

联邦学习通信优化方法综述

杨智凯¹, 刘亚萍¹, 张硕¹, 孙哲¹, 严定宇²

(1. 广州大学网络空间安全学院, 广东 广州 510006; 2. 可信分布式计算与服务教育部重点实验室 (北京邮电大学), 北京 100876)

摘要: 随着以深度学习为代表的人工智能技术的发展和普及, 其不断暴露的安全问题已经成为影响网络空间安全的巨大挑战。传统的以云为中心的分布式机器学习通过收集参与方的数据来训练模型或优化模型表现, 在数据交换过程中容易受到安全攻击和隐私攻击等网络安全风险的影响, 进而造成整体系统效能下降或隐私数据泄露等后果。联邦学习作为一种具有隐私保护能力的分布式机器学习模式, 通过客户端与参数服务器的频繁通信进行模型参数交换, 在原始数据不离开本地的情况下训练联合模型, 这极大降低了隐私数据泄露风险, 在一定程度上保证了数据安全。但是随着深度学习模型的规模越来越大, 联邦学习任务越来越复杂, 通信开销也逐渐增加, 最终成为联邦学习应用落地的阻碍。因此, 针对联邦学习通信优化方法的探索成为联邦学习研究中的热点。介绍了联邦学习的技术背景和工作流程、分析了其通信瓶颈的来源和造成的影响。根据影响通信效率的因素, 从模型参数压缩、模型更新策略、系统架构和通信协议等优化目标对现有联邦学习通信优化方法进行全面梳理和分析, 并给出了该研究领域的发展脉络。最后总结了现有联邦学习通信优化方法所面临的问题, 并对未来的发展趋势与研究方向进行展望。

关键词: 联邦学习; 边缘计算; 通信优化; 模型压缩

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024077

Review of communication optimization methods in federated learning

YANG Zhikai¹, LIU Yaping¹, ZHANG Shuo¹, SUN Zhe¹, YAN Dingyu²

1. Guangzhou University, Cyberspace Institute of Advanced Technology, Guangzhou 510006, China

2. Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education, Beijing 100876, China

Abstract: With the development and popularization of artificial intelligence technologies represented by deep learning, the security issues they continuously expose have become a huge challenge affecting cyberspace security. Traditional cloud-centric distributed machine learning, which trains models or optimizes model performance by collecting data from participating parties, is susceptible to security attacks and privacy attacks during

收稿日期: 2024-05-27; 修回日期: 2024-11-05

通信作者: 刘亚萍, ypliu@gzhu.edu.cn

基金项目: 国家自然科学基金 (62372124); 广东省重点领域研发计划 (2019B010137005)

Foundation Items: The National Natural Science Foundation of China (62372124), Key-Area Research and Development Program of Guangdong Province (2019B010137005)

引用格式: 杨智凯, 刘亚萍, 张硕, 等. 联邦学习通信优化方法综述[J]. 网络与信息安全学报, 2024, 10(6): 1-23.

Citation Format: YANG Z K, LIU Y P, ZHANG S, et al. Review of communication optimization methods in federated learning [J]. Chinese Journal of Network and Information Security, 2024, 10(6): 1-23.

the data exchange process, leading to consequences such as a decline in overall system efficiency or the leakage of private data. Federated learning, as a distributed machine learning paradigm with privacy protection capabilities, exchanges model parameters through frequent communication between clients and parameter servers, training a joint model without the raw data leaving the local area. This greatly reduces the risk of private data leakage and ensures data security to a certain extent. However, as deep learning models become larger and federated learning tasks more complex, communication overhead also increases, eventually becoming a barrier to the application of federated learning. Therefore, the exploration of communication optimization methods for federated learning has become a hot topic. The technical background and workflow of federated learning were introduced, and the sources and impacts of its communication bottlenecks were analyzed. Then, based on the factors affecting communication efficiency, existing federated learning communication optimization methods were comprehensively sorted out and analyzed from optimization objectives such as model parameter compression, model update strategies, system architecture, and communication protocols. The development trend of this research field was also presented. Finally, the problems faced by existing federated learning communication optimization methods were summarized, and future development trends and research directions were looked forward to.

Keywords: federated learning, edge computing, communication optimization, model compression

0 引言

随着人工智能 (artificial intelligence, AI) 技术的发展以及大模型的兴起与普及, 以数据为驱动的深度学习的模型质量与数据的质量及数量紧密相关。传统的以云为中心的深度学习通过广泛收集公开或私人的数据用以模型训练和评估, 个人用户或组织开发专用模型时往往需要将私有数据上传至云服务器上进行处理和训练, 这些含有隐私的信息在网络中传播、存储和处理的过程中会带来隐私泄露风险, 保护用户隐私成为云计算、大数据技术等领域的研究重点。李风华等^[1]强调了隐私保护的重要性, 并于 2016 年率先提出了隐私计算概念, 其是面向隐私信息全生命周期保护的计算理论和方法, 旨在促进和指导隐私保护理论和技术体系的建立与发展。欧盟在 2018 年出台了《通用数据保护条例》(general data protection regulation, GDPR)^[2], 旨在规范企业在对用户数据进行收集、处理和使用时应遵守的规则。GDPR 的实施让人们重新审视数据隐私保护的必要性, GDPR 合规性检查技术^[3]也推动了隐私数据使用技术的发展, 将带有隐私保护属性的深度学习范式推向前台。

2016 年 Google 提出了联邦学习 (federated learning, FL)^[4]的概念并将其应用于 Google 输

入法^[5]的候选词推荐功能。在保护用户隐私的前提下, 联邦学习可以使多个用户共享数据潜在分布等非隐私信息, 将模型训练和推理任务卸载到边缘终端设备, 缓解云上计算压力, 联合建立起比单一用户模型更加精准、泛用的联邦模型。联邦学习在金融^[6]、医疗^[7-9]、物联网^[10-12]以及政府公益项目等领域将会有着不可替代的作用。

然而, 参与联邦学习的边缘终端设备不尽相同, 它们拥有不同的软硬件资源和网络环境, 大量的移动设备电量和网络带宽有限, 不同设备所持有的数据并非完全独立同分布 (independent and identically distribution, IID)。复杂的边缘环境造成了联邦学习系统效能的下降, 由于现有网络通信技术逐渐无法满足边缘终端设备计算资源与数据量剧烈增长所带来的网络需求, 通信逐步成为制约联邦学习发展的瓶颈。因此通信优化是联邦学习中重要研究方向之一, 这对推进通信和电力资源有限的边缘终端设备参与到联合学习、贡献个人非隐私数据至关重要, 同时也对联邦学习应用在复杂边缘环境的落地具有现实意义。

本文对近些年联邦学习通信优化方法和技术进行了较为全面系统的梳理和总结, 利用对通信效率及其影响因素的形式化阐述, 从优化目标着手对相关研究进行分类, 涵盖了从底层

硬件设备、网络拓扑组织、通信协议到应用层通信策略、参数压缩等诸多方面，并提供了该研究领域的发展脉络图，提出了研究展望，为未来的联邦学习通信优化方法研究提供了新的见解。

1 联邦学习

联邦学习将深度学习模型训练任务卸载至边缘终端设备，各参与方利用自有数据和本地计算资源在本地训练模型，然后将本地模型上传至参数服务器（parameter server, PS），由参数服务器将其聚合为一个全局模型，从而达到在不上传原始数据的前提下联合训练一个含有不同参与方数据潜在信息的模型的目的。联邦学习通信模式是典型的客户端-服务端形式，以下将参与方称为客户端，将参数服务器称为服务端。

中心化架构^[13]的联邦学习在实际应用中较为普遍，工作流程如图1所示，具体如下：①初始化全局模型：服务端初始化待训练全局模型；②选择客户端，并分发模型：服务端根据客户端选择策略选择满足参与联邦学习条件的客户端，然后将全局模型广播至选定的客户端；③本地训练：客户端下载全局模型并利用本地数据训练模型，训练结束后将模型参数上传至服务端；④全局模型聚合与更新：服务端在收到客户端的模型参数更新后，使用模型聚合策略将其聚合为新的全局模型，并将更新后的全局模型参数再次广播至客户端。上述流程中的模型参数可以是模型的梯度、权重或对应的更新（即本地模型与全局模型的差值）等。重复步骤②~④，直到达到设置的目标（预定的模型精度或者训练时间）。全局模型的分发和本地模型的上传是联邦学习通信过程的主要组成部分，这两阶段直接影响了联邦学习系统的效率，如何降低这两阶段的通信时间和带宽开销成为联邦学习通信优化的关键问题之一。去中心化架构的联邦学习同样是研究热点，该架构可以避免恶意服务器对联邦学习造成威胁。例如 BrainTorrent^[14]框架允许在每轮本地训练结束后随机选取一个客户端作为临时参数服务器。

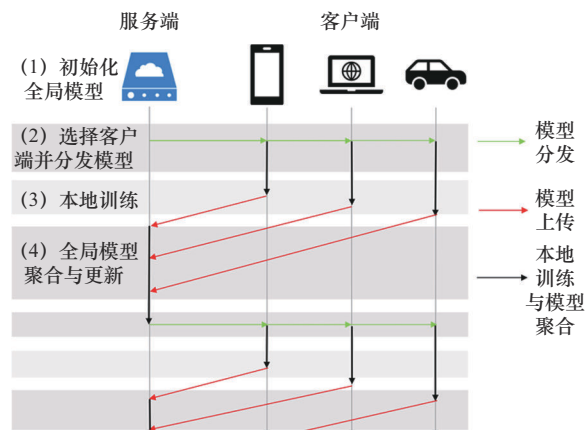


图1 联邦学习工作流程

Figure 1 The workflow of federated learning

无论中心化架构还是去中心化架构的联邦学习，其过程中客户端与服务端都需要频繁交换模型参数等数据。现如今模型规模不断扩大，参数数量不断增加，即便是计算机视觉中常用的 ResNet18^[15]模型的参数数量也达到了 1 169 万，因此，模型参数成为联邦学习通信的主要内容。联邦学习一般要求客户端同时上传模型参数并进行聚合，视参与的客户端数量和模型规模而定，客户端和服务端在短时间内的交互可能带来大量的通信和计算开销。联邦学习作为边缘计算类任务，面临复杂的边缘网络通信环境，作为边缘计算节点的客户端的网络延迟较高，网卡、基站等内外部网络数据处理硬件的性能差异较大，客户端侧的网络实为异构网络，异构网络下各计算节点的同步问题和一致性问题导致联邦学习相较于数据中心内部的分布式机器学习的性能有着大幅下降。通信能力、信道质量与联邦学习的效能紧密相关，因此，通信能力的强大与否和稳定性，成为影响联邦学习能否实际且广泛落地应用的核心因素之一。

以下3个因素影响联邦学习的通信效率，成为联邦学习发展的挑战。

1) 庞大的客户端数量^[16]：海量的智能终端设备为联邦学习的发展提供了硬件基础，丰富的算力资源和数据也有助于提升模型的精度，但庞大的客户端群体也造成了潜在的通信压力，同步更新模式下的联邦学习会出现短时间内模型参数同时上传而造成通信链路拥塞的情况，网络内各处需要时间来处理拥塞，服务端聚合流程将被

迫延长, 继而造成联邦系统效率的下降。

2) 异构设备与异构网络^[17]: 不同边缘节点使用的软硬件大多具有一定的异构性, 例如智能手机大多采用 ARM 平台, 而计算机大多采用 x86 平台, 且各平台 CPU 与 GPU 的性能也相差过大等, 这些因素影响模型训练速度与精度。硬件异构性使并行训练方式变得不再可行, 一些客户端模型训练时间长、网络带宽低、通信延迟高, 甚至存在突然离线和掉队现象, 无法及时上传模型参数, 造成服务端长时等待, 严重影响联合学习效率。

3) 数据的统计异质性^[18]: 由于受到用户所处环境、个人行为以及上述异构设备等的实际情况影响, 不同客户端数据实际上是非独立同分布 (non-independent and identically distribution, Non-IID) 的, 存在特征漂移、标签漂移、时间漂移等现象, 本地数据的潜在分布与全体数据的分布并非完全一致。Non-IID 数据一定程度上降低了全局模型收敛效率, 进而需要更长的通信时间和更多的通信数据总量来使模型收敛到预定目标, 最终通信效率下降。

2 联邦学习通信优化方法

联邦学习通信优化的总体目标是减少模型参数传输过程的数据总量或提高模型参数的传输速

率, 减少全局模型分发和本地模型上传的时间。可以将联邦学习通信效率用如下公式近似表达:

$$E \propto \frac{\text{Acc}}{C \times f \times P \times T}$$

其中, E 表示通信效率, 该值越大表示联邦系统通信效率越高; Acc 表示全局模型精度; C 表示客户端选择比例, C 越大表示参与本轮联邦学习聚合的客户端数量越多; $f = N/N_{\text{iter}}$ 表示客户端与服务端的通信频率, N 为达到预定训练目标时需要的本地训练总轮次, N_{iter} 为每轮联邦学习中客户端在本地的训练轮次; P 表示单次通信的模型参数总比特数; $T = t_{\text{train}} + t_{\text{trans}} + t_{\text{agg}}$ 表示单个轮次联邦学习通信花费的总时间, 由客户端模型训练时间、模型参数传输时间和服务端参数处理时间构成。

依据优化目标侧重不同, 可将联邦学习通信优化方法分为如下 4 类, 如图 2 所示。

1) 基于参数压缩的优化方法: 该类优化方法的目标是压缩单次通信中传输的模型参数数量, 当全局模型精度不变的情况下, 传输的参数越少, 通信效率越高。

2) 面向模型更新策略的优化方法: 模型更新策略优化方法主要优化目标是模型更新的频率和方式, 优化模型更新策略可以降低通信次数或者间接降低通信总体时间 T 。

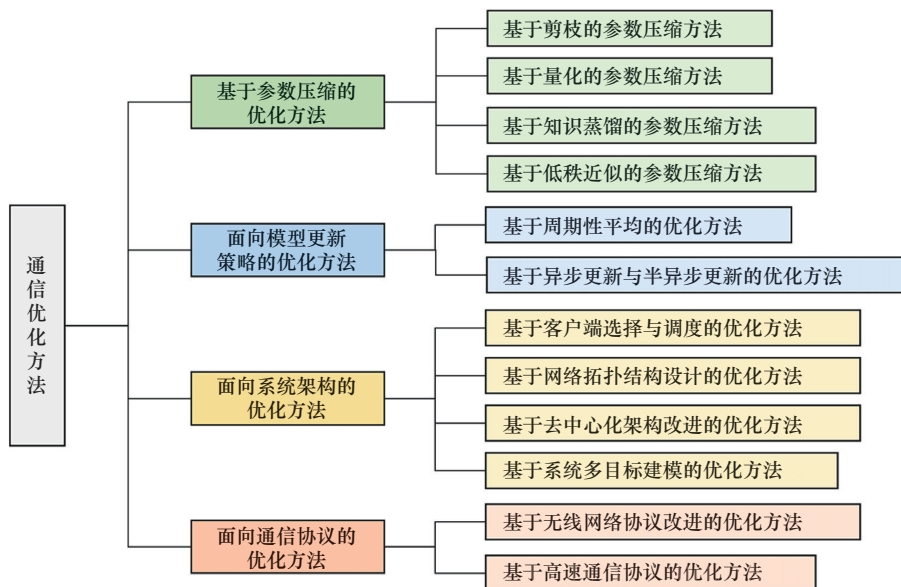


图2 联邦学习通信优化方法分类

Figure 2 Classification of federated learning communication optimization method

3) 面向系统架构的优化方法：面向系统架构的优化方法指的是对系统架构、客户端调度方式（如客户端选择比例 C ）和网络拓扑结构进行优化，或者对系统整体进行最优建模和组合优化，进而从整体上提高通信效率。

4) 面向通信协议的优化方法：通信协议优化方法的主要目标是客户端与服务端之间的信息传输机制与通信内容，最终降低通信延迟和参数传输时间，即通信总时间 T 。

3 基于参数压缩的优化方法

联邦学习中客户端和服务端通信的主要内容是模型参数，包括模型权重、梯度以及更新等。深度学习模型中的部分信息是冗余的^[19-20]，如果压缩模型参数，减少冗余信息的传输，那么就可以显著减少单次通信数据量，最终减少通信数据总量。基于参数压缩的优化方法主要包括剪枝、量化、知识蒸馏和低秩分解。

3.1 基于剪枝的参数压缩方法

剪枝是深度学习模型压缩领域中应用最广泛的技术之一，Hanson等^[21]提出的基于有偏权重衰减的剪枝方法是对网络稀疏化最早的研究之一。剪枝的核心思想是去除模型中信息量较少的冗余权重。剪枝分为非结构化剪枝和结构化剪枝。非结构化剪枝直接对模型权重进行删减，实现参数矩阵的稀疏化，而结构化剪枝直接对神经网络结构进行遮掩或删除，如在卷积神经网络（convolutional neural network, CNN）对卷积核和通道进行剪枝。

3.1.1 基于结构化剪枝的参数压缩方法

结构化剪枝方法基于显著性原则，对非重要的结构进行删减，如文献[22-23]中的剪枝技术对卷积核等卷积神经网络结构进行删减。在优化目标中加入稀疏性限制也是结构化剪枝常用的方法之一，Zhou等^[24]向目标函数中加入稀疏性限制，在训练阶段限制神经元数量，最终得到稀疏神经网络。Han等^[25]提出了三步剪枝方法：训练权重、模型剪枝、稀疏网络重训练，将低于阈值的模型权重批量删除，该三步法被广泛应用于模型剪枝技术中。联邦学习中的结构化剪枝通常用于客户端本地训练中，虽然该方法需要消耗时间探索剪枝结构，但是一旦确定剪枝结构后，将加

速本地训练，减少计算开销，而且结构化剪枝相较非结构化剪枝从整体结构上克服了无法对稀疏矩阵进行高效计算的问题，更有利于数据的读取和计算^[26]。

结构化更新^[27]是联邦学习中最早的结构化剪枝方法之一，其从训练过程着手，将可学习参数矩阵限制为低秩矩阵，或利用随机掩码遮掩部分参数，从而达到直接训练稀疏模型的目的。结构化更新仅考虑了上行通信的参数压缩，文献[28]对服务端发往客户端的模型参数进行有损压缩以减少下行通信开销，并使用联邦丢弃来对本地模型进行结构化剪枝，在每个全连接层中随机选择一定比例的激活值进行丢弃，创建并训练全局模型的子模型。FEDOBD^[29]是一种更粗粒度的结构化剪枝方法，其将神经网络模型分解成多个块，这些块通常对应网络中的一些结构，如卷积核、池化层、归一化层和激活层，每个块在训练结束后根据平均块差（mean block difference, MBD）指标赋予重要性分数，客户端根据MBD分数将重要性高的块上传。CS（complement sparsification）^[30]要求服务端和客户端协作剪枝，服务端根据显著性原则对聚合模型剪枝，生成全局稀疏模型，然后将对应的稀疏掩码一同下发给客户端，客户端使用反转掩码来进行训练，生成与全局稀疏模型互补的本地稀疏模型，并上传至服务端，这些本地稀疏模型与前一次的全局稀疏模型进行聚合，生成新的密集模型，重复这个过程直至收敛。该方法通过稀疏模型互补过程能够避免对全局稀疏模型非零权重的过拟合。

一些研究也在探索不同的剪枝准则，剪枝准则指的是根据何种规则和目标进行剪枝，不同的剪枝准则会产生不同的效果。Flado^[31]通过调整客户端模型中的每个神经元的稀疏性实现剪枝，其剪枝目标是使客户端的模型更新轨迹（即梯度方向）与全局模型更新轨迹更好地对齐，令客户端的计算资源向那些有助于协同训练的神经元倾斜。FedDST^[32]方法中服务端会额外初始化一个稀疏掩码，客户端依据稀疏掩码训练指定部分的权重，在特定轮次时客户端会依据本地数据和模型调整稀疏掩码以探索更有

效的稀疏子模型。FLuID 框架^[33]则考虑了客户端资源限制问题,其核心在于动态地为联邦学习中的落后客户端创建子模型,以减轻计算负担,FLuID 根据实时的客户端性能调整丢弃比例和子模型的大小,确保更新子模型结构后的落后客户端的训练时间接近目标训练时间。FedResCuE^[34]通过渐进式学习来实现结构化剪枝,该学习模式将神经网络模型垂直分割成多列,每列代表模型的一部分,如一个或多个模型通道,允许在学习的不同阶段逐步增加新列,从而逐渐丰富其特征表示。

3.1.2 基于非结构化剪枝的参数压缩方法

非结构化剪枝直接作用于训练后的模型参数,例如, Srinivas 等^[35]提出的一种不依赖训练数据的神经网络剪枝框架,计算神经元的冗余程度并删除相似神经元中的一个,尽量使得剪枝后的模型与原模型的输出保持一致。联邦学习中非结构化剪枝主要实现方式是在本地模型训练结束后对其权重、梯度等参数进行稀疏化,减少通信开销、降低通信时延。

草图更新 (sketched updates)^[27]是联邦学习中最早使用非结构化剪枝技术的参数压缩方法之一,其流程如图 3 所示,在模型训练完成后,首先使用下采样方法随机采样参数,随机采样保证了稀疏化矩阵是原矩阵的无偏估计;然后利用旋转和量化压缩采样后的稀疏矩阵;最后进行编码和传输,最终降低了需要传输的数据总量。这种模型压缩方式为之后以剪枝和量化为主的模型压缩方法提供了路径参考,剪枝-旋转-量化-编码方法成为联邦学习模型参数压缩类方法的主要技术路线之一。

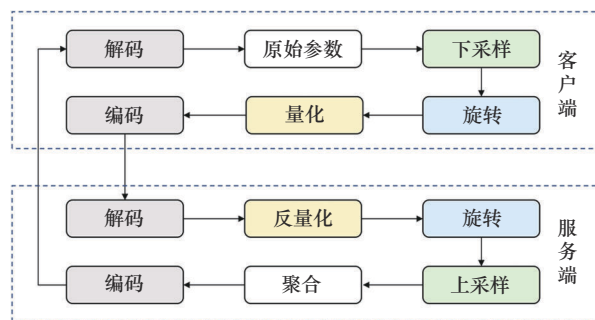


图3 草图更新类算法流程

Figure 3 Sketched updates algorithm process

深度梯度压缩 (deep gradient compression, DGC)^[36]结合了动量校正、局部梯度裁剪、动量因子遮蔽和热身训练等技术,并只上传超过设定阈值的梯度。CANITA^[37]有条件地进行非结构化剪枝,将随机稀疏化或随机量化的期望与方差限制在阈值之内,并引入动量机制来加速收敛。FetchSGD^[38]使用计数草图作为稀疏化算子来压缩客户端模型参数,服务端将其重建为稀疏模型参数的无偏估计,并用 Top-K 方法广播压缩模型,实现了上下行双向压缩,并采用误差反馈机制,在服务端上累计动量和误差,克服客户端无法在本地多次累计误差的弊端。

针对非结构化剪枝后模型的聚合策略,文献^[39]提出了利用空间相关性(不同客户端上传的参数之间的相似性)和时间相关性(单个客户端在不同通信轮次中上传的参数之间的相似性)来对上传的模型参数进行处理: Rand-k-Spatial 算法利用显著性原则调整不同客户端的模型参数中每个参数的权重,如果某个位置上的参数被不同客户端频繁采样,意味着该参数相对重要,在聚合时为其分配更大的权重; Rand-k-Temporal 算法则在服务端处根据客户端以往上传的稀疏化模型参数来填补当轮次模型的稀疏化空缺。

Top-K 是最常见的基于显著性原则的非结构化剪枝方法,许多研究基于此方法并进行改进^[40-43],优化 Top-K 方法将有助于提升现有的基于该方法的通信优化算法的表现。

3.1.3 基于自适应剪枝的参数压缩方法

固定的剪枝结构或剪枝比例无法完全适应动态变化的联邦学习环境,单一采用结构化剪枝或非结构化剪枝无法完全处理客户端硬件与网络的异构性和数据的统计异质性造成的模型差异问题。自适应剪枝考虑了复杂环境下的剪枝方法,采取动态剪枝策略,自适应联邦学习流程或者客户端环境。PruneFL^[44]是一种二阶段自适应剪枝策略:初始阶段,服务端在选定的客户端上对全局模型进行剪枝,使联邦学习以一个较小的模型开始;自适应剪枝阶段,持续跟踪一个足够小的便于传输和计算的模型,通过间隔性的重配置(遮掩参数或者回添参数)来保留重要参数,并可以通过参数保持量来控制模型大小,在不影响

模型效果和训练效率的情况不断减小模型。

为具有不同计算资源的客户端自适应采取不同的剪枝方法可以有效缓解设备异构性带来的掉队等问题。FedMP 框架^[45]采用一种基于多臂老虎机 (multi-armed bandit, MAB) 的在线学习算法学习剪枝比例、资源消耗和模型性能之间的关系, 为异构客户端确定不同的剪枝比例。文献^[46]对于计算资源充足的客户端针对性地训练一个二进制掩码来实现非结构化剪枝; 对于资源受限的客户端进行结构化剪枝。Helios 框架^[47]中服务端会识别弱计算能力设备, 并确定其在每次训练轮次中预期的模型训练计算量, 保证资源受限的客户端在预定时间内能够完成联邦协作, 减少服务端的通信等待时间。

剪枝技术作为模型压缩领域最常用的压缩技术之一, 其优点除了能够减小模型大小外, 还可

以提高推理速度和存储效率, 并且在一定情况下可以提高模型泛化能力, 该特性对全局模型在客户端上的表现非常重要。剪枝技术必然会降低模型精度和表达能力, 且大部分剪枝方法对网络结构有着不可逆的改变, 有可能损失潜在信息, 因此仍需要对联邦学习中的剪枝技术进一步优化, 降低其模型造成的不可逆损失。表1总结了联邦学习中基于剪枝的参数压缩方法。

3.2 基于量化的参数压缩方法

现有的深度神经网络模型大都使用高精度的32位浮点值 (FP32) 数值格式的参数来训练和推理, 但是并非所有的模型都需要如此高精度的数值表示, 因此可以使用将高精度数值转换为低精度数值的量化技术, 通过牺牲一定的精度来降低训练和推理消耗的计算资源, 加快训练和推理速度, 减少模型存储空间。模型量化后不但减少

表1 基于剪枝的参数压缩方法

Table 1 Parameter compression methods based on pruning

年份	方法	主要压缩技术	压缩方向	特点
2016	文献 ^[27]	结构化剪枝/非结构化剪枝	上行	结构化更新, 草图更新
2017	DGC ^[36]	非结构化剪枝	上行	动量校正, 局部梯度裁剪, 动量因子遮蔽, 热身训练
2018	文献 ^[28]	结构化剪枝	上行/下行	草图更新, 联邦丢弃
2019	PruneFL ^[44]	自适应剪枝	上行	两阶段自适应剪枝
2020	FetchSGD ^[38]	非结构化剪枝	上行/下行	计数草图, Top-K 稀疏化, 误差反馈
2021	CANITA ^[37]	非结构化剪枝	上行	限制稀疏化或量化的期望与方差
2021	文献 ^[39]	非结构化剪枝	上行	空间相关性, 时间相关性
2021	FL-TOP ^[43]	非结构化剪枝	上行/下行	在服务端公共数据集上训练稀疏掩码
2021	Helios ^[47]	自适应剪枝	上行	根据计算资源自适应调整剪枝比例
2022	FedDST ^[32]	结构化剪枝	上行	服务端指定稀疏掩码
2022	FedResCuE ^[34]	结构化剪枝	上行/下行	渐进式学习, 动态增加网络结构, 允许下载非完整模型
2022	SmartIdx ^[41]	非结构化剪枝	上行	改进 Tok-K 稀疏化, Golomb 编码
2022	ZeroFL ^[42]	非结构化剪枝	上行/下行	调整不同客户端的 Top-K 稀疏化参数, 传输模型参数更新的 Top-K 稀疏化矩阵
2022	FedMP ^[45]	自适应剪枝	上行	强化学习, 自适应剪枝比例
2023	FEDOBD ^[29]	结构化剪枝	上行	模型分解成块, 根据重要性评分选择性保留
2023	CS ^[30]	结构化剪枝	上行/下行	训练反转稀疏模型
2023	Flado ^[31]	结构化剪枝	上行	以本地模型更新对齐全局模型为目标剪枝
2023	FLuID ^[33]	结构化剪枝	上行	为落后客户端动态创建特定子模型
2023	RandTopk ^[40]	非结构化剪枝	上行	增加 Top-K 稀疏化方法的随机性
2023	文献 ^[46]	自适应剪枝	上行	根据计算资源确定结构化剪枝和非结构化剪枝

了计算资源的消耗, 每轮通信的数据量和通信时间也相应减少, 既提高了计算速度, 又提高了通信性能。

量化方法一般分为两类: 量化感知训练 (quantization-aware training, QAT)^[48]和训练后量化 (post-training quantization, PTQ)^[49]。量化感知训练直接使用已量化模型进行训练, 其再训练过程可以弥补一定的精度误差。训练后量化则在训练结束后使用校准数据集对模型参数进行量化, 可以快速部署和推理、减少存储空间和计算资源的消耗。常见的量化方法有: 定点量化, 使用特定的比例因子将原始值映射到定点值, 如将 FP32 量化到 INT8; 二元量化^[50-51]将参数量化到 $\{-1, +1\}$ 或者 $\{0, 1\}$ 中, 二元量化契合硬件特性, 可以加速训练和推理; 三元量化^[52]将量化后参数限制在 $\{-1, 0, +1\}$ 中, 表征能力相较二元量化得到加强; 混合精度量化将不同量化技术结合使用, 为重要的参数分配较大的量化位宽, 其余参数则分配较小的量化位宽。量化位宽的参数数值表示所占的二进制位数, 代表了量化程度的高低, 量化位宽越大表示精度越高。

3.2.1 基于固定量化的参数压缩方法

草图更新^[27]中使用了概率量化压缩被剪枝后的稀疏矩阵, 同时提出了在量化前对参数矩阵进行旋转操作的方法。当量化目标矩阵的参数上下限差距过大时, 量化产生的误差会更大, 旋转使参数矩阵在各个维度差异不会过大, 矩阵更加平滑, 有助于减小量化误差。许多研究在此基础上使用不同形式的固定量化位宽的方法压缩模型参数。

稀疏二元压缩 (sparse binary compression, SBC)^[53]利用二元量化和 Top-K 稀疏化来压缩参数, 并使用 Golomb 位置编码方法最优编码压缩后参数中非零元素的位置信息, 该作者在后续工作中提出了稀疏三元压缩 (sparse ternary compression, STC)^[54], 将上述 SBC 中的二元量化替换为三元量化, 并将其扩展到下行通信阶段, 实现了双向通信压缩。FTTQ 算法和 T-FedAvg 算法^[55]同样使用了三元量化方法实现了双向压缩。

FedPAQ^[56]使用随机量化方法来解决通信瓶颈问题, 显著减小了上行带宽。LFL^[57]算法则考

虑了双向通信链路上的随机量化。CE-FedAvg^[58]在分布式 Adam 优化方法中引入两种量化算法: 均匀量化和指数量化。文献[59]提出了一种适用于联邦学习的减法抖动格子量化, 该方法的优势在于其普适性, 即它不依赖于模型更新的具体分布, 有助于降低压缩误差。

3.2.2 基于自适应量化的参数压缩方法

固定量化位宽的量化方法未充分考虑客户端复杂的硬件与数据环境造成的模型差异, 也未充分使用量化算法之外的额外信息, 这可能会对全局模型贡献较大的客户端模型无法得到充分利用, 不利于全局模型的收敛。自适应量化方法为不同客户端或模型的不同部分设置不同的量化位宽, 有助于减少量化误差。QD-Compressor^[60]中客户端上传的是模型参数更新的量化, 考虑了 DNN 中不同层参数分布的差异性, 为每一层分配不同的量化器和量化位宽。FedCOMGATE^[61]在量化压缩之外还跟踪全局模型梯度与本地模型梯度之间的差异, 在更新全局模型时考虑该信息, 优化量化算法表现。

一些研究根据客户端资源来配置不同的量化方法或量化位宽。DAdaQuant^[62]是一种双自适应量化算法, 其结合了时间自适应量化和客户端自适应量化两种方法, 前者随着训练进程的发展逐渐增加量化位宽, 在训练前期用低量化位宽来减少模型参数量, 而在后期用高量化位宽来进一步降低训练损失; 后者则根据客户端持有数据总量来分配不同的量化位宽, 拥有样本较多的客户端所训练的模型参数在全局模型中占据更大比重。该算法通过结合并调节这两种自适应量化方法有效降低了联邦学习的带宽使用、能源消耗和训练时间。AdaGQ^[63]是一种结合了自适应量化和异构量化的综合量化方法, 根据训练过程中梯度范数的变化动态调整量化位宽, 并考虑到客户端的通信能力不同, AdaGQ 为训练较慢的客户端分配较少的量化位宽以减少其传输时间, 而为训练较快的客户端分配更多的量化位宽以保持全局模型的准确性。EDEN^[64]允许客户端根据自己的通信预算使用不同的量化值集合和压缩比例。LAQ^[65]量化目标为梯度的更新, 只有当梯度的更新超过某个阈值时才上传量化的梯度更新。

量化技术为有损压缩技术, 量化模型无法准

确表示原始模型能够表达的动态范围，限制了模型的高精度表示，不适合一些需要高精度的应用或任务，且引入的非线性效应也会掩盖模型的真实行为，因此需要进一步探索和改进联邦学习中的量化技术，以适应更复杂、更广泛的应用场景。表2总结了联邦学习中基于量化的参数压缩方法。

3.3 基于知识蒸馏的参数压缩方法

知识蒸馏（knowledge distillation, KD）是一种知识迁移方法，基本思想是将大型的教师神经网络模型中的知识迁移至小型的学生神经网络模型中。知识蒸馏最早由Hinton等^[66]提出，他们认为Softmax分类器输出的分类置信概率相较于信息熵极低的独热编码更具有丰富的潜在数据信息，这些信息有着相当高的学习价值，最终提出了教师-学生模型，将教师网络经Softmax分类器输出的知识蒸馏至学生网络中，从而达到压缩模型的目的。

联邦蒸馏（federated distillation, FD）^[67]是联邦学习中首个采用知识蒸馏技术对不同客户端的异构模型进行知识聚合的模型压缩方法，每个客户端保存模型训练期间的每个标签的logit向量，服务端收集这些logit向量并平均，然后广播这些含有其他客户端模型信息的logit向量，由此完成联合模型训练。文献^[68]在理论上证明了分

布式蒸馏算法下，所有参与方蒸馏正则化损失函数的梯度收敛到零的概率为1，为联邦学习中的知识蒸馏的应用提供了理论保证。FedKD^[69]引入了自适应的相互蒸馏框架，本地教师模型和学生模型能够从彼此蒸馏的知识中学习。DS-FL蒸馏框架^[70]使用了半监督学习方式，客户端共享的是未标记的公共数据集在本地模型的输出，并且通过熵减少聚合（entropy reduction aggregation, ERA）来使输出更加“尖锐”，所含信息更少，减少输出不确定性，加速训练过程。

采用知识蒸馏模型压缩方法的联邦学习，其通信负载依赖于输出的学生模型的大小，而且由于知识蒸馏技术的特性，该方法可以利用学生模型屏蔽异构本地模型，有助于缓解设备异构性和数据的统计异质性。

FedGTK算法^[71]在客户端上训练小型模型，定期将知识蒸馏至服务端的大型模型中，这样中央大模型可以获得客户端潜在数据特征，并且通过知识蒸馏减小了通信带宽。FedBoost^[72]在集成学习上进行联邦化，中央服务器仅发送基础学习器给不同的客户端，客户端随后将模型参数更新返回至服务端，服务端使用集成学习的方式聚合基础学习器，形成全局集成模型。

知识蒸馏需要两个规模不同的模型进行协作，

表2 基于量化的参数压缩方法

Table 2 Parameter compression methods based on quantization

年份	方法	主要压缩技术	压缩方向	特点
2016	文献 ^[27]	固定量化	上行	结构化更新,草图更新,概率量化
2019	SBC ^[53]	固定量化	上行	二元量化,Top-K稀疏化
2019	STC ^[54]	固定量化	上行/下行	三元量化,Top-K稀疏化
2020	LFL ^[57]	固定量化	上行/下行	双向随机量化
2020	CE-FedAvg ^[58]	固定量化	上行	均匀量化,指数量化,Golomb编码
2020	文献 ^[59]	固定量化	上行	适用于联邦学习的减法抖动格子量化
2020	FedCOMGATE ^[61]	自适应量化	上行	跟踪全局模型梯度与本地模型梯度之间的差异
2021	FedPAQ ^[56]	固定量化	上行	随机量化
2022	FTTQ和T-FedAvg ^[55]	固定量化	上行/下行	三元量化
2022	DAdaQuant ^[62]	自适应量化	上行	客户端自适应量化,时间自适应量化
2022	AdaGQ ^[63]	自适应量化	上行	自适应量化,异构客户端适应量化
2022	EDEN ^[64]	自适应量化	上行	根据通信预算自适应量化
2022	LAQ ^[65]	自适应量化	上行	量化梯度的更新,自适应量化
2023	QD-Compressor ^[60]	自适应量化	上行	为DNN中不同层分配不同的量化器和量化位宽

维持两个模型的训练和推理会消耗大量资源,因此知识蒸馏类的模型压缩方法并不适用于计算资源有限的边缘终端设备。但是知识蒸馏提供了一个有价值的思路:可以将模型参数转换为高信息熵的形式来减少通信开销,这类方法将有较大的研究价值。表3总结了联邦学习中基于知识蒸馏的参数压缩方法。

3.4 基于低秩近似的参数压缩方法

低秩分解或者张量分解^[73]是一种基于矩阵分解的模型压缩技术,神经网络的参数是矩阵化的,这意味着可以通过多个低秩矩阵来近似原参数矩阵,对于高维的张量也可以使用张量分解来压缩。如图4所示,原矩阵 $W \in \mathbb{R}^{m \times n}$, 使用矩阵分解方法使其等于 $X \in \mathbb{R}^{m \times r}$ 与 $Y^T \in \mathbb{R}^{r \times n}$ 的乘积,且使 $(m+n) \times r < m \times n$, 从而实现参数压缩的目的。联邦学习中常常将其用于客户端模型参数上传前的压缩,并在服务端进行重建和聚合,从而减少通信过程中的带宽消耗。

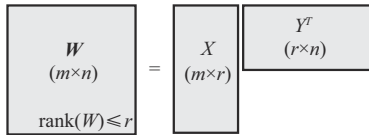


图4 低秩分解

Figure 4 Low-Rank Factorization

PowerSGD^[74]基于幂迭代方法压缩梯度为低秩矩阵,降低了通信数据量。FedPara^[75]使用低秩权重矩阵和Hadamard乘积来重新参数化模型参数,具体来说是将参数矩阵表示为两个低秩矩阵乘积的Hadamard积。作者证明了FedPara在给定相同数量的参数时,可以比传统的低秩参数化方法达到更高的秩,从而具有更好的模型表达能力。文献[76]采用了基于SVD的低秩分解方案,并且在本地模型训练过程中只学习低秩部分的参数。SVDFed^[77]在

SVD方案基础上,额外引入了时间和空间上的梯度相关性来优化算法。LBGM^[78]探索了梯度空间的低秩特性,作者认为梯度空间通常由少数几个主成分构成,这意味着训练过程中生成的梯度可以由这些主成分低秩近似表示。

类似于低秩分解思想,一些研究将高维的模型参数转换为低维数据,然后在服务端进行近似重建。FedDCS^[79]是一种基于分布式压缩感知理论的模型参数压缩方法,在联邦学习中相邻轮次之间的梯度具有相关性,作者使用前一轮的梯度作为侧信息来提高稀疏重建的性能。文献[80]中客户端使用草图矩阵将原始模型的高维梯度向量映射到一个低维空间,在服务端利用草图矩阵的转置将低维梯度映射回原始维度空间。Rand-Proj-Spatial^[81]使用随机投影矩阵将原始向量由高维投影到低维,服务端收到编码向量后恢复出原始向量的近似值,并在重建时考虑了客户端向量之间的相关性。

低秩分解类方法的缺陷是需要消耗一定的计算资源进行矩阵的分解和重建,在参与学习的客户端数量较少的情况下,边缘终端设备较贫乏的计算资源成为该类方法性能的瓶颈,而在大量客户端的情况下服务端的模型参数重建任务也会给服务端带来计算压力,最终增加了通信延迟。计算高效的低秩分解与重建算法的研究是该类方法的重要突破方向之一。表4总结了近年来联邦学习中基于低秩分解的参数压缩方法。

4 面向模型更新策略的优化方法

模型更新策略指的是客户端与服务端之间以何种方式在何时传递模型参数。联邦学习一般有3种模型参数传递方式:①同步方式,服务端需要等待所有选定的客户端上传模型参数结束后才开始聚合,现有的联邦学习方法大多数是同步方

表3 基于知识蒸馏的参数压缩方法

Table 3 Parameter compression methods based on knowledge distillation

年份	方法	压缩方向	特点
2018	FD ^[67]	上行/下行	联邦蒸馏,联邦增强
2020	文献[68]	上行/下行	为联邦学习中知识蒸馏的应用提供了理论保证
2020	FedGTK 算法 ^[71]	上行	将客户端知识定期蒸馏至服务端中的大模型
2022	FedKD ^[69]	上行/下行	自适应相互蒸馏,低秩近似模型
2023	DS-FL ^[70]	上行/下行	半监督,熵减少聚合

表4 基于低秩分解的参数压缩方法

Table 4 Parameter compression methods based on low-rank factorization

年份	方法	压缩方向	特点
2019	PowerSGD ^[74]	上行	幂迭代, 误差反馈, 并行聚合
2022	FedPara ^[75]	上行	将模型参数分解为低秩矩阵乘积的 Hadamard 积
2022	文献[76]	上行	基于 SVD 分解
2022	LBGM ^[78]	上行	利用梯度空间的低秩特性
2023	SVDFed ^[77]	上行	引入时间相关性和空间相关性
2023	FedDCS ^[79]	上行	分布式压缩感知
2023	Rand-Proj-Spatial ^[81]	上行	随机投影矩阵, 利用客户端编码向量的相关性

式; ②异步方式, 允许客户端在训练结束后立刻上传模型参数, 服务端收到后迅速返回更新的参数, 或者允许客户端在未及时上传或下载模型的情况下仍在本地保持训练; ③半异步方式是同步和异步方式的结合。现有一些研究关注并优化模型参数更新策略, 以减少服务端和客户端因等待对方的下一步行为而产生的时间空耗, 进而降低通信延迟。

4.1 基于周期性平均的优化方法

在分布式均值估计类算法中引入周期性平均是减少通信开销的有效手段, 周期性平均指的是允许客户端在多次本地训练后再上传参数, 而非每轮本地训练结束后均需上传。FedAvg 算法相较 FedSGD 算法增加了本地训练的轮次, 以通信延迟作为代价减少了通信轮次, 从而减少了通信的总数据量。DGA^[82]延迟了梯度平均步骤, 并容许客户端在上传梯度的时候依然进行本地训练。FT-LSGD-DB^[83]使用批量增加方法, 即在训练过程中动态增加数据批量大小, 延长本地训练时间, 在保持收敛率的情况下减少通信轮次。Cooperative SGD 框架^[84]系统讨论了多种局部更新 SGD 算法, 并提供了统一的收敛性分析, 该框架允许客户端在本地执行多次更新并定期同步至服务端来降低通信频率。文献[85]对 DNN 的不同层进行分类, 根据 DNN 的特性将参数更新频率进行差异化处理, 不同深度的层进行不同周期的全局平均, 此外还引入了时间加权聚合策略, 以更有效地整合之前的模型信息。

周期性平均已经成为联邦学习中的默认选择, 其合理性来源于随着时间发展, 深度学习训练损失下降和精度提升的程度会降低, 相近轮次

间的模型差异逐渐减小, 因此在联邦学习后期通过周期性平均来降低通信数据量被认为是可以接受的策略。周期性平均策略的缺陷是客户端无法及时从其余参与方中获得新的知识, 对于数据更新较快和对实时性要求较高的任务并不适用。

4.2 基于异步更新与半异步更新的优化方法

同步更新方式存在面对掉队客户端时系统陷入等待参数上传的问题, 这极大地降低了联邦学习效率, 该问题可以通过异步更新或半异步更新的方式解决, 这两种方式允许在不等待掉队客户端的情况下继续学习。异步更新方式允许客户端在训练结束后立刻上传参数, 并且允许在等待服务端更新参数之前继续本地训练, 而服务端只要收集到足够数量的客户端的参数就进行聚合, 不必等待所有选定的客户端上传完模型参数。

SAFA^[86]是一种异步更新方法, 通过在模型广播、客户端选择和全局聚合阶段引入如下几种机制来减轻通信延迟、客户端掉队和模型陈旧性对全局学习进度的影响: ①滞后容忍性模型分发, 引入滞后容忍性超参数来控制客户端可以接受的本地模型的最大滞后版本, 客户端可以在模型广播不及时的情况下继续使用本地模型进行训练; ②补偿性客户端选择, 服务端优先选择那些在上一轮训练中未被选中的客户端; 区分性聚合, 未被选中的客户端的更新将放入缓存中, 在下一轮聚合中使用这些更新。在 PFL 框架^[87]中, 服务端在接收已完成的客户端更新后立即广播存在服务端缓冲区中上一轮聚合的陈旧的全球模型, 然后进行当前模型的全球聚合, 这使聚合、通信和训练可以并行执行。AOCC-FL^[88]通过校准补偿来对齐本地模型与全局模型, 可以在不恶

化收敛性能的情况下隐藏通信延迟：方向校准利用历史梯度变化信息来校准本地梯度，使其在方向上接近全局梯度；幅度适应使用优化策略自动调整补偿梯度的大小，以最小化近似全局梯度的误差。FedCR^[89]使用对比学习来使本地模型的表示接近全局模型的表示。Overlap-FedAvg^[90]使用了基于费舍尔信息矩阵的数据补偿机制来解决训练和通信并行导致的梯度陈旧性问题。DoCoFL 框架^[91]研究了下行链路的模型陈旧性问题。

FedMDS 框架^[92]是一种半异步更新框架，利用梯度相似性、计算能力、数据量和网络条件将客户端进行聚类为数个客户端组，所有客户端执行异步更新策略。如果某个客户端与组内其他客户端的差异超过阈值，则组内强制执行同步更新策略，减少了客户端掉队现象的发生。Qsparse-local-SGD 算法^[93]结合了稀疏化、量化和误差反馈等技术，并分别实现了其同步和异步形式。

异步更新和半异步更新方式的优势在于允许联邦学习系统中存在因各种原因而掉队的客户端，提升了系统的鲁棒性，并通过减少服务端的等待时间进而减少了通信时间；其缺陷则是该类策略对服务端的调度能力是一个挑战，核心是如何减缓或解决模型陈旧性问题，一些解决模型陈旧性的方法要求服务端有足够空间来缓存陈旧模型信息，对服务端的计算资源也有一定要求。总

体而言，异步和半异步更新方式相对同步更新方式更加适合计算资源和网络异构的边缘环境。表 5 总结了联邦学习中现有的面向模型更新策略的优化方法。

5 面向系统架构的优化方法

联邦学习系统架构的优化主要指对客户端组织与调度策略、网络拓扑结构、系统资源配置等方面进行优化。相较细粒度的模型参数压缩和模型更新优化，系统架构级的优化从更高的角度来描述和理解联邦学习，试图从整体设计上提高联邦学习效率。

5.1 基于客户端选择与调度的优化方法

客户端选择是一种有效管理异构客户端并提升联邦系统运行效率的方法，可以降低部分客户端因为有限的计算资源和波动的网络状态产生的掉队问题对系统效率的影响。

FedCS^[94]利用客户端资源信息主动管理异构客户端，促使联邦系统高效运行，FedCS 设置了客户端模型下载、更新和上传的截止时间，并在这种约束下尽可能选择更多的客户端参与模型聚合，从而在更短的通信时间内获得与原始联邦学习相同精度的模型。FedMral 框架^[95]利用多智能体强化学习（multi-agent reinforcement learning, MARL）来选择客户端以最大化系统奖励，奖励

表 5 面向模型更新策略的优化方法

Table 5 Optimization method for model update strategy

年份	方法	优化方法	特点
2020	文献 ^[85]	周期性平均	时间相关性,空间相关性,模型分层分周期平均
2020	Qsparse-local-SGD ^[93]	异步更新与半异步更新	稀疏化,量化,误差反馈,分别实现其同步和异步形式
2021	DGA ^[82]	周期性平均	周期性平均,加入动量机制
2021	FT-LSGD-DB ^[83]	周期性平均	动态增加批量大小以增加本地训练轮次
2021	Cooperative SGD ^[84]	周期性平均	讨论了多种局部更新 SGD 算法
2021	SAFA ^[86]	异步更新与半异步更新	滞后容忍性模型分发,补偿性客户端选择,区分性聚合
2022	Overlap-FedAvg ^[90]	异步更新与半异步更新	模型训练与模型的上传下载并行
2023	FedMDS ^[92]	异步更新与半异步更新	客户端聚类,组内强制同步更新,组间异步更新
2023	DoCoFL ^[91]	异步更新与半异步更新	研究了下行链路的模型陈旧性问题;服务端保存锚点模型,利用锚点模型间的差值重建全局模型
2023	PFL ^[87]	异步更新与半异步更新	接收客户端更新后立即广播陈旧模型
2023	AOCC-FL ^[88]	异步更新与半异步更新	校准补偿,方向校准和幅度适应
2023	FedCR ^[89]	异步更新与半异步更新	类对比正则化,对比学习

函数反映了模型精度、训练延迟和通信成本，在提高模型准确性的同时，显著降低了训练延迟和通信成本。文献[96]将客户端采样问题视为一个在线学习任务，使用带有多臂老虎机反馈的在线随机梯度下降算法来解决客户端选择问题，目标是最小化采样方差（该方差用于衡量不同客户端采样分布的效果）。FedCor框架^[97]采用基于高斯过程（gaussian process, GP）的客户端选择策略，倾向于选择与其他客户端相关性较大的客户端，能够在每一轮通信中显著降低全局损失。FedCG^[98]是一种基于次模函数最大化和线性规划的迭代算法，以优化客户端选择和参数压缩比的决策表现。聚集采样^[99]对客户端进行基于样本数量的聚类 and 基于模型相似性的聚类，并在集群中对客户端进行采样以减少聚合时模型参数的方差。FedSCP^[100]通过最大均值差异衡量本地模型与全局模型的差异，过滤相关性较低的本地模型，并在服务端根据相关性对本地模型进行加权聚合。QSFL框架^[101]通过双层优化方法来减少上行流量：资格判断算法，服务端只允许对全局模型有正向贡献的客户端上传模型参数；稀疏滑动分段算法，将每个客户端的模型平铺并分割成多个分段，在多轮模型更新轮次中逐段上传，最终确保所有分段都能够上传至服务端。两种方法的结合可以加快模型收敛，减少上行通信成本。

除了对客户端进行选择外，还可以根据客户端的任务承载能力分配不同工作负载来避免掉队问题。FedSAE框架^[102]包含两种自适应客户端资源调度算法：FedSAE-Ira和FedSAE-Fassa。FedSAE-Ira基于客户端在最新一轮中完成任务的历史情况来预测下一轮的可承受工作负载，如果客户端在当前轮次成功完成了分配的工作负载，则下一轮的预测工作负载将增加一个增量，如果客户端未能完成工作负载，则下一轮的工作负载将减半；FedSAE-Fassa则利用过去多轮中的训练完成情况来自适应地预测下一轮的可承受工作负载。Min Average Cost算法^[103]则利用贪心策略分配工作负载来减少掉队现象的发生。

客户端选择和调度虽然降低了掉队问题的影响，但是有可能导致潜在的不公平性：选择和调度策略往往倾向于可以迅速完成学习任务的客户

端，进一步提高了这些客户端数据及模型在全局模型中的权重，虽然可以加速模型的收敛和降低通信开销，但是其产生的后果是模型存在偏向性，这对部分“弱小”的客户端并不公平。因此开发可以保证公平性的客户端选择和调度策略仍然是重要的研究方向之一。

5.2 基于网络拓扑结构设计的优化方法

优化联邦学习中的网络拓扑结构设计有助于对网络流量进行管理和调度，进而最大化通信链路的利用率。

将客户端根据地理位置或者计算资源进行聚类并管理可以高效调度网络上的联邦学习流量。FL+HC^[104]引入聚类方法，根据客户端本地模型更新的相似性将客户端分组，在进行多个轮次的初期联合学习后，不同的客户端集群独立并行训练差异化模型，这种为客户端子集训练专有模型的方式加快了整体联邦学习系统收敛速度，缓减了Non-IID数据对系统通信效率的负面影响。HybridFL^[105]在客户端聚类的基础上考虑了客户端掉队问题，提出了区域松弛因子机制，动态调整不同客户端集群内的客户端采样率，保证每个区域都有足够的客户端参与学习。

在传统的客户端-服务端通信模式下增加边缘服务器形成客户端-边缘服务器-服务端层次化架构是网络拓扑优化的常用方法之一，该架构被称为层次联邦学习。服务端将一些计算和管理功能卸载到边缘服务器上，边缘服务器并行处理这些任务从而提高系统效率，可以有效缓解服务端的通信压力。HFEL^[106]将部分模型聚合的任务卸载至边缘服务器上，实现了低延迟的联邦学习，并制定了以全局学习成本最小化为目标的计算与通信资源分配方案，实现了高效的资源调度。RFL-HA^[107]将客户端聚类为 K 个簇，每个簇内的节点通过同步方式将本地更新转发给簇头进行聚合，所有簇头通过异步方式进行全局聚合，并提出了一种算法来确定最优的簇结构，且将其扩展、应用到动态网络条件下。PSP方法^[108]利用基站进行天然的层次划分，用正交频分多址技术将频带划分为多个子载波，同步进行模型广播和下载，并使用贪心算法来调整载波分配，以最小化通信时间。Hier-Local-QSGD^[109]在客户端-

边缘服务器、边缘服务器-服务端这两次模型聚合阶段采用不同的通信轮次间隔时间,根据训练损失和传播延迟来自适应优化各自通信轮次间隔时间,以减少整体训练时间。HiFlash 框架^[110]进一步优化层次联邦学习,在客户端-边缘服务器利用局域网高带宽低延迟优势采用同步更新方法,提高模型的精度和收敛速度,在边缘服务器-服务端采用异步更新方法,以减少掉队现象和广域网传输的流量。

利用层次联邦学习中的边缘服务器,可以依照不同的准则灵活分配客户端至不同边缘服务器,实现计算和流量上的负载均衡。但是现有的方法大多数是启发式的,对于设计进行理论分析的研究极少。文献^[111]对网络拓扑设计理论进行了探索,定义了跨数据孤岛的联邦学习中的拓扑设计问题,使用最大加线性系统理论来计算系统吞吐量,并提出了两种算法用于在不同网络条件下的拓扑结构设计,该设计可以显著加快跨数据孤岛的联邦学习训练速度。

对客户端进行聚类管理以及层次联邦学习同客户端选择和调度方法一样,容易引发潜在的不公平性,而且边缘服务器的加入使联邦系统的优化问题变得更加复杂,需要综合考虑负载均衡和公平性问题。除此之外,层次联邦学习类方法未充分利用边缘服务器带来的额外优势,除了承担来自服务端的卸载任务,还存在为客户端代理计算任务的可能,以实现系统级的计算和通信等方面的负载均衡,然而类似的研究极少。

5.3 基于去中心化架构改进的优化方法

去中心化架构下通信过程主要发生在客户端与客户端之间,而客户端的稳定性相对服务端有所欠缺,通信优化方法也需要进行适当的改进。QuanTimed-DSGD 算法^[112]结合两种方法来优化通信效率:为每个客户端的本地训练设定截止时间,客户端之间交换量化后的模型参数,前者降低了劣质节点的通信延迟,后者节省了通信开销。GossipFL^[113]使用随机掩码稀疏化模型参数,根据客户端之间的连接带宽动态选择通信对,只与一个同伴进行通信,有效提高了共识效率和带宽利用率。DeLi-CoCo 算法^[114]将单一高精度 Gossip 协议替换为多次压缩形式的 Gossip 协议。

HLSGD^[115]考虑了设备间(device-to-device, D2D)和设备到服务器(device-to-server, D2S)的通信, D2D 为高速链路,而 D2S 为带宽受限链路,将客户端分为数个集群,将集群内部视为去中心化架构,采用 D2D 进行通信。服务端通过 D2S 来广播全局模型给客户端,当训练完成后客户端通过 D2D 链路执行 Gossip 协议进行模型更新,最后每个集群选择部分客户端通过 D2S 链路上传模型参数,由服务端完成全局模型的聚合。FedMigr^[116]允许客户端在本地训练结束后利用深度强化学习自适应地将本地模型迁移到另一个客户端进行聚合和训练,有效利用了客户端之间的局部通信,减少了客户端对服务端的通信依赖。

去中心化架构缺少服务端对联邦系统进行实时调控,缺少灵活性,无法在系统运行时自适应调整,因此去中心化架构需要更多的工作来优化通信效率和提高策略的灵活性。

5.4 基于系统多目标建模的优化方法

将联邦学习系统所要考虑的优化目标进行联合建模,并通过控制参数来调整优化目标是一种常见的面向系统架构的综合优化方法。Nguyen 等^[117]提出了名为 FEDL 的非凸的联合优化问题,使用帕累托效率模型来平衡无线网络中本地计算时间与通信延迟、联邦学习与客户端能源消耗,该方法为理解客户端计算时间、通信延迟和能耗如何影响系统学习时间提供了参考,为提高该环境下的通信效率提供了理论工具。Luo 等^[118]关注联邦学习系统中的成本效益设计,提出了一种自适应调整这两个变量的算法以最小化学习总成本。Cui 等^[119]分析了使用强凸和非凸损失函数的联邦学习中压缩误差对于收敛率的影响,通过这种分析可以确定模型参数压缩率与模型准确率之间的最佳平衡点,并提出了一个根据压缩误差和学习率动态调整压缩率的框架,目标是在有限通信带宽下最大化最终模型的精度。

虽然基于系统多目标建模的优化方法可以给出理论上的通信优化调整方向,但随着联邦学习的不断发展,联邦系统的构成也变得更加复杂,现有的多目标优化方法往往只能关注其中几个目标,组合优化问题的构建也越来越困难,问题的

求解也逐渐从精确解过渡到近似解。如何优化理论分析手段,并充分利用理论指导实践也是目前多目标优化所要面对的问题。表6总结了联邦学习中现有的面向系统架构的优化方法。

6 面向通信协议的优化方法

通信协议是联邦学习模型参数传输机制的重要基础之一,在联邦学习中引入高速通信协议或对现有通信协议进行优化和改进,可以有效降低通信延迟、提升通信效率。

6.1 基于无线网络协议改进的优化方法

处于边缘侧的移动终端设备大多以无线方式

接入网络,无线环境下的联邦学习通信优化更加复杂,需要充分考虑移动设备能源消耗敏感、算力弱、网络环境时刻变动等特性,所提出的通信优化方法需要平衡通信效率提升程度和为优化付出的功耗与计算资源代价。

无线网络以波为载体传输信息的物理特性有助于降低通信开销,宽带模拟聚合(broadband analog aggregation, BAA)^[120]根据无线多接入信道的特性,使用了波形叠加方法在接入点直接对模型进行聚合,这种技术称为载波空中聚合,这种学习方式也称为空中联邦学习,该方法显著降低了通信延迟。BLCD算法^[121]则使用载波空中

表6 面向系统架构的优化方法

Table 6 Optimization methods for system architecture

年份	方法	优化方法	特点
2018	FedCS ^[94]	客户端选择与调度	设置模型下载/训练/上传截止时间,最大化参与学习的客户端数量
2019	QuanTimed-DSGD ^[112]	去中心化架构改进	设定训练截止时间,客户端交换量化后的模型参数
2019	FEDL ^[117]	系统多目标建模	利用帕累托效率模型平衡本地计算时间与通信延迟、联邦学习与客户端能源消耗
2020	FL+HC ^[104]	网络拓扑结构设计	客户端聚类,多任务并行
2020	HFEL ^[106]	网络拓扑结构设计	层次联邦学习,将模型聚合任务卸载至边缘服务器
2020	文献[111]	网络拓扑结构设计	网络拓扑设计优化理论
2021	聚集采样 ^[99]	客户端选择与调度	基于样本数量的聚类采样,基于模型相似性聚类采样
2021	FedSAE ^[102]	客户端选择与调度	自适应客户端资源调度,动态分配工作负载
2021	Min Average Cost ^[103]	客户端选择与调度	利用贪心算法分配工作负载
2021	RFL-HA ^[107]	网络拓扑结构设计	层次联邦学习,客户端聚类,最优化簇结构
2021	HybridFL ^[105]	网络拓扑结构设计	客户端聚类,区域松弛因子,动态采样
2021	文献[118]	系统多目标建模	建模优化参与客户端的数量和每轮本地迭代次数两个变量
2022	FedMral ^[95]	客户端选择与调度	利用多智能体强化学习最大化奖励函数
2022	FedCor ^[97]	客户端选择与调度	基于高斯过程对客户端进行选择
2022	QSFL ^[101]	客户端选择与调度	两层优化方法,资格判断,稀疏滑动分段算法
2022	PSP ^[108]	网络拓扑结构设计	层次联邦学习,无线信道划分,贪心算法分配资源
2022	DeLi-CoCo ^[114]	去中心化架构改进	多次压缩形式的Gossip协议
2022	HLSGD ^[115]	去中心化架构改进	客户端聚类,集群内去中心化通信
2022	FedMigr ^[116]	去中心化架构改进	深度强化学习,客户端间迁移模型
2022	文献[119]	系统多目标建模	根据每次全局迭代的压缩误差和学习率动态调整压缩率
2023	文献[96]	客户端选择与调度	强化学习
2023	FedCG ^[98]	客户端选择与调度	基于次模函数最大化和线性规划的迭代算法
2023	Hier-Local-QSGD ^[109]	网络拓扑结构设计	层次联邦学习,差异化两阶段聚合的间隔时间
2023	HiFlash ^[110]	网络拓扑结构设计	层次联邦学习,同步更新和异步更新结合
2023	GossipFL ^[113]	去中心化架构改进	随机掩码稀疏化模型参数,根据带宽动态选择通信对象
2024	FedSCP ^[100]	客户端选择与调度	根据本地模型与全局模型的相关性过滤低相关性客户端

聚合的方法上传压缩参数,进一步压缩通信数据量。文献[122]中的方法将空中联邦学习和非正交多址接入(non-orthogonal multiple access, NOMA)整合到同一框架内,提出一个非线性、非凸的涉及通信速率的混合优化问题,通过交替优化算法解决复杂的资源配置问题,支持按需通信,为无线环境下的空中联邦学习通信效率优化问题提供了系统的理论分析。

空中联邦学习对无线网络环境下的通信进行了优化,但是联邦学习中仍存在大量的有线连接设备,两类网络环境下的通信优化方法并非完全兼容,空中联邦学习的效率也严重依赖于边缘无线环境的稳定程度。在有线网络中引入类似空中载波聚合技术的基于网络编码的优化方法可以有效利用物理信道特点,此类研究目前仍然较少。

6.2 基于高速通信协议的优化方法

现有研究极少会针对底层的通信协议进行优化,在联邦学习中引入高带宽、低延迟的通信协议或对现有协议进行适应性调整是潜在研究方向之一。高速通信协议通常指的是能够支持高数据传输速率的通信标准或技术,它们能够实现在网络或设备之间快速、高效地传输大量数据,具有高带宽、低延迟、可靠传输等特性。例如远程直接地址访问(remote direct memory access, RDMA)技术,使用了零复制、内核旁路和CPU offload等技术,允许网络接口直接在应用程序的内存中操作,绕过操作系统网络协议栈,从而提供更高的带宽和更低的延迟,同时节省了CPU资源。文献[123]针对纵向联邦学习中的内部通信效率问题提出了基于RDMA技术的通信优化方案,用来替代现有的基于gRPC技术的通信方案。

一些研究考虑了联邦学习中网络拥塞或丢包问题,通过设置合理的拥塞控制方法可以避免频繁重传带来的高延迟状态。NAC-FL^[124]利用动态调整客户端有损压缩比例以适应网络拥塞变化,在每轮训练中,服务端根据当前的网络拥塞状态来调整每个客户端的压缩参数,以最小化整体的通信延迟。FedLC^[125]对网络丢包问题进行了研究,在传输模型参数时使用前向纠错机制,在传输参数时附加含有冗余信息的数据包,即使在丢包的情况下,接收方也能够

恢复出原始的数据;当服务端检测到模型更新的数据包丢失时会通知客户端重传丢失的数据包。Sun等^[125]使用凸优化和整数规划方法找到最优的压缩率、冗余率和重传次数,这些机制保证了FedLC在面对不稳定的网络条件时仍然可以完成联合建模。

在联邦学习中引入高速通信协议、灵活的拥塞控制方法和重传机制并非简单的事,一些技术需要特定的硬件支持,例如RDMA技术需要特定型号的网卡支持,由于专用硬件的成本和功耗限制,往往只能在边缘服务器和服务端之间有限部署,然而联邦学习中的通信延迟大多发生在靠近客户端的一侧,因此仍需对客户端侧的高速通信协议优化方法进行探究以兼容现有的边缘计算环境。一类可行的方法是对云环境和物联网环境中的高效通信协议进行联邦学习环境的改造,以适应分布式模型训练,如寇文龙等^[126]提出的支持差异化可协商的数据通信机制可以根据接收端能力的差异进行动态自适应的高效并行通信,其技术对联邦学习中的通信协议改进有着启发意义。表7总结了联邦学习中现有的面向通信协议的通信优化方法。

7 研究展望

联邦学习作为一个新兴的面向隐私保护的分布式机器学习概念,为分布式机器学习提供了一定的数据安全保证,但其通信优化方法的研究刚刚起步,图5展示了联邦学习中通信优化研究发展脉络。由于联邦学习系统及其应用环境的复杂性,通信优化方法仍需要进一步完善,下面这些研究方向值得深入探索。

(1) 面向网络安全场景的通信优化方法

随着针对深度学习的攻击手段不断发展和丰富,作为更加复杂的联邦学习系统同样面临网络安全威胁,比如来自半诚实敌手的模型反演攻击和成员推理攻击,来自恶意敌手的身份仿冒攻击、通信窃听与篡改等,现有的研究较少关注网络安全保证下的通信优化方法,该方向具有较高的研究价值。联邦学习中主要关注的安全问题为隐私数据泄露,李凤华等^[127]对联邦学习作出深刻见解,认为其本质是分布式模型训练架构,其交换的信息仍然存在数据泄露问题,无法有效保

表7 面向通信协议的优化方法

Table 7 Optimization methods for communication protocols

年份	方法	优化方法	特点
2020	宽带模拟聚合 ^[120]	无线网络协议改进	载波空中聚合
2020	文献[123]	高速通信协议	利用RDMA技术传输模型参数,动态分配RDMA带宽
2021	BLCD ^[121]	无线网络协议改进	载波空中聚合,模型参数压缩
2021	文献[122]	无线网络协议改进	载波空中聚合,建立混合优化问题,动态资源配置
2023	NAC-FL ^[124]	高速通信协议	优化网络拥塞下的通信表现,动态调整参数压缩比例
2023	FedLC ^[125]	高速通信协议	使用前向纠错机制与选择重传机制,优化压缩率、冗余率和重传次数,优化网络丢包状况下的通信性能

护数据和隐私安全。现有联邦学习研究通常采用差分隐私、同态加密和多方安全计算等技术手段加强隐私数据的机密性,但是这些安全技术由于会产生额外噪声或占用计算资源进而导致联邦学习的通信效率下降,因此通信高效的隐私保护技术是当前较为迫切的研究方向之一。

(2) 面向数据中心和大规模联邦学习场景的通信优化方法

现有研究缺乏对大规模客户端场景下的服务端计算和通信承载能力进行评估,大多数研究默认服务端的网络环境是稳定的且能承受足够大的负载,然而随着参与联邦学习的客户端的数量、计算能力和数据量的急剧增长,服务端所处的数据中心需要承担的計算和通信任务可能比预期要多得多,任务也随着需求的变化而变得更加复杂,因此需要对服务端所处的数据中心面对大规模联邦学习任务流量时采取的通信优化方法做出前瞻性研究。要实现该目标首先需要大规模联邦学习场景下的数据中心通信任务进行梳理,确认可能存在的瓶颈,建立通信效率评估机制,并进一步探索面向服务端的通信优化方法。服务端侧的通信压力可能出现在边缘服务器与服务端之间、数据中心内部不同服务端之间。对于前者需要研究和优化高速通信协议来提升带宽、降低延迟,同时也可以从拥塞控制策略或流量控制策略着手,优先放行对全局模型贡献较大的流量;对于后者,需要研究面向联邦学习任务的负载均衡机制,并设计面向服务端的低延迟、高效率的联邦学习任务负载均衡框架。

(3) 面向公平性保证场景的通信优化方法

Non-IID数据场景是联邦学习实际落地必然

要面对的问题,Non-IID数据场景下不同客户端之间数据的统计异质性将严重降低联合建模效率,联邦学习系统需要使用更多的通信轮次来保证全局模型的质量。一些使用基于客户端选择策略和基于客户端模型参数相似性进行聚类的通信优化方法导致了潜在的不公平性,全局模型将偏向于被选中的客户端或聚类中成员数量较多的客户端集群。不公平性还体现在最终的参与方的收益上,联邦学习有效促进了数据的增值和衍生数据的产生^[128],联邦学习组织者需要公平分配模型与数据收益以促进联合学习。现有研究普遍不重视这种不公平性,因此在保证公平性、模型准确性和高效通信的前提下寻找平衡的通信优化方法是极具研究价值的方向之一。

8 结束语

本文对联邦学习中的通信优化方法进行了全面的阐述与总结。通信优化的总体目标覆盖范围较广,除了包含基础的降低通信数据总量和通信延迟等目标,还有减少通信轮次和总体通信时间等目标。本文将优化方法分为基于参数压缩的优化方法、面向模型更新策略的优化方法、面向系统架构的优化方法和面向通信协议的优化方法4类,对现有工作进行了简要介绍和分析,在最后总结了现有研究的缺陷,并对未来的研究工作进行了展望。

大部分通信优化方法需要付出额外代价,例如,以通信延迟的增加换取通信轮次的减少,各类参数压缩方法带来了额外计算负担,潜在地降低了模型训练速度等。大多参与联邦学习的边缘设备需要考虑能量消耗和网络带宽分配等资源配置问题,对通信优化所需耗费的计算资源较为敏

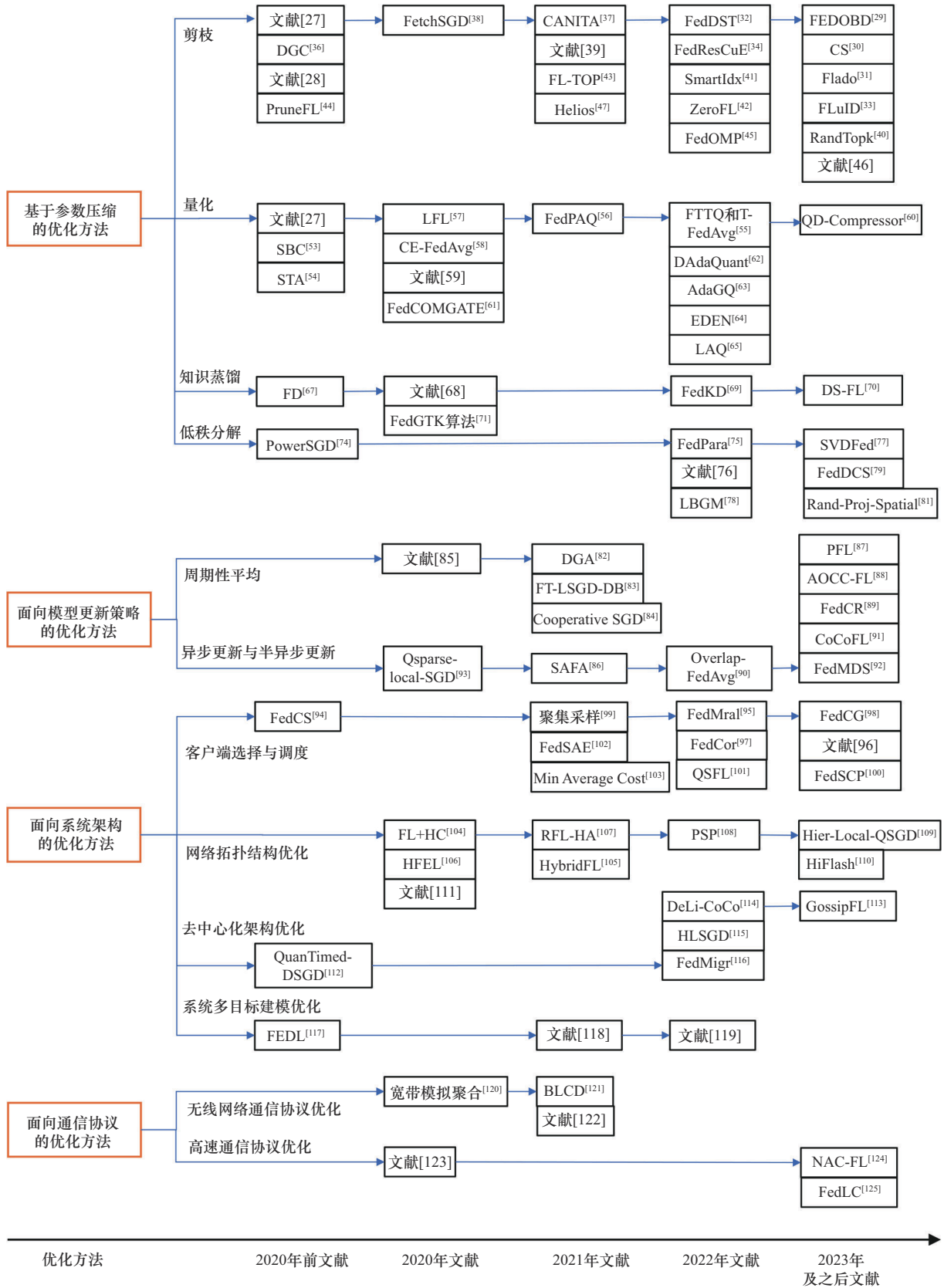


图5 联邦学习中通信优化研究发展脉络

Figure 5 Development of communication optimization research in federated learning

感, 这些额外的代价可能在边缘侧被放大, 研究人员需要全面且慎重考虑边缘环境的特殊性。

总体而言, 联邦学习的通信优化是一个复杂的研究问题, 需要结合实际环境综合考虑各类优化方法的应用场景及其之间的平衡与结合, 通信高效的联邦学习离真正的实际落地和广泛应用还有一段路要走。

参考文献:

- [1] 李风华, 李晖, 贾焰, 等. 隐私计算研究范畴及发展趋势[J]. 通信学报, 2016, 37(4):1-11.
LI F H, LI H, JIA Y, et al. Privacy computing: concept, connotation and its research trend[J]. Journal on Communications, 2016, 37(4): 1-11.
- [2] CUSTERS B, SEARS A M, DECHESNE F, et al. EU personal data protection in policy and practice[J]. Information Technology and Law Series, 2019.
- [3] 李昕, 唐鹏, 张西珩, 等. 面向GDPR隐私政策合规性的智能化检测方法[J]. 网络与信息安全学报, 2023, 9(6):127-139.
LI X, TANG P, ZHANG X Y, et al. GDPR-oriented intelligent checking method of privacy policies compliance[J]. Chinese Journal of Network and Information Security, 2023, 9(6):127-139.
- [4] MCMAHAN H B, MORRE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]//International Conference on Artificial Intelligence and Statistics, 2017, 54: 1273-1282.
- [5] HARD A, RAO K, MATHEWS R, et al. Federated learning for mobile keyboard prediction[J]. Computing Research Repository, 2018, abs/1811.03604.
- [6] ZHENG W, YAN L, GAO C, et al. Federated meta-learning for fraudulent credit card detection[C]//International Joint Conference on Artificial Intelligence, 2020: 4654-4660.
- [7] STEFANIE W, HARTMUT S, KRISHNAPRASAD L S, et al. Swarm learning for decentralized and confidential clinical machine learning[J]. Nature, 2021, 594(7862): 265-270.
- [8] SHELLER M J, EDWARDS B, REINA G A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data[J]. Scientific Reports, 2020, 10(1): 12598.
- [9] DAYAN I, ROTH H R, ZHONG A, et al. Federated learning for predicting clinical outcomes in patients with COVID-19[J]. Nature Medicine, 2021, 27(10): 1735-1743.
- [10] NGUYEN T D, MARCHAL S, MIETTINEN M, et al. Dfot: a federated self-learning anomaly detection system for IoT[C]//2019 IEEE 39th International Conference on Distributed Computing Systems, 2019: 756-767.
- [11] ZHANG D, KOU Z, WANG D. FedSens: a federated learning approach for smart health sensing with class imbalance in resource constrained edge computing[C]//IEEE Conference on Computer Communications, 2021: 1-10.
- [12] YE Y, ZHAO W, WEI T, et al. FedLight: federated reinforcement learning for autonomous multi-intersection traffic signal control[C]//Design Automation Conference, 2021: 847-852.
- [13] YANG Q, LIU Y, TONG Y, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 12:1-12:19.
- [14] ROY A G, SIDDIQUI S, POLSTERL S, et al. BrainTorrent: a peer-to-peer environment for decentralized federated learning[J]. arXiv: Learning, 2019, abs/1905.06731.
- [15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, abs/1512.03385(1): 770-778.
- [16] MOTHUKURI V, PARIZI R M, POURIYEH S, et al. A survey on security and privacy of federated learning[J]. Future Generation Computer Systems, 2021, 115: 619-640.
- [17] SHI Y, YANG K, JIANG T, et al. Communication-efficient edge AI: algorithms and systems[J]. IEEE Communications Surveys and Tutorials, 2020, 22(4): 2167-2191.
- [18] ZHU H, XU J, LIU S, et al. Federated learning on non-IID data: a survey[J]. Neurocomputing, 2021, 465: 371-390.
- [19] YANN L, DENKER J S, SOLLA S A. Optimal brain damage[C]//Conference on Neural Information Processing Systems, 1989: 598-605.
- [20] DENIL M, SHAKIBI B, DINH L, et al. Predicting parameters in deep learning[C]//Conference on Neural Information Processing Systems, 2014, abs/1306.0543: 2148-2156.
- [21] HANSON S J, PRATT L Y. Comparing biases for minimal network construction with back-propagation[J]. Advances in Neural Information Processing Systems, 1988, 1: 177-185.
- [22] LI H, KADAV A, DURDANOVIC L, et al. Pruning filters for efficient ConvNets[C]//International Conference on Learning Representations, 2017, abs/1608.08710.
- [23] LUO J, WU J, LIN W. ThiNet: a filter level pruning method for deep neural network compression[C]//2017 IEEE International Conference on Computer Vision, 2017, abs/1707.06342: 5068-5076.
- [24] ZHOU H, ALVAREZ J M, PORIKIL F. Less is more: Towards compact CNNs[J]. Lecture Notes in Computer Science, 2016, 9908: 662-677.
- [25] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural network[J]. Advances in Neural Information Processing Systems, 2015, 28: 1135-1143.
- [26] ZHOU A, MA Y, ZHU J, et al. Learning N:M fine-grained structured sparse neural networks from scratch[C]//International Conference on Learning Representations, 2021, abs/2102.04010.
- [27] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency[J]. Computing Research Repository, 2016, abs/1610.05492.
- [28] CALDAS S, KONEČNÝ J, MCMAHAN H B, et al. Expanding the reach of federated learning by reducing client resource requirements[J]. Computing Research Repository, 2018, abs/1812.07210.
- [29] CHEN Y, CHEN Z, YU P, et al. FEDOBD: opportunistic block dropout for efficiently training large-scale neural networks through federated learning[C]//International Joint Conference on Artificial Intelligence, 2023: 3541-3549.
- [30] JIANG X, BORCEA C. Complement sparsification: low-overhead model pruning for federated learning[C]//AAAI Conference on Arti-

- ficial Intelligence, 2023, 37(7): 8087-8095.
- [31] LIAO D, GAO X, ZHAO Y, et al. Adaptive channel sparsity for federated learning under system heterogeneity[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 20432-20441.
- [32] BIBIKAR S, VIKALO H, WANG Z, et al. Federated dynamic sparse training: computing less, communicating less, yet learning better[C]//AAAI Conference on Artificial Intelligence, 2022, abs/2112.09824: 6080-6088.
- [33] WANG I, NAIR P J, MAHAJAN D. FLuID: mitigating stragglers in federated learning using invariant dropout[C]//Annual Conference on Neural Information Processing Systems, 2023, abs/2307.02623.
- [34] ZHU Z, HONG J, DREW S, et al. Resilient and communication efficient learning for heterogeneous federated systems[C]//International Conference on Machine Learning, 2022, 162: 27504-27526.
- [35] SRINIVAS S, BABU R V. Data-free parameter pruning for deep neural networks[C]//British Machine Vision Conference, 2015, abs/1507.06149.
- [36] LIN Y, HAN S, MAO H, et al. Deep gradient compression: reducing the communication bandwidth for distributed training[C]//International Conference on Learning Representations, 2017, abs/1712.01887.
- [37] LI Z, RICHTÁRIK P. CANITA: faster rates for distributed convex optimization with communication compression[C]//Conference on Neural Information Processing Systems, 2021, abs/2107.09461: 13770-13781.
- [38] ROTHCHILD D, PANDA A, ULLAH E, et al. FetchSGD: communication-efficient federated learning with sketching[C]//International Conference on Machine Learning, 2020, abs/2007.07682: 8253-8265.
- [39] JHUNJHUNWALA D, MALLICK A, GADHIKAR A, et al. Leveraging spatial and temporal correlations in sparsified mean estimation[C]//Conference on Neural Information Processing Systems, 2021, abs/2110.07751: 14280-14292.
- [40] ZHENG F, CHEN C, LYU L, et al. Reducing communication for split learning by randomized Top- k sparsification[C]//International Joint Conference on Artificial Intelligence, 2023, abs/2305.18469: 4665-4673.
- [41] WU D, ZOU X, ZHANG S, et al. SmartIdx: reducing communication cost in federated learning by exploiting the CNNs structures[C]//AAAI Conference on Artificial Intelligence, 2022, 36(4): 4254-4262.
- [42] QIU X, FERNANDEZ-MARQUES J, GUSMAO P P, et al. ZeroFL: efficient on-device training for federated learning with local sparsity[C]//International Conference on Learning Representations, 2022.
- [43] KERKOUCHER, ÁCS G, CASTELLUCCIAC, et al. Constrained differentially private federated learning for low-bandwidth devices[C]//Conference on Uncertainty in Artificial Intelligence, 2021, abs/2103.00342: 1756-1765.
- [44] JIANG Y, WANG S, KO B J, et al. Model pruning enables efficient federated learning on edge devices[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, abs/1909.12326(12): 10374-10386.
- [45] JIANG Z, XU Y, XU H, et al. FedMP: federated learning through adaptive model pruning in heterogeneous edge computing[J]. Proceedings. International Conference on Data Engineering, 2022: 767-779.
- [46] SHAH S M, LAU V K N. Model compression for communication efficient federated learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(9): 5937-5951.
- [47] XU Z, YU F, XIONG J, et al. Helios: Heterogeneity-aware federated learning with dynamically balanced collaboration[C]//Design Automation Conference, 2021, 2021: 997-1002.
- [48] ZHANG D, YANG J, YE D, et al. Lq-nets: Learned quantization for highly accurate and compact deep neural networks[J]. Computer Vision - ECCV 2018, PT VIII, 2018, 11212: 373-390.
- [49] BANNER R, NAHSHAN Y, SOUDRY D. Post training 4-bit quantization of convolutional networks for rapid-deployment[J]. Advances in Neural Information Processing Systems, 2019, 32: 7948-7956.
- [50] COURBARIAUX M, BENGIO Y, DAVID J P. Binary connect: training deep neural networks with binary weights during propagations[J]. Advances in Neural Information Processing Systems, 2015, 28: 3123-3131.
- [51] COURBARIAUX M, HUBARA I, SOUDRY D, et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to 1 or 0[J]. arXiv: Learning, 2016.
- [52] LI F, LIU B. Ternary weight networks[J]. IEEE International Conference on Acoustics, Speech and Signal Processing, 2016, abs/1605.04711: 1-5.
- [53] SATTLER F, WIEDEMANN S, MÜLLER K, et al. Sparse binary compression: towards distributed deep learning with minimal communication[J]. 2019 International Joint Conference on Neural Networks, 2018, abs/1805.08768: 1-8.
- [54] SATTLER F, WIEDEMANN S, MULLER K, et al. Robust and communication-efficient federated learning from Non-I. I. D. data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, (99):1-14.
- [55] XU J, DU W, JIN Y, et al. Ternary compression for communication-efficient federated learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(3): 1162-1176.
- [56] REISIZADEH A, MOKHTARI A, HASSANI H, et al. FedPAQ: a communication-efficient federated learning method with periodic averaging and quantization[C]//International Conference on Artificial Intelligence and Statistics, 2019, 108: 2021-2031.
- [57] AMIRI M M, GUNDUZ D, KULKARNI S R, et al. Federated learning with quantized global model updates[J]. Openalex, 2020, abs/2006.10672.
- [58] MILLS J, HU J, MIN G. Communication-efficient federated learning for wireless edge intelligence in IoT[J]. IEEE Internet of Things Journal, 2020, 7(7): 5986-5994.
- [59] SHLEZINGER N, CHEN M, ELDAR Y C, et al. Federated learning with quantization constraints[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing, 2020: 8851-8855.
- [60] JIN H, WU D, ZHANG S, et al. Design of a quantization-based

- DNN delta compression framework for model snapshots and federated learning[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2023, 34(3): 923-937.
- [61] HADDADPOUR F, KAMANI M M, MOKHTARI A, et al. Federated learning with compression: unified analysis and sharp guarantees[C]//*International Conference on Artificial Intelligence and Statistics*, 2020, 130: 2350-2358.
- [62] HOENIG R, ZHAO Y, MULLINS R. DAdaQuant: doubly-adaptive quantization for communication-efficient federated learning[C]//*International Conference on Machine Learning*, 2022, abs/2111.00465: 8852-8866.
- [63] LIU H, HE F, CAO G. Communication-efficient federated learning for heterogeneous edge devices based on adaptive gradient quantization[C]. *IEEE Conference on Computer Communications*, 2022: 1-10.
- [64] VARGAFTIK S, BASAT R B, PORTNOY A, et al. EDEN: communication-efficient and robust distributed mean estimation for federated learning[C]//*International Conference on Machine Learning*, 2022: 21984-22014.
- [65] SUN J, CHEN T, GIANNAKIS G B, et al. Lazily aggregated quantized gradient innovation for communication-efficient federated learning[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(4): 2031-2044.
- [66] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. *arXiv: Machine Learning*, 2015, abs/1503.02531.
- [67] JEONG E, OH S, KIM H, et al. Communication-efficient on-device machine learning: federated distillation and augmentation under Non-IID private data[J]. *Computing Research Repository*, 2018, abs/1811.11479.
- [68] BISTRITZ I, MANN A, BAMBOS N. Distributed distillation for on-device learning[C]//*Conference on Neural Information Processing Systems*, 2020, 33: 22593-22604.
- [69] WU C, WU F, LYU L, et al. Communication-efficient federated learning via knowledge distillation[J]. *Nature Communications*, 2022, 13(1): 2032-8.
- [70] ITAHARA S, NISHIO T, KODA Y, et al. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with Non-IID private data[J]. *IEEE Transactions on Mobile Computing*, 2023, 22(1): 191-205.
- [71] HE C, ANNAVARAM M, AVESTIMEHR S. Group knowledge transfer: federated learning of large CNNs at the edge[C]//*Conference on Neural Information Processing Systems*, 2020, 33: 14068-14080.
- [72] HAMER J, MOHRI M, SURESH A T. FedBoost: communication-efficient algorithms for federated learning[C]//*International Conference on Machine Learning*, 2020, 119: 3973-3983.
- [73] ALMEIDA A, FAVIER G, MOTA J C M. Tensor decompositions and applications to wireless communication systems[C]//*IEEE International Telecommunications Symposium*, 2006.
- [74] VOGELS T, KARINIREDDY S P, JAGGI M. Powersgd: practical low-rank gradient compression for distributed optimization[J]. *Advances in Neural Information Processing Systems*, 2019, 32: 14236-14245.
- [75] NAM H, MOON Y, OH T. FedPara: low-rank hadamard product for communication-efficient federated learning[C]//*International Conference on Learning Representations*, 2022.
- [76] CHEN B, BAKHSI ALI, BATISTA G, et al. Update compression for deep neural networks on the edge[C]//*Computer Vision and Pattern Recognition*, 2022: 3075-3085.
- [77] WANG H, LIU X, NIU J, et al. SVD Fed: enabling communication-efficient federated learning via singular-value-decomposition[C]//*IEEE Conference on Computer Communications*, 2023: 1-10.
- [78] AZAM S S, HOSSEINALIPOUR S, QIU Q, et al. Recycling model updates in federated learning: are gradient subspaces low-rank[C]//*International Conference on Learning Representations*, 2022.
- [79] GUAN Y, LIU X, REN T, et al. Enabling communication-efficient federated learning via distributed compressed sensing[C]//*IEEE Conference on Computer Communications*, 2023: 1-10.
- [80] SONG Z, WANG Y, YU Z, et al. Sketching for first order method: efficient algorithm for low-bandwidth channel and vulnerability[C]//*International Conference on Machine Learning*, 2023: 32365-32417.
- [81] JIANG S, SHARMA P, JOSHI G. Correlation aware sparsified mean estimation using random projection[C]//*Conference on Neural Information Processing Systems*, 2023, abs/2310.18868.
- [82] ZHU L, LIN H, LU Y, et al. Delayed gradient averaging: tolerate the communication latency for federated learning[C]//*Conference on Neural Information Processing Systems*, 2021, 34: 29995-30007.
- [83] LI L, SHI D, HOU R, et al. To talk or to work: flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices[C]//*IEEE Conference on Computer Communications*, 2021, abs/2012.11804: 1-10.
- [84] WANG J, JOSHI G. Cooperative SGD: A unified framework for the design and analysis of local-update SGD algorithms[J]. *Journal of Machine Learning Research*, 2021, 22(1): 1-50.
- [85] CHEN Y, SUN X, JIN Y. Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(10): 4229-4238.
- [86] WU W, HE L, LIN W, et al. SAFA: A Semi-asynchronous protocol for fast federated learning with low overhead[J]. *IEEE Transactions on Computers*, 2021, 70(5): 655-668.
- [87] ZHANG F, LIU X, LIN S, et al. No one idles: efficient heterogeneous federated learning with parallel edge and server computation[C]//*International Conference on Machine Learning*, 2023: 41399-41413.
- [88] WANG H, XU W, FAN Y, et al. AOCC-FL: Federated learning with aligned overlapping via calibrated compensation[C]//*IEEE Conference on Computer Communications*, 2023: 1-10.
- [89] CHEN R, WAN Q, PRAKASH P, et al. Workie-Talkie: accelerating federated learning by overlapping computing and communications via contrastive regularization[C]//*IEEE International Conference on Computer Vision*, 2023: 16953-16963.
- [90] ZHOU Y, QING Y, LV J. Communication-efficient federated learning with compensated overlap-FedAvg[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2022, 33(1): 192-205.
- [91] DORFMAN R, VARGAFTIK S, BEN-ITZHAK Y, et al. DoCoFL: downlink compression for cross-device federated learning[C]//*International Conference on Machine Learning*, 2023: 8356-8388.

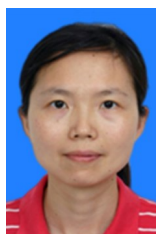
- [92] ZHANG Y, LIU D, DUAN M, et al. FedMDS: an efficient model discrepancy-aware semi-asynchronous clustered federated learning framework[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2023, 34(3): 1007-1019.
- [93] BASU D, DATA D, KARAKUS C, et al. Qsparse-local-SGD: distributed SGD with quantization, sparsification, and local computations[J]. *IEEE Journal on Selected Areas in Information Theory*, 2020, 1(1): 14668-14679.
- [94] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge[C]//2019 IEEE International Conference on Communications, 2018, abs/1804.08333: 1-7.
- [95] ZHANG S Q, LIN J, ZHANG Q. A Multi-agent reinforcement learning approach for efficient client selection in federated learning[C]//AAAI Conference on Artificial Intelligence, 2022: 9091-9099.
- [96] ZHAO B, LIU Z, CHEN C, et al. Adaptive client sampling in federated learning via online learning with bandit feedback[C]. *International Conference on Learning Representations*, 2021, abs/2112.14332.
- [97] TANG M, NING X, WANG Y, et al. FedCor: correlation-based active client selection strategy for heterogeneous federated learning[J]. *Semanticscholar*, 2022(1): 10092-10101.
- [98] JIANG Z, XU Y, XU H, et al. Heterogeneity-aware federated learning with adaptive client selection and gradient compression[C]//IEEE Conference on Computer Communications, 2023: 1-10.
- [99] FRABONI Y, VIDAL R, KAMENI L, et al. Clustered sampling: low-variance and improved representativity for clients selection in federated learning[C]//International Conference on Machine Learning, 2021, 139: 3407-3416.
- [100] 李群, 陈思光. 基于选择性通信策略的高效联邦学习研究[J]. *小型微型计算机系统*, 2024, 45(3):549-554.
- LI Q, CHEN S G. FedSCP: efficient federated learning based on selective communication policy[J]. *Journal of Chinese Computer Systems*, 2024, 45(3):549-554.
- [101] YI L, WANG G, LIU X. QSFL: A two-level uplink communication optimization framework for federated learning[C]//International Conference on Machine Learning, 2022: 25501-25513.
- [102] LI L, DUAN M, LIU D, et al. FedSAE: A novel self-adaptive federated learning framework in heterogeneous systems[C]//IEEE International Joint Conference on Neural Network, 2021, abs/2104.07515: 1-10.
- [103] WANG C, YANG Y, ZHOU P. Towards Efficient scheduling of federated mobile devices under computational and statistical heterogeneity[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2021, 32(2): 394-410.
- [104] BRIGGS C, FAN Z, ANDRAS P. Federated learning with hierarchical clustering of local updates to improve training on non-IID data[C]. *IEEE International Joint Conference on Neural Network*, 2020, abs/2004.11791: 1-9.
- [105] WU W, HE L, LIN W, et al. Accelerating federated learning over reliability-agnostic clients in mobile edge computing systems[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2021, 32(7): 1539-1551.
- [106] LUO S, CHEN X, WU Q, 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.
- [107] WANG Z, XU H, LIU J, et al. Resource-efficient federated learning with hierarchical aggregation in edge computing[C]. *IEEE Conference on Computer Communications*, 2021: 1-10.
- [108] QU Z, GUO S, WANG H, et al. Partial synchronization to accelerate federated learning over relay-assisted edge networks[J]. *IEEE Transactions on Mobile Computing*, 2022, 21(12): 4502-4516.
- [109] LIU L, ZHANG J, SONG S, et al. Hierarchical federated learning with quantization: convergence analysis and system design[J]. *IEEE Transactions on Wireless Communications*, 2023, 22(1): 2-18.
- [110] WU Q, CHEN X, TAO O, et al. HiFlash: communication-efficient hierarchical federated learning with adaptive staleness control and heterogeneity-aware client-edge association[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2023, 34(5): 1560-1579.
- [111] MARFOQ O, XU C, NEGLIA G, et al. Throughput-optimal topology design for cross-silo federated learning[C]//Conference on Neural Information Processing Systems, 2020, 33: 19478-19487.
- [112] REISIZADEH A, TAHERI H, MOKHTARI A, et al. Robust and communication-efficient collaborative learning[J]. *Advances in Neural Information Processing Systems*, 2019, 32: 8386-8397.
- [113] TANG Z, SHI S, LI B, et al. GossipFL: a decentralized federated learning framework with sparsified and adaptive communication [J]. *IEEE Transactions on Parallel and Distributed Systems*, 2023, 34(3): 909-922.
- [114] HASHEMI A, ACHARYA A, DAS R, et al. On the benefits of multiple gossip steps in communication-constrained decentralized federated learning[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2022, 33(11): 2727-2739.
- [115] GUO Y, SUN Y, HU R, et al. Hybrid local SGD for federated learning with heterogeneous communications[C]//International Conference on Learning Representations, 2022.
- [116] LIU J, XU Y, XU H, et al. Enhancing federated learning with intelligent model migration in heterogeneous edge computing[C]//IEEE International Conference on Data Engineering, 2022: 1586-1597.
- [117] NGUYEN H T, BAO W, NGUYEN M N, et al. Federated learning over wireless networks: optimization model design and analysis[C]//IEEE Conference on Computer Communications, 2019: 1387-1395.
- [118] LUO B, LI X, WANG S, et al. Cost-effective federated learning design[C]//IEEE Conference on Computer Communications, 2021, abs/2012.08336: 1-10.

- [119] CUI L, SU X, ZHOU Y, et al. Optimal rate adaption in federated learning with compressed communications[C]//IEEE Conference on Computer Communications, 2022, abs/2112.06694: 1459-1468.
- [120] ZHU G, WANG Y, HUANG K. Broadband analog aggregation for low-latency federated edge learning[J]. IEEE Transactions on Wireless Communications, 2020, 19(1): 491-506.
- [121] ZHANG J, LI N, DEDEOGLU M. Federated learning over wireless networks: a band-limited coordinated descent approach[C]//IEEE Conference on Computer Communications, 2021, abs/2102.07972 (2021): 1-10.
- [122] NI W, LIU Y, YANG Z, et al. Over-the-air federated learning and non-orthogonal multiple access unified by reconfigurable intelligent surface[C]//IEEE Conference on Computer Communications, 2021: 1-6.
- [123] LIU D. Accelerating intra-party communication in vertical federated learning with RDMA[C]//Conference on Emerging Network Experiment and Technology, 2020: 14-20.
- [124] HEGDE P, VECIANA G D, MOKHTARI A. Network adaptive federated learning: congestion and lossy compression[C]//IEEE Conference on Computer Communications, 2023: 1-10.
- [125] SU X, ZHOU Y, CUI L, et al. On model transmission strategies in federated learning with lossy communications[J]. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(4): 1173-1185.
- [126] 寇文龙, 李凤华, 董秀则, 等. 支持差异化可协商的数据通信机制[J]. 通信学报, 2021, 42(10):55-66.
KOU W L, LI F H, DONG X Z, et al. Differentiated and negotiable mechanism for data communication[J]. Journal on Communications, 2021, 42(10):55-66.
- [127] 李凤华, 李晖, 牛犇, 等. 隐私计算的学术内涵与研究趋势[J]. 网络与信息安全学报, 2022, 8(6):1-8.
LI F H, LI H, NIU B, et al. Academic connotation and research trends of privacy computing[J]. Chinese Journal of Network and Information Security, 2022, 8(6):1-8.
- [128] 李凤华, 李晖, 牛犇, 等. 数据要素流通与安全的研究范畴与未来发展趋势[J]. 通信学报, 2024, 45(5):1-11.
LI F H, LI H, NIU B, et al. Research category and future development trend of data elements circulation and security[J]. Journal on Communications, 2024, 45(5):1-11.

[作者简介]



杨智凯 (1996-), 男, 山西晋城人, 广州大学博士生, 主要研究方向为网络安全、联邦学习。



刘亚萍 (1973-), 女, 湖南常德人, 广州大学教授、博士生导师, 主要研究方向为网络安全、数据中心网络、高性能网络、联邦学习。



张硕 (1984-), 男, 湖北枣阳人, 广州大学副教授、硕士生导师, 主要研究方向为网络安全、数据中心网络、云计算、分布式系统。



孙哲 (1987-), 男, 安徽安庆人, 广州大学副教授、硕士生导师, 主要研究方向为数据安全、隐私保护、人工智能安全。



严定宇 (1997-), 男, 湖南常德人, 北京邮电大学博士生, 主要研究方向为网络安全、RDMA网络、数据中心网络。