦学习的后门攻击
I	良性与恶意客户端样本集，三种实验场景
P	1. 利用HDBSCAN算法动态聚类各客户端的参数更新，确定其中的良性客户端更新 2. 对选定良性客户端参数进行自适应裁剪和加噪
O	免于后门攻击的全局模型

P	现有基于检测和过滤的防御方法对应的只考虑特定的攻击者模型（恶意客户端的攻击策略和潜在的数据分布） 基于差分隐私的方法影响正常任务的精度
C	以恶意客户端在每轮次参与训练客户端占比作为先验知识
D	正确地聚类良性与恶意客户端的分布 确定梯度裁剪和加噪的比例
L	NDSS 2022

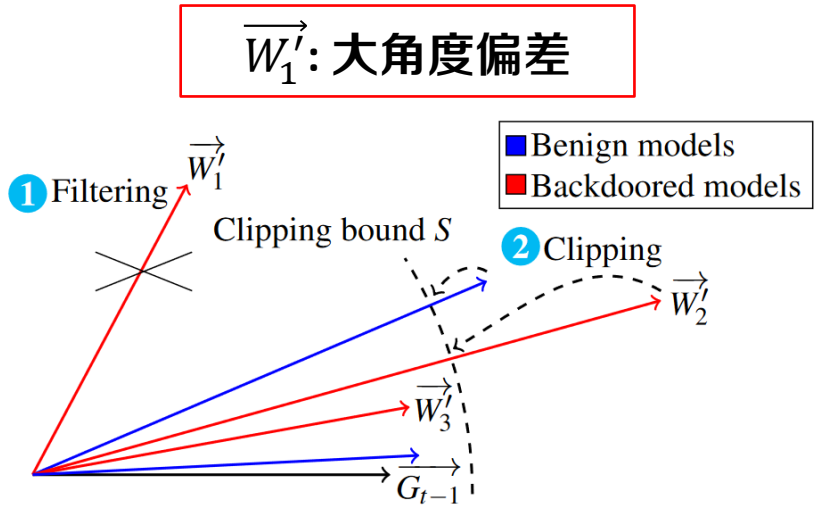
算法原理——FLAME

- 算法思想

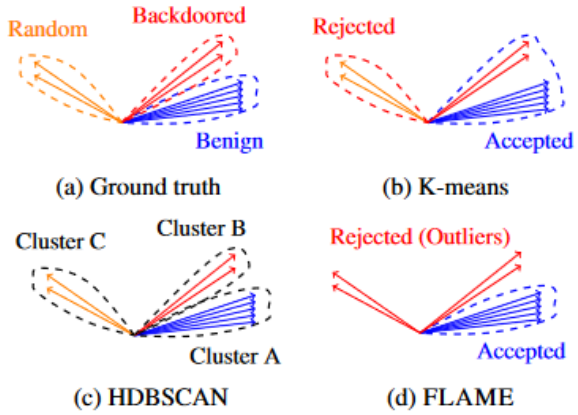


算法原理——FLAME

- 动态模型过滤
  - 基于HDBSCAN动态聚类方法，识别并移除优化方向偏离的参数更新
  - 通过成对余弦距离计算各参数更新的角度偏差，避免幅值影响聚类结果



(a) Applying clustering and clipping.

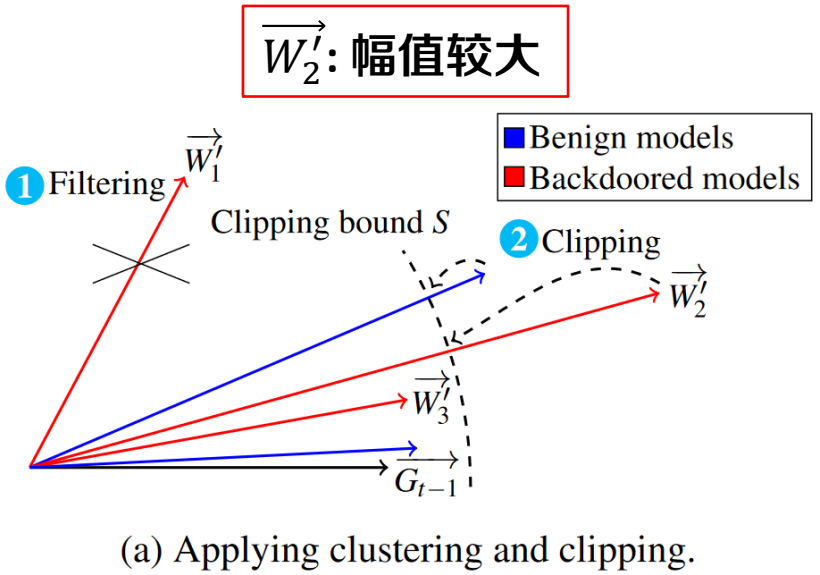


不同聚类质量对比

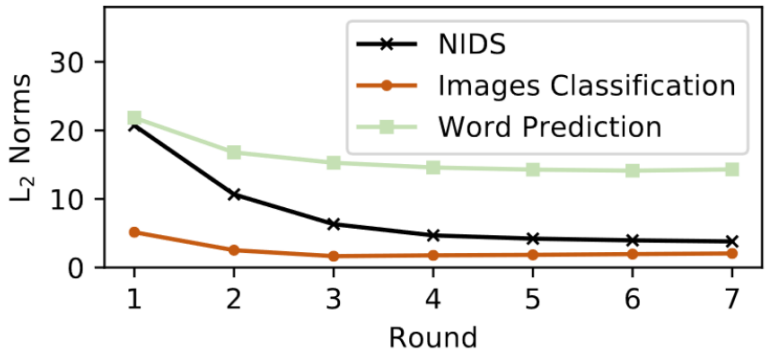
$$W_i \leftarrow \text{Client\_Update}\ (G_{t-1})$$
$$(c_{11}, \ldots, c_{nn}) \leftarrow \text{Cosine\_Distance}\ (W_1, \ldots, W_n)$$
$$(b_1, \ldots, b_L) \leftarrow \text{Clustering}\ (c_{11}, \ldots, c_{nn})$$

算法原理——FLAME

- 自适应裁剪
  - 特性：随迭代次数增加，良性模型更新与全局模型的L_2范数减小
  - 裁剪阈值：每轮次n个模型对应L_2范数的中位数



(a) Applying clustering and clipping.

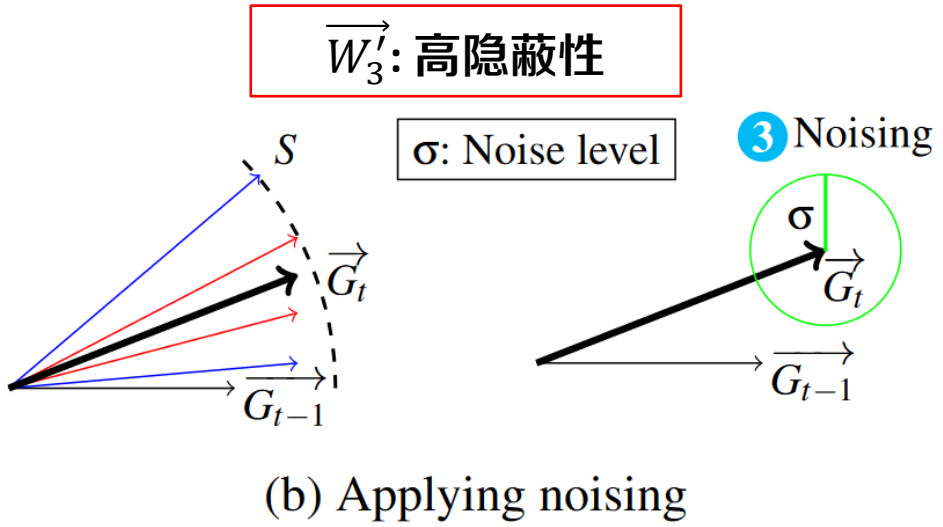


模型更新的L_2范数

$$(e_1, \dots, e_n) \leftarrow \text{Euclidean\_Distance}\left(G_{t-1}, (W_1, \dots, W_n)\right)$$
$$S_t \leftarrow \text{Median}(e_1, \dots, e_n)$$
$$w_j \leftarrow G_{t-1} + \left(W_j - G_{t-1}\right) * \text{Min}\left(1, \frac{S_t}{e_j}\right) \forall \in \{b_1, \dots, b_L\}$$

算法原理——FLAME

- 自适应加噪
  - 通过理论证明(ε, σ)差分隐私抵御联邦学习下的后门攻击
  - 基于本地模型的距离差异动态估计噪声量，噪声水平σ与裁剪阈值S_t成比例
  - 通过裁剪和过滤操作，降低所需噪声，最小化对模型正常性能的影响



(b) Applying noising

$$G_t \leftarrow \sum_{j\in\{b_1,\ldots,b_L\}} \frac{W_j}{L}$$

$$G_t \leftarrow G_t + N(0, \sigma_t^2) \text{ where } \sigma_t^2 \leftarrow \frac{S_t \cdot \sqrt{2\ln\left(\frac{1.25}{\delta}\right)}}{\varepsilon}$$

燥起来

实验数据——FLAME

- 实验数据
  - 三种实验场景：单词预测（WP）、图片分类（IC）、IoT入侵检测（NIDS）
  - 5种对比攻击方法，6种对比防御方法
- 评测指标
  - 后门任务精度 BA
    - 模型预测后门样本为指定目标类别的准确率
  - 主任务精度 MTA
    - 模型预测良性样本的准确率
  - 真阳性率 TPR
    - 正确归类为投毒模型数量的比率，表示防御方法识别投毒模型的能力
  - 真阴性率 TNR
    - 正确归类为良性模型数量的比率

实验数据源与模型

Application	Datasets	#Records	Model	#params
WP	Reddit	20.6M	LSTM	~20M
NIDS	IoT-Traffic	65.6M	GRU	~507k
IC	CIFAR-10	60k	ResNet-18 Light	~2.7M
	MNIST	70k	CNN	~431k
	Tiny-ImageNet	120k	ResNet-18	~11M

多场景多指标多对比方法，实验充分！

实验结果——FLAME

- FLAME有效性实验
  - 三种数据集上FLAME完全抵御Constrain-and-scale攻击（BA=0%）
  - 防御方法MA影响较小，最大下降约为5%

不同攻击方法下FLAME有效性

Attack	Dataset	No Defense		FLAME	
		BA	*MA*	*BA*	*MA*
Constrain-and-scale [7]	Reddit	100	22.6	0	22.3
	CIFAR-10	81.9	89.8	0	91.9
	IoT-Traffic	100.0	100.0	0	99.8
DBA [61]	CIFAR-10	93.8	57.4	3.2	76.2
Edge-Case [59]	CIFAR-10	42.8	84.3	4.0	79.3
PGD [59]	CIFAR-10	56.1	68.8	0.5	65.1
Untargeted Poisoning [20]	CIFAR-10	-	46.72	-	91.31

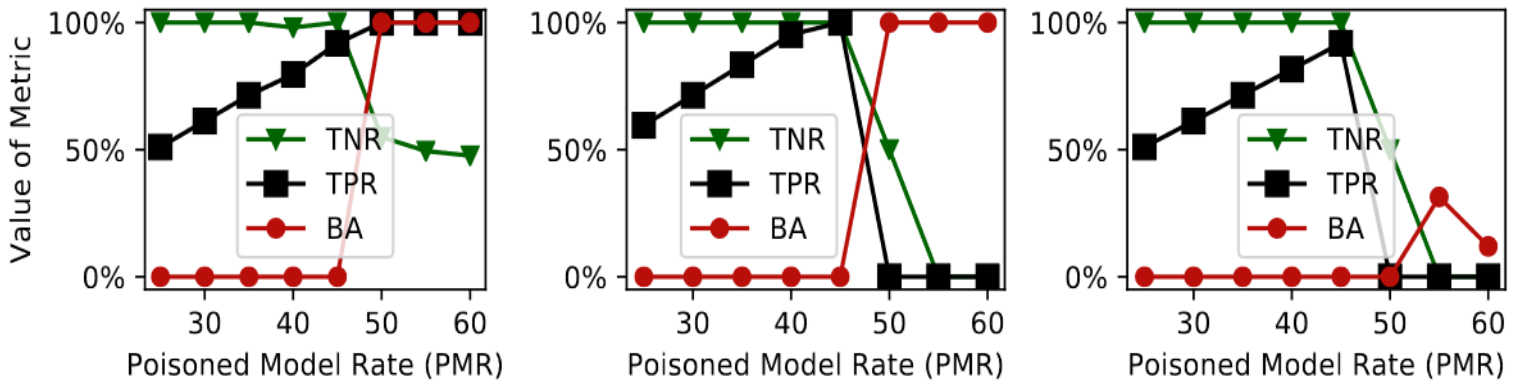
- 对比现有防御方法
  - FLAME在3种数据集上效果更优，保持MA的前提下，降低BA
  - Krum、FoolsGold、Auror和AFA方法抵御后门攻击效果差
  - DP和Median方法对MA影响较大

不同防御方法与FLAME对比

Defenses	Reddit		CIFAR-10		IoT-Traffic	
	BA	*MA*	*BA*	*MA*	*BA*	*MA*
Benign Setting	-	22.7	-	92.2	-	100.0
No defense	100.0	22.6	81.9	89.8	100.0	100.0
Krum [9]	100.0	9.6	100.0	56.7	100.0	84.0
FoolsGold [23]	**0.0**	**22.5**	100.0	52.3	100.0	99.2
Auror [53]	100.0	**22.5**	100.0	26.1	100.0	96.6
AFA [43]	100.0	22.4	**0.0**	91.7	100.0	87.4
DP [18]	14.0	18.9	**0.0**	78.9	14.8	82.3
Median [64]	**0.0**	22.0	**0.0**	50.1	0.0	87.7
FLAME	**0.0**	22.3	**0.0**	**91.9**	**0.0**	**99.8**

实验结果——FLAME

- 恶意客户端数目的影响
  - 条件：投毒模型比率PMR小于50%，否则影响聚类规则判定
  - IC、NIDS、WP三种场景下BA、TPR和TNR随PMR（25%-60%）的变化情况
  - 当PMR小于50%时，聚类规则正确，TNR = 100%，BA = 0%
  - PMR > 50%，TNR下降，良性模型分类错误

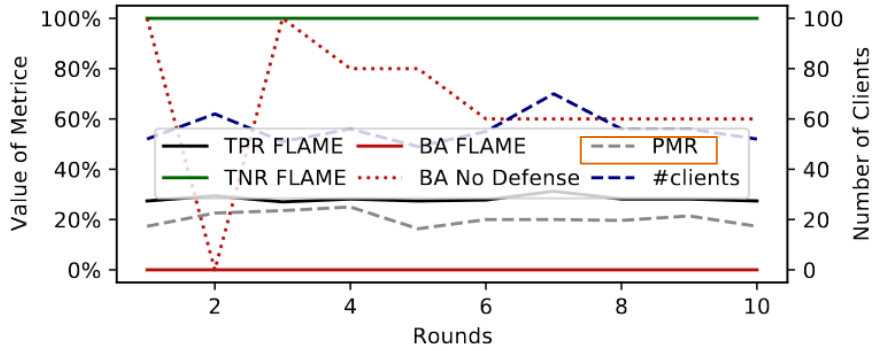


(a) IC (b) NIDS (c) WP

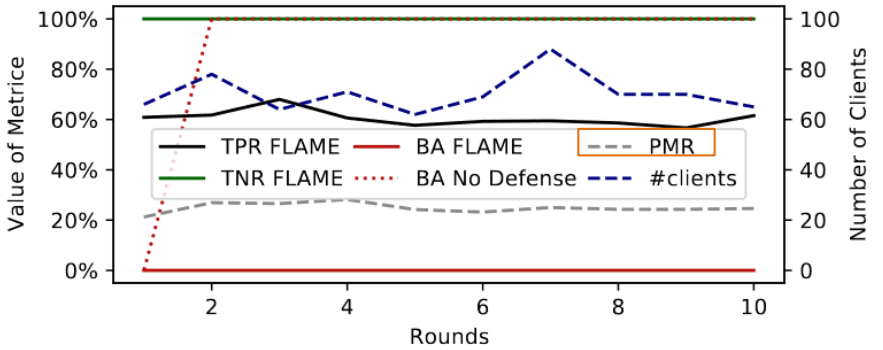
投毒模型比率PMR的影响

实验结果——FLAME

- 不同训练轮次客户端数量变化
  - 实验设置：100个候选客户端，存在25个恶意客户端；每轮次随机选取60~90客户端
  - FLAME不考虑训练前后轮次关联，如不记录历史可疑或良性客户端信息
  - PMR的变化不影响FLAME方法的有效性，后门任务BA始终为0%



(a) Image Classification



(b) Network Intrusion Detection

参与训练客户端数量对效果的影响

FLAME独立于训练前后轮次，不引入额外知识，鲁棒性强！

优劣分析——FLAME

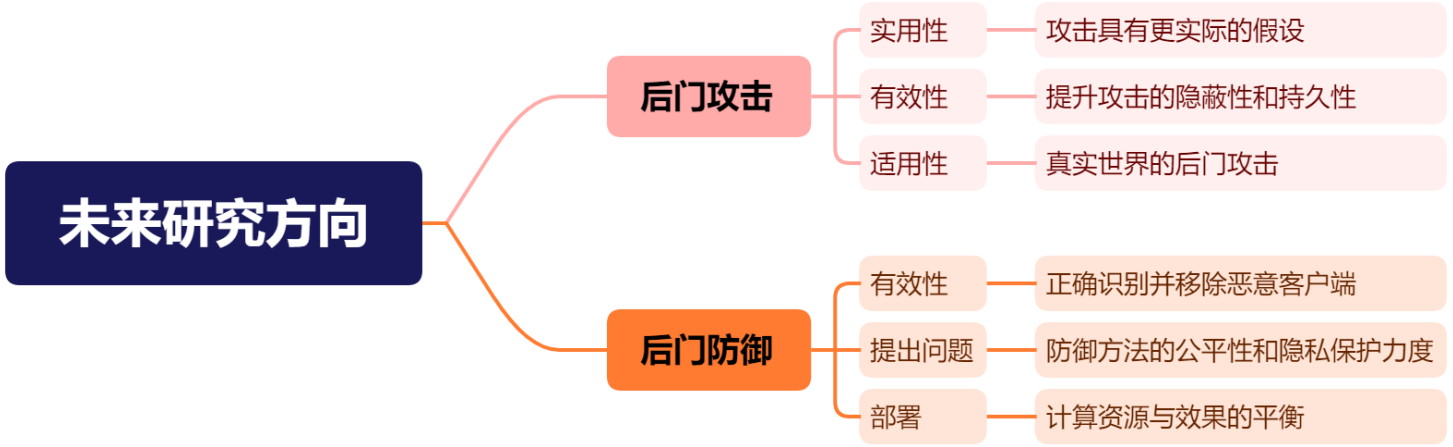
- 算法优势
  - 对参与方数据分布不做假设，且不限制恶意参与方的攻击策略
  - 自适应权重裁剪和加噪，降低对主任务精度的影响
- 算法不足
  - 投毒模型比率PMR影响良性客户端聚类规则，算法假设恶意客户端数量至多为50%
  - 服务器检测各参与方参数更新，不支持安全聚合

牛掰

总结

总结

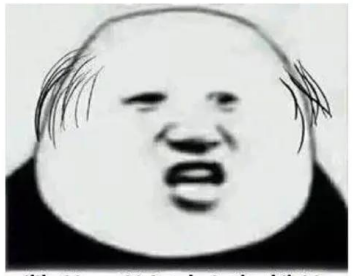
- 发展前景
  - 后门攻击
    - 假设更实际的后门攻击：FL中大多数后门攻击都依赖于不同的假设，包括关于恶意客户端的百分比、FL客户端的总数和训练数据分布的假设
    - 提升后门攻击的额外效果：关注隐蔽性和持久性
  - 后门防御
    - 联邦学习的差分隐私
    - 防御方法的公平性：鲁棒性聚合方法会影响数据或参数分布相差大的正常客户端聚合



总结

- 创新思路
  - 根据自身课题任务的特殊性，在深度学习领域引入适合的传统算法
    - 传统算法本身可解释性强，且相较于同等深度学习方法效率高
    - HDBSCAN动态聚类，不需要指定聚类数，方法的鲁棒性强
  - 创新点的理论依据不怕简单，但要充分可靠（创新度足够）
    - 联邦学习下客户端的参数更新符合中心极限定理
  - 论文中公式支撑尽量多
    - 密码学和安全相关顶刊顶会，如NDSS，基本都有大篇幅公式推导，证明方法的合理性
    - 隐私保护相关，差分隐私推导会是不错的切入点，对密码学相关的知识要求低

α阿花，β憋它，γ甲马，
δ呆它，ε爱破戏楼。



数学，学起来很轻松的
只是头冷

模型安全领域许多想法是相通的！

参考文献

[1] RODRÍGUEZ-BARROSO N, MARTÍNEZ-CÁMARA E, LUZÓN M V, et al. Backdoor attacks-resilient aggregation based on Robust Filtering of Outliers in federated learning for image classification[J]. Knowledge-Based Systems, 2022(245), 245: 108588.

[2] Nguyen T D, Rieger P, De Viti R, et al. {FLAME}: Taming backdoors in federated learning[C]//31st USENIX Security Symposium (USENIX Security 22). 2022: 1415-1432.

[3] Nguyen T D, Nguyen T, Nguyen P L, et al. Backdoor Attacks and Defenses in Federated Learning: Survey, Challenges and Future Research Directions[J]. arXiv preprint arXiv:2303.02213, 2023.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！