面向智能通信和计算的移动边缘分布式学习：现状、挑战与方法*

谢雨良，田雨晴，张朝阳

（浙江大学，浙江 杭州 310013）

【摘　要】　分布式机器学习被视为发展下一代智能通信网的基石。然而，在无线网络上部署分布式学习面临若干挑战，包括不确定的无线环境、有限的通信资源等。如何在无线边缘网络上高效地部署分布式学习成为研究热点。将调研论述无线人工智能在分布式架构中的国内外研究现状及挑战，重点介绍几种新兴的分布式学习范式，包括联邦学习、分布式推理和多智能体强化学习。最后从全局的角度，描述了移动分布式学习减少通信代价的研究现状和未来发展。

【关键词】　分布式机器学习；无线边缘计算；联邦学习；分布式推理；多智能体强化学习；通信效率

doi:10.3969/j.issn.1006-1010.20230424-0001　　中图分类号：TN929.5
文献标志码：A　　文章编号：1006-1010(2023)06-0048-08
OSID：扫描二维码与作者交流
引用格式：谢雨良,田雨晴,张朝阳. 面向智能通信和计算的移动边缘分布式学习：现状、挑战与方法[J]. 移动通信, 2023,47(6): 48-55.
XIE Yuliang, TIAN Yuqing, ZHANG Zhaoyang. Mobile Edge Distributed Learning for Intelligent Communication and Computing: Methods, Challenges and Opportunities[J]. Mobile Communications, 2023,47(6): 48-55.

Mobile Edge Distributed Learning for Intelligent Communication and Computing: Methods, Challenges and Opportunities

XIE Yuliang, TIAN Yuqing, ZHANG Zhaoyang

(Zhejiang University, Hangzhou 310013, China)

[Abstract]　Distributed machine learning is envisioned as the cornerstone of developing next-generation intelligent networks. However, deploying distributed learning over wireless networks faces several challenges including the uncertain wireless environment and limited wireless resources, and how to efficiently deploy distributed learning on wireless edge networks has become a research hotspot. This paper provides an overview of state-of-art and challenges of wireless artificial intelligence in distributed architecture and focuses on several emerging distributed learning paradigms, including federated learning, distributed inference, and multi-agent reinforcement learning. Finally, the research status and future development of mobile distributed learning to reduce communication costs are described from a global perspective.

[Keywords]　distributed machine learning; wireless edge networks; federated learning; distributed inference; multi-agent reinforcement learning; communication efficiency

0　引言

（1）移动边缘分布式机器学习的发展动机

下一代无线通信网络中，边缘设备收集海量的异构数据，使得基于大数据的人工智能（AI, Artificial Intelligence）推理和决策成为可能。然而，由于资源限制、延迟限制和隐私问题，边缘设备无法将其收集的数据全部卸载到云服务器，传统的依托于数据中心进行集中数据分析与处理的模式逐渐无法满足业务需求。与此同时，移动设备的发展使得网络的边缘节点具有较强的计算能力与储存能力，可以胜任针对本地的小规模数据量的分析与计算。基于此提出的分布式学习技术，使边缘设备能够在无需原始数据交换的情况下协同训练机器学习模型，从而减少通信开销和延迟，并改善数据隐私。

传统的机器学习一般是基于中央控制器集中进行数据处理，所有的原始数据由边缘节点上传到中央控制器，由控制器进行模型训练，再把模型参数下发给终端节点。

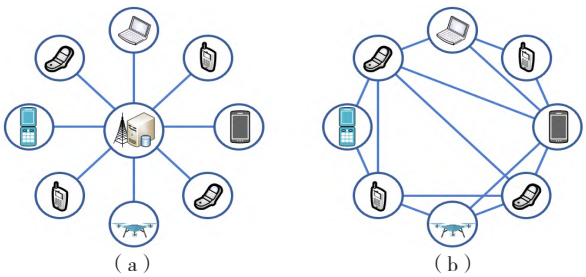
收稿日期：2023-04-24
*基金项目：国家重点研发项目“无线智能传输系统与技术”（2020YFB1807100）；大规模移动边缘网络智能协同感知-接入-处理理论与方法联合基金（U20A20158）

这样的模型有许多弊端，比如通信时延、用户数据隐私泄露、计算速度慢等问题。近几年，随着智能服务需要处理的数据量井喷式爆发，去中心化的分布式计算模型已经成为发展的必然趋势[1-3]。这种去中心化的分布式计算模型，把大量的数据分散到边缘节点分散处理，这极大程度上提高了模型计算速度。基于这种范式的分布式计算模型，得以拥有类人智能的实时反应能力[4-9]。

分布式机器学习也存在很多问题和挑战。首先，分布式学习中如何保证用户数据隐私是一个难题，不交换原始数据与训练出精确模型参数是相矛盾的，这就需要在模型准确性和隐私之间进行折衷。第二，由于移动分布式机器学习是依托于无线通信环境进行训练的，无线通信网络中传输性能会直接影响训练效率，比如干扰、噪声和信道衰落等因素。文献[10]中指出，通信时延和比特错误会很大程度影响分布式计算的收敛速度和模型准确度。第三，分布式计算中需要多次交换庞大的模型参数，这会给无线通信网带来很大的负担，在保证模型准确性的同时如何减少通信代价是需要解决的问题。最后，分布式计算需要合适的分布式优化方案，把复杂的总体优化问题分拆解决[11]。常见的分布式优化算法包括交替方向乘子法[12-13]和分布式随机梯度下降法[14]。

（2）分布式机器学习概览

分布式机器学习的模型架构一般可以根据是否存在中央控制器被分为两类。第一种架构如图1（a）中所示，它由一个中央控制器和若干个边缘节点构成。常见的例子有联邦学习（FL, Federated Learning），中央控制器可以和所有边缘节点进行通信[15-16]。第二种架构是完全去中心化的，它不包含任何中央控制器，完全由边缘节点构成，相邻的两个节点之间可以相互通信，如图1（b）所示。



（a）　　（b）

图1　分布式机器学习的模型架构

（3）文章框架

本文对移动边缘分布式学习模型的挑战、现状和发展方向进行系统的介绍，其中会重点讨论联邦学习，分布式推理和多智能体协同强化学习这几个当下重点关注的技术。最后从全局的角度，从减少通信代价的角度讨论当下的研究现状和未来发展方向，包括减少通信次数、信息的压缩和量化、通信资源分配和博弈论。

1　联邦学习

这部分首先对联邦学习的基本概念和过程进行介绍，然后讨论基于无线通信网络环境的联邦学习算法性能的四个参数以及无线通信环境对它们的影响。最后对联邦学习目前存在的挑战和研究方向进行阐述，并介绍空中计算联邦学习和联邦蒸馏两种联邦学习的优化模型。

1.1　联邦学习基本概念

联邦学习是谷歌提出的一种分布式机器学习模型[17]。它包括一个中央控制器和若干个边缘节点，中央控制器可以与每个边缘节点进行通信。中央控制器的功能是聚合每个边缘节点的模型参数，整个过程中不交换原始数据，边缘节点只把模型参数传给中央控制器。

联邦学习整个计算过程包含两部分，分别是本地训练和全局模型融合。下面将以联邦学习中最基本的平均算法（FedAvg, Federated Averaging）为例描述整个过程[18]。

在本地模型训练阶段，中央控制器首先发起任务，选择 k 个边缘节点执行任务，同时屏蔽其他节点。中央控制器首先初始化模型参数 w_0 并传输给终端节点开始训练并更新模型参数 $w_t \leftarrow w_0$。边缘节点一般采用随机梯度下降法，利用本地的数据集进行模型参数训练。把第 k 个设备训练得到的参数记作 $w_t^k \leftarrow w_t - \alpha \nabla F(w_t;k)$，其中 α 是表示学习率，$\nabla F(w_t;k)$ 表示第 k 个设备损失函数的梯度。最后，每个设备把更新后的模型参数 w_t^k 传送给中央控制器。

在全局模型融合阶段，中央控制器把各个设备传来的参数进行融合 $w_G = \frac{1}{k}\sum_{k=1}^{k} w_t^k$，得到新的全局模型参数，并传送给边缘节点作为下一轮的参数初始值。整个过程重复若干遍，直到模型收敛，达到预期的精度。

这种基本的联邦学习算法最终目标是从不同设备中得到一个统一模型，当遇到每个边缘设备上的数据非独立同分布时，重点设备的训练模型偏差会很大。由此提出了两种个性化联邦学习，分别是多任务联邦学习[19]和基于模型的非相关元学习的联邦学习算法[20]。

1.2　联邦学习的性能参数

这里将介绍四种衡量基于无线通信网络的联邦学习性能指标：训练损失函数、收敛时间、能量消耗和可靠性。

（1）训练损失函数

在联邦学习过程中，训练损失函数的数值由所有边缘设备的训练模型参数共同决定。在无线环境中，模型参数通过无线通信网传输，因此会影响最终全局模型参数。

（2）收敛时间

无线环境下的联邦学习收敛时间可定义为下式：$T=(T_c+T_T)\times N_T$。其中 T_c 表示边缘节点进行单次本地模型训练的时间，T_T 表示每一轮学习过程中模型参数的传输时间，N_T 表示达到全局收敛需要的训练轮数。需要注意的是，T_c 和 N_T 不是相互独立的。在随机梯度下降算法中，T_c 的增加可以使得全局收敛所需要的轮数 N_T 减小。

（3）能量消耗

无线联邦学习中，每个参与计算的设备能耗可以表示为：$E=(E_c+E_T)\times N_T$。其中，E_c 是终端设备进行单次本地训练的能耗，E_T 是每一轮计算终端设备和中央控制器传输模型参数的能耗，N_T 是达到全局收敛所需要的轮数。由于增加本地训练次数可以减少全局融合总轮数，所以要在 E_c 和 N_T 之间折衷。

（4）可靠性

无线联邦学习的可靠性定义为达到训练损失函数标准的概率。在无线联邦学习模型中，由于可分配的通信资源有限，每次训练中只包含一部分边缘设备的子集可以参与训练。每一轮参与模型融合的边缘设备可能不同，提供的模型参数也就不同，这会对收敛时间和损失函数都造成影响。同时，不稳定的无线通信环境还会造成模型参数传输错误，这会降低模型的准确率。

1.3 无线通信网络参数对联邦学习性能的影响

无线通信网络参数对上一节提到的四个联邦学习性能有一定的影响，比如频谱、传输功率以及计算能力等。具体解释如下。

频谱资源分配决定了每个设备的信号噪声干扰比、传输速率和传输错误发生概率。因此频谱分配对训练损失、传输时间 T_T、传输能耗 E_T 和可靠性都有影响。

计算能力决定了每次本地训练时进行随机梯度下降所需要的次数。因此它会影响本地训练的时间和能耗。

传输功率决定了信干噪比、传输速率、传输错误概率。当每个终端设备的传输能量增加时，训练损失、传输时间 T_T、训练轮次 N_T 和可靠性都会下降，但传输能耗 E_T 会增加。

当每轮训练参与的设备数量增加时，训练轮次 N_T 和训练损失都会减少，与此同时可靠性和传输时间 T_T 会增加。

当边缘设备训练的模型参数数量增加时，联邦学习的训练损失、可靠性和训练总轮次都会减少，同时能耗和训练时间都会增加。

1.4 联邦学习的挑战和研究方向

本节首先概述无线网络环境中联邦学习的挑战和研究现状，然后介绍空中计算联邦学习和联邦蒸馏两种改进方案。

（1）联邦学习的挑战和研究现状

1）通信资源

无线网络环境中，由于可分配的传输资源有限，且联邦学习每一轮需要传输的模型参数数据量庞大，这造成了传输瓶颈，特别是对于深度学习等复杂的大规模神经网络架构。缓解这一问题的方法有压缩和量化等。文献 [21] 中提出了一种在本地训练过程中对模型参数进行压缩和稀疏化的方法，并测试了这种方法的收敛性能。文献 [22] 中，作者采用有损压缩的方法对全局模型参数在传输前进行压缩。为了进一步压缩数据量，文献 [23] 提出一种三元量化方法，运用在训练和推理阶段。文献 [24] 中，作者设计了基于随机线性编码的新型联邦学习优化方法，提高了传输能耗效率。文献 [25-28] 中，提出了更多基于概率标量量化方法的联邦学习优化法。以上的研究都是基于上行链路传输，即边缘设备把本地模型参数上传给中央控制器的过程，然而在带限网络中，中央控制器通过广播传输全局模型给边缘设备的过程也会存在传输瓶颈。针对此问题，文献 [29] 中研究了下行传输过程中收敛速度受噪音影响的程度。

2）无线资源分配

如表 1 所示，频谱、传输速率等无线资源对联邦学习的性能有直接影响。因此如何有效分配无线传输资源，高效地完成联邦学习是研究的重要方向。文献 [30-31] 中，作者研究了如何权衡本地训练次数和全局融合的关系，以最小化能耗，训练损失和传输时间。文献 [32] 中提出中央控制器只对信干噪比低于某个阈值的边缘节点传输的模型进行融合，并且评估了这种方法对收敛速度和可靠性的影响。文献 [33] 提出了一种联合优化边缘设备调度和资源分配的方法，在有限训练时间内提高模型准确性。

3）联邦学习训练方法优化

除上述两种从无线网络角度调整的方法，还可以从优化联邦学习训练方法的角度，调整算法使其更适应无线网络环境。文献 [34] 中提出基于错误反馈的随机梯度

下降法，并证明它可以优化收敛速度和泛化能力。文献[35]中提出了一种分簇的联邦学习模型，把原本的架构分为若干子集，每一个子集由一个基站负责聚合本子集的参数模型，再由基站把融合后的参数传送给中央控制器进行全局融合。这一架构在文献[36]中得到进一步发展，提出了一种多层级联邦学习的训练方法。文献[37]和[38]提出了一种只对必要边缘节点进行全局融合的筛选聚合方案，只可以缓解联邦学习的通信负担。文献[39]提出基于参数贝叶斯推断架构的联邦学习，由此减少达到收敛需要的训练次数。文献[40]提出一种基于聚类算法的无监督联邦学习算法，进一步优化算法。

（2）基于空中计算的联邦学习

前文提到，基于无线通信网络的联邦学习（FEEL, Federated Edge Learning）面临的一个主要挑战是如何突破通信资源瓶颈。研究者们尝试通过不同的方式减小通信时延，比如在训练过程中淘汰速度过慢的边缘设备[41-42]，或者只选择对全局模型融合影响显著的模型参数参与全局融合[43]。近期提出了一种把空中计算（OAC, Over the Air Computation）和FEEL相结合的方案，它可以提高联邦学习架构的可扩展性。文献[44-45]从信息论的角度对空中计算的局限性进行了论述。文献[46-48]提出了OAC和无线通信网的结合方案。文献[49-50]中提出了OAC和FEEL的结合架构设计，基于空中计算，全局融合可以实现同步多信道传输模型参数，这便解决了通信瓶颈的问题，使得FEEL能够运用在更大规模的环境下。

目前这种架构的的研究方向主要是解决数据隐私问题和实现宽带环境下的训练。尽管联邦学习已经不需要用户之间交换原始数据，但传输梯度向量依然有可能泄露隐私信息，这被称为梯度泄露[51-52]。文献[53]提出了在基于数字信号传输的联邦学习过程中，边缘设备在本地训练模型参数中适当加入造噪声后再传输给中央控制器。对于宽带传输的问题，如果每次传输本地模型参数不进行预编码，则需要信道资源和模型的维度相当，占用很大带宽，往往超出边缘节点可利用的带宽资源。要解决这个问题，可以通过对模型参数进行压缩缓解。

（3）联邦蒸馏

联邦蒸馏（FD, Federated Distillation）是把只是知识蒸馏的概念运用到联邦学习中[54]。知识蒸馏（KD, Knowledge Distillation）是近年来发展起来的一类模型压缩与加速技术，主要是利用一个已经训练好的复杂模型（作为教师），将其学习到的决策信息（知识）迁移到另一个轻量级模型（作为学生）中，帮助和指导学生模型的训练。知识蒸馏旨在通过将知识从深度网络转移到小型网络来压缩和改进模型。在文献[55-56]中，联邦蒸馏模型被运用在无线衰落信道中，测试表明联邦蒸馏比传统的联邦学习算法鲁棒性和抗干扰性更强。关于联邦蒸馏的更详细论述可参见文献[57]。

2 分布式推理

目前分布式学习领域的研究大部分都关注的是模型训练阶段，但由于终端设备计算能力有限和任务的及时性要求，推理过程也存在很多挑战。本节将介绍分布式推理的挑战和目前研究现状。

2.1 分布式推理的挑战

分布式推理是指用训练得到的模型参数来对新的数据进行推断（比如分类、回归问题）。第一个挑战是终端设备计算能力有限，特别是对于深度神经网络这样的大模型，会造成计算速度过慢。许多推理任务对实时性要求较高[58]，这使得满足时延要求的分布式推理更加困难。其二，即使不考虑计算能力和实时性要求，用作推理的数据集往往分布在不同的终端设备上，这也给分布式推理造成困难。例如文献[59]中，智能监控设备需要获取所有终端设备上的数据。

从信息论的角度，无线网络环境中的分布式推理可以等效为同信源编码的率失真问题[60-61]。目前的研究方向一方面是通过压缩和量化等方式简化模型参数，另一个思路是把大模型拆分到不用的设备上，进行联合推理。下面将介绍这两类优化分布式推理方法。

2.2 神经网络压缩和量化

为解决边缘设备计算能力有限的问题，考虑可以尽可能简化神经网络模型的规模和复杂度。最常用的方法之一是采用剪枝和量化的方式去掉对模型整体影响不大的冗余模型参数。适当的简化模型还可以避免模型过拟合的问题[62-64]。文献[65]详细介绍了当下常用的模型剪枝方法。另一种有效的方案是在训练阶段对模型加入稀疏正则条件，直接得到稀疏模型[66-67]。

模型压缩旨在尽可能减少权重参数，而量化则是减少表征每个参数的比特数。文献[68-69]提出了定点表征方式，并证明了这种方法可以保证模型的一定精确度。一些研究试图用单比特二进制来表示模型权重值，并实验得出这样的深度神经网络依然有较好的性能[70-71]。

深度神经网络的压缩还可以被看成是一个典型信源压缩问题。文献 [72] 提出了哈希编码的方法，文献 [73] 运用了向量量化法，文献 [74] 中，作者使用哈夫曼编码来进一步简化量化后的冗余模型参数。

2.3 协同边缘推理

除了上面提到的压缩和量化，还可以考虑把模型拆到给不同的设备上，让多个边缘设备协同进行推理，这可以缓解单个设备的计算负担。常见的一种方法是把深度神经网络拆分成两部分，第一部分分配给终端设备，剩余的多层计算由中央控制器完成 [75]。文献 [76] 进一步提出把深度神经网络的推理计算过程看成一个计算图模型，目标是获得联合模型分割和模型搜索的优化结果。在此基础上，文献 [77-79] 提出了把模型剪枝和协同推理结合深度神经网络模型，这可以进一步减少每个设备的计算压力。文献 [80-81] 具体讨论了在无线环境中的协同边缘推理，考虑到了无线传输中的时延、可靠性、信噪比等问题。文献 [82] 提出了一种基于 MIMO 空中计算的分割学习系统，将预编码器和组合器设计与隐式 MIMO 信道矩阵一起构成了一个神经网络中的可训练层，显著地提高了系统通信效率。

3 多智能体强化学习

之前介绍的内容均为基于无线网的监督式学习。本节将介绍将强化学习运用在无线网络的控制和优化中。

3.1 多智能体强化学习基本概念

强化学习（RL, Reinforcement Learning）通过实时与当前的无线环境学习反馈，可实现网络控制和资源分配等功能 [83]。基本的强化学习可以分成三类。第一种是单智能体强化学习，它可以描述为一个简单的马尔可夫过程。第二种是独立多智能体强化学习，它是最简单的多智能体算法（MARL, Multi-Agent Reinforcement Learning），其中每个智能体按单智能体的算法独立运行。第三种为多智能体协同强化学习，这种算法需要智能体之间相互交换反馈、状态等信息。不同的场景下主要智能体相互交换的信息会有差异。比如文献 [84] 中提到的多智能体协同算法需要智能体间相互交换状态信息和行为信息。而文献 [85] 提出的中值分解网络则通过智能体交换反馈信息。交换信息的不同会影响多智能体协同架构的复杂度。文献 [86] 中，作者对比研究了不同多智能体协同算法的模型复杂度和性能。

3.2 研究现状与挑战

文献 [87-88] 中，作者调研总结了多智能体强化学习在无线通信网控制和优化中的具体应用。文献 [89] 中，作者设计了独立多智能体架构运用在基站频谱效率优化上，并证明了协同多智能体算法效果更优。文献 [90] 提出一种新型架构，中央控制器收集多个智能体的经验信息来训练机器学习模型。文献 [91] 中把多智能体协同算法运用在调制和解调中。文献 [92] 中设计了一种多层级的多智能体联邦学习策略，运用在协同优化和调度无线网络资源中。

多智能体算法面临一些挑战，比如收敛性的严格证明以及影响收敛性的参数。另一个方面就是无线网络的性能参数对算法效果的影响。这些亟待解决的问题决定了多智能体强化学习算法是否能在无线网络的优化控制中得到更广泛的普及，也是未来的研究发展方向。

4 移动边缘分布式学习的研究方向

本节从分布式学习算法如何在无线网络中实现最优性能的角度，讨论其未来的发展方向。目前的研究已经通过减少通信次数、压缩和量化等方法解决了一部分难题，但这些方法依然有待解决难题和挖掘的潜力，下面讲具体说明。

4.1 减少分布式学习的通信次数

一种方法是在进行多次本地计算后进行一次全局融合，比如文献 [93] 中提出的改进后的分布式算法和文献 [94] 中提出的联邦学习改进算法。

另一种思路是通过事件触发机制减少通信次数，也就是在特定场景下设备之间和控制器才进行参数传输。文献 [95] 中提出的事件触发机制改进的分布式梯度下降算法就是一个典型的例子。

此方法依然有可以改进的余地。首先是这个方法研究的问题大多都是针对单一任务的分布式学习，仅有少数文献研究多任务和个性化的机器学习问题，比如文献 [96-97]。可以进一步结合元学习，根据任务和算法的需求，采用动态的方式灵活制定每两次通信之间的本地更新次数。

4.2 压缩与量化

前文中已经提到过压缩和量化的概念和一些研究现状。压缩过程难以避免地会使得每次分布式学习过程产生微小误差，每次的误差在多次训练后会累积起来，影响模型的整体性能。可以考虑每一次通信后，反馈给边缘设备误差值，让它做相应的误差补偿，由此避免错误的累积 [98]。

除此之外，还可以设计动态压缩和量化方式，根据

任务特性和当前无线通信资源状况实时调整压缩的程度，以此权衡模型精确度和通信速度。

4.3 通信资源分配

分布式计算过程中需要消耗大量的通信资源，比如带宽和能量，这些资源往往是有限的。虽然目前有很多工作已经提出了很多优化资源分配的方案，但依然有很多可以继续探索的空间。

其中一个重要问题是数据隐私和传输效率的权衡问题。分布式计算虽然不传输原始数据，但模型参数依然可能泄露重要的信息。传输过程中的加入噪音可以一定程度上避免窃听问题。由信噪比的定义可知，当分配到功率较大时，信噪比则偏大，容易泄露隐私，反之则信噪比小，用户信息不易被解析出来。这和传输效率是相反的，故如何同时保证通信效率和隐私保护是需要探索的方向。

4.4 博弈论与分布式计算

近来，有研究提出把博弈论的机制运用在分布式学习算法资源调配上，并鼓励各边缘设备间的互相协作配合。

文献 [99] 中，作者提出了一种基于联邦学习的多维鼓励机制，旨在权衡模型训练损耗、通信时延和用户数据隐私等因素，达到最优平衡。文献 [100] 中加入的信誉（Reputation）的概念，用信誉度来衡量每个边缘节点的可靠性和可信度。进一步，文献 [101] 提出了基于联邦学习的竞争机制，刺激单用户尽可能节省通信资源，保证数据隐私。

目前的研究主要集中在边缘设备和中央控制器组成的架构上，对完全去中心化的分布式架构，尚未有典型的运用和研究。同样地，目前大多数工作针对的是单任务分布式学习，而对于个性化分布式学习，仍然有可以挖掘的空间。个性化学习中，单个设备往往只关注自己的模型精确度，最大化利用通信资源，而不关心邻居节点的训练模型准确度。若想要达到全局最优，可以尝试把博弈论与个性化分布式学习相结合。

5 结束语

本文对基于无线通信网络的分布式机器学习进行了全面的综述。首先介绍了分布式机器学习的基本概念和架构，之后，重点介绍了基于无线通信网的联邦学习、分布式推理和多智能体强化学习三种重要技术的研究现状和挑战。最后，基于权衡通信资源利用效率，从四个方面讨论了无线分布式学习的未来研究方向。

参考文献：

[1] Letaief K B, Chen W, Shi Y, et al. The Roadmap to 6G: AI Empowered Wireless Networks[J]. IEEE Communications Magazine: Articles, News, and Events of Interest to Communications Engineers, 2019,57(8): 84-90.

[2] Akyildiz I F, Kak A, Nie S. 6G and Beyond: The Future of Wireless Communications Systems[J]. IEEE Access, 2020,8: 133995 - 134030.

[3] Dang S, Amin O, Shihada B, et al. What should 6G be?[J]. Nature Electron., 2020,3(1): 20-29.

[4] Posner J, Tseng L, Aloqaily M, et al. Federated Learning in Vehicular Networks: Opportunities and Solutions[J]. IEEE Network, 2021,35(2): 152-159.

[5] Wang X, Han Y, Wang C, et al. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning[J]. IEEE Netw., 2019,33(5): 156-165.

[6] Niknam S, Dhillon H S, Reed J H. Federated learning for wireless communications: Motivation, opportunities, and challenges[J]. IEEE Commun. Mag., 2020,58(6): 46-51.

[7] Liu Y, Yuan X, Xiong Z, et al. Federated learning for 6G communications: Challenges, methods, and future directions[J]. China Commun., 2020,17(9): 105-118.

[8] Zhao Z, Feng C, Yang H H, et al. Federated-learningenabled intelligent fog radio access networks: Fundamental theory, key techniques, and future trends[J]. IEEE Wireless Commun., 2020,27(2): 22-28.

[9] Kang J, Xiong Z, Niyato D, et al. Reliable federated learning for mobile networks[J]. IEEE Wireless Commun., 2020,27(2): 72-80.

[10] Chen M, Yang Z, Saad W, et al. A joint learning and communications framework for federated learning over wireless networks[J]. IEEE Trans. Wireless Commun., 2021,20(1): 269-283.

[11] Bertsekas D, Tsitsiklis J. Parallel and Distributed Computation: Numerical Methods[M]. Prentice Hall, 1989.

[12] Huang Z, Hu R, Guo Y, et al. DP-ADMM: ADMM-based distributed learning with differential privacy[J]. IEEE Trans. Inf. Forensics Secur., 2019,15: 1002-1012.

[13] Kumar C, Rajawat K. Network dissensus via distributed ADMM[J]. IEEE Trans. Signal Process., 2020,68: 2287-2301.

[14] Nedic A. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization[J]. IEEE Signal Process. Mag., 2020,37(3): 92-101.

[15] Letaief K B, Shi Y, Lu J, et al. Edge artificial intelligence for 6G: Vision, enabling technologies, and applications[J]. IEEE Journal on Selected Areas in Communications, 2022,40(1): 5-36.

[16] Gafni T, Shlezinger N, Cohen K, et al. Federated learning: A signal processing perspective[J]. IEEE Signal Processing Magazine, 2022,39(3): 14-41.

[17] Mcmahan H B, Moore E, Ramage D, et al. Communication-Efficient Learning of Deep Networks from Decentralized Data[C]// Proc. 20th Int. Conf. Artif. Intell. Statist., Fort Lauderdale, Florida, USA, 2016: 1273-1282.

[18] Hong C S, Khan L U, Chen M, et al. Federated Learning for Wireless Networks[M]. Berlin, Germany: Springer, 2021.

[19] Smith V, Chiang C K, Sanjabi M, et al. Federated Multi-Task Learning:, 10.48550/arXiv.1705.10467[P]. 2017.

[20] Fallah A, Mokhtari A, Ozdaglar A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach[C]//Proc. Adv. Neural Inf. Process. Syst., 2020,33: 3557-3568.

[21] Haddadpour F, Kamani M M, Mokhtari A, et al. Federated learning with compression: Unified analysis and sharp guarantees[C]// Proc. Int. Conf. Artif. Intell. Statist., 2021,130: 2350-2358.

[22] Caldas S, J Konečny, Mcmahan H B, et al. Expanding the Reach of Federated Learning by Reducing Client Resource Requirements: 10.48550/arXiv.1812.07210[P]. 2018.

[23] Xu J, Du W, Cheng R, et al. Ternary Compression for Communication-

Efficient Federated Learning[J]. IEEE Trans.Neural Netw. Learn. Syst., early access, 2021-09.

[24] Abdi A, Saidutta Y M, Fekri F. Analog compression and communication for federated learning over wireless MAC[C]//Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun.(SPAWC). Atlanta, GA, USA: IEEE, 2020.

[25] Wen W, Xu C, Yan F, et al. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning[C]//Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017: 1-13.

[26] Alistarh D, Grubic D, Li J, et al. QSGD: Communication-efficient SGD via gradient quantization and encoding[C]//Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017: 1709-1720.

[27] Horvath S, Ho C Y, Horvath L, et al. Natural compression for distributed deep learning[J]. 2019, arXiv:1905.10988. [Online]. Available: http://arxiv.org/abs/1905.10988.

[28] Reisizadeh A, Mokhtari A, Hassani H, et al. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization[C]//Proc. Int. Conf. Artif. Intell. Statist., Palermo, Italy, 2020.

[29] Amiri M M, Gündüz D, Kulkarni S R, et al. Convergence of federated learning over a noisy downlink[J]. IEEE Trans. Wireless Commun., early access, 2021-08-17.

[30] Tran N H, Bao W, Zomaya A, et al. Federated learning over wireless networks: Optimization model design and analysis[C]// Proc. IEEE Conf. Comput. Commun. (IEEE INFOCOM). IEEE, Paris, France, 2019.

[31] Tian Y Q, Zhang Z Y, Yang Z H, et al. Hierarchical Federated Learning with Adaptive Clustering on Non-IID Data[C]//IEEE Global Communications Conference(IEEE GLOBECOM). IEEE, 2022.

[32] Yang H H, Liu Z, Quek T Q S, et al. Scheduling policies for federated learning in wireless networks[J]. IEEE Trans. Commun., 2020,68(1): 317-333.

[33] Shi W, Zhou S, Niu Z, et al. Joint device scheduling and resource allocation for latency constrained wireless federated learning[J]. IEEE Trans. Wireless Commun., 2021,20(1): 453-467.

[34] Karimireddy S P, Rebjock Q, Stich S, et al. Error feedback fixes SignSGD and other gradient compression schemes[C]//Proc. Int. Conf. Mach. Learn., Long Beach, CA, USA, 2019.

[35] Abad M S H, Ozfatura E, Gündüz D, et al. Hierarchical federated learning ACROSS heterogeneous cellular networks[C]//Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP). IEEE, Barcelona, Spain, 2020: 8866-8870.

[36] Hosseinalipour S, Azam S S, Brinton C, et al. Multi-Stage Hybrid Federated Learning over Large-Scale D2D-Enabled Fog Networks[J]. 2020, arXiv:2007.09511. [Online]. Available: http://arxiv.org/abs/2007.09511.

[37] Chen T, Giannakis G, Sun T, et al. LAG: Lazily aggregated gradient for communication-efficient distributed learning[C]//Proc. Adv. Neural Inf. Process. Syst., Montreal, QC, Canada, 2018.

[38] Sun J, Chen T, Giannakis G, et al. Communication-efficient distributed learning via lazily aggregated quantized gradients[C]//Proc. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, 2019: 1-20.

[39] Kassab R, Simeone O. Federated generalized Bayesian learning via distributed stein variational gradient descent[J]. 2020, arXiv:2009.06419. [Online]. Available: http://arxiv.org/abs/2009.06419.

[40] Dennis D K, Li T, Smith V. Heterogeneity for the win: One-shot federated clustering,[C]//Proc. Int. Conf. Mach. Learn., 2021.

[41] Chen J, Pan X, Monga R, et al. Revisiting distributed synchronous SGD[J]. 2016, arXiv:1604.00981. [Online]. Available: http://arxiv.org/abs/1604.00981.

[42] Xu J, Huang S L, Song L, et al. Gradient coding: Avoiding stragglers in distributed learning[C]/Proc. Int. Conf. Mach. Learn., Sydney, NSW, Australia, 2021.

[43] Kamp M. Efficient decentralized deep learning by dynamic model averaging[J]. 2018, arXiv:1807.03210. [Online]. Available: http://arxiv.org/abs/1807.03210.

[44] Nazer B, Gastpar M. Computation over multiple-access channels[J]. IEEE Trans. Inf. Theory, 2007,53(10): 3498-3516.

[45] Soundararajan R, Vishwanath S. Communicating linear functions of correlated Gaussian sources over a MAC[J]. IEEE Trans. Inf. Theory, 2012,58(3): 1853-1860.

[46] Goldenbaum M, Stanczak S. Robust analog function computation via wireless multiple-access channels[J]. IEEE Trans. Commun., 2012,61(9): 3863-3877.

[47] Chen L, Zhao N, Chen Y, et al. Over-the-air computation for IoT networks: Computing multiple functions with antenna arrays[J]. IEEE Internet Things J., 2018,5(6): 5296-5306.

[48] Zhu G, Xu J, Huang K, et al. Over-the-air computing for wireless data aggregation in massive IoT[J]. 2020, arXiv:2009.02181. [Online]. Available: http://arxiv.org/abs/2009.02181.

[49] Zhu G, Wang Y, Huang K. Broadband analog aggregation for low-latency federated edge learning[J]. IEEE Trans. Wireless Commun., 2020,19(1): 491-506.

[50] Amiri M M, Gündüz D. Federated learning over wireless fading channels[J]. IEEE Trans. Wireless Commun., 2020,19(5): 3546-3557.

[51] Zhu L, Liu Z, Han S. Deep leakage from gradients[C]//Proc.Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, 2019.

[52] Melis L, Song C, Cristofaro E D, et al. Exploiting unintended feature leakage in collaborative learning[C]//Proc. IEEE Symp. Secur. Privacy (SP). IEEE, San Francisco, CA, USA, 2019.

[53] Abadi M, et al. Deep learning with differential privacy[C]//Proc.Conf. Comput. Commun. Secur. (ACM SIGSAC), Vienna, Austria,Oct. 2016.

[54] Jeong E, Oh S, Kim H, et al. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data[C]//Proc. Adv. Neural Inf. Process. Syst. Workshop Mach. Learn. Phone Consum. Devices, Montreal, QC, Canada, 2018.

[55] Ahn J H, Simeone O, Kang J. Cooperative learning VIA federated distillation OVER fading channels[C]//Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP). IEEE, Barcelona, Spain, 2020.

[56] Ahn J H, Simeone O, Kang J. Wireless federated distillation for distributed edge learning with heterogeneous data[C]//Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.(PIMRC). IEEE, Istanbul, Turkey, 2019.

[57] Seo H, Park J, Oh S, et al. Federated knowledge distillation[J]. 2020, arXiv:2011.02367. [Online]. Available: http://arxiv.org/abs/2011.02367.

[58] Ramjee S, Ju S, Yang D, et al. Fast deep learning for automatic modulation classification[J]. 2019, arXiv:1901.05850. [Online]. Available: http://arxiv.org/abs/1901.05850.

[59] Jankowski M, Gündüz D, Mikolajczyk K. Wireless image retrieval at the edge[J]. IEEE J. Sel. Areas Commun., 2021,39(1): 89-100.

[60] Dobrushin R, Tsybakov B. Information transmission with additional noise[J]. IRE Trans. Inf. Theory, 1962,8(5): 293-304.

[61] Wolf J K, Ziv J. Transmission of noisy information to a noisy receiver with minimum distortion[J]. IEEE Trans. Inf. Theory, 1970.IT-16(4): 406-411.

[62] Hanson S, Pratt L. Comparing biases for minimal network construction with back-propagation[C]//Proc. Adv. Neural Inf. Process. Syst., Denver, CO, USA, 1989.

[63] LeCun Y, Denker J, Solla S. Optimal brain damage[C]//Proc. Adv. Neural Inf. Process. Syst., Denver, CO, USA, 1990.

[64] Hassibi B, Stork D G, Wolff G, et al. Optimal Brain Surgeon: Extensions and performance comparison[C]//Proc. Adv. Neural Inf. Process. Syst., San Francisco, CA, USA, 1993.

[65] Liu J, Tripathi S, Kurup U, et al. Pruning algorithms to accelerate convolutional neural networks for edge applications: A survey[J]. 2020, arXiv:2005.04275. [Online]. Available: http://arxiv.org/abs/2005.04275.

[66] Lebedev V, Lempitsky V. Fast ConvNets using group-wise brain damage[C]//Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). IEEE, Las Vegas, NV, USA, 2016.

[67] Wen W, Wu C, Wang Y, et al. Learning structured sparsity in deep neural networks[C]//Proc. Adv. Neural Inf. Process. Syst., Barcelona, Spain, 2016.

[68] Gupta S, Agrawal A, Gopalakrishnan K, et al. Deep learning with limited numerical precision[C]//Proc. Int. Conf. Mach. Learn., Lille, France, 2015.

[69] Courbariaux M, Bengio Y, David J P. Training deep neural networks with low precision multiplications[C]//Proc. Int. Conf. Learn. Represent. Workshop, San Diego, CA, USA, 2015.

[70] Courbariaux M, Bengio Y, David J. P. BinaryConnect: Training deep neural networks with binary weights during propagations[C]//Proc. Adv. Neural Inf. Process. Syst., Montreal, QC, Canada, 2015.

[71] Courbariaux M, Hubara I, Soudry D, et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1[J]. 2016, arXiv:1602.02830. [Online]. Available: http://arxiv.org/abs/1602.02830.

[72] Chen W, Wilson J T, Tyree S, et al. Compressing neural networks with the hashing trick[C]//Proc. Int. Conf. Mach. Learn., Lille, France, 2015.

[73] Gong Y, Liu L, Yang M, et al. Compressing deep convolutional networks using vector quantization[J]. 2014, arXiv:1412.6115.[Online]. Available: http://arxiv.org/abs/1412.6115.

[74] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding[C]// Proc. Int. Conf. Learn. Represent., San Juan, Puerto Rico, 2016.

[75] Teerapittayanon S, McDanel B, Kung H T. Distributed deep neural networks over the cloud, the edge and end devices[C]//Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. IEEE, (ICDCS).Atlanta, GA, USA, 2017.

[76] Tian Y, Zhang Z, Yang Z, et al. JMSNAS: Joint Model Split and Neural Architecture Search for Learning over Mobile Edge Networks[J]. IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2022.

[77] Jankowski M, Gündüz D, Mikolajczyk K. Joint device-edge inference over wireless links with pruning[C]//Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC). IEEE, 2020.

[78] Shi W, Hou Y, Zhou S, et al. Improving device-edge cooperative inference of deep learning via 2-step pruning[C]//Proc. IEEE Conf. Comput. Commun. IEEE, Workshops, Paris, France, 2019.

[79] Shao J, Zhang J. Communication-computation trade-off in resource-constrained edge inference[J]. IEEE Commun. Mag., 2020,58(12): 20-26.

[80] Gamal A El, Kim Y H. Network Information Theory[M]. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[81] Van Dyck R E, Miller D J. Transport of wireless video using separate, concatenated, and joint source-channel coding[C]//IEEE Proc. IEEE, 1999,87(10): 1734-1750.

[82] Yang Y. Over-the-air split machine learning in wireless mimo networks[J]. IEEE Journal on Selected Areas in Communications, 2023: 1007-1022.

[83] Busoniu L, Babuska R, Schutter B D. A comprehensive survey of multiagent reinforcement learning[J]. IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., 2008,38(2): 156-172.

[84] Chen M, Saad W, Yin C. Virtual reality over wireless networks: Quality-of-service model and learning-based resource management[J]. IEEE Trans. Commun., 2018,66(11): 5621-5635.

[85] Hu Y, Chen M, Saad W, et al. Distributed multi-agent meta learning for trajectory design in wireless drone networks[J]. IEEE J. Sel. Areas Commun., 2021,39(10): 3177-3192.

[86] Pytorch Implementations of the Multi-Agent Reinforcement Learning Algorithms[J]. Accessed: Apr. 2021. [Online]. Available: https://github.com/starry-sky6688/StarCraft.

[87] Luong N C, Hoang D.T, Gong S, et al. Applications of deep reinforcement learning in communications and networking: A survey[J]. IEEE Commun. Surveys Tuts., 2019,21(4): 3133-3174.

[88] Feriani A, Hossain E. Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial[J]. IEEE Commun. Surveys Tuts., 2021,23(2): 1226-1252.

[89] Bennis M, Perlaza S M, Blasco P, et al. Self-organization in small cell networks: A reinforcement learning approach[J]. IEEE Trans. Wireless Commun., 2013,12(7): 3202-3212.

[90] Nasir Y S, Guo D. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks[J]. IEEE J. Sel. Areas Commun., 2019,37(10): 2239-2250.

[91] Vrieze C de, Barratt S, Tsai D, et al. Cooperative multi-agent reinforcement learning for low-level wireless communication[J]. 2018, arXiv:1801.04541. [Online]. Available: http://arxiv.org/abs/1801.04541.

[92] Hu F, Deng Y, Aghvami A H. Correlation-aware cooperative multigroup broadcast 360◦ video delivery network: A hierarchical deep reinforcement learning approach[J]. 2020, arXiv:2010.11347. [Online]. Available: http://arxiv.org/abs/2010.11347.

[93] Wang S, 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.

[94] McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. 2017: 1273-1282.

[95] Liu C, Li H, Shi Y, et al. Distributed event-triggered gradient method for constrained convex minimization[J]. IEEE Transactions on Automatic Control, 2019,65(2): 778-785.

[96] Fallah A, Mokhtari A, Ozdaglar A, Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach[J]. Proceedings of the Advances in Neural Information Processing Systems, 2020,33: 3557-3568.

[97] Ozkara K, Singh N, Data D, et al. QuPeD: Quantized personalization via distillation with applications to federated learning[J]. Advances in Neural Information Processing Systems, 2021,34.

[98] Koloskova A, Stich S, Jaggi M. Decentralized stochastic optimization and gossip algorithms with compressed communication[C]// Proceedings of the International Conference on Machine Learning. 2019: 3478-3487.

[99] Ding N, Fang Z, Huang J. Optimal contract design for efficient federated learning with multi-dimensional private information[J]. IEEE Journal on Selected Areas in Communications, 2020,39(1): 186-200.

[100] Kang J, Xiong Z, 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.

[101] Jiao Y, Wang P, Niyato D, et al. Toward an automated auction framework for wireless federated learning services market[J]. IEEE Transactions on Mobile Computing, 2021,20(10): 3034-3048. ★

作者简介

谢雨良：浙江大学信息与电子工程学院在读博士研究生，研究方向为人工智能与无线通信的融合领域、分布式算法等。

田雨晴：浙江大学信息与通信工程专业在读博士研究生，研究方向为机器学习、分布式算法和神经架构搜索。

张朝阳：博士毕业于浙江大学，现任浙江大学求是特聘教授，主要研究新一代无线通信、智能协同感知-通信-计算、无线人工智能等。承担和完成国家杰出青年科学基金等国家级项目和课题二十余项。获得ICC 2019、GlobeCom 2020等国际学术会议最佳论文奖8项，获省部级科技奖励多项和中国通信学会、中国发明协会一等奖各1项。现任国家IMT-2030（6G）推进组无线AI任务组组长。