

面向智能物联网的双层级联邦安全学习架构

郑诚波¹, 闫皓楠¹, 傅彩利², 张栋³, 李晖¹, 王滨^{1,4}

(1. 西安电子科技大学网络与信息安全学院, 陕西 西安 710126;

2. 杭州海康威视数字技术股份有限公司, 浙江 杭州 310051;

3. 国家计算机网络应急技术处理协调中心, 北京 100029;

4. 浙江省全省智能物联网与数据安全重点实验室, 浙江 杭州 310051)

摘要: 联邦学习作为一种分布式机器学习架构, 可在保护数据隐私的条件下完成模型共同训练, 被广泛应用于智能物联网中。然而, 联邦学习中往往存在隐私泄露、投毒攻击等安全威胁。为了克服智能物联网场景下, 多机构间使用联邦学习进行联合训练时存在的性能和安全的挑战, 提出了面向智能物联网的双层级联邦安全学习架构。该架构将整个智能物联网安全学习系统分为底层和顶层的双层级架构, 底层架构由机构内各个物联网设备和一台服务器组成, 不同物联网设备之间通过区块链网络连接, 服务器通过设备上传的历史梯度进行投毒攻击检测并剔除恶意设备, 避免投毒攻击造成的收敛缓慢、全局模型精度下降等问题; 顶层架构由不同机构的服务器组成, 采用基于秘密共享的安全多方计算进行安全聚合, 保护梯度隐私的同时实现了去中心化聚合。实验结果表明, 该架构对4种常见投毒攻击的检测准确率均达到85%以上, 极大地提高了系统的安全性, 并在线性时间复杂度内实现梯度隐私保护的的去中心化安全聚合。

关键词: 智能物联网; 联邦学习; 投毒攻击; 安全多方计算; 区块链; 隐私保护

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024081

Double layer federated security learning architecture for artificial intelligence of things

ZHENG Chengbo¹, YAN Haonan¹, FU Caili², ZHANG Dong³, LI Hui¹, WANG Bin^{1,4}

1. School of Cyber Engineering, Xidian University, Xi'an 710126, China

2. Hangzhou Hikvision Digital Technology Co., Ltd, Hangzhou 310051, China

3. National Computer Network Emergency Response Technical Team/Coordination Center of China(CNCERT/CC), Beijing 100029, China

4. Zhejiang Key Laboratory of Artificial Intelligence of Things(AIoT)Network and Data Security, Hangzhou 310051, China

Abstract: Federated learning, as a distributed machine learning architecture, can complete model co-training while

收稿日期: 2024-04-28; 修回日期: 2024-09-30

通信作者: 张栋, zhangdong@cert.org.cn

基金项目: 国家自然科学基金(92167203, 61932015, 62402373), 中国博士后科学基金(374102), 浙江省博士后科研择优项目(ZJ2024009)

Foundation Items: The National Natural Science Foundation of China(92167203, 61932015, 62402373), The China Postdoctoral Science Foundation(374102), The Zhejiang Province Postdoctoral Research Priority Funding Program(ZJ2024009)

引用格式: 郑诚波, 闫皓楠, 傅彩利, 等. 面向智能物联网的双层级联邦安全学习架构[J]. 网络与信息安全学报, 2024, 10(6): 71-80.

Citation Format: ZHENG C B, YAN H N, FU C L, et al. Double layer federated security learning architecture for artificial intelligence of things[J]. Chinese Journal of Network and Information Security, 2024, 10(6): 71-80.

protecting data privacy, and is widely used in Artificial Intelligence of Things. However, there are often security threats such as privacy breaches and poisoning attacks in federated learning. In order to overcome the performance and security challenges of using federated learning for joint training among multiple institutions in the context of intelligent Internet of Things, a two-level federated security learning architecture was proposed for intelligent Internet of Things. The entire security learning system was divided into a two-level architecture of bottom and top layers. The bottom architecture consisted of various IoT devices and a server within the organization. Different devices were connected through blockchain networks, and the server detected and eliminated malicious devices through the historical gradient uploaded by the devices, avoiding slow convergence and decreased global model accuracy caused by poisoning attacks. The top-level architecture consisted of servers from different institutions, using secure multi-party computation based on secret sharing for secure aggregation, protecting gradient privacy while achieving decentralized gradient aggregation. The experimental results show that the architecture achieves detection accuracy of over 85% for four common poisoning attacks, greatly improving the security of the system and achieving decentralized security aggregation with gradient privacy protection within linear time complexity.

Keywords: artificial intelligence of things, federated learning, poisoning attack, secure multi-party computation, blockchain, privacy protection

0 引言

智能物联网 (artificial intelligence of thing, AIoT) 正日益融入人们生活的方方面面, 从智能家居的便捷体验, 到智慧医疗的精准服务, 再到智慧城市的全面升级, AIoT 的应用领域正在以惊人的速度增长。为了能够充分利用 AIoT 中物联网设备生成的数据, 机器学习 (machine learning, ML) 技术被广泛应用在 AIoT 中。传统的 ML 需要中央服务器存储数据, 然而随着物联网设备大量增加, 数据量的快速增长给 ML 带来严重的性能限制。同时, 《通用数据保护条例》(general data protection regulation, GDPR) 中认为传输原始数据易造成隐私泄露。因此, 联邦学习 (federated learning, FL) 被提出构建 AIoT 系统。与传统的 ML 不同, FL 作为一种分布式的机器学习框架, 核心思想是在不共享原始数据的情况下, 通过多方协作共同训练一个全局模型, 其分布式的特点与 AIoT 十分契合, 二者结合则可以在不侵犯数据隐私的情况下, 充分挖掘终端设备上数据的巨大潜力。

目前, 一些工作已经将 FL 应用到 AIoT 实际场景中。例如, 在智慧医疗领域, Liu 等^[1]使用联邦学习技术对 COVID-19 的胸部 X 光图像进行分类模型的训练; Sheller 等^[2]采用联邦学习技术对脑部扫描图像进行语义分割模型训练, 最终得到的模型具有与共享数据下训练的模型相近的性能。在个人健康设备方面, Chen 等^[3]提出了准确

且个性化的医疗保健模型; FedPer^[4]和 pFedMe^[5]也可用于物联网设备进行协作训练, 同时保证模型本身的个性化特点。在车载网络领域, Saputra 等^[6]采用基于物联网设备集群的联合能源需求学习方法, 对区域能源需求进行预测。

然而, FL 本身也带来了一些新的挑战, 限制了其在现实场景中的应用: ①FL 容易受到投毒攻击的威胁^[7], 恶意客户端可以通过操纵、注入或篡改训练数据的方式污染本地模型, 进而降低全局模型的准确率和收敛速度。②虽然 FL 避免传输原始数据, 但是梯度数据依旧会造成隐私泄露。现有的梯度泄露 (deep leakage gradient, DLG) 攻击^[8]仍然可以从梯度参数恢复出原始训练数据。③传统的 FL 需要一个中央服务器来执行聚合等操作, 服务器性能成为限制 FL 训练的瓶颈, 一旦中央服务器出现问题, 整个 FL 学习系统就会随之崩溃。

当前已有一些研究致力于防范 FL 中潜在的挑战。针对投毒攻击的安全威胁, Blanchard 等^[9]提出了一种拜占庭容错的聚合算法名为 Krum, 通过最接近的 L2 范数测量每个上传梯度的分数并选择最高得分的梯度作为聚合结果; 文献^[10-13]提出了几何中位数、边缘中位数、坐标中位数等基于中位数的投毒攻击判别方法; Xia 等^[14]提出了针对分布式网络中拜占庭攻击的快速聚合算法, 并命名为 FABA, 他们采用欧几里得距离代替中位数进行判断, 以此删除异常梯度;

Prakash等^[15]提出了基于方向和长度相似度的判别方法；文献[16]提出对训练的数据进行清洗，这是针对数据投毒的一种常用方法；文献[17]提出根据中值和余弦相似性，采用隐马尔可夫模型排除恶意的局部更新；Cao等^[18]使用未被投毒的小数据集计算加噪梯度，并将每个本地更新与加噪梯度进行比较来过滤恶意更新；Yan等^[19]提出使用精心构建的聚合梯度查询每个客户端，根据恶意客户端的响应进行检测，并使用信任分数来更加稳健聚合梯度。然而上述投毒攻击防御方法并未能对投毒梯度进行追踪溯源，无法得知上传投毒梯度的具体设备。针对隐私泄露问题，一些隐私保护算法，例如差分隐私（differential privacy, DP）^[20-21]，同态加密^[22-23]，安全多方计算（secure multi-party computation, SMPC）^[24-25]等也在联邦学习中被广泛应用，然而此类密码算法往往对客户端的设备算力有一定的要求。

针对上述存在的问题，本文提出了一种面向智能物联网的双层级联邦安全学习架构。该架构将整个智能物联网安全学习系统分为底层和顶层的双层级架构，其中双层级架构能够缓解中央服务器性能瓶颈带来的问题。底层架构由每个机构内各个物联网设备和服务器组成，设备之间采用区块链网络进行连接，区块链技术能够对每个物联网设备上传的参数进行存储，保证了设备上传的梯度数据的可溯源性与不易篡改性，服务器通过设备上传的历史梯度进行投毒检测溯源并剔除恶意设备；顶层架构由不同机构的服务器组成，采用基于秘密共享的SMPC进行安全聚合，保护梯度隐私的同时实现了去中心化梯度安全聚合。本文的主要贡献归纳如下。

1) 提出了一种新型的双层级联邦安全学习架构，缓解了联邦学习中央服务器性能瓶颈带来的限制。

2) 提出了在区块链网络中基于历史梯度的投毒攻击防御方法，实现对潜在投毒梯度的检测与溯源。

3) 提出了基于秘密共享的SMPC安全聚合方案，避免聚合过程中梯度隐私泄露安全威胁的同时实现去中心化的模型聚合。

4) 为了验证架构方案的有效性，本文分别在

MNIST、CIFAR10和FEMNIST数据集上进行实验。实验结果表明，该架构对4种常见投毒攻击的防御成功率均达到85%以上，并在线性时间复杂度内实现梯度隐私保护的分布式聚合。

1 预备知识

1.1 联邦学习

由于传统的ML存在隐私泄露、数据孤岛等问题，谷歌在2016年提出了一种新的分布式机器学习框架——联邦学习，其核心思想在于多个客户端协作，联合训练一个泛用性较高的全局模型，实现数据共享和计算的同时避免直接交换客户端本地数据，确保了数据的隐私。常规的FL场景中存在一个中央服务器和 N 个客户端，第 i 个客户端用 c_i 表示， D_i 表示每个客户端 c_i 拥有的数据集， D 代表全部数据。FL的目标就是训练一个适配 D 的全局模型：

$$m^* = \arg \min \sum_{i=1}^n p_i F_i(m, D_i) \quad (1)$$

其中 m 代表客户端的本地模型， m^* 代表最终得到的全局模型， $p_i = |D_i|/|D|$ 表示本地数据集占全部数据的比例，满足 $\sum_{i=1}^n p_i = 1$ ， $F(\cdot)$ 表示本地模型训练的损失函数。

1.2 区块链

区块链作为一种去中心化、不易篡改的数字账本技术，能够通过共识算法在无信任的环境下构建服务，在物联网、边缘计算等领域得到了广泛的应用。在区块链中，所有参与节点通过共识机制维护一个全节点一致的账本，账本中每个区块均有一系列交易和前一个区块的哈希散列值构成，这种链式存储让数据不易被篡改的同时，确保了进行交易等任务时的可审计性，同时其去中心化特点完美契合了物联网和FL的思想，为构建更加安全的联邦智能物联网架构提供了新的思路。

1.3 安全多方计算

理想场景中往往存在一个可信第三方（trusted third party, TTP）与多个数据提供方 P_i （ $i \in [1, n]$ ）， P_i 将数据发送给TTP进行计算，TTP完成后将结果返回。然而在现实场景中，这样的可信第三方基本是不存在的，因此SMPC的目的是使各方能够以安全的方式执行分布式计算

任务, 定义如下:

n 个参与方 P_i ($i \in [1, n]$) 希望利用各自的秘密输入共同计算 n 元 $f(\cdot)$ 函数, $f(\cdot)$ 函数的功能由参与方共同决定, 表达式如下:

$$f(x_1, x_2, \dots, x_n) = (y_1, y_2, \dots, y_n) \quad (2)$$

其中 x_i 代表参与方 P_i 的秘密输入, y_i 代表最后的输出, 且每个参与方除了自己的输入输出以及可推导出的信息外, 无法获得任何额外的信息。

1.4 k -means 聚类算法

k -means 聚类算法^[26]是一种经典的无监督学习算法, 能够将不同的数据点分割成 k 个不同的簇。其核心思想是通过迭代的方式, 将数据点分配到最近的簇中心, 并更新簇中心以最小化数据点与簇中心的距离之和。给定数据集 $X = (x_1, x_2, \dots, x_n)$, k -means 聚类算法的目标是最小化以下损失函数。

$$F(C, \mu) = \sum_{i=1}^n \min_j \|x_i - \mu_j\|^2 \quad (3)$$

其中, $C = (C_1, C_2, \dots, C_k)$ 代表不同的分簇, $\mu = (\mu_1, \mu_2, \dots, \mu_k)$ 代表对应 C_i 的簇中心。

k -means 聚类算法通过迭代更新簇中心来最小化损失函数, 直到簇中心不再变化或变化可以忽略不计。

2 系统设计

2.1 架构设计思路

本文考虑了在训练过程中可能存在的外部和内部威胁。外部威胁主要考虑在聚合过程中上传梯度参数时, 来自外网的攻击者进行窃听攻击, 进而利用窃听到的梯度数据进行 DLG 等攻击推理出训练时使用的隐私数据; 内部威胁主要考虑部分设备被攻击者控制进行数据投毒攻击, 进而造成收敛缓慢、模型精度下降等情况。

根据上述可能存在的威胁, 本文系统设计旨在实现以下目标。

1) 确保数据与模型的准确性: 内部设备可能被攻击者控制进行投毒。因此为了应对内部设备被挟持的情况, 需要确保上传聚合参数的准确性, 防止模型准确率受到影响。

2) 保障上传数据的不可篡改性及可追溯性: 由于投毒攻击会上传恶意梯度, 服务器需要根据

设备历史上传的梯度判断是否是恶意设备, 同时需要在下一轮聚合时剔除被挟持的设备。因此服务器需要对设备上传的参数进行记录, 并在需要时可以追溯恶意梯度参数的来源。

3) 强化数据的隐私性: 针对外部攻击者的窃听行为, 需要对聚合的梯度参数进行隐私保护。

2.2 系统架构设计

如图 1 所示, 本系统采用双层级架构设计。假设有 m 个机构需协作训练一个全局模型, 每个机构拥有 n_i 个智能物联网设备和一个边缘服务器 S_i , 其中 $i \in [1, m]$ 。底层架构由服务器各自维护一个区块链网络组成, 不同区块链网络中每一个节点都是一个 AIoT 设备, 服务器和物联网设备间通过局域网进行通信; 而顶层架构由各个不同机构的服务器组成, 通过互联网连接形成一个用于模型聚合的去中心化网络。底层模型聚合之前, 服务器先对每个设备上传的梯度进行投毒检测, 剔除恶意梯度后再进行聚合; 顶层采用 SMPC 算法实现安全的模型聚合。下面描述架构中每个部分的详细流程。

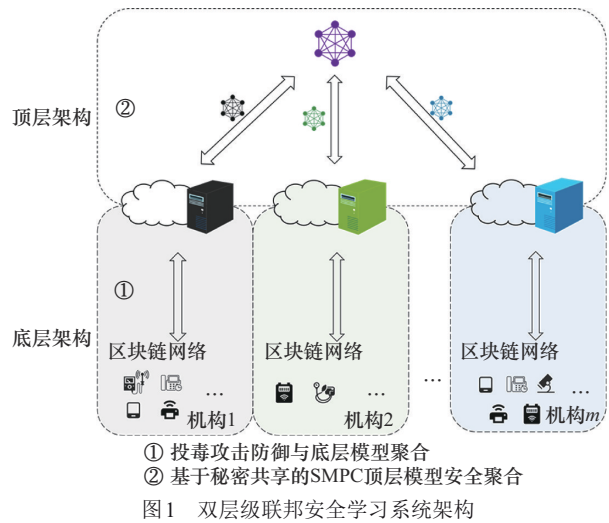


图 1 双层级联邦安全学习系统架构
Figure 1 Architecture of a two-level federated security learning system

(1) 设备本地训练

为了简化讨论, 本文在底层架构描述中只对一个机构内部架构进行阐述, 其余机构内部结构同理。不妨假设机构中有 n 个设备和一台服务器 S , 每个设备均具有数据收集和模型训练的能力。在训练启动之前, 各个机构采用相同的初始模

型，并将模型分发至各自的设备中。服务器 S 维护一个区块链网络，每个设备均可作为区块链节点。

随着训练周期的进行，每个设备自动采集数据并进行模型训练。当训练完成，它们将需要聚合的梯度参数上传至区块链网络，区块链技术能够确保每个设备上传梯度的可追溯性和不易篡改性，为服务器进行投毒检测提供了基础。

(2) 底层架构投毒攻击检测

如图2所示，服务器从区块链网络接收各个设备的梯度参数后，在聚合前需要对每个设备上传的梯度数据进行投毒攻击检测。以第 t 轮训练为例，对于良性设备 i ，根据柯西中值定理，第 t 轮上传的梯度数据 g_i^t 应该满足：

$$g_i^t = g_i^{t-1} + H_i^t(w_t - w_{t-1}) \quad (4)$$

其中， g_i^{t-1} 代表第 i 个设备在第 $t-1$ 轮训练上传的梯度， H_i^t 表示设备 i 在第 t 轮的海森矩阵，计算方式为：

$$H_i^t = \int_0^1 H_i(w_{t-1} + x(w_t - w_{t-1})) dx \quad (5)$$

其中， w_t 和 w_{t-1} 代表第 t 轮和 $t-1$ 轮中迭代的全局模型。根据 g_i^t 计算式，当设备 i 在 $t-1$ 轮上传梯度参数 g_i^{t-1} 后， S 会计算出第 t 轮的预测梯度 $\hat{g}_i^t$ 。

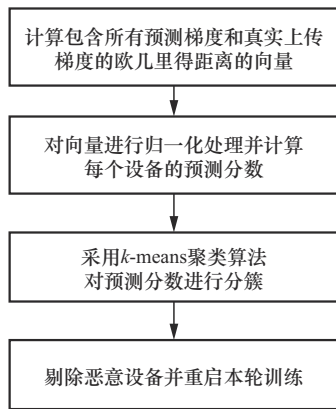


图2 投毒攻击检测流程

Figure 2 Poisoning attack detection process

由于用被投毒的设备来上传的梯度参数会与全局模型收敛方向有较大的差异，因此真实上传的梯度与估算的梯度之间往往存在较大的欧几里得距离差异。假设 n 个设备在第 t 轮预测梯度和真实上传梯度的欧几里得距离差异为向量 d^t ，表示为：

$$d^t = [\|\hat{g}_1^t - g_1^t\|_2, \|\hat{g}_2^t - g_2^t\|_2, \dots, \|\hat{g}_n^t - g_n^t\|_2] \quad (6)$$

其中， $\hat{g}_i^t$ 是第 i 个设备真实上传的梯度， g_i^t 是预测的梯度。随后服务器 S 将 d^t 进行向量归一化处理 $\hat{d}^t = d^t / \|d^t\|_1$ ，其中每个分量 d_i^t 为对应设备的归一化后欧几里得距离。计算完成后，服务器 S 设置可疑分数 s ，第 i 个设备在第 t 轮训练中的 s_i^t 计算式为：

$$s_i^t = \frac{1}{N} \sum_{j=0}^{N-1} \hat{d}_i^{t-j} \quad (7)$$

其中， $\hat{d}_i^t$ 为第 i 个设备对应的预测分数， N 为迭代轮数，为了避免所有设备均为被投毒而导致的预测分数普遍偏大，本文未采取固定分数阈值，而是采用最近 N 轮迭代的平均分数 s_i^t 作为评估指标。

服务器 S 接收到所有设备的可疑评分 s_i^t 后，采用 k -means聚类算法对设备进行分簇。 S 首先使用基于差距统计(gap statistic)算法^[27]对设备的评分进行集群划分，若差距统计算法最终将设备划分为1个以上的集群，说明确实存在可疑设备。随后， S 根据 s_i^t 使用 k -means将客户端划分为2个聚类，最后将平均 s_i^t 偏大的设备聚类视作恶意设备并剔除后，重新进行训练轮次。

(3) 底层模型聚合

完成投毒攻击检测后，服务器 S 采用聚合算法对所有良性梯度参数进行聚合，这里以联邦平均(FedAvg)算法^[28]为例。假设投毒攻击检测后最终有 p 个设备参与聚合，每个设备第 t 轮上传的梯度参数为 g_i^t ，那么最终本地模型聚合后的梯度为：

$$g_{\text{local}}^t = \frac{1}{p} \sum_{i=1}^p g_i^t \quad (8)$$

其中， g_{local}^t 为该机构内部完成本地模型聚合后的模型梯度参数。

(4) 顶层模型聚合

完成底层模型聚合后，如图3所示，为了确保模型梯度数据的隐私，顶层架构中不同服务器采用基于秘密共享的SMPC协议进行安全的模型聚合。在该协议下，假设有 m 个机构，每个机构有各自对应的编号 i ， $i \in [1, m]$ ， S_i 代表每个机构的服务器， g_i^t 代表每个机构内部完成底层聚合后的模型梯度参数。以第 i 个机构为

例, 在顶层聚合之前, 服务器 S_i 模型梯度 g_i 会划分成 m 份, 分割后的梯度表示为 $\hat{g}_i = (g_{i1}, g_{i2}, \dots, g_{ii}, \dots, g_{im})$, 确保每份 g 都包含了模型梯度一部分信息。

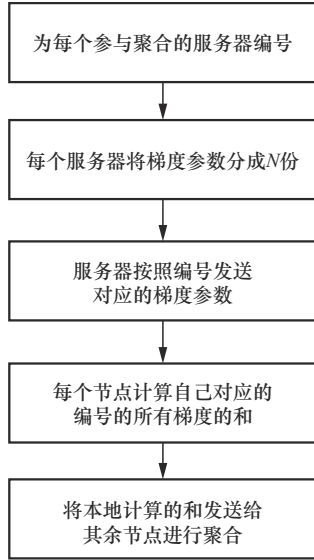


图3 SMPC 安全聚合流程

Figure 3 SMPC security aggregation process

在顶层聚合开始后, 服务器 S_i 将分割后的梯度 $\hat{g}_i = (g_{i1}, g_{i2}, \dots, g_{ii}, \dots, g_{im})$ 按照编号依次分发给其余不同机构 $(1, 2, \dots, i, \dots, m)$, 每个机构均能获得 g_i 的一部分参数而非完整的 g_i 数据。随后, 服务器 S_i 负责计算所有机构的第 i 份梯度参数之和 $\tilde{g}_i = g_{1i} + g_{2i} + \dots + g_{mi}$, 并将结果发送给所有机构的服务器。其余机构的服务器 S 均进行相同操作, 最终, 每个服务器 S 均能获得梯度之和并计算出最终用于聚合的梯度 $\bar{g} = \sum_{i=1}^m \tilde{g}_i / m$, 并不会知晓任何一个服务器 S 的完整梯度信息。通过这种方式, 每个机构都能贡献自己的力量, 同时保护数据隐私, 最终实现了去中心化的模型梯度安全聚合。

3 实验结果与分析

3.1 实验设置

(1) 硬件与框架设置

本文的所有实验在显卡为 NVIDIA GeForce RTX3090, CUDA 版本 12.1, 系统为 Ubuntu 22.04.4 LTS 的服务器上运行。使用 Go 语言处理

底层区块链相关部分, 模型训练相关内容使用 Pytorch 2.1.0 框架实现, 并采用 Go-Python 库实现 Python 和 Go 语言之间的交互。

(2) 数据集与模型选择

实验中训练用的模型采用传统的卷积神经网络 (包含 2 个 5×5 的卷积层, 2 个全连接层)。实验的数据集采用广泛使用的 MNIST、CIFAR10 和 FEMNIST。其中, MNIST 是一个手写数字数据集, 包含 60 000 个训练样本和 10 000 个测试样本, 每个样本是一个标签为 0~9 的 28×28 灰度图像; CIFAR10 是一个彩色图像数据集, 包含 50 000 个训练样本和 10 000 个测试样本, 每个样本是一个 32×32 的彩色图像 (3 通道), 标签包含 “飞机” “鸟” “卡车” 等 10 类图片; FEMNIST 是一个专用于异构联邦学习的手写数字字母数据集, 来源于 leaf 框架, 完整的数据集包含 805 263 个样本, 采用训练测试比为 9: 1 的方法对数据集进行划分, 90% 作为训练样本, 10% 作为测试样本, 每个样本也是一个 28×28 的灰度图像, 标签包含数字 0~9、小写字母 a~z 和大写字母 A~Z 共 62 类。

(3) 实验环境与参数设置

为了模拟联邦学习的分布式环境, 本文实验假设共有 5~20 个机构, 每个机构中有 20 个物联网设备, 并将数据集随机划分给设备作为本地的数据集。为了模拟投毒攻击, 实验随机选取 30% 的设备作为被控制的恶意设备, 并测试了投毒攻击中常见的 4 种攻击类型: 非定向模型投毒攻击^[29]、缩放攻击^[30]、分布式后门攻击^[31]和 ALE (a little is enough, ALE) 攻击^[32], 聚合算法采用 4 种常见的算法: FedAvg、Krum、Trimmed-Mean 和 Median^[33]。

(4) 4 种投毒攻击简介

非定向模型投毒是攻击者向训练数据中注入恶意样本, 从而影响到全局模型针对所有标签类别的分类效果, 引起全局模型精度下降; 缩放攻击是攻击者利用视觉图像处理技术对图片进行压缩, 对目标攻击的图像进行微小扰动, 使图片在视觉上正常, 但在图像压缩后, 模型会将其识别成攻击者期望的目标类别; 分布式后门攻击与常规的中心化后门攻击的区别, 在

于把一个全局触发器 **trigger** 的模式分解为数个独立的局部 **trigger**，并且相应地嵌入不同的恶意攻击方的训练集中，使安全威胁更加持久和隐蔽；ALE 攻击属于模型投毒方法，攻击者对模型中少量参数进行扰动，随着训练的进行，微小扰动的影响会逐步累积，造成模型性能下降。

(5) 4种聚合算法简介

FedAvg 通过计算各个客户端本地梯度的平均值聚合全局模型；**Krum** 是一种抵抗拜占庭错误的聚合算法，它在多个梯度中挑选一个与其余梯度欧几里得距离之和最小的梯度作为聚合后的梯度；**Trimmed-Mean** 是一种鲁棒的计算平均方法，通过剔除梯度中的一些比例的极值后计算平均值，减少异常梯度的影响；**Median** 则是将梯度排序后取位于中间位置的梯度作为聚合结果，异常梯度相较于计算平均方法更加不敏感。

3.2 全局模型性能

为了验证投毒攻击检测模块的有效性，本文在4种常见攻击和 **FedAvg** 聚合算法下，分别检测了无攻击、未部署和部署投毒检测模块情况，底层模型聚合完成后的全局模型精度对比，如表1所示，其中“无投毒聚合”代表服务器直接采用未被投毒的70%良性设备进行聚合，“无防御聚合”代表未部署投毒检测的全设备聚合，

包括良性和恶意设备，“防御聚合”代表部署投毒攻击检测模块后，将检测出的恶意设备剔除训练后的聚合。在实验参数选取上，由于FL训练中前几十轮迭代存在不稳定性，因此本文的实验是在训练的第50轮迭代开始检测恶意客户端，投毒检测迭代轮数 $N = 10$ ，**k-means** 最大聚类数 $k = 10$ 。

由表1可知，在部署了投毒攻击检测模块并将被投毒的恶意设备剔除后，在“非定向模型投毒攻击”的攻击模式下，模型的准确性相比“全设备聚合”有非常明显的提升，与全是良性设备的“无投毒聚合”相当。然而，其余3种攻击通常属于有针对性的投毒攻击，因此，表2测试了3种投毒攻击针对标签为“0”的圖片的攻击成功率，较低的攻击成功率意味着投毒检测模块防御有针对性的投毒攻击较为成功。

由表2可知，投毒检测模块部署后，3种有针对性的投毒攻击的成功率极大地降低。由此可见，通过历史梯度进行投毒检测的模块能够良好地防御不同类型的投毒攻击。此外，投毒检测模块有良好的可扩展性，可以与不同的聚合算法结合，例如 **Median**、**Krum** 等本身具有一定鲁棒性的聚合算法，能够抵抗少量被投毒检测模块遗漏的恶意客户端，进一步保证模型聚合的安全性。

表1 不同场景下全局模型精度对比

Table 1 Comparison of global model accuracy in different scenarios

数据集	攻击模式	精度		
		无投毒聚合	无防御聚合	防御聚合
MNIST	非定向模型投毒	97.6%	69.5%	97.4%
	缩放攻击	97.6%	97.6%	97.6%
	分布式后门攻击	97.6%	97.4%	97.5%
	ALE 攻击	97.6%	97.8%	97.9%
CIFAR10	非定向模型投毒	65.8%	27.8%	65.9%
	缩放攻击	65.8%	66.6%	65.7%
	分布式后门攻击	65.8%	66.1%	65.2%
	ALE 攻击	65.8%	62.1%	64.3%
FEMNIST	非定向模型投毒	64.4%	14.3%	63.2%
	缩放攻击	64.4%	66.4%	64.5%
	分布式后门攻击	64.4%	67.5%	64.3%
	ALE 攻击	64.4%	66.7%	65.0%

表2 有针对性的投毒攻击成功率对比
Table 2 Comparison of success rates of targeted poisoning attacks

数据集	攻击模式	成功率	
		无防御聚合	防御聚合
MNIST	缩放攻击	0.5%	0.5%
	分布式后门攻击	0.5%	0.4%
	ALE 攻击	100.0%	0.3%
CIFAR10	缩放攻击	91.2%	2.4%
	分布式后门攻击	93.5%	1.9%
	ALE 攻击	95.2%	1.8%
FEMNIST	缩放攻击	57.9%	1.7%
	分布式后门攻击	53.2%	2.1%
	ALE 攻击	59.6%	1.6%

3.3 投毒攻击检测能力

为了验证投毒检测模块对投毒攻击的防御能力, 本文分别检测了在4种常见的攻击模式和4种聚合算法下的检测精度、假阳率、假阴率。其中假阳率即良性设备被错误分类成恶意设备的比率; 假阴率即恶意设备被错误分类成正常设备的比率。该实验在数据集FEMNIST上进行, 投毒检测迭代轮数 $N = 10$, k -means最大聚类数 $k = 10$ 。

由表3可得, 在FEMNIST数据集上, 投毒检测模块的检测精度始终都在一个较高的水平。尽管在部分情形下, 假阳率存在大于0的情况, 但假阴率始终保持为0, 并不存在将恶意客户端识别成良性客户端的情况。总体来说, 投毒检测模块可以确保参与聚合的客户端均为良性客户端。此外, 投毒检测模块在Krum、Trimmed-Mean和Median聚合算法上实现了比FedAvg更高的检测准确率, 这是因为这3种聚合算法具有鲁棒性的特点, 本身具有一定的防御能力, 可以帮助投毒检测模块处理遗漏的恶意设备, 也验证了3.2节中提出的观点。

3.4 SMPC 安全聚合时间消耗

为了测试顶层模型聚合时采用的基于秘密共享的SMPC的时间成本, 分别测试了在不同数据集下, 采用FedAvg聚合不同数量的服务器所消耗的时间。

图4显示了顶层采用SMPC进行安全模型聚合的时间消耗。从结果来看, 聚合的时间消耗基本随着参与聚合的服务器数量增加呈线性增长。这说明系统能够相对简单地扩展, 而无须担心非线性效应导致时间成本不可预测地增长, 同时线

表3 针对不同攻击和聚合算法的防御性能指标
Table 3 Defense performance indicators for different attacks and aggregation algorithms

攻击模式	聚合算法	检测精度	假阳率	假阴率
非定向模型投毒	FedAvg	100.0%	0.0	0.0
	Krum	100.0%	0.0	0.0
	Trimmed-Mean	100.0%	0.0	0.0
	Median	100.0%	0.0	0.0
缩放攻击	FedAvg	85.0%	20.0%	0.0
	Krum	100.0%	0.0	0.0
	Trimmed-Mean	98.0%	3.0%	0.0
	Median	100.0%	0.0	0.0
分布式后门攻击	FedAvg	92.0%	11.0%	0.0
	Krum	100.0%	0.0	0.0
	Trimmed-Mean	100.0%	0.0	0.0
	Median	100.0%	0.0	0.0
ALE 攻击	FedAvg	93.0%	10.0%	0.0
	Krum	100.0%	0.0	0.0
	Trimmed-Mean	100.0%	0.0	0.0
	Median	100.0%	0.0	0.0

性关系也使得规划和资源分配变得简单，能够更加有效地利用有限的资源。

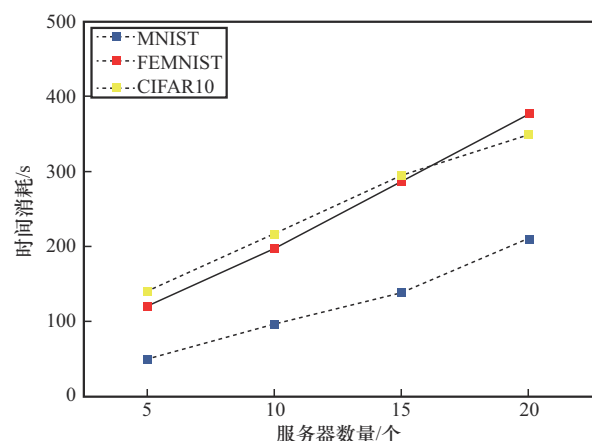


图4 SMPC安全聚合时间消耗

Figure 4 SMPC security aggregation time consumption

4 结束语

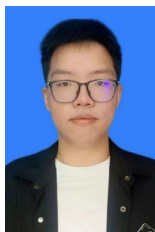
本文设计了面向智能物联网的双层级联邦安全学习架构，避免了恶意设备投毒干扰全局模型和梯度隐私泄露等问题。在底层架构聚合阶段，服务器通过历史梯度数据对上传梯度进行预测和分簇，避免了投毒攻击，在顶层架构聚合阶段通过基于秘密共享的SMPC保护了梯度的隐私，实现了安全可靠的聚合。实验结果表明：该架构设计方案能够很好地防御投毒攻击并保护梯度隐私。本文的重点在于对使用FL的AIoT场景进行投毒攻击防御，并以隐私保护的方式进行模型聚合，然而未充分利用区块链网络中的奖惩机制，下一步将考虑对区块链中良性节点设备进行奖励，并对恶意节点设备进行惩罚的措施，鼓励更多节点设备参与到训练中。

参考文献：

- [1] LIU B Y, YAN B J, ZHOU Y Z, et al. Experiments of federated learning for COVID-19 chest X-ray images[EB]. arXiv Preprint arXiv: 2007.05592, 2020.
- [2] SELLER M J, REINA G A, EDWARDS B, et al. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation[J]. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries BrainLes, 2019, 11383: 92-104.
- [3] CHEN Y Q, QIN X, WANG J D, et al. FedHealth: a federated transfer learning framework for wearable healthcare[J]. IEEE Intelligent Systems, 2020, 35(4): 83-93.
- [4] ARIVAZHAGAN M G, AGGARWAL V, SINGH A K, et al. Federated learning with personalization layers[EB]. arXiv Preprint arXiv: 1912.00818, 2019.
- [5] SELLER M J, EDWARDS B, REINA G A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data[J]. Scientific Reports, 2020, 10: 12598.
- [6] SAPUTRA Y M, HOANG D T, NGUYEN D N, et al. Energy demand prediction with federated learning for electric vehicle networks [C]//Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2019: 1-6.
- [7] 高莹, 陈晓峰, 张一余, 等. 联邦学习系统攻击与防御技术研究综述[J]. 计算机学报, 2023, 46(9): 1781-1805.
- [8] GAO Y, CHEN X F, ZHANG Y Y, et al. A survey of attack and defense techniques for federated learning systems[J]. Chinese Journal of Computers, 2023, 46(9): 1781-1805.
- [9] ZHU L, LIU Z, HAN S. Deep leakage from gradients[C]//Proceedings of Advances in Neural Information Processing Systems, 2019, 32.
- [10] BLANCHARD P, EL MHAMDI E M, GUERRAOU R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent [C]// Proceedings of Advances in Neural Information Processing Systems, 2017, 30.
- [11] XIE C, KOYEJO O, GUPTA I. Generalized byzantine-tolerant SGD [EB]. arXiv preprint arXiv:1802.10116, 2018.
- [12] YIN D, CHEN Y D, RAMCHANDRAN K, et al. Byzantine-robust distributed learning: towards optimal statistical rates[EB]. arXiv Preprint arXiv: 1803.01498, 2018.
- [13] CHEN Y D, SU L L, XU J M. Distributed statistical machine learning in adversarial settings[J]. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2017, 1(2): 1-25.
- [14] ALISTARH D, ALLEN-ZHU Z, LI J. Byzantine stochastic gradient descent[EB]. arXiv Preprint arXiv: 1803.08917, 2018.
- [15] XIA Q, TAO Z Y, HAO Z J, et al. FABA: an algorithm for fast aggregation against Byzantine attacks in distributed neural networks [C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019.
- [16] PRAKASH S, AVESIMEHR A S. Mitigating byzantine attacks in federated learning[EB]. arXiv Preprint arXiv: 2010.07541, 2020.
- [17] XIA G M, CHEN J, YU C D, et al. Poisoning attacks in federated learning: a survey[J]. IEEE Access, 2023, 11: 10708-10722.
- [18] MUÑOZ-GONZÁLEZ L, CO K T, LUPU E C. Byzantine-robust federated machine learning through adaptive model averaging[EB]. arXiv Preprint arXiv: 1909.05125, 2019.
- [19] CAO X Y, LAI L F. Distributed gradient descent algorithm robust to an arbitrary number of Byzantine attackers[J]. IEEE Transactions on Signal Processing, 2019, 67(22): 5850-5864.
- [20] YAN H, ZHANG W, CHEN Q, et al. Recess vaccine for federated learning: proactive defense against model poisoning attacks[C]// Proceedings of Advances in Neural Information Processing Systems, 2024, 36.

- [20] DWORK C. Differential privacy automata, languages and programming[M]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006: 1-12.
- [21] 李敏, 肖迪, 陈律君. 兼顾通信效率与效用的自适应高斯差分隐私个性化联邦学习[J]. 计算机学报, 2024, 47(4): 924-946.
LI M, XIAO D, CHEN L. Adaptive Gaussian differential privacy personalized federated learning considering communication efficiency and utility[J]. Chinese Journal of Computers, 2024, 47(4): 924-946.
- [22] ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes[J]. ACM Computing Surveys, 2019, 51(4): 1-35.
- [23] CHEN Q, WANG Z L, CHEN J W, et al. Dap-FL: federated learning Flourishes by adaptive tuning and secure aggregation[J]. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(6): 1923-1941.
- [24] TRUEX S, BARACALDO N, ANWAR A, et al. A hybrid approach to privacy-preserving federated learning[C]//Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security. 2019: 1-11.
- [25] 肖雄, 唐卓, 肖斌, 等. 联邦学习的隐私保护与安全防御研究综述[J]. 计算机学报, 2023, 46(5): 1019-1044.
XIAO X, TANG Z, XIAO B, et al. A survey on privacy and security issues in federated learning[J]. Chinese Journal of Computers, 2023, 46(5): 1019-1044.
- [26] AHMED M, SERAJ R, ISLAM S M S. The k-means algorithm: a comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.
- [27] TIBSHIRANI R, WALTHER G, HASTIE T. Estimating the number of clusters in a data set via the gap statistic[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 2001, 63(2): 411-423.
- [28] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[EB]. arXiv Preprint arXiv: 1602.05629, 2016.
- [29] FANG M, CAO X, JIA J, et al. Local model poisoning attacks to byzantine-robust federated learning[C]//Proceedings of the 29th Usenix Conference on Security Symposium. 2020: 1623-1640.
- [30] XIE C, HUANG K, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning[C]//Proceedings of International Conference on Learning Representations. 2019.
- [31] BAGDASARYAN E, VEIT A, HUA Y Q, et al. How to backdoor federated learning[EB]. arXiv Preprint arXiv: 1807.00459, 2018.
- [32] BARUCH M, BARUCH G, GOLDBERG Y. A little is enough: circumventing defenses for distributed learning[EB]. arXiv Preprint arXiv: 1902.06156, 2019.
- [33] YIN D, CHEN Y D, RAMCHANDRAN K, et al. Byzantine-Robust distributed learning: towards optimal statistical rates[EB]. arXiv Preprint arXiv: 1803.01498, 2018.

[作者简介]



郑健波 (1999-), 男, 江苏南通人, 西安电子科技大学硕士生, 主要研究方向为联邦学习、人工智能安全。



闫皓楠 (1996-), 男, 山西忻州人, 西安电子科技大学博士生, 主要研究方向为可信人工智能、隐私计算。



傅彩利 (1986-), 男, 浙江杭州人, 杭州海康威视数字技术股份有限公司高级工程师, 主要研究方向为物联网安全。



张栋 (1991-), 男, 河北衡水人, 国家计算机网络应急技术处理协调中心工程师, 主要研究方向为人工智能安全、区块链、下一代互联网。



李晖 (1968-), 男, 河南灵宝人, 西安电子科技大学教授、博士生导师, 主要研究方向为密码信息安全、隐私计算、信息论与编码理论。



王滨 (1978-), 男, 山东泗水人, 浙江省全省智能物联网与数据安全重点实验室研究员, 主要研究方向为物联网安全、人工智能安全、密码学。