Journal of Computer Applications
计算机应用, 2022, 42(2): 333 - 342
ISSN 1001-9081
CODEN JYIIDU
2022-02-10
http://www.joca.cn

文章编号:1001-9081(2022)02-0333-10
DOI:10.11772/j.issn.1001-9081.2021020232

联邦学习通信开销研究综述

邱鑫源[1,2],叶泽聪[1,2],崔翛龙[2,3*],高志强[2]

(1.武警工程大学 研究生大队,西安 710086; 2.武警工程大学 反恐指挥信息工程研究团队,西安 710086;
3.武警工程大学乌鲁木齐校区,乌鲁木齐 830049)

(*通信作者电子邮箱787942392@qq.com)

摘 要:为了解决数据共享需求与隐私保护要求之间不可调和的矛盾,联邦学习应运而生。联邦学习作为一种分布式机器学习,其中的参与方与中央服务器之间需要不断交换大量模型参数,而这造成了较大通信开销;同时,联邦学习越来越多地部署在通信带宽有限、电量有限的移动设备上,而有限的网络带宽和激增的客户端数量会使通信瓶颈加剧。针对联邦学习的通信瓶颈问题,首先分析联邦学习的基本工作流程;然后从方法论的角度出发,详细介绍基于降低模型更新频率、模型压缩、客户端选择的三类主流方法和模型划分等特殊方法,并对具体优化方案进行深入的对比分析;最后,对联邦学习通信开销技术研究的发展趋势进行了总结和展望。

关键词:联邦学习;通信开销;模型压缩;并行计算;客户端选择策略

中图分类号:TP181; TP309 **文献标志码**:A

Survey of communication overhead of federated learning

QIU Xinyuan[1,2], YE Zecong[1,2], CUI Xiaolong[2,3*], GAO Zhiqiang[2]

(1. *Postgraduate Brigade, Engineering University of PAP, Xi'an Shaanxi* 710086, *China*;
2. *Anti-Terrorism Command Information Engineering Research Team, Engineering University of PAP, Xi'an Shaanxi* 710086, *China*;
3. *Urumqi Campus of Engineering University of PAP, Urumqi Xinjiang* 830049, *China*)

Abstract: To solve the irreconcilable contradiction between data sharing demands and requirements of privacy protection, federated learning was proposed. As a distributed machine learning, federated learning has a large number of model parameters needed to be exchanged between the participants and the central server, resulting in higher communication overhead. At the same time, federated learning is increasingly deployed on mobile devices with limited communication bandwidth and limited power, and the limited network bandwidth and the sharply raising client amount will make the communication bottleneck worse. For the communication bottleneck problem of federated learning, the basic workflow of federated learning was analyzed at first, and then from the perspective of methodology, three mainstream types of methods based on frequency reduction of model updating, model compression and client selection respectively as well as special methods such as model partition were introduced, and a deep comparative analysis of specific optimization schemes was carried out. Finally, the development trends of federated learning communication overhead technology research were summarized and prospected.

Key words: federated learning; communication overhead; model compression; parallel computing; client selection strategy

0 引言

众所周知,机器学习的性能依赖于大量可用的训练数据:数据越丰富,机器学习所得模型的性能往往会越好。然而人们越来越重视数据隐私安全,法规制定者和监管机构也出台了很多规范数据管理和使用的法律。面对数据共享需求与隐私保护要求之间不可调和的矛盾,联邦学习这一解决方案应运而生[1-3]。

联邦学习是一种借助多个参与方的本地数据,联合训练一个全局模型的分布式机器学习架构。具体地,每个参与方的数据存储在本地,在中央服务器的协调下,多个参与方联合完成机器学习任务(如图1),其工作流程描述如下。

收稿日期:2021-02-09;**修回日期**:2021-04-13;**录用日期**:2021-04-20。

基金项目:国家自然科学基金资助项目(U1603261);武警工程大学基础研究基金资助项目(WJY202124)。

作者简介:邱鑫源(1999—),女,江西南昌人,硕士研究生,主要研究方向:联邦学习、深度学习; 叶泽聪(1997—),男,广东东莞人,硕士研究生,主要研究方向:模型压缩、目标检测; 崔翛龙(1973—),男,安徽长丰人,教授,博士,主要研究方向:指挥信息系统、大数据分析; 高志强(1989—),男,河北故城人,讲师,博士,主要研究方向:联邦学习、边缘智能。

1)参与方选择:中央服务器从满足条件的参与方集中选择合适的参与方;

2)初始化:被选择的参与方从中央服务器下载初始模型的参数;

3)本地训练:每一个被选择的参与方利用自己的本地数据训练初始化模型,把更新的参数传给中央服务器;

4)聚合:中央服务器收集各个参与方更新的参数;

5)模型更新:中央服务器根据聚合结果更新全局模型的参数,并下发至参与方。

重复步骤3)~5),直到全局模型满足既定的要求,即达到预设的性能指标或达到预设的时间。

图2体现了联邦学习各节点可采用的降低联邦学习通信开销的几类方法。

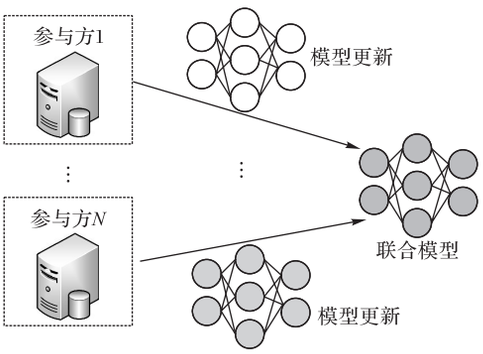


图1 联邦学习架构

Fig. 1 Architecture of federated learning

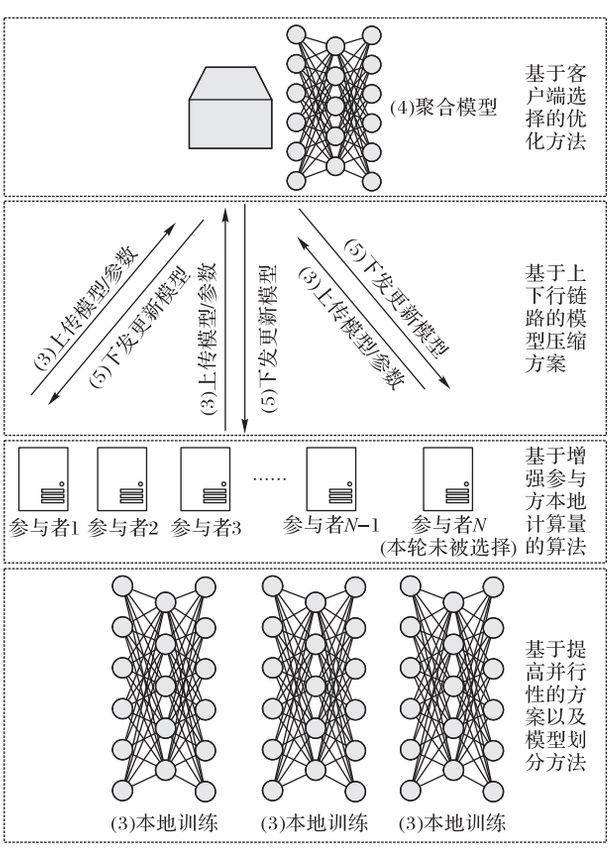


图2 联邦学习工作流程的第3)~5)步

Fig. 2 Steps 3)-5) of federated learning workflow

从工作流程可以看出,参与方与中央服务器需要不断交换大量模型参数,通信时间、通信次数、传送数据的总比特数较高,造成了较高的通信开销;其次,联邦学习越来越多地部署在通信带宽有限、电量有限的移动设备上,加之有限的网络带宽和大量的客户端增加了连接受限的客户端掉队的概率,延长了通信时间。因此,通信开销和通信效率成为了联邦学习的关键瓶颈之一,研究如何降低联邦学习的通信开销变得十分必要。然而目前国内学者主要对其隐私保护[4-6]、应用场景[7-9]等进行了综述研究,却鲜见与联邦学习通信开销方面有关的研究[10]。研究如何降低联邦学习通信开销,对联邦学习的落地应用,尤其是在电量有限、通信带宽受限的移动设备上的应用[11],具有很重要的现实意义。

通信开销一般包含两层含义:一是通信数据总量;二是通信总时耗。由于联邦学习是一个不断交互更新的通信过程,本文研究的通信开销特指联邦学习达到预设性能指标(如特定精度值)所需传输的数据总量和通信轮次。因此,降低通信开销通常可以从减少通信总次数、降低通信频率以及减少单轮通信回合的通信总比特数入手。减少通信总次数主要依靠降低模型更新频率和选择更少的客户端进行通信;而减少单轮通信回合数据量则主要依靠进行适当的模型压缩,以降低通信占用的带宽。

McMahan等[12]提出的联邦平均算法将每个客户端上的局部随机梯度下降(Stochastic Gradient Descent, SGD)与执行模型平均的服务器相结合,是一种通过增加本地计算能力降低通信频率的典型算法。模型压缩,比如模型稀疏化、量化、联邦蒸馏、低秩与子抽样等技术,则是采用减少单轮通信回合的数据量的思路,可以大幅压缩要传输的局部模型,从而节省通信开销。

本文的主要工作如下:

1)对近几年降低联邦学习通信开销的基本方法,进行整理,如图3~4所示:图3将各类方法进行了归纳分类,图4对目前一些主流算法按照发表时间进行了罗列,体现了降低通信开销方法的研究进展。

2)明确了几类主流方法作用原理,并对比其作用节点(如图2所示),详细介绍、分析了如图3所示的几种典型算法。

3)由于目前还没有标准化、统一化、权威性的指标来衡量联邦学习的通信开销,本文从优化角度、应用场景角度出发,对文献中的几种典型算法进行了对比分析。

4)对联邦学习通信开销技术研究的发展趋势进行了总结和展望。

1 基于降低模型更新频率的优化方法

起初,在联邦学习工作流程的本地训练中,客户端都是在本地运行SGD等算法后生成本地模型。而联邦学习随机梯度下降(Federated SGD, FedSGD)算法是每一轮通信都在随机选择的客户端上进行单个批次梯度计算,这种方法计算高效,但需要再将梯度计算结果传给中央服务器,通信代价较高。针对这一问题,降低通信代价的一种行之有效的办法就是降低通信频率,即降低模型更新频率。部分学者通过牺牲计算代价换取通信开销,即增加参与方的计算量或提高并行性以减少训练模型所需的通信次数:

1)增加参与方的计算量:每个参与方在每个通信回合之

间执行更复杂的计算。具体地,每个参与方执行随机梯度下降的多次迭代以计算权重更新,而不是在每次迭代后进行权重更新进行通信。

2)提高并行性:引入更多的参与方在每个通信回合之间独立工作,使得计算更快,减少通信时间,不过更多参与方可能导致相对更多通信开销。

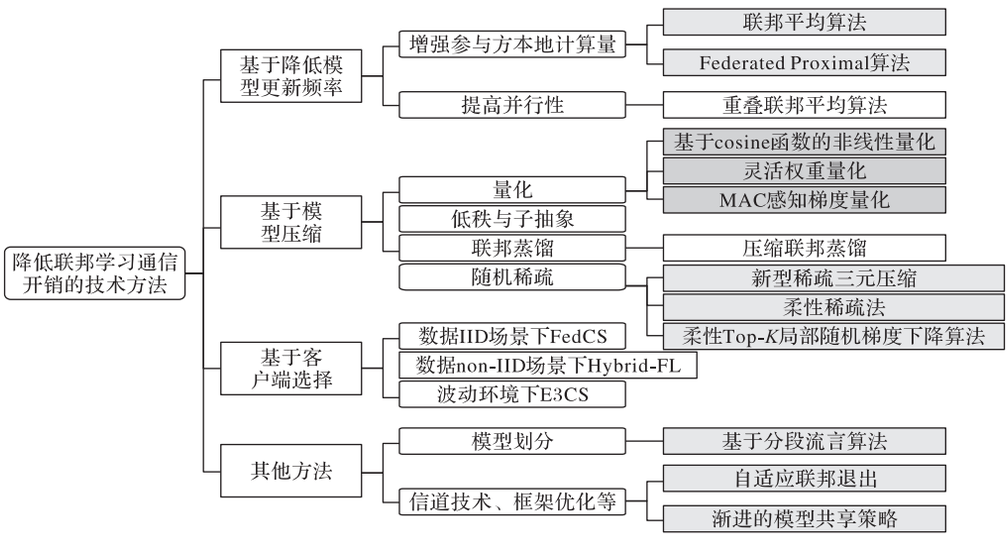


图3　降低通信开销的典型算法

Fig. 3　Typical algorithms of reducing communication overhead

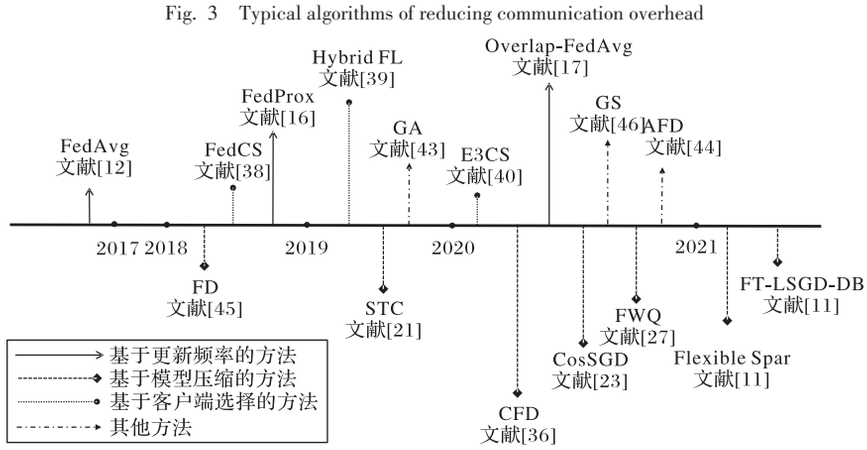


图4　降低通信开销的研究方法时间轴

Fig. 4　Methods of reducing communication overhead in chronological order

1.1　增强参与方计算量

为了克服FedSGD通信代价较高的缺陷,很多学者提出一些牺牲本地计算代价换取通信代价的算法,其中包括联邦平均(Federated Averaging, FedAvg)算法等[12-15],其在CIFAR-10测试集上性能对比如表1所示。CIFAR-10测试集是用于识别普适物体的小型数据集,一共包含10个类别的尺寸为32×32的RGB彩色图片,数据集中一共有50 000张训练图片和10 000张测试图片。CIFAR-10测试集获取地址:https://github.com/tensorflow/models。

表1　CIFAR-10测试集上同一目标精度下不同算法的通信轮次

Tab. 1　Communication rounds of different algorithms with same target accuracy on CIFAR-10 test set

目标精度/%	不同算法的通信轮次		
	SGD	FedSGD	FedAvg
80	18 000	3 750	280
82	31 000	6 600	630
85	99 000	—	2 000

McMahan等[12]提出的联邦平均算法将局部随机梯度下降与执行模型平均的服务器相结合,通过客户端先多次迭代本地更新再将本地迭代结果发送给服务器。

实验结果如表2~3和图5所示。表2~表3中E表示每个客户端在每一通信轮次上对数据集进行本地训练的次数;B表示用于客户端更新所需的本地最小批次量的大小;u表示每个用户每轮预计更新的数量,$u=\left(\frac{\mathrm{E}[n_k]}{B}\right)E=\frac{nE}{KB}$($n_k$为客户端$k$拥有的数据样本数,$K$为客户端集合中客户端总数量,$\mathrm{E}[n_k]$为$n_k$的期望值,$n$为客户端集合中样本总数)。在FedSGD中,$E=1$,$B=\infty$。表2的MNIST测试集是手写数字数据集,来自美国国家标准与技术研究所,由250个志愿者手写数字构成。该数据集图像是固定大小(28×28像素),包含60 000个用于训练的图片和10 000个用于测试的图片。MNIST测试集获取地址:http://yann.lecun.com/exdb/mnist/。表3的SHAKESPEARE测试集是语言模型测试集,采集了莎士比亚戏剧作品中各角色的台词,常用于字符预测,以莎士比亚作品集前80%行(3 564 579个字符)作为训练集,后20%

行(870 014 个字符)作为测试集。SHAKESPEARE 测试集获取地址:https://www. gutenberg. org/ebooks/100。由表 2~3 和图 5 可以看出,不论是在卷积神经网络(Convolutional Neural Network, CNN),还是在长短期记忆(Long Short Term Memory, LSTM)网络上,为了达到相同目标精度,该方法所需通信轮次明显少于随机梯度下降,但 FedAvg 仅在数据独立同分布(Independently Identically Distribution, IID)时,优化效果明显,数据非独立同分布(non-Independent Identically Distribution, non-IID)时性能较差。

表 2 MNIST 测试集上 99% 目标精度下 FedSGD 与 FedAvg 所需通信轮次[12]

Tab. 2 FedSGD and FedAvg communication rounds under 99% target accuracy on MNIST test set[12]

模型	E	B	u	通信轮次	
				IID	non-IID
FedSGD	1	∞	1. 0	626	483
FedAvg	5	∞	5. 0	179	1 000
	1	50	12. 0	65	600
	20	∞	20. 0	234	672
	1	10	60. 0	34	350
	5	50	60. 0	29	334
	20	50	240. 0	32	426
	5	10	300. 0	20	229
	20	10	1 200. 0	18	173

表 3 SHAKESPEARE 测试集上 54% 目标精度下 FedSGD 与 FedAvg 所需通信轮次[12]

Tab. 3 FedSGD and FedAvg communication rounds under 54% target accuracy on SHAKESPEARE test set[12]

模型	E	B	u	通信轮次	
				IID	non-IID
FedSGD	1	∞	1. 0	2 488	3 906
FedAvg	1	50	1. 5	1 635	549
	5	∞	5. 0	613	597
	1	10	7. 4	460	164
	5	50	7. 4	401	152
	5	10	37. 1	192	41

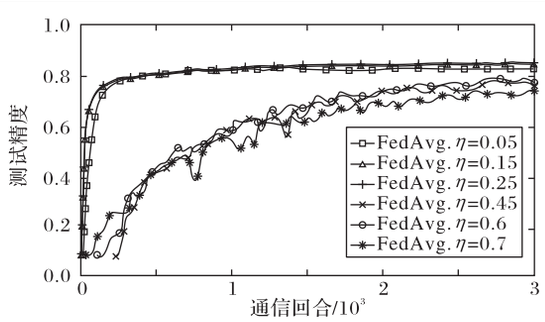


图 5 CIFAR-10 测试集上 FedSGD 与 FedAvg 测试精度对比[12]

Fig. 5 Test accuracy comparison of FedSGD and FedAvg on CIFAR-10 test set[12]

Alistarh 等[13-15]在 McMahan 的基础上优化了 FedAvg 算法,增加每一轮迭代在每个客户端本地更新参数的计算次数,将该方法与 FedSGD 算法进行对比。通过 MNIST 卷积神经网络测试,结果表明:当数据 IID 时,该算法可以明显降低通信成本;但当数据 non-IID 时,算法依旧只能轻微地减少通信开销。显然,联邦学习的数据基本都呈 non-IID,因此 FedAvg 算法及 Alistarh 提出的优化算法虽然较 FedSGD 算法通信成本更低,但其实应用场景有限,需要进一步探究针对 non-IID 数据的优化算法。

因此,Li 等[16]提出了更通用的 FedProx(Federated Proximal)算法,在每一轮中,只对一部分客户端进行采样以执行更新,这种算法在数据为 non-IID 时优化效果更明显。FedProx 算法可以动态地更新不同客户端每一轮需要本地计算的次数,不需要参与方在每次更新时统一运算次数,因此该算法更适用于非独立同分布的联合建模场景。

1.2 提高并行性

并行计算分为同步并行和异步并行,引入更多参与方后,可以显著减少整个联邦学习系统的通信时间和单个参与方的通信量。但同步并行计算中存在显著的"短板效应":当某个参与方出错需要重新计算时,该节点计算所需时间比其他所有节点都多很多,但此时其他节点依然需要一直等待该节点完成计算方可进行下一步,这样空转时间长,工作效率低。

为了解决这种"短板效应",Shi 等在文献[17]中提出了一种设备调度方案,以平衡训练轮与每轮之间的关系。Zhou 等[18]则从算法框架的角度出发,将通信与训练并行,基于集合分层计算策略、数据补偿机制和 NAG(Nesterov Accelerated Gradient)算法,提出重叠联邦平均(Overlap FedAvg, Overlap-FedAvg)算法,该算法可与许多其他压缩方法正交,以最大限度利用集群,以 FedAvg 算法为基线对比组,在数据 non-IID 场景下分别使用 MLP 等模型在不同数据集上训练,Overlap-FedAvg 算法单次迭代训练需要的时间都短于普通 FedAvg 算法,实验结果如表 4 所示。

表 4 Overlap-FedAvg 与 FedAvg 平均每次迭代耗时对比[18]

Tab. 4 Comparison of average wall-clock time of Overlap-FedAvg and FedAvg for one iteration[18]

模型	数据集	参数量	单次迭代耗时/s	
			FedAvg	Overlap-FedAvg
MLP	MNIST	199 210	31. 20	28. 85
MnistNet	FMNIST	1 199 882	32. 96	28. 31
MnistNet	EMNIST	1 199 882	47. 19	42. 15
CNNCifar	CIFAR-10	878 538	48. 07	45. 33
VGG^R	CIFAR-10	2 440 394	64. 40	49. 33
$ResNet^R$	CIFAR-10	11 169 162	156. 88	115. 31
$ResNet^R$	CIFAR-100	11 169 162	156. 02	115. 30
Transformer	WIKITEXT-2	13 828 478	133. 19	87. 90

从表 4 可看出,该重叠 FedAvg 框架具有并行性,能够在保持与 FedAvg 几乎相同的最终精度的前提下,大大加快联邦学习过程,非常适用于模型相对较大且客户端的网络连接缓慢或不稳定的场景,对不平衡和 non-IID 数据分布具有鲁棒性,可以减少在分散数据上训练深度网络所需的通信轮次。表 4 中 MLP 为多层感知机(MultiLayer Perception),也称作人工神经网络(文献[12]用 MLP 和 CNNCifar 验证了

FedAvg 的有效性,文献[18]则对普通 FedAvg 和 Overlap-FedAvg进行性能对比)。

2　基于模型压缩的优化方法

模型压缩也称为稀疏化,更新的模型结构用更少的变量刻画,压缩方案可以是随机稀疏模式、概率量化、梯度量化、子抽样、低秩等方法的一种或多种组合。如图2所示,压缩方案可以在联邦学习的不同阶段执行:参与方训练本地模型之前(下行链路),即中央服务器压缩全局模型的规模后广播给各参与方;参与方上传更新模型之前(上行链路),各参与方压缩本地训练模型参数的规模后上传给中央服务器。

Konečný 等[19]为了减少上行链路的通信消耗,考虑通过结合低秩、稀疏化、随机分散和概率量化,设计结构化更新和压缩更新的方法。结构化更新即直接在受限空间学习更新,使用较少数量的变量进行参数化;压缩更新即学习完整的更新模型后,进行压缩再发送给服务器。在卷积网络和递归网络上实验结果表明:该算法与传统 FedAvg算法相比,可实现通信回合次数减少两个数量级,不过其收敛速度略有下降。Dinh 等[20]的实验结果表明,所有参与者的梯度稀疏程度共同影响了全局收敛性和通信复杂性。下面给出随机稀疏、量化、知识蒸馏等基本策略。

2.1　随机稀疏

随机稀疏是根据预先设定的随机稀疏模式,由稀疏矩阵刻画本地更新的模型 $\boldsymbol{H}$,该模式在每一轮中为每个客户端独立重新生成矩阵。

Shi 等[11]将训练算法与本地计算、梯度稀疏相结合,提出更灵活的柔性稀疏法(Flexible Sparsification, Flexible Spar):对参与方施加误差补偿,本地计算允许在每两个全局模型更新之间对5G 移动设备执行更多的本地计算,从而减少通信回合的总次数;梯度稀疏允许参与者只上传一小部分具有显著特性的梯度,从而减少每一轮的通信有效载荷。在5G 移动设备上进行实验,结果如图6~7所示,表明该方法能耗更低,适用于异质移动设备,与统一稀疏化(Unified Sparsification, Unified Spar)在收敛速度和最终精度方面表现出非常相似的性能特征,但二者的最终精度都略低于FedAvg算法,这也反映了模型压缩的缺点:在降低通信开销的前提下,不可避免地牺牲部分精度,造成最终模型性能下降。

Sattler 等[21]基于非独立同分布、不平衡和小规模 batch 的本地数据,提出一种新型稀疏三元压缩(Spatio-Temporal Context, STC)框架,其中STC通过稀疏化、三元化、错误累积和最佳 Golomb 编码扩展当前的 top-K 梯度稀疏化的上行和下行压缩方法,在减少每一通信轮次传输数据量的同时还可以降低通信频率。然后,Li 等[22]运用了与文献[11]和文献[21]类似的思想,集成局部计算和梯度稀疏,提出了具有动态批处理大小 FT-LSGD-DB(Flexible Top-K Local Stochastic Gradient Descent with Dynamic Batch size)的柔性 Top-K 局部随机梯度下降算法,通过允许参与方执行不同"K"值的梯度稀疏化,实现了灵活压缩。与文献[11]较为相近,文献[22]在进行性能评估时同样以 FedAvg作为基准,并加入了贪婪压缩法(Greedy Sparsification, Greedy Spar)作对比,实验结果如图8所示:图8(a)~(b)表示在 CIFAR-10 数据集上使用 ResNet20 模型进行训练时,随着参与方数量增大、参与方异构性水平更高时,FT-LSGD-DB 算法相较其他算法节省的能耗更多;图8(c)~(d)为在 MNIST 数据集上使用 LeNet5-Caffe 模型进行训练,体现了 FT-LSGD-DB 算法在节省通信消耗方面的优势,该方法在适应异质移动边缘设备和提高联邦学习边缘的能量效率方面具有很大潜力。

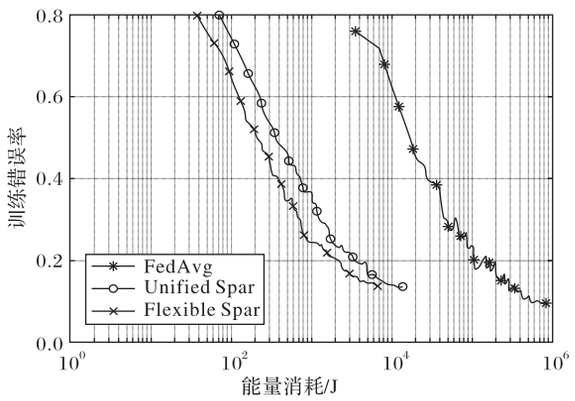


图6　同一目标精度下 Flexible Spar、Unified Spar 和 FedAvg能耗对比[11]

Fig. 6　Energy consumption comparison of Flexible Spar, Unified Spar and FedAvg under same target accuracy[11]

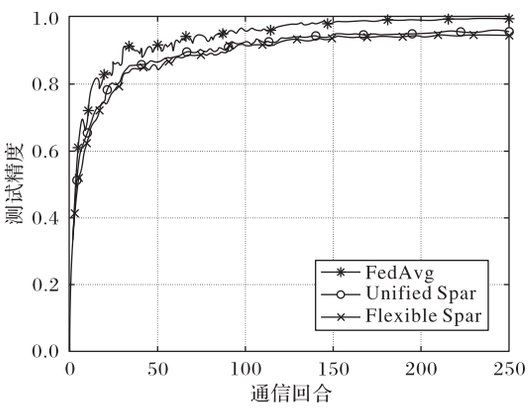


图7　同一目标精度下 Flexible Spar、Unified Spar 和 FedAvg所需通信次数对比[11]

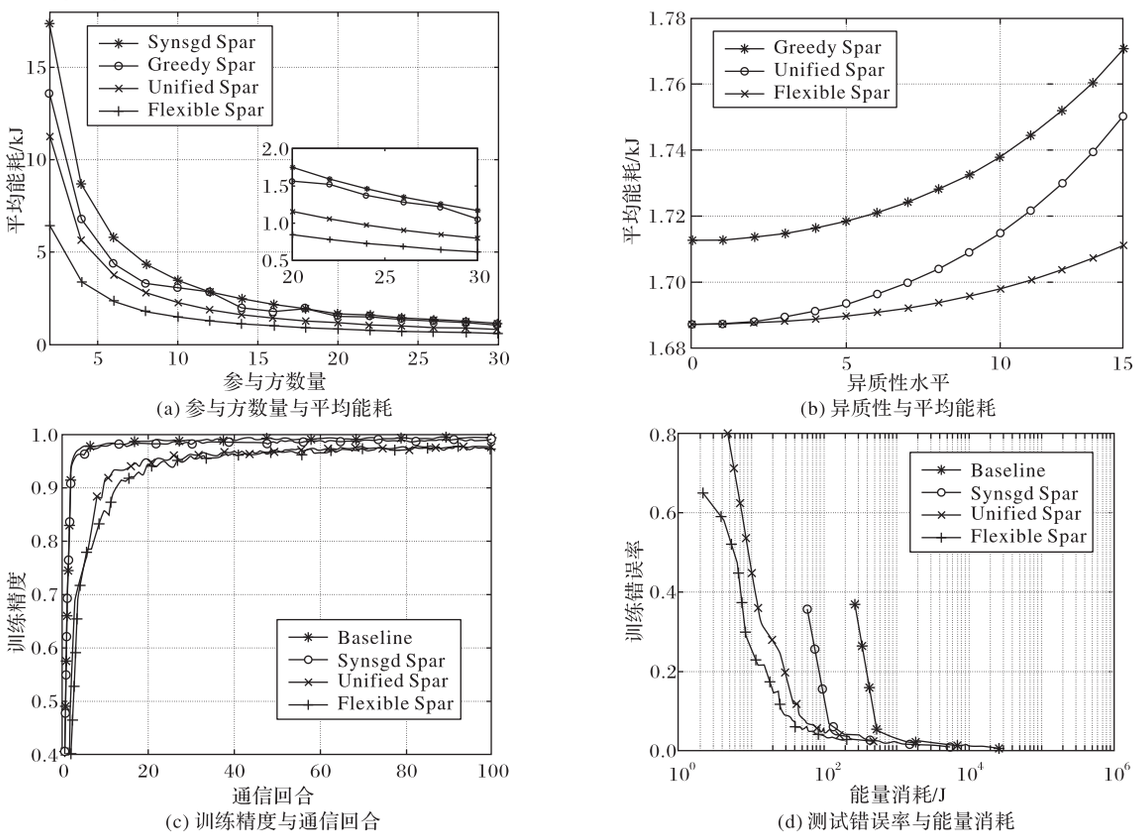
Fig. 7　Communication times comparison of Flexible Spar, Unified Spar and FedAvg under same target accuracy[11]

2.2　量化

量化最初用于数据压缩,对需要数百万参数的深度学习至关重要,能够显著降低通信成本,但依旧有损模型性能。量化一般分为概率量化与梯度量化。前者是本地更新模型向量化后,对其权重量化;后者是将梯度量化成低精度值以降低通信带宽,应用更为广泛。通过量化本地计算梯度,将梯度量化为低精度值而非直接上传原始梯度值,能降低每回合通信代价、通信比特数,但这样会降低精度,反而增加总体计算能耗。

最开始提出的量化方案是线性的,但最基本的线性量化方法,性能往往表现得不够好。因此,Ye 等[23]以非线性的方

式划分空间，提出了一种基于cosine函数的非线性量化方案cosSGD（cosine SGD），不需要误差反馈等额外梯度恢复信息[24]来调整梯度，与之前的线性量化、文献[24-26]中的低比特压缩方案相比，能够在更新客户端梯度时将数据量压缩至原来的0.1%，极大地节省了通信开销。此外，Chen等[27]将能量最小化问题描述为混合整数非线性规划问题，融合无线传输和权重量化，以最小化全局模型的损失函数为目标，应用广义弯曲分解（Generalized Benders' Decomposition，GBD）算法，提出不同5G移动设备的带宽分配和灵活权重量化（Flexible Weight Quantification，FWQ）的压缩策略。在CIFAR-100、CIFAR-10测试集上实验，结果如图9所示，得出FWQ与随机量化（Rand Quantification，RandQ）、全精度（Full Precision）、统一量化（Unified Quantification，UnifiedQ）策略相比，实现了保证精度的前提下，总体计算和通信能耗最小化。同样地，Chang等[28]结合多个接入信道（Multiple Access Channel，MAC）技术，提出了MAC感知梯度量化方案：根据各用户梯度信息性和底层信道条件，基于MAC的容量区域优化进行参数优化，这种信道感知量化与均匀量化相比，能够更加充分利用信道，但未来需要与随机稀疏等策略[29-31]相结合，降低其通信开销，进一步提升性能。



(a) 参与方数量与平均能耗　(b) 异质性与平均能耗

(c) 训练精度与通信回合　(d) 测试错误率与能量消耗

图8　Flexible Spar等算法能耗、精度、通信次数对比[22]

Fig. 8　Comparison of Flexible Spar and other algorithms on energy consumption, precision and communication times[22]

2.3　联邦蒸馏

2015年，Hinton等[32]提出知识蒸馏法（Knowledge Distillation，KD）：先利用大规模数据训练得到一个教师网络，将教师网络的知识迁移到学生网络上，使得学生网络的性能表现和教师网络相似；并以手写数字识别和语音识别为例，验证了知识蒸馏方法的有效性及模型的泛化能力。而后，Jeong等[33]提出了联邦蒸馏（Federated Distillation），其基础是只交换局部模型输出而非交换传统联邦学习采用的模型参数，这些输出的尺寸通常比模型尺寸小得多，因此可以减少通信消耗。联邦蒸馏与联邦平均有着完全不同的通信轮廓，更适用于异构客户端，颇具新颖性，但其基本原理较为复杂，只有少数著作[32,34-35]试图分析其收敛性。联邦蒸馏的工作流程如下：

1）在本地训练期间，每个工作节点存储每个标签的平均***logit***向量。

2）每个工作节点定期将其本地平均***logit***向量上传到参数服务器，并对接收到的其他工作节点的本地平均***logit***向量进行平均。

3）每个工作人员从服务器下载构建所有标签的全局平均***logit***向量。

4）在基于知识蒸馏的本地训练中，每个工作节点选择其教师网络的***logit***作为全局平均***logit***，标记为与当前训练样本的基本事实（ground-truth）相同的标签。

Sattler等[36]利用知识蒸馏的协同蒸馏（Cooperated Distillation，CD）的关键原理，提出压缩联邦蒸馏方法（Compressed Federated Distillation，CFD），可以将实现固定性能目标所需的累积通信量从8 570 MB减少到0.81 MB，相当于通信量减少至原来的0.009%。目前，联邦蒸馏可以大幅减少通信代价，适用于缺少标签的异质数据、异构模型的场景，但囿于方法要求较为苛刻（如当两个网络模型大小相差

太大时,知识蒸馏会失效)以及交换输出还可能增加用户隐私泄露的风险,联邦蒸馏的收敛性和应用性研究需要进一步研究。

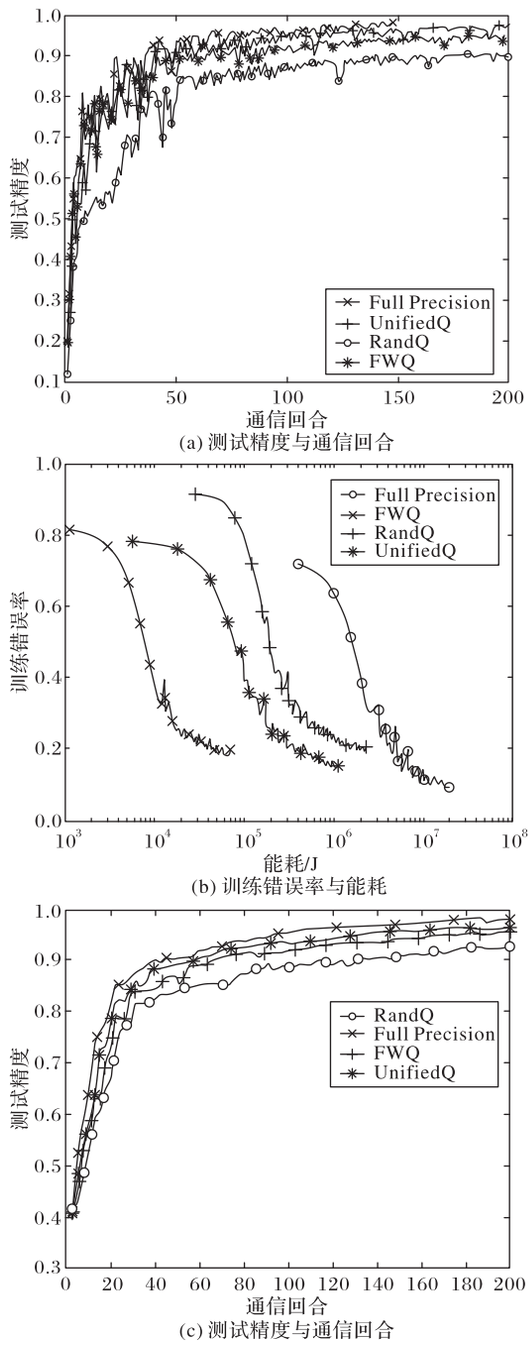


图9　CIFAR-100以及CIFAR-10测试集上FWQ等算法的精度、能耗对比[27]

Fig. 9　Comparison of accuracy and energy overhead of FWQ and other algorithms on CIFAR-100 and CIFAR-10 test sets[27]

2.4　低秩与子抽样

目前主流的压缩方法是随机稀疏和量化,子抽样和低秩等方法研究尚少。其中,子抽样的方法是本地更新模型由其随机子矩阵刻画;低秩是本地更新模型 $\boldsymbol{H} \in \mathbf{R}^{d_1 \times d_2}$ 由秩至多是 k 的矩阵刻画,其中 k 小于本地更新模型的秩,与3.1节的随机稀疏方法相类似,低秩中每一通信轮次均为每个客户端独立生成刻画矩阵。该方法在文献[16]中也有所应用。Yang等[37]基于MAC的自然信号叠加,针对模型聚合问题,提出了一种稀疏和低秩建模方法。

3　基于客户端选择的优化方法

在联邦学习中,客户端的数量可能非常大,但由于模型分发和重新上传的带宽相当有限,一般只选取一部分参与方参与训练过程。因此客户选择策略对于联邦学习过程计算效率、通信效率、最终模型的质量以及公平性等至关重要。客户端选择算法需要根据数据集是否IID、是否有用户退出等实际情况选择最优方案。

3.1　数据IID场景下的FedCS算法

Nishio 等[38]提出了一种 FedCS (Federated Client Selection)算法,根据累计有效参与值(Cumulative Effective Participation, CEP)选择模型迭代效率最高的客户端进行聚合更新,以此提高整个联邦学习算法的收敛效率,进而降低通信代价;但该算法只有在基础的动态神经网络等典型网络性能较好或数据IID时,精度与通信开销性能较好,对于拓扑结构或参数较为复杂的情况,该方法客户端选择公平性和客户端聚合效率会更低,反而会造成通信次数增加。

3.2　数据non-IID场景下的Hybrid-FL算法

针对FedCS算法只能在数据IID时同时保证高精度和降低通信开销,但数据non-IID时降低通信代价却无法保证高精度的问题,Yoshida 等[39]在启发式算法(heuristic algorithms)的基础上提出了一种Hybrid-FL(Hybrid Federated Learning)的协议,该协议可以处理数据non-IID的客户端数据,解决在non-IID数据上FedAvg、FedCS算法精度、准确度等性能不高的问题,文献[39]在数据non-IID场景下仿真边缘计算环境,在CIFAR-10和Fashion MNIST数据集上通过执行分类任务进行性能测试,结果表明non-IID数据场景下,为了达到较高准确率时,该方法所需通信代价小于FedAvg、FedCS算法,但Hybrid-FL协议一定程度上增加了通信损耗:服务器需要通过额外的资源请求选择部分客户端,从而在本地建立一种近似独立同分布的数据集用于联邦学习的训练和迭代。因此,下一步可以研究如何综合运用Hybrid-FL和FedCS方法,平衡精度与通信代价之间的关系,在保证高精度的同时使通信开销足够低。

3.3　波动训练环境下E3CS算法

在真实的联邦学习中,被选中的客户往往有机会退出,不会返回经过训练的模型,也不会通知服务器他们的退出,这种情况将会形成一种波动的训练环境。Huang等[40]针对更接近现实的波动的训练环境和数据的non-IID分布,研究了客户端选择问题,在文献[38]的基础上,提出了FedCS的改进方法E3CS(Exp3-based Client Selection),这一研究扩展了Exponential-weight算法的应用领域。对该算法进行性能评估时,以随机选择客户端和FedCS为基准组,对EMNIST和CIFAR-10数据集未带标签的图片进行分类,实验结果表明虽然该方法的CEP低于FedCS,但为了达到相同最终精度,所需通信次数更少。而后,Wu等[31]在研究波动环境下的客

户选择问题时，为了提高训练收敛速度和最终模型精度，也运用了E3CS随机选择算法，并进一步设计了“公平配额”设置，该方法在减少通信时间的同时能够保证最终模型精度的损失很小。

4 模型划分等其他方法

此外，还有模型划分的分割方法等，在2.2节中Chang等研究之后，Xia等[41]也同样基于MAC信道考虑联邦学习，提出了一种联邦分割算法：边缘服务器通过空中计算[42]聚合由多个终端设备传输的本地模型，该算法采用基于阈值的设备选择方案实现可靠的本地模型上传，鲁棒性更强，可实现快速收敛、通信回合更少，不过该算法只在目标函数具有强凸和光滑的假设下线性收敛到最优解。Hu等[43]设计了一种基于分段流言算法（Gossip Algorithm，GA）的分布式联邦学习，将模型进行划分，划分后各部分包含相同数量的彼此不重叠的模型参数，各个参与方通过将本地细分与来自其他参与方的相应细分进行汇总，来执行细分级别更新，该方法可通过以点对点（Point to Point，P2P）方式传输划分的模型来充分利用节点到节点之间的带宽，通过形成动态同步流言组实现了良好的训练收敛性。Bouacida等[44]将自适应联邦退出（Adaptive Federated Dropout，AFD）和联邦退出（Federated Dropout，FD）[45]与深度梯度压缩（Deep Gradient Compression，DGC）[16]相结合，允许客户端在本地训练全局模型的特定子集，以减少下载和上传，进而降低服务器-客户端通信代价。在SHAKESPEARE测试集上训练时，该方案收敛时间仅为文献[12]中FedAvg算法的原来的1.8%，另外，由于某些子模型往往比其他子模型更具代表性，AFD能够构建最适合每个客户数据的子模型，与不涉及压缩的场景相比，精度提高了0.9个百分点到1.7个百分点。该实验结果表明有选择地删除模型的部分子集可以在保证全局模型的质量的前提下，显著减少需要与服务器交换的权值数量，降低通信开销。

此外，还有一些从策略、框架设计角度出发的解决方案：Li等[46]引入了一种渐进的模型共享（Gradually Sharing，GS）策略和双头设计（Double Head，DH），在TTC（TCP Traffic Classification）上的实验如表5所示。当逐步共享频率设置为80轮时，该方法可以比标准FedAvg与HDAFL（Heterogeneity Dynamic Adopted Federated Learning）分别节省60%和56%的通信量。表5中：IID指各客户端的数据集分布是相同的，即每个客户端都拥有一个与其他客户端的样本数量相同的数据集，且单个客户端无法覆盖整个标签；non-IID指数据集在客户端上的分布是不同的，但是每个客户端的数据可以覆盖整个标签；dispatch指不同的客户端拥有不同类的数据，即分布不同。Tran等[47]考虑到参与方在自身数据规模、信道增益、计算和通信能力方面的差异性，提出无线网络下联合学习问题的解决方法：使用Pareto效率模型探究学习时间与参与方能耗之间的平衡，通过找寻最优准确率参数来探究计算与通信时间的平衡。

表5 TTC数据集上DH+GS等算法的模型精度比较[46] 单位：%

Tab. 5 Comparison of model precision of DH+GS and other algorithms on TTC dataset[46] unit: %

种类	IID	non-IID	dispatch
FedAvg	83.75	83.41	79.94
HDAFL	83.70	80.21	44.20
DH	85.44	85.12	74.95
DH+GS	84.30	84.44	85.62

5 结语

研究如何降低联邦学习通信开销，对联邦学习的落地应用，尤其是在电源有限的移动设备上的应用，具有很重要的现实意义。本文首先针对联邦学习的工作流程和发展现状，重点关注了联邦学习框架中的通信开销研究进展。目前，大多数文献都从压缩的角度出发解决通信开销问题，如随机稀疏化、量化、联邦蒸馏等，这些方法的思路都是通过减少上行、下行传递的数据量来减轻通信开销，而降低通信频率则一般是通过增加计算开销来降低通信开销，优化通信开销时最好综合考量性能，不能一味增加计算开销换取更低的通信开销。因此，降低通信频率的另一种方法是考虑使用并行计算，但是这种方法会引入更多参与方，虽然可以减少通信时间，但是一定程度上会导致更多参与方与中央服务器之间进行通信，从而增加通信成本；此外，同步并行中的“短板效应”也在一定程度上降低了其通信效率，对参与方稳定性有较高要求。值得注意的是，目前一部分自适应的灵活压缩方案以及基于客户端选择和模型划分等方案，对参与方要求相对要更低，可以针对实时情况动态更新改变通信策略，十分具有创新性，拓宽了研究思路，但是使用这些方法要注意将质量损耗控制在可接受范围内。

尽管联邦学习作为一种新兴技术，有很多自身优势，应用场景越来越普遍，如与区块链等新兴技术领域，但仍然存在一些值得改进的地方：

1）面向5G移动设备场景下的研究。目前，随着高通信速率的5G技术的发展，越来越多联邦学习的应用场景扩展部署到了5G移动边缘等设备上[11,20,48]，这类设备不仅通信带宽有限，且电源有限，希望系统能耗尽可能小，因此对移动终端的联邦学习通信代价技术的研究需要进一步深化到综合考量总能耗的研究。

2）non-IID数据和异构终端场景下的研究。联邦学习中参与方的数据通常以非独立同分布、非对齐、多噪声等形式存在，同时存在跨模式（如跨视频与文本数据的联邦学习）、跨语言等带来的数据异质问题，然而，目前很多研究方法还是仅在数据IID等理想状态下性能较好、行之有效，下一步需要重点研究如何在数据异质、系统异构、波动环境的真实状态下以及保证准确性和公平性前提下，提升模型性能，降低联邦学习的通信开销。

3）通信开销与计算开销的综合优化。在实际中，应用联邦学习的系统需要通盘考虑整体性能，一味牺牲计算代价或模型精度以获取低通信开销是不可取的，目前一些自适应的

灵活压缩方案等为研究提供了思路,下一步需要针对各工作节点的特点,继续深化综合考量系统的整体性能、优化通信机制。

参考文献 (References)

[1] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): No. 12.

[2] KAIROUZ P, MCMAHAN H B, AVENT B, et al. Advances and open problems in federated learning [EB/OL]. (2021-03-09) [2021-03-26]. https://arxiv.org/pdf/1912.04977.pdf.

[3] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions [J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[4] 王健宗,孔令炜,黄章成,等. 联邦学习隐私保护研究进展[J]. 大数据, 2021, 7(3): 130-149. (WANG J Z, KONG L W, HUANG Z C, et al. Research advances on privacy protection of federated learning[J]. Big Data Research, 2021, 7(3): 130-149.)

[5] 陈兵,成翔,张佳乐,等. 联邦学习安全与隐私保护综述[J]. 南京航空航天大学学报, 2020, 52(5): 675-684. (CHEN B, CHENG X, ZHANG J L, et al. Survey of security and privacy in federated learning [J]. Journal of Nanjing University of Aeronautics and Astronautics, 2020, 52(5): 675-684.)

[6] 周俊,方国英,吴楠. 联邦学习安全与隐私保护研究综述[J]. 西华大学学报(自然科学版), 2020, 39(4): 9-17. (ZHOU J, FANG G Y, WU N. Survey on security and privacy-preserving in federated learning [J]. Journal of Xinhua University (Natural Science Edition), 2020, 39(4): 9-17.)

[7] LI L, FAN Y X, TSE M, et al. A review of applications in federated learning [J]. Computers and Industrial Engineering, 2020, 149: No. 106854.

[8] 刘耕,赵立君,陈庆勇,等. 联邦学习在5G云边协同场景中的原理和应用综述[J]. 通讯世界, 2020, 27(7): 50-52. (LIU G, ZHAO L J, CHEN Q Y, et al. Summary of principles and applications of federated learning in 5G cloud-edge collaboration scenarios[J]. Telecom World, 2020, 27(7): 50-52.)

[9] 王亚珅. 面向数据共享交换的联邦学习技术发展综述[J]. 无人系统技术, 2019, 2(6): 58-62. (WANG Y S. A survey on federated learning for data sharing and exchange [J]. Unmanned Systems Technology, 2019, 2(6): 58-62.)

[10] 王健宗,孔令炜,黄章成,等. 联邦学习算法综述[J]. 大数据, 2020, 6(6): 64-82. (WANG J Z, KONG L W, HUANG Z C, et al. Research review of federated learning algorithms [J]. Big Data Research, 2020, 6(6): 64-82.)

[11] SHI D, LI L, CHEN R, et al. Towards energy efficient federated learning over 5G+ mobile devices[EB/OL]. (2021-01-13)[2021-01-26]. https://arxiv.org/pdf/2101.04866.pdf.

[12] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]// Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. New York: JMLR.org, 2017: 1273-1282.

[13] ALISTARH D, GRUBIC D, LI J Z, et al. QSGD: communication-efficient SGD via gradient quantization and encoding [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 1707-1718.

[14] KONEČNÝ J. Stochastic, distributed and federated optimization for machine learning [D/OL]. (2017-07-04) [2021-01-26]. https://arxiv.org/pdf/1707.01155.pdf.

[15] KONEČNÝ J, MCMAHAN H B, YU F X, et al. Federated learning: strategies for improving communication efficiency [EB/OL]. (2017-10-30) [2021-01-26]. https://arxiv.org/pdf/1610.05492.pdf.

[16] LI T, SAHU A K, ZAHEER M, et al. Federated optimization for heterogeneous networks [EB/OL]. [2021-01-26]. https://arxiv.org/pdf/1812.06127.pdf.

[17] SHI W Q, ZHOU S, NIU Z S, et al. Joint device scheduling and resource allocation for latency constrained wireless federated learning [J]. IEEE Transactions on Wireless Communications, 2021, 20(1): 453-467.

[18] ZHOU Y H, YE Q, LV J C. Communication-efficient federated learning with compensated Overlap-FedAvg [J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(1): 192-205.

[19] KONEČNÝ J, MCMAHAN H B, RAMAGE D, et al. Federated optimization: distributed machine learning for on-device intelligence [EB/OL]. (2016-10-08) [2021-01-26]. https://arxiv.org/pdf/1610.02527.pdf.

[20] DINH C T, TRAN N H, NGUYEN M N H, et al. Federated learning over wireless networks: convergence analysis and resource allocation [J]. IEEE/ACM Transactions on Networking, 2021, 29(1): 398-409.

[21] SATTLER F, WIEDEMANN S, MÜLLER K R, et al. Robust and communication-efficient federated learning from non-IID data [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3400-3413.

[22] LI L, SHI D, HOU R H, et al. To talk or to work: flexible communication compression for energy efficient federated learning over heterogeneous mobile edge devices [EB/OL]. (2020-12-22) [2021-01-26]. https://arxiv.org/pdf/2012.11804.pdf.

[23] HE Y, ZENK M, FRITZ M. CosSGD: nonlinear quantization for communication-efficient federated learning [EB/OL]. (2020-12-15) [2021-01-26]. https://arxiv.org/pdf/2012.08241.pdf.

[24] KARIMIREDDY S P, REBJOCK Q, STICH S, et al. Error feedback fixes signSGD and other gradient compression schemes [C]// Proceedings of the 36th International Conference on Machine Learning. New York: JMLR.org, 2019: 3252-3261.

[25] LIN Y J, HAN S, MAO H Z, et al. Deep gradient compression: reducing the communication bandwidth for distributed training [EB/OL]. (2020-06-23) [2021-01-26]. https://arxiv.org/pdf/1712.01887.pdf.

[26] BERNSTEIN J, WANG Y X, AZIZZADENESHELI K, et al. signSGD: compressed optimisation for non-convex problems [C]// Proceedings of the 35th International Conference on Machine Learning. New York: JMLR.org, 2018: 560-569.

[27] CHEN R, LI L, XUE K P, et al. To talk or to work: energy efficient federated learning over mobile devices via the weight quantization and 5G transmission co-design[EB/OL]. (2020-12-21)[2021-01-26]. https://arxiv.org/pdf/2012.11070.pdf.

[28] CHANG W T, TANDON R. Communication efficient federated learning over multiple access channels[EB/OL]. (2020-01-23)[2021-01-26]. https://arxiv.org/pdf/2001.08737.pdf.

[29] AJI A F, HEAFIELD K. Sparse communication for distributed gradient descent [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2017: 440-445.

[30] WANGNI J Q, WANG J L, LIU J, et al. Gradient sparsification for communication-efficient distributed optimization [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2018: 1306-1316.

[31] WU H D, WANG P. Fast-convergent federated learning with adaptive weighting [J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(4): 1078-1088.

[32] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2021-01-26]. https://arxiv.org/pdf/1503.02531.pdf.

[33] JEONG E, OH S, KIM H, et al. Communication-efficient on-device machine learning: federated distillation and augmentation under non-IID private data[EB/OL]. (2018-11-28)[2021-01-26]. https://arxiv.org/pdf/1811.11479.pdf.

[34] RAHBAR A, PANAHI A, BHATTACHARYYA C, et al. On the unreasonable effectiveness of knowledge distillation: analysis in the kernel regime — long version[EB/OL]. (2020-09-25)[2021-01-26]. https://arxiv.org/pdf/2003.13438.pdf.

[35] PHUONG M, LAMPERT C. Towards understanding knowledge distillation [C]// Proceedings of the 36th International Conference on Machine Learning. New York: JMLR.org, 2019: 5142-5151.

[36] SATTLER F, MARBAN A, RISCHKE R, et al. Communication-efficient federated distillation[EB/OL]. (2020-12-01)[2021-01-26]. https://arxiv.org/pdf/2012.00632.pdf.

[37] YANG K, JIANG T, SHI Y M, et al. Federated learning based on over-the-air computation [C]// Proceedings of the 2019 IEEE International Conference on Communications. Piscataway: IEEE, 2019: 1-6.

[38] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge [C]// Proceedings of the 2019 IEEE International Conference on Communications. Piscataway: IEEE, 2019: 1-7.

[39] YOSHIDA N, NISHIO T, MORIKURA M, et al. Hybrid-FL for wireless networks: cooperative learning mechanism using non-IID data [C]// Proceedings of the 2020 IEEE International Conference on Communications. Piscataway: IEEE, 2020: 1-7.

[40] HUANG T S, LIN W W, LI K Q, et al. Stochastic client selection for federated learning with volatile clients[EB/OL]. (2020-11-17)[2021-01-26]. https://arxiv.org/pdf/2011.08756.pdf.

[41] XIA S H, ZHU J Y, YANG Y H, et al. Fast convergence algorithm for analog federated learning [C]// Proceedings of the 2021 IEEE Conference on Computer Communications. Piscataway: IEEE, 2021: 1-6.

[42] TANG S H, ZHANG C, OBANA S. Multi-slot over-the-air computation in fading channels[EB/OL]. (2020-10-23)[2021-01-26]. https://arxiv.org/pdf/2010.13559.pdf.

[43] HU C, JIANG J, WANG Z. Decentralized federated learning: a segmented gossip approach[EB/OL]. [2021-01-26]. https://arxiv.org/pdf/1908.07782.pdf.

[44] BOUACIDA N, HOU J H, ZANG H, et al. Adaptive federated dropout: improving communication efficiency and generalization for federated learning [C]// Proceedings of the 2021 IEEE Conference on Computer Communications Workshops. Piscataway: IEEE, 2021: 1-6.

[45] CALDAS S, KONEČNÝ J, MCMAHAN H B, et al. Expanding the reach of federated learning by reducing client resource requirements[EB/OL]. (2019-01-08)[2021-01-26]. https://arxiv.org/pdf/1812.07210.pdf.

[46] LI D W, CHANG Q L, PANG L X, et al. More industry-friendly: federated learning with high efficient design[EB/OL]. (2020-12-16)[2021-01-26]. https://arxiv.org/pdf/2012.08809.pdf.

[47] TRAN N H, BAO W, ZOMAYA A, et al. Federated learning over wireless networks: optimization model design and analysis [C]// Proceedings of the 2019 IEEE Conference on Computer Communications. Piscataway: IEEE, 2019: 1387-1395.

[48] AMIRI M M, GÜNDÜZ D. Federated learning over wireless fading channels[J]. IEEE Transactions on Wireless Communications, 2020, 19(5): 3546-3557.

This work is partially supported by National Natural Science Foundation of China (U1603261), Fundamental Research Funds for Engineering University of PAP (WJY202124).

QIU Xinyuan, born in 1999, M. S. candidate. Her research interests include federated learning, deep learning.

YE Zecong, born in 1997, M. S. candidate. His research interests include model compressing, object detection.

CUI Xiaolong, born in 1973, Ph. D., professor. His research interests include command information system, big data analysis.

GAO Zhiqiang, born in 1989, Ph. D., lecturer. His research interests include federated learning, edge intelligence.