计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202111100
60(6): 1276−1295, 2023

面向边缘智能的联邦学习综述

张雪晴[1,2] 刘延伟[1] 刘金霞[3] 韩言妮[1]
[1](中国科学院信息工程研究所 北京 100093)
[2](中国科学院大学网络空间安全学院 北京 100049)
[3](浙江万里学院 浙江宁波 315100)
(zxq20141213@163.com)

An Overview of Federated Learning in Edge Intelligence

Zhang Xueqing[1,2], Liu Yanwei[1], Liu Jinxia[3], and Han Yanni[1]
[1] (*Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100093)
[2] (*School of Cyber Security*, *University of Chinese Academy of Sciences*, *Beijing* 100049)
[3] (*Zhejiang Wanli University*, *Ningbo*, *Zhejiang* 315100)

Abstract With the increasing demand of edge intelligence, federated learning (FL) has been now of great concern to the industry. Compared with the traditionally centralized machine learning that is mostly based on cloud computing, FL collaboratively trains the neural network model over a large number of edge devices in a distributed way, without sending a large amount of local data to the cloud for processing, which makes the compute-extensive learning tasks sunk to the edge of the network closed to the user. Consequently, the users' data can be trained locally to meet the needs of low latency and privacy protection. In mobile edge networks, due to the limited communication resources and computing resources, the performance of FL is subject to the integrated constraint of the available computation and communication resources during wireless networking, and also data quality in mobile device. Aiming for the applications of edge intelligence, the tough challenges for seeking high efficiency FL are analyzed here. Next, the research progresses in client selection, model training and model updating in FL are summarized. Specifically, the typical work in data unloading, model segmentation, model compression, model aggregation, gradient descent algorithm optimization and wireless resource optimization are comprehensively analyzed. Finally, the future research trends of FL in edge intelligence are prospected.

Key words federated learning; edge computing; edge intelligence; model aggregation; resource constraints

摘 要 随着边缘智能需求的快速增长,联邦学习(federated learning, FL)技术在产业界受到了极大的关注.与传统基于云计算的集中式机器学习相比,边缘网络环境下联邦学习借助移动边缘设备共同训练机器学习模型,不需要把大量本地数据发送到云端进行处理,缩短了数据处理计算节点与用户之间的距离,在满足用户低时延需求的同时,用户数据可以在本地训练进而实现数据隐私保护.在边缘网络环境下,由于通信资源和计算资源受限,联邦学习的性能依赖于无线网络状态、终端设备资源以及数据质量的综合限制.因此,面向边缘智能应用,首先分析了边缘智能环境下高效联邦学习面临的挑战,然后综述联邦学习在客户端选择、模型训练与模型更新等关键技术方面的研究进展,最后对边缘智能联邦学习的发展趋势进行了展望.

收稿日期: 2021-11-08; **修回日期**: 2022-09-16
基金项目: 国家自然科学基金项目(61771469); 重庆市属本科高校与中国科学院所属院所合作项目(HZ2021015)
This work was supported by the National Natural Science Foundation of China (61771469) and the Cooperation Project Between Chongqing Municipal Undergraduate Universities and Institutes Affiliated to CAS (HZ2021015).
通信作者: 刘延伟(liuyanwei@iie.ac.cn)

关键词 联邦学习；边缘计算；边缘智能；模型聚合；资源受限

中图法分类号 TP3

随着移动通信技术的快速发展和智能终端的普及，连接到网络的边缘设备数量和智能应用持续增长，人类将迎来边缘智能[1]时代. 边缘智能应用大多基于机器学习技术,例如视频推荐[2]、人脸识别[3]、自动驾驶[4]与无人机[5]等. 像自动驾驶和增强现实这样的智能应用需要更多的计算和数据资源以及更短的处理时延需求. 传统的机器学习基于云计算平台对数据进行集中处理训练. 由于边缘智能终端产生的数据量大、计算任务分散以及数据的隐私保护需求，将所有数据发送到云端进行处理是不切实际的. 近年来，5G 系统中引入了移动边缘计算（mobile edge computing, MEC）[6]架构，将计算、存储和网络资源与基站集成，将计算能力从云端下沉到网络边缘，缩短了数据处理计算节点与用户之间的距离，能够满足用户低时延的需求.

MEC 是一项快速发展的技术，旨在通过利用边缘设备未充分利用的计算和通信资源，在无线网络边缘部署移动应用. 作为传统集中式云计算的补充，MEC 在降低核心网络流量负载、缓解中央服务器处理压力、缩短端到端操作响应延迟以及提高无线网络整体系统性能方面表现出巨大潜力. MEC 提供了分布式计算环境，可用于部署应用程序和服务. 但是，多个终端想彼此分享各自的数据集和学到的知识，面临着监管约束、隐私以及安全问题. 而且，相比于使用所有终端数据进行训练，只使用一个终端的数据训练获得的模型不够精确. 面对这样的形势，FL（federated learning, FL）[7-8]技术应运而生. 由于不需要共享和传输原始数据，采用类似集群的通信结构，FL 更适合于移动终端等大规模、广分布的部署环境，得到了广泛认可.

FL 采用分布式学习架构，使得神经网络模型在 MEC 架构下可以进行分布式训练，参与学习的客户端无需上传本地数据，只需将训练后的模型参数更新上传，再由边缘服务器节点聚合、更新参数并下发给参与学习的客户端. 图 1 给出了面向无人机和车联网边缘智能应用环境下 FL 的经典部署架构. 由于边缘智能应用独特的环境特性，包括其动态的无线信道状态、广泛变化的本地数据集大小、设备处理能力和设备电量有限等，边缘智能环境下的 FL 面临着诸多挑战. 首先，在边缘智能应用中，FL 能够从每个终端的本地数据集中提取有用的信息，而不需要将数据传送到一个中心位置，在本地设备保留原始数据的同时，训练多个终端共享的神经网络模型，解决了以往智能网络模型只能通过云端下发，而无法在本地训练的问题[9]. 但挑战在于，对于 MEC 来说，FL 相当耗费资源. 尽管原始数据不再需要发送到中心服务器，但由于高维度的模型训练需要大量的计算资源，因此优化模型也是 FL 需要考虑的问题之一. 其次，FL 通过平均局部随机梯度下降（stochastic gradient descent, SGD）[10]来更新模型，参与学习的终端设备与中央参数服务器之间需要频繁地进行参数交换，高频次的模型更新过程必然会占用相当多的带宽资源，较高的通信成本是 FL 在实际应用中面临的另一个关键问题. 再次，在无线资源受限的边缘网络下，由于参与设备在数据质量、通信网路、计算能力和参与意愿等方面的异构性，训练跨终端数据的共享模型是一个具有挑战性的任务.

针对这些挑战，研究人员进行了深入研究，并取得了一定的进展，但还存在一些值得深入的剖析的问题. 经过文献调研分析表明，如表 1 所示，以往的 FL 综述缺少对上述问题的深入讨论. 基于这一点，本文从 FL 如何应对边缘智能应用环境挑战为主线，首先简要概括 FL 基本原理，然后从客户端选择方法、模型训练优化技术、模型更新技术几个方面详细综述现有的边缘智能 FL 关键技术，并讨论了未来边缘智能系统下 FL 的研究趋势.

1 FL 基本原理

FL 是一种使用分布式机器学习方法来保护多方合作数据隐私的技术. FL 的中心节点负责神经网络模型参数聚合与参数配置功能. 每个终端根据自己的数据进行模型训练. 每个客户端模型都有在每个数据样本 j 的参数向量 $\boldsymbol{w}$ 上定义的损失函数. 损失函数捕获训练中模型的误差，并且模型学习过程是将训练数据样本集合上的损失函数最小化. 样本 j 的损失函数定义为 $f(\boldsymbol{w}, \boldsymbol{x}_j, y_j)$,其中，向量 $\boldsymbol{x}_j$ 和标量 y_j 是 1 个训练数据样本 j 的 2 个组成部分. $\boldsymbol{x}_j$ 被视为学习模型的输入，y_j 是模型的期望输出.

假设有 K 个终端，它们的本地数据分别表示为 $D_1, D_2, \cdots, D_i, \cdots, D_K$. 对于每个终端 K 的数据集 D_k，收集的损失函数为

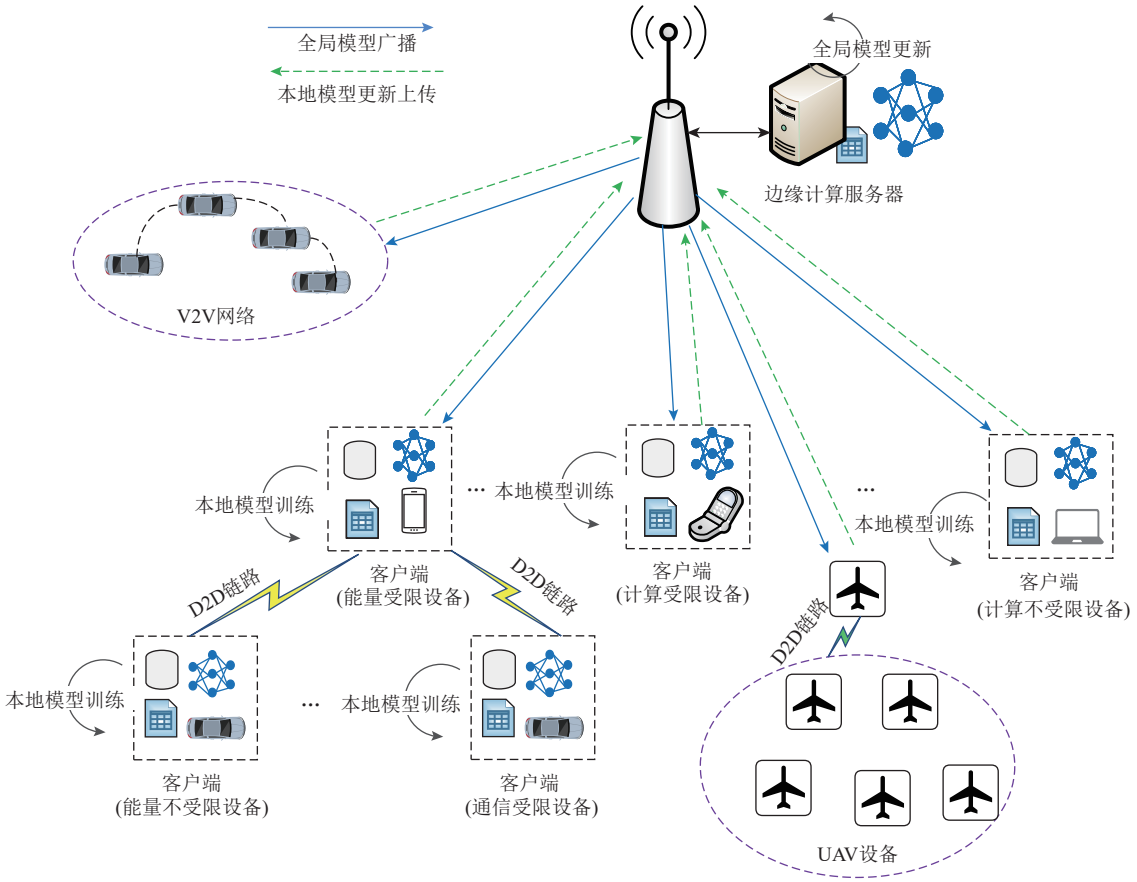

Fig. 1 Edge intelligent federated learning architecture

图 1 边缘智能联邦学习架构

Table 1 Comparison of Studies on Existing Federated Learning Reviews

表 1 现有联邦学习综述研究对比

研究工作	资源优化	激励机制	算法设计	从无线网络角度优化	无线应用	说明
文献 [11]	√	√	√	×	×	考虑边缘网络的 FL
文献 [12]	×	×	√	×	×	
文献 [13]	×	×	√	×	×	
文献 [14]	×	×	×	×	×	主要考虑用于缓存和计算卸载的 FL
文献 [15]	×	×	√	×	×	
本文	√	√	√	√	√	

注: “√”表示文献中完成了该项工作, “×”表示文献中未完成该项工作.

$$F_k(\boldsymbol{w})=\frac{1}{|D_k|}\sum_{j\in D_k}f_j(\boldsymbol{w}), \tag{1}$$

其中 $f_j(\boldsymbol{w})$ 是 $f(\boldsymbol{w},\boldsymbol{x}_j,y_j)$ 的简写，因此在所有分布式数据集上定义的损失函数为

$$F(\boldsymbol{w})=\frac{\sum\limits_{j\in U_k|D_k|}f_j(\boldsymbol{w})}{\left|\bigcup_{k=1}^{k}\times\left|D_k\right|\right|}=\frac{\sum\limits_{k=1}^{K}|D_k|\times F_k(\boldsymbol{w})}{\sum\limits_{k=1}^{K}|D_k|}, \tag{2}$$

其中 $|D_k|$ 表示客户端 k 在 D_k 里的元素数量，$D=\cup_{k=1}^{K}D_k$，$|D|=\sum\limits_{k=1}^{K}|D_k|$. 该模型被训练成最小化损失函数，例如使用梯度下降法，以找到最佳参数集学习的目标最小化损失函数 $F(\boldsymbol{w})$:

$$\boldsymbol{w}^*=\arg\min F(\boldsymbol{w}). \tag{3}$$

由于原始数据分布在不同的客户端，FL 不能像集中式机器学习一样在中央服务器上直接找到梯度. 如果使用梯度下降法来最小化全局损失函数，即

$$\boldsymbol{w}(t):=\boldsymbol{w}(t-1)-\mu\nabla F(\boldsymbol{w}(t-1))=\frac{\sum\limits_{k=1}^{K}|D_k|\times\boldsymbol{w}_k(t)}{\sum\limits_{k=1}^{K}|D_k|}, \tag{4}$$

其中 μ 是一个小的正数，即学习率. $\nabla F(\cdot)$ 是损失函数局部梯度. $\boldsymbol{w}(t)$ 是中央服务器在时间 t 的全局聚集参数集，$\boldsymbol{w}_k(t)$ 是客户端 k 在时间 t 的本地参数集，可以表示为

$$\boldsymbol{w}_k(t)=\boldsymbol{w}(t-1)-\mu\nabla F_k(\boldsymbol{w}(t-1)). \tag{5}$$

之后，只要获得局部梯度 $\nabla F_k(\boldsymbol{w}(t-1))$，中央服务器就可以计算 $\boldsymbol{w}(t)$. 因此，只需要将本地梯度发送到

中央服务器，就可以节省通信资源，特别是当使用梯度压缩时，能够一定程度地减少传输的梯度数据量.

2 FL 客户端选择技术

在边缘智能应用中，移动设备并不总是用于训练数据. 一方面，边缘设备的存储和计算资源有限，网络中的边缘设备并不能都用于参加每一轮 FL 训练. 此外，边缘设备采集的实际数据往往是非独立同分布的，这也会影响训练效率. 另一方面，参与学习的设备将模型状态信息更新并上传到边缘服务器的能力高度依赖于各自的无线信道状态. 当边缘设备处于糟糕的无线信道条件下或边缘设备掉队[16]时，将导致更长的模型更新时间，进而耽误后续训练. 由于边缘智能这种独特的环境特性，在资源限制下为 FL 每轮训练选择合适的参与者就变得尤为重要[12].

过度的训练迭代和模型转换会占用大量的计算和通信资源. 一些研究人员提出通过优化资源使用的方式来选择参与学习的客户端. Jin 等人[17]提出选择适当的客户端设备并排除不必要的模型更新以帮助节省资源，并设计了一个在线学习算法，以在线方式共同控制参与者的选择. 但是该算法不同于常用的 FedAvg[7]算法，不能体现出部分客户端参与训练从而对模型更新产生的影响. Chai 等人[18]根据客户端的训练性能将客户端划分为不同的层，并在每轮训练中从同一层中选择客户端，以缓解由于资源和数据量的异质性而导致的模型偏离问题. Chai 等人[18]提出的 TiFL 是一种同步 FL 方法. 这种方法的一个明显的缺点是：在每次全局迭代时，当 1 个或多个客户端遭受较高的网络延迟，或者客户端有更多的数据，需要更长的训练时间时，其他客户端必须等待模型更新. 由于参数服务器通常在所有客户端完成 1 次迭代训练后进行聚合，同步优化协议中延长的等待时间会导致计算资源的浪费. 文献 [19]提出了一种用于 FL 的分级在线速度控制框架，它通过一种节能的方式来平衡训练时间和模型精度. 文献 [20] 提出一种基于社交知识的聚类算法. 首先，通过考虑社会关系和计算能力，将一组密集的设备组成一个集群，然后选择簇头（中央设备），簇头节点执行与传统 FL 中的参数服务器相同的功能，实现自组织 FL. 该学习算法利用较长的电池寿命、与其他设备较好的连接性能，以及更多的计算资源等关键参数来选择簇头. 在无线 FL 网络中，学习性能取决于在每一轮迭代训练中如何选择客户端以及如何在选定的客户端之间进行带宽分配. 以往的研究方法试图通过分配有限的无线资源来优化 FL，但它们关注的是单次学习迭代的问题. Xu 等人[21]从一个新的视角来看待无线 FL 网络中的资源配置，认识到迭代学习不仅在时间上相互依赖，而且对最终的学习结果有着不同的意义，并针对长期能量约束下的联合客户选择和带宽分配的随机优化问题，提出利用当前可用的无线信道信息来获得长期的性能保证. 因为有些客户端比其他客户端慢，所以提供异步[22-23]和半同步[24]学习机制.

在客户端选择协议方面，如图 2 所示. Nishio 等人[25]提出了一个 FL 客户端选择协议，即 FedCS. FedCS 为客户端在 FL 协议中下载、更新和上传机器学习模型设置了一个期限，以保证中央参数服务器在该期限内聚合尽可能多的客户端更新，从而使整个训练过程高效，减少了训练所需要的时间. FedCS 解决了 FL 参与者之间资源异构的问题，但忽略了数据分布异构的特性. 为了解决这个问题，Yoshida 等人[26]将 FedCS 扩展成处理参与者之间数据分布差异的混合 FL（hybrid federated learning, Hybrid-FL）协议. Hybrid-FL 协议中，中央参数服务器在资源请求阶段询问随机参与者是否允许上传数据. 在参与者选择阶段，除了考察其计算能力外，还要考察其上传的数据是否可以在中央参数服务器中形成一个近似独立同分布的数据集. 实验结果表明，与 FedCS 相比，即使只有 1% 的参与者共享它们的数据，Hybrid-FL 的分类精度也有显著的提高. 然而，Hybrid-FL 要上传客户端的数据分布信息，可能会侵犯用户的隐私和安全，特别是如果参与者是恶意的，将引入严重的安全问题.

显而易见，具有大量数据样本的设备对全局训练的贡献更大. 在不提供补偿的情况下，这种设备不太愿意与拥有少量数据样本的其他设备联合. 因此，除了资源与数据方面的考量，客户端选择也需要通过激励机制鼓励参与者对 FL 做出贡献[27]. Kang 等人[28]考虑了高质量移动设备的选拔和可靠模型训练的激励问题. 为了提高 FL 任务的性能，每个任务发布者都选择具有高精度和可靠本地数据的高信誉客户端候选者作为参与者. 每个任务发布者通过主观逻辑模型计算参与交互的客户端的信誉分数，将以往交互产生的直接信誉分数和来自其他任务发布者的间接信誉分数集成到一个综合信誉中进行评价. 这些客户端信誉分数由第三方区块链维护，并在任务发布者之间共享，信誉越高的客户端可以从任务发布者那里获得更多的奖励. 而且，边缘智能设备可能会参与并中断训练过程；恶意设备可能对本地训练过程没有

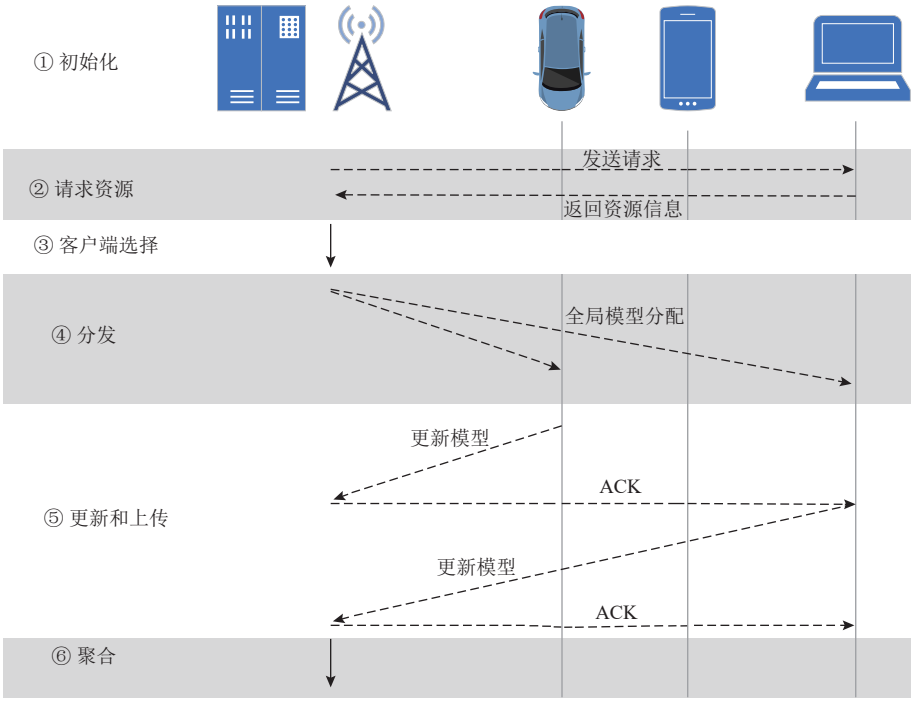

Fig. 2 FedCS protocol overview

图 2 FedCS 协议概述

贡献，而只接收其他设备计算的全局训练结果. 使用分布式分类账技术(distributed ledger technology, DLT)记录训练过程有助于缓解这些问题. 例如，当交换本地模型状态信息时，每个设备交叉验证模型状态信息，并将接受的模型状态信息存储在其本地分布式分类账中. 本地分布的分类账通过 DLT 与其他设备的分类账同步，例如借助于区块链技术，训练过程提高了对恶意的、有故障的设备的鲁棒性[29]. 考虑到数据分布对 FL 性能的影响，Li 等人[30]提出一个类似的加权激励方法来选择客户端，即 q-FedAvg. q-FedAvg 通过为性能较差的终端设备分配比性能较好的终端设备更高的权重来修改 FedAvg 的目标函数，将损失函数中的更高权重分配给损耗更高的设备，鼓励在联邦训练上下文中跨设备的更公平的准确性分布. 文献[31]提出一个众包框架，以一种通信高效的方式支持无线物联网环境中的 FL，并引入了一种基于 Stackelberg 博弈模型的激励机制，以吸引客户参与 FL.

客户端之间的通信，尤其是通过无线信道的通信，可能是不对称、缓慢和不稳定的. 而假设具有高信息传输速率和可忽略的数据包丢失的通信环境是不现实的. 例如，移动互联网的上传速度通常比下载速度慢得多. 一些参与者可能会因为与互联网断开连接而退出，尤其是在拥挤的无线通信的场景下使用手机[15]. 此外，本地实际数据往往是非独立同分布的，会显著影响学习效率. 而且，大量参与训练的客户端可能会加剧通信拥塞，因此需要合理分配有限的无线频谱资源. 文献 [32] 提出了一个在蜂窝连接无线系统中进行 FL 的通信和计算模型. 考虑到上行带宽的限制，基站需要选择合适的客户端来执行 FL 算法，以最小化成本. 在给定预定义的子信道束和本地精度的情况下，客户端优化传输功率和 CPU 周期频率，以在满足 FL 延迟要求的同时最小化能耗. 不同于文献 [32] 仅仅考虑了无线信道需求，文献 [33] 还考虑了每个客户的数据大小、数据分布，重点研究了在无线通信场景中的 FL，并提出了一种基于深度 RL 的拍卖机制，以鼓励和选择数据所有者参与 FL. 与上文提到的 Stackelberg 博弈和契约理论不同，拍卖机制允许客户端主动报告其类型，并已应用于各种应用场景[34].

选择合适的客户端能够剔除训练过程中一些计算资源有限(即需要更长的更新时间)或无线信道条件差(即上传时间更长)的客户端，这样有利于 FL 持续训练. 无论是采用资源优化，还是通过激励机制或者从安全的角度考虑数据分布的方式，这些方法关注的都是客户端的性能对训练带来的影响. 然而，使用超大规模的数据，训练一个具有数十亿参数的复杂模型，单单从选择性能较好的客户端这一单一角度，无法保证 FL 在边缘智能应用环境下的整体性能，因此还需要对 FL 训练过程进行优化. 表 2 归纳了现有的 FL 客户端选择方案.

Table 2 Comparison of Federated Learning Client Selection Schemes

表 2 联邦学习客户端选择方案比较

方案类型	方案思路	客户端目标	服务器端目标
计算与通信资源优化	剔除不必要的模型更新[17]、客户端分层[18]、控制学习节奏[19]、基于聚类实现自组织学习[20]、长期能耗约束下的带宽分配[21]、设置学习期限[25]、基于设备的计算能力进行选择[26]		
激励机制	契约理论[28]：基于信誉进行激励反馈鼓励可靠的终端设备参与学习	奖励和能耗的平衡	最大化由全局迭代时间与补偿给客户端的报酬之间的差异所获得的利润
	Stackelberg 博弈[31]：实现高质量无线通信效率的全局模型	奖励（即准确率等级）和成本（即通信和计算成本）的平衡	获取不同准确率的凹函数
	拍卖理论[32-33]：最大限度地降低客户端投标的成本	奖励和成本的平衡	最小化投标成本
	修订目标函数权重[30]	为了引入潜在的公平性并降低训练精度方差，通过在 q-FedAvg 中分配更高的相对权重来强调具有高经验损失的本地设备	

3 模型训练优化方法

由于终端的算力限制、通信资源限制、用户隐私需求，边缘智能环境下零散分布的大量本地数据往往使 FL 的模型训练变得困难. 面对这些困难，以往的研究在模型训练过程中关注对数据的处理，采用卸载数据到其他设备，或直接在本地对数据进行筛选的方法，去除无用的数据等技术，来解决算力资源不足的问题. 除了卸载数据进而转移其关联的计算外，还可以对模型进行压缩处理，减少 FL 过程中需要交互的模型参数规模，降低通信资源的消耗.

3.1 数据卸载方法

由于隐私保护原则和通信带宽限制，跨个体组织边界共享数据非常困难. 数据摘要[35]是一种减少共享数据量的技术，同时保留数据中对训练机器学习模型有用的特征. 目前的数据摘要研究主要有 3 类方法：1)统计摘要. 这类方法源于对数据进行汇总以有效地探索和分析大量数据的需要. 此外，这类方法生成摘要信息只需要少量的时间和空间，通常只需对整个数据集进行一次遍历即可创建，并且占用较少的内存. 但是，这种类型的摘要只适用于特定类型数据集的查询. 2)降维. 通过将高维数据映射到低维空间，使得原始数据集的某些特征属性保留在映射空间中，不影响学习的效率，降维具有减少数据总量的效果. 3)数据降采样. 与前 2 种方法相比，基于降采样的方法在原始数据集的样本空间内构造一个小的数据样本集，因此可以在 FL 任务中使用小数据集作为原始数据集的代理. 数据摘要通过多种方式辅助 FL，例如，当不同客户端的数据集是非独立同分布时，可以与其他客户共享本地原始数据集的摘要，以提高训练效率.

数据从终端设备卸载到边缘计算服务器，有助于利用边缘计算服务器的强大算力加快 FL 速度，减少回程拥塞[36]. 数据通常由终端设备保存，必须通过无线链路传输到边缘计算节点. FL 任务要求在一定的时间限制内执行，这可能导致传输不完整的数据集. 考虑到每个数据包传输的开销以及计算率和通信率之间的关系，文献 [37] 通过优化数据包的有效负载大小来寻求计算延迟和准确性之间的折中，提出使用优化的块大小进行通信和计算，实现了 FL 中数据与计算卸载. 一些边缘智能学习方法通过对网络边缘的原始数据进行处理和压缩来减少数据传输时间，但是，同时也带来了学习精度降低的问题. 文献 [38] 研究了一个兼顾学习精度的模型参数传输优化的任务调度问题，通过来自于云端的调度，实现了提高学习精度和减少通信流量之间的最优折中. 但是文献 [38] 中提出的架构由一个主节点和多个工作节点组成. 工作节点识别特定领域的对象，并通过管道为主节点提供训练实例. 这种 FL 架构在私有场景，例如在家里，所有设备都有内在动力，协作为其主节点可以创建更智能的模型. 然而，在公共场景中，它并不能很好地工作. 在公共场景中，主节点初始化一个任务并将子任务分配给不熟悉的参与者. 在这种情况下，会出现额外的激励问题.

3.2 模型分割迁移方法

不仅数据卸载可以转移模型训练计算，转移模型同样可以卸载其相关联的计算. 当神经网络模型尺寸过大时，可以将单个神经网络结构分割成分布在多个设备上的多个段，即模型分割. 模型分割迁移就是一种实现计算卸载的方法. 在移动设备和边缘计算服务器之间划分深度神经网络模型，通过将深度

神经网络的浅层部分部署在移动设备上, 而复杂的神经网络深层部分则转移到边缘计算服务器上[39-41]. 首先对本地输入数据进行快速转换, 然后, 将转换后的数据表示发送给边缘计算服务器以进行需要大量时间和计算的推断任务. 但是这种模型分割迁移的方式使用户无法控制数据在边缘计算服务器的处理过程, 隐私得不到保障. 现有的模型分割方式假设网络模型条件不变, 通过划分网络模型操作, 将部分计算转移到云或边缘服务器上. 然而, 边缘智能应用中的网络模型因上下文而异, 深度神经网络模型分割策略的空间有限. 文献 [42] 在端边云协同的场景下, 提出了一种成本驱动型卸载策略, 降低了学习成本, 该方法在一定的场景下表现良好. 然而, 这种策略算法计算复杂度高、执行时间长, 在实时系统中并不适用, 特别是当通信环境动态变化时, 这种策略需要更长的时间来确定新的最佳协作决策, 难以满足实时数据分析的时延要求; 而文献 [43] 考虑模型结构的灵活性, 使其能够实时根据上下文信息动态地做出模型压缩和分割的决策.

基于模型分割后进行部分模型转移需要兼顾模型的隐私保护问题, 为了在没有隐私风险的情况下利用云数据中心的海量计算能力, 文献 [44] 在移动设备和云数据中心之间分割了深度神经网络模型, 提出了 ARDEN 框架来保护隐私. ARDEN 在移动设备上执行简单的数据转换, 然后将需要大量资源的训练转移到云数据中心, 并引入了一种轻量级的隐私保护机制, 不仅对转移的模型部分提供了一定的隐私保护, 而且提高了推断的准确性, 并减少了资源消耗. 边缘智能应用中, 多个用户通过共享一个深度神经网络模型来实现 FL, 模型的隐私保护更为重要. Zhang 等人[45] 采用模型分割技术和差分隐私方法, 提出了一种利用移动边缘计算的 FL 框架(federated learning scheme in mobile edge computing, FedMEC), 该框架是一种典型的模型分割迁移带动计算转移的边缘智能环境的 FL 架构, 具体学习框架如图 3 所示. FedMEC 框架将一个深层神经网络分为 2 部分: 预先训练的客户端神经网络模型和边缘服务器端神经网络模型, 复杂的计算可以通过模型迁移转移给边缘服务器. 同时, 通过差异私有数据扰动机制, 防止局部模型参数隐私泄露.

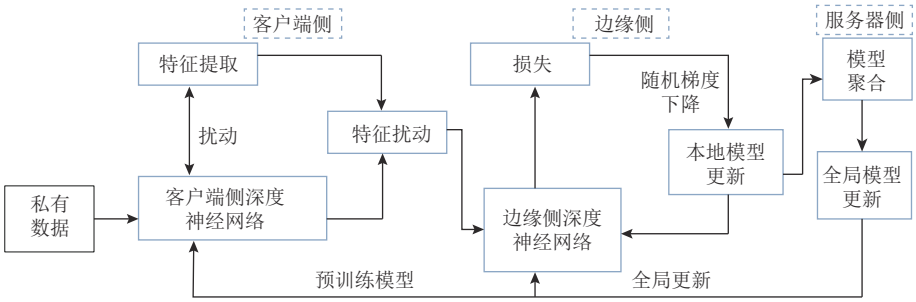

Fig. 3 Model segmentation migration framework

图 3 模型分割迁移框架

在边缘计算环境中, 模型分割技术不仅可以实现高效的 FL 服务, 而且可以降低移动边缘设备上的计算消耗, 即端边协作将深度神经网络分为 2 部分, 其中大部分繁重的计算工作卸载到边缘服务器. 此外, 在部分模型上传到边缘服务器之前, 使用差分隐私保护机制来保护数据隐私. 目前的模型分割迁移技术虽然可以做到模型训练的计算卸载, 但是, 虽然基于差分隐私的保护机制防止模型分割迁移过程中的数据隐私泄露, 却带来了模型训练精度的损失问题. 然而文献 [39−45] 中的研究并没有对其带来的这一负面影响给出解决方案.

3.3 模型压缩方法

虽然用户终端的计算能力在过去十几年中大大提升, 但这些终端仍然受到电池电量和存储空间的限制, 使得大规模部署 FL 成为一个难点问题. 原因主要有 2 个: 1)一个深层的神经网络往往由大量的激活单元和相互连接的节点组成, 因此训练这样一个模型必然会产生大量的能量消耗和内存占用. 2)局部模型的反馈不仅需要高发射功率, 而且需要足够宽的移动频谱以保证训练效率. 为了克服 FL 范式中局部训练和反馈方面的困难, 一种有效的方法是对学习模型进行压缩, 例如将草图应用到 FL 中[46]. 压缩模型大小可以使模型适应更小、更快的内存, 从而实现低延迟的推理和训练. 此外, 模型压缩提高了能效, 因为内存访问是神经网络能耗的主要来源, 与模型大小成正比[47]. 最后, 在分布式训练中, 模型

压缩最小化模型状态信息有效负载，从而减少通信延迟.

表 3 总结了现有的模型压缩技术特点. 文献 [48] 提出结构化和草图更新技术，以减少参与者在每轮通信中发送到 FL 服务器的模型更新数据. 结构化更新方式限制参与者更新预先指定的结构，即低秩和随机掩码. 低秩结构更新是指每次更新都被强制为一个低秩矩阵，表示为 2 个矩阵的乘积. 一个矩阵是随机生成的，并且在每一轮通信中保持不变，而另一个矩阵则被优化. 因此，只需要将优化后的矩阵发送到服务器. 草图更新方式是指在与服务器通信之前以压缩形式对更新数据进行编码，服务器随后在聚合之前对更新数据进行解码. 在文献 [48] 基础上，文献 [49] 进行了扩展研究，提出了 2 种新的策略来减少服务器到客户端的通信负载：1)在发送服务器到客户端的全局模型上使用有损压缩；2)采用联邦退出，允许用户在全局模型的最小子集上进行局部训练，减少客户到服务器的通信和本地计算量. 由于压缩而产生的误差需要在每个客户端保存，同时每一轮都需要大量的客户端参与，这对于 FL 来说是不实用的. 文献 [50] 对文献 [46−49] 方案进行进一步的改进，直接检索最新的梯度值，而不要求更新向量中的位置. 文献 [50] 这种方案更加有效，因为它需要的通信轮数更少. 文献 [48−50] 的研究提出了实用的模型压缩方法，可以减少服务器和参与者之间的通信成本，但通信成本的降低往往伴随着模型精度的牺牲. 因此，对压缩粒度进行形式化刻画将非常有用，尤其是当面临不同的任务或者涉及到不同数量的 FL 参与者时，采用不同的压缩强度尤为重要. FL 通过每隔一段时间交换模型参数来降低通信成本. 基于周期性模型信息交换，Jeong 等人[51] 所提出的联邦蒸馏方法交换的不是模型参数而是模型输出，允许终端设备采用规模较大的局部模型. 在联邦蒸馏基础上，为了解决非独立同分布数据问题，文献 [51] 提出了一种基于生成对抗网络的数据增强方法，即联邦增强. 联邦增强可以提高联邦蒸馏降低的精度，而不会引起严重的通信开销. 联邦蒸馏是建立在无噪声且理想的通信信道假设基础上. 事实上，由于存在噪声和无线传输的叠加特性，无线通信链路给联邦蒸馏方法造成了新的挑战. Ahn 等人[52] 考虑无线网络环境下的联邦蒸馏实现，提出了一种基于分离信道编码和无线计算的混合联邦蒸馏(hybrid-federated distillation, HFD)方案. 该方案在信源信道编码中采用了带有误差累积的稀疏二进制压缩方法. 对于通过高斯多路访问通道进行的数字和模拟实现，HFD 可以在恶劣的通信环境中优于传统的 FL，这一原理与边缘人工智能模型自适应的降维和量化有一些共同之处，但 HFD 减少了数据传输源的特征尺寸，它为 FL 框架和数据编码的协同设计提供了新的研究思路. 表 4 给出了目前 FL 模型训练优化方法及特点.

Table 3 Summary of Model Compression Techniques

表 3 模型压缩技术总结

方法	优化手段	优缺点
结构化和草图更新机制[48]	压缩传输模型，提升客户端到服务器的通信效率	客户端到服务器参数压缩；代价是复杂的模型结构可能出现收敛问题
服务端-客户端更新[49]	压缩传输模型，提升服务器到客户端的通信效率	服务器到客户端参数压缩；代价是准确性降低，可能有收敛问题
草图[50]	使用计数草图压缩模型更新，然后利用草图的可合并性来组合来自客户端的模型更新	解决了客户端参与稀少而导致的收敛问题，建立在假设网络；已经尽了最大努力使通信效率最大化；可能遇到网络瓶颈
Adam[1]	通过使用 Adam 优化和压缩方案改进了 FedAvg 算法	Adam 优化加快了收敛速度，压缩方案降低了通信开销
模型蒸馏[51-52]	交换模型输出模型状态信息，即其有效载荷大小仅取决于输出维度的标签数量；然后使用联邦蒸馏实现权重更新规则	解决了数据独立同分布的问题；代价是无线信道对模型训练精度的影响

Table 4 Optimization Methods and Characteristics of Model Training

表 4 模型训练优化方法及特点

优化方法	特点	方法来源
数据卸载	利用边缘计算服务器的强大算力加快模型训练	文献 [36−38]
模型分割迁移	分割模型和隐私保护技术	文献 [42−44]
模型压缩	采用不同压缩粒度对模型输出值或者中间值梯度进行压缩	文献 [48−52]

4 无线网络下的模型更新技术

在 FL 过程中，模型更新过程主要涉及 FL 客户端本地更新过程和客户端向服务器更新上传模型参数过程，即全局聚合过程. 在模型更新过程中，每次局部更新消耗终端的计算资源，每次全局聚合消耗网络的通信资源. 消耗的资源量可能会随着时间的

推移而变化,并且全局聚合的频率、模型训练的准确性和资源消耗之间存在复杂的关系.因为人工智能模型的训练通常是资源密集型的,而学习任务的非优化操作可能会浪费大量的资源.现有的模型更新的研究工作主要从3个方面进行资源优化:1)通过全局聚合和本地更新两者之间的最优折中,保证在一定的资源预算下最小化模型的损失函数;2)优化梯度下降算法来降低通信资源开销;3)通过合理并动态地调整资源分配进一步降低对资源的盲目消耗.此外,由于FL中的模型更新严重依赖于网络,越来越多的研究致力于开发高效的无线通信FL技术,利用无线多址信道的叠加特性,以及无线资源优化技术来加速FL的全局模型更新过程.

4.1 自适应模型聚合技术

在端边协同的FL框架下,每个边缘节点执行梯度下降以调整局部模型参数,从而最小化在自己本地数据集上定义的损失函数.然后不同终端节点获得的模型参数被发送到参数聚合器,该参数聚合器可以是远程云、网络元素或边缘节点上允许的逻辑组件.参数聚合器对收到的参数进行全局聚合后,将更新后的参数发送回终端节点进行下一轮迭代训练[7].全局聚合频率可以根据一个或多个本地更新的间隔进行动态调整[53].文献[7]提出FedAvg模型聚合算法,将客户端上的本地SGD与执行模型平均的服务器相结合,显著减少模型聚合的通信次数.模型聚合算法协调全局模型参数的学习,它包含的异常机制确保了全局模型的收敛性[54]和异构客户端的公平性[55-56].文献[55]基于FedAvg提出了一个异构网络的联合优化框架,称为FedProx.FedProx通过设置一个修正项,使本地模型更接近全局模型,解决了不同设备的统计异质性问题.但是,FedProx未能正确配置和维护健壮的聚合算法,将使全局模型变得脆弱和不可信.目前聚合算法[57-58]在鲁棒性方面有广泛研究,这些算法可以在训练期间检测和丢弃错误或恶意的更新.此外,健壮的聚合方法应该能够承受通信不稳定性、客户端丢失、恶意参与者的错误模型聚合[59-62].然而,这些聚合算法大多数都没有考虑移动边缘网络下FL模型聚合所面临的问题.

考虑FL在边缘计算环境中的计算和通信资源受限的独特挑战,Wang等人[63]提出一种自适应模型聚合控制算法,并分析了具有非独立同分布数据的FL收敛界限.在当前资源受限状态下,这种自适应控制方案在全局模型聚合和局部模型更新之间提高了一种理想的折中,以最小化具有资源预算约束的损失函数.自适应模型聚合实质上是一种终端之间异步的非固定频率的模型聚合方式.图4给出了固定频率聚合和自适应聚合的区别固定频率聚合是为了固定全局聚合的频率,在固定的资源预算下最小化学习损失;自适应模型聚合是为了动态地调整全局聚合的频率,在固定的资源预算下最小化学习损失.文献[63]研究表明,在相同的时间预算内,自适应聚合方案在损失函数最小化和精度方面均优于终端之间同步的固定频率聚合方案,实现了计算资源和通信资源之间的权衡,降低了边缘服务器的负载.然而,文献[63]工作只是为了服务器的权衡,而不是考虑移动设备的资源限制.此外,在不可靠的网络下,传输的数据包丢失、移动设备突然断开连接等,也可能对FL产生不可预测的影响.文献[64]考虑客户端动态资源优化,为了有效地利用带宽资源,提出ACFL算法,ACFL可以根据网络条件自适应地压缩共享信

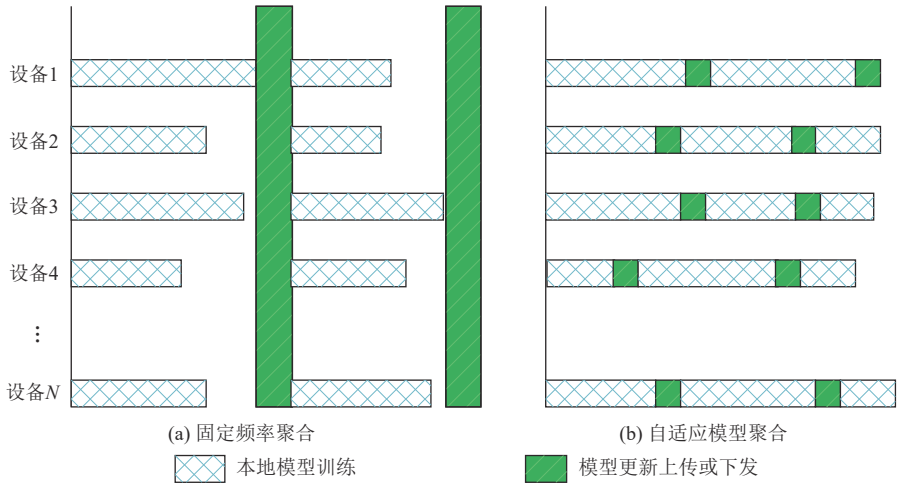

Fig. 4 Comparison of adaptive model aggregation and fixed frequency aggregation

图4 自适应模型聚合与固定频率聚合的比较

息. 此外, 在没有数据分布假设的情况下, 考虑了通信压缩和信息丢失, 分析了算法的收敛性. 针对基于云的 FL 模型的训练导致通信资源的显著开销, Liu 等人[65] 进一步将移动边缘计算平台作为中间结构, 提出一种基于客户端—边缘—云分层架构的联邦学习 (hierarchical federated averaging, HierFAVG)算法, 通过在边缘服务器和云服务器执行 2 级聚合, 为大量用户解决了与基于云的 FL 模型的训练相关的高通信资源消耗问题. 与传统的基于云的 FL 相比, 由于引入了边缘服务器聚合, HierFAVG 能更有效地利用通信资源. HierFAVG 是在移动边缘网络上实现 FL 的一种有效的方法, 它可以同时减少模型训练时间和终端设备的能量消耗.

尽管文献 [7, 53−65] 研究能够在资源受限的条件下优化模型聚合频率, 但忽略了不同终端节点的计算能力和链路通信能力的内在异质性. 这种异质性对优化不同学习者的任务分配、选择学习模型、提高学习精度、最小化局部和全局循环时间, 以及最小化能量消耗, 具有重要意义. 文献 [66] 建立了一个在相邻的异构无线边缘节点上高效执行分布式学习任务的优化框架, 首次协同分布式学习和分层移动边缘计算的新趋势, 提出了移动边缘学习概念框架. 通过考虑具有异构计算能力和到异构无线链接的边缘节点, 自适应地调整分布式学习的任务分配. 文献 [67] 考虑到移动边缘网络下异构设备通常在个性化的精度目标下被分配不同的任务, 提出了 CuFL 算法来加速 FL 过程, 同时确保所有终端设备都能满足其特定的任务要求. 为了进一步加快终端设备的本地模型训练, 提出了一个提前终止方案, 通过减少汇总轮次来缩短培训时间. 在早期终止方案中, 当终端设备满足精度要求时, 它们可以提前退出 FL 过程. 其结果是降低了能量成本, 并且剩余设备的通信资源是丰富的. 在 MEC 服务器端, 优化了全局聚合方法. 为了在 MEC 服务器上实现公平的参数聚合, 引入了一个公平系数来最小化当前精度和目标精度之间的差异. 从理论上严格分析了 CuFL 算法的收敛性, 还验证了 CuFL 在车辆分类任务中的有效性. 评价结果表明, CuFL 算法在准确率、训练时间和聚合过程的公平性方面具有优势. 表 5 列举了主要的 FL 模型局和技术.

Table 5　A Comparative Summary of Major Federated Learning Mode Aggregation Technologies

表 5　主要联邦学习模型聚合技术的比较总结

聚合技术	优化角度	主要思想	特点
FedAvg[7]	统计异构性	客户端对其本地数据执行多个批处理更新，并与服务器传输更新的权重，而不是梯度.	从统计的角度看，FedAvg 已被证明设备间数据分布不一致的情况下开始发散；从系统的角度看，FedAvg 不允许参与学习的设备根据其底层系统限制执行可变数量的本地更新.
FedProx[55]	统计异构性	在每个客户端上的本地训练子问题中添加一项，以限制每个本地模型更新对全局模型的影响.	FedProx 的提出是为了提高统计异质性数据的收敛性. 与 FedAvg 类似，在 FedProx 中，所有设备在全局聚合阶段的权重相等，因为没有考虑设备功能（例如硬件、电量）的差异.
FedPAQ[53]	通信	在与服务器共享更新之前，允许客户端在模型上执行多个本地更新.	与 FedAvg 类似，FedPAQ 中的新全局模型为局部模型的平均值，但这在强凸和非凸设置中都需要很高的复杂性.
FedMA[54]	统计异构性	在执行聚合前考虑神经元的排列不变性，并允许全局模型大小自适应.	使用贝叶斯非参数机制根据数据分布的异构性调整中心模型的大小；FedMA 中的贝叶斯非参数机制容易受到模型中毒攻击，在这种情况下，对手可以很容易地欺骗系统扩展全局模型，以适应任何中毒的本地模型.
Turbo-Aggregate[62]	通信和安全	一种多组策略，其中客户端被分成几个组，模型更新以循环方式在组之间共享和一种保护用户隐私数据的附加秘密共享机制.	Turbo-Aggregate 非常适合无线拓扑，在这种拓扑中，网络条件和用户可用性可能会快速变化. Turbo-Aggregate 中嵌入的安全聚合机制虽然能有效处理用户流失，但无法适应加入网络的新用户. 因此，通过重新配置系统规范（即多组结构和编码设置）以确保满足弹性和隐私保证，开发一种可自我配置的协议来扩展它的.
自适应聚合[63]	通信和统计异构性	在给定的资源预算下确定局部更新和全局参数聚合之间的最佳折中的自适应控制算法.	改变了全局聚合频率，以确保期望的模型性能，同时确保在 FL 训练过程中有效利用可用资源，例如能量，可用于边缘计算中的 FL. 自适应聚合方案的收敛性保证目前只考虑凸损失函数.
HierFAVG[65]	通信	一种分层的客户端—边缘—云聚合体系结构，边缘服务器聚合其客户端的模型更新，然后将它们发送到云服务器进行全局聚合.	这种多层结构能够在现有的客户端—云架构上实现更高效的模型交换. HierFAVG 仍然容易出现掉队和终端设备掉线的问题.
自适应任务分配[66]	设备异构性、通信、计算	在保证异构信道上的数据分发/聚合总次数和异构设备上的本地计算，在延迟约束下最大化学习精度.	自适应任务分配方案，该方案将最大化分布式学习者的本地学习迭代次数（从而提高学习精度），同时遵守时间限制. 该方案没考虑动态参数，如变化的信道状态和数据到达时间.
公平聚合[67]	设备异构性、任务异构性、通信、计算	一种具有自适应学习率的定制学习算法，以适应不同的精度要求，并加快本地训练过程. 为边缘服务器提出了一个公平的全局聚合策略，以最小化异构终端设备之间的精度差异.	一种学习率自适应的 CuFL 算法，以最小化总学习时间. 考虑到终端设备的任务异质性，CuFL 允许终端设备在满足其独特的精度要求后提前退出训练. 该方案没考虑动态参数，如变化的信道状态和数据到达时间.

4.2 梯度下降算法优化技术

通过调整模型聚合次数可以降低终端的计算资源. 在 FL 设置中, 快速的算法收敛同样可以减少通信轮数, 降低上传的梯度量也同样可以减少每轮更新的数据量, 从而降低通信资源开销[68].

在模型更新过程中, 每个终端根据其局部训练数据独立计算梯度, 对学习模型做出贡献. 现有的研究只利用了一阶梯度下降. 一阶梯度下降方法中每一次迭代只依赖于当前梯度, 并没有考虑到之前的迭代梯度更新可能加速训练的收敛[69]. 由于动量梯度法可以改善收敛性, 有许多研究工作将动量随机梯度下降应用于分布式机器学习领域. Liu 等人[69] 考虑与最后一次迭代相关的动量项, 提出动量 FL 系统, 并采用动量梯度下降的方法进行局部更新. 减轻 FL 系统中的通信负载问题已经被广泛研究, 主要是在无噪声、速率受限链路和星形拓扑的假设下进行. 这些解决方案的关键要素是压缩和降维操作, 这些操作将原始模型参数或梯度向量映射到由有限数量的位和/或稀疏性定义的表示中. 重要的解决方案类别包括无偏压缩[70-72] 和带有误差反馈机制的偏压压缩[73-76]. 一个众所周知的结合 SGD 和一致性的协议是分布式随机梯度下降(decentralized stochastic gradient descent, DSGD), 它已经通过梯度跟踪算法[77-78] 和减少代理之间大数据异质性的方差减少方案[79] 得到了进一步的扩展和改进. 此外, 在文献 [80–82] 中对空中计算(over-the-air computation, AirComp)进行了研究, 它是一种有前途的解决方案, 可通过利用无线介质的叠加特性来支持大规模 FL 中的同时传输. 与使用标准数字信号处理模块的传统实现相比, 基于模拟的 AirComp 直接从接收的基带样本中估计聚合统计. 文献 [80] 研究了有限带宽的高斯多址信道(multiple access channel, MAC)上的 FL, 并提出了新的数字和模拟 SGD. 在数字 SGD 中, 无线设备采用梯度量化和误差累计, 并通过 MAC 将它们的梯度估计传输到参数服务器, 模拟 SGD 利用无线媒体访问控制的加法性质进行空中梯度计算.

对于每一轮通信, 梯度量化减少了表示模型更新的位数, 从而有减少了分布式学习中的通信有效载荷大小. 由于量化引入了误差, 模型更新的算法精度降低, 这可能阻碍学习算法的收敛. 因此, 应该仔细设计量化[71, 83] 及其量化级, 以保证高精度的收敛性. Shokri 等人[84] 提出分布式选择性随机梯度下降(distributed selective stochastic gradient descent, DSSGD)方法, 依据不同参数或者不同特征对训练收敛的贡献不同, 选择性地进行梯度参数更新. DSSGD 方法达到了与传统 SGD 相当的精度, 但在每次学习迭代中更新的参数减少了 1~2 个数量级. 文献 [85] 的方法 Q-GADMM, 将随机量化与分组交替方向乘子法(group-based alternating direction method of multipliers, GADMM)[86] 的空间稀疏化相结合, 其中权重更新以概率 p 和 $1-p$ 分别向上和向下舍入, 而 p 被自适应地调整以最小化通信成本, 同时保证 GADMM 收敛. L-FGADMM[87] 对 GADMM 应用分层联邦, 而不像在 Q-GADMM 中那样量化, L-FGADMM中的节点分为头尾组, 只与邻近的节点交流. 与 GADMM 相比, L-FGADMM 通过 2 种方式进一步提高通信效率. 首先, 与 GADMM 中的每次迭代通信不同, L-FGADMM 中的节点定期进行通信. 其次, 针对每一层分别调整 L-FGADMM 的通信周期, 与交换整个模型的 GADMM 不同, L-FGADMM 可以增加大规模层的通信周期, 同时减小通信有效载荷的大小.

考虑到同步训练会丢弃模型更新后到达的所有延迟结果, 从而浪费相应设备的电池电量和它们潜在的有用数据. 因此, 现有研究采用异步更新取代了标准 FL 的同步方案. 然而, 异步更新带来了梯度值过时的问题, 因为多个用户可以在任意时间自由地执行学习任务, 当在过时的模型版本上计算学习任务时, 会出现过时的结果, 与此同时, 全局模型已经发展到一个新版本, 过时的结果会给训练过程增加噪声, 减缓甚至阻止 FL 模型收敛[88]. 考虑到这些问题, 文献 [22] 提出了新的 SGD 算法, 即 ADASGD. 提出了一种预测移动设备上每个学习任务的计算时间和能耗的分析工具, 用于防止在服务器聚集本地模型的截止日期之后出现不必要的计算. 文献 [22, 88] 的方法在加速收敛方面有一定的优势, 但是它们并没有考虑到移动边缘智能场景下的独特挑战.

为适配边缘计算场景, Tao 等人[89] 提出边缘随机梯度下降(edge stochastic gradient descent, eSGD)算法, 在梯度下降过程中, 某些参数对神经网络的目标函数贡献更大, 因此在给定的训练迭代过程中会经历更大的更新. 梯度值取决于训练样本, 并且随样本的不同而变化. 此外, 输入数据的某些特征比其他特征更为重要, 而帮助计算这些特征的参数在学习过程中更为关键, 并经历更大的变化. 因此, eSGD 算法只选择一小部分重要梯度在每一轮通信过程中与 FL 聚合服务器进行更新. 与标准 SGD 方法相比, eSGD 仍然存在精度损失. 在 Tao 等人[89] 研究梯度的选择性通信的同时, Wang 等人[90] 提出了 CMFL 算法. 该

算法保证了只上传相关的局部模型更新，以降低通信成本，同时保证全局收敛. 在每次迭代中，首先将参与者的本地更新与全局更新进行比较，以确定更新是否相关. 通过消除不相关的、损害训练的异常更新，CMFL 可以获得稍高的精度. 文献 [91] 研究了边缘网络的 DSGD 实现问题. 通过考虑数字和模拟传输方案，提出了在无线 D2D 网络上实现 DSGD 的协议，模拟实现利用 AirComp. 为了应对无线干扰，将基于图着色的调度策略应用到数字和模拟实现的设计中.

边缘智能环境下的 FL 与网络通信技术的发展密切相关，在第 2 节和第 3 节讨论的研究工作中，大多忽略了无线通信链路的特性. 无线通信链路承载了 FL 的参数更新过程. 无线链路的资源分配也是智能边缘系统中 FL 优化的一个重点方向.

4.3 无线资源优化技术

通常来说，在移动边缘网络环境下的 FL 是动态的、不确定的，具有时变的约束条件. 基于无线网络实现 FL 架构，客户端必须通过无线链路传输其本地训练结果，目前 FL 有很多无线应用，例如：无人机（unmanned aerial vehicle, UAV）[4-5, 92-99]、车联网 [82, 100-104] 和目标定位 [104] 等.

1）FL 在无人机系统中的应用研究. 表 6 描述了边缘网络下无人机 FL 应用主要组件，如客户端、服务器和数据、FL 的预期结果. 无人机可以作为边缘内容缓存，这种范式的主要挑战是通过预测无人机内容的流行度来有效地确定每个缓存中应该存储哪些内容. 然而，这需要直接访问私人无人机信息，以进行内容区分，这在实践中是不可能的. FL 是基于内容流行度预测天然的匹配方案，因为它支持本地训练模型，从而保护用户数据隐私. 例如，增强现实应用程序需要访问用户的隐私数据，以便获得增强的流行元素 [105].

Table 6 Unmanned Aerial Vehicle Application Based on Federated Learning in Edge Network

表 6 边缘网络下基于联邦学习的无人机应用

挑战	联邦学习				结果
	客户端	服务器	数据特征	本地和全局模型	
边缘内容缓存 [95-96]	UAVs	边缘服务器	内容特征（新鲜度、位置、占用内存、内容请求历史等）	内容受欢迎度预测	有效地确定哪些内容应该存储在每个缓存中
无人机作为基站 [93]	地面用户	边缘服务器	关于地面用户可移动性的信息（位置、方向、速度等）	地面用户模式（移动性和内容负荷）的预测	优化无人机基站部署、提高网络覆盖和连通性、有效提供热门内容.
无人机轨迹规划 [92]	UAVs	边缘服务器或云	源、目的点位置、无人机机动性信息（速度、方向、位置、高度等）、无人机能量消耗、物理障碍、服务需求等.	每条潜在路径的性能预测	无人机选择最优轨迹、优化服务性能、优化无人机能耗

无人机由于其固有的属性，如机动性、灵活性和自适应高度，一方面，无人机可以用作空中基站 [93]，无人机基站可以有效地补充现有的蜂窝系统，为热点地区提供额外的容量，并在难以到达的偏远地区提供网络覆盖，以提高无线网络的覆盖范围、容量、可靠性和能效. 另一方面，无人机可以在蜂窝网络中作为飞行移动终端运行，这种蜂窝网络连接的无人机可以实现视频流、物品交付等多种应用. 与传统的地面基站相比，使用无人机作为飞行基站的优势是能够调整高度、避开障碍物，并提高与地面用户建立视影通信链路的可能性 [106].

装有不同类型传感器（如摄像机、全球定位系统和湿度传感器）的无人机通过收集周围环境的传感数据来执行传感任务. 由于风和其他随机因素，大规模无人机控制变得具有挑战性，以避免碰撞并快速到达目的地. 基于无线网络的 FL 可以实现对无人机机群的飞行路线控制 [92].

2）FL 在车联网中的应用研究. 图 5 显示了智能交通下的 FL 用例. 文献 [100] 研究了车联网中超可靠低时延通信的联合功率和资源分配问题，FL 用于估计反映网络状态的网络范围队列长度的尾部分布. 文献 [101] 讨论了车联网中使用 FL 进行图像分类的问题. 车辆客户端配备有各种传感器来捕获图像，通过考虑局部图像质量和每辆车的计算能力，引入选择性模型聚集方法来选择在车辆处计算的局部模型. 考虑到无线资源的有效利用和低时延，在车辆附近进行学习是很重要的，为了将 FL 应用于分散网络，可以结合车辆聚类的方法，即选择一些车辆作为 FL 的参数服务器. 在文献 [107] 中提出了将一种联合分配发射功率和资源分配方法用于在车辆网络中实现超可靠的低时延通信. 在传统的同步 FL 中，每个车辆从服务器获取全局模型，并将更新推送到服务器. 然后，服务器同步所有更新，并将更新聚合到全局模型中. 同步学习会导致较高的通信成本，同时还会导致等待较慢节点的空闲时间较长. 一些研究探索了异步学习机制以提高学习效率. 例如，文献 [108]

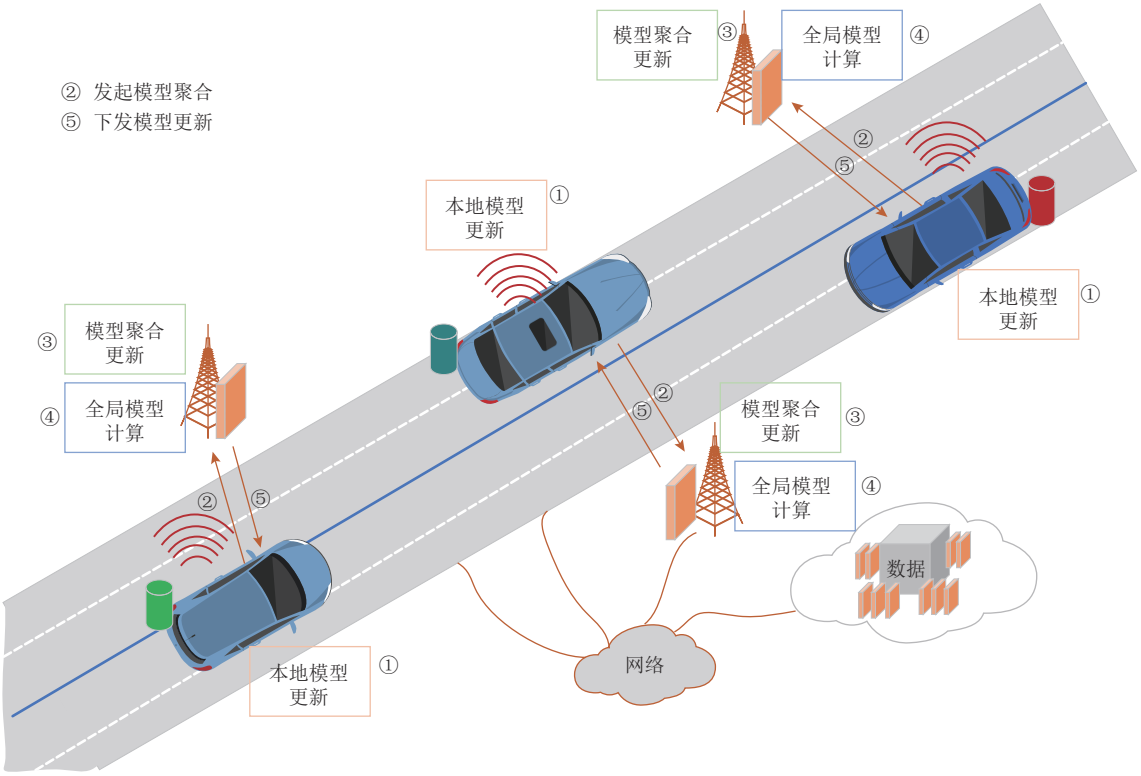

Fig. 5 Intelligent transportation

图 5 智能交通

中提出了一种异步小批量算法，利用多个处理器来解决正则化随机优化问题. 为了提高 FL 的效率，文献 [109] 提出一种基于节点选择和异步聚合算法的异步 FL 方案，为了提高共享数据的可靠性，通过将学习的模型集成到区块链并执行 2 阶段验证. 文献 [4] 提出一种基于 FL 的无人机飞行自组网干扰攻击检测机制，基于 Dempster-Shafer 理论的客户端选择方法可以提高 FL 的学习效率.

3)FL 在目标定位方面的研究. 在新冠肺炎大流行期间，需要共享跟踪感染传播和预测高风险区域的同时，迫切需要保护移动用户的位置隐私. FL 框架是一个出色的解决方案，可以提高无线定位的准确性，同时保护用户之间的安全合作. 文献 [110] 中使用 FL 来训练机器学习模型进行本地化，称为联邦本地化. 作为本地客户端，每个移动设备收集关于无线电特征和位置的本地数据，在本地更新模型参数集，并将其发送到中央服务器. 基站或聚合中心作为中心服务器，将接收到的局部参数集合进行聚合，得到全局参数集合. 在对 2 种机器学习模型进行局部化比较之后，基于真实数据的测试，具有最大似然损失函数的高斯过程模型优于具有最小二乘损失函数的神经网络模型.

4)FL 在无线环境下存在巨大的应用需求. 但由于无线通信资源有限，这可能会影响 FL 的性能. 因此，有必要根据模型更新的上下文信息来调整资源分配优化 FL. FL 模型更新时间包括用户设备计算的时间(取决于用户设备的 CPU 类型和本地数据集大小)，还包括所有用户设备的通信时间(取决于用户设备信道增益和更新数据集大小). 由于参与者的电池电量有限，如何分配用户设备资源(如计算和传输功率)以最小化能耗是主要关注的问题. 即如何在最小化 FL 模型更新时间和用户设备能耗这 2 个相互冲突的目标之间取得平衡？为此，Merluzzi 等人[111] 提出面向边缘学习的一种通信和计算资源分配的动态调整策略，探索系统能耗、系统服务延迟和学习精度之间的最佳权衡. 这种方法为确保在特定应用程序所施加的指定延迟约束内保证 FL 精度的方法铺平了道路. 相似地，Yang 等人[112] 考虑本地计算和无线传输的时延和总消耗能量的折中，提出了一种低复杂度的迭代学习算法. 在该算法的每一步，都得到了时间分配、带宽分配、功率控制、计算频率和学习精度的新的闭式解，解决了一个以完成时间和总消耗能量的最小加权和为目标的联合传输和计算的优化问题. 面向边缘智能环境下天然的端边云应用场景，Luo 等人[113] 提出一种端边云分层的联邦边缘学习框架，制定了一个整体联邦计算、通信资源分配和边缘

关联的模型用于全局学习成本最小化，该框架在低延迟和高能效的 FL 中具有巨大的潜力. Abad 等人[114]进一步考虑在异构蜂窝网络中实现联邦边缘学习，利用梯度稀疏化提出了一种优化的同步梯度更新资源分配方案来最小化训练的延迟.

文献 [111–114] 主要从整体训练的角度来进行无线资源分配，而文献 [115–118] 则从用户调度的角度实现了用户设备偏好的资源分配. 文献 [115] 将重要度感知的无线资源管理的设计原则应用于改进用户调度，根据信道状态和数据统计对模型训练的重要性，将无线资源分配给终端设备. 文献 [116] 通过降低训练组中速度较快的移动设备的 CPU 循环频率来提高 FL 的能量效率. 文献 [117] 为了降低设备的能量消耗，提出高效的带宽分配和调度策略，导出的调度优先权函数能适应设备的信道状态和计算能力，为具有较好信道状态和计算能力的设备提供了偏好. 文献 [118] 提出一种概率用户选择方案，选择本地模型对基站连接以及全局模型具有高概率影响的用户，为他们分配上行资源块. 文献 [119–120] 重点关注了通信资源的分配. 文献 [119] 则对全局聚合的通信资源分配和局部更新模型参数的计算资源分配进行了联合优化. 特别地，分别基于非正交多址和时分多址，提出了 2 种用于边缘设备向边缘服务器上传机器学习参数的传输协议. 在这 2 种协议下，通过联合优化全局聚合上传参数过程中的终端设备传输功率和速率以及本地更新过程中的 CPU 频率，从而在有限时间内最小化所有终端设备的总能量消耗. 文献 [120] 提出了对数据批量大小和无线资源的优化来加速 FL.

文献 [111–120] 研究工作在随机梯度下降算法的基础上，侧重于增加时间和能耗的约束来进行资源分配，主要通过构建能耗模型来优化能效，或者从无线资源管理的角度对一些设备状态较好的客户端进行偏好设置. 文献 [121–122] 通过引入深度强化学习技术与动态环境的交互，来优化模型训练的资源分配. Anh 等人[121]提出一个以训练服务器为主体，状态空间包括移动设备的 CPU 和能量状态，动作空间包括从移动设备获取的数据单元和能量单元的数量的随机优化问题. 奖励被定义为累积数据、能量消耗和训练延迟的函数，然后采用双深度 Q 网络来解决该优化的问题. 作为对文献 [121] 的扩展，文献 [122] 提出一种使用深度强化学习的资源分配方法，考虑了 FL 参与者的移动性. 在没有移动网络先验知识的情况下，FL 参数服务器能够优化参与者之间的资源分配. 类似地，文献 [123] 也提出了通过 D2D 通信结合 FL 来构建 D2D-FedAvg 算法. 该算法利用状态较好的设备作为 D2D 学习组的簇头，从无线资源的角度降低 FL 蜂窝网络的通信负载.

尽管移动设备的计算能力迅速增长，但许多设备仍然面临无线资源短缺的问题. 针对这个问题，越来越多的研究致力于开发面向 FL 的高效无线通信技术[73, 80, 114]. Zhu 等人[73]研究了宽带无线衰落 MAC 上的 FL，其中设备在完全了解信道状态信息（channel state information, CSI）的情况下执行信道反转，以在参数服务器处对齐它们的信号，并提出一种用于无线网络 FL 的多址宽带模拟聚合（broadband analog aggregation, BAA），以减少 FL 中的通信延迟，而不是在服务器的全局聚合期间分别执行通信和计算，BAA 方案基于空中计算的概念，通过利用多址信道的信号叠加特性来集成计算和通信. BAA 方案允许整个带宽的重用，而传统的正交频分多址是正交化带宽分配. 文献 [73] 的研究表明，BAA 方案可以达到与正交频分多址方案相当的测试精度，同时降低延迟 10~1 000 倍. Amiri 等人[74]进一步扩展，在空中计算基础上引入了误差积累和梯度稀疏化，能更有效地利用带宽，显著降低通信负载，同时可以获得比空中计算更高的测试精度. 与文献 [74] 相似，文献 [124] 针对 AirComp 过程中产生的聚集误差会导致模型精度下降的问题，提出一种参与者选择算法用于训练的设备数量最大化，以提高统计学习性能，同时将信号失真保持在一定的阈值以下. 图 6 展示了该算法的原理.

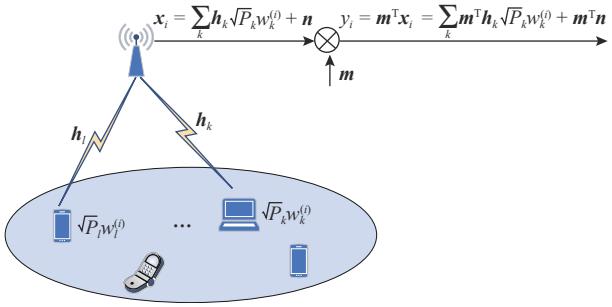

Fig. 6 The parameters are aggregated by air calculation and spatial freedom[125]

图 6 通过空中计算并利用空间自由度进行参数聚合[125]

由图 6 可知，本地客户端通过无线信道同时发送本地参数集的第i个元素，这些元素具有功率比例$\sqrt{P_k} w_k^{(i)}, k = 1, 2, \cdots, K$，则在中央服务器接收的信号向量将是$\boldsymbol{x}_i = \sum_k \boldsymbol{h}_k \sqrt{P_k} w_k^{(i)} + \boldsymbol{n}$，其中$\boldsymbol{h}_k$和$\boldsymbol{n}$分别对应于本地客户端$k$的信道增益向量和噪声向量. 之后结合

波束形成向量$\boldsymbol{m}$得到$y_i=\boldsymbol{m}^{\mathrm{T}}\boldsymbol{x}_i=\boldsymbol{m}^{\mathrm{T}}\sum_k \boldsymbol{h}_k\sqrt{P_k}w_k^{(i)}+\boldsymbol{m}^{\mathrm{T}}\boldsymbol{n}$, 这里的波束形成矢量为高效传输提供了自由度. 如果每个本地客户端没有噪声或最大功率限制, 则应选择$\boldsymbol{m}$和P_k, 以使$\boldsymbol{m}^{\mathrm{T}}\boldsymbol{h}_k\sum_k\sqrt{P_k}=\frac{|D_k|}{\sum_l|D_l|}$根据式(4)进行聚合. 实际上, 由于信道失真和每个本地客户端的最大传输功率限制, 尤其当存在大量的客户端时, 可能没有足够的自由度来实现聚合. 文献[124]提出了一个稀疏低秩化问题来解决这个问题. 在文献[124–126]中分别介绍了多天线参数服务器处的波束形成技术, 用于增加参与设备的数量和克服设备处的CSI匮乏. 文献[127]研究了无线信道上的FL的跨设备的资源分配, 设备的参与频率作为设备调度度量标准引入[128]. 文献[129–131]提供了在各种资源分配方案下无线网络上的FL的收敛性分析. 文献[132]考虑在资源有限的块衰落无线网络中从边缘设备到基站的数字传输, 设计了新的设备调度策略和跨设备的资源分配, 以执行正交(无干扰)传输. 数值结果说明了在设备调度中同时考虑信道条件和本地模型更新的优势, 而不是基于2个度量中的任何一个单独进行调度.

上述资源优化方法的目标是提高FL的训练效率, 然而, 这可能会导致一些设备因资源有限而被排除在聚合阶段之外. 这种不公平的资源分配的一个后果是FL模型将被拥有更高计算能力设备的参与者所拥有的数据的分布所过度代表. 因此, 面向FL的无线资源分配还需要结合数据统计分布特性进一步优化.

5 结论与展望

从目前边缘智能FL的研究现状可见, FL在客户端选择、模型训练与模型更新等方面都取得了大量的进展, 基本能够满足边缘智能应用的实际需求. 但是面向未来大规模多样化的边缘智能应用, FL技术还存在很大的发展空间.

1)FL过程需要更细粒度的隐私保护. 目前的FL架构采用了差分隐私[84]或者多方安全计算[133]等技术来实现模型聚合传递参数的隐私保护. 这些技术能够提供系统全局粒度的隐私保护. 在未来的边缘智能应用中, 异构终端、异构网络、异构数据等天然的异构应用环境需要更细粒度的隐私保护方法. 例如不同设备之间、不同样本集合之间需要不同粒度的隐私保护. 设计不同粒度混合的隐私保护方法是边缘智能FL技术的一个发展方向.

2)FL需要与无线网络深度融合[39, 41], 提升学习收敛速度. FL能够大规模实际应用的一个重要方面是学习算法在有限的通信和计算资源下能够快速收敛. 为实现该目标, FL除了算法方面的优化, 还需要网络技术的协同优化来解决资源受限问题. 目前, 分布式边缘智能应用需求已经驱动了无线通信技术与网络架构的革新与发展. 未来面向6G无线通信系统, FL技术需要更紧密地与无线通信技术耦合, 享受无线通信技术发展带来的红利, 实现AirComp与空口通信的有机融合, 进而突破通信与计算资源对学习性能的限制.

3)FL需要结合迁移学习、强化学习等技术, 满足边缘智能应用的多样化需求. 迁移学习与强化学习已经取得了长足的进步. 在实际应用中, FL各个参与方可能只有少量的标注数据, 而且数据在统计上可能高度异构. 为了帮助只有少量数据和弱监督的应用建立有效且精确的机器学习模型, 并且不违背用户的数据隐私原则, FL可以与迁移学习结合, 形成联邦迁移学习, 以适用于更广的业务范围. 同样, FL可以对分布式强化学习进行扩展, 形成强化学习的隐私保护版本——联邦强化学习, 解决边缘智能环境下的序列决策问题.

4)FL需要有效的参与激励机制. FL目前的大多数研究侧重于提升性能, 但忽略了学习参与者的意愿问题. 在边缘智能应用环境下, 如何鼓励数据拥有者积极参与联邦训练是一个非常现实的问题. 特别是如何刻画数据质量, 并激励拥有高质量数据的客户端参与FL是未来需要深入探索的一个潜在方向.

作者贡献声明: 张雪晴负责论文的整体文献调研、整理及撰写; 刘延伟辅助调研、提出论文整体思路、设计全文框架和审核最终论文; 刘金霞、韩言妮对论文结构与内容进行讨论、修改, 并提出了指导意见.

参 考 文 献

[1] Mills J, Hu Jia, Min Geyong. Communication-efficient federated learning for wireless edge intelligence in IoT[J]. IEEE Internet of Things Journal, 2019, 7(7): 5986−5994

[2] Covington P, Adams J, Sargin E. Deep neural networks for YouTube recommendations[C] //Proc of the 10th ACM Conf on Recommender Systems. New York: ACM, 2016: 191−198

[3] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[C] //Proc of the 15th IEEE Int Conf on Computer Vision Workshop.

Piscataway, NJ: IEEE, 2015: 258−266

[4] Mowla N I, Tran N H, Doh I, et al. Federated learning-based cognitive detection of jamming attack in flying ad-hoc network[J]. IEEE Access, 2020, 8: 4338−4350

[5] Brik B, Ksentini A, Bouaziz M. Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems[J]. IEEE Access, 2020, 8: 53841−53849

[6] Abbas N, Zhang Yan, Taherkordi A, et al. Mobile edge computing: A survey[J]. IEEE Internet of Things Journal, 2017, 5(1): 450−465

[7] 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. New York: PMLR, 2017 : 1273−1282.

[8] Yang Qiang, Liu Yang, Chen Tianjian, et al. Federated machine learning: Concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1−19

[9] Zhou Zhi, Yang Song, Pu Lingjun, et al. CEFL: Online admission control, data scheduling, and accuracy tuning for cost-efficient federated learning across edge nodes[J]. IEEE Internet of Things Journal, 2020, 7(10): 9341−9356

[10] Ruder S. An overview of gradient descent optimization algorithms[J]. arXiv preprint, arXiv: 1609.04747, 2016

[11] Lim W Y B, Luong N C, Hoang D T, et al. Federated learning in mobile edge networks: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2020, 22(3): 2031−2063

[12] Li Tian, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50−60

[13] Li Qinbin, Wen Zeyi, Wu Zhaomin, et al. A survey on federated learning systems: Vision, hype and reality for data privacy and protection[J]. arXiv preprint, arXiv: 1907.09693, 2019

[14] Wang Xiaofei, Han Yiwen, Wang Chenyang, et al. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning[J]. IEEE Network, 2019, 33(5): 156−165

[15] Kairouz P, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning[J]. arXiv preprint, arXiv: 1912.04977, 2019

[16] Wang Yan, Li Nianshuang, Wang Xiling, et al. Coding-based performance improvement of distributed machine learning in large-scale clusters[J]. Journal of Computer Research and Development, 2020, 57(3): 542−561 (in Chinese)
(王艳, 李念爽, 王希龄, 等. 编码技术改进大规模分布式机器学习性能综述[J]. 计算机研究与发展, 2020, 57(3): 542−561)

[17] Jin Yibo, Jiao Lei, Qian Zhuzhong, et al. Resource-efficient and convergence-preserving online participant selection in federated learning[C] //Proc of the 40th IEEE Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE, 2020: 606−616

[18] Chai Z, Ali A, Zawad S, et al. TiFL: A tier-based federated learning system[C] //Proc of the 29th Int Symp on High-Performance Parallel and Distributed Computing. New York: ACM, 2020: 125−136

[19] Li Li, Xiong Haoyi, Guo Zhishan, et al. SmartPC: Hierarchical pace control in real-time federated learning system[C] //Proc of the 40th IEEE Real-Time Systems Symp (RTSS). Piscataway, NJ: IEEE, 2019: 406−418

[20] Khan L U, Alsenwi M, Han Zhu, et al. Self organizing federated learning over wireless networks: A socially aware clustering approach[C] //Proc of the 34th Int Conf on Information Networking (ICOIN). Piscataway, NJ: IEEE, 2020: 453−458

[21] Xu Jie, Wang Heqiang. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective[J]. IEEE Transactions on Wireless Communications, 2020, 20(2): 1188−1200

[22] Damaskinos G, Guerraoui R, Kermarrec A M, et al. Fleet: Online federated learning via staleness awareness and performance prediction[C] //Proc of the 21st Int Middleware Conf. New York: ACM, 2020: 163−177

[23] Sprague M R, Jalalirad A, Scavuzzo M, et al. Asynchronous federated learning for geospatial applications[C] //Proc of the Joint European Conf on Machine Learning and Knowledge Discovery in Databases. Cham, Switzerland: Springer, 2018: 21−28

[24] Wu Wentai, He Ligang, Lin Weiwei, et al. Safa: A semi-asynchronous protocol for fast federated learning with low overhead[J]. IEEE Transactions on Computers, 2020, 70(5): 655−668

[25] Nishio T, Yonetani R. Client selection for federated learning with heterogeneous resources in mobile edge[C/OL] //Proc of the 53rd IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2019[2022-09-05].https://ieeexplore.ieee.org/document/8761315

[26] Yoshida N, Nishio T, Morikura M, et al. Hybrid-FL for wireless networks: Cooperative learning mechanism using non-IID data[C/OL] //Proc of the 54th IEEE Int Conf on Communications (ICC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9149323

[27] Khan L U, Pandey S R, Tran N H, et al. Federated learning for edge networks: Resource optimization and incentive mechanism[J]. IEEE Communications Magazine, 2020, 58(10): 88−93

[28] Kang Jiawen, Xiong Zehui, Niyato D, et al. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory[J]. IEEE Internet of Things Journal, 2019, 6(6): 10700−10714

[29] Kim H, Park J, Bennis M, et al. Blockchained on-device federated learning[J]. IEEE Communications Letters, 2019, 24(6): 1279−1283

[30] Li Tian, Sanjabi M, Beirami A, et al. Fair resource allocation in federated learning[J]. arXiv preprint, arXiv: 1905.10497, 2020

[31] Pandey S R, Tran N H, Bennis M, et al. A crowdsourcing framework for on-device federated learning[J]. IEEE Transactions on Wireless Communications, 2020, 19(5): 3241−3256

[32] Le T H T, Tran N H, Tun Y K, et al. Auction based incentive design for efficient federated learning in cellular wireless networks[C/OL] // Proc of the IEEE Wireless Communications and Networking Conf (WCNC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9120773

[33] Jiao Yutao, Wang Ping, Niyato D, et al. Toward an automated

auction framework for wireless federated learning services market[J]. IEEE Transactions on mobile Computing, 2020, 20(10): 3034−3048

[34] Gao Xiaozheng, Wang Ping, Niyato D, et al. Auction-based time scheduling for backscatter-aided RF-powered cognitive radio networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(3): 1684−1697

[35] Ko BongJun, Wang Shiqiang, He Ting, et al. On data summarization for machine learning in multi-organization federations[C] //Proc of the 7th IEEE Int Conf on Smart Computing (SMARTCOMP). Piscataway, NJ: IEEE, 2019: 63−68

[36] Valerio L, Passarella A, Conti M. Optimal trade-off between accuracy and network cost of distributed learning in mobile edge Computing: An analytical approach[C/OL] //Proc of the 18th Int Symp on a World of Wireless, Mobile and Multimedia Networks (WoWMoM). Piscataway, NJ: IEEE, 2017[2022-09-05].https://ieeexplore.ieee.org/abstract/document/7974310

[37] Skatchkovsky N, Simeone O. Optimizing pipelined computation and communication for latency-constrained edge learning[J]. IEEE Communications Letters, 2019, 23(9): 1542−1546

[38] Huang Yutao, Zhu Yifei, Fan Xiaoyi, et al. Task scheduling with optimized transmission time in collaborative cloud-edge learning[C/OL] //Proc of the 27th Int Conf on Computer Communication and Networks (ICCCN). Piscataway, NJ: IEEE, 2018[2022-09-05].https://ieeexplore.ieee.org/abstract/document/8487352

[39] Dey S, Mukherjee A, Pal A, et al. Partitioning of CNN models for execution on fog devices[C] //Proc of the 1st ACM Int Workshop on Smart Cities and Fog Computing. New York: ACM, 2018: 19−24

[40] Zhang Shigeng, Li Yinggang, Liu Xuan, et al. Towards real-time cooperative deep inference over the cloud and edge end devices[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2020, 4(2): 1−24

[41] Dey S, Mukherjee A, Pal A. Embedded deep inference in practice: Case for model partitioning[C] //Proc of the 1st Workshop on Machine Learning on Edge in Sensor Systems. New York: ACM, 2019: 25−30

[42] Lin Bing, Huang Yinhao, Zhang Jianshan, et al. Cost-driven off-loading for DNN-based applications over cloud, edge, and end devices[J]. IEEE Transactions on Industrial Informatics, 2019, 16(8): 5456−5466

[43] Wang Lingdong, Xiang Liyao, Xu Jiayu, et al. Context-aware deep model compression for edge cloud computing[C] //Proc of the 40th Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE, 2020: 787−797

[44] Wang Ji, Zhang Jianguo, Bao Weidong, et al. Not just privacy: Improving performance of private deep learning in mobile cloud[C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2407−2416

[45] Zhang Jiale, Wang Junyu, Zhao Yanchao, et al. An efficient federated learning scheme with differential privacy in mobile edge computing[C] //Proc of the Int Conf on Machine Learning and Intelligent Communications. Berlin: Springer, 2019: 538−550

[46] Ivkin N, Rothchild D, Ullah E, et al. Communication-efficient distributed SGD with sketching[J]. Advances in Neural Information Processing Systems, 2019, 32: 13144−13154

[47] Zhang Boyu, Davoodi A, Hu Yuhen. Exploring energy and accuracy tradeoff in structure simplification of trained deep neural networks[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2018, 8(4): 836−84

[48] Konen J, Mcmahan H B, Yu F X, et al. Federated learning: Strategies for improving communication efficiency[J]. arXiv preprint, arXiv: 1610.05492, 2016

[49] Caldas S, Konečny J, Mcmahan H B, et al. Expanding the reach of federated learning by reducing client resource requirements[J]. arXiv preprint, arXiv: 1812.07210, 2018

[50] Rothchild D, Panda A, Ullah E, et al. FetchSGD: Communication-efficient federated learning with sketching[C] //Proc of the 37th Int Conf on Machine Learning. New York: PMLR, 2020: 8253−8265

[51] Jeong E, Oh S, Kim H, et al. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data[J]. arXiv preprint, arXiv: 1811.11479, 2018

[52] Ahn J H, Simeone O, Kang J. Wireless federated distillation for distributed edge learning with heterogeneous data[C/OL] //Proc of the 30th Annual Int Symp on Personal, Indoor and Mobile Radio Communications (PIMRC). Piscataway, NJ: IEEE, 2019[2022-09-05]. https://ieeexplore.ieee.org/abstract/document/8904164

[53] Reisizadeh A, Mokhtari A, Hassani H, et al. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization[C] //Proc of the 23rd Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2021−2031

[54] Karimireddy S P, Kale S, Mohri M, et al. SCAFFOLD: Stochastic controlled averaging for federated learning[C] //Proc of the 37th Int Conf on Machine Learning. New York: PMLR, 2020: 5132−5143

[55] Li Tian, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine Learning and Systems, 2020, 2: 429−450

[56] Wang Hongyi, Yurochkin M, Sun Yuekai, et al. Federated learning with matched averaging[J]. arXiv preprint, arXiv: 2002.06440, 2020

[57] Pillutla K, Kakade S M, Harchaoui Z. Robust aggregation for federated learning[J]. IEEE Transactions on Signal Processing, 2022, 70: 1142−1154

[58] Grama M, Musat M, Muñoz-González L, et al. Robust aggregation for adaptive privacy preserving federated learning in healthcare[J]. arXiv preprint, arXiv: 2009.08294, 2020

[59] Ang Fan, Chen Li, Zhao Nan, et al. Robust federated learning with noisy communication[J]. IEEE Transactions on Communications, 2020, 68(6): 3452−3464

[60] Lu Yanyang, Fan Lei. An efficient and robust aggregation algorithm for learning federated CNN[C/OL] //Proc of the 3rd Int Conf on Signal Processing and Machine Learning. New York: ACM, 2020[2022-09-05].https://dl.acm.org/doi/abs/10.1145/3432291.3432303

[61] Chen Zhou, Lv Na, Liu Pengfei, et al. Intrusion detection for wireless

edge networks based on federated learning[J]. IEEE Access, 2020, 8: 217463−217472

[62] So J, Güler B, Avestimehr A S. Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning[J]. IEEE Journal on Selected Areas in Information Theory, 2021, 2(1): 479−489

[63] Wang Shiqiang, Tuor T, Salonidis T, et al. Adaptive federated learning in resource constrained edge computing systems[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1205−1221

[64] Zhang Xiongtao, Zhu Xiaomin, Wang Ji, et al. Federated learning with adaptive communication compression under dynamic bandwidth and unreliable networks[J]. Information Sciences, 2020, 540(5): 242−262

[65] Liu Lumin, Zhang Jun, Song Shenghui, et al. Client-edge-cloud hierarchical federated learning[C/OL] //Proc of the 54th IEEE Int Conf on Communications (ICC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9148862

[66] Mohammad U, Sorour S. Adaptive task allocation for mobile edge learning[C/OL] //Proc of the Wireless Communications and Networking Conf Workshop (WCNCW). Piscataway, NJ: IEEE, 2019[2022-09-05].https://ieeexplore.ieee.org/abstract/document/8902527

[67] Jiang Hui, Liu Min, Yang Bo, et al. Customized federated learning for accelerated edge computing with heterogeneous task targets[J]. Computer Networks, 2020, 183(12): 107569−107569

[68] Lin Yujun, Han Song, Mao Huizi, et al. Deep gradient compression: Reducing the communication bandwidth for distributed training[J]. arXiv preprint, arXiv: 1712.01887, 2017

[69] Liu Wei, Chen Li, Chen Yunfei, et al. Accelerating federated learning via momentum gradient descent[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(8): 1754−1766

[70] Abdi A, Saidutta Y M, Fekri F. Analog compression and communication for federated learning over wireless MAC[C/OL] // Proc of the 21st Int Workshop on Signal Processing Advances in Wireless Communications (SPAWC). Piscataway, NJ: IEEE, 2020[2022-09-05]. https://ieeexplore.ieee.org/abstract/document/9154309

[71] Alistarh D, Grubic D, Li J, et al. QSGD: Communication-efficient SGD via gradient quantization and encoding[J]. Advances in Neural Information Processing Systems, 2017, 30: 1709−1720

[72] Bernstein J, Wang Yuxiang, Azizzadenesheli K, et al. signSGD: Compressed optimisation for non-convex problems[C] //Proc of the 35th Int Conf on Machine Learning. New York: PMLR, 2018: 560−569

[73] Zhu Guangxu, Wang Yong, Huang Kaibin. Broadband analog aggregation for low-latency federated edge learning[J]. IEEE Transactions on Wireless Communications, 2019, 19(1): 491−506

[74] Amiri M M, Gündüz D. Federated learning over wireless fading channels[J]. IEEE Transactions on Wireless Communications, 2020, 19(5): 3546−3557

[75] Wu Jiaxiang, Huang Weidong, Huang Junzhou, et al. Error compensated quantized SGD and its applications to large-scale distributed optimization[C] //Proc of the 35th Int Conf on Machine Learning. New York: PMLR, 2018: 5325−5333

[76] Basu D, Data D, Karakus C, et al. Qsparse-local-SGD: Distributed SGD with quantization, sparsification, and local computations[J]. arXiv preprint, arXiv: 1906.02367, 2019

[77] Xin Ran, Kar S, Khan U A. An introduction to decentralized stochastic optimization with gradient tracking[J]. arXiv preprint, arXiv: 1907.09648, 2019

[78] Haddadpour F, Kamani M M, Mokhtari A, et al. Federated learning with compression: Unified analysis and sharp guarantees[C] //Proc of the 24th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2021: 2350−2358

[79] Tang Hanlin, Lian Xiangru, Yan Ming, et al. D^2: Decentralized training over decentralized data[C] //Proc of the 35th Int Conf on Machine Learning. New York: PMLR, 2018: 4848−4856

[80] Amiri M M, Gündüz D. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air[J]. IEEE Transactions on Signal Processing, 2020, 68(1): 2155−2169

[81] Zhu Guangxu, Du Yuqing, Gündüz D, et al. One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis[J]. IEEE Transactions on Wireless Communications, 2020, 20(3): 2120−2135

[82] Lu Yunlong, Huang Xiaohong, Dai Yueyue, et al. Differentially private asynchronous federated learning for mobile edge computing in urban informatics[J]. IEEE Transactions on Industrial Informatics, 2019, 16(3): 2134−2143

[83] Sun Jun, Chen Tianyi, Giannakis G B, et al. Communication-efficient distributed learning via lazily aggregated quantized gradients[J]. arXiv preprint, arXiv: 1909.07588, 2019

[84] 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

[85] Elgabli A, Park J, Bedi A S, et al. Q-GADMM: Quantized group ADMM for communication efficient decentralized machine learning[J]. IEEE Transactions on Communications, 2020, 69(1): 164−181

[86] Elgabli A, Park J, Bedi A S, et al. GADMM: Fast and communication efficient framework for distributed machine learning[J]. Journal of Machine Learning Research, 2020, 21(76): 1−39

[87] Elgabli A, Park J, Ahmed S, et al. L-FGADMM: Layer-wise federated group ADMM for communication efficient decentralized deep learning[C/OL] //Proc of the IEEE Wireless Communications and Networking Conf(WCNC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9120758

[88] Zhang Wei, Gupta S, Lian Xiangru, et al. Staleness-aware async-SGD for distributed deep learning[J]. arXiv preprint, arXiv: 1511.05950, 2015

[89] Tao Zeyi, Li Qun. eSGD: Communication efficient distributed deep learning on the edge[C/OL] //Proc of the 1st USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18). Berkeley, CA: USENIX Association, 2018[2022-09-05].https://www.usenix.org/conference/hotedge18/presentation/tao

[90] Wang Luping, Wang Wei, Li Bo. CMFL: Mitigating communication overhead for federated learning[C] //Proc of the 39th Int Conf on Distributed Computing Systems (ICDCS). Piscataway, NJ: IEEE: 954−964

[91] Xing Hong, Simeone O, Bi Suzhi. Decentralized federated learning via SGD over wireless D2D networks[C/OL] //Proc of the 21st Int Workshop on Signal Processing Advances in Wireless Communications (SPAWC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9154332

[92] Shiri H, Park J, Bennis M. Communication-efficient massive UAV online path control: Federated learning meets mean-field game theory[J]. IEEE Transactions on Communications, 2020, 68(11): 6840−6857

[93] Zeng Tengchan, Semiari O, Mozaffari M, et al. Federated learning in the sky: Joint power allocation and scheduling with UAV swarms[C/OL] //Proc of the 54th IEEE Int Conf on Communications (ICC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9148776

[94] Pham Q V, Zeng Ming, Ruby R, et al. UAV communications for sustainable federated learning[J]. IEEE Transactions on Vehicular Technology, 2021, 70(4): 3944−3948

[95] Fadlullah Z M, Kato N. HCP: Heterogeneous computing platform for federated learning based collaborative content caching towards 6G networks[J]. IEEE Transactions on Emerging Topics in Computing, 2020, 10(1): 112−123

[96] Chen Mingzhe, Mozaffari M, Saad W, et al. Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(5): 1046−1061

[97] Lahmeri M A, Kishk M A, Alouini M S. Artificial intelligence for UAV-enabled wireless networks: A survey[J]. IEEE Open Journal of the Communications Society, 2021, 2: 1015−1040

[98] Wang Yuntao, Su Zhou, Zhang Ning, et al. Learning in the air: Secure federated learning for UAV-assisted crowdsensing[J]. IEEE Transactions on Network Science and Engineering, 2020, 8(2): 1055−1069

[99] Lim W Y B, Huang Jianqiang, Xiong Zehui, et al. Towards federated learning in UAV-enabled Internet of vehicles: A multi-dimensional contract-matching approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(8): 5140−5154

[100] Samarakoon S, Bennis M, Saad W, et al. Distributed federated learning for ultra-reliable low-latency vehicular communications[J]. IEEE Transactions on Communications, 2019, 68(2): 1146−1159

[101] Ye Dongdong, Yu Rong, Pan Miao, et al. Federated learning in vehicular edge computing: A selective model aggregation approach[J]. IEEE Access, 2020, 8: 23920−23935

[102] Lu Yunlong, Huang Xiaohong, Dai Yueyue, et al. Federated learning for data privacy preservation in vehicular cyber-physical systems[J]. IEEE Network, 2020, 34(3): 50−56

[103] Du Zhaoyang, Wu Celimuge, Yoshinaga T, et al. Federated learning for vehicular Internet of things: Recent advances and open issues[J]. IEEE Open Journal of the Computer Society, 2020, 1: 45−61

[104] Deveaux D, Higuchi T, Uçar S, et al. On the orchestration of federated learning through vehicular knowledge networking[C/OL] //Proc of IEEE Vehicular Networking Conf (VNC). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9318386

[105] Chen Mingzhe, Semiari O, Saad W, et al. Federated echo state learning for minimizing breaks in presence in wireless virtual reality networks[J]. IEEE Transactions on Wireless Communications, 2019, 19(1): 177−191

[106] Mozaffari M, Saad W, Bennis M, et al. A tutorial on UAVs for wireless networks: Applications, challenges, and open problems[J]. IEEE Communications Surveys & Tutorials, 2019, 21(3): 2334−2360

[107] Samarakoon S, Bennis M, Saad W, et al. Federated learning for ultra-reliable low-latency V2V communications[C/OL] //Proc of the IEEE Global Communications Conf (GLOBECOM). Piscataway, NJ: IEEE, 2018[2022-09-05].https://ieeexplore.ieee.org/abstract/document/8647927

[108] Feyzmahdavian H R, Aytekin A, Johansson M. An asynchronous mini-batch algorithm for regularized stochastic optimization[J]. IEEE Transactions on Automatic Control, 2016, 61(12): 3740−3754

[109] 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

[110] Yin Feng, Lin Zhidi, Kong Qinglei, et al. FedLoc: Federated learning framework for data-driven cooperative localization and location data processing[J]. IEEE Open Journal of Signal Processing, 2020, 1: 187−215

[111] Merluzzi M, Di Lorenzo P, Barbarossa S. Dynamic resource allocation for wireless edge machine learning with latency and accuracy guarantees[C] //Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 9036−9040

[112] Yang Zhaohui, Chen Mingzhe, Saad W, et al. Energy efficient federated learning over wireless communication networks[J]. IEEE Transactions on Wireless Communications, 2020, 20(3): 1935−1949

[113] Luo Siqi, Chen Xu, Wu Qiong, et al. Hfel: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning[J]. IEEE Transactions on Wireless Communications, 2020, 19(10): 6535−6548

[114] Abad M S H, Ozfatura E, Gunduz D, et al. Hierarchical federated learning across heterogeneous cellular networks[C] //Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 8866−8870

[115] Liu Dongzhu, Zhu Guangxu, Zhang Jun, et al. Data-importance aware user scheduling for communication-efficient edge machine learning[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 7(1): 265−278

[116] Zhan Yufeng, Li Peng, Guo Song. Experience-driven computational resource allocation of federated learning by deep reinforcement learning[C] //Proc of the 34th 2020 IEEE Int Parallel and Distributed Processing Symp (IPDPS). Piscataway, NJ: IEEE, 2020: 234−243

[117] Zeng Qunsong, Du Yuqing, Huang Kaibin, et al. Energy-efficient

radio resource allocation for federated edge learning[C/OL] //Proc of the 54th 2020 IEEE Intl Conf on Communications Workshops (ICC Workshops). Piscataway, NJ: IEEE, 2020[2022-09-05]. https://ieeexplore.ieee.org/abstract/document/9145118

[118] Chen Mingzhe, Poor H V, Saad W, et al. Convergence time optimization for federated learning over wireless networks[J]. IEEE Transactions on Wireless Communications, 2020, 20(4): 2457−2471

[119] Mo Xiaopeng, Xu Jie. Energy-efficient federated edge learning with joint communication and computation design[J]. Journal of Communications and Information Networks, 2021, 6(2): 110−124

[120] Ren Jinke, Yu Guanding, Ding Guangyao. Accelerating DNN training in wireless federated edge learning systems[J]. IEEE Journal on Selected Areas in Communications, 2020, 39(1): 219−232

[121] Anh T T, Luong N C, Niyato D, et al. Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach[J]. IEEE Wireless Communications Letters, 2019, 8(5): 1345−1348

[122] Nguyen H T, Luong N C, Zhao J, et al. Resource allocation in mobility-aware federated learning networks: A deep reinforcement learning approach[C/OL] //Pro of the 6th World Forum on Internet of Things (WF-IoT). Piscataway, NJ: IEEE, 2020[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9221089

[123] Zhang Xueqing, Liu Yanwei, Liu Jinxia, et al. D2D-assisted federated learning in mobile edge computing networks [C/OL] //Pro of the 2021 IEEE Wireless Communications and Networking Conf (WCNC). Piscataway, NJ: IEEE, 2021[2022-09-05].https://ieeexplore.ieee.org/abstract/document/9417459

[124] Yang Kai, Jiang Tao, Shi Yuanming, et al. Federated learning via over-the-air computation[J]. IEEE Transactions on Wireless Communications, 2020, 19(3): 2022−2035

[125] Qin Zhijin, Li G Y, Ye Hao. Federated learning and wireless communications[J]. IEEE Wireless Communications, 2021, 28(5): 134−140

[126] Amiria M M, Dumanb T M, Gündüzc D, et al. Collaborative machine learning at the wireless edge with blind transmitters[C/OL] //Proc of the 7th IEEE Global Conf on Signal and Information Processing. Piscataway, NJ: IEEE, 2019[2022-09-05].https://iris.unimore.it/handle/11380/1202665

[127] Chen Mingzhe, Yang Zhaohui, Saad W, et al. A joint learning and communications framework for federated learning over wireless networks[J]. IEEE Transactions on Wireless Communications, 2020, 20(1): 269−283

[128] Yang H H, Arafa A, Quek T Q, et al. Age-based scheduling policy for federated learning in mobile edge networks[C] //Proc of the 45th IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE: 8743−8747

[129] Dinh C, Tran N H, Nguyen M N, et al. Federated learning over wireless networks: Convergence analysis and resource allocation[J]. IEEE/ACM Transactions on Networking, 2020, 29(1): 398−409

[130] Yang Hao, Liu Zuozhu, Quek T Q, et al. Scheduling policies for federated learning in wireless networks[J]. IEEE Transactions on Communications, 2019, 68(1): 317−333

[131] Shi Wenqi, Zhou Sheng, Niu Zhisheng. Device scheduling with fast convergence for wireless federated learning[C/OL] //Proc of the 54th IEEE Int Conf on Communications (ICC). Piscataway, NJ: IEEE, 2020[2022-09-05]. https://ieeexplore.ieee.org/abstract/document/9149138

[132] Amiri M M, Gündüz D, Kulkarni S R, et al. Update aware device scheduling for federated learning at the wireless edge[C] //Proc of the 2020 IEEE Int Symp on Information Theory (ISIT). Piscataway, NJ: IEEE, 2020: 2598−2603

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

Zhang Xueqing, born in 1994. Master. Her main research interest includes machine learning.

张雪晴, 1994 年生. 硕士. 主要研究方向为机器学习.

Liu Yanwei, born in 1976. PhD, associate professor. Member of CCF. His main research interests include wireless communication, intelligent multimedia processing, and cyber security.

刘延伟, 1976 年生. 博士, 副研究员. CCF 会员. 主要研究方向为无线通信、智能多媒体信息处理和网络安全.

Liu Jinxia, born in 1969. Master, professor. Her main research interests include wireless communication and edge intelligence.

刘金霞, 1969 年生. 硕士, 教授. 主要研究方向为无线通信和边缘智能.

Han Yanni, born in 1981. PhD, associate professor. Her main research interests include wireless communication and intelligent data analysis.

韩言妮, 1981 年生. 博士, 副研究员. 主要研究方向为无线通信和智能数据分析.