计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330894
61(9): 2199−2212, 2024

面向车联网数据持续共享的安全高效联邦学习

乐俊青　谭州勇　张　迪　刘　高　向　涛　廖晓峰
（重庆大学计算机学院　重庆　400044）
（信息物理社会可信服务计算教育部重点实验室（重庆大学）　重庆　400044）
（junqingle@cqu.edu.cn）

Secure and Efficient Federated Learning for Continuous IoV Data Sharing

Le Junqing, Tan Zhouyong, Zhang Di, Liu Gao, Xiang Tao, and Liao Xiaofeng
（*College of Computer Science*, *Chongqing University*, *Chongqing* 400044）
（*Key Laboratory of Dependable Service Computing in Cyber Physical Society* (*Chongqing University*), *Ministry of Education*, *Chongqing* 400044）

Abstract　The combination of the Internet of vehicles (IoV) and artificial intelligence (AI) has driven the rapid development of autonomous vehicles. Sharing IoV data distributed across different vehicles for training AI models enables more efficient and reliable intelligent driving services. Autonomous vehicles can continuously gather real-time vehicle information, road images and videos among other IoV data, through onboard cameras and sensors. This data are then utilized to optimize and update intelligent traffic models, addressing issues where changes in IoV data result in decreased model accuracy. We propose an efficient and secure federated learning scheme (named as SEFL) for continuous data sharing in an IoV environment to address the problems related to inefficient data collection, catastrophic forgetting problems due to dynamic data updates and privacy leakage from model training parameters. In SEFL, to enable the automatic collection of training samples, each vehicle is based on the global model to only collect IoV data with lower recognition accuracy, and the output with the highest probability is used as the label for that sample. Since vehicle storage space is limited and new samples can overwrite old ones, the data on vehicles are dynamically changing, making traditional fine-tuning training methods prone to catastrophic forgetting. Thus, a dual-knowledge distillation-based training algorithm is proposed in SEFL to ensure that the model learns the knowledge of each sample, maintaining high accuracy. Besides, to prevent privacy leakage from the model parameters between vehicles and servers, an adaptive differential privacy strategy is proposed to achieve client-level privacy protection. Simultaneously, this strategy minimizes the negative impact of differential privacy noise on the accuracy of the global model. Finally, a security analysis and performance evaluation of SEFL scheme are conducted using the GTSRB dataset and vehicle identification dataset. The analysis and experimental results indicate that the proposed SEFL scheme can provide strong privacy protection and efficient data collection. Furthermore, SEFL outperforms existing federated learning-based algorithms in terms of model accuracy.

Key words　Internet of vehicles; continuous sharing; federated learning; differential privacy; privacy protection

收稿日期：2023-11-01；**修回日期**：2024-05-20

基金项目：国家重点研发计划项目(2022YFB3103500)；国家自然科学基金项目(61932006, 62202071, 62302072)；中国博士后科学基金项目(2022M710518, 2022M710520)；重庆市自然科学基金项目(CSTB2022NSCQ-MSX0358, CSTB2022NSCQ-MSX1217)
The work was supported by the National Key Research and Development Program of China (2022YFB3103500), the National Natural Science Foundation of China (61932006, 62202071, 62302072), the China Postdoctoral Science Foundation (2022M710518, 2022M710520), and the Natural Science Foundation of Chongqing (CSTB2022NSCQ-MSX0358, CSTB2022NSCQ-MSX1217).

通信作者：张迪(dizhang@cqu.edu.cn)

摘 要 车联网与人工智能结合推动了自动驾驶汽车的快速发展.分散于不同车辆中的车联网数据共享并用于训练人工智能模型可实现更高效、更可靠的智能驾驶服务.自动驾驶汽车可通过车载摄像头、传感器等持续采集车辆实时信息、道路图像和视频等车联网数据,并用于优化更新智能交通模型,弥补车联网数据变化导致的模型准确度下降问题.提出面向车联网环境下数据持续共享的高效安全联邦学习方案SEFL,以解决车联网数据采集低效、数据动态更新导致的灾难性遗忘、模型训练参数导致的隐私泄露等问题.在方案SEFL中,车辆基于全局模型,只采集模型识别率较低的车联网数据,并以最大概率对应的输出作为该样本的标签,完成训练样本自动采集.由于车辆存储空间有限,采集的新样本会覆盖旧样本,导致车辆上数据是动态变化的,传统微调训练方式容易引起灾难性遗忘问题.为此,方案中设计了一种基于双重知识蒸馏的训练算法,确保模型学习到每个样本的知识,使模型保持较高的准确度.此外,为了防止车辆与服务器之间传播的模型参数泄露用户隐私,提出了一种自适应的差分隐私策略来实现客户端级的强隐私保护,同时该方案能最大限度地减少差分隐私噪声对全局模型准确度的负面影响.最后,进行了安全性分析并结合交通标志数据集GTSRB和车辆识别数据集对SEFL方案进行了性能评估.实验结果表明所提出的SEFL方案能提供可靠的强隐私保护和高效的采集策略,并且在模型准确度方面要优于现有基于联邦学习的算法.

关键词 车联网;持续共享;联邦学习;差分隐私;隐私保护

中图法分类号 TP309; TP181

车联网(Internet of vehicles, IoV)是一个分布式的汽车网络,它是实现智能交通系统和自动驾驶的关键技术之一[1].通过按照共同制定的标准将车辆连接到互联网,车联网允许车辆之间和车辆与交通系统之间的数据互通.例如,车辆与云端的通信(vehicle-to-cloud, V2C)可实现不同车辆交通数据的共享.基于车辆共享的交通数据,人工智能模型可以汇集来自各个车辆的学习经验,从而提高整个自动驾驶系统的智能水平.车联网与人工智能的协同作用使得自动驾驶汽车能够更好地应对复杂的交通情况,提高了驾驶的安全性和效率,不断推动自动驾驶技术的进步.

智能交通识别模型是自动驾驶汽车实现安全、合规和高效行驶的关键组成部分之一[2-3].各类交通数据可被收集用于训练识别模型,基于该模型的准确识别和解释交通状况,自动驾驶汽车可以更好地融入现实道路交通环境,为驾驶员和行人提供更安全的出行体验.然而,车联网数据,如交通标志数据,存在视觉色差、扭曲破损等变化会导致当前模型的识别准确度下降.为解决该问题,可通过自动驾驶汽车的车载摄像头捕捉道路实时图像和视频等视觉数据,进行持续共享,实现对识别模型的动态更新,提高模型识别准确度.

基于车联网数据持续共享优化智能识别模型,面临着样本采集低效、灾难性遗忘、隐私泄露等问题.具体而言,由于车辆内存空间有限,通常采用“循环存储”的策略进行数据管理,即覆盖最早的旧数据,从而确保存储空间一直被最新的数据所占用.如果自动驾驶汽车无差别地采集道路遇到的所有交通数据,会出现大量交通数据被重复采集,导致有限存储空间内利于提升模型泛化能力的样本占比过小.此外,采集的样本若采用手动标签设置会显著降低采集效率.

若将采集到的车联网数据直接发送给服务器或者第三方机构,由于车联网数据能关联到司机或乘客的出行路线,容易导致个人隐私泄露.为此,常采用联邦学习的方式,在保证样本不离开车辆的前提下协同训练源于不同车辆的车联网数据.然而,在联邦学习(federated learning, FL)训练过程中,循环存储会使得车辆上的样本数据动态变化,基于新数据的模型训练会遗忘原先模型学习到的知识,导致灾难性遗忘(catastrophic forgetting)问题,使得新模型无法学习到本地旧数据对应的知识以及来自全局模型的旧知识.此外,虽然联邦学习能避免直接泄露车辆上的原始数据,但攻击者依旧能通过模型反转攻击、推理攻击等手段从模型参数中推断车辆用户的敏感信息.

为了实时优化智能识别模型,增强模型训练的效率、准确度与安全性,本文结合知识蒸馏、差分隐私等技术,设计了一个面向车联网数据持续共享的安全高效联邦学习架构.该架构能支持车辆动态更新交通识别模型,确保识别模型具备较高的泛化能力,辅助车辆准确辨识各类交通信息,提高自动驾驶

的安全性和可靠性，为推动自动驾驶技术进一步发展提供助力. 本文的主要贡献包括 3 个方面：

1）提出了一个高效安全的联邦学习架构，旨在提高智能交通识别模型训练的准确度和隐私安全性，并基于全局模型设计高效的样本采集方法，避免对无效样本的采集，提高了存储空间利用率与模型学习效率.

2）在本地端设计了一种双重知识蒸馏的训练方法，有效避免车辆数据动态变化引起的灾难性遗忘问题，确保车辆采集到每个样本都能参与到模型的学习中，以提升训练模型的准确度.

3）设计了一种自适应的差分隐私（differential privacy, DP）预算分配算法，对每个车辆的训练模型参数提供了客户级的隐私保护强度，同时能自适应地为每个模型参数分配合理的隐私预算，以减少差分隐私噪声对模型参数可用性造成的负面影响，进而确保全局模型具备更高的识别准确度.

自动采集方法基于已训练好的模型进行采样，提高了采样效率，而基于该采集方法得到的样本又能提升模型训练的收敛速度. 此外，基于知识蒸馏的方式，是解决灾难性遗忘问题最有效的方法之一，本文提出的双重知识蒸馏训练算法在提高模型准确度方面效果显著. 最后，选择基于差分隐私（differential privacy, DP）技术进行隐私保护，有效地可避免过多的通信与计算开销，且本文所设计的自适应差分隐私策略，能减少差分噪声导致的模型准确度下降问题. 其中，自动采样方法和双重知识蒸馏训练算法相互结合，克服了车辆存储有限对共享数据学习造成的负面影响. 自适应差分隐私策略则作用于学习过程中，防止了模型训练参数泄露原始数据信息. 这 3 种方案彼此关联呼应、共同作用，实现了面向车联网数据持续共享的安全高效联邦学习.

1 相关工作

在车联网中，车辆可以通过车载设备采集各类交通与车辆行为数据，这些数据可以用于改善驾驶安全性、自动驾驶等方面的应用. 然而，将这些数据直接传输到中心服务器进行训练可能涉及隐私问题和高成本的数据传输. 联邦学习作为一种隐私安全的分布式机器学习技术，在 2016 年由 Konečný等人[4]提出，其可以确保车辆的原始数据不离开车辆，只有模型参数在各个车辆之间共享，从而保护了用户隐私，并减少了数据传输的需求. McMahan 等人[5]提出了一种通信高效的联邦学习方法 FedAvg，它进一步降低了模型参数传输的通信成本. 随后，文献 [6–8] 中相继提出了一些可用于训练非独立同分布数据的高性能联邦学习方案.

目前，联邦学习已被广泛用于智能交通系统、车联网等领域[9-10]. Xie 等人[11]结合脉冲神经网络设计了一种用于交通标志识别的高效联邦学习方案. Stergiou 等人[12]提出一种增强自动驾驶汽车图像识别的联邦学习方法. 此外，一些面向车联网的高性能联邦学习方案也相继被提出. Liang 等人[13]提出了一种具有动态聚合的半同步联邦学习协议来提高机器学习模型在车联网中的学习效率. Zhou 等人[14]设计了一种基于异构模型聚合的 6G 车联网双层联邦学习，旨在提高模型的收敛速度和准确度. 文献 [15] 设计了一个面向车联网的鲁棒分层联邦学习，可防止来自恶意客户端的中毒攻击. 然而这些方案并未考虑车辆数据动态变化导致的灾难性遗忘问题，以及模型参数导致的隐私泄露问题.

灾难性遗忘问题是指在神经网络或机器学习模型中，当学习新任务或新数据时，导致已经学习的旧任务或数据性能显著下降的现象. 为了解决灾难性遗忘问题，Li 等人[16]提出了无遗忘学习（learning without forgetting, LwF），使用知识蒸馏生成旧任务的数据和标签的方式传递旧任务的知识给新模型. Kirkpatrick 等人[17]提出弹性权重共享（elastic weight consolidation, EWC）方法，通过对模型参数引入正则化项，来保护已学习任务的参数免受新任务学习的干扰. Rebuffi 等人[18]设计了一种基于数据重放的增量分类器和表示学习（incremental classifier and representation learning, iCaRL）. 该方案使用一个存储缓冲区，用于存储旧任务的样本，并将旧任务的样本与新任务的样本一起用于训练，以此来缓解灾难性遗忘问题. 然而文献 [16–18] 所述方案无法直接用于解决车辆数据动态变化导致本地模型以及全局模型同时存在的灾难性遗忘问题.

联邦学习架构中，本地客户端传递的模型参数可能遭受来自诚实但好奇的服务器的攻击，如模型反转攻击[19]、推理攻击[20-21]等，进而导致隐私信息泄露[22]. 为增强联邦学习的隐私保护能力、防止模型参数泄露用户隐私信息，许多隐私保护的联邦学习方案被相继提出. 其中，基于安全多方计算（secure multi-party computation, SMC）[23]、同态加密（homomorphic encryption, HE）[24]、区块链（blockchain）[25]和差分隐私[26–28]的策略是最流行和高效的解决方案. 然而，基

于SMC的方案通信代价较高;基于HE和区块链的方案计算开销过大;基于差分隐私的方案在通信和计算开销上具有明显优势,但添加的差分隐私噪声会扰动模型参数真实值,并且噪声会随着模型训练的进行叠加,影响模型收敛或者模型训练的准确度.为此,设计一种兼顾强隐私保护和高准确度联邦学习架构是亟待解决的难题.

表1总结了上述相关工作在隐私保护、学习效率、模型准确度方面的优势与不足.通过对比可知,设计兼顾强隐私保护、高效学习与高准确度的联邦学习架构是当前亟待解决的难题.本文设计的SEFL方案则致力于解决该难题.

Table 1 Comparison of Related Work

表1 相关工作对比

性能	文献				SEFL(本文)
	[11−15]	[16−18]	[23−25]	[26−28]	
强隐私保护	×	×	√	√	√
高效学习	√	√	×	√	√
高准确度	×	√	√	×	√

注:×表示不支持,√表示支持.

2 基础知识

在本节中,本文主要介绍联邦学习、差分隐私和蒸馏知识相关定义与基础知识.

2.1 联邦学习

联邦学习是一种分布式机器学习,多个分散的客户端在中央服务器的协调下共同训练一个模型[29].在联邦学习中,原始数据集保留在客户端本地设备上,客户端只需要上传模型参数(梯度或者模型权重)给中央服务器.联邦学习过程一般可以分解为多个训练轮次.在每轮训练中,服务器首先将初始模型或全局模型更新发送给客户端;然后每个客户端使用自己的数据在本地训练模型;最后将经过训练的模型更新上传到服务器进行聚合,得到新的全局模型.多次进行上述操作,直至模型收敛,则训练结束.

假设有n个客户,样本索引集为$\mathcal{U}=\{u_1,u_2,\cdots,u_n\}$,其中集合$u_i$存储客户端$i$的样本索引.假设$(x_k,y_k)$是客户端$i$的一个样本,则客户端$i$拥有的样本可以表示为$(X_i,Y_i)=(x_k,y_k)_{k\in u_i}$,客户端$i$的样本数表示为$n_i=|u_i|$.单一样本对应的损失函数通常定义为$f_k(\omega)=l(\omega;x_k,y_k)$.在联邦学习中,优化模型可表示为

$$\min_{\omega} f(\omega)=\sum_{i=1}^{N}\frac{n_i}{n}F_i(\omega),$$

其中$n=|\mathcal{U}|$代表训练中的总样本数,且

$$F_i(\omega)=\frac{1}{n_i}\sum_{k\in u_i}f_k(\omega).$$

联邦平均算法(federated averaging algorithm, FedAvg)旨在利用客户端进行更多计算来减少模型训练所需的通信轮数,提高训练效率.在FedAvg中,客户端上传给中心服务器的是模型权重而不是模型梯度.服务器接收到本地上传的模型权重后进行加权聚合,即

$$\omega_{t+1}\leftarrow\sum_{i=1}^{N}\frac{n_i}{n}\omega_{t+1}^{i},$$

获得新一轮的全局模型权重ω_{t+1}.经过多轮上述操作,直至模型收敛.目前,FedAvg作为一种高效、鲁棒的技术已被广泛用于分布式学习.本文设计的联邦学习架构采用的是FedAvg的更新方式.

2.2 差分隐私

Dwork等人[30]在2014年提出了差分隐私保护模型,差分隐私可以抵抗任意背景的知识攻击,并可通过隐私预算ε实现隐私保护水平的定量分析.差分隐私保护模型最初应用在数据统计查询领域,目的在于保护查询时用户的隐私信息,之后经过扩展,广泛应用到数据发布领域和数据挖掘领域,而且,交互式的统计查询和非交互式的数据发布均可以采用差分隐私模型进行隐私信息保护.

定义1. 差分隐私.设D为一组数据集,一个具有定义域D和值域$\mathbb{R}$的机制$M:D\rightarrow\mathbb{R}$满足差分隐私,则对于任意2个相邻的数据集$d\in D$和$d'\in D$,以及任何输出$O\subseteq\mathbb{R}$,符合以下条件:

$$Pr[M(d)\in O]\leqslant\exp(\varepsilon)Pr[M(d')\in O]+\delta,$$

其中$\varepsilon>0$为隐私预算,用于确定隐私保护程度,即ε越小,隐私保护越强.

在联邦学习中,d和d'可以被定义为2种不同类型的邻接数据集,即样本级邻接数据集[27,31-32]和客户级邻接数据集[33-35].

1)样本级邻接数据集.假设d和d'是2个包含训练样本的数据集.如果d和d'仅在1个样本上不同,它们被称为样本级邻接数据集.

2)客户级邻接数据集.假设d和d'是2个训练样本的数据集,每一个训练样本都归属唯一客户端.如果d和d'仅相差1个客户级的样本,那么d和d'就是客户级相邻.

例如,假设$d=\{d_1,d_2,d_3,d_4\}$,$d'=\{d_1,d_2,d_4\}$是2个

数据集，其中这 2 个数据集只在d_3上不同. 如果d_3代表 1 个样本，那么d和d'是样本级相邻的；如果d_3代表客户 3 的样本，那么d和d'是客户级相邻的.

查询函数$Q: D \to \mathbb{R}^m$的隐私保护近似的标准范例为添加与查询函数Q的敏感度S_Q成比例的噪声，其中，敏感度$S_Q = \max\limits_{d,d' \in D} \|Q(d) - Q(d')\|_2$. 本文采用高斯噪声机制$M$作为提供隐私保护的近似方法，可表示为$M(d) = Q(d) + \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right)$，其中$\mathcal{N}$是均值为 0、标准差为$\sigma$的正态分布，$\boldsymbol{I}$是单位矩阵，$\sigma^2\boldsymbol{I}$表示协方差矩阵. 此外，差分隐私还具有顺序组合这一重要性质.

定理 1. 顺序组合[36]. 令 m 个机制$M_1, M_2, \cdots, M_m$分别满足(ε_i, δ)-DP. 那么，由$M_1(d), M_2(d), \cdots, M_m(d)$形成的机制满足

$$\left(\sum_i^m \varepsilon_i, m\delta\right)\text{-DP}.$$

2.3 知识蒸馏

知识蒸馏（knowledge distillation）[16,37] 旨在将教师模型的知识传递给学生模型，是一种可用于解决迁移学习和遗忘问题的方法. 蒸馏网络的优化模型可表示为

$$\min_{\omega_{\text{stu}}} F_{\text{KD}}(D, \omega_{\text{tech}}, \omega_{\text{stu}}, T') = \mathcal{L}_{\text{KD}} + \mathcal{L}_{EC} = EC\left(\bar{Y}, \hat{Y}\right) + EC(Y, O),$$

其中，$\hat{Y}$，$\bar{Y}$分别表示教师模型预测输出的软标签和学生模型预测输出的软标签，Y表示学生模型的输出，O表示硬标签，是训练样本的真实输出. 损失函数$F_{\text{KD}}(\cdot)$由蒸馏损失$\mathcal{L}_{\text{KD}}$和交叉熵（cross-entropy）损失$\mathcal{L}_{EC}$这 2 部分组成. 蒸馏损失是指学生模型的预测与教师模型的预测之间的差异；交叉熵损失为学生模型预测与硬标签之间的差异. 蒸馏损失可转换为软标签$\bar{Y}$与软标签$\hat{Y}$之间的交叉熵损失. 软标签通常是指类别的概率分布，在神经网络中多为基于温度参数T'处理后的 Softmax 输出. 假设学生模型的 Softmax 输出为$Y = \{y_1, y_2, \cdots, y_l\}$，$\bar{Y} = \{\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_l\}$，则有

$$\bar{y}_i = \frac{(y_1)^{\frac{1}{T'}}}{\sum\limits_{j=1}^{l} (y_j)^{\frac{1}{T'}}}.$$

交叉熵是一种用于度量 2 个概率分布之间差异的度量方式，常用于衡量模型的预测分布与真实分布之间的差异. 假设存在 2 个概率分布$\bar{Y} = \{\bar{y}_1, \bar{y}_2, \cdots, \bar{y}_l\}$和$\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_l\}$，则交叉熵计算表示为

$$EC\left(\bar{Y}, \hat{Y}\right) = -\bar{Y}\log\hat{Y} = -\sum_{j=1}^{l} \bar{y}_j \log \hat{y}_j.$$

3 安全高效的联邦学习架构

本节介绍了用于交通信息识别动态更新的联邦学习系统模型、基于全局模型的自动采样方法、基于双重知识蒸馏的本地训练算法以及自适应差分隐私策略 4 个部分.

3.1 系统模型

本节介绍了基于差分隐私技术的联邦学习模型总体架构和 1 轮联邦训练的实现过程.

为提升交通信息识别模型的泛化能力，保证训练数据的隐私安全，本文提出了安全高效的联邦学习架构 SEFL. 如图 1 所示，该框架主要由 1 个中心服务器端与多个客户端组成. 其中，服务器端假设为诚实但好奇的，即诚实地遵守联邦学习聚合协议，但对客户端的敏感信息好奇，通过各种攻击来窃取用户隐私信息. 该服务器端在现实场景中可能是第三方机构或企业，主要目标是借助客户端数据来完善交通信息识别模型，以提升其产品（如自动驾驶汽车、识别系统等）的性能；参与训练的客户端主要为行驶在道路上的各类智能汽车，本文中假设客户端为诚实节点，主要负责对交通信息样本进行有选择性地采集、完成识别模型的训练与更新、对更新后的模型参数进行差分隐私处理.

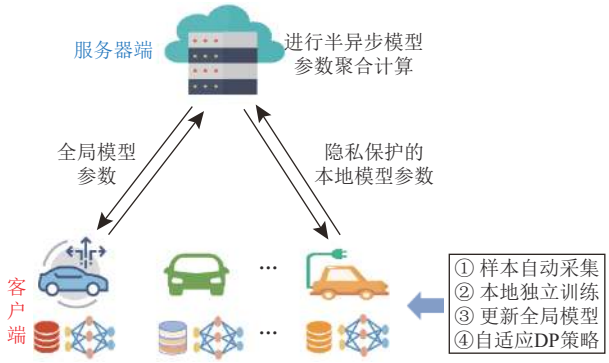


Fig. 1 Architecture of efficient and secure federated learning

图 1 高效安全的联邦学习架构

在 SEFL 的训练过程中，服务器端进行半异步模型参数聚合计算，即当服务器端接收到预设数量的本地模型参数时才进行聚合操作. 该聚合方式能避免同步聚合导致的低效问题以及异步聚合引起的参数过时问题. 在进行联邦训练之前，假设服务器端基于公开的交通数据样本训练得到一个初始全局模型. 算法 1 给出了方案 SEFL 完成 1 轮联邦学习训练过程的具体实现过程，算法 1 主要包含 6 个步骤：

1)本地客户端下载当前全局模型,并基于全局模型进行样本采集与保存;

2)基于采集的数据对本地模型进行单独训练,直至本地模型收敛;

3)基于本地模型与当前客户端保存的样本数据,对全局模型进行训练;

4)根据自适应差分隐私策略,对更新后的全局模型参数进行隐私处理;

5)将隐私保护后的全局模型参数上传至服务器端;

6)服务器端在接收到约定数量的客户端模型参数后进行聚合计算,得到新的全局模型.

重复上述6个步骤,以实现对交通信息识别模型的动态更新,提升识别模型的泛化能力,确保能更准确地识别各类交通,优化自动驾驶服务.

算法 1. SEFL 方案 1 轮全局更新的主要步骤.

输入:客户端 i 当前数据为 D_t^i,前一轮全局模型参数 ω_{T-1},全局模型参数 ω_T, $\bar{\omega}_T^i$;

输出:全局模型参数 ω_{T+1};

初始化:客户端数量为 n,半异步聚合阈值 m.

① 基于上一轮全局模型 ω_{T-1} 收集新的交通数据样本,表示为 D_{t+1}^i;(见 3.2 节)

② 基于本地模型参数 $\tilde{\omega}_t^i$ 和 D_{t+1}^i 完成本地独立训练,得到 $\tilde{\omega}_{t+1}^i$;(算法 2 的行①②)

③ 基于 $\tilde{\omega}_{t+1}^i$, D_{cur}^i, ω_T 进行训练,更新 ω_T,得到模型参数 ω_{T+1}^i;(见算法 2 的行③~⑧)

④ $\bar{\omega}_{T+1}^i = ADP(\omega_T, \bar{\omega}_T^i, \omega_{T+1}^i)$,即自适应差分隐私处理;(见算法 3)

⑤ 客户端将 $\bar{\omega}_{T+1}^i$ 上传至服务器端,当服务器端接收到 m 个本地模型参数,则进行聚合计算,得到新的全局模型参数 ω_{T+1}.

3.2 基于全局模型的自动采集方法

本地车辆下载服务器的全局模型,并通过全局模型对路途中的交通目标进行取照识别,若模型的 Softmax 输出的最大概率值小于 a 或者大于 b 时,则不进行采集,其中阈值下限 a 是为了限制采集中标签自动设置的错误率,减少错误样本对模型学习的负面影响,阈值上限 b 则是避免再次耗费资源对已掌握较好的知识进行重复学习.若模型的 Softmax 输出的最大概率值在预设的阈值范围内,即 $[a, b]$,则该图片被采集,其中图片为样本的特征,最大概率对应的输出为该样本的标签,作为最新样本保存在车载内存中,内存中最早被保存的样本被删除.在每次样本采集过程中,本地车辆都是基于当前下载的最新全局模型进行操作,可持续进行样本采集,如图 2 所示.假设当前采集的样本表示为 D_t^i,当新采集的数据完全覆盖 D_t^i 时,新样本集表示为 D_{t+1}^i.

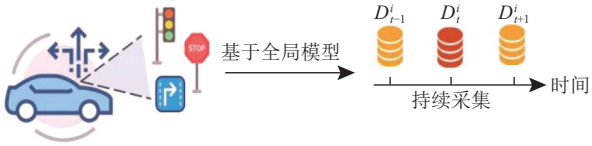


Fig. 2 Vehicles automatically collecting based on the global model

图 2 车辆基于全局模型自动采集

通过该方式采集到的样本能提升模型的泛化能力,使得模型训练后能更准确识别出不同类型的交通对象.同时避免了对已学习较好知识对应的样本进行重复采集.此外,该样本基于全局模型进行自动采集,显著减少了对样本手工设置标签的工作量.随着全局模型识别率的逐渐提升,能用于优化模型的样本也会相应减少,进一步缩小了样本采集工作量.

3.3 基于双重知识蒸馏的本地训练算法

为保证本地车辆收集到的样本都参与到模型学习中,本文设计了基于双重知识蒸馏的本地训练算法,如算法 1 所示.在算法 1 中主要包括本地独立训练和全局模型更新 2 个部分,在客户端完成.其中本地独立训练将收集过程中被覆盖的样本对应的知识保存在本地模型中;全局模型更新则将本地模型学习到的知识和当前收集到的数据融合到全局模型中.

3.3.1 本地独立训练

本文设计基于知识蒸馏的方式来处理数据动态变化导致的灾难性遗忘问题.为了防止先前数据未被遗忘,将学习到的知识存储在模型 $\tilde{\omega}_t^i$ 中,并在后续的训练当中,该 $\tilde{\omega}_t^i$ 作为教师模型,将学习到的知识传递给学生模型 $\tilde{\omega}_{t+1}^i$,如图 3 所示.通过该方式来确保以往样本与当前样本都参与了学生模型 $\tilde{\omega}_{t+1}^i$ 的学习.具体实现如下所示:

1)初始化.初始本地模型为本地车辆首次从服务器下载到的全局模型.当车载内存首次被采集到的

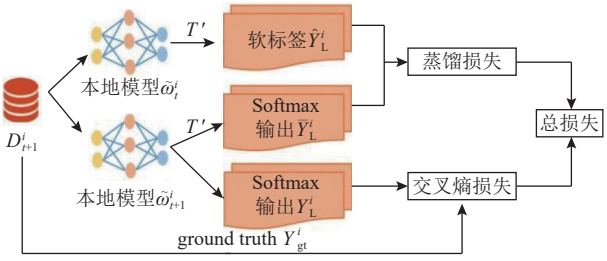


Fig. 3 Local independent training based on knowledge distillation

图 3 基于知识蒸馏的本地独立训练

样本占满时，则在初始本地模型上进行微调操作，直至模型收敛，得到新的模型M_{L}，模型参数表示为$\tilde{\omega}_1^i$.

2）本地模型动态更新. 如算法 2 的行①所述，当新收集到的本地样本完全覆盖了上一轮本地模型训练时的样本，并且此时未收到全局模型参数时，则对本地模型进行更新训练. 更新过程如图 3 所示. 首先，样本数据D_{t+1}^i基于本地模型$\tilde{\omega}_t^i$生成软标签$\hat{Y}_{\mathrm{L}}^i$；其次，本地模型$\tilde{\omega}_{t+1}^i$基于数据D_{t+1}^i生成蒸馏后的 Softmax 输出$\bar{Y}_{\mathrm{L}}^i$以及常规 Softmax 输出Y_{L}^i；模型训练的总损失函数为蒸馏损失与交叉熵损失的总和，则优化模型可表示为

$$\min_{\tilde{\omega}_{t+1}^i} F_{\mathrm{L}}\left(D_{t+1}^i,\tilde{\omega}_t^i,\tilde{\omega}_{t+1}^i,T'\right)=\mathcal{L}_{\mathrm{KD}}(D_{t+1}^i,\tilde{\omega}_t^i,\tilde{\omega}_{t+1}^i,T')+$$
$$\mathcal{L}_{EC}(D_{t+1}^i,\tilde{\omega}_{t+1}^i)=EC\left(\hat{Y}_{\mathrm{L}}^i,\bar{Y}_{\mathrm{L}}^i\right)+EC\left(Y_{\mathrm{gt}}^i,Y_{\mathrm{L}}^i\right)=$$
$$-\hat{Y}_{\mathrm{L}}^i\log\bar{Y}_{\mathrm{L}}^i-Y_{\mathrm{gt}}^i\log Y_{\mathrm{L}}^i,$$

其中Y_{gt}^i为样本D_{t+1}^i对应的真实标签 ground truth，$\mathcal{L}_{\mathrm{KD}}$为蒸馏损失，$\mathcal{L}_{EC}$为交叉熵损失，$EC$函数表示交叉熵计算. 当模型收敛，则本地模型更新结束，如算法 2 的行②所示.

3.3.2 全局模型更新（本地端）

如算法 2 的行③所述，当本地端接收到全局模型参数ω_T时，则基于当前的本地模型$\tilde{\omega}_{t+1}^i$以及样本D_{cur}^i进行模型更新. 具体过程如图 4 所示，结合知识蒸馏的方式，在微调全局模型的同时，保存本地模型学习到的知识. 优化模型可表示为

$$\min_{\omega_T} F_{\mathrm{G}}\left(D_{\mathrm{cur}}^i,\omega_T,\tilde{\omega}_{t+1}^i,T'\right)=\mathcal{L}_{\mathrm{KD}}\left(D_{\mathrm{cur}}^i,\tilde{\omega}_t^i,\tilde{\omega}_{t+1}^i,T'+\right.$$
$$\left.(D_{\mathrm{cur}}^i,\tilde{\omega}_{t+1}^i)\right)=EC\left(\hat{Y}_{\mathrm{G}}^i,\bar{Y}_{\mathrm{G}}^i\right)+EC\left(Y_{\mathrm{gt}}^i,Y_{\mathrm{G}}^i\right)=$$
$$-\hat{Y}_{\mathrm{G}}^i\log\bar{Y}_{\mathrm{G}}^i-Y_{\mathrm{gt}}^i\log Y_{\mathrm{G}}^i.$$

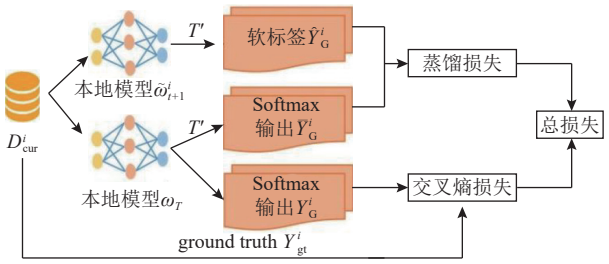


Fig. 4 Global model updates on vehicle clients based on knowledge distillation

图 4 在车辆客户端基于知识蒸馏进行全局模型更新

全局模型的更新参考 FedAvg 方案，在本地端采用小批量梯度下降的更新方式，见算法 2 的行④~⑥. 为了便于后续的隐私保护处理，对训练得到的梯度进行了裁剪操作，如算法 2 的行⑦所示. 最后，计算得到更新后的ω_T，表示为ω_{T+1}^i，如算法 2 的行⑧所示.

算法 2. 双重知识蒸馏的本地训练算法 *LT*.

输入：D_{t+1}^i，全局模型ω_T，M_{L}的模型参数$\tilde{\omega}_t^i$；

输出：全局模型参数ω_{T+1}^i；

初始化：客户端存储样本上限为 S，梯度裁剪门限值为 C.

① 当客户端 i 未接收到ω_T；

② $\tilde{\omega}_{t+1}^i=\tilde{\omega}_t^i-\eta_{\mathrm{L}}\nabla F_{\mathrm{L}}\left(D_{t+1}^i|\tilde{\omega}_t^i\right)$，直至$M_{\mathrm{L}}$收敛，$t=t+1$；

③ 当客户端下载得到ω_T时，当前本地端保存的样本表示为D_{cur}^i；

④ $\mathcal{B}\leftarrow D_{\mathrm{cur}}^i$划分为 b 个 batches；

⑤ $w_l^i=\omega_T, l=E(T-1)b$ and $k=0$；

⑥ for 本地迭代次数从 1 到 E do

  for batch $B\in\mathcal{B}$ do

    $g_{l+k}^i=\nabla F_{\mathrm{G}}\left(D_T^i|\tilde{\omega}_{t+1}^i,w_{l+k}^i\right)$；

    $k=k+1$；

    $w_{l+k}^i=w_l^i-\eta_T g_{l+k-1}^i$；

  end for

end for

⑦ $g_T^i=\dfrac{g_{l+k-1}^i}{\max\left(1,\dfrac{\|g_{l+k-1}^i\|_2}{C}\right)}$；

⑧ return $\omega_{T+1}^i=\omega_T-\eta_T g_T^i$.

3.4 自适应差分隐私策略

自适应差分隐私策略是确保客户端数据得到强隐私保护的同时尽可能地减少差分噪声对全局模型识别准确度的影响. 为实现这一目标，本文设计了客户级隐私保护措施，并开发了一种模型参数置换规则，以实现对模型参数ω_{T+1}^i的自适应差分噪声的添加. 客户端 i 的自适应差分隐私策略设计如算法 3 所示. 在算法 3 中，行①和行②实现了客户级差分隐私设置，计算当前全局训练轮数对应的敏感度和拟添加的噪声尺度. 行③是设计的置换规则，以选择接近模型参数ω_{T+1}^i且不泄漏客户隐私的模型参数作为最终上传给服务器的值. 下面将详细介绍客户级隐私保护和置换规则的实现过程.

算法 3. 自适应差分隐私处理策略 ADP.

输入：ω_{T+1}^i，ω_T，$\bar{\omega}_T^i$，学习率η_T；

输出：全局模型参数$\bar{\omega}_{T+1}^i$.

① 计算在第 T 轮训练的全局敏感度：$S_T=2\eta_T C$；

② 计算噪声尺度$\sigma_T=\dfrac{S_T}{\varepsilon}$；

③ if $\min(\|\omega_{T+1}^i-\bar{\omega}_T^i\|_2,\|\omega_{T+1}^i-\omega_T\|_2)\leqslant\alpha\mathcal{N}(0,\boldsymbol{I}\sigma_T^2)$

  if $\|\omega_{T+1}^i-\bar{\omega}_T^i\|_2\leqslant\|\omega_{T+1}^i-\omega_T\|_2$

    $\bar{\omega}_{T+1}^i=\bar{\omega}_T^i$；

  else

$\tilde{\omega}_{T+1}^{i}=\omega_T$;

end if

else

$\tilde{\omega}_{T+1}^{i}=\omega_{T+1}^{i}+\mathcal{N}(0,\boldsymbol{I}\sigma_T^2)$;

end if

④ return $\tilde{\omega}_{T+1}^{i}$.

3.4.1 客户级隐私保护

为了保证客户级隐私保护，本文通过使用客户相邻数据集的概念，将差分隐私应用于模型训练. 在本地客户端，1 个全局训练轮次内的模型更新训练过程被定义为客户级别差分隐私的查询函数 Q:

$$Q\left(D^i|\omega_T\right)=\omega_T-\eta_T\nabla F_{\mathrm{G}}\left(D^i|\omega_T,\tilde{\omega}_t^i\right).$$

该查询函数对应的隐私敏感度计算表示为

$$S_Q(\cdot|\omega_T)=\max_{D^i,D^j\in D_T}\|Q(D^i|\omega_T,\tilde{\omega}_t^i),Q(D^j|\omega_T,\tilde{\omega}_t^i)\|_2,$$

其中，D_T为第 T 轮全局模型训练的客户端当前样本的集合，D^i和D^j表示不同客户端的样本. 为了量化敏感度，本文对训练得到的梯度进行了裁剪操作，梯度裁剪门限值为 C，如算法 2 的行⑦所示. 根据上述设置，可以计算出在全局训练第 T 轮时的敏感度 $S_Q^T=2\eta_T C$.

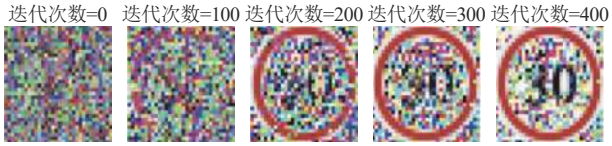
3.4.2 模型参数替换规则

在添加差分噪声的步骤中，本文采用一种自适应策略[38]，主要思想是本地客户端自适应地选择上传差分隐私处理后的模型参数、前一轮本地上传的模型参数或者当前的全局模型参数这三者之一至服务器. 由于，差分隐私处理后的模型无法恢复出原始信息，并且前一轮本地上传的模型参数和当前的全局模型参数都是已知隐私安全的. 因此，本文设计的自适应策略保证了上传的模型参数都不会引起隐私泄露. 选择最利于模型训练的模型参数是所设计的替换规则需要达到的目标.

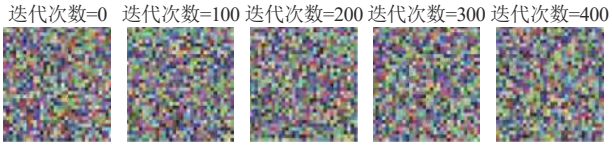
为实现该目标，本文将上述 3 类模型参数中最接近当前本地训练得到的模型参数的模型参数上传给服务器，以保证全局模型的准确度，具体实现过程如算法 3 所示. 以模型参数ω_{T+1}^i为例，若ω_{T+1}^i与$\tilde{\omega}_T^i$或ω_T之间的任一差距小于ω_{T+1}^i添加差分噪声导致的偏差 $\alpha\times\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)$，则最终上传给服务器的模型参数为$\omega_{T+1}^i$和$\tilde{\omega}_T^i$二者之间与$\omega_{T+1}^i$距离更近的一个，否则，上传差分隐私处理后的模型参数$\omega_{T+1}^i+\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)$. 替换规则具体过程见算法 3 的行③. 此外，由于差分隐私噪声与模型参数之间的分布不一致，在差距对比过程中，设置了一个超参数α以平衡不同分布之间的差异.

4 安全性分析

在联邦学习中，客户端会传递模型参数给服务器或者第三方机构进行聚合操作. 若服务器端是诚实且好奇的，则其可能通过反转攻击等方式来挖掘出原始数据信息. 为验证本文所提差分隐私保护策略的有效性，分别对无隐私保护处理的模型参数和本文隐私策略处理的模型参数进行反转攻击. 实验结果如图 5 所示.



(a) 无隐私保护



(b) 差分隐私保护

Fig. 5 Model inversion attack

图 5 模型反转攻击

图 5(a)展示了无隐私保护处理的模型参数可通过反转攻击恢复样本原始信息. 图 5(b)展示了本文隐私策略处理的模型参数可以有效抵御反转攻击，不透露原始样本信息. 因此，本文的 SEFL 方案为客户端数据提供了隐私保护. 此外，通过理论分析可得到本文所设计的隐私保护策略能实现客户级隐私保护，具体分析如下所示：

假设$Q:D\to\mathbb{R}^m$表示一个 m 维查询函数，且其敏感度为$S_Q=\max_{d,d'\in D}\|Q(d)-Q(d')\|_2$. 高斯噪声机制 M 向每个客户端的输出(即模型权重)都添加噪声尺度为 $\mathcal{N}\left(0,\boldsymbol{I}\sigma^2\right)$的噪声，满足$Pr[M(d)\in O]\leqslant\mathrm{e}^{\varepsilon}Pr[M(d')\in O]+\delta$.

定理 2. 假设隐私预算 ε 为任意值，当$c^2>2\ln\left(\frac{1.25}{\delta}\right)$且$\sigma\geqslant\frac{c\times S_Q}{\varepsilon}$，则本文所提出的 SEFL 满足客户级$\left(\tilde{T}\varepsilon,\tilde{T}\delta\right)$-DP. 其中，条件$c^2>2\ln\left(\frac{1.25}{\delta}\right)$为常数 c 的上限约束，$S_Q=2\eta_T C$，$\tilde{T}\leqslant qT$，T 是总的全局训练轮数，q 是每个客户端每轮参与训练的概率，$\tilde{T}$是分配了隐私预算的轮数.

证明. 在本文的 SEFL 方案中，每个用户的客户端训练过程为查询函数 Q，则在第 T 轮全局模型训练时，客户端 i 的查询函数表示为 $Q\left(D^i|\omega_{T-1},\tilde{\omega}_t^i\right)$，查询

函数的输出为客户端基于本地数据训练得到模型权重ω_T^i. 为此, 高斯噪声机制$M_T(D^i)$将返回$Q\left(D^i|\omega_{T-1},\tilde{\omega}_t^i\right)+\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)$.

因此, 查询函数 Q 的敏感度S_Q可表示为

$$
\begin{aligned}
&S_Q(\cdot|\omega_{T-1})=\\
&\max_{D^i,D^j\in D}\left\|Q\left(D^i|\omega_{T-1},\tilde{\omega}_t^i-Q\left(D^j|\omega_{T-1},\tilde{\omega}_t^j\right)\right)\right\|_2=\\
&\max_{\omega_T^i,\omega_T^j\in\omega_T^{\text{all}}}\left\|\omega_T^i-\omega_T^j\right\|_2=\\
&\max_{g_T^i,g_T^j\in g_T^{\text{all}}}\left\|(\omega_{T-1}-\eta_T g_T^i)-(\omega_{T-1}-\eta_T g_T^j)\right\|_2=\\
&\max_{g_T^i,g_T^j\in g_T^{\text{all}}}\left\|\eta_T g_T^i-\eta_T g_T^j\right\|_2.
\end{aligned}
$$

由于g_T^i是经过梯度裁剪之后得到的梯度, 则有$\|g_T^i\|_2\leqslant C$, 其中 C 为梯度裁剪门限值. 因此, 可计算得到$S_Q=2\eta_T C$. 噪声尺度$\sigma_T=\dfrac{c\times S_Q}{\varepsilon}$, 则隐私损失可以计算为

$$
\begin{aligned}
&\frac{Pr\left(M_T\left(D^i\right)=o\right)}{Pr\left(M_T\left(D^j\right)=o\right)}=\frac{Pr\left(Q\left(D^i|\omega_{T-1},\tilde{\omega}_t^j\right)+\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)=o\right)}{Pr\left(Q\left(D^i|\omega_{T-1},\tilde{\omega}_t^j\right)+\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)=o\right)}=\\
&\frac{Pr\left(\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)=Q\left(D^i|\omega_{T-1},\tilde{\omega}_t^i\right)-o\right)}{Pr\left(\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)=Q\left(D^j|\omega_{T-1},\tilde{\omega}_t^j\right)-o\right)}=\\
&\frac{Pr\left(\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)=a_i\right)}{Pr\left(\mathcal{N}\left(0,\boldsymbol{I}\sigma_T^2\right)=a_i+S_Q\left(\cdot|\omega_{T-1}\right)\right)}=\\
&\left|\exp\left((-1/2\sigma_T^2)(\|a_i\|_2^2-\|a_i+S_Q(\cdot|\omega_{T-1})\|_2^2)\right)\right|,
\end{aligned}
$$

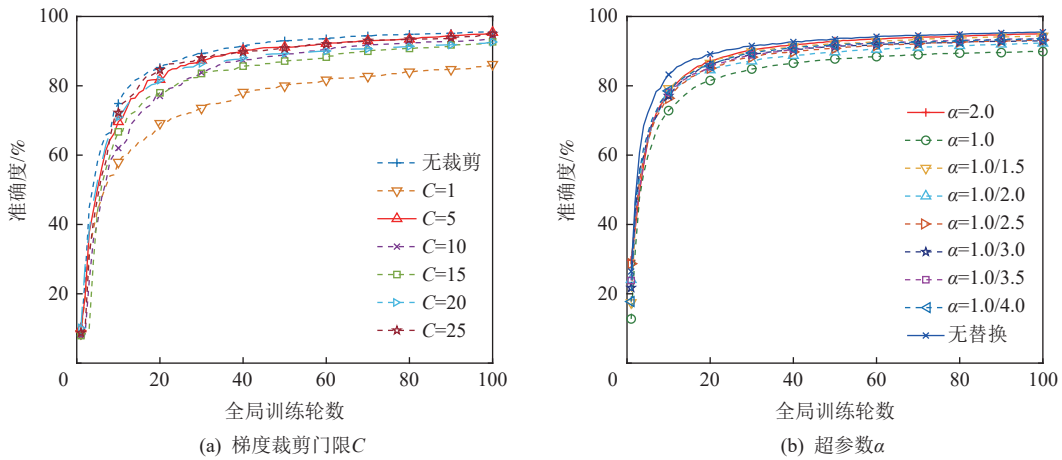
其中$o\in O$表示一个模型输出, a_i表示$Q\left(D^j|\omega_{T-1},\tilde{\omega}_t^j\right)$和 o 的差异. 因此, 隐私损失可重写为$\ln\dfrac{Pr\left(M_T\left(D^i\right)=o\right)}{Pr\left(M_T\left(D^j\right)=o\right)}=\left|(-1/2\sigma_T^2)(\|a_i\|_2^2-\|a_i+S_Q(\cdot|\omega_{T-1})\|_2^2)\right|$. 根据参考文献 [30] 中的定理 A.1.9 可得, 当$\sigma\geqslant\dfrac{c\times S_Q}{\varepsilon}$时, 机制 M_T满足 (ε,δ)-DP, 其中$S_Q=2\eta_T C$且$c^2>2\ln\left(\dfrac{1.25}{\delta}\right)$.

假设高斯噪声机制为$M=\{M_1,M_2,\cdots,M_T,\cdots\}$, 表示整个全局训练过程, M_T为第 T 轮的全局训练. 每轮的查询函数 Q 只与之前的全局模型权重ω_{T-1}和其本地模型的$\tilde{\omega}_t^i$相关. 由于, 全局模型权重ω_{T-1}在每轮都是不同的, 因此, $M_1,M_2,\cdots,M_T,\cdots$是相互独立的. 其中, 全局模型权重$\omega_{T-1}$为公开的. 若假设攻击者已知$\tilde{\omega}_t^i$信息, 每轮分配相同的隐私预设 ε 和δ, 且客户端已知分配隐私预算的训练轮数为$\tilde{T}$, 则根据定理 1 可得, 高斯噪声机制 M 满足$(\tilde{T}\varepsilon,\tilde{T}\delta)$-DP, 定理 2 成立. 在本文所提出的 SEFL 方案中, 由于本地模型$\tilde{\omega}_t^i$是私密的. 为此, SEFL 实际上能提供比$(\tilde{T}\varepsilon,\tilde{T}\delta)$-DP 更高一点的隐私保护. 证毕.

5 实验评估

本文主要在交通标志识别数据集(German traffic sign recognition benchmark, GTSRB)上对 SEFL 方案进行了性能评估实验. GTSRB 数据库拥有 50 000 多张图片, 覆盖 43 种不同类别的交通标志图片, 同种类别下包含形状不同、色彩差异大、图标模糊的交通标志. 在实验中, 设置每张图片大小为 30×30×3, 每个客户端随机分配相同数量的训练样本, 采用卷积神经网络(convolutional neural network, CNN)训练交通标志识别模型. 此外, 还基于车辆识别数据集对主要算法进行了通用性验证, 该数据集包含 1 600 张车辆图片和 10 种不同车辆类别. 实验主要基于联邦学习架构进行仿真模拟, 客户端采用小批量随机梯度下降的训练方法, 批量大小设置为 50, 本地迭代次数为 10. 实验使用 Python 3.10 编程语言, 实验环境配置为 RTX 3070Ti、32GB 内存、Windows 11 操作系统.

首先对 SEFL 方案中涉及到的 2 个重要超参数



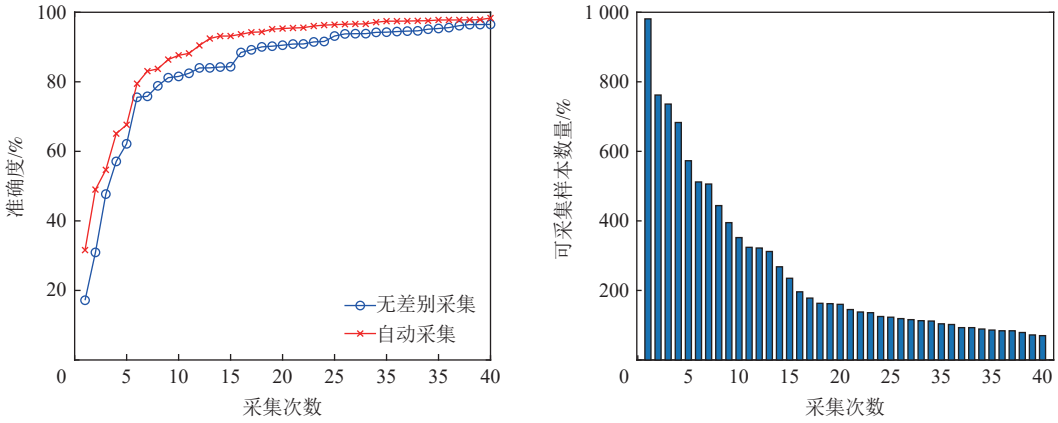
(a) 梯度裁剪门限C (b) 超参数α

Fig. 6 Model hyperparameter configuration

图 6 模型超参数设置

(梯度裁剪门限值 C 和模型参数替换规则中的超参数α)进行了实验设置. 通过图 6(a)可知, 当 C=5 时, 与未设置梯度裁剪的模型训练得到的训练准确度接近. 由于在差分隐私处理时, 阈值 C 的值越小, 添加的差分噪声越少, 对确保模型高准确度越有利. 为此, 在 SEFL 方案中, 设置梯度裁剪门限值 C=5. 为了减少模型参数替换对模型训练准确度的负面影响, 设置合理的超参数α至关重要. 由图 6(b)可知, 当 α=2 时, 模型训练受到的影响最小, 为此, 在后续的实验过程中, 设置超参数α=2.

为了验证基于全局模型的自动采集方法的高效性, 对比了本文所提出的自动采集方法与无差别采集方法对模型训练的影响. 在该对比实验中, 模型 Softmax 输出的最大概率值阈值范围设置为 [0.5, 0.9]. 通过图 7(a)可知, 自动采集的方式有针对性地采集未学习好的交通标志样本, 能更快地促使模型得到较高的准确度, 提升了模型收敛速度. 而无差别采集方式, 存在重复采集已学习过的交通标志样本的情况, 会降低模型的泛化性能. 由此可知, 设置阈值的方法保证了采集的有效性. 此外, 实验选取了 1 000 个交通标志样本来模拟可被采集的交通标志样本集. 若采取无差别采集, 意味着车辆在运行中遇到任何交通标志都会进行采集, 容易造成存储空间浪费. 本文提出的基于全局模型的自动采集方法, 随着模型泛化能力的提升, 可以采集的样本会越来越少, 如图 7(b)所示. 该方式有效释放了车辆存储空间, 无需采集冗杂重复的样本进行模型训练, 进一步优化了车辆存储与计算资源.



(a) 采集方式对比 (b) 可采集样本数量变化

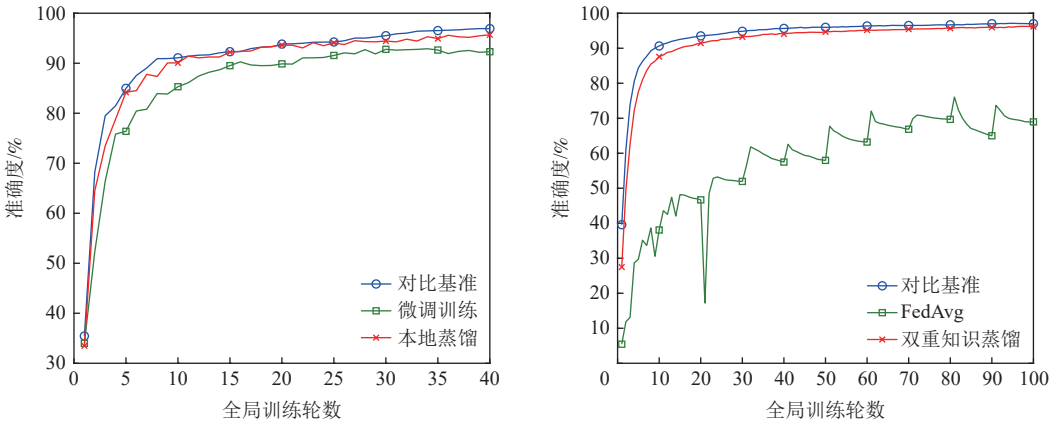
Fig. 7 Comparison of collecting methods and changes in collecting range on GTSRB

图 7 GTSRB 上的样本采集方式对比及采集范围变化

基于 GTSRB 和车辆识别数据集对本文提出的双重知识蒸馏算法进行了性能评估. 为了能更好地体现所提算法的可行性与优越性, 分别基于不同假设场景完成实验对比.

为了验证本地独立训练的性能, 针对单个本地用户进行实验验证. 在该实验中, 为了模拟用户本地数据动态变化情形, 用户每轮参与训练随机选取 775 个样本数据. 对比了基准算法、微调训练以及本文所提出的本地独立训练方法(本地蒸馏), 其中对比基准基于所有被选取的数据进行训练, 微调训练则是每轮基于用户当前数据直接进行模型更新. 实验结果如图 8(a)所示, 基于蒸馏知识的本地训练方式学习到的知识与基准方案接近, 且高于基于微调训练的准确度, 表明基于蒸馏知识训练的本地模型能学习到较多样本信息, 能有效缓解数据变化对模型训练的影响. 此外, 基于联邦学习架构, 验证了数据动态变化导致的灾难性遗忘问题, 以及本文所提出的双重知识蒸馏算法在处理灾难性遗忘问题的效果. 该实验中, 20 个用户共同参与训练, 每个用户每轮训练样本数量为 500. 由于车辆存储空间有限, 会出现新数据覆盖旧数据的情形. 为模拟该场景, 假设每隔 10 轮训练, 用户样本数据完全替换为与先前不重复的 500 个新样本数据. 实验结果如图 8(b)所示, 传统的 FedAvg 算法在处理动态数据时容易出现灾难性遗忘问题, 导致模型准确度下降. 本文提出的基于双重知识蒸馏的训练算法, 能有效解决灾难性遗忘问题, 并得到较高训练准确度.

本文所提出的双重知识蒸馏算法主要由本地独立训练和全局模型更新 2 部分组成, 为了验证每个组成部分的合理性与必要性, 基于 GTSRB 数据集和车辆识别数据集进行了相关对比实验. 实验对比了 5 种不同的联邦训练方法: 训练方式 1, 即本地独立训

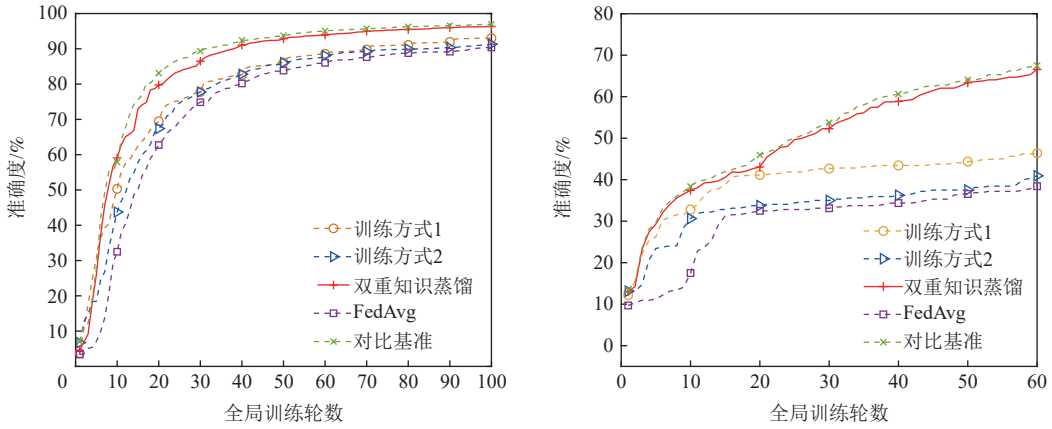


(a) 本地蒸馏 (b) 双重知识蒸馏

Fig. 8 Local independent training and training based on dual knowledge distillation

图 8 本地独立训练和基于双重知识蒸馏训练

练与不基于蒸馏的全局模型更新结合；训练方式 2，即无本地独立训练与基于蒸馏的全局模型更新结合；双重知识蒸馏的本地训练方式；FedAvg 训练方式；对比基准方案，即基于所有已有数据进行模型训练. 在基于 GTSRB 的实验中，参与训练的用户数量为 20，每个用户最大存储样本量为 100，每轮训练都有 10 个新样本数据被采集. 图 9(a)为实验对比结果，基于双重知识蒸馏算法得到的模型准确度与对比基准接近，且高于其他训练方式得到的模型准确度. 在基于车辆识别数据集的实验中，由于数据量较少，设置参与训练的用户数量为 10，每个用户最大存储样本量为 20，每轮训练都有 2 个新样本数据被采集. 图 9(b)是实验对比结果，同样表明基于双重知识蒸馏算法表现优越，且该算法中的本地独立训练与基于蒸馏知识的全局更新方式都有助于解决灾难性遗忘问题，确保高模型准确度.



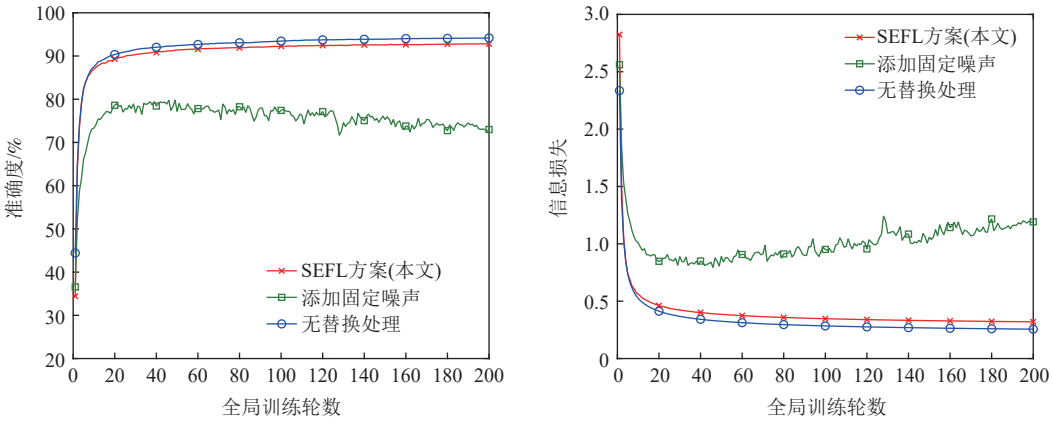
(a) 基于GTSRB (b) 基于车辆识别数据集

Fig. 9 Comparison of different local training methods

图 9 不同本地训练方式的对比

对比了不同的差分隐私保护策略，包括对每轮训练添加固定大小的差分噪声、添加可变隐私噪声但不进行模型替换、SEFL 方案中的自适应差分隐私等策略. 如图 10 所示，对每轮添加固定噪声，随着训练的进行，噪声会叠加导致模型无法收敛. 为了保证模型收敛，要求添加的噪声随着训练逐渐减少. 为此，本文的噪声尺寸设置为 $\frac{2c\eta_T C}{\varepsilon}$，其中 $\eta_T=\frac{lr_0}{\rho\times epcoh+1}$，$lr_0$ 为初始学习率，ρ 为常数，$epcoh$ 为全局训练轮数，$c=4$，$\varepsilon=5$. 由第 4 节可知，分配隐私预算轮数越少，提供的差分隐私保护越强.

在关于模型参数替换率的实验中，对模型进行了 200 轮训练，当发生模型替换时，标注为 0，不发生替换则用 1 表示，图 11 展示了 200 轮训练中发生模型替换的情况. 其中，进行模型替换的次数高达 181 次，占比为 90.5%. 大幅度缩减了参与预算分配的轮数，显著提升了隐私保护强度. 结合图 10 和图 11 可



(a) 准确度对比 (b) 信息损失对比

Fig. 10 Comparison of accuracy and information loss with the same privacy protection

图 10 相同隐私保护程度下的准确度与信息损失对比

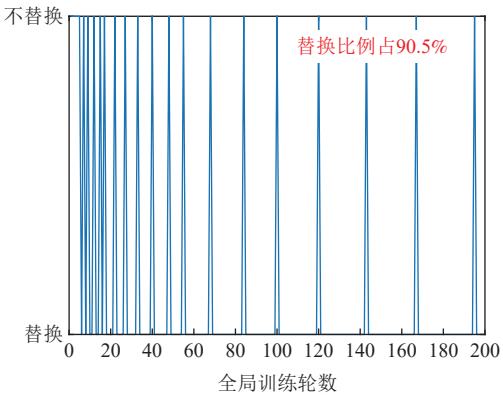


Fig. 11 Model parameter replacement during training

图 11 训练过程中的模型参数替换

总结得到, 本文设计的 SEFL 方案能保证模型收敛, 在隐私保护强度显著提升的情况下只牺牲少量的模型准确度.

本文设计的 SEFL 方案能有效处理车联网数据持续共享中存在的采集低效、灾难性遗忘等问题, 所提出的双重知识蒸馏训练算法保证了识别模型的准确度, 自适应差分隐私策略显著提升了隐私保护强度. 为车联网中分布式模型学习提供了安全高效的训练方法, 对促进自动驾驶、智慧车联网等领域发展具有积极效应.

6 总 结

本文致力于解决车联网数据持续共享中存在的采集低效、灾难性遗忘以及隐私泄露等问题, 设计了一种安全高效的联邦学习方案 SEFL. 该方案中, 提出了基于全局模型的自动采集方法、基于双重知识蒸馏的训练算法以及自适应差分隐私策略, 确保了车联网数据可安全持续地用于交通信息识别等人工智能任务, 并训练得到高模型准确度. 最后, 通过安全性分析和实验评估对 SEFL 方案性能进行了评估. 实验结果表明, 本文所提出的 SEFL 方案在处理存储空间有限、数据动态变化、具有隐私需求的车联网数据上表现优越.

作者贡献声明: 乐俊青完成方案设计、实验设计、论文编写; 谭州勇完成方案设计、实验仿真、论文编写及修改工作; 张迪完成方案设计、论文编写与修改工作; 刘高完成方案论证、论文修改; 向涛与廖晓峰完成论文指导与修改工作.

参 考 文 献

[1] Yang Fangchun, Wang Shangguang, Li Jinglin, et al. An overview of Internet of vehicles[J]. China Communications, 2014, 11(10): 1−15

[2] Muhammad K, Ullah A, Lloret J, et al. Deep learning for safe autonomous driving: Current challenges and future directions[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(7): 4316−4336

[3] Liu Zhanwen, Zhao Xiangmo, Li Qiang, et al. A traffic sign recognition method based on graph models and convolutional neural networks[J]. Journal of Transportation Engineering, 2016, 16(5): 122−131 (in Chinese)
(刘占文, 赵祥模, 李强, 等. 基于图模型与卷积神经网络的交通标志识别方法[J]. 交通运输工程学报, 2016, 16(5): 122−131)

[4] Konečný J, McMahan H B, Ramage D, et al. Federated optimization: Distributed machine learning for on-device intelligence[J]. arXiv preprint, arXiv: 1610.02527, 2016

[5] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proc of the 20th Int Conf on Artificial Intelligence and Statistics. Cambridge, MA:

MIT, 2017: 1273−1282

[6] Zhao Yue, Li Meng, Lai Liangzhen, et al. Federated learning with non-IID data[J]. arXiv preprint, arXiv: 1806.00582, 2018

[7] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design[C/OL]//Proc of the 2nd Conf on Machine Learning and Systems. 2019[2024-03-12].https://proceedings.mlsys.org/paper_files/ paper/2019/file/7b770da633baf74895be22-a8807f1a8f-Paper.pdf

[8] Liu Biao, Zhang Fangjiao, Wang Wenxin, et al. Byzantine-robust federated learning algorithm based on matrix mapping[J]. Journal of Computer Research and Development, 2021, 58(11): 2416−2429 (in Chinese)
(刘飚, 张方佼, 王文鑫, 等. 基于矩阵映射的拜占庭鲁棒联邦学习算法[J]. 计算机研究与发展, 2021, 58(11): 2416−2429)

[9] Manias M D, Shami A. Making a case for federated learning in the Internet of vehicles and intelligent transportation systems[J]. IEEE Network, 2021, 35(3): 88−94

[10] Xing Ling, Zhao Pengcheng, Gao Jianping, et al. A survey of the social Internet of vehicles: Secure data issues, solutions, and federated learning[J]. IEEE Intelligent Transportation Systems Magazine, 2022, 15(2): 70−84

[11] Xie Kan, Zhang Zhe, Li Bo, et al. Efficient federated learning with spike neural networks for traffic sign recognition[J]. IEEE Transactions on Vehicular Technology, 2022, 71(9): 9980−9992

[12] Stergiou K D, Psannis K E, Vitsas V, et al. A federated learning approach for enhancing autonomous vehicles image recognition[C]// Proc of the 4th Int Conf on Computer Communication and the Internet. Berlin: Springer, 2022: 87−90

[13] Liang Feiyuan, Yang Qinglin, Liu Ruiqi, et al. Semi-synchronous federated learning protocol with dynamic aggregation in Internet of vehicles[J]. IEEE Transactions on Vehicular Technology, 2022, 71(5): 4677−4691

[14] Zhou Xiaokang, Liang Wei, She Jinhua, et al. Two-layer federated learning with heterogeneous model aggregation for 6G supported Internet of vehicles[J]. IEEE Transactions on Vehicular Technology, 2021, 70(6): 5308−5317

[15] Zhou Hongliang, Zheng Yifeng, Huang Hejiao, et al. Toward robust hierarchical federated learning in Internet of vehicles[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(5): 5600−5614

[16] Li Zhizhon, Hoiem D. Learning without forgetting[J]. IEEE Transactions on Pattern Analysis and Aachine Intelligence, 2017, 40(12): 2935−2947

[17] Kirkpatrick J, Pascanu R, Rabinowitz N, et al. Overcoming catastrophic forgetting in neural networks[J]. Proceedings of the National Academy of Sciences, 2017, 114(13): 3521−3526

[18] Rebuffi S A, Kolesnikov A, Sperl G, et al. iCaRL: Incremental classifier and representation learning[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2001−2010

[19] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1322−1333

[20] Shokri R, Stronati M, Song Congzheng, et al. Membership inference attacks against machine learning models[C]//Proc of the 38th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2017: 3−18

[21] He Zecheng, Zhang Tianwei, Lee R B. Model inversion attacks against collaborative inference[C]//Proc of the 35th Annual Computer Security Applications Conf. New York: ACM, 2019: 148−162

[22] Zhou Chunyi, Chen Dawei, Wang Shang, et al. Research progress and challenges in privacy and security attacks on distributed deep learning[J]. Journal of Computer Research and Development, 2021, 58(5): 927−943 (in Chinese)
(周纯毅, 陈大卫, 王尚, 等. 分布式深度学习隐私与安全攻击研究进展与挑战[J]. 计算机研究与发展, 2021, 58(5): 927−943)

[23] Bonawitz K, Ivanov V, Kreuter B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//Proc of the 24th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 1175−1191

[24] Fang Chen, Guo Yuanbo, Wang Na, et al. Highly efficient federated learning with strong privacy preservation in cloud computing[J]. Computers & Security, 2020, 96: 101889

[25] Lu Yunlong, Huang Xiaohong, Zhang Ke, et al. Blockchain empowered asynchronous federated learning for secure data sharing in Internet of vehicles[J]. IEEE Transactions on Vehicular Technology, 2020, 69(4): 4298−4311

[26] Shokri R, Shmatikov V. Privacy-preserving deep learning[C]//Proc of the 22nd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2015: 1310−1321

[27] McMahan H B, Ramage D, Talwar K, et al. Learning differentially private recurrent language models[J]. arXiv preprint, arXiv: 1710.06963, 2017

[28] Wei Kang, Li Jun, Ding Ming, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3454−3469

[29] 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

[30] Dwork C, Roth A. The algorithmic foundations of differential privacy[J]. Foundations and Trends® in Theoretical Computer Science, 2014, 9(3/4): 211−407

[31] Dwork C. Differential privacy: A survey of results[C]//Proc of the 5th Int Conf on Theory and Applications of Models of Computation. Berlin: Springer, 2008: 1−19

[32] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy[C]//Proc of the 23rd ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308−318

[33] Amin K, Kulesza A, Munoz A, et al. Bounding user contributions: A bias-variance trade-off in differential privacy[C]//Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 263−271

[34] Andrew G, Thakkar O, McMahan B, et al. Differentially private learning with adaptive clipping[J]. Advances in Neural Information

Processing Systems, 2021, 34: 17455−17466

[35] Le Junqing, Zhang Di, Lei Xinyu, et al. Privacy-preserving federated learning with malicious clients and honest-but-curious servers[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4329−4344

[36] Dwork C, Lei Jing. Differential privacy and robust statistics[C]//Proc of the 41st Annual ACM Symp on Theory of Computing. New York: ACM, 2009: 371−380

[37] Gou Jianping, Yu Baosheng, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789−1819

[38] Fan Liyue, Li Xiong. An adaptive approach to real-time aggregate monitoring with differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2094−2106



Le Junqing, born in 1991. PhD, research assistant. Member of CCF. His main research interests include privacy protection, federated learning, and information security.

乐俊青, 1991 年生. 博士，助理研究员. CCF 会员. 主要研究方向为隐私保护、联邦学习、信息安全.



Tan Zhouyong, born in 1997. Master. His main research interests include privacy protection and federated learning.

谭州勇, 1997 年生. 硕士. 主要研究方向为隐私保护、联邦学习.



Zhang Di, born in 1993. PhD, research assistant. Her main research interests include privacy protection, blockchain, and cryptography.

张　迪, 1993 年生. 博士，助理研究员. 主要研究方向为隐私保护、区块链、密码学.



Liu Gao, born in 1991. PhD, research assistant. His main research interests include Internet of vehicles, blockchain, and privacy protection.

刘　高, 1991 年生. 博士，助理研究员. 主要研究方向为车联网、区块链、隐私保护.



Xiang Tao, born in 1980. PhD, professor, PhD supervisor. Member of CCF. His main research interests include privacy protection, information security, and machine learning.

向　涛, 1980 年生. 博士，教授，博士生导师. CCF 会员. 主要研究方向为隐私保护、信息安全、机器学习.



Liao Xiaofeng, born in 1964. PhD, professor, PhD supervisor. Member of CCF. His main research interests include neural network, privacy protection, and cryptography.

廖晓峰, 1964 年生. 博士，教授，博士生导师. CCF 会员. 主要研究方向为神经网络、隐私保护、密码学.