

面向边缘智能的移动算力网络

Research on Key Technologies of Mobile Computing Force Network for Edge Intelligence

关键技术研究

郑远鹏¹,张天魁¹,庞 博²,陈 杲²(1. 北京邮电大学,北京 100876;2. 中国联通研究院,北京 100191)
Zheng Yuanpeng¹,Zhang Tiankui¹,Pang Bo²,Chen Gao²(1. Beijing University of Posts and Telecommunications, Beijing 100876, China;2. China Unicom Research Institute, Beijing 100191, China)

摘 要:

随着人工智能计算与应用的不断发展,基于深度学习的智能服务对算力资源需求也逐渐提高,因此利用移动算力网络为边缘智能提供泛在异构、高效可靠的算力服务支撑。针对边缘智能和移动算力网络的研究现状和关键技术进行了讨论,提出了面向边缘智能的移动算力网络设计实例,给出了面向智能任务的网络部署方案和多维资源分配方案。

关键词:

边缘智能;移动算力网络;结合部署;资源分配
doi:10.12045/j.issn.1007-3043.2023.05.015
文章编号:1007-3043(2023)05-0088-05
中图分类号:TN915
文献标识码:A
开放科学(资源服务)标识码(OSID):



Abstract:

With the continuous development of artificial intelligence computing and applications, the demand for computing resources in intelligent services based on deep learning has gradually increased. Therefore, using mobile computing force networks to provide ubiquitous, heterogeneous, efficient and reliable computing force services for edge intelligence is essential. It discusses the research status and key technologies of edge intelligence and mobile computing force network, proposes a design example of a mobile computing network for edge intelligence, and provides a deployment scheme for intelligent tasks and a multi-dimensional resource allocation scheme.

Keywords:

Edge intelligence; Mobile computing force network; Combined deployment; Resource allocation

引用格式:郑远鹏,张天魁,庞博,等. 面向边缘智能的移动算力网络关键技术研究[J]. 邮电设计技术,2023(5):88-92.

1 概述

近年来,随着从云数据服务器到个人端的计算和存储设备的激增,计算普遍化和计算服务下沉的信息中心的时代到来。Machina Research报告显示,预计到2025年全球联网设备会超过270亿台;同时,根据Gartner预测,2025年将会有超过75%的数据需要分流到网络边缘侧^[1]。边缘计算作为一种将计算任务尽量靠近数据源和用户终端处理的技术^[2],具有缓解骨干网压力、服务响应灵活和任务完成时延低的优势。

随着人工智能技术的不断发展和广泛应用,基于深度学习的智能服务对算力资源需求大幅增加,在网络边缘部署智能服务并提供高效、泛在和安全的计算服务获得广泛关注,相关研究提出了边缘智能(Edge Intelligence, EI)概念。边缘计算和智能服务相结合,形成了边缘智能的概念,边缘与智能的结合能使两者相互受益^[3]。同时,为了充分挖掘异构广域的可用计算资源,相关研究机构提出了计算与通信资源深度融合的算力网络(Computing Force Network)概念^[1]。算力网络将不同地理位置分布的云计算服务器、边缘计算服务器等算力节点组网,通过计算任务和计算结果在不同算力节点间传输实现了异构算力协同,并具备

收稿日期:2023-03-16

计算资源状态感知、面向计算任务的动态路由、计算任务统筹调度等能力。结合边缘计算和算力网络,移动算力网络概念被提出。移动算力网络是指由移动设备和边缘服务器组成的计算网络,通过在网络中相

互协作、共享计算资源和任务负载提供边缘计算服务。显然,移动算力网络将为边缘智能提供泛在异构的算力资源服务。面向边缘智能的移动算力网络架构如图1所示。

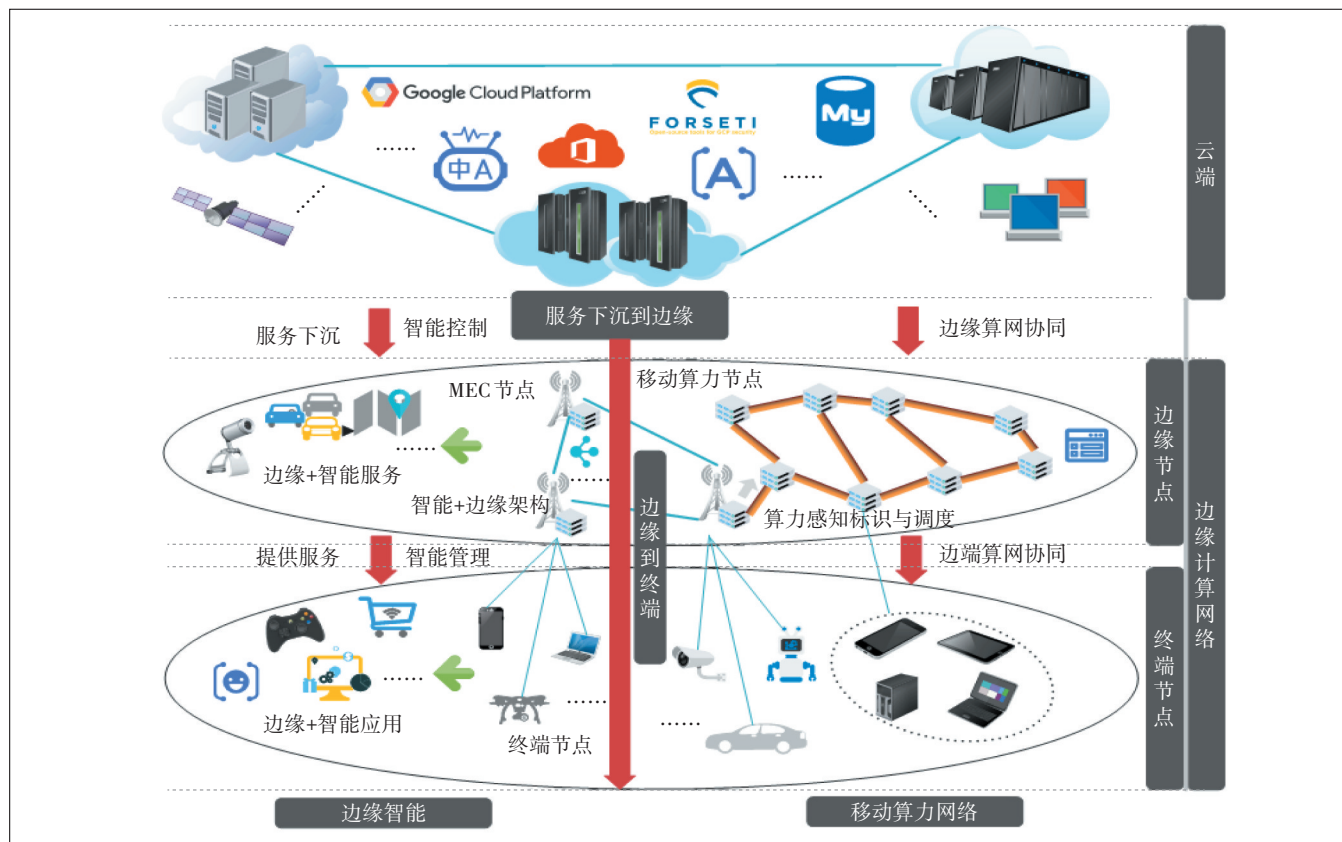


图1 面向边缘智能的移动算力网络架构

边缘智能技术通过边缘与智能的相互结合,为边缘计算带来更多的服务种类和更灵活的网络部署,以满足日益增长的业务需求。边缘智能概念包含以下2个方面:第一,致力于通过边缘计算部署人工智能服务,形成边缘+智能技术;第二,将人工智能技术整合进边缘计算网络,打造动态自适应的边缘网络管理和维护,形成智能+边缘技术。边缘+智能和智能+边缘之间并非相互独立,在边缘智能中,边缘+智能作为边缘智能的目标,将深度学习计算从云推向边缘,提供多样化、分布式、低时延、可靠的智能服务;智能+边缘中的深度学习服务也是边缘智能的一部分,通过深度学习的方法对终端-边缘-云这一动态性网络架构进行管理和维护,为边缘智能提供更高的服务吞吐量和资源利用率。

算力是对数据处理能力及服务的统称,由多种芯片、部件和封装形成的上层服务组成^[4]。而算力网络

是一种以泛在算力资源为基础、网络通信为纽带、智能化调度为核心,实现网云边端和业务的高效协同与适配,从而满足差异化算力服务需求的技术^[5]。目前算力网络研究尚处于起步阶段,主要面向需求日益增长的计算和网络相互感知、协同、调度问题,推进网络和计算的深度融合。

移动算力网络则面向多级算力资源(云计算、边缘计算和端侧计算)并存情况下的资源统一供给问题。移动算力网络通过云网协同、云边协同和边边协同等算力协同技术,重点解决边缘计算中资源泛在化后的用户体验一致性以及服务灵活动态部署问题^[6],实现资源利用的最优化,从而推进云、网、边的深度融合,在移动网络中提供包括人工智能、区块链、云、数据、网络、边缘计算、终端计算和安全保障的整合服务。

因此,综合上述边缘智能和移动算力网络的概

念,下面针对边缘智能和移动算力网络的研究现状和关键技术进行了讨论,提出了面向边缘智能的移动算力网络设计实例,并给出了面向智能任务的网络部署方案和多维资源分配方案,为后续边缘智能和移动算力网络结合技术的实现和应用部署提供有益参考。

2 面向边缘智能的移动算力网络技术挑战

边缘智能和移动算力网络两者互为依托,边缘智能关键技术包括边缘训练、边缘推理和智能边缘管理,需要移动算力网络提供灵活的算力调度和优化;移动算力网络关键技术包括算力感知、算力度量与标识和算力路由,可以依靠边缘智能进行服务扩展和智能部署。因此,在智能应用逐渐增多的背景下,在移动网络的基础上逐渐形成了面向边缘智能的移动算力网络技术,移动算力网络不断促进边缘智能技术发展与应用部署。

2.1 边缘智能技术研究现状

在边缘+智能方面,部分研究仍在起步阶段。相关研究首先考虑了如何将深度学习服务靠近用户端侧部署,云端只参与所需的额外处理^[7],这样可以显著降低传输数据至云端的时延和开销;然后考虑了将深度学习所需的原始数据存储的边缘或用户端侧而非云端,增强用户的隐私保护^[8];部分研究考虑了部署适用于深度学习的分层计算架构,提供更为可靠的深度学习计算,同时还具有自学习特性,能高效处理不同的智能任务需求;而后更多研究开始考虑边缘拥有的丰富数据和应用场景,从而适配深度学习应用,实现无处不在的人工智能服务的前景;最后,在商业应用方面,部分研究通过在边缘部署人工智能以降低系统开销,并通过多种类的深度学习应用部署,提高系统效用,从而提升整个系统的商业价值。

在智能+边缘方面,相关研究起步较早。首先提出的是边缘计算架构能为端侧和云端提供可靠和持续性的连接,将云-边-端整合进一个共享资源池中,但是对这一整合了无线通信、网络、计算和存储等功能的庞大架构进行管理和维护十分复杂,典型的基于固定数学模型的网络优化方法很难适配快速变化的边缘网络环境和系统。并且,边缘节点的计算和存储能力以及云边协同,使得通过深度学习来优化边缘计算网络和系统成为可能^[3]。因此提出引入深度学习方法解决这类面向复杂网络信息的问题,通过深度学习进行自适应的决策、智能管理和维护。最后,部分研

究开始对自适应边缘缓存,边缘任务卸载以及包括边缘通信、安全和联合边缘优化的管理和维护等具体场景的研究,主要优化移动边缘计算系统的资源管理。

2.2 移动算力网络技术研究现状

移动算力网络目前主要从RAN通信计算融合、算网协同调度和移动算网一体化3个阶段出发,从架构、协议、度量等方面协同演进。边缘计算服务器作为兼具通信计算功能的边缘主体,在RAN整合通信和计算功能中发挥着重要作用,该类研究主要通过研究多个边缘节点计算资源的分配和调度,满足用户的业务体验并提高资源利用率。而随着云边算力的泛在发展,网络变得扁平灵活,算网协同调度阶段研究首先面向业务算力的协同感知技术,设计算力感知技术架构,通过服务自动化部署、最优路由和负载均衡等技术构建协同感知能力。而后通过网络感知的应用、计算等多维度资源,对算力路由进行了研究,能够将算力和业务按需调度到合适的节点,实现自适应的算力调度功能。在移动算网一体化阶段,首先面向云、网、边融合的资源利用,部署多级算力节点,对计算、存储和网络算法资源进行多级部署和灵活调度,初步提出算网一体化构想^[6],而后部分研究对算网资源的统一运维和优化部署方案进行讨论,提出算网融合需要云边端多级算力的泛在和互联,减少边缘计算节点的管控复杂度,以达到网络资源和计算资源的全面融合^[4]。

2.3 面向边缘智能的移动算力网络技术挑战

当前,已有研究关注了边缘智能与移动算力网络相结合需要解决的关键技术难题,具体包括以下2个方面。

a) 面向边缘智能的算网资源管控。目前,已有从具体应用如图像传输到更一般的智能模型性能和资源分配结合的研究,该研究主要结合智能服务的精度特性和无线接入网络资源,讨论包括算力在内的资源对智能服务的模型推理精度的影响。而后提出在边缘网络中部署分布式学习和联邦学习方法,通过灵活分配资源的方法支持分布式模型训练的部署,初步提出计算负载量化的概念。

b) 智能技术辅助的移动算力网络设计。目前,相关研究首先从移动边缘计算的计算卸载问题出发,讨论了在边缘部署智能的需求以及所需的算力网络架构,并结合算力感知和调度功能设计算力网络框架^[6];而后从智能服务辅助算力网络的角度出发,通过部署分布式学习的方法解决动态的计算卸载和算力调度

问题。

然而,需要指出的是,面向边缘智能的移动算力网络部署目前仍存在诸多挑战,一方面需要部署多种类的定制化轻量智能模型,在包括边缘训练和边缘推理方面减轻边缘的资源压力,并能提供更多种智能服务,对模型的参数量和精度有非常高的要求;另一方面需要移动算力网络进行高效率的算力感知,适配智能模型的多业务部署,在算力度量和路由的设计方面存在很大的挑战。

3 面向边缘智能的移动算力网络设计实例

在边缘智能关键技术中,存在很多智能模型的轻量化部署方法,包括模型的优化、模型的分割部署以及模型的提前退出,这些技术为面向边缘智能与移动算力网络部署提供了基础。在移动算力网络关键技术中,算力感知度量与路由为部署提供了资源保障。因此,结合部署的关键在于整合2类技术,根据智能服务的场景进行不同层次的网络部署与资源管控。

3.1 网络部署设计方案

模型轻量化部署可以显著减轻算力网络资源压力。在深度学习模型分割中,模型可以被分解成多个部分并按照算网的适配性分配至终端设备中异构的本地处理器、分布式算力节点或合作式的“端-边-云”架构,其中分配至“端-边-云”架构的被称为水平分割,分配至分布式算力节点的被称为垂直分割^[3]。垂直分割方法按照输入维度对模型进行分割,将模型各层分成独立的分布式计算任务进入各个算力节点,为移动算力网络提供了很好的适配条件,而移动算力网络的分层算力感知架构^[4]为智能模型的部署提供了底层条件。

面向边缘智能与移动算力网络部署需要结合算力感知技术架构,对边缘智能服务进行垂直分割。通过合适的模型分割算法将智能模型分割成适合部署在算力网络中的分布式计算模式,由算力编排层进行统一管理,通过算力服务层承载,由算力感知功能对分割后的计算任务进行算力需求的感知,而后算力路由层将分割后的子任务分配算力资源,即路由调度到不同的算力节点中,由网络资源层和算力资源层提供网络连接和算力资源进行传输和计算,最后将业务部署完成并反馈给用户。可以看出,面向边缘智能与移动算力网络部署需要高度融合的算网架构和成熟的模型分割算法,以确保网络资源的利用和智能服务的

质量,同时还需要云-边-端的协同以应对不同场景下不同类型的智能任务,将边缘算力节点无法处理的智能任务回传至云端进行处理,结合部署架构如图2所示。

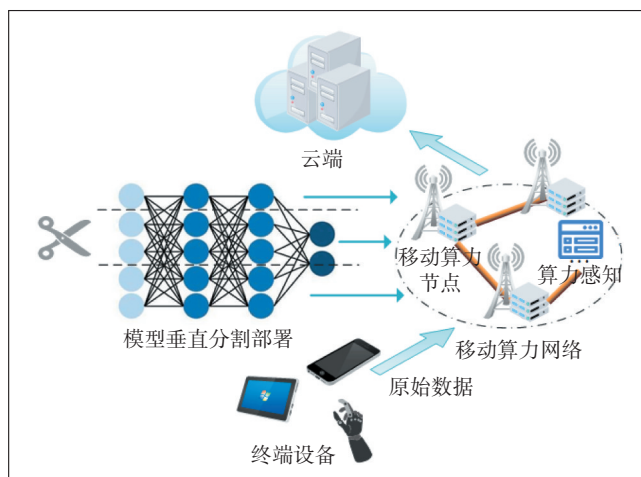


图2 面向边缘智能的移动算力网络部署

3.2 面向边缘智能的资源分配设计方案

基于上述部署架构,需要考虑面向边缘智能任务的多维资源分配方法,以满足边缘智能服务对通信和计算资源的需求。本文提出了一种面向智能任务的资源分配方法。智能任务需要考虑训练精度参数,系统采用OFDMA多任务到达,计算任务并行处理。故在原始数据压缩回传时,考虑数据量和训练精度的关系,同时考虑多计算任务并行对计算时延的影响。考虑将移动算力设备部署至基站侧,设接入基站 k 的终端为 u_k ,考虑终端任务 m 存在的时延和精度需求,由边缘供给计算资源,将原始数据压缩上传至边缘服务器进行处理,并将结果回传给终端执行。考虑用户数据量 α_{u_i} 对智能任务处理精度 $y(\alpha_{u_i})$ 的影响,如式(1)所示。

$$y(\alpha_{u_i}) = p - q \times \alpha_{u_i}^{-r} \quad (1)$$

其中, p 、 q 和 r 均为拟合参数。考虑保证时延内多计算任务并行对任务处理时延的影响,假设任务 m 处理时并行处理的任务数为 i_m ,那么有:

$$T_{u,m} = T_{u_i}^0 \times (1 + d)^{i_m - 1} \quad (2)$$

其中, d 为退化因子,即一个虚拟机在与另一个虚拟机复用时经历的预期服务时间增加的百分比,用于表征并行任务对移动算力节点计算产生的影响。考虑在精度限制和时延限制条件下,优化通信与计算资源分配和压缩比,并以精度和时延的结合作为系统效

用,通过连续凸近似的方法进行求解。

随后设计固定信道(FC)、平均计算(AC)和不考虑压缩比并行(WCR)的对比算法进行仿真性能分析,图3和图4为该场景下系统效用和计算容量以及并行参数的变化关系。根据系统效用在不同带宽(即10 MHz和50 MHz)下随计算容量的变化,本文提出的算法系统效用在200 Gigacycle/s时达到最大值,在并行参数的影响中可以看出系统能效呈下降趋势,且并行参数越大趋势越缓和。此外,可以看出更高的带宽对变化趋势的影响不显著。

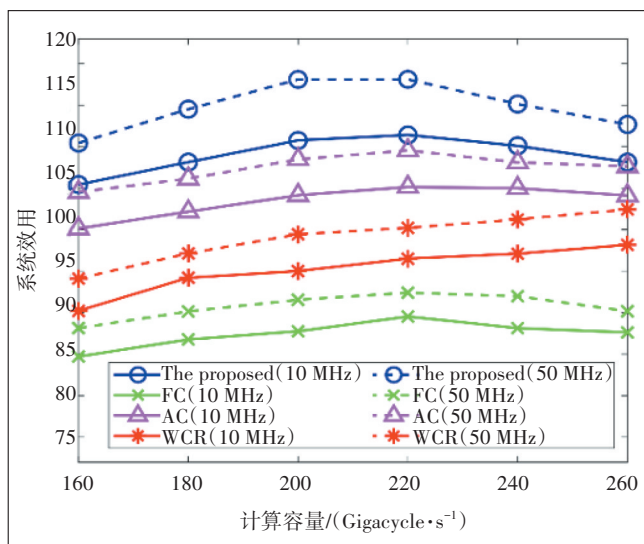


图3 系统效用和计算容量变化关系

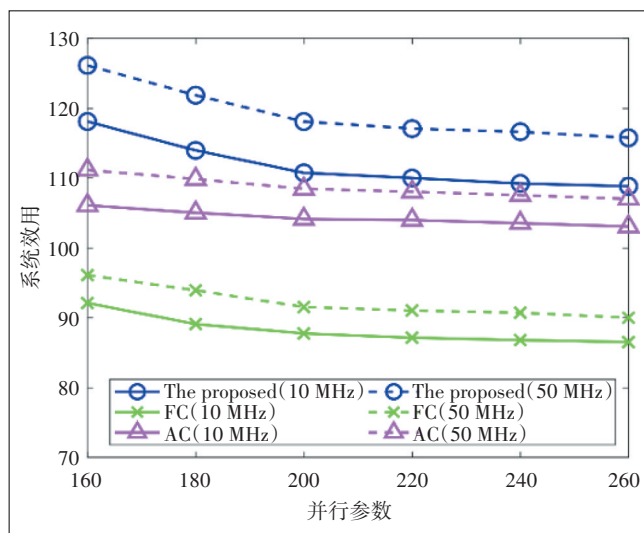


图4 系统效用和并行参数变化关系

该场景模型对移动算力网络中的边缘节点在面向智能任务时的性能进行了建模分析,通过传统优化

的方法提出了结合任务精度的资源分配的算法;通过仿真分析可以看出,在面向智能任务时,不同需求量和不同资源量对系统性能影响的不同,对面向边缘智能的移动算力网络结合部署的资源分配有着参考意义。

4 结束语

边缘智能和移动算力网络作为5G时代演进至6G未来移动通信的关键技术,在针对智能服务和移动算力利用方面都有重要作用。本文针对新兴的边缘智能和移动算力网络的概念、发展和关键技术进行了讨论,并提出了一种面向边缘智能的移动算力网络技术,分析了面向智能任务的边缘资源分配实例,为未来边缘智能和移动算力网络的结合部署提供了参考。

参考文献:

- [1] Gartner. Gartner research; top 10 strategic technology trends for 2018: cloud to the edge [EB/OL]. [2023-01-08]. <https://www.gartner.com/en/documents/3865403>.
- [2] SHI W S, CAO J, ZHANG Q, et al. Edge computing: vision and challenges [J]. IEEE Internet Of Things Journal, 2016, 3(5): 637-646.
- [3] WANG X F, HAN Y W, LEUNG V C M, et al. Convergence of edge computing and deep learning: a comprehensive survey [J]. IEEE Communications Surveys & Tutorials, 2020, 22(2): 869-904.
- [4] 段晓东, 姚惠娟, 付月霞, 等. 面向算网一体化演进的算力网络技术 [J]. 电信科学, 2021, 37(10): 76-85.
- [5] 张宏科, 权伟, 刘康. 算力网络研究与探索 [J]. 中兴通讯技术, 2023, 29(1): 1-5.
- [6] 雷波, 刘增义, 王旭亮, 等. 基于云、网、边融合的边缘计算新方案: 算力网络 [J]. 电信科学, 2019, 35(9): 44-51.
- [7] KANG Y P, HAUSWALD J, GAO C, et al. Neurosurgeon: collaborative intelligence between the cloud and mobile edge [C]//Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems. Xi'an, China: Association for Computing Machinery, 2017: 615-629.
- [8] JEANS D. Related's Hudson Yards: smart city or surveillance city? [EB/OL]. [2023-01-15]. <https://therealdeal.com/new-york/2019/03/15/hudson-yards-smart-city-or-surveillance-city/>.

作者简介:

郑远鹏, 北京邮电大学, 博士在读, 主要从事无线网络方向的研究工作; 张天魁, 北京邮电大学教授, 博士生导师, 主要从事通信与信息系统专业的科研与教学工作; 庞博, 硕士, 主要从事无线接入网和边缘计算相关领域的研究工作; 陈杲, 高级工程师, 博士, 主要从事边缘算网、6G等相关领域的研究工作。