计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(1): 179−206, 2025

面向 AIoT 的协同智能综述

罗宇哲[1,2] 李 玲[1] 侯朋朋[1] 于佳耕[1] 程丽敏[1,2] 张常有[1] 武延军[1] 赵 琛[1]

[1](中国科学院软件研究所 北京 100190)

[2](中国科学院大学 北京 100049)

(yuzhe2019@iscas.ac.cn)

Survey of AIoT-Oriented Collaborative Intelligence

Luo Yuzhe[1,2], Li Ling[1], Hou Pengpeng[1], Yu Jiageng[1], Cheng Limin[1,2], Zhang Changyou[1], Wu Yanjun[1], and Zhao Chen[1]

[1] (*Institute of Software*, *Chinese Academy of Sciences*, *Beijing* 100190)

[2] (*University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract The fusion of deep learning and the Internet of things has significantly promoted the development of the AIoT ecosystem. On the one hand, the huge amounts of multi-modal data collected by AIoT devices provide deep learning with abundant training data resources, which plays a more important role in the era of big models. On the other hand, the development of deep learning makes AIoT devices smarter, which shows great potential for promoting social development and the convenience of human life. As major support for the usage of deep learning in AIoT, federated learning effectively makes use of the training data provided by AIoT devices to train deep learning models with data privacy protection while collaborative inference overcomes the obstacles in the deployment of deep learning brought by the limited computation resource of AIoT devices. We induce the concept of AIoT-oriented collaborative intelligence. Aiming at implementing knowledge transmission and computation resource supply with high efficiency and security, we review the related works, published in the past 10 years, about the architecture, algorithm, privacy, and security of federated learning and collaborative inference, and introduce the inner connection of federated learning and collaborative inference. The algorithm part summarizes the federated learning and collaborative inference algorithm related to AIoT use cases and their optimization goals. The architecture part introduces the related works about deep learning accelerators, deep learning compilation, deep learning frameworks, communication among devices, and collaboration among devices from the view of AI computing systems. The privacy and security part introduces the privacy and security threats faced by AIoT-oriented collaborative intelligence and the defense methods against them. We also provide insights into the future development of AIoT-oriented collaborative intelligence in the aspect of equipment sharing, model sharing, collaboration of privacy and security mechanisms, and collaboration of incentive mechanisms.

Key words collaborative intelligence; federated learning; collaborative inference; artificial intelligence of things; AI computing system

摘 要 深度学习和物联网的融合发展有力地促进了 AIoT 生态的繁荣. 一方面 AIoT 设备为深度学习提供了海量数据资源, 另一方面深度学习使得 AIoT 设备更加智能化. 为保护用户数据隐私和克服单个 AIoT 设备的资源瓶颈, 联邦学习和协同推理成为了深度学习在 AIoT 应用场景中广泛应用的重要支撑. 联邦学

收稿日期: 2023-12-05; **修回日期**: 2024-08-19

基金项目: 广东省重点领域研发计划项目(2019B010154004)

This work was supported by the Key-Area Research and Development Program of Guangdong Province(2019B010154004).

通信作者: 李玲(liling@iscas.ac.cn)

习能在保护隐私的前提下有效利用用户的数据资源来训练深度学习模型,协同推理能借助多个设备的计算资源来提升推理的性能. 引入了面向 AIoT 的协同智能的基本概念,围绕实现高效、安全的知识传递与算力供给,总结了近十年来联邦学习和协同推理算法以及架构和隐私安全 3 个方面的相关技术进展,介绍了联邦学习和协同推理在 AIoT 应用场景中的内在联系. 从设备共用、模型共用、隐私安全机制协同和激励机制协同等方面展望了面向 AIoT 的协同智能的未来发展.

关键词 协同智能;联邦学习;协同推理;智能物联网;智能计算系统

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202330975 **CSTR:** 32373.14.issn1000-1239.202330975

智能物联网(artificial intelligence of things, AIoT)来源于物联网和人工智能的技术融合[1]. 随着智慧工厂[2]、智慧家居[3]、智慧农业[4]和智慧交通[5]等物联网智慧应用的迅速发展, AIoT 的设备数目和市场规模日益增长①. 与此同时,华为、小米、微软和英伟达等头部企业相继推出其 AIoT 战略或产品线[6-7]加速了 AIoT 产业的扩张,显示出 AIoT 在推动社会经济发展和人类生活方式智能化方面的巨大潜力.

AIoT 既具有传统物联网的一般特征,也具有因融合了人工智能而带来的特殊性. 传统物联网采用感知层、传输层和应用层 3 层逻辑架构[8]: 感知层由大量传感器设备构成,负责采集各种数据信息;传输层通过通信网络将感知层收集的数据信息传递给边缘服务器或云服务器;应用层则通过边缘服务器或云服务器分析和处理来自感知层的数据并为用户提供服务. 在 AIoT 中,端侧设备除了能通过传感器收集数据信息外,往往也具有本地计算能力,可以为用户提供智能计算服务[9-10]. 这些 AIoT 端侧设备运行的应用可以通过人工智能算法在端侧进行智能数据处理. 深度学习是目前最先进的人工智能技术之一[1, 11],在图像识别[12]、语音识别[13]、自然语言处理[14]和自动控制[15]等领域都取得了显著的效果. 深度学习在 AIoT 中的部署进一步提升了 AIoT 的智能化水平[16],成为未来发展的重要方向[1].

随着 AIoT 设备的普及,其收集的海量多模态数据成为深度学习等数据驱动型人工智能的宝贵资源. 由于这些数据可能包含用户隐私信息,根据欧盟的《通用数据保护条例》[17](General Data Protection Regulation, GDPR)和《中华人民共和国数据安全法》等相关法律,服务商不能像在数据中心中一样直接收集原始数据样本进行分布式训练[18],而只能在用户的数据隐私得到充分保障的前提下训练深度学习模型. 此外,在 AIoT 应用场景中端侧设备在地理上的广泛分布[19]、训练样本在不同的端侧设备上存在的统计异构[20]、高延迟与不稳定的广域网传输[21]和对激励机制的需求[22]也使得传统分布式训练方法[23]并不适用,联邦学习提供了解决这些问题的一系列方法[20, 24-25].

然而在算力层面,随着深度学习模型的复杂化, AIoT 端侧设备在计算和存储方面难以满足训练和推理任务对资源的需求[26-27],日益成为制约深度学习在 AIoT 中应用落地的限制性因素. 基础模型[28]的发展与 ChatGPT 等人工智能内容生成[29](AI-generated content, AIGC)现象级应用的出现一方面为 AIoT 开辟了更为广阔的应用前景;另一方面,基础模型也大幅提升了深度学习模型对算力的需求(例如 GPT-3 最多有 1 750 亿个参数,训练需要 3 640 PFLOPS-days 的算力),使得其训练只能在云端数据中心进行[29]. 为了弥补算力方面的不足, AIoT 端侧设备通过通信网络将部分或全部深度学习任务卸载到计算和存储资源较为丰富的边缘服务器或云服务器上执行的方法成为近年来的研究热点[16, 30]. 智能计算系统[31]为端边云各层级的深度学习算力平台提供了统一的理论分析框架,因此本文将端侧设备、边缘服务器和云服务器视为单设备智能计算系统. 通过通信网络,分布在不同层级的单设备智能计算系统可以合作完成联邦学习和协同推理等深度学习任务.

在 AIoT 应用场景中的联邦学习和协同推理都涉及多个设备协同完成深度学习任务的过程,从中可以引申出面向 AIoT 的协同智能的概念. 面向 AIoT 的协同智能可以从广义和狭义 2 个方面来解释: 从广义上来讲,面向 AIoT 的协同智能包含 AIoT 应用场景中可以使用的协同训练和协同推理 2 个相对独

① 根据 IoT-Analytics 的报告,近年来 AIoT 的设备数目和市场规模均保持年均 15% 以上的增长,详见 https://iot-analytics.com/number-connected-iot-devices/和 https://iot-analytics.com/iot-market-size/

立的技术领域；从狭义上来说，面向 AIoT 的协同智能一方面基于协同训练使用 AIoT 端侧设备收集的海量多模态数据训练深度学习模型（包括基础模型等大模型），另一方面以 AIoT 端侧设备为媒介使用协同训练得到的深度学习模型为用户提供基于协同推理的智能服务，形成了协同训练和协同推理相结合的应用新形态. 协同训练包含联邦学习、分离式学习[32]、迁移学习[33]、元学习[34]、模型补丁[35]和双向协同[36]等方式[1, 37]. 在监督学习的情况下，协同训练所需的数据标签可以来自用户反馈[37]、众包数据标注[38]和云端大模型[39]等. 由于在训练数据中包含用户隐私的情况下联邦学习是进行协同训练的主要方式[37]，而在 AIoT 应用场景中保护用户数据隐私十分重要[1, 40]，本文将联邦学习作为 AIoT 应用场景中的主要协同训练方式①.

为更好地利用 AIoT 端侧设备收集的海量多模态数据、弥补 AIoT 端侧设备在算力方面的不足和通过 AIoT 端侧设备为用户提供优质的智能服务，本文归纳了近年来 AIoT 应用场景中联邦学习和协同推理这 2 种主要的协同智能[41-42]方法的主要进展，介绍了这两者之间的内在联系并基于这种联系对它们的未来进行了展望. 已有的相关综述主要涉及 AIoT、联邦学习和协同推理等领域[1, 16, 20, 22, 25, 30, 37, 43]，如表 1 所示. 部分已有综述[16, 20, 22, 25, 30, 43]主要单独总结联邦学习或协同推理的相关工作，没有在 AIoT 应用场景中将两者联系起来. 一些综述[1, 30, 37]同时涉及了联邦学习和协同推理，但没有从算法、架构和隐私安全等角度归纳出两者在 AIoT 应用场景中的内在联系，而本文正是基于这一联系引出了面向 AIoT 的协同智能的概念并组织全文内容. 在架构方面，本文从智能计算系统的角度对面向 AIoT 的协同智能在架构方面相关工作所在的架构层级进行了分类，据我们所知目前没有本领域相关综述使用这种分类方法. 此外，本文还总结了面向 AIoT 的协同智能近年来在联邦持续学习、联邦多模态学习、联邦强化学习、协同推理的隐私安全以及大模型等新兴领域的技术进展.

Table 1 A Brief Summary of Related Surveys

表 1 相关综述简介

相关综述	AIoT	大模型	联邦学习								协同推理			
			定义	架构	异构	多模态	FCL	FRL	P&S	优化	定义	架构	P&S	优化
文献 [1]	●			◐					◐			◐		◐
文献 [16]	●										●	●		●
文献 [20]			●	●					●	●				
文献 [22]			●	●					●					
文献 [25]	●		●	●	◐				●					
文献 [43]											●	●		●
文献 [30]				●	◐				◐	●		●		●
文献 [37]	◐	◐	◐	●	◐			◐	◐	◐		◐		
本文	●	◐	●	●	●	●	●	◐	●	●	●	●	●	●

注：隐私安全（privacy and pecurity，P&S）；联邦持续学习（federated continual learning，FCL）；联邦强化学习（federated reinforcement learning，FRL）. ◐ 简略介绍；● 详细介绍.

在协同训练和协同推理的过程中，多个设备共享的资源主要有知识和算力 2 种类型，因此面向 AIoT 的协同智能的主要设计目标是通过高效、安全的知识传递和算力供给来实现智能任务的多设备协同计算. 本文总结了可用于实现该目标的相关技术的进展，主要贡献有 3 个方面：

1）基于 AIoT 应用场景中协同训练与协同推理的内在联系归纳出面向 AIoT 的协同智能的概念，提出两者相结合的应用新形态；

2）围绕知识传递和算力供给从算法、架构和隐私安全 3 个方面总结了联邦学习和协同推理近十年来的相关进展，介绍了联邦学习和协同推理的内在联系；

3）从设备共用、模型共用、隐私安全机制协同和激励机制协同 4 个方面对面向 AIoT 的协同智能未来的发展方向进行了展望.

① 当本文分别论述联邦学习和协同推理这 2 个领域中与 AIoT 应用场景相关的技术进展时，是从广义的角度来介绍面向 AIoT 的协同智能；当本文论述这 2 种技术的联系或涉及两者联合起作用的新的应用形态时，则是从狭义的角度来介绍面向 AIoT 的协同智能.

1 研究背景

本节介绍了联邦学习、协同推理和智能计算系统的相关背景知识.

1.1 联邦学习

为解决传统分布式机器学习中直接传递训练样本可能带来的隐私安全问题, Google 公司在 2017 年提出了联邦平均(FedAvg)算法[44]. FedAvg 算法被提出后, 联邦学习成为学术界的热点研究领域. 2021 年, Kairouz 等人[45] 总结了联邦学习的研究进展和开放问题并给出了联邦学习的定义, 提出联邦学习是在一个中心服务器或服务提供商的协调下多个参与者合作解决一个机器学习问题的过程, 在这一过程中, 参与者将原始数据保留在本地, 而通过传递模型参数达到学习目标. 2019 年, Yang 等人[22] 将联邦学习归纳为多个数据拥有者合作训练机器学习模型的过程. 在联邦学习过程中, 数据拥有者的数据并不暴露给彼此, 同时使用来自多个数据拥有者的数据进行训练能取得较高精度.

近年来, 联邦学习在模型训练方式上又有了新的发展, 主要体现在拓展了参与者共享的知识类型. 知识在机器学习中指的是模型的输入向量和输出向量之间的映射关系[46]. FedAvg 算法的参与者实际上是通过共享模型参数来共享知识; 除了模型参数外, 知识还可以通过模型输入输出数据、模型中间层的输出和梯度等进行传递[47-48]. 例如, 文献 [48] 基于端侧设备本地模型对公共样本的标签进行预测, 然后通过共享预测结果在参与者之间传递知识. 文献 [47] 将分离式学习的思想引入联邦学习, 通过前向推理过程中的模型中间层输出和反向传播过程中的模型中间层梯度来传递知识. 本文从知识共享的角度归纳了联邦学习的定义:

定义 1. 联邦学习是在保证隐私安全的前提下, 参与者通过共享知识合作训练机器学习模型以提升模型精度并共同获益的过程. 其中:

1)知识共享是联邦学习参与者合作训练机器学习模型的技术途径;

2)保证隐私安全是对知识共享过程的要求;

3)提升模型精度是联邦学习参与者通过知识共享要达到的目标.

1.2 协同推理

协同推理是多个设备共同完成机器学习模型推理的过程. 深度学习模型的协同推理[16] 一般包含模型切分、模型切片下发、模型切片推理、中间数据传递和推理结果回传等步骤. 在这个过程中, 深度学习模型被切分为若干模型切片, 这些切片被下发至参与协同推理的设备上, 随后发起设备将输入样本输入模型输入层所在的模型切片, 模型切片的输出作为中间数据被传递至下一级模型切片所在的设备作为继续推理的输入数据, 模型切片推理和中间数据传递在设备间依次进行直到推理过程完成, 最终推理结果被回传至推理任务的发起设备. 除了基于模型切分的协同推理外, 在设备间传递协同推理所需的输入参数也是协同推理的一种方式[36-37]. 近年来, 协同推理相关研究通过优化模型分割算法提升了推理的性能[49] 和延迟的鲁棒性[50], 降低了推理过程的能耗[21]. 协同推理的隐私安全保护也成为新兴研究的方向[51-53].

1.3 智能计算系统

智能计算系统[31] 是面向智能算法的专用计算系统, 它包含智能算法、智能编程框架、智能芯片和智能编程语言. 例如, 在深度学习盛行的当下, 智能计算系统一般包含深度学习算法(如多层感知机、卷积神经网络和循环神经网络等)、深度学习编程框架(如 PyTorch 和 TensorFlow 等)、深度学习加速器(GPU 和深度学习处理器等)和深度学习的领域专用语言(如寒武纪的 BANG 语言).

深度学习模型的训练和推理是智能计算系统的 2 个基本功能, 而面向 AIoT 的协同智能包含多种不同的单设备智能计算系统(AIoT 端侧设备、边缘服务器和云服务器等), 能通过协同训练和协同推理在 AIoT 应用场景中实现这 2 个基本功能, 从而通过多设备协同实现了深度学习在 AIoT 应用场景中的训练和推理 2 个基本功能. 由于面向 AIoT 的协同智能还涉及隐私安全等传统智能计算系统尚未涉及的领域, 它与协同训练、联邦学习、协同推理和智能计算系统 4 个领域之间的关系如图 1 所示. 其中, 本文将联邦学习作为面向 AIoT 的协同智能所采用的主要

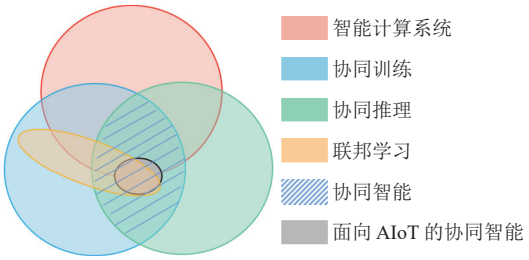


Fig. 1 The relationship among the domains

图 1 各领域之间的相互关系

协同训练方法，但 AIoT 应用场景中也可能存在分离式学习[54]等其他协同训练方法.

2 算　　法

联邦学习算法和协同推理算法是面向 AIoT 的协同智能实现知识传递和算力供给的基本方式，2 种算法通过传递中间数据（模型参数和模型切片推理结果等）来进行知识传递，通过将深度学习任务在不同设备间分配实现算力供给. 本节主要从联邦学习和协同推理 2 个方面来介绍面向 AIoT 的协同智能的算法.

2.1 联邦学习的算法

通过高效的知识提取来提升模型的精度是联邦学习的主要目标[22]. 然而，广泛分布的 AIoT 端侧设备在训练数据的统计分布和设备配置 2 个方面表现出明显的异构特征，这 2 种异构分别被称为统计异构和设备异构. 在包含统计异构和设备异构的环境中进行联邦学习会对模型训练效率和精度产生显著的影响[55]，因此有必要对异构环境中的联邦学习进行研究.AIoT 端侧设备能产生海量的多模态数据，为了更好地使用这些数据进行模型训练，联邦多模态学习成为需要进一步研究的重要领域[56]. 由于 AIoT 端侧设备不断产生新的数据，这些新的数据可能会有与以往旧的训练数据不同的分布（如新用户的加入），为了使联邦学习训练的模型既能学习新知识又不产生对旧知识的灾难性遗忘[57]，联邦持续学习[58]应运而生. 此外，智能机器人等 AIoT 端侧设备需要具备自动决策能力，联邦强化学习[59]为在保障设备数据隐私的前提下使用多个 AIoT 端侧设备的数据训练决策模型提供了可行途径. 本节总结了联邦学习在上述 4 个方面近年来的相关进展.

2.1.1 异构环境中的联邦学习

异构环境中的联邦学习主要涉及统计异构环境中的联邦学习和设备异构环境中的联邦学习. 近年来兴起的模型异构联邦学习则可以同时用于解决统计异构问题和设备异构问题.

统计异构可能造成联邦学习过程中的权重发散，即联邦学习过程中的模型参数更新方向偏离在单机上使用所有样本进行随机梯度下降（stochastic gradient descent, SGD）[60]训练时的模型参数更新方向，从而导致所训练的深度学习模型的精度下降. 解决统计异构的主要方法有改变训练样本分布[61-62]、纠正梯度方向[63-64]和个性化训练[65-66]等. Fed-CBS[61]使端侧设备被选择的概率与其本地训练数据集的类别标签平衡程度呈正相关，从而降低了联邦学习训练数据的不平衡程度，提升了全局模型的收敛速度和精度. Astraea[62]使用平衡参与者样本数量和平衡参与者样本分布 2 种方法来使训练样本分布接近测试样本分布. 在纠正梯度方向方面，FedProx[63]通过在损失函数中加入正则项来约束端侧设备更新原全局模型的范围，以减轻统计异构带来的精度下降. SCAFFOLD[64]则通过控制参量来对端侧设备的梯度方向进行纠偏，从而使得聚合后的全局模型在更新方向上与单机训练时的更新方向基本一致. 然而，改变训练样本分布和纠正梯度方向 2 种方法都在训练过程中损失了部分端侧设备本地训练数据分布的个性化信息[62, 66]，可能造成全局模型在个性化数据分布上的精度下降，个性化训练为解决这一问题提供了方案. FEDPER[65]通过将端侧设备本地模型分为基础层和个性化层来达到在统计异构环境中训练个性化端侧设备本地模型的目的. 通过在损失函数中加入正则项，Ditto[66]使得端侧设备本地模型既能在训练过程中从全局模型里提取任务的通用知识又能从端侧设备本地数据中提取个性化的知识，从而能在统计异构环境中取得公平性和模型健壮性之间的良好平衡.

设备异构可能导致 AIoT 应用场景中不同的参与设备进行本地模型训练的速度不同，从而导致不同设备的模型难以进行同步更新. 解决设备异构的主要方法有权重选择[67]和异步联邦学习[68-70]等. 为解决设备异构造成的同步难题，FedAsync[67]在服务器的聚合过程中根据端侧设备所提交的模型参数更新的陈旧程度来选择其聚合时所占权重，从而在设备间难以进行模型同步时减少了陈旧的模型参数更新对训练过程的影响. 在异步联邦学习方面，APFL[70]通过交换参数聚合和全局模型下发的顺序实现了异步联邦学习的流水线并行，使得训练过程中端侧设备和中心服务器的空闲时间大幅度降低. PermiDAG[68]在车联网中使用异步联邦学习来减少因车辆终端本地训练速度不同而造成的等待开销.

近年来，研究者提出了模型异构联邦学习[71-75]（model-heterogeneous federated learning），它允许参与联邦学习的不同设备根据自身训练数据分布和设备资源状况选择结构合适的模型，而且能在不同结构的模型之间实现知识共享，从而既能用于统计异构环境也能用于设备异构环境. 模型异构联邦学习在端侧设备和服务器之间进行知识传递的媒介主要有

模型参数、模型输出和无标签样本3种类型，由此可以将模型异构联邦学习分为参数类[71]、输出类[73]和样本类[74]三类. 在参数类模型异构联邦学习方面，HeteroFL[71]根据端侧设备的计算能力分配复杂度不同的模型并使简单模型成为复杂模型的组成部分，从而实现了在复杂度不同的模型之间的参数聚合. FedRolex[72]使用轮转式子模型生成方式对HeteroFL进行了改进，使得全局模型的各部分在聚合过程中能被均匀地更新，在统计异构场景下取得了比HeteroFL更好的效果. 在输出类模型异构联邦学习方面，FedGKT[73]允许端侧设备和服务器拥有结构不同的模型，通过在服务器和端侧设备之间传递模型输出来使得各自的本地模型拥有相似的输出，从而实现知识的传递，但该方法所训练模型在统计异构的数据集上可能有显著的精度下降. 在样本类模型异构联邦学习方面，DS-FL[74]使用中心服务器聚合各端侧设备本地模型对无标注数据集的模型输出，然后使用聚合结果对无标注数据集进行标注，再使用经过标注的新数据训练各端侧设备的本地模型，从而通过数据扩增效应提升各端侧设备本地模型的精度.

2.1.2 联邦多模态学习

AIoT既为多模态深度学习模型的训练提供了海量的多模态训练数据，也为多模态模型的应用提供了丰富的场景. 通过联邦学习训练多模态深度学习模型是近几年的新兴研究方向[76]，有着广阔的发展空间.

实现联邦多模态学习的关键在于综合使用来自不同模态训练数据的知识. 联邦多模态学习可以通过从不同模态数据中学习共享隐藏表示来实现. MmFL[56]基于FedAvg算法训练能从各端侧设备的多模态数据中学习共享隐藏表示的自编码器，然后使用共享隐藏表示在中心聚合服务器上训练分类器，从而能使用多模态数据进行行为识别. MMFed[77]通过联邦学习训练注意力模块和分类器. 注意力模块使用共同注意力机制[78]来学习不同模态之间的内在联系，然后将融合特征向量输入分类器以得到分类结果. Liu等人[79]提出了面向视觉问答任务的联邦学习框架aimNet，该框架使用联邦学习训练特征融合的神经网络和解码器. 其中，特征融合网络基于多头注意力机制将CNN提取的视觉特征和词嵌入提取的文本特征进行融合，解码器则基于融合特征向量得到任务输出. 在联邦多模态学习中，不同端侧设备可能通过传感器收集模态类型不同的数据，从而带来模态不一致(modality incongruity)问题. 为此，FedMSplit[80]在端侧设备本地不同结构的多模态模型之间通过共享的模型参数块来传递知识，从而减轻模态不一致带来的负面影响. HA-Fedformer[81]通过单模态参数聚合和跨模态参数聚合来分别训练各种模态输入数据的编码器和基于多模态信息进行预测的解码器，实现了单模态训练多模态推理的联邦学习. 基础模型等大模型的兴起使得多模态深度学习模型的训练过程更为复杂，与大模型相关的联邦多模态学习目前还是较少涉及的领域.

2.1.3 联邦持续学习

在联邦学习过程中，由于AIoT端侧设备的加入和退出、应用场景的变化以及用户在使用过程中产生新数据等原因，来自AIoT端侧设备的训练数据的分布是动态变化的[25, 82-83]，因此模型训练过程需要不断适应新的训练数据分布. 然而，深度学习模型在适应新的数据分布的过程中，可能会经历灾难性遗忘[57]，从而在旧的数据分布上表现不佳. 这一挑战可以通过联邦持续学习予以应对.

联邦持续学习的目标是在训练过程中既提升模型在新任务上的表现又使模型在旧任务上的表现下降得不过于明显[84]，它可以通过知识蒸馏、参数分离、损失函数调和、输出聚类等方法实现这一目标. 在知识蒸馏方面，FLwF[85]在联邦学习过程中通过知识蒸馏同时从服务器端的全局模型和端侧设备的旧模型中学习知识，从而在学习新知识的同时保留旧知识. 然而知识蒸馏需要不断在端侧设备和服务器之间传递全局模型参数，通信开销较大. FedWeIT[86]通过在任务之间共享参数来学习多个任务的共有知识，然后通过个性化掩码从共享参数中提取适合端侧设备当前任务的参数，再使用任务独有参数进一步提升模型对当前任务的适应性，从而提升了联邦持续学习在多个任务上的表现并通过稀疏化参数减少了通信开销. Cross-FCL[58]基于文献[86]中的参数分离方法实现了联邦持续学习，使得智能汽车能在不同的联邦学习系统间移动时既能学习新任务的知识又能保留旧任务的知识. 在联邦持续学习中还可能出现不同任务的特征空间不同的情况. 为解决这个问题，CHFL[87]训练公共神经网络来处理任务的公共特征向量，训练任务独有的神经网络来处理任务独有的特征向量. 输出聚类也能实现联邦持续学习，Protonet[88]根据一个样本对应的神经网络输出与各类别样本对应神经网络输出的质心的距离来确定该样本所属类别，通过联邦学习不断修正神经网络对各类样本所输出向量的质心来提升模型对新类别样本的泛化能力. FedReg[89]通过损失函数约束使得本地模型在更

新后对本地样本的输出改变不至于过大，从而加速了联邦持续学习的收敛速度并减少了通信开销. 损失函数调和可以减轻统计异构对联邦持续学习带来的影响. GLFC[90] 基于新类别的样本和旧类别的样本在训练过程中的梯度大小来对这 2 种样本的损失值进行加权平均，从而使模型训练过程同时适应新旧类别的样本.

由于基础模型可能用于一系列下游任务，其联邦持续学习需要使大模型能持续在下游任务上取得良好的表现. 然而，大模型的灾难性遗忘机制目前尚不明确，而且在大模型中难以确定需要更新和保留的模型参数以引入新知识并保存旧知识[28]，因此如何通过联邦持续学习提升大模型的表现依然需要进一步研究.

2.1.4 联邦强化学习

联邦强化学习[59] 可以为 AIoT 设备训练决策模型并保护设备的数据隐私，被广泛应用于智能机器人[82]、智慧交通[91] 和智慧电网[92] 等 AIoT 应用场景. FedRL[59] 基于联邦学习为多个智能体训练一个共享决策神经网络来提升智能体任务执行的成功率. 为提升云机器人系统在导航任务中的表现，LFRL[82] 通过终身联邦强化学习为不同环境中的机器人训练共享决策神经网络，使机器人能更快速地适应不同的环境. 为了改善交叉路口的信号灯控制，FL-A2C[91] 使用 FedAvg 训练多智能体强化学习决策模型来进行信号灯控制决策.FDRL[93] 使用联邦强化学习来控制建筑物的能源管理系统以减少能耗.

联邦强化学习也在缓存分配[94]、负载均衡[95] 和路由协议[96] 等方面优化了 AIoT. 为提升动态移动通信网络环境中的边缘缓存访存效率，FADE[94] 使用 FedAvg 算法来为 AIoT 端侧设备训练缓存策略决策模型以优化其访问基站中缓存数据的平均延迟和缓存命中率. 动态网络环境和设备异构使得边缘服务器难以为 5G 超密组网中的多个物联网设备提供稳定和低延迟的边缘计算服务，2Ts-DRL[95] 使用联邦学习在 5G 边缘计算场景中为多个物联网端侧设备训练任务划分和资源分配决策模型，有效降低了任务执行延迟. 5G 工业物联网中的大量设备进行数据传输可能带来巨大的通信压力，QoSPR[96] 路由协议通过联邦强化学习对 5G 工业物联网中的网关部署进行优化来减少通信延迟和实现良好的负载均衡.

联邦强化学习可以在包含通信传输波动和多个异构设备的复杂动态场景中为多个 AIoT 设备训练决策模型并保护设备本地数据隐私[95]，未来将对智能机器人和无人机群等包含多智能体决策的 AIoT 场景产生更加深远的影响.

2.2 协同推理的算法

协同推理相关的算法主要包括模型切分算法和任务调度算法，其主要优化目标包括性能、动态环境中推理延迟的鲁棒性和能耗等.

在性能方面，DeepThings[49] 将神经网络的卷积层切分为互不重叠的部分来实现并行化的 CNN 协同推理. 该研究还允许空闲设备通过任务窃取来进行任务调度，从而有效利用空闲设备来加速协同推理过程. 然而，DeepThings 的模型切分策略主要适合链式 CNN，而不适合 GoogleNet 和 ResNet 等块状堆叠深度学习模型. 为此，DeepSlicing[26] 的模型切分策略将支持的 CNN 类型扩展到了所有包含有向无环图结构的 CNN. 对于 CNN 中任意神经网络层的给定输出范围，DeepSlicing 根据得到该输出范围数据所需的来自前继神经网络层的最小输入数据进行模型切分，从而减少重复计算开销和设备间通信开销. 在进行任务调度时，DeepSlicing 综合考虑了任务的计算量和设备的计算能力，使得设备完成各自计算任务的时间接近，从而减少了设备间的同步开销. DeepThings 和 DeepSlicing 考虑的主要是封闭场景中协同推理任务在不同 AIoT 设备之间的调度，IONN[97] 则研究了开放场景下可穿戴式设备将协同推理任务通过移动通信网络分配给边缘服务器时的任务调度问题. 由于在开放场景下参与协同推理的可穿戴式设备不断变化，IONN 考虑了新加入设备将模型切片上传至边缘服务器的通信开销. 该研究将端边协同推理中神经网络的执行过程转化为执行图，执行图中的点代表不同的神经网络层，边的权重则表示神经网络层的预估执行时间或通信时间，然后通过最短路径搜索算法选择执行时间最短的执行方案.

近年来对协同推理性能优化相关研究逐渐转向网络结构更复杂（如通用的有向无环图结构）和包含多个异构设备的应用场景，动态规划和线性规划是优化模型切分和任务调度的主要方法[98-100]. OFL[98] 基于层融合对模型进行切分来减少设备间的通信开销. 在进行任务调度时，该研究使用线性回归来预测卷积层的执行时间，然后根据设备的计算和通信能力使用动态规划来将模型切片分配给设备执行，以适应包含异构设备的环境. PICO[99] 将模型依次划分为多个结束片，每一个结束片的输入全部来自上一个结束片的输出且其输出全部作为下一个结束片的输入. 然后该研究通过动态规划将结束片分配给多个

异构设备, 实现了协同推理在多个异构设备上的流水线并行, 提升了协同推理的吞吐量. EdgeFlow[100] 基于线性规划将深度学习模型切分为互相独立的执行单元来使得协同推理过程中各设备的负载更加均衡, 然后通过并行执行协同推理任务降低了延迟. 根据延迟预测模型的预测延迟, IAO[50] 在本地协同推理中通过调整本地服务器为多个异构端侧设备分配的资源和端侧设备在本地执行的任务量来最小化所有协同推理中的最大延迟.

动态规划和线性规划一般依赖用于评估执行方案的开销模型, 而运行时变化因素 (网络波动和设备负载变化等) 和异构设备计算能力的非线性特征可能导致不准确的评估结果[101-102]. 运行时变化因素还可能导致协同推理的延迟不稳定, 从而影响服务质量[103]. 为此, ICE[103] 基于服务质量感知的任务批调度和任务重划分改善了端云协同推理中通过执行图最短路径搜索生成的任务调度方案, 增强了推理延迟对任务负载变化、模型结构变化和网络状况变化的鲁棒性. 强化学习可以根据环境的改变作出新的决策并根据反馈不断更新决策依据, 从而不断适应环境的动态变化. DistrEdge[102] 使用 DDPG 强化学习算法[104] 对存在多个异构 AIoT 设备和网络波动的协同推理场景中的模型切分和任务调度进行优化, 有效提升了动态环境中协同推理的性能. MTS[105] 以模型树的形式来表达模型压缩、模型切分和任务调度的联合策略. 该研究进一步通过强化学习在动态网络环境中为协同推理选择合适的模型树分支, 从而有效提升了协同推理对网络波动的适应性. 由于 AIoT 应用场景具有设备多样、网络波动和设备执行状态动态变化等特征, 强化学习将在其协同推理的模型切分与任务调度优化中发挥更大的作用.

由于对配备电池的 AIoT 设备来说, 电池的容量和寿命是有限的[106], 因此能耗也是协同推理在 AIoT 应用场景中需要考虑的一个重要因素. CoEdge[21] 建立了估计不同执行方案的延迟和能耗的开销模型, 然后通过面向模型并行的切分和线性规划来求解调度优化问题的最优解, 以在满足延迟要求的情况下最小化推理能耗. AutoScale[101] 将深度学习模型的结构特征和端侧设备的设备状态 (CPU 利用率、内存占用情况和网络传输质量) 作为环境的状态特征来训练 Q-Learning 决策模型, 以选出满足推理延迟和精度等服务质量要求的前提下最小化能耗的推理任务执行方案. Neurosurgeon[107] 基于预测模型分别对性能和能耗 2 个优化目标下的模型切分点进行选取, 有效减少了端云协同推理中的端到端延迟和端侧设备能耗.

2.3 小 结

本节主要介绍了 AIoT 应用场景中联邦学习和协同推理相关的算法. 在联邦学习方面, 本节从异构环境中的联邦学习、联邦多模态学习、联邦持续学习和联邦强化学习 4 个方面总结了联邦学习近年来的相关进展, 相关内容如表 2 所示. 在大模型时代, 联邦学习可以用于训练大模型, 大模型也可以用于产生联邦学习所用的训练数据, 两者的联系将更加紧密[108]. 展望未来, 模型异构联邦学习为实现大小模型之间的知识传递提供了可行方案, 但在参数量巨大的模型中确定每部分参数的作用是一件困难的事情[28], 因此在参数量巨大的大模型与小模型之间实现知识传递依然是一件需要研究的事情.

Table 2 Summary of Related Works About the Algorithm of Federated Learning

表 2 联邦学习的算法相关工作总结

联邦学习算法		适用情形	参考文献
异构环境中的联邦学习	统计异构	不同 AIoT 端侧设备上样本训练样本分布不同	[61−66]
	设备异构	不同 AIoT 端侧设备进行本地模型训练的速度不同	[67−70]
	模型异构	统计异构或设备异构环境	[71−75]
联邦多模态学习		AIoT 端侧设备收集的数据具有多模态特性	[56, 77, 79−81]
联邦持续学习		AIoT 端侧设备的数据分布随时间变化	[58, 85−90]
联邦强化学习		基于联邦学习为 AIoT 设备训练决策模型	[59, 82, 91, 93−96]

在协同推理方面, 本节面向性能、能耗和延迟鲁棒性 3 个优化目标总结了相关的算法, 相关内容如表 3 所示. 近年来, 一些研究工作将分离式学习的思路引入联邦学习, 提出了联邦分离式学习[47] 的概念. 文献 [27, 47] 将部分模型切片的训练任务从资源受限的端侧设备分配给服务器, 来缓解端侧设备计算资源的紧张. 由于联邦分离式学习的训练过程也采用模型切片的方式进行, 因此也可以通过模型切分和任务调度对其进行优化[27], 这体现了联邦学习和协同推理的内在联系, 而这类研究目前还比较少. 未来研究人员可以借鉴协同推理领域的模型切分优化算法来对联邦分离式学习的性能和能耗等指标进行优化.

3 架 构

在面向 AIoT 的协同智能中, 分布在端、边和云等不同层级的多个单设备智能计算系统通过算力供给

Table 3 Summary of Related Works About the Algorithm of Collaborative Inference

表 3 协同推理的算法相关工作总结

主要优化目标	相关工作	模型切分方法	任务调度方法
性能	DeepThings[49]	卷积层并行	任务窃取
	DeepSlicing[26]	通信量优化、模型并行	同步开销优化
	IONN[97]	执行图生成与最短路径搜索	
	OFL[98]	基于层融合的模型切分	动态规划
	PICO[99]	基于结束片的模型切分	动态规划
	EdgeFlow[100]	模型并行	线性规划
	IAO[50]	延迟预测	
	Neurosurgeon[107]	基于延迟预测的模型切分	
延迟鲁棒性	DistrEdge[102]	强化学习	
	ICE[103]	服务质量感知、执行图最短路径搜索	
	MTS[105]	强化学习	
能耗	CoEdge[21]	模型并行	线性规划
	Neurosurgeon[107]	基于能耗估计的模型切分	
	AutoScale[101]		强化学习

和知识传递相互协作，共同完成联邦学习和协同推理等智能计算任务. 本节参考智能计算系统的架构层级将面向 AIoT 的协同智能的架构分为深度学习加速器、深度学习编译[109]、深度学习框架、设备间通信和多设备协同方式 5 个层级，分别总结相关工作.

3.1 深度学习加速器

深度学习加速器为智能计算系统提供了重要的硬件支持[31]，它可以用来加速联邦学习的训练过程和协同推理的模型切片推理过程.

对于计算资源较为丰富的 AIoT 端侧设备，联邦学习的训练过程和协同推理可以在 CPU 和本地深度学习加速器上进行[110-111]. 文献 [110] 在包含 CPU 和 GPU 的异构端侧设备上实现了联邦学习本地训练的数据并行加速. 文献 [111] 在 HiKey 970 的嵌入式开发板上实现了包含 CPU、GPU 和深度学习处理器等多种设备的协同推理. 资源较为贫乏的 AIoT 端侧设备可以将联邦学习本地训练任务或协同推理任务分配给计算资源更为丰富的边缘服务器和云服务器. 文献 [27, 47] 通过联邦分离式学习将部分模型的训练任务分配给边缘服务器或云服务器上的 GPU 执行. 协同推理任务也可以通过边缘服务器或云服务器上的 GPU 进行加速[41, 97, 101]. 由于 GPU 需要兼顾通用计算任务面积较大且执行深度学习任务的能效比（性能与能耗的比值）较低，DaDianNao[112] 和 TPU[113] 等服务器级深度学习处理器为深度学习任务设计了专门的硬件加速结构，较好地提升了能效比. 文献 [114−115] 使用 TPU 加速联邦学习中端侧设备的本地训练过程. Villarrubia 等人[116] 基于性能评估切分深度学习模型，然后分配给多个 TPU 进行协同推理，从而降低了单个 TPU 访问主机内存的开销并使得协同推理过程流水线化，提升了推理的性能.

近年来，为进一步提升联邦学习的效率，研究人员提出了联邦学习加速器[117-118]. 联邦学习加速器既加速联邦学习的一般计算过程（如梯度计算和参数聚合等）也加速联邦学习的安全机制（如同态加密和噪声添加等）. PipeFL[117] 为联邦学习提供了基于 FPGA 的硬件加速. 在联邦学习过程中 RSA 算法和 Paillier 同态加密算法[119] 可以分别用于实现安全的数据对齐和参数聚合，这 2 种算法主要涉及模幂和模乘 2 种操作. PipeFL 基于蒙哥马利乘法单元实现了模幂和模乘操作的 FPGA 硬件加速，从而提升了联邦学习的效率.FLAIRS[118] 对联邦学习中安全聚合算法 FLAME[120] 的参数聚合和噪声添加过程进行了 FPGA 硬件加速，在防止后门植入的同时提升了参数聚合的效率.

展望未来，芯片敏捷设计[121]、开源硬件[122] 和 FPGA 加速器[123] 将为快速开发 AIoT 端侧深度学习加速器提供重要支撑，从而使加速器的开发能适应不断更新的 AIoT 应用需求. 由于联邦学习和协同推理都涉及设备间交互，通过减少硬件碎片化来提升设备的交互效率成为一个重要的发展方向[124]. 为适应大模型对算力的需求，数据中心对深度学习加速器的需求将进一步提升，智能计算集群的规模将进一步扩大，由此对集群在性能、能耗和可扩展性等方面的优化提出了更高的要求[28]. 随着联邦学习的进一步发展，联邦学习加速器将得到更加广泛的应用，是一个值得进一步探索的研究方向.

3.2 深度学习编译

深度学习编译器是智能计算系统软件栈的重要组成部分，它可以将深度学习框架中对深度学习模型的描述转换为在硬件上可以执行的代码[31, 125]. 深度学习编译一般可以分为即时编译和预编译 2 种类型：即时编译在程序运行时产生可执行代码并执行这些代码，因此可以使用运行时信息进行编译优化；预编译则先产生可执行的二进制程序再进行程序执行，需要基于静态分析对可执行程序进行优化[109]. 深度学习编译可以用于单设备智能计算系统上训练和推理任务的优化，因此也可以用于联邦学习和协同推理的优化.

XLA（accelerated linear algebra）是一款支持

TensorFlow 和 PyTorch 等常用深度学习框架及 CPU, TPU, GPU 等不同硬件平台的深度学习编译器[116, 126], 可以对深度学习训练和推理任务进行性能优化[126]. FedJAX[114] 通过 XLA 编译器为联邦学习端侧设备的本地训练过程生成 GPU 或 TPU 的高性能可执行代码, 从而实现本地训练过程的加速. JaxPruner[115] 进一步为联邦学习中端侧设备的本地稀疏化训练生成 XLA 高性能可执行代码来提升训练过程的效率. FedJAX 和 JaxPruner 的编译过程均通过即时编译实现. 预编译也可以用来加速模型训练, 从而可以用来加速联邦学习的本地训练过程, 成为未来可能的发展方向之一. PyTorch 支持使用预编译对模型训练进行算子融合和激活值重计算等优化, 在提升训练速度的同时减少了内存开销[127].

在协同推理方面, 深度学习编译器可以通过预编译为深度学习加速器生成高性能可执行代码. Edge TPU 编译器[128] 可以将模型编译为 TPU 可执行代码. 文献 [116] 使用 Edge TPU 编译器对模型切片的访存和计算行为进行分析, 从而能减少协同推理过程中 TPU 访问主机内存的开销和实现推理过程的流水线并行. TensorRT 编译器可以为计算图中的算子选择合适的预编译算子和可用的 GPU, 从而为 TensorRT 运行时生成高性能可执行代码[129]. TensorRT-LLM[130] 基于 TensorRT 编译器实现了大语言模型在多个 GPU 上的协同推理过程. 为了提升推理性能, TensorRT-LLM 还对协同推理进行了数据并行和流水线并行的优化.

展望未来, AIoT 生态的碎片化特征和协同智能等需要多种异构设备共同参与的应用场景, 可能使得编译过程需要考虑任务在多个异构设备上的任务划分和执行优化, 该领域仍需要在未来进行进一步的研究[109]. ONNX[131] 和 MLIR[132] 分别为深度学习模型和碎片化的软硬件提供了统一的模型表示和统一的编译中间表示, 将 ONNX 模型通过 MLIR 进行深度学习编译优化成为一种多异构设备应用场景中富有潜力的编译优化方案[133].

3.3 深度学习框架

深度学习框架可以为智能算法的编程、调试和部署提供软件平台, 它是智能计算系统的重要组成部分[31]. 在面向 AIoT 的协同智能中, 深度学习框架也能为联邦学习和协同推理任务提供重要的软件支持, 主要用于联邦学习中的模型训练和参数聚合以及协同推理中的模型切片推理.

联邦学习可以通过 TensorFlow 和 PyTorch 等端边云通用深度学习框架实现[72, 77, 85, 134]. 在对 AIoT 应用场景的支持方面, TensorFlow Lite 提供了在嵌入式设备上执行的轻量级运行时, 可以执行设备本地训练任务[135], 从而可以用于 AIoT 应用场景中的联邦学习任务[136]. 开源联邦学习框架 FedML[137] 基于 PyTorch 提供了对 AIoT 设备本地训练的支持, 通过 MPI 和 MQTT 等协议提供了对物联网分布式通信的支持, 基于拓扑结构管理器提供了对多种联邦学习拓扑结构(中心化、层次化和去中心化拓扑结构等)的支持和基于 FedML-IoT 提供了对 Raspberry PI 4 和 NVIDIA Jetson Nano 等真实 AIoT 设备的实验平台支持. 开源联邦学习框架 Flower[138] 支持 Raspberry PI 和 NVIDIA Jetson 等 AIoT 设备, 而且提供了对大量异构端侧设备和各种网络通信状况的模拟功能, 便于将实验室环境中的联邦学习系统部署于实际应用场景.

深度学习框架也被广泛用于协同推理过程, IAO[50], CoEdge[21], MOSAIC[111], EdgeFlow[100], PICO[99], ICE[103] 等前述协同推理工作所用的深度学习框架包括 TensorFlow, TensorFlow Lite, Pytorch 等. JEDI[139] 在 NVIDIA Jatson AGX Xavier 上使用 TensorRT 实现了 GPU 和深度学习处理器的协同推理并对其进行了流水线优化. 文献 [140] 比较了 MNN, TensorFlow Lite, NCNN, CoreML, MACE 等不同深度学习框架的性能, 其实验结果表明, MNN 在 GPU 上的推理延迟与 TensorFlow Lite 相比具有显著优势. 为了进一步支持异构端侧设备上运行的多种机器学习任务(计算机视觉、自然语言处理和推荐系统等), Walle[141] 基于 MNN 实现了一个端云协同机器学习系统.Walle 以大小模型协同的方式为各种任务取得性能与服务质量之间的平衡, 它在端侧设备使用小模型为用户提供高性能的机器学习服务, 但如果端侧设备上的小模型不足以为用户提供高质量的机器学习服务, 则使用云端大模型来为用户提供更高质量的机器学习服务.

随着深度学习模型日趋复杂和大模型等参数量巨大的模型的出现, 未来的联邦学习任务和协同推理任务将需要服务器端深度学习框架拥有适应大模型的能力. 在联邦学习方面, 模型异构联邦学习提供了在大小模型间进行知识传递的有效方法, 联邦学习框架可以通过集成模型异构联邦学习算法来适应大模型的发展. 在协同推理方面, 单个深度学习加速器的计算能力和设备内存容量已经无法满足大模型推理任务的需求, 服务器级深度学习框架需要具备良好的可拓展能力. DeepSpeed[142] 采用了 GPU 显存、DRAM 和 NVMe 组成的混合存储系统来满足大模型

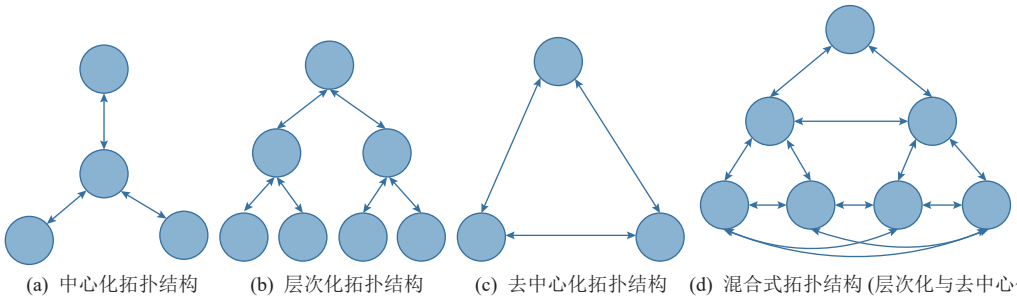
参数对内存的需求. 在推理延迟和吞吐量的优化方面, DeepSpeed 将大模型推理任务分配给多个 GPU 执行. 在推理过程中, DeepSpeed 采用流水线、算子融合、GPU 算子优化、张量并行和通信并行等方法进行了多层次的性能优化. 此外, DeepSpeed 对混合专家模型还进行了专家并行的推理优化.

3.4 设备间通信

在面向 AIoT 的协同智能中, 知识传递和算力供给都是通过单设备智能计算系统间的通信来完成的, 因此设备间通信是进行其架构设计时需要考虑的重要因素. 设备间通信主要涉及通信拓扑结构的选取以及通信开销的优化, 通信拓扑结构对通信效率和系统管理方式有着重要影响. 文献 [137] 将联邦学习的通信拓扑结构分为中心化拓扑结构、去中心化拓扑结构和层次化拓扑结构等类型, 文献 [30] 提出不同的拓扑结构类型可以混合, 如中心化拓扑结构和去中心化拓扑结构可以混合. 如图 2 所示, 本文将上述通信拓扑结构类型从联邦学习扩展至协同推理, 将面向 AIoT 的协同智能的通信拓扑结构分为中心化拓扑结构[44, 97]、层次化拓扑结构[143-144]、去中心化拓扑结构[49, 145] 和混合式拓扑结构[68]. 在中心化拓扑结构中, 多个端侧设备将深度学习任务分配至作为中心聚合服务器的边缘服务器或云服务器. 例如, FedAvg[44] 将联邦学习中的参数聚合任务分配给云服务器; IONN[97] 将协同推理任务分配给边缘服务器. 中心化拓扑结构的优点是系统结构简单、便于管理, 但中心服务器易成为通信和计算的瓶颈[146].

为了缓解中心服务器资源紧张的情况, 可以在中心服务器和端侧设备之间增加中间服务器来分担中心服务器的计算任务, HierFAVG[143] 和 CNNPC[144] 所采用的端边云架构就是典型的层次化拓扑结构. 无论是中心化拓扑结构还是层次化拓扑结构都可能依赖于第三方提供的计算服务(如服务商提供的边缘服务器和云服务器)和系统管理, 这带来了隐私安全的风险, 去中心化拓扑结构一定程度上减轻了这种风险. 在联邦学习领域, 参与者可以通过区块链和共识协议实现模型训练过程的自治和拜占庭容错[147], 但共识过程带来了额外的系统管理开销; 在协同推理领域, AIoT 端侧设备可以使用 P2P 的通信架构实现任务调度, 以提升推理任务的性能[49], 但其任务调度过程较为复杂. 值得注意的是, 不同的通信拓扑结构可以混合使用. 例如, 在文献 [68] 的混合式车联网联邦学习中设计者采用了层次化和去中心化的混合通信拓扑结构, 既可以通过逐层参数聚合加快全局模型的生成, 也为车辆的参数聚合过程提供了拜占庭容错的验证.

通信开销的优化可以减少设备间数据传递的延迟, 是提升联邦学习和协同推理任务执行性能的重要方法. 通过降低数据传输量可以减少通信开销, 主要的方法有减少通信的次数和减少每次通信的数据量. 减少通信的次数通常用于减少联邦学习过程中的设备间通信开销. HierFAVG[143] 通过将梯度传递给通信延迟较小的边缘服务器进行聚合, 再由多个边缘服务器将各自的聚合结果上传至云端服务器进行二次聚合来减少高延迟的云端服务器聚合的次数. CMFL[148] 基于模型参数更新的大小和符号来判断其与全局模型更新趋势的契合程度, 继而通过过滤与全局模型更新趋势相关性不高的模型参数更新来减少通信频率. 通信次数的减少还可以通过加速收敛过程来实现. Yu 等人[149] 允许联邦学习中计算速度快的参与者比计算速度慢的参与者进行更多次的本地更新, 从而在加快模型收敛速度的同时减少了模型训练过程中参与者和中心聚合服务器的通信次数. AutoFL[150] 从理论上分析了基于模型不可知元学习(model-agnostic meta-learning, MAML)[34] 的联邦学习为达到目标精度所需的通信轮数上界, 为基于 MAML 的联邦学习选择最佳通信轮数提供了理论依据. 一次性联邦学习可以将通信轮数限制为 1 轮, 从



(a) 中心化拓扑结构 (b) 层次化拓扑结构 (c) 去中心化拓扑结构 (d) 混合式拓扑结构 (层次化与去中心化)

Fig. 2 The different communication topology structure[68,137]

图 2 不同的通信拓扑结构[68,137]

而大幅度减少通信开销[151]. FedCAVE[151]在端侧设备本地训练条件变分自动编码器, 然后将标签分布和解码器参数一次性传递给中心服务器, 从而使中心服务器可以使用来自各端侧设备的标签分布和解码器生成的样本来训练分类器.

减少每次通信的数据量可以用于联邦学习和协同推理2个过程中的通信优化. 在联邦学习方面, 文献[152−153]对联邦学习过程中参与者向中心聚合服务器传输的梯度和中心聚合服务器向参与者传输的全局模型都进行压缩来减少通信量. 为了减少联邦学习的上行带宽压力, Hönig等人[154]提出一种双层自适应压缩方法DAdaQuant, 它可以在不同的时间为不同的端侧设备选择合适的量化水平, 从而在提升压缩率的同时保持了全局模型的高精度. HCFL [155]先使用自编码器对端侧设备的模型进行压缩, 然后在服务器上重建端侧设备模型, 从而提升了联邦学习的通信效率. DisPFL[156]通过去中心化通信拓扑结构和稀疏化训练减轻了联邦学习中繁忙节点的通信压力. FD[157]使用知识精炼来训练全局模型, 通过传输端侧设备本地模型的输出而不是模型参数来减少通信开销. 在协同推理方面, CNNPC[144]同时考虑了模型压缩对推理精度的影响, 通过实际采样的数据建立了压缩率和精度之间的数学关系模型, 然后根据该模型在压缩率和精度之间取得了良好的平衡. RtCoInfer[158]通过实时选择不同的神经网络量化分支来实现动态网络环境中的自适应压缩率控制, 减少了任务执行超时的可能性.

在共享通信网络中, 多个同时进行数据传输的设备可能会互相干涉, 从而降低通信效率[159-160], 因此需要对通信干涉进行管理. Chen等人[159]为多个共享移动通信网络的端侧设备求解通信通道选择策略和任务分配策略的纳什均衡, 有效减少了任务执行的总体开销. Yi等人[160]基于排队理论对多通道数据传输过程的网络拥塞问题进行分析的结果表明, 对于任意一种给定的任务分配比例策略, 数据在通道队列中的等待时间所造成费用的最小化问题可以归结为凸优化问题, 进而可以求解最优传输调度策略. Ale等人[161]使用深度强化学习为端侧设备在通信状况不断变化的情况下从多个边缘服务器中选择合适的对象来分配任务, 从而提升准时完成的任务数量和降低能耗. 近年来, 5G[96]和Wi-Fi 6[162]等现代通信技术的发展有力地支持了大量端侧设备共享通信网络进行高速通信, 从而为面向AIoT的协同智能提供了新的发展机遇.

3.5 多设备协同方式

多设备协同方式主要涉及面向AIoT的协同智能中多个单设备智能计算系统的组织方式, 它与AIoT应用场景有较强的关系. AIoT的应用场景可以分为封闭场景和开放场景2种类型. 前者包括智慧仓储[163]、智慧工厂[2]和智慧家居[164]等, 参与的设备包括智能机器人、智能摄像头和智能语音助手等计算资源较为贫乏的AIoT端侧设备, 其特点是地理分布范围较小、可以通过局域网进行通信; 后者包括智慧农业[4]、智慧交通[5]和智慧城市[165]等, 参与的AIoT端侧设备既包括智能无人机和智能摄像头等计算资源较为贫乏的AIoT端侧设备, 也包括智能汽车等计算资源较为丰富的AIoT端侧设备, 其特点是地理分布较为广泛, 一部分AIoT端侧设备(如智能汽车和智能无人机)等可以在较大的地域范围内移动, 一般采用移动互联网进行通信. 文献[16]将协同推理中的多设备协同方式分为端云协同推理、端边协同推理、端边云协同推理和设备到设备的协同推理4种类型. 本文将这4种类型扩展至面向AIoT的协同智能, 将其多设备协同方式分为端云协同、端边协同、端边云协同和本地协同4种方式. 其中, 本地协同既包含封闭场景中多个设备之间的协同, 也包括单设备中CPU与深度学习加速器之间的协同.

在上述2种应用场景中, AIoT端侧设备都可以通过广域网将深度学习模型的训练和推理任务分配至云服务器, 从而形成端云协同的架构[101, 107, 166-167]. 然而, 广域网数据传输的延迟较高且波动较大[21], 难以满足一些对实时性要求较高的AIoT应用场景的性能要求[16]. 边缘计算通过在靠近AIoT设备的地方提供计算服务减少了数据传输的延迟和网络波动, 成为实现面向AIoT的协同智能的重要选择[30]. 在封闭场景中, AIoT端侧设备(或异构设备中的CPU)可以通过局域网(或高速总线和片上网络)等将训练和推理任务分配给本地其他计算设备(或异构设备中的深度学习加速器), 以实现本地设备的分布式训练[23, 168]和本地协同推理[21, 26, 49, 111]. 在开放场景中, AIoT端侧设备可以通过移动通信网络借助边缘服务器的算力进行深度学习模型的训练和推理[58, 97]. 在需要大量算力和长期存储的情况下, 云服务器依然对边缘服务器具有优势, 因此成为边缘计算的有益补充[143-144, 169]. 端边云协同成为面向AIoT的协同智能在架构上的长期发展趋势.

展望未来, 随着大模型时代的到来, 使用AIoT设备收集的海量多模态数据训练多模态大模型来为

用户提供种类丰富的智能服务成为发展趋势[28]. 然而，大模型参数量巨大，难以部署在边缘服务器和 AIoT 端侧设备等资源受限的设备上，但仅在云服务器上进行大模型的训练和推理可能会造成网络拥塞和较高的通信延迟[169]. 大小模型协同[72, 141, 170]成为一种架构设计新范式. Lu 等人[170]提出云服务器可以通过模型压缩将包含云端大模型知识的小模型下发至端侧设备，由小模型为用户提供低延迟的推理服务. 端侧设备使用本地数据对小模型进行个性化训练，而大小模型之间则通过知识精炼进行知识传递，使得大模型能获取来自小模型的个性化知识以实现持续更新，而小模型则可以通过获取来自大模型的知识来防止过拟合. 除知识精炼外，大小模型之间还可以通过模型压缩和模型异构联邦学习等多种方式传递知识[72, 141]. 大小模型协同在未来将会得到更广泛的应用，因此是一个重要的研究方向.

3.6 小 结

本节从智能计算系统的角度将架构方面的相关工作分为深度学习加速器、深度学习编译、深度学习框架、设备间通信和多设备协同方式等不同层级，表 4 展示了各架构层级的相关工作. 联邦学习和协同推理在上述 5 个架构层级均能进行优化且优化方法

Table 4 Summary of Related Works at Different Levels of AIoT-Oriented Collaborative Intelligence Achitecture

表 4 面向 AIoT 的协同智能架构各层次相关工作总结

架构层级	分类		优势	劣势	参考文献	
					联邦学习	协同推理
深度学习加速器	GPU		高性能、软件栈成熟、兼顾通用计算任务	面积大、能耗高	[27, 47]	[41, 97, 101]
	深度学习处理器		面积较小、能效比高	任务类型相对单一	[114−115]	[111, 116]
深度学习编译	即时编译		可以获取运行时信息[109]	增加启动开销[171]	[114−115]	
	预编译		更大的静态搜索空间、支持交叉编译等[109]	无法获取运行时信息	[127]*	[116, 130]
深度学习框架	AIoT 联邦学习框架	FedML	基于 MPI 和 MQTT 的分布式通信协议支持、支持多种通信拓扑结构、对真实 AIoT 设备的实验平台支持	没有对推理任务提供专门支持和优化	[73, 137]	
		Flower	支持大量异构端侧设备和各种通信状况的模拟		[138]	
	轻量级端侧框架	TensorFlow Lite	支持嵌入式设备的轻量级运行时	一般只用于端侧设备	[136]	[21, 111]
		MNN	基于半自动搜索的最佳执行策略搜索			[141]
	端边云通用框架	PyTorch	编程风格简洁、多进程并行计算和通信优化	嵌入式设备等资源受限设备难以支持	[72, 77, 134]	[99−100, 103]
		TensorFlow	良好的可扩展性、调度策略优化		[85]	[50]
		TensorRT	高性能、高吞吐量推理	没有对训练任务提供支持		[139]
设备间通信	通信拓扑结构[30, 137]	中心化	结构简单、易于管理	中心节点通信瓶颈，可能依赖第三方提供的计算服务	[44]	[97]
		层次化	缓解中心节点通信瓶颈	增加额外通信层级，可能依赖第三方提供的计算服务	[143]	[144]
		去中心化	P2P 直接通信、系统自治、拜占庭容错	共识开销，系统管理复杂	[147]	[49]
		混合式	可以兼具多种通信拓扑结构的优点	结构和系统管理较为复杂	[68]	
	减少通信的次数		降低通信开销	一般只用于联邦学习场景	[143, 148, 150−151, 172]	
	减少每次通信的数据量			可能降低模型精度	[152−157]	[144, 158]
	通信干涉管理		减少通信干涉的负面影响	对 Wi-Fi 6 等新通信网络需要进一步研究	[159−161]	
多设备协同[16]	端云协同		云服务器计算、存储资源充足，有利于数据的长期存储	云服务器带宽受限、广域网不稳定、隐私安全问题	[166−167]	[41, 101]
	端边协同		降低通信延迟	边缘服务器计算和存储资源受限、隐私安全问题	[58]	[97]
	端边云协同		减轻云服务器计算和通信负担	隐私安全问题	[143]	[144]
	本地协同		高速和稳定的数据传输、隐私安全保障	只适用于封闭场景，不适用于开放场景		[21, 26, 49, 111]
	大小模型协同		既可以使用大模型中包含的丰富知识来提供高质量的服务，又可以使用小模型来提升服务的响应速度	大小模型之间的知识传递需要进一步研究	[72, 170]	[141]

注：“*”表示潜在解决方案.

具有相似性,这体现了两者在架构上的内在联系.在软硬件方面,联邦学习加速器、联邦学习编译优化和联邦学习框架为联邦学习提供了专用支持,而协同推理也可以通过专用的深度学习编译器对多个深度学习加速器进行任务划分优化,这体现了对两者的软硬件支持有往专用化方向的发展趋势.展望未来,随着联邦学习和协同推理在实际应用场景中的推广和配合使用,可能会出现综合支持联邦学习和协同推理的专用软硬件.

4 隐私安全

在面向 AIoT 的协同智能中,联邦学习和协同推理都在传递可能包含用户隐私信息的知识,部署在参与设备上的模型参数也可能包含用户隐私信息,而根据 GDPR 和《中华人民共和国数据安全法》等相关法律,用户的隐私信息必须得到保护.本节将总结面向 AIoT 的协同智能面临的不同攻击类型和防御方法并对大模型相关攻击的发展趋势进行展望.

4.1 数据隐私相关的攻击

包含隐私信息的数据主要有 2 类:一是用户的训练样本,如患者的医疗记录可以用于训练疾病诊断模型,但训练数据中包含着患者健康状况有关的隐私信息[166];二是模型参数,由于模型参数包含来自训练样本的用户隐私信息,而且模型可用于提升生产效率和便利社会生活并为拥有者带来经济利益[173],大模型的出现更是大幅度提升了模型的实用价值和训练成本[28],因此模型参数也是需要保护的隐私数据.与数据隐私相关的攻击主要包含窃取用户训练样本中隐私信息的攻击和窃取有价值的模型参数的攻击.前者主要包含模型反演攻击、成员推断攻击和性质推断攻击;后者主要包括模型提取攻击和 free-rider 攻击.

1)模型反演攻击.模型反演攻击可以通过获取联邦学习过程中端侧设备上传的模型参数更新来重构模型的训练样本[119],也可以通过获取协同推理过程中端侧设备上传的模型切片输出来重构输入样本[174].Phong 等人[119]证明了攻击者可以在使用 MNIST 数据集训练的神经网络时从未加密的梯度中重构出可识别的手写数字训练样本.CAFE[175]通过基于 SGD 的迭代优化过程从纵向联邦学习中大批量训练数据的梯度中恢复高质量的训练数据.模型反演攻击还可以用于协同推理场景[174].He 等人[52]将协同推理场景中的模型反演攻击分为白盒、黑盒和不提供模型切片访问 3 种情形.在白盒情形下攻击者可以使用梯度下降法调整输入样本来减少重构误差;在黑盒情形下攻击者可以根据模型切片的训练数据分布来训练逆向网络以重构输入样本;在不提供模型切片访问的情况下,攻击者可以先训练目标切片的影子模型再实施白盒情形下的模型反演攻击.Yin 等人[174]通过在样本上加入高斯噪声来产生一系列目标模型切片的输入样本,然后使用遗传算法来估计逆向网络的梯度,使得逆向网络向着重构误差不断减少的方向优化,实现了不依靠额外分布信息的黑盒攻击方法.

2)成员推断攻击.成员推断攻击的目的是确认某一样本是否存在于目标模型的训练数据集中.在联邦学习场景下按攻击者是否改变联邦学习过程的行为,成员推断攻击可以分为主动攻击和被动攻击.在主动攻击方面,Nguyen 等人[176]提出不诚实的中心聚合服务器可以通过观察联邦学习参与者提交的梯度是否激活特定神经元来判定能激活该神经元的训练样本是否存在于参与者的训练集中.在被动攻击方面,Li 等人[177]根据端侧设备训练集中的样本与非训练集中样本所生成的模型参数更新在余弦相似性上服从均值不同的高斯分布这一现象来进行被动成员推理攻击.在协同推理方面,Chen 等人[53]训练二分类器来判断一个样本是否属于目标模型切片的影子模型的训练集,进而使用该二分类器对原目标模型切片实施成员推断攻击.

3)性质推断攻击.性质推断攻击的主要目的是推断目标模型的训练样本是否具有某一性质,这种攻击一般用于联邦学习场景.Melis 等人[178]根据具有某一性质的训练样本与不具有该性质的训练样本在模型梯度上的差异性,通过训练二分类器来判定产生模型梯度的训练样本是否具有某一性质.为了加强基于二分类器的性质推断攻击,Wang 等人[179]提出攻击者可以将训练数据中具有待推断性质的训练数据进行标签翻转,从而迫使其他参与者投入更多具有该性质的训练数据来恢复模型精度.

4)模型提取攻击.模型提取攻击的目的在于窃取黑盒模型的模型参数等,以获取经济利益或为进一步的攻击做准备[173-174].该种攻击既可以用于联邦学习场景也可以用于协同推理场景.

在联邦学习的部署阶段,使用基于参与者本地数据训练的全局模型可能通过机器学习即服务(ML-as-a-service, MLaaS)对外提供服务,以获取经济利益[180].然而,攻击者可能通过模型访问 API 窃取全局模型的参数,从而能达到支付较少的费用就能享有模型

提供的预测服务的目的[173]. Truong 等人[181] 使用 GAN 生成目标模型和影子模型的输入样本，然后以目标模型和影子模型输出的 KL 距离为损失函数使用零阶梯度估计[182] 来训练能模仿目标模型的影子模型. 在协同推理方面，文献 [52, 174] 中攻击者训练目标模型切片的影子模型的方法即为典型的模型提取攻击，其目的在于为进一步发起模型反演攻击和成员推断攻击做准备.

5）free-rider 攻击. free-rider 攻击[183-184] 的目的在于窃取有价值的模型参数. 该种攻击的发起者通过在联邦学习过程中提供无价值或低价值的模型参数更新来获取高价值的全局模型，从而以不公平的方式获益. Fraboni 等人[183] 的研究表明直接上传原始模型参数的 free-rider 攻击和在原始模型基础上添加噪声的 free-rider 攻击 2 种方式都可能在模型收敛精度不受显著影响①的情况下对模型收敛速度带来较大的负面影响. Lin 等人[184] 将前后 2 轮训练过程中的模型权重做差得到模型的权重差，然后将权重差或加入高斯噪声的权重差上传给中心聚合服务器聚合，从而在没有任何训练数据的情况下成功实施 free-rider 攻击.

4.2 数据隐私相关攻击的防御方法

从 4.1 节对窃取数据隐私的攻击方式的介绍中可以看出，这些攻击方法依赖目标模型的可及性，一般需要能获取目标模型的模型参数或模型输出，降低目标模型的模型参数或模型输出的可及性能够降低攻击的效果[52, 185]. 因此，防御这些攻击的一个基本思路就是降低目标模型的可及性.

1）混淆（数据隐私）. 混淆可以增加攻击者获取真实模型参数和模型输出的难度. 添加噪声[186] 可以改变神经网络层的输出从而实现混淆，而神经元无效化[187] 可以随机选择神经网络层中一定比例的神经元并将其输出置为 0 来实现混淆[185, 187].

差分隐私可以用来度量神经网络添加噪声后对训练数据的隐私保护能力[185]. PPeFL[186] 使用差分隐私方法向联邦学习的训练数据和 AIoT 端侧设备的模型参数更新中加入噪声来防止训练数据的隐私信息泄露. FLScheduler[188] 通过联邦学习任务的调度优化来实现隐私预算内端侧设备训练数据的充分利用，从而在隐私资源有限的情况下提升任务完成数量和减少服务水平目标的违反率. 文献 [27] 将噪声加在 CNN 模型切片的激活函数上，然后使用 Rényi 差分隐私[189] 分析了协同推理过程的隐私预算. 文献 [187] 对比了添加噪声和神经元无效化的防御效果，其实验结果表明，对模型切片的输入层或中间层进行随机神经网络无效化能在取得对模型反演攻击同等隐私保护效力的情况下比添加噪声更好地保留模型精度. 混淆虽然能一定程度地防止训练数据的隐私泄露，但并不能提供完全的隐私安全保障，攻击者依然可能从梯度或推理中间数据中获取训练数据的隐私信息[119, 190].

2）同态加密. 同态加密[119] 可以防止攻击者获取目标模型的模型参数和神经网络层输出. 同态加密的基本原理是使加密过程满足对密文进行的计算操作的结果在解密后与在对应明文上进行相同计算操作的结果相同或基本相同[191]. 在联邦学习方面，文献 [119] 使用加性同态加密对参与者提交给中心聚合服务器的梯度进行加密，避免了中心聚合服务器在梯度聚合过程中获取参与者的隐私信息.BatchNet[192] 将一批经过量化的梯度编码为一个长整数且经过编码后的梯度满足加性同态加密的性质，显著降低了同态加密对联邦学习通信过程和训练过程带来的额外开销. 近年来，通过硬件加速来优化基于同态加密的联邦学习也成为一种新兴的解决方案[193-194]. 在协同推理方面，Gazelle[195] 基于单指令多数据流对同态加密的基本操作（如加法、标量乘法和排列等）进行加速，实现了用于同态加密场景下进行快速的矩阵向量乘和卷积的线性代数算子，以实现神经网络推理的安全协议. 同态加密能为模型参数和神经网络层输出提供强有力的加密保护，但同态加密的延迟开销较大，会降低联邦学习或协同推理的速度.

3）多方安全计算. 多方安全计算将一个隐私数据分为 n 份，每一份分给一个计算方进行计算，需要至少 $m(0<m\leq n)$ 个计算方的计算结果才能恢复正确计算结果，如果获取的计算结果数目少于 m 则为无意义的信息. 多方安全计算可以用于保护联邦学习场景中的模型参数和协同推理场景中的模型输出. 在联邦学习方面，文献 [196] 中车联网内若干个雾计算节点将量化后的本地模型梯度通过多方安全计算进行安全聚合，在诚实但好奇的共谋者数目小于门槛值时能防止梯度中包含的隐私信息泄露. 在协同推理方面，PRICURE[197] 基于秘密分享将模型参数分配给不同的工作者来为客户端执行协同推理任务. 但该研究依赖一个可信的聚合者来对不同工作者的推理结

① 攻击者保持模型收敛精度不受显著影响是为了获得有价值的全局模型参数同时防止被发现.

果进行聚合, 以得到最终的聚合结果. SecureTVM[51]基于 TVM 和 ABY 库为协同推理任务生成两方安全计算的高性能可执行代码. 多方安全计算能取得比同态加密更快的速度[198], 但它可能需要拜占庭参与者的数目小于一定门槛值.

4)可信执行环境(trusted executed environment, TEE). TEE 是一种通过软硬件强制的访问控制手段将代码、数据和堆栈等敏感数据进行保护并证明程序正确运行的技术[199-200]. TEE 可用于防止攻击者获取模型输入数据、模型参数和模型输出等隐私数据, 从而可用于联邦学习和协同推理 2 种场景的隐私保护. 在联邦学习方面, Kalapaaking 等人[201]提出了一种基于 TEE 的去中心化联邦学习框架, 该框架使用 TEE 保护模型聚合过程中模型参数包含的隐私信息和保证模型聚合结果的正确性. SecureFL[202]将部分模型切片的训练任务放在 ARM Trust Zone 中执行, 以分离式学习[54]的方式进行端侧设备的本地模型训练, 然后将经过加密的模型参数上传至服务器, 再在 SGX 中解密模型参数并进行参数聚合. 在协同推理方面, Li 等人[203]将经过加密的模型参数和输入数据读入 SGX 的安全内存空间, 然后在 SGX 中对模型参数和输入数据进行解密并进行推理, 推理结果经过加密后写回非安全内存. Tramèr 等人[199]通过将部分计算复杂性较高的任务从 TEE 中加密后外包给计算资源较为丰富的宿主机环境来加速推理过程. 虽然 TEE 能在一定程度上同时保障数据隐私和深度学习任务的执行效率, 但 TEE 需要额外的硬件支持且面临着侧信道攻击的威胁[204].

5)异常检测(数据隐私). 异常检测可以通过检查推理请求的输入数据或联邦学习参与者提交的参数来防御模型提取攻击或 free-rider 攻击. PRADA[205]通过检测连续接收的样本之间的距离是否偏离正常分布来判断样本是否来自模型提取攻击.DAGMM[206]使用自编码器对高维模型参数更新进行降维压缩, 然后使用高斯混合模型来估计降维后模型参数更新的能量, 高能量的模型参数更新可能来自 free-rider 攻击. RFFL[207]基于参数向量的余弦相似度来检测 free-rider 攻击. 根据检测结果, RFFL 建立了联邦学习参与者的声誉系统并将低声誉参与者从联邦学习过程中排除. 异常检测能一定程度防止模型提取攻击和 free-rider 攻击, 但异常检测也可能因为逃逸攻击而失效[205].

6)改变输出. 改变输出可以防止攻击者获取有价值的模型(或模型切片)输出, 从而可用于联邦学习的部署阶段和协同推理场景来防御模型提取攻击. Zhang 等人[208]通过在输入样本中增加噪声使模型输出拥有相似的置信度向量, 以防止模型输出泄露模型内部信息. Tan 等人[209]使用触发数据集训练深度学习模型来为模型打上模型水印, 通过检测一个模型对触发数据集的模型输出是否达到一定预测精度可以判定该模型是否通过模型提取攻击获取. 考虑到改变输出可能带来模型精度下降, 文献 [210] 首先通过异常检测来为输入样本进行分类, 对判定为可能来自模型提取攻击的输入样本进行模型输出扰动, 从而在防御模型提取攻击的同时减少输出扰动对模型精度的影响.

7)区块链. 区块链[211]是一种具有拜占庭容错和防篡改等特性的分布式账本, 为联邦学习参与者提供了一个可靠的去中心化合作平台. Lu 等人[68]将区块链用于记录车联网中车辆所提交模型参数质量的验证结果, 提供模型参数的质量越高的参与者会被分配更高的声誉值, 积累的声誉值越多的参与者在聚合过程中所占的权重越高, 从而防止参与者通过提交低质量的本地模型参数来进行 free-rider 攻击. Lyu 等人[147]使用区块链构建了一种用于联邦学习的去中心化梯度交易系统来防止 free-rider 攻击. 在该系统中, 区块链根据联邦学习参与者的历史表现记录了各参与者的信用度, 各参与者优先与信用度更高的参与者进行梯度交易, 从而激励各参与者保持良好的信用记录. 区块链需要多节点之间的共识过程来生成区块进行记录, 这带来了冗余存储和共识开销.

4.3 模型安全相关的攻击

模型安全相关的攻击主要有投毒攻击和逃逸攻击[212-213]两种类型, 攻击者的主要目的是让模型在特定任务上精度下降或失效. 其中, 投毒攻击会对模型参数产生影响, 一般在联邦学习的模型训练过程中发生; 逃逸攻击则主要对模型的输入样本产生影响, 主要在模型推理过程中发生, 可以发生在联邦学习的部署阶段和协同推理过程中.

1)投毒攻击. 在联邦学习中, 投毒攻击[214-215]通过提供受污染的训练样本或模型参数来降低模型在特定任务上的表现, 它可以分为数据投毒攻击[215]和模型投毒攻击[214]. 前者的攻击面[216]为训练数据, 后者的攻击面为模型参数. PoisonGAN[215]使用 GAN 生成与正常联邦学习参与者的训练数据在分布上一致的人造样本, 然后对特定种类的人造数据进行标签翻转来生成受污染的训练数据. 攻击者使用这些受污

染的训练数据进行本地模型训练，从而在本地模型参数中植入后门，使其对特定种类的输入样本进行错误分类. 相比于数据投毒攻击，模型投毒攻击更注重对全局模型的影响. Bagdasaryan 等人[214]在全局模型接近收敛时使用被植入后门的模型替换全局模型，使得后续训练过程对后门的影响较小.

2）逃逸攻击. 逃逸攻击通过生成对抗样本[212]来使模型失效. 在联邦学习中，逃逸攻击可以分为外部攻击和内部攻击 2 种类型[217]. 当外部攻击的发起者能够获取黑盒全局模型的输出时，可以直接生成对抗样本来实行逃逸攻击，这种攻击可以发生在联邦学习的部署阶段（如通过 MLaaS 向用户提供服务时）. Tang 等人[213]训练了一个自编码器来基于干净的样本生成能将其变为对抗样本的扰动向量. FEAT[218]首先按照高维离散输入特征向量中的特征对神经网络分类结果的影响来对特征进行排序，然后使用组合搜索在对分类结果有重要影响的较少代表特征中选择特征进行扰动，从而兼顾了攻击的效率和有效性. 如果外部攻击者无法直接获取黑盒全局模型的输出，则可以通过代理数据集训练一个与目标模型行为相似的代理模型，通过代理模型生成对抗样本来实施逃逸攻击[219]. 内部攻击者则可以直接根据白盒全局模型使用梯度下降生成对抗样本来实施逃逸攻击[217]. 在 AIoT 应用场景中，参与协同推理的端侧设备也可能获取模型的最终推理结果，因此它们也能生成对抗样本，从而成为逃逸攻击的潜在发起者.

4.4 模型安全相关攻击的防御方法

防御投毒攻击和逃逸攻击的基本方法都包括异常检测[220-223]，而防御逃逸攻击的方法还有对抗学习[217, 224]和混淆[47]等.

1）异常检测（模型安全）. 在投毒攻击的防御方面，FLAME[120]通过过滤方向偏离程度较大的权重向量、修剪过大的权重向量和向模型参数中加入适量噪声来降低后门植入的负面影响. BayBFed[225]根据参与者所提交模型参数的统计特征对模型参数进行聚类，再根据新到来的模型参数与各类参数的差异程度判断新到来的模型参数是否来自恶意参与者. FedRecover[226]首先通过异常检测排除攻击者，然后基于排除攻击者前良性参与者提交模型参数的历史信息重新估计排除攻击者后各参与者提交的模型参数来重新生成全局模型，从而将全局模型从投毒攻击中恢复，避免了从头开始重新训练全局模型带来的高计算和通信开销. 异常检测方法可以用于防御投毒攻击，但该方法依赖于白盒模型参数，而出于数据隐私保护的考虑，参与者提交的模型参数可能经过加密（如同态加密）. 对于加密后的模型参数如何进行异常检测仍然有待研究.

异常检测还可以用来防御逃逸攻击. Wen 等人[227]发现由于对抗样本往往靠近决策边界，其判别结果的置信度一般比同一类别的良性样本的置信度低. 基于这种差异，该研究训练了一个可以根据模型输出向量判断输入样本是否为良性样本的神经网络分类器. TAD[228]将多个神经网络逃逸攻击检测器的检测结果使用迁移学习进行融合来提升网络入侵检测的检测率.

2）对抗学习. 对抗学习可以通过在模型训练过程中额外加入对抗样本来提升模型对逃逸攻击的鲁棒性[217, 224]. FAT[224]在端侧设备的本地训练过程中加入对抗样本来提升联邦学习全局模型对逃逸攻击的鲁棒性，这种方法可以用于防御外部逃逸攻击，但对内部逃逸攻击的防御效果较差[217]. 为提升联邦学习对内部逃逸攻击的防御效果，文献 [217] 将对抗学习与个性化学习结合起来，使每个端侧设备中对抗样本占本地训练样本的比例与设备资源成正相关，然后基于个性化联邦学习进行模型训练，从而增加了内部攻击者获取其他端侧设备模型信息的难度，有效提升了联邦学习对内部逃逸攻击的防御效果.

3）混淆（模型安全）. 混淆机制也能用来防御逃逸攻击. 文献 [47] 提出了一种联邦分离式学习框架 SplitFed，该框架将全局模型切分为 2 部分分别部署在端侧设备和服务器上，通过在端侧设备和服务器之间传递模型中间层输出、样本标签和反向传播过程中的梯度来完成训练过程. SplitFed 基于 PixelDP[229]在端侧设备的模型切片中加入差分隐私噪声层来提升模型对逃逸攻击的鲁棒性. 实验结果表明，SplitFed 能有效减轻波动幅度在一定范围内的对抗样本带来的负面影响. 这种防御方法也有望用于提升协同推理对逃逸攻击的鲁棒性.

4.5 小 结

本节总结了面向 AIoT 的协同智能在隐私安全方面的相关工作，表 5 分别展示了联邦学习和协同推理在数据隐私和模型安全 2 方面的攻击和相关的防御方法. 其中，表 5 中的攻击面指的是攻击者在联邦学习或协同推理的过程中可以访问且用以发动攻击的漏洞[216]. 可以看到，联邦学习和协同推理在隐私安全方面拥有许多共同的攻击方法和防御机制，呈现出紧密的内在联系.

攻击面为模型梯度和模型参数更新的相关攻击

Table 5 Summary of the Attacks That the AIoT-Oriented Collaborative Intelligence Faces and the Corresponding Defense Methods

表 5 面向 AIoT 的协同智能面临的攻击和对应防御方法总结

攻击类型			攻击面	攻击场景	参考文献	防御机制
数据隐私	训练样本相关	模型反演攻击	模型参数、模型输出	联邦学习、协同推理	[52, 119, 174−175]	混淆[186-188]、同态加密[119, 192, 195]、多方安全计算[51, 197-198, 230-231]、可信执行环境[199, 201-203]
		成员推断攻击			[53, 176−177]	
		性质推断攻击	模型参数	联邦学习	[178−179]	
	模型参数相关	模型提取攻击	模型输出	联邦学习、协同推理	[52, 173−174, 181]	异常检测[205]、改变输出[208-210]
		free-rider 攻击	模型参数	联邦学习	[183−184]	异常检测[206-207]、区块链[68, 147]
模型安全	投毒攻击		训练数据、模型参数	联邦学习、协同推理	[214−215]	异常检测[120, 226]
	逃逸攻击		模型输出、模型参数		[213, 217−219]	异常检测[227-228]、对抗学习[224]、混淆[229]

都归为攻击面为模型参数的一类.

展望未来, 随着大模型逐渐在生产生活场景中落地, 与大模型相关的攻击将会更加频繁. 模型提取攻击在大模型时代的影响将会进一步增强, 这一方面是因为大模型的复杂度决定了很多基于大模型的服务需要通过 MLaaS 向用户提供; 另一方面是因为大模型训练所需的资源和强大的服务能力使其具有较高的经济价值[28]. He 等人[232]的研究表明, 基于 BERT 的自然语言大模型通过 MLaaS API 对外提供服务的场景下模型提取攻击能以较低成本获取精度较高的模型. Keskar 等人[233]的研究表明, 攻击者可以使用从单语言模型中提取的知识来微调多语言模型以提升其精度. 除了模型提取攻击, 对于部署在云服务器上的大模型可能还会遭遇拒绝服务攻击[28], 相关防御方法依然需要进一步研究.

5 总结与展望

本文探讨了面向 AIoT 的协同智能的概念, 围绕通过高效、安全的知识传递和算力供给来实现智能任务的多设备协同计算这一主要设计目标, 归纳了联邦学习和协同推理在算法、架构和隐私安全 3 个方面近十年来的技术进展, 揭示了两者在 AIoT 应用场景中的内在联系. 展望未来, 随着面向 AIoT 的协同智能的进一步发展, 联邦学习和协同推理将在 AIoT 应用场景下呈现进一步融合的趋势. 具体体现在 4 个方面:

1) 设备共用. 在 AIoT 应用场景中, 联邦学习和协同推理将共同使用端、边和云各层级设备的计算资源和通信资源, 在任务调度层面相互影响, 需要统筹考虑. 在真实场景中, 不同设备的联邦学习任务和协同推理任务在系统中可能同时存在, 从而带来复杂的动态任务调度场景. 如何在复杂动态任务调度场景中合理利用设备资源来满足不同任务的需求是一个富有挑战性和具有实用价值的问题. 排队理论[234]可以用于分析包含各种异构设备的复杂动态任务调度环境, 提供了解决问题的一个可能方向.

2) 模型共用. 由 AIoT 端侧设备收集的海量数据训练而成的模型将由协同推理通过各 AIoT 端侧设备为媒介向用户提供服务. 由于设备共用的存在, 服务于不同任务的模型可能同时存在于系统之中. 已有研究已经证明, 不同模态任务可以彼此促进[28]. 因此面向 AIoT 的协同智能为具有不同知识的模型相互传递知识, 为实现协同进化提供了可能.

3) 隐私安全机制协同. 在 AIoT 应用场景中, 联邦学习和协同推理面临相似的隐私安全威胁, 防御机理高度相关, 可以在防御方式和防御设备(如 TEE)上进行协同. 由于隐私安全机制可能造成任务执行效率的降低, 通过硬件加速等方式提高隐私安全机制工作的效率成为一种可行方案[193-194]. 设计者需要综合考虑联邦学习和协同推理对安全的需求, 以在安全保障、任务执行效率和成本之间实现平衡.

4) 激励机制协同. 在 AIoT 应用场景中, 参与联邦学习和协同推理的设备可能来自不同参与者, 如端侧设备可能来自个人用户而边缘服务器和云服务器则可能来自运营商. 设计合理的激励机制对在这些设备之间高效和公平地共享知识和算力 2 种资源具有重要意义. 目前激励机制相关研究主要集中在联邦学习领域[235], 协同推理相关的激励机制研究还比较少. 由于协同推理所使用的算力可能来自不止 1 个参与方, 因此也需要设计激励机制来根据各参与方的贡献给予合理的报酬.

随着 AIoT 设备的普及和 AIoT 应用场景的进一步丰富, 面向 AIoT 的协同智能将被更广泛地使用. 举例说明, 自动驾驶服务提供商可以通过联邦学习收集车辆行驶和用户操作信息来训练自动驾驶模型,

然后通过路边基站上部署的边缘服务器或车辆本地的深度学习加速器为用户提供高性能的协同推理服务以更好地实现智能汽车的自动驾驶[236]. 面向 AIoT 的协同智能还将为大模型的应用落地提供重要的途径, 例如智能助手服务商可以基于移动设备上的智能助手和用户的对话记录使用联邦学习来训练大语言模型, 然后将大模型(或部分大模型)部署在移动基站服务器上或压缩后部署在用户的移动设备上, 通过端边协同推理或本地协同推理为用户提供服务[237-238]. 可以预见, 面向 AIoT 的协同智能将对社会经济发展和人类生活的智能化产生更加深远的影响.

作者贡献声明: 罗宇哲负责收集材料和撰写论文; 李玲负责论文的整体架构规划; 侯朋朋、于佳耕、程丽敏、张常有、武延军和赵琛参与了论文的修改与讨论.

参 考 文 献

[1] Chang Zhuoqing, Liu Shubo, Xiong Xingxing, et al. A survey of recent advances in edge-computing-powered artificial intelligence of things[J]. IEEE Internet of Things Journal, 2021, 8(18): 13849−13875

[2] Wang Wenbo, Zhang Yingfeng, Gu Jinan, et al. A proactive manufacturing resources assignment method based on production performance prediction for the smart factory [J]. IEEE Transactions on Industrial Informatics, 18(1): 46−55

[3] Yu Liang, Xie Weiwei, Xie Di, et al. Deep reinforcement learning for smart home energy management[J]. IEEE Internet of Things Journal, 2020, 7(4): 2751−2762

[4] Shaikh F K, Karim S, Zeadally S, et al. Recent trends in Internet-of-things-enabled sensor technologies for smart agriculture[J]. IEEE Internet of Things Journal, 2022, 9(23): 23583−23598

[5] Zhao Jianxin, Chang Xinyu, Feng Yanhao, et al. Participant selection for federated learning with heterogeneous data in intelligent transport system[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 24(1): 1106−1115

[6] Analitics IoT. IoT 2020 in review: The 10 most relevant IoT developments of the year [EB/OL]. (2021-01-12)[2024-07-16]. https://iot-analytics.com/iot-2020-in-review/

[7] Analitics IoT. IoT 2021 in review: The 10 most relevant IoT developments of the year [EB/OL]. (2022-01-11)[2024-07-16]. https://iot-analytics.com/iot-2021-in-review/

[8] Zhang Yuqing, Zhou Wei, Peng Anni. Survey of Internet of things security[J]. Journal of Computer Research and Development, 2017, 54(10): 2130−2143(in Chinese)
(张玉清, 周威, 彭安妮. 物联网安全综述[J]. 计算机研究与发展, 2017, 54(10): 2130−2143)

[9] Dong Yudi, Yao Yudong. Secure mmwave-radar-based speaker verification for IoT smart home[J]. IEEE Internet of Things Journal, 2021, 8(5): 3500−3511

[10] Liu Yangyang, Chang Shuo, Wei Zhiqing, et al. Fusing mmwave radar with camera for 3-D detection in autonomous driving[J]. IEEE Internet of Things Journal, 2022, 9(20): 20408−20421

[11] Zhang Chaoyun, Patras P, Haddadi H, et al. Deep learning in mobile and wireless networking: A survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21((3):): 2224−2287

[12] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]// Proc of the 2016 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[13] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End-to-end speech recognition in English and Mandarin[C]// Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 173−182

[14] Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning for natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 604−624

[15] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning[C]// Proc of the 30th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2016: 2094−100

[16] Ren Weiqing, Qu Yuben, Dong Chao, et al. A survey on collaborative DNN inference for edge intelligence[J]. Machine Intelligence Research, 2023, 20(3): 370−395

[17] EU. Regulation (EU) 2016/679 of the European parliament and of the council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [EB/OL]. (2018-05-25) [2024-07-16]. https://gdpr-info.eu/

[18] Li Mu, Andersen D G, Park J W, et al. Scaling distributed machine learning with the parameter server[C]// Proc of the 11th USENIX Conf on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2014: 583–598

[19] Teerapittayanon S, Mcdanel B, Kung H T. Distributed deep neural networks over the cloud, the edge and end devices[C]// Proc of the 37th IEEE Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2017: 328−339

[20] Lim W Y B, Luong N C, Hoang D T, et al. Federated learning in mobile edge networks: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2019, 22: 2031−2063

[21] Zeng Liekang, Chen Xu, Zhou Zhi, et al. CoEdge: Cooperative DNN inference with adaptive workload partitioning over heterogeneous edge devices[J]. IEEE/ACM Transactions on Networking, 2021, 29(2): 595−608

[22] Yang Qiang, Liu Yang, Chen Tianjian, et al. Federated machine learning: Concept and applications [J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): Article 12

[23] Zhu Hongrui, Yuan Guojun, Yao Chengji, et al. Survey on network of distributed deep learning training[J]. Journal of Computer Research and Development, 2021, 58(1): 98−115 (in Chinese)

(朱泓睿, 元国军, 姚成吉, 等. 分布式深度学习训练网络综述[J]. 计算机研究与发展, 2021, 58(1): 98−115)

[24] Nguyen D C, Ding Ming, Pathirana P N, et al. Federated learning for internet of things: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1622−1658

[25] Khan L U, Saad W, Han Zhu, et al. Federated learning for Internet of things: Recent advances, taxonomy, and open challenges[J]. IEEE Communications Surveys & Tutorials, 2021, 23(3): 1759−1799

[26] Zhang Shuai, Zhang Sheng, Qian Zhuzhong, et al. DeepSlicing: Collaborative and adaptive CNN inference with low latency[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(9): 2175−2187

[27] Mao Yunlong, Hong Wenbo, Wang Heng, et al. Privacy-preserving computation offloading for parallel deep neural networks training[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(7): 1777−1788

[28] Bommasani R, Hudson D, Adeli E, et al. On the opportunities and risks of foundation models [J]. arXiv preprint, arXiv: 2108.07258, 2021

[29] Cao Yihan, Li Siyu, Liu Yixin, et al. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT [J]. arXiv preprint, arXiv: 2303.04226, 2023

[30] Zhou Zhi, Chen Xu, Li En, et al. Edge intelligence: Paving the last mile of artificial intelligence with edge computing[J]. Proceedings of the IEEE, 2019, 107: 1738−1762

[31] Chen Yunji, Li Ling, Li Wei et al. AI Computing System [M] Beijing: China Machine Press, 2020(in Chinese)
(陈云霁, 李玲, 李威, 等. 智能计算系统[M]. 北京: 机械工业出版社, 2020)

[32] Poirot M G, Vepakomma P, Chang Ken, et al. Split learning for collaborative deep learning in healthcare [J]. arXiv preprint, arXiv: 1912.12115, 2019

[33] Zhuang Fuzhen, Qi Zhiyuan, Duan Keyu, et al. A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2021, 109(1): 43−76

[34] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]// Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 1126–1135

[35] Yao Jiangchao, Wang Feng, Jia Kunyang, et al. Device-cloud collaborative learning for recommendation[C]// Proc of the 27th ACM SIGKDD Conf on Knowledge Discovery & Data Mining. New York: ACM, 2021: 3865−3874

[36] Chen Zeyuan, Yao Jiangchao, Wang Feng, et al. Mc^2-SF: Slow-fast learning for mobile-cloud collaborative recommendation [J]. arXiv preprint, arXiv: 2109.12314, 2021

[37] Yao Jiangchao, Zhang Shengyu, Yao Yang, et al. Edge-cloud polarization and collaboration: A comprehensive survey for AI[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 6866−6886

[38] Zhao Yuxi, Gong Xiaowen, Mao Shiwen. Truthful incentive mechanism for federated learning with crowdsourced data labeling[C]// Proc of the 2023 IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2023: 1−10

[39] Zhang Tuo, Feng Tiantian, Alam S, et al. GPT-FL: Generative pre-trained model-assisted federated learning [J]. arXiv preprint, arXiv: 2306.02210, 2023

[40] Guo Bin, Liu Sicong, Liu Yan, et al. AIoT: The concept, architecture and key techniques[J]. Chinese Journal of Computers, 2023, 46(11): 2259−2278 (in Chinese)
(郭斌, 刘思聪, 刘琰, 等. 智能物联网: 概念、体系架构与关键技术[J]. 计算机学报, 2023, 46(11): 2259−2278)

[41] Kang Yiping, Hauswald J, Gao Cao, et al. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge[C] //Proc of the 22nd Int Conf on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2016: 615−629

[42] Pan Jingyu, Chang C C, Xie Zhiyao, et al. Towards collaborative intelligence: Routability estimation based on decentralized private data[C] //Proc of the 59th ACM/IEEE Design Automation Conf. New York: ACM, 2017: 961−966

[43] Wang Rui, Qi Jianpeng, Chen Liang, et al. Survey of collaborative inference for edge intelligence[J]. Journal of Computer Research and Development, 2023, 60(2): 398−414 (in Chinese)
(王睿, 齐建鹏, 陈亮, 等. 面向边缘智能的协同推理综述[J]. 计算机研究与发展, 2023, 60(2): 398−414)

[44] Mcmahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Proc of the 20th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2017: 1273−1282

[45] Kairouz P, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundation Trends in Machine Learning, 2021, 14(1): 1−210

[46] Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. arXiv preprint, arXiv: 1503.02531, 2015

[47] Thapa C, Chamikara M A P, Camtepe S, et al. SplitFed: When federated learning meets split learning[C]// Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 8485−8493

[48] Lu Ying, Luo Lingkun, Huang Di, et al. Knowledge transfer in vision recognition: A survey [J]. ACM Computing Surveys, 2020, 53(2): Article 37

[49] Zhao Zhuoran, Barijough K M, Gerstlauer A. DeepThings: Distributed adaptive deep learning inference on resource-constrained IoT edge clusters[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(11): 2348−2359

[50] Tang Xin, Chen Xu, Zeng Liekang, et al. Joint multiuser dnn partitioning and computational resource allocation for collaborative edge intelligence[J]. IEEE Internet of Things Journal, 2021, 8(12): 9511−9522

[51] Huang P H, Tu C H, Chung S M, et al. SecureTVM: A TVM-based compiler framework for selective privacy-preserving neural inference[J]. ACM Transactions on Design Automation of

Electronic Systems, 2023, 28(4): 1−28

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

[53] Chen Hanxiao, Li Hongwei, Dong Guishan, et al. Practical membership inference attack against collaborative inference in industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2022, 18(1): 477−487

[54] Ayad A, Renner M, Schmeink A. Improving the communication and computation efficiency of split learning for IoT applications[C/OL]// Proc of the 2021 IEEE Global Communications Conf. Piscataway, NJ: IEEE, 2021[2024-08-17]. https://ieeexplore.ieee.org/document/9685493

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

[56] Zhao Yuchen, Barnaghi P, Haddadi H. Multimodal federated learning on IoT data[C]// Proc of the 7th IEEE/ACM Int Conf on Internet-of-Things Design and Implementation. Piscataway, NJ: IEEE, 2022: 43−54

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

[58] Zhang Zhouyangzi, Guo Bin, Sun Wen, et al. Cross-FCL: Toward a cross-edge federated continual learning framework in mobile edge computing systems[J]. IEEE Transactions on Mobile Computing, 2022, 23(1): 313−326

[59] Zhuo H H, Feng Wenfeng, Lin Yufeng, et al. Federated deep reinforcement learning [J]. arXiv preprint, arXiv: 1901.08277, 2019

[60] Kingma D P, Ba J. Adam: A method for stochastic optimization [C/OL]// Proc of the 3rd Int Conf on Learning Representations. Washington: ICLR, 2015[2024-08-16]. https://www.semanticscholar.org/reader/a6cb366736791bcccc5c8639de5a8f9636bf87e8

[61] Zhang Jianyi, Li Ang, Tang Minxue, et al. Fed-CBS: A heterogeneity-aware client sampling mechanism for federated learning via class-imbalance reduction[C]// Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: Article 1734

[62] Duan Moming, Liu Duo, Chen Xianzhang, et al. Self-balancing federated learning with global imbalanced data in mobile systems[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 32(1): 59−71

[63] Li Tian, Sahu A K, Zaheer M, et al. Federated optimization in heterogeneous networks[C/OL]// Proc of the 3rd Conf on Machine Learning and Systems. Indio, CA: MLSys. org, 2020[2024-08-16]. https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html

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

[65] Arivazhagan M G, Aggarwal V, Singh A K, et al. Federated learning with personalization layers [J]. arXiv preprint, arXiv: 1912.00818, 2019

[66] Li Tian, Hu Shengyuan, Beirami A, et al. Ditto: Fair and robust federated learning through personalization[C]// Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 6357−6368

[67] Xie Cong, Koyejo O, Gupta I. Asynchronous federated optimization [J]. arXiv preprint, arXiv: 1903.03934, 2019

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

[69] Sun Yuchang, Shao Jiawei, Mao Yuyi, et al. Semi-decentralized federated edge learning with data and device heterogeneity[J]. IEEE Transactions on Network and Service Management, 2023, 20(2): 1487−1501

[70] Zhang Feilong, Liu Xianming, Lin Shiyi, et al. No one idles: Efficient heterogeneous federated learning with parallel edge and server computation[C]// Proc of the 40th Int Conf on Machine Learning. New York: ACM, 2023: 41399−41413

[71] Diao Enmao, Ding Jie, Tarokh V. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients[C] // Proc of the 2021 Int Conf on Learning Representations. Washington: ICLR, 2021: 1−24

[72] Alam S, Liu Luyang, Yan Ming, et al. FedRolex: Model-heterogeneous federated learning with rolling sub-model extraction[C] // Proc of the 36th Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 29677−29690

[73] He Chaoyang, Annavaram M, Avestimehr S. Group knowledge transfer: Federated learning of large CNNs at the edge[C]// Proc of the 34th Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2020: Article 1180

[74] 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

[75] Lin Tao, Kong Lingjing, Stich S U, et al. Ensemble distillation for robust model fusion in federated learning[C]// Proc of the 34th Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2020: Article 198

[76] Lin Yiming, Gao Yuan, Gong Maoguo, et al. Federated learning on multimodal data: A comprehensive survey[J]. Machine Intelligence Research, 2023, 20(4): 539−553

[77] Xiong Baochen, Yang Xiaoshan, Qi Fan, et al. A unified framework for multi-modal federated learning [J]. Neurocomputing, 2022, 480: 110−118

[78] Lu Jiasen, Yang Jianwei, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]// Proc of the 30th Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2016: 289−297

[79] Liu Fenglin, Wu Xian, Ge Shen, et al. Federated learning for vision-and-language grounding problems[C]// Proc of the 34th AAAI Conf

on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11572−11579

[80] Chen Jiayi, Zhang Aidong. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks[C]// Proc of the 28th ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2022: 87–96

[81] Zhang Rongyu, Chi Xiaowei, Liu Guiliang, et al. Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation [J]. arXiv preprint, arXiv: 2303.15486, 2023

[82] Liu Boyi, Wang Lujia, Liu Ming. Lifelong federated reinforcement learning: A learning architecture for navigation in cloud robotic systems[J]. IEEE Robotics and Automation Letters, 2019, 4(4): 4555−4562

[83] Jiang Ziyue, Ren Yi, Lei Ming, et al. FedSpeech: Federated text-to-speech with continual learning[C]// Proc of the 30th Int Joint Conf on Artifical Intelligence. Berlin: Springer, 2021: 3829−3835

[84] Hung S C Y, Tu Chenghao, Wu Chengen, et al. Compacting, picking and growing for unforgetting continual learning[C]// Proc of the 33rd Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2019: Article 1225

[85] Usmanova A, Portet F, Lalanda P, et al. Federated continual learning through distillation in pervasive computing[C]// Proc of the 2022 IEEE Int Conf on Smart Computing. Piscataway, NJ: IEEE, 2022: 86−91

[86] Yoon J H, Jeong W Y, Lee G W, et al. Federated continual learning with weighted inter-client transfer[C]// Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 12073−12086

[87] Mori J, Teranishi I, Furukawa R. Continual horizontal federated learning for heterogeneous data[C/OL]// Proc of the 2022 Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2022[2024-08-16].https://www.semanticscholar.org/reader/3674cbf1900f748e5d1e981f296790256989a62e

[88] Hendryx S M, Kc D R, Walls B, et al. Federated reconnaissance: Efficient, distributed, class-incremental learning [J]. arXiv preprint, arXiv: 2109.00150, 2021

[89] Xu Chencheng, Hong Zhiwei, Huang Minlie, et al. Acceleration of federated learning with alleviated forgetting in local training [C/OL]// Proc of the 10th Int Conf on Learning Representations. Washington:ICLR,2022[2024-07-30].https://openreview.net/pdf?id=541PxiEKN3F

[90] Dong Jiahua, Wang Lixu, Fang Zhen, et al. Federated class-incremental learning[C]// Proc of the 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10154−10163

[91] Wang Tianyu, Liang Teng, Li Jun, et al. Adaptive traffic signal control using distributed MARL and federated learning[C]// Proc of the 20th IEEE Int Conf on Communication Technology. Piscataway, NJ: IEEE, 2020: 1242−1248

[92] Liu Haotian, Wu Wenchuan. Federated reinforcement learning for decentralized voltage control in distribution networks[J]. IEEE Transactions on Smart Grid, 2022, 13(5): 3840−3843

[93] Rezazadeh F, Bartzoudis N. A federated DRL approach for smart micro-grid energy control with distributed energy resources[C]// Proc of the 27th IEEE Int Workshop on Computer Aided Modeling and Design of Communication Links and Networks. Piscataway, NJ: IEEE, 2022: 108−114

[94] Wang Xiaofei, Wang Chenyang, Li Xiuhua, et al. Federated deep reinforcement learning for Internet of things with decentralized cooperative edge caching[J]. IEEE Internet of Things Journal, 2020, 7(10): 9441−9455

[95] Yu Shuai, Chen Xu, Zhou Zhi, et al. When deep reinforcement learning meets federated learning: Intelligent multitimescale resource management for multiaccess edge computing in 5G ultradense network[J]. IEEE Internet of Things Journal, 2021, 8(4): 2238−2251

[96] Wang Xiaoding, Hu Jia, Lin Hui, et al. QoS and privacy-aware routing for 5G-enabled industrial Internet of things: A federated reinforcement learning approach[J]. IEEE Transactions on Industrial Informatics, 2022, 18(6): 4189−4197

[97] Jeong H J, Lee H J, Shin C H, et al. IONN: Incremental offloading of neural network computations from mobile devices to edge servers[C] //Proc of the 2018 ACM Symp on Cloud Computing. New York: ACM, 2018: 401−411

[98] Zhou Li, Samavatian M H, Bacha A, et al. Adaptive parallel execution of deep neural networks on heterogeneous edge devices[C] // Proc of the 4th ACM/IEEE Symp on Edge Computing. New York: ACM, 2019: 195−208

[99] Yang Xiang, Xu Zikang, Qi Qi, et al. PICO: Pipeline inference framework for versatile CNNs on diverse mobile devices[J]. IEEE Transactions on Mobile Computing, 2023, 23(4): 2712−2730

[100] Hu Chenghao, Li Baochun. Distributed inference with deep learning models across heterogeneous edge devices[C]// Proc of the 2022 IEEE Conf on Computer Communications. Piscataway, NJ: IEEE, 2022: 330−339

[101] Kim Y G, Wu C J. AutoScale: Energy efficiency optimization for stochastic edge inference using reinforcement learning[C]// Proc of the 53rd Annual IEEE/ACM Int Symp on Microarchitecture. Piscataway, NJ: IEEE, 2020: 1082−1096

[102] Hou Xueyu, Guan Yongjie, Han Tao, et al. DistrEdge: Speeding up convolutional neural network inference on distributed edge devices[C]// Proc of the 2022 IEEE Int Parallel and Distributed Processing Symp. Piscataway, NJ: IEEE, 2022: 1097−1107

[103] Fu Kaihua, Shi Jiuchen, Chen Quan, et al. QoS-aware irregular collaborative inference for improving throughput of DNN services [C/OL] //Proc of the 2022 Int Conf for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ: IEEE, 2022[2024-08-16]. https://dl.acm.org/doi/10.5555/3571885.3571976

[104] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[C/OL]// Proc of the 4th Int Conf on Learning Representations. Washington: ICLR, 2016[2024-07-30]. https://www.semanticscholar.org/reader/024006d4c2a89f7acacc6e4438d156525b60a98f

[105] Wang Lingdong, Xiang Liyao, Xu Jiayu, et al. Context-aware deep

model compression for edge cloud computing[C]// Proc of the 40th IEEE Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2020: 787−797

[106] Molina M, Muñoz O, Pascual-Iserte A, et al. Joint scheduling of communication and computation resources in multiuser wireless application offloading[C]// Proc of the 25th IEEE Annual Int Symp on Personal, Indoor, and Mobile Radio Communication. Piscataway, NJ: IEEE, 2014: 1093−1098

[107] Kang Yiping, Hauswald J, Gao Cao, et al. Neurosurgeon: Collaborative intelligence between the cloud and mobile edge[C] // Proc of the 22nd Int Conf on Architectural Support for Programming Languages and Operating Systems. New York: ACM, 2017: 615−629

[108] Zhuang Weiming, Chen Chen, Lyu Lingjuan. When foundation model meets federated learning: Motivations, challenges, and future directions [J]. arXiv preprint, arXiv: 2306.15546, 2023

[109] Li Mingzhen, Liu Yi, Liu Xiaoyan, et al. The deep learning compiler: A comprehensive survey[J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(3): 708−727

[110] Zeng Qunsong, Du Yuqing, Huang Kaibin, et al. Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing[J]. IEEE Transactions on Wireless Communications, 2021, 20(12): 7947−7962

[111] Han M, Hyun J, Park S, et al. MOSAIC: Heterogeneity-, communication-, and constraint-aware model slicing and execution for accurate and efficient inference[C]// Proc of the 28th Int Conf on Parallel Architectures and Compilation Techniques. Piscataway, NJ: IEEE, 2019: 165−177

[112] Chen Yunji, Luo Tao, Liu Shaoli, et al. DaDianNao: A machine-learning supercomputer[C]// Proc of the 47th Annual IEEE/ACM Int Symp on Microarchitecture. Piscataway, NJ: IEEE, 2014: 609−622

[113] Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit[C/OL] // Proc of the 44th ACM/IEEE Annual Int Symp on Computer Architecture. New York: ACM, 2017[2024-08-16]. https://dl.acm.org/doi/10.1145/3079856.3080246

[114] Ro J H, Suresh A T, Wu Ke. FedJAX: Federated learning simulation with JAX [J]. arXiv preprint, arXiv: 2108.02117, 2021

[115] Lee J Y, Park W P, Mitchell N, et al. JaxPruner: A concise library for sparsity research [J]. arXiv preprint, arXiv: 2304.14082, 2023

[116] Villarrubia J, Costero L, Igual F D, et al. Improving inference time in multi-TPU systems with profiled model segmentation[C]// Proc of the 31st Euromicro Int Conf on Parallel, Distributed and Network-Based Processing. Piscataway, NJ: IEEE, 2023: 84−91

[117] Wang Zixiao, Che Biyao, Guo Liang, et al. PipeFL: Hardware/software co-design of an FPGA accelerator for federated learning[J]. IEEE Access, 2022, 10: 98649−98661

[118] Li H M, Rieger P, Zeitouni S, et al. FLAIRS: FPGA-accelerated inference-resistant & secure federated learning [J]. arXiv preprint, arXiv: 2308.00553, 2023

[119] Phong L T, Aono Y, Hayashi T, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(5): 1333−1345

[120] Nguyen T D, Rieger P, Chen Huili, et al. FLAME: Taming backdoors in federated learning[C]// Proc of the 31st USENIX Security Symp USENIX Association. Berkeley, CA, 2022: 1415−1432

[121] Bao Yungang, Chang Yisong, Han Yinhe, et al. Agile design of processor chips: Issues and challenges[J]. Journal of Computer Research and Development, 2021, 58(6): 1131−1145 (in Chinese)
(包云岗, 常轶松, 韩银和, 等. 处理器芯片敏捷设计方法: 问题与挑战[J]. 计算机研究与发展, 2021, 58(6): 1131−1145)

[122] Wang Kaifan, Xu Yinan, Yu Zihao, et al. XiangShan open-source high performance RISC-V processor design and implementation[J]. Journal of Computer Research and Development, 2023, 60(3): 476−493 (in Chinese)
(王凯帆, 徐易难, 余子濠, 等. 香山开源高性能 RISC-V 处理器设计与实现[J]. 计算机研究与发展, 2023, 60(3): 476−493)

[123] Dhilleswararao P, Boppu S, Manikandan M S, et al. Efficient hardware architectures for accelerating deep neural networks: Survey[J]. IEEE Access, 2022, 10: 131788−131828

[124] Zhao Yongwei, Du Zidong, Guo Qi, et al. Cambricon-F: Machine learning computers with fractal von neumann architecture[C]// Proc of the 46th ACM/IEEE Annual Int Symp on Computer Architecture. Piscataway, NJ: IEEE, 2019: 788−801

[125] Chen Tianqi, Moreau T, Jiang Ziheng, et al. TVM: An automated end-to-end optimizing compiler for deep learning[C]// Proc of the 13th USENIX Conf on Operating Systems Design and Implementation. Berkeley, CA: USENIX Association, 2018: 579–594

[126] Pytorch. Pytorch on XLA devices [EB/OL]. (2024-07-17)[2023-10-21]. https://pytorch.org/xla/master/

[127] Pytorch. Aot autograd —How to use and optimize?[EB/OL]. (2023-10-25)[2024-07-17]. https://pytorch.org/functorch/stable/notebooks/aot_autograd_optimizations.html

[128] Coral. Edge TPU compiler [EB/OL]. (2020-05-15)[2024-07-16]. https://coral.ai/docs/edgetpu/compiler/#help

[129] Nvidia. Optimizing inference on large language models with NVIDIA tensorrt-LLM, now publicly available [EB/OL]. (2023-10-19)[2024-07-16]. https://developer.nvidia.com/blog/optimizing-inference-on-llms-with-tensorrt-llm-now-publicly-available/

[130] NVIDIA. TensorRT-LLM [EB/OL]. (2023-10-24)[2024-07-17]. https://github.com/NVIDIA/TensorRT-LLM

[131] ONNX. ONNX [EB/OL]. (2024-05-24)[2024-07-17]. https://onnx.ai/

[132] MLIR. Multi-level intermediate representation overview [EB/OL]. (2017-07-17)[2024-07-17]. https://mlir.llvm.org/

[133] Jin Tian, Bercea G T, Tung D L, et al. Compiling ONNX neural network models using MLIR [J]. arXiv preprint, arXiv: 2008.08272, 2020

[134] Gao Liang, Li Li, Chen Yingwen, et al. FIFL: A fair incentive mechanism for federated learning[C]// Proc of the 50th Int Conf on Parallel Processing. New York: ACM, 2021: Article 82

[135] Team Tensorflow Lite. On-device training in tensorflow lite [EB/OL]. (2021-11-09)[2024-07-17]. https://blog.tensorflow.org/2021/11/on-device-training-in-tensorflow-lite.html

[136] Space Ts2. A comprehensive guide to tensorflow lite's federated learning [EB/OL]. (2023-04-07)[2024-07-17]. https://ts2.space/en/a-comprehensive-guide-to-tensorflow-lites-federated-learning/

[137] He Chaoyang, Li Songze, So Jinhyun, et al. FedML: A research library and benchmark for federated machine learning [J]. arXiv preprint, arXiv: 2007.13518, 2020

[138] Beutel D J, Topal T, Mathur A, et al. Flower: A friendly federated learning research framework [J]. arXiv preprint, arXiv: 2007.14390, 2020

[139] Jeong E J, Kim J R, Ha S H. TensorRT-based framework and optimization methodology for deep learning inference on jetson boards [J]. ACM Transactions on Embedded Computing Systems, 2022, 21(5): Article 51

[140] Jiang Xiaotang, Wang Huan, Chen Yiliu, et al. MNN: A universal and efficient inference engine[C/OL]// Proc of the 3rd Conf on Machine Learning and Systems. Austin, Texas: MLSys. org, 2020 [2024-07-30]. https://proceedings.mlsys.org/paper_files/paper/2020/file/bc19061f88f16e9ed4a18f0bbd47048a-Paper.pdf

[141] Lv Chengfei, Niu Chaoyue, Gu Renjie, et al. Walle: An end-to-end, general-purpose, and large-scale production system for device-cloud collaborative machine learning[C/OL]// Proc of the 16th USENIX Symp on Operating Systems Design and Implementation. Berkeley, CA:USENIXAssociation,2022[2024-08-16].https://www.usenix.org/conference/osdi22/presentation/lv

[142] Aminabadi R Y, Rajbhandari S, Awan A A, et al. DeepSpeed-inference: Enabling efficient inference of transformer models at unprecedented scale[C/OL]// Proc of the 2022 Int Conf for High Performance Computing, Networking, Storage and Analysis. Piscataway, NJ: IEEE, 2022[2024-08-16]. https://dl.acm.org/doi/abs/10.5555/3571885.3571946

[143] Liu Lumin, Zhang Jun, Song S H, et al. Client-edge-cloud hierarchical federated learning[C/OL]// Proc of the 2020 IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2020[2024-08-17]. https://ieeexplore.ieee.org/document/9148862

[144] Yang Shusen, Zhang Zhanhua, Zhao Cong, et al. CNNPC: End-edge-cloud collaborative cnn inference with joint model partition and compression[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(12): 4039−4056

[145] Korkmaz C, Kocas H E, Uysal A, et al. Chain FL: Decentralized federated machine learning via blockchain[C]// Proc of the 2nd Int Conf on Blockchain Computing and Applications. Piscataway, NJ: IEEE, 2020: 140−146

[146] Du Jiangsu, Shen Minghua, Du Yunfei. A distributed in-situ CNN inference system for IoT applications[C]// Proc of the 38th Int Conf on Computer Design. Piscataway, NJ: IEEE, 2020: 279−287

[147] Lyu L, Yu Jiangshan, Nandakumar K, et al. Towards fair and privacy-preserving federated deep models[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(11): 2524−2541

[148] Luping Wang, Wei Wang, Li Bo. CMFL: Mitigating communication overhead for federated learning[C]// Proc of the 39th IEEE Int Conf on Distributed Computing Systems. Piscataway, NJ: IEEE, 2019: 954−964

[149] Yu Hao, Yang Sen, Zhu Shenghuo. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 5693−5700

[150] You Chaoqun, Guo Kun, Feng Gang, et al. Automated federated learning in mobile-edge networks—Fast adaptation and convergence[J]. IEEE Internet of Things Journal, 2023, 10(15): 13571−13586

[151] Heinbaugh C E, Luz-Ricca E, Shao Huajie. Data-free one-shot federated learning under very high statistical heterogeneity[C/OL] //Proc of the 11th Int Conf on Learning Representations. Washington: ICLR, 2023[2024-07-30]. https://openreview.net/forum?id=_hb4vM3jspB

[152] Sattler F, Wiedemann S, Müller K R, et al. Robust and communication-efficient federated learning from non-I. I. D. data[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3400−3413

[153] Gao Hongchang, Xu An, Huang Heng. On the convergence of communication-efficient local SGD for federated learning[C]// Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 7510−7518

[154] Hönig R, Zhao Yiren, Mullins R D. DAdaQuant: Doubly-adaptive quantization for communication-efficient federated learning[C] //Proc of the 39th Int Conf on Machine Learning. New York: ACM, 2022: 8852−8866

[155] Nguyen M D, Lee S M, Pham Q V, et al. HCFL: A high compression approach for communication-efficient federated learning in very large scale IoT networks[J]. IEEE Transactions on Mobile Computing, 2023, 22(11): 6495−6507

[156] Dai Rong, Shen Li, He Fengxiang, et al. DisPFL: Towards communication-efficient personalized federated learning via decentralized sparse training[C]// Proc of the 39th Int Conf on Machine Learning. New York: ACM, 2022: 4587−4604

[157] Wen Hui, Wu Yue, Li Jingjing, et al. Communication-efficient federated data augmentation on non-IID data[C]// Proc of the 2022 IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops Piscataway, NJ: IEEE, 2022: 3376−3385

[158] Zhang Zhanhua, Yang Shusen, Zhao Cong, et al. RtCoInfer: Real-time collaborative CNN inference for stream analytics on ubiquitous images[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(4): 1212−1226

[159] Chen Xu, Jiao Lei, Li Wenzhong, et al. Efficient multi-user computation offloading for mobile-edge cloud computing[J]. IEEE/ACM Transactions on Networking, 2016, 24(5): 2795−2808

[160] Yi Changyan, Cai Jun, Su Zhou. A multi-user mobile computation offloading and transmission scheduling mechanism for delay-sensitive applications[J]. IEEE Transactions on Mobile Computing,

2020, 19(1): 29−43

[161] Ale L H, Zhang Ning, Fang Xiaojie, et al. Delay-aware and energy-efficient computation offloading in mobile-edge computing using deep reinforcement learning[J]. IEEE Transactions on Cognitive Communications and Networking, 2021, 7(3): 881−892

[162] Mozaffariahrar E, Theoleyre F, Menth M. A survey of wi-fi 6: Technologies, advances, and challenges[J]. Future Internet, 2022, 14(10): 293−345

[163] Das A K, Roy S, Bandara E, et al. Securing age-of-information (AoI)-enabled 5G smart warehouse using access control scheme[J]. IEEE Internet of Things Journal, 2023, 10(2): 1358−1375

[164] Mehr H D, Polat H. Human activity recognition in smart home with deep learning approach[C]// Proc of the 7th Int Istanbul Smart Grids and Cities Congress and Fair. Piscataway, NJ: IEEE, 2019: 149−153

[165] Qi Lianyong, Hu Chunhua, Hu Chunhua, et al. Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment[J]. IEEE Transactions on Industrial Informatics, 2021, 17(6): 4159−4167

[166] Chen Yiqiang, Wang Jindong, Yu Chaohui, et al. FedHealth: A federated transfer learning framework for wearable healthcare[J]. IEEE Intelligent Systems, 2020, 35(4): 83−93

[167] Lee S, Choi D H. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources[J]. IEEE Transactions on Industrial Informatics, 2022, 18(1): 488−497

[168] Wang Shuai, Li Dan. Research progress on network performance optimization of distributed machine learning system[J]. Chinese Journal of Computers, 2022, 45(7): 1384−1412 (in Chinese)
(王帅, 李丹. 分布式机器学习系统网络性能优化研究进展[J]. 计算机学报, 2022, 45(7): 1384−1412)

[169] Martinez I, Hafid A S, Jarray A. Design, resource management, and evaluation of fog computing systems: A survey[J]. IEEE Internet of Things Journal, 2021, 8(4): 2494−2516

[170] Lu Yan, Shu Yuanchao, Tan Xu, et al. Collaborative learning between cloud and end devices: An empirical study on location prediction[C]// Proc of the 4th ACM/IEEE Symp on Edge Computing. New York: ACM, 2019: 139–151

[171] Encora. Ahead-of-time compilation vs just-in-time compilation—Part 1 of understanding angular [EB/OL]. (2024-07-14)[2024-07-16]. https://www.encora.com/insights/ahead-of-time-compilation-vs-just-in-time-compilation-part-1

[172] Yu Hao, Yang Sen, Zhu Shenghuo. Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning[C]// Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: Article 698

[173] Tramer F, Zhang Fan, Juels A, et al. Stealing machine learning models via prediction APIs[C]// Proc of the 25th USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2016: 601–618

[174] Yin Yupeng, Zhang Xianglong, Zhang Huanle, et al. Ginver: Generative model inversion attacks against collaborative inference[C]// Proc of the 2023 ACM Web Conf. New York: ACM, 2023: 2122–2131

[175] Jin Xiao, Chen Pinyu, Hsu Chiayi, et al. CAFE: Catastrophic data leakage in vertical federated learning[C/OL] //Proc of the 35th Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2021[2024-07-30]. https://papers.nips.cc/paper/2021/hash/08040837089cdf46631a10aca5258e16-Abstract.html

[176] Nguyen T D T, Lai P, Tran K, et al. Active membership inference attack under local differential privacy in federated learning[C]// Proc of the 26th Int Conf on Artificial Intelligence and Statistics. Piscataway, NJ: IEEE, 2023: 5714−5730

[177] Li Jiacheng, Li Ninghui, Ribeiro B. Effective passive membership inference attacks in federated learning against overparameterized models[C/OL] //Proc of the 11th Int Conf on Learning Representation. Washington: ICLR, 2023[2024-07-30]. https://openreview.net/pdf?id=QsCSLPP55Ku

[178] Melis L, Song C, Cristofaro E D, et al. Exploiting unintended feature leakage in collaborative learning[C]// Proc of the 40th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 691−706

[179] Wang Zhibo, Huang Yuting, Song Mengkai, et al. Poisoning-assisted property inference attack against federated learning[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 20(4): 3328−3340

[180] Kourtellis N, Katevas K, Perino D. FLaaS: Federated learning as a service[C] //Proc of the 1st Workshop on Distributed Machine Learning. New York: ACM, 2020: 7−13

[181] Truong J B, Maini P, Walls R J, et al. Data-free model extraction[C]// Proc of the 2021 IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 4769−4778

[182] Liu Sijia, Chen Pinyu, Kailkhura B, et al. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications[J]. IEEE Signal Processing Magazine, 2020, 37(5): 43−54

[183] Fraboni Y, Vidal R, Lorenzi M. Free-rider attacks on model aggregation in federated learning[C]// Proc of the 24th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2021: 1846−1854

[184] Lin Jierui, Du Min, Liu Jian. Free-riders in federated learning: Attacks and defenses [J]. arXiv preprint, arXiv: 1911.12560, 2019

[185] Abadi M, Chu Andy, Goodfellow I J, et al. Deep learning with differential privacy[C]// Proc of the 2016 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2016: 308–318

[186] Wang Baocang, Chen Yange, Jiang Hang, et al. PPeFL: Privacy-preserving edge federated learning with local differential privacy[J]. IEEE Internet of Things Journal, 2023, 10(17): 15488−15500

[187] He Zecheng, Zhang Tianwei, Lee R. B. Attacking and protecting data privacy in edge–cloud collaborative inference systems[J]. IEEE Internet of Things Journal, 2021, 8(12): 9706−9716

[188] Jiang Bin, Li Jianqiang, Wang Huihui, et al. Privacy-preserving federated learning for industrial edge computing via hybrid

differential privacy and adaptive compression[J]. IEEE Transactions on Industrial Informatics, 2023, 19(2): 1136−1144

[189] Mironov I. Rényi differential privacy [C]//Proc of the 30th IEEE Computer Security Foundations Symp. Piscataway, NJ: IEEE 2017: 263−275

[190] Ryu Jihyeon, Zheng Yifeng, Gao Yansong, et al. Can differential privacy practically protect collaborative deep learning inference for IoT?[J/OL]. Wireless Networks, 2022[2024-07-30]. https://link.springer.com/article/10.1007/s11276-022-03113-7

[191] Cheon J H, Kim A, Kim M, et al. Homomorphic encryption for arithmetic of approximate numbers[C/OL]// Proc of the 2017 Int Conf on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2017 [2024-07-30]. https://link.springer.com/chapter/10.1007/978-3-319-70694-8_15

[192] Zhang Chengliang, Li Suyi, Xia Junzhe, et al. BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning[C]// Proc of the 2020 USENIX Annual Technical Conf. Berkeley, CA: USENIX Association, 2020: 493−506

[193] Zhu Yilan, Wang Xinyao, Ju Lei, et al. FxHENN: FPGA-based acceleration framework for homomorphic encrypted CNN inference[C]// Proc of the 29th IEEE Int Symp on High-Performance Computer Architecture. Piscataway, NJ: IEEE, 2023: 896−907

[194] Yang Zhaoxiong, Hu Shuihai, Chen Kai. FPGA-based hardware accelerator of homomorphic encryption for efficient federated learning [J]. arXiv preprint, arXiv: 2007.10560, 2020

[195] Juvekar C, Vaikuntanathan V, Chandrakasan A P. Gazelle: A low latency framework for secure neural network inference[C]// Proc of the 27th USENIX Security Symp. Berkeley, CA: USENIX Association, 2018: 1651−1668

[196] Li Yiran, Li Hongwei, Xu Guowen, et al. Practical privacy-preserving federated learning in vehicular fog computing[J]. IEEE Transactions on Vehicular Technology, 2022, 71(5): 4692−705

[197] Jarin I, Eshete B. PRICURE: Privacy-preserving collaborative inference in a multi-party setting[C]// Proc of the 2021 ACM Workshop on Security and Privacy Analytics. New York: ACM, 2021: 25–35

[198] Liu Yang, Kang Yan, Xing Chaoping, et al. A secure federated transfer learning framework[J]. IEEE Intelligent Systems, 2020, 35(4): 70−82

[199] Tramèr F, Boneh D. Slalom: Fast, verifiable and private execution of neural networks in trusted hardware [C/OL]//Proc of the 7th Int Conf on Learning Representations. Washington: ICLR, 2019 [2024-07-30]. https://openreview.net/pdf?id=rJVorjCcKQ

[200] Intel. Innovative technology for CPU based attestation and sealing [EB/OL].(2013-08-14)[2024-07-17].https://www.intel.com/content/www/us/en/developer/articles/technical/innovative-technology-for-cpu-based-attestation-and-sealing.html

[201] Kalapaaking A P, Khalil I, Rahman M S, et al. Blockchain-based federated learning with secure aggregation in trusted execution environment for Internet-of-things[J]. IEEE Transactions on Industrial Informatics, 2023, 19(2): 1703−1714

[202] Kuznetsov E, Chen Yitao, Zhao Ming. SecureFL: Privacy preserving federated learning with SGX and trustzone [C]//Proc of the 2021 IEEE/ACM Symp on Edge Computing. Piscataway, NJ: IEEE, 2021: 55−67

[203] Li Yuepeng, Zeng Deze, Gu Lin, et al. Efficient and secure deep learning inference in trusted processor enabled edge clouds[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(12): 4311−4325

[204] Law A, Leung C, Poddar R, et al. Secure collaborative training and inference for xgboost[C]// Proc of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice. New York: ACM, 2020: 21–26

[205] Juuti M, Szyller S, Marchal S, et al. PRADA: Protecting against dnn model stealing attacks[C]// Proc of the 2019 IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 512−527

[206] Lin Jierui, Du Min, Liu Jian. Free-riders in federated learning: Attacks and defenses [J]. arXiv prerprint, arXiv: 1911.12560, 2019

[207] Xu Xinyi, Lyu Lingjuan. A reputation mechanism is all you need: Collaborative fairness and adversarial robustness in federated learning [C/OL]// Proc of the 2021 Int Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML. New York: ACM, 2020 [2024-08-16]. https://www.semanticscholar.org/reader/329734fdbb35faab89e14eb9b105a665d7a5f079

[208] Zhang Jiliang, Peng Shuang, Gao Yansong, et al. APMSA: Adversarial perturbation against model stealing attacks[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1667−1679

[209] Tan Jingxuan, Zhong Nan, Qian Zhenxing, et al. Deep neural network watermarking against model extraction attack[C]// Proc of the 31st ACM Int Conf on Multimedia. New York: ACM, 2023: 1588−1597

[210] Zhang Haitian, Hua Guang, Wang Xinya, et al. Categorical inference poisoning: Verifiable defense against black-box DNN model stealing without constraining surrogate data and query times[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1473−1486

[211] Dai Hongning, Zheng Zibin, Zhang Yan. Blockchain for Internet of things: A survey[J]. IEEE Internet of Things Journal, 2019, 6(5): 8076−8094

[212] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time [J]. arXiv preprint, arXiv: 1708.06131, 2013

[213] Tang Pengfei, Wang Wenjie, Lou Jian, et al. Generating adversarial examples with distance constrained adversarial imitation networks [J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(6): 4145−4155

[214] Bagdasaryan E, Veit A, Hua Yiqing, et al. How to backdoor federated learning[C]// Proc of the 23rd Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2020: 2938−2948

[215] Zhang Jiale, Chen Bing, Cheng Xiang, et al. PoisonGAN: Generative poisoning attacks against federated learning in edge computing

systems[J]. IEEE Internet of Things Journal, 2021, 8(5): 3310−3322

[216] Qammar A, Ding Jianguo, Ning Huansheng. Federated learning attack surface: Taxonomy, cyber defences, challenges, and future directions[J]. Artificial Intelligence Review, 2022, 55(5): 3569−3606

[217] Kim T, Singh S, Madaan N, et al. Characterizing internal evasion attacks in federated learning[C]// Proc of the 26th Int Conf on Artificial Intelligence and Statistics. New York: PMLR, 2023: 907−921

[218] Bao Hongyan, Han Yufei, Zhou Yujun, et al. Towards effcient and domain-agnostic evasion attack with high-dimensional categorical inputs[C]// Proc of the 37th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2023: 6753−6761

[219] Demontis A, Melis M, Pintor M, et al. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks[C]// Proc of the 28th USENIX Conf on Security Symp. Berkeley, CA: USENIX Association, 2019: 321–338

[220] Blanchard P, Mahdi E, Guerraoui R, et al. Machine learning with adversaries: Byzantine tolerant gradient descent[C]// Proc of the 31st Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2017: 118–128

[221] Lugan S, Desbordes P, Brion E, et al. Secure architectures implementing trusted coalitions for blockchained distributed learning[J]. IEEE Access, 2019, 7: 181789−181799

[222] Bao Hongyan, Han Yufei, Zhou Yujun, et al. Towards understanding the robustness against evasion attack on categorical data[C/OL]// Proc of the 10th Int Conf on Learning Representations. Washington: ICLR, 2022[2024-07-30]. https://openreview.net/pdf?id=BmJV7ky-Amg

[223] Cao Xiaoyu, Gong N Z Q. Mitigating evasion attacks to deep neural networks via region-based classification[C/OL]// Proc of the 33rd Annual Computer Security Applications Conf. New York: ACM, 2017: 278−287

[224] Zizzo G, Rawat A, Sinn M, et al. FAT: Federated adversarial training [J]. arXiv preprint, arXiv: 2012.01791, 2020

[225] Kumari K, Rieger P, Fereidooni H, et al. BayBFed: Bayesian backdoor defense for federated learning[C]// Proc of the 2023 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2023: 737−754

[226] Cao Xiaoyu, Jia Jinyuan, Zhang Zaixi, et al. FedRecover: Recovering from poisoning attacks in federated learning using historical information[C]// Proc of the 44th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2023: 1366−1383

[227] Wen Jing, Hui L C K, Yiu S M, et al. DCN: Detector-corrector network against evasion attacks on deep neural networks[C]// Proc of the 48th Annual IEEE/IFIP Int Conf on Dependable Systems and Networks Workshops. Piscataway, NJ: IEEE, 2018: 215−221

[228] Debicha I, Bauwens R, Debatty T, et al. TAD: Transfer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems[J]. Future Generation Computer Systems, 2023, 138: 185−197

[229] Lecuyer M, Atlidakis V, Geambasu R, et al. Certified robustness to adversarial examples with differential privacy[C]// Proc of the 40th IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2019: 656−672

[230] Byrd D, Polychroniadou A. Differentially private secure multi-party computation for federated learning in financial applications[C]// Proc of the 1st ACM Int Conf on AI in Finance. New York: ACM, 2021: Article 16

[231] Rathee D, Rathee M, Kumar N, et al. Cryptflow2: Practical 2-party secure inference[C]// Proc of the 2020 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2020: 325–342

[232] He Xuanli, Lyu L, Xu Qiongkai, et al. Model extraction and adversarial transferability, your bert is vulnerable![C]// Proc of the 2021 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2021: 2006–2012

[233] Keskar N S, Mccann B, Xiong Caiming. The thieves on sesame street are polyglots-extracting multilingual models from monolingual APIs[C]// Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 6203–6207

[234] Wu Huaming, Wolter K. Stochastic analysis of delayed mobile offloading in heterogeneous networks[J]. IEEE Transactions on Mobile Computing, 2018, 17(2): 461−474

[235] Tu Xuezhen, Zhu Kun, Luong N C, et al. Incentive mechanisms for federated learning: From economic and game theoretic perspective [J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(3): 1566−1593

[236] Liu Shaoshan, Liu Liangkai, Tang Jie, et al. Edge computing for autonomous driving: Opportunities and challenges[J]. Proceedings of the IEEE, 2019, 107(8): 1697−1716

[237] Li Yuanchun, Wen Hao, Wang Weijun, et al. Personal LLM agents: Insights and survey about the capability, efficiency and security [J]. arXiv preprint, arXiv: 2401.05459, 2024

[238] Yu Sixing, Muñoz J P, Jannesari A. Federated foundation models: Privacy-preserving and collaborative learning for large models [J]. arXiv preprint, arXiv: 2305.11414, 2023



Luo Yuzhe, born in 1995. PhD candidate. His main research interest includes distributed machine learning.

罗宇哲, 1995 年生. 博士研究生. 主要研究方向为分布式机器学习.



Li Ling, born in 1982. PhD, professor.Senior member of CCF. Her mian research interest includes intelligent computing.

李　玲, 1982 年生. 博士, 研究员. CCF 高级会员. 主要研究方向为智能计算.



Hou Pengpeng, born in 1985. PhD, associate professor. His main research interest includes operating system.

侯朋朋, 1985 年生. 博士, 副研究员. 主要研究方向为操作系统.



Yu Jiageng, born in 1983. PhD, associate professor. His main research interests include operating system, cloud computing and intelligent software.

于佳耕, 1983 年生. 博士, 副研究员. 主要研究方向为操作系统、云计算、智能软件.



Cheng Limin, born in 1988. PhD candidate. Her main research interests include intelligent system software, machine learning and video analysis.

程丽敏, 1988 年生. 博士研究生. 主要研究方向为智能系统软件、机器学习和视频分析.



Zhang Changyou, born in 1970. PhD, professor. His main research interests include parallel distributed software and software engineering.

张常有, 1970 年生. 博士, 研究员. 主要研究方向为并行与分布式软件、软件工程.



Wu Yanjun, born in 1979. PhD, professor. Senior member of CCF. His main research interests include operating system and system security.

武延军, 1979 年生. 博士, 研究员. CCF 高级会员. 主要研究方向为操作系统、系统安全.



Zhao Chen, born in 1967. PhD, professor. His main research interests include logic and automatic program analysis, and design and implementation of programming languages.

赵　琛, 1967 年生. 博士, 研究员. 主要研究方向为逻辑、自动程序分析和编程语言设计与实现.