

# 边缘计算下的轻量级联邦学习隐私保护方案

张海超<sup>1</sup>, 赖金山<sup>2</sup>, 刘东<sup>3</sup>, 张凤荔<sup>2</sup>

- (1. 四川公安厅科技信息化总队, 四川 成都 610015;
2. 电子科技大学 信息与软件工程学院, 四川 成都 610054;
3. 电子科技大学 计算机科学与工程学院, 四川 成都 611731)

**摘要:**随着物联网和大数据技术的高速发展,以传统云计算模式为代表的集中式学习效率低下,且易受到单点攻击、共谋攻击、中间人攻击等一系列攻击手段,造成数据安全的隐患。边缘计算模式使得分布式联邦学习成为了可能,然而,联邦学习虽然能够保证数据在本地安全和隐私,但是也面临众多安全威胁,如梯度泄露攻击,此外,效率问题也是联邦学习的痛点所在。为了保障边缘计算场景下的模型训练安全,提出了一种边缘计算下的轻量级联邦学习隐私保护方案(Lightweight Federated Learning Privacy Protection Scheme Under Edge Computing, LFLPP)。首先,提出一种云-边-端分层的联邦学习框架;其次,对不同层进行隐私保护;最后,提出一种周期性更新策略,极大地提高了收敛速度。使用乳腺癌肿瘤数据集和 CIFAR10 数据集在 LR 模型和 Resnet18 残差模型上进行训练和测试,同时使用 CIFAR10 数据集与 FedAvg 和 PPFLEC (Privacy-Preserving Federated Learning for Internet of Medical Things under Edge Computing) 两种方案进行对比实验,得出准确率和训练效率的差距,并进行准确率、效率以及安全性等方面的分析,该方案在 CIFAR-10 数据集上能达到 84.63% 的准确率。

**关键词:**联邦学习;边缘计算;同态加密;差分隐私;隐私保护

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2023)09-0161-07

doi:10.3969/j.issn.1673-629X.2023.09.024

## Lightweight Federated Learning Privacy Protection Scheme under Edge Computing

ZHANG Hai-chao<sup>1</sup>, LAI Jin-shan<sup>2</sup>, LIU Dong<sup>3</sup>, ZHANG Feng-li<sup>2</sup>

- (1. Science and Technology Informatization Corps of Sichuan Public Security Department, Chengdu 610015, China;
2. School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China;
3. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

**Abstract:** With the rapid development of the Internet of Things and big data technology, centralized learning represented by the traditional cloud computing model is inefficient and vulnerable to a series of attacks such as single point attack, collusion attack, man in the middle attack, resulting in hidden dangers of data security. The edge computing model makes distributed federated learning possible. However, although federated learning can ensure the security and privacy of data locally, it also faces many security threats, such as gradient disclosure attacks. In addition, the efficiency is also the pain point of federated learning. In order to ensure the security of model training in the edge computing scenario, a lightweight federated learning privacy protection scheme under edge computing (LFLPP) is proposed. Firstly, a cloud-edge-end layered federated learning framework is proposed. Secondly, privacy protection for different layers. Finally, a periodic updating strategy is proposed, which greatly improves the convergence speed. The breast cancer tumor data set and CIFAR10 data set were used for training and testing on LR model and Resnet18 residual model. At the same time, CIFAR10 data set was used to conduct comparative experiments with FedAvg and PPFLEC (Privacy Preserving Federated Learning for Internet of Medical Things under Edge Computing), to find out the gap between accuracy and training efficiency, and to conduct accuracy analysis, efficiency analysis and security analysis. This scheme can achieve 84.63% accuracy on CIFAR-10 dataset.

**Key words:** federated learning; edge computing; homomorphic encryption; differential privacy; privacy protection

收稿日期:2022-10-28

修回日期:2023-02-28

基金项目:四川省科技计划项目(2021YFS0391);四川省重大科技专项(22DZX0046);国家自然科学基金重点项目(61133016)

作者简介:张海超(1985-),男,研究方向为信息化系统建设应用;通讯作者:赖金山(1998-),男,博士,CCF会员(I4152G),研究方向为网络信息安全。

## 0 引言

物联网时代,智能终端产生的大量数据促进了机器学习的高速发展。由于连接设备数和产生的数据量都呈现指数级增长,传统的云计算模式已经不能跟上物联网时代的脚步,边缘计算模式<sup>[1]</sup>的出现使得分布式机器学习成为可能。同时,谷歌提出的联邦学习<sup>[2]</sup>不仅完美地契合边缘计算模式,还能将数据保留在终端设备上<sup>[3]</sup>,降低数据泄露的风险,很好地解决了数据孤岛问题<sup>[4]</sup>。此外,终端设备算力和内存的提升也使得模型训练成为了可能。针对不同的领域,终端设备使用不同的数据集训练一个好的神经网络模型,能够高效地解决各个领域的决策和分类等问题<sup>[5]</sup>。但是,随着梯度泄露攻击<sup>[6]</sup>、中间人攻击<sup>[7]</sup>和共谋攻击<sup>[8]</sup>等攻击手段的兴起以及参与训练的终端设备的不可信,使得攻击者可以通过模型梯度推演出本地数据集,导致各个终端设备不愿将本地训练结果上传给边缘服务器。同时边缘服务器到云端的数据传输通过核心网,传输过程中可能遭遇不同的攻击。此外,经典的FedAvg<sup>[9]</sup>模型聚合方法存在通信时延高的问题。为了解决通信时延问题,王等人<sup>[10]</sup>提出将模型训练任务卸载到边缘服务器上,将训练数据集从终端上传到边缘服务器,在边缘服务器上训练模型,减少终端设备和边缘服务器之间的通信次数,减少通信时间。但是数据从终端设备上传到边缘服务器的过程中容易遭到隐私泄露。

为了解决隐私泄露问题,文献[11]提出了分层联邦学习同态加密方法,对本地数据集进行同态加密后再进行神经网络模型的训练,但是对本地大量的数据集进行同态加密需要耗费大量的时间和空间,这并不适用于边缘计算下实时性的场景。文献[12]提出了使用多方安全计算来保证数据安全,但是攻击者可以通过获得密钥来获得数据或梯度,同时多方安全计算不适用于分布式场景。文献[13]提出了差分隐私方案,对模型梯度添加噪声或者在本地对数据添加噪声。为了平衡安全和精度,文献[14]提出了一种自适应的差分隐私方案,通过对梯度进行自适应裁剪来提高精度,同时还可以降低吞吐量,减少时延。文献[15]提出利用网络不同层敏感度来进行模型压缩,解决权重参数冗余的问题,以达到模型训练效率和模型复杂度之间的平衡。文献[16]提出使用参数稀疏化来传输与掩码相与之后不为0的参数,能防止模型参数泄露。

以上方法虽然能够在一定程度上保护用户隐私,但是应用在边缘计算场景下的联邦学习需要耗费大量的通信成本,同时面临各种攻击手段。为此,该文提出了一种边缘计算下的轻量级联邦学习隐私保护方案(Lightweight Federated Learning Privacy Protection

Scheme Under Edge Computing, LFLPP),使用加性同态加密来保护在云服务器和边缘服务器上的模型参数。所有模型参数都被加密并存储在云服务器和边缘服务器上,该方案有以下贡献:

(1)提出了一种云-边-端分层的联邦学习架构:在云-边-端分层架构中,不同的终端设备负责进行模型训练,而边缘服务器和云服务器进行模型参数聚合。

(2)提出了一种基于差分隐私和同态加密的两层隐私保护方案:在终端设备训练得到模型参数后,对其进行差分扰动,再上传给边缘服务器,边缘服务器迭代训练得到模型参数后,对模型参数进行同态加密后传输到云服务器进行聚合,保证参数在传输过程中的安全。

(3)提出了一种本地和边缘服务器端多次迭代更新的策略:在终端设备上设置训练轮数阈值,当更新次数达到阈值,将本地模型上传到边缘服务器,同时在边缘服务器设置另一个阈值,当训练次数达到阈值时,将模型参数上传到云服务器进行聚合更新。

基于此,边缘计算下的轻量级联邦学习隐私保护方案能够高效率地进行模型训练,同时实现不同层之间的数据隐私保护。

## 1 系统架构

所提出的架构分为云边端三层,一共有3个实体,分别是:(1)终端设备:负责数据采集和预处理,从边缘服务器获取模型框架,在本地对神经网络模型进行周期性训练,并对模型参数进行微分扰动;(2)边缘服务器:从云端获取将需要训练的模型框架发送给参与训练的终端设备,对终端设备上传的模型参数进行聚合,对聚合后的模型参数进行同态加密后上传到中央云服务器进行聚合;(3)中央云服务器(Central Cloud Server, CS):聚合各个边缘服务器上传的本地模型参数,并将聚合后的全局模型参数发送到边缘服务器,为下一轮训练更新模型参数。总体框架如图1所示,相关实体参数如表1所示。

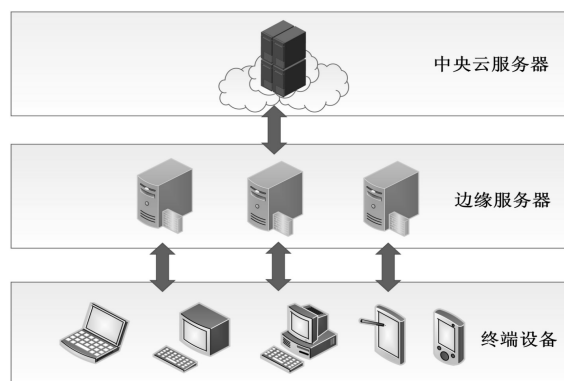


图1 边缘计算下的联邦学习架构

表1 相关参数

名称	说明
$d_i$	第 $i$ 个终端设备
$D_i$	第 $i$ 个终端设备的训练集
$T_i$	第 $i$ 个终端设备的标签集
$E$	同态加密算子
$g_i^t(k)$	第 $i$ 个终端设备上第 $t$ 轮训练中迭代 $k$ 次得到的梯度
$\omega_i^t(k)$	第 $i$ 个终端设备上第 $t$ 轮训练中迭代训练 $k$ 次得到的模型参数
$\omega_j^t(k)$	第 $i$ 个边缘服务器上第 $t$ 轮训练中迭代训练 $k$ 次得到的模型参数
$Y_i$	模型输出

## 2 方案

### 2.1 问题定义

该方案旨在保证模型精度和训练效率的前提下,对数据集和模型参数进行隐私保护。在对方案进行评测时,相关定义如下:

**差分隐私:**给定一个数据集  $D$  和相邻数据集  $D'$ ,对于查询函数  $f$ ,如果满足以下式子,则  $f$  满足差分隐私:

$$\Pr[f(D) \in R] \leq \exp(\epsilon) * \Pr[f(D') \in R] + \delta \quad (1)$$

其中,  $\epsilon$  代表隐私预算,  $\epsilon$  越大代表数据可用性越高,越小代表隐私保护程度越高,添加的噪声越大,当  $\epsilon$  为 0 时,代表没有添加差分隐私;  $\delta$  代表置信度参数,在严格差分隐私中,  $\delta$  为 0,当  $\delta > 0$  时,为近似差分隐私,在实际工业中,近似差分隐私广泛使用。

**拉普拉斯机制:**给定一个数据集  $D$  和查询函数  $f$ ,则提供差分隐私的机制  $M$  满足:

$$M(D) = f(D) + \text{lap}(\frac{\Delta f}{\epsilon}) \quad (2)$$

其中,  $\Delta f$  代表全局敏感度,计算公式如下:

$$\Delta f = \max \|f(D) - f(D')\| \quad (3)$$

代表相邻两个数据集查询结果差值的 1 范数的最大值。

**加法同态性质:**对于任意明文  $m_1, m_2$  和随机数  $r_1, r_2$ ,对应密  $c_1 = E[m_1, r_1], c_2 = E[m_2, r_2]$ ,满足:

$$c_1 * c_2 = E[m_1, r_1] * E[m_2, r_2] = g^{m_1+m_2} * (r_1 * r_2)^n \bmod n^2 \quad (4)$$

### 2.2 云-边-端分层联邦学习架构

在本方案中,将传统的两层联邦学习应用到边缘计算场景下,形成了云-边-端三层联邦学习架构。首先,终端设备使用本地数据集进行本地模型训练,同时使用周期性策略进行更新。同样,模型迭代聚合后,边

缘服务器将模型参数同态加密后上传到中心云服务器,用于聚合后更新,再进行下一轮模型训练,系统整体框架如图 2 所示,周期性更新策略如图 3 所示。

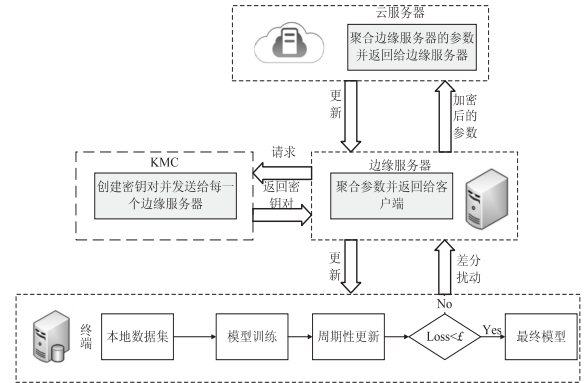


图2 边缘计算下的联邦学习架构

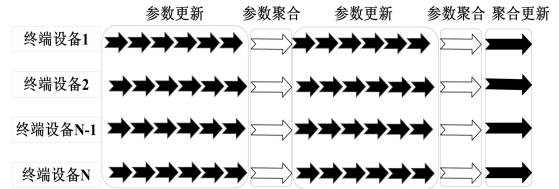


图3 模型更新策略

其中,参数更新部分箭头表示在终端设备上进行的模型参数更新,参数聚合部分箭头表示在边缘服务器将终端设备上传的模型参数进行聚合,聚合更新部分箭头表示迭代到一定次数之后,云服务器收集来自边缘服务器的模型参数并聚合更新,完成这一轮的训练。可以看到,针对不同算力和数据集的终端设备,在每一轮的模型更新次数是不同的,极大地避免了同步联邦学习造成的资源浪费。

### 2.3 基于差分隐私的模型参数保护机制

设终端设备  $d_i$  拥有数据集  $D_i$ ,包括数据  $X_i$  和标签  $T_i$ ,将数据集按照 7:3 的比例划分为训练集  $D_{\text{train}}$  和测试集  $D_{\text{test}}$ ,然后将训练集作为神经网络模型的输入进行模型训练。设模型的输出为  $Y$ ,则:

$$Y = f(X_i) \quad (5)$$

对比  $Y$  和  $T_i$ ,计算出准确率和损失值。在每一轮训练完成之后,将测试数据集输入到神经网络中,计算准确率和损失值,当达到本地训练轮数的阈值时,将模型参数进行差分扰动并上传给边缘服务器。

本地梯度计算公式如下:

$$g_i^t(k) = \nabla[f(X_i) - T_i] \quad (6)$$

对模型梯度进行裁剪,裁剪公式如下:

$$g_i^t(k) = \frac{g_i^t(k)}{\max(1, \|g_i^t(k)\|)} \quad (7)$$

在得到第  $k$  次训练的模型梯度后,判断  $k$  是否小于  $k_1$ ,若  $k$  小于阈值  $k_1$ ,则在本地对模型参数进行更新,更新公式如下:



$$\omega_i'(k+1) = \omega_i'(k) - \eta * g_i'(k) \quad (8)$$

若  $k$  等于  $k_1$ , 则将模型参数  $\omega_i^k$  发送给边缘服务器进行聚合, 由于边缘服务器并不完全可信, 因此在本地对模型参数添加差分隐私, 采用随机扰动机制。设模型参数被攻击者窃取的概率为  $p$ , 数据集到模型参数的映射函数为  $h$ , 映射公式如下:

$$\omega_i'(k) = h_i'(D_k) \quad (9)$$

其中,  $D_k$  表示第  $k$  次训练时划分的训练数据集, 令隐私预算  $\epsilon = \ln \frac{p}{1-p}$ , 噪声添加如下:

$$M(D_k) = h_i'(D_k) + \text{lap}(\frac{\Delta g}{\epsilon}), k \mid k_1 = 0 \quad (10)$$

其中,  $\text{lap}$  为拉普拉斯分布,  $\Delta g = \|\omega_i' - (\omega_i')'\|$  为查询敏感度, 当本地训练次数达到  $k_1$  时添加噪声, 实现本地数据集和模型参数的隐私保护。本地差分隐私机制如图 4 所示。

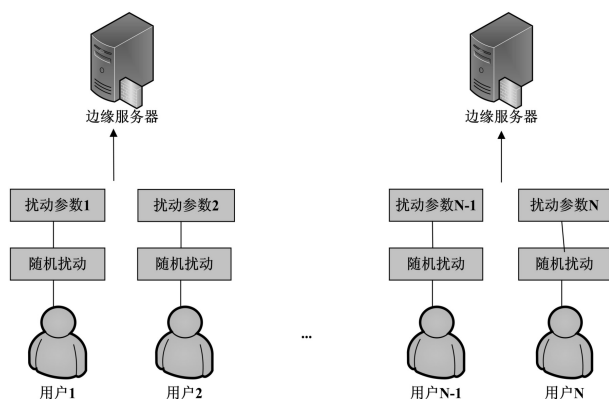


图 4 本地差分隐私模型

## 2.4 面向边云模型传输的同态加密隐私保护

密钥管理中心 (KMC) 收到来自边缘服务器的请求之后, 创建密钥对并分发给边缘服务器, 密钥生成采用满足加法同态的 Paillier<sup>[17]</sup> 算法, 密钥生成与分配算法如算法 1 所示。

算法 1: 密钥生成与分配算法

输入: 边缘服务器集合  $E$

For  $j$  in  $E$ :

1. 随机选取大素数  $p$  和  $q$  且满足  $p * q$  和  $(p-1) * (q-1)$  的最大公因数为 1;
2. 令  $p * q$  为  $n$ ,  $(p-1)$  和  $(q-1)$  的最小公倍数为  $\lambda$ ;
3. 随机选取  $G$  且满足  $G$  的阶 (mod  $n^2$ ) 为  $n$  的倍数生成公私钥对, 公钥为  $(n, g)$ , 私钥为  $\lambda$ ;
4. 将公私钥对发送给边缘服务器  $j$ 。

为了安全考虑, 使用同态加密算法加密上传的参数来达到保护隐私数据的目的, 模型更新公式如下, 其中  $E$  表示同态加密,  $t$  表示训练轮数,  $e$  表示边缘服务器的数量,  $d$  表示终端设备的数量。

$$E(\omega_i'(k)) =$$

$$\begin{cases} E(\sum_{i=1}^d \omega_i'(k) * \frac{1}{d}), (k \mid k_1 = 0, k \mid k_1 * k_2 \neq 0) \\ E(\sum_{j=1}^e \omega_j'(k) * \frac{1}{e}), (k \mid k_1 * k_2 = 0) \end{cases} \quad (11)$$

其中,  $k_1$  表示在每轮训练中, 终端设备上迭代次数的阈值,  $k_2$  表示在边缘服务器上迭代次数的阈值, 在训练轮数  $t$  中, 当迭代次数未到达  $k_1 * k_2$  次时, 边缘服务器对  $d$  个终端设备上传的模型参数进行聚合并进行同态加密得到  $E(\omega_i'(k))$ ; 当迭代次数达到  $k_1 * k_2$  次时, 则将边缘服务器解密得到的模型参数更新, 同时进行下一轮训练。对于模型参数, 具体的加密算法如算法 2 所示。

算法 2: 模型参数加解密算法

输入: 模型参数  $\omega'$

1. 使用  $\text{random}()$  生成随机向量  $r$
2. 计算密文

$$c = E(\omega', r) = g^{\omega'} r^n \bmod n^2$$

3. 计算  $L(u) = \frac{u-1}{N}$
4. 解密

$$D(c, \lambda) = \frac{L(c^\lambda \bmod n^2)}{L(g^\lambda \bmod n^2)} \bmod n$$

## 3 实验验证及分析

在本节中, 对 LFLPP 方案进行实验验证。利用笔记本电脑作为终端设备, PC 机作为边缘服务器进行仿真实验。分别就模型准确率、模型收敛效率和模型安全性进行实验验证。

### 3.1 实验配置

使用 Python 语言进行仿真实验, 分别使用乳腺癌肿瘤数据集<sup>[18]</sup>和 CIFAR10<sup>[19]</sup>数据集在 LR 模型和 Resnet18<sup>[20]</sup>残差模型上进行训练和测试, 同时使用 CIFAR10 数据集与 FedAvg 和 PPFLEC<sup>[10]</sup> (Privacy - Preserving Federated Learning for Internet of Medical Things under Edge Computing) 两种方案进行对比实验, 得出准确率和训练效率的差距。其中 PPFLEC 是一种在三层联邦学习架构下进行掩码添加的隐私保护方案。

### 3.2 实验数据

本次实验中使用了两个有代表性的数据集 (良/恶性乳腺癌肿瘤数据集和 CIFAR-10 数据集) 进行实验。

良/恶性乳腺癌肿瘤数据集是将病人数据格式化之后的带标签的数据集, 该数据集将肿瘤细胞分为两类: 良性肿瘤和恶性肿瘤, 根据肿瘤细胞的外观特征以及细胞核的特征来划分。该数据集收集了 699 条病人样本, 共 11 列数据。每一轮训练时随机选择 500 条作

为训练数据,199 条作为测试数据。

CIFAR-10 数据集是一个包含 60 000 张图片的数据集。其中每张照片为  $32 \times 32$  的彩色照片,每个像素点包括 RGB 三个数值,数值范围为  $0 \sim 255$ 。所有照片分属 10 个不同的类别,分别是 ‘airplane’ ‘automobile’ ‘bird’ ‘cat’ ‘deer’ ‘dog’ ‘frog’ ‘horse’ ‘ship’ ‘truck’。每一轮训练时随机选择 50 000 条作为训练数据,10 000 条作为测试数据。

### 3.3 评价指标

通过自定义正类和负类,通过与基本事实进行比较,可以获得表 2,从而计算准确率。

表 2 正负类结果判定

正类	负类
真阳性(true positive, TP)	假阳性(false positive, FP)
假阴性(false negative, FN)	真阴性(true negative, TN)

准确率定义:

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN}) * 100\%$$

损失值定义:

$$\text{Loss} = \sum_{i=1}^N \text{loss}_i(\hat{y}, y)$$

其中,  $\hat{y}$  是测试数据集样本  $x$  的预测输出,  $y$  是测试数据集的样本标签。

### 3.4 结果分析

首先使用乳腺癌肿瘤数据集进行实验,采用逻辑回归模型输出结果,得到的准确率如图 5 所示。

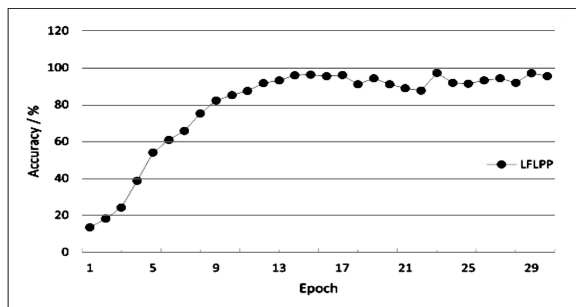


图 5 乳腺癌数据集准确率

从图 5 可以看出,训练一开始 LFLPP 可以达到 75% 的准确率,随后快速提升,在第 10 轮达到 90% 左右并在第 10 轮到第 20 轮缓慢增加,在第 25 轮左右达到收敛,收敛时能达到 95% 的准确率。该实验结果表明 LFLPP 能够很好地应用于医疗领域,用于病情诊断,这将大大减少医疗误诊,同时缓解医疗专家不足带来的压力。

为了验证 LFLPP 在影像识别领域的精度,使用 CIFAR10 数据集进行实验测试,为了使结果具有参考性,使用经典的联邦聚合方法 FedAvg 和隐私保护方案 PPFLC 进行对比,得到的准确率对比如图 6 所示,

损失值对比如图 7 所示。

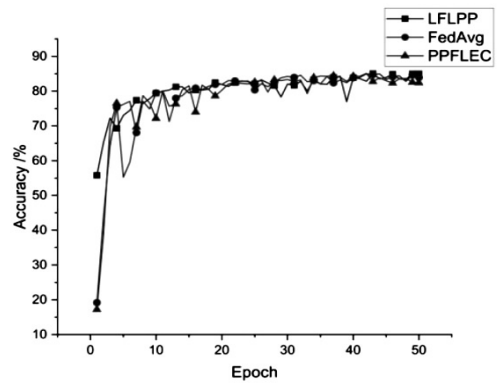


图 6 三种方案在 CIFAR10 数据集上的准确率对比

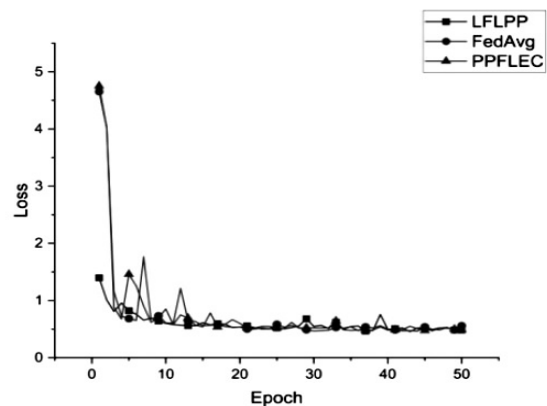


图 7 三种方案在 CIFAR10 数据集上的损失值对比

从图 6 可以看到,FedAvg 作为经典的联邦平均聚合方案,收敛时能达到 85% 的准确率,LFLPP 在收敛时能达到 84.63% 的准确率,而 PPFLC 在收敛时的准确率为 82.95%。具体来说,LFLPP 由于在终端设备上添加了噪声,会在一定程度上影响模型精度,但是在残差网络 Resnet18 中仍能达到不错的准确率,略低于 FedAvg,明显高于 PPFLC。训练开始时,由于噪声的存在,LFLPP 的准确率只有 20% 左右,经过 10 轮左右的训练,准确率能达到 75% 左右,随后缓慢提升,从图 6 中可以看到,LFLPP 的稳定性明显高于 PPFLC,与 FedAvg 大致持平,这是由于随着训练轮数的增加,模型参数趋于稳定,噪声对其的影响减弱。

从图 7 也开始看出,训练开始的前几轮,LFLPP 的损失值高达 4,经过 10 轮左右的训练,损失值降低到 1 以下,在训练轮数达到 30 时,LFLPP 的损失值趋于平稳,为 0.48 左右,与 FedAvg 大致相同,而 PPFLC 在近 50 轮时才接近平稳,同时高于 FedAvg 和 LFLPP。由此可以看出,LFLPP 在保护隐私的同时对于图像识别也能达到不错的准确率。

为了测试 LFLPP 的效率,同样对比了 FedAvg 和 PPFLC,在同一设备上不同方案的训练,训练时间对比如表 3 所示。

表 3 三种方案训练时间对比

训练时间	LFLPP	FedAvg	PPFLEC
每轮平均时间/s	447.95	431.08	468
收敛平均时间/s	21 950	21 123	23 400

从表 3 可以看到, LFLPP 的模型训练时间高于 FedAvg, 但低于 PPFLEC, 相较于 FedAvg, LFLPP 需要对模型参数进行噪声添加以及同态加解密操作, 这会导致训练时间的增加, 但是轻量级的 LFLPP 在模型收敛时所用时间与 FedAvg 相差不大, 远低于 PPFLEC。这说明 LFLPP 可以适用于需要隐私保护的实时场景下。为了验证这一结论的普遍性, 在 50 个不同的终端设备上进行了实验, 得到每一轮训练所需的时间分布如图 8 所示。

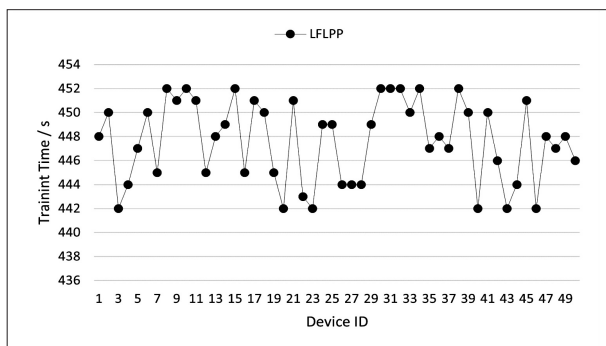


图 8 不同终端设备训练时间

从图 8 可以看到, 针对不同的终端设备, 训练时间分布在 442 秒到 452 秒之间, 且集中分布在 446 秒到 448 秒之间, 说明 LFLPP 在训练时间上具有一定的稳定性, 如果使用 GPU 对图像进行计算, 训练时间可以缩短至 1/20, 可见 LFLPP 可以适配于不同的终端设备, 满足边缘计算下实时场景的需求。

## 4 安全分析

在本节中, 主要对隐私保护方案进行理论上的分析, 通过给出隐私证明以及分析其相关流程, 可以得出该隐私保护方案能够很好地保护数据和模型。

本节分别就 LFLPP 的安全性方面进行理论分析, 主要就数据的隐私性和抵抗攻击的能力进行分析。

### 4.1 数据安全分析

参与训练的终端设备在本地进行模型训练, 然后将参数进行差分扰动后发送给边缘服务器; 之后, 边缘服务器在聚合完成后将参数进行同态加密后再发送给云服务器。在整个过程中, 终端设备不必将数据集发送给任何实体, 保护了用户数据的隐私性。基于拉普拉斯的差分隐私机制证明如下:

设数据集  $D$  经过  $h(D)$  变化为  $n$  维模型参数  $\omega'_i(k)$ , 设  $\omega'_i(k) = (w_1, w_2, \dots, w_n)^T$ , 结果集  $S = (s_1, s_2, \dots, s_n)^T$ , 则:

$$\Delta f = \max \|h(D) - h(D')\| = \max \left( \sum_{i=1}^n |\Delta w_i| \right) \quad (12)$$

$$\frac{\Pr(M(D) \in S)}{\Pr(M(D') \in S)} =$$

$$e^{\frac{\epsilon}{\Delta f} \sum_{i=1}^n (|\Delta w_i - s_i| - |s_i|)} < e^{\frac{\epsilon}{\Delta f} \sum_{i=1}^n |\Delta w_i|} = e^\epsilon \quad (13)$$

### 4.2 抵抗恶意攻击分析

由于成员推理攻击的威胁, 以纯文本传输梯度数据可能被恶意用户利用来训练他自己的阴影模型。其他终端设备的隐私相关数据安全将受到侵犯。为了抵抗梯度泄露攻击和共谋攻击, 在对模型参数进行差分扰动后再进行传输。此外, 在边缘服务器中进行同态运算, 即使中心云服务器存在安全漏洞, 也能保证加密后的模型参数不会泄露。在边缘服务器与密钥管理中心的交互中, 密钥管理中心只负责密钥生成, 不请求访问任何模型参数, 对于密钥管理中心来说, 它甚至不知道边缘服务器使用密钥的维数, 因此无法与其他方进行共谋攻击来窃取模型参数。

为了验证 LFLPP 的隐私保护效果, 使用梯度泄露攻击<sup>[6]</sup>进行仿真测试, 同时对 LFLPP 和 FedAvg 进行梯度泄露攻击, 使用 CIFAR10 图片的变化作为对比, 如图 9 和图 10 所示。

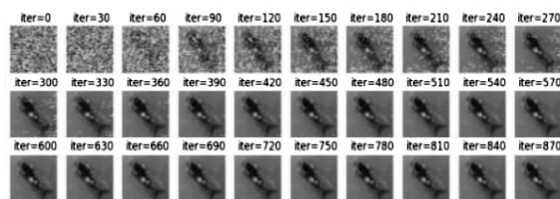


图 9 梯度泄露

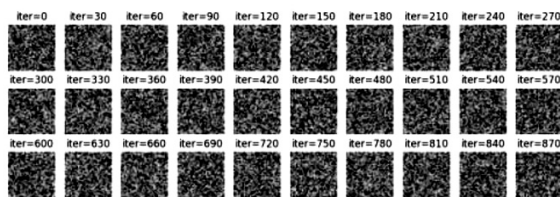


图 10 梯度未泄露

在针对 FedAvg 的梯度泄露攻击中, 攻击者首先使用虚假数据和标签来参与神经网络的训练, 并且通过训练所得梯度来推断真实训练数据集。对于一个  $N$  维的梯度向量, 攻击者至多需要  $N+1$  次参与就可以推断出真实数据。从图 9 可以看出, 当迭代次数增加时, 图像的特征向量被推断出来, 图像也逐渐被还原出来。而在 LFLPP 中, 如图 10 所示, 由于差分隐私和同态加密机制, 导致攻击者得到的梯度是被加密或者被添加了噪声的, 无法由梯度得到原始图像。由此可见, LFLPP 可以很好地抵抗梯度泄露攻击, 保护用户数据的隐私。

## 5 结束语

在提出分层联邦学习架构的基础上,针对每一层设计不同的隐私保护方法,并在不同的数据集上取得了不错精度,能够适配智慧医疗等各种适合应用边缘计算的场景。未来将考虑使用边缘服务器将本地模型进行迁移来实现不同终端设备之间互相学习,以提高模型的准确率。

### 参考文献:

- [1] 马富齐,王 波,董旭柱,等. 电力视觉边缘智能:边缘计算驱动下的电力深度视觉加速技术[J]. 电网技术,2020,44(6):2020-2029.
- [2] ALAZAB M, RM S P, PARIMALA M, et al. Federated learning for cybersecurity: concepts, challenges and future directions[J]. IEEE Transactions on Industrial Informatics, 2022, 18(15):3501-3509.
- [3] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning[J]. Foundations and Trends® in Machine Learning, 2021, 14(1-2):1-210.
- [4] KHAN L U, SAAD W, HAN Z, et al. Federated learning for internet of things: recent advances, taxonomy, and open challenges[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3):1759-1799.
- [5] MUSBAH H, ALI G, ALY H H, et al. Energy management using multi-criteria decision making and machine learning classification algorithms for intelligent system[J]. Electric Power Systems Research, 2022, 203:107645.
- [6] ZHU L, LIU Z, HAN S. Deep leakage from gradients[J]. arXiv:1906.08935, 2019.
- [7] 胡 杨,韩增杰,叶幅华,等. 基于无证书签名的抗 DNS 中间人攻击方案[J]. 网络与信息安全学报, 2021, 7(6):167-177.
- [8] CRAWFORD B, KEEN F. The hanau terrorist attack: how race hate and conspiracy theories are fueling global far-right violence[J]. CTC Sentinel, 2020, 13(3):118-123.
- [9] LI X, HUANG K, YANG W, et al. On the convergence of fedavg on non-iid data[J]. arXiv:1907.02189, 2019.
- [10] WANG R, LAI J, ZHANG Z, et al. Privacy-preserving federated learning for internet of medical things under edge computing[J]. IEEE Journal of Biomedical and Health Informatics, 2022, 3(2):1-1.
- [11] ZHANG C, LI S, XIA J, et al. BatchCrypt: efficient homomorphic encryption for cross-silo federated learning[C]//USENIX annual technical conference (USENIX ATC 20). Boston: Ada Gavrilovska, 2020:493-506.
- [12] KANAGAVELU R, LI Z, SAMSUDIN J, et al. Two-phase multi-party computation enabled privacy-preserving federated learning[C]//2020 20th IEEE/ACM international symposium on cluster, cloud and internet computing (CC-GRID). Melbourne: IEEE, 2020:410-419.
- [13] WEI K, LI J, DING M, et al. Federated learning with differential privacy: algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15:3454-3469.
- [14] WU X, ZHANG Y, SHI M, et al. An adaptive federated learning scheme with differential privacy preserving[J]. Future Generation Computer Systems, 2022, 127:362-372.
- [15] YU R, LI A, CHEN C F, et al. Nisp: pruning networks using neuron importance score propagation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE, 2018:9194-9203.
- [16] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3):50-60.
- [17] PAILLIER P. Public-key cryptosystems based on composite degree residuosity classes[C]//International conference on the theory and applications of cryptographic techniques. Heidelberg: Springer, 1999:223-238.
- [18] MUSHTAQ Z, YAQUB A, SANI S, et al. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets[J]. Journal of the Chinese Institute of Engineers, 2020, 43(1):80-92.
- [19] HINTON G E. Learning multiple layers of representation[J]. Trends in Cognitive Sciences, 2007, 11(10):428-434.
- [20] CHEN Z, JIANG Y, ZHANG X, et al. ResNet18DNN: prediction approach of drug-induced liver injury by deep neural network with ResNet18[J]. Briefings in Bioinformatics, 2022, 23(1):bbab503.