

基于边缘智能的沉浸式元宇宙关键技术与展望

王智¹, 夏树涛¹, 毛睿²

1. 清华大学深圳国际研究生院, 广东 深圳 518055;

2. 深圳大学, 广东 深圳 518060

摘要

近年来, 360度视频、增强现实、虚拟现实等应用蓬勃发展, 并逐渐形成元宇宙沉浸式体验新模式。这些新服务和应用具有高保真、沉浸式交互等共同特征。而作为一种新兴架构范式, 边缘计算有效支持了这些需求。阐述释放边缘计算潜力支持沉浸式元宇宙体验的关键, 包括AI驱动的内容生成; 提出一种精度与延迟感知的模型推理框架, 实现自适应深度学习模型部署和数据流推理, 支撑用户体验感知的元宇宙服务与应用。

关键词

边缘计算; 人工智能; 元宇宙; 沉浸式体验

中图分类号: TP37

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2024001

Edge-intelligence-based immersive metaverse: key technologies and prospects

WANG Zhi¹, XIA Shutao¹, MAO Rui²

1. Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China

2. Shenzhen University, Shenzhen 618060, China

Abstract

In recent years, applications such as 360 video, augmented reality, virtual reality, etc. have been rapidly developed and gradually formed a new mode of immersive metaverse experience. These new services and applications have common features such as high fidelity and immersive interaction. As a new paradigm, edge computing is increasingly ready to support these features. This paper firstly elucidates that the key to releasing the power of edge computing to support immersive metaverse experience is the processing of AI models and data for content generation. Then, this paper presents a generic framework to realize adaptive deep learning model deployment and data flow to support metaverse services and applications.

Key words

edge computing, artificial intelligence, metaverse, immersive experience

0 引言

内容服务和用户交互的元宇宙化是互联网在线服务发展的趋势,并逐渐形成多元虚拟空间^[1]、用户生成场景^[2]、可计算IP等相互交互的新型在线服务和体验。在这一趋势的推动下,国际科技巨头纷纷开始在多元宇宙、全真互联网领域布局^[3]。2021年Facebook改名Meta正式宣布打造元宇宙服务;微软收购了视频游戏公司动视暴雪,以游戏为基点向元宇宙扩张。作为下一代互联网的支撑技术,元宇宙吸引了包括网络游戏制造商、互联网金融企业、社交网络和其他垂直产业布局。根据Bloomberg Intelligence分析,全球元宇宙收入将从2020年的5 000亿美元增加到2024年的8 000亿美元。

(1) 核心技术挑战

元宇宙场景中需要进行高实时、高码率、高动态的场景创作、用户交互以及内容分发,给传统内容服务模式带来巨大挑战。以下核心技术瓶颈将阻碍元宇宙相关产业的发展。

迫切需要面向多元世界的场景与内容生成技术。元宇宙时空是一个与现实世界映射与交互的虚拟世界与协作空间^[4],如支持用户交互的虚拟博物馆、虚拟街区、虚拟教室等。现有虚拟环境生成大多源于特定区域(如大学、景区)或者特定领域(如教育、工业检测)。这样的场景通常由专业人员创建,导致制作场景的成本高。此外,缺乏稳定的内容质量与场景扩展制约了元宇宙应用的发展,迫切需要快速进入人工智能(AI)生成内容阶段^[5]。形成智能驱动的多元世界,需要从复杂的场景解析、多模态数字内容生成,以及内容生成多样化3个方面进行考虑。

迫切需要面向稀疏数据的个性与虚像重建技术。虚拟化身是用户在元宇宙时空中的身份象征^[6]。进入元宇宙时空通常需要对应的虚拟化形象、千人千面。利用智能方式建模虚拟数字人的前提是获取足够的数据,例如:对大量照片、视频进行学习,提取对应用户的肢体与行为特征,进而模拟其外观和行为。在这个过程中样本越多,算法生成的模型就越优秀。智能生成的虚拟数字人所呈现的效果受到自然语言处理、语音生成、情感驱动面部动画等技术的共同影响^[7]。目前人工智能建模尚不能完全取代三维采集、人工建模的结果,尤其是对于超写实和影视级别的虚拟数字人需求。人工智能驱动的个性化虚拟形象生成需要解决算法效率低、形象刻板、情感缺失等问题。

迫切需要面向端云协同的扩展与算力支撑技术。在主流的端云协同架构中,用户通过网络上传交互等行为数据,云服务器运行相应的计算流程,包括机器学习模型接收数据并进行推理服务,并通过网络将反馈结果传输给用户^[8]。这一架构面临以下一系列挑战:

- 模型的日益增大与端侧资源偏弱之间的矛盾,导致传统分布式训练框架效率偏低^[9];
- 涉及异构化硬件资源及多方参与,使得传统中心化预训练模型部署方法难以满足层次化部署需求;
- 在智能推理任务的执行中,传统推理服务难以适应元宇宙应用复杂多变的决策需求。作为一种新兴的计算架构和服务模式^[10],边缘计算为元宇宙场景和形象智能化生成提供了新的思路。

(2) 沉浸式元宇宙发展依赖边缘智能
随着360度视频、增强现实、虚拟现实等新应用的快速发展,一个沉浸式元宇宙服务体系正在形成,这些应用特点是高保真度和沉浸式交互^[11]。在基于云的解决方

案中,多媒体内容从全球部署的云服务器进行流式传输;而在基于设备(端侧)的解决方案中,内容处理工作流程由用户的终端设备(如头戴式显示器)完成,此类设备通常有本地计算和存储支持,可以在本地渲染和显示虚拟现实场景。

这两种范式在高质量沉浸式元宇宙应用方面存在局限性。基于云的解决方案假定所有或大部分多媒体功能可以在云中进行远程部署,但这必须满足严格数据隐私要求的服务或应用是不可行的^[12];同时,传播延迟和云基础设施(如负载均衡和重定向机制)造成的延迟,使得实现低延迟服务的目标变得困难;此外,由于专用基础设施的安装和维护成本,云部署更加昂贵。以设备为中心的解决方案最大的限制是用户端设备上的能源、计算和存储资源有限,导致其无法为资源密集型元宇宙应用提供质量保证;另外,这种解决方案假定在某些设备上独立安装,限制了应用扩展的灵活性。

由于这些限制,新的范例逐渐出现,其中边缘计算赋能沉浸式元宇宙前景广阔^[13]。作为一种新兴的基础设施设计范例,边缘计算通常部署在离用户很近的地方,作为云和用户设备之间的“中间层”。一方

面,用户在沉浸式元宇宙应用中生成交互数据,可由边缘设备处理、以“用户设备到边缘”的方式卸载(offloading)^[14];另一方面,云服务器也可以将某些任务卸载到边缘,如执行轻量化深度学习模型,并在本地提供推理服务,实现更低的成本和延迟、高可扩展性和更好的隐私保护^[15]。参考文献[16]对元宇宙中的数据管理技术进行了讨论,将元宇宙分解为云、边、端3个层次,分析了元宇宙中的大数据特点和数据管理特点。参考文献[17]从边缘计算赋能人工智能的维度,针对深度学习模型在网络边缘侧的部署,提出了基于边端协同的深度学习按需加速框架,实现时延约束下的高精度模型推理。与以设备为中心的解决方案相比,边缘计算更具资源性和可扩展性,并随着时间的推移更易适应应用部署。目前,针对沉浸式元宇宙技术的研究主要集中在图1所示的范畴,包括场景和内容的自主生成,虚实融合的数据表达与处理,以及高能量低延迟计算传输。

互联网核心技术是我国“网络强国”战略中最大的“命门”,而沉浸式元宇宙可能成为未来互联网发展的新方向,沉浸式元宇宙有望成为新一代数字内容产业、互联网信息服务等重大应用的共性基础。目

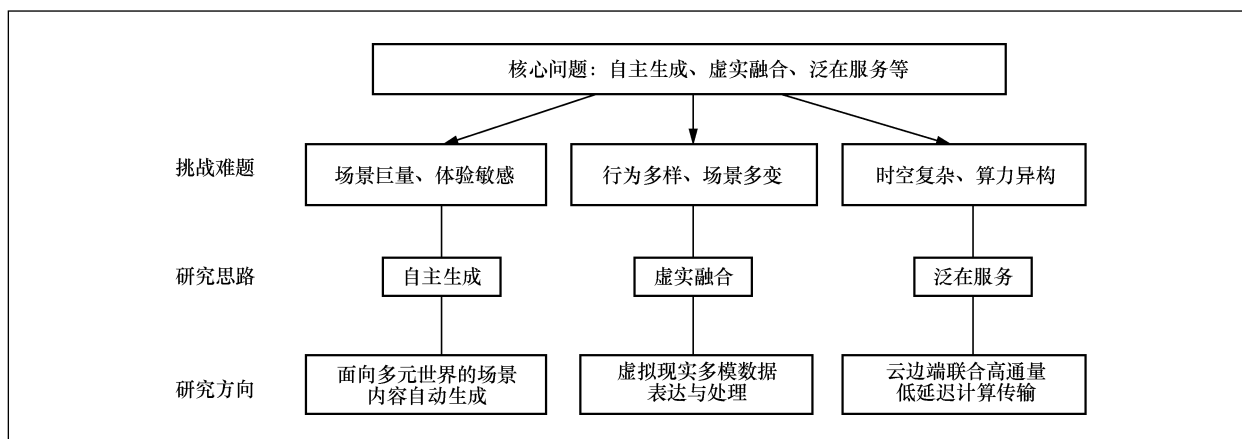


图1 沉浸式元宇宙关键技术

前，沉浸式元宇宙应用已积累了足够的前序发展时间，而随着爆发式发展窗口的到来，迫切需要自主生成、虚实融合、泛在服务等关键核心技术支撑。

1 边缘智能与沉浸式元宇宙服务

1.1 沉浸式元宇宙服务中的核心任务

1.1.1 数据驱动的场景内容生成

在沉浸式元宇宙应用中，虚拟环境提供基本应用场景，而元宇宙场景可以分为两类。①大型多人在线游戏式场景：支持多人在高复杂度的三维环境中互动、交流与内容创建等，如在《我的世界》沙盒游戏中，玩家可与由方块和实体组成的可修改三维环境交互、构建建筑和关卡。②去中心化的虚拟世界：形成虚拟世界的产权和交易属性，利用区块链等技术支持，创建去中心的众筹世界。在面向沉浸式元宇宙体验的服务与应用中，用户通常佩戴虚拟现实、混合现实设备，虚拟场景的大规模实时构建是关键技术难题。

1.1.2 数字角色与虚拟人生成

得益于人工智能技术的发展，近年来虚拟数字人的制作流程得到有效简化，且成本进一步降低，虚拟数字人行业开始快速发展。目前数字虚拟人主流制作技术分为两条路径。①计算机图形学（computer graphics, CG）通过计算机软件绘制出来的各种图像，被用于视觉设计和生产的相关行业，推动了虚拟场景、虚拟人物等内容的创作^[18]。②AI技术是生成虚拟数字人的第二条路径，大幅降低了制作成本、简化

了制作流程。然而，AI仍处于发展阶段，一些细微的不自然变化很难用传统手工方式实现，而市场上巨头公司推出的虚拟数字人仍有很多缺陷。如何快速生成虚拟数字人、降低成本且提高精度，仍是一个迫切需要解决的技术瓶颈。

1.2 面向沉浸式体验的边缘智能

为了改善用户的沉浸式体验，需要高质量的网络环境，而边缘拓扑结构和内容缓存是重要的网络优化策略。研究者们正在从3个方面努力提升用户沉浸式体验：第一，构建能确保用户体验的高质量网络拓扑和基础架构，让多媒体内容和数据以服务质量保证的带宽流向用户端；第二，构建有效的数据预取和缓存策略，让用户可以就近获取所需的内容^[19]；第三，根据不同的系统拓扑和通信模式，用户交互的响应质量可以通过本地处理（如预先下载的游戏场景）得到改善，边缘基础架构可以使元宇宙应用在边缘设备上处理特定任务，从而减少远程通信开销，进一步提升用户的沉浸式体验。

1.3 边缘计算赋能沉浸式元宇宙

边缘计算是一种分布式计算模式，能够将计算、存储和网络服务从云端推向靠近终端设备的边缘，以缩短数据传输的延迟，提高数据处理速度。在沉浸式元宇宙中，边缘计算具有以下重要作用。

- 低延迟：边缘计算能够在离用户最近的地方处理大量数据，缩短延迟，提高数据传输速度。这对于沉浸式元宇宙应用而言是至关重要的，因为任何延迟都会影响到用户的视觉体验。
- 实时性：边缘计算可以实时分析和处理用户设备产生的数据，对于沉浸式元

宇宙应用而言,实时性是保证用户体验流畅的关键。

- 高并发:通过在边缘节点上部署计算资源,边缘计算可以提供更强的并发能力。这对于支持大量用户的沉浸式元宇宙应用尤为关键,因为它们需要同时处理成千上万的用户请求。

- 省带宽:边缘计算可以在数据源的地方进行数据处理和存储,从而减少不必要的带宽消耗。这会降低网络传输成本,为沉浸式元宇宙提供更好的服务支持。

- 数据安全:边缘计算可以将用户数据处理和分析的过程控制在本地,从而保证用户数据的隐私性和安全性。这对于涉及个人隐私的沉浸式元宇宙应用具有重要意义。

- 离线功能:边缘计算允许元宇宙应用在没有云端连接的情况下继续运行,提供离线功能。这样用户在网络不稳定的环境中仍然可以体验到沉浸式元宇宙。

总之,边缘计算具有低延迟、高并发和数据安全等优势,极大地提升了沉浸式元宇宙中的用户体验。随着移动技术和物联网的快速发展,边缘计算在沉浸式元宇宙领域将发挥越来越重要的作用。

1.4 影响边缘智能性能的关键因素

处理深度学习模型是边缘智能的关键部分,能够保证在推理阶段达到用户期望的精度和延迟^[20]。沉浸式元宇宙应用有多种不同的推理性能指标,包括推理延迟、推理复杂度以及推理精度,并且这些指标会随时间变化。

- 模型对推理质量的影响。模型结构和训练状态会对推理性能产生重要影响,一般来说,更大的模型能够提供更高的精度,但同时会带来更高的计算成本和推理延迟,特别是在资源有限的边缘设备上。因

此,通常使用知识蒸馏、模型剪枝和模型量化等技术形成一系列“轻量化”模型版本。

- 数据对推理质量的影响。边缘智能的质量受到输入数据的影响。输入数据(如视频流、图像和音频)从用户上传到边缘设备和/或云服务器,上传延迟由数据体积决定,而推理准确性则取决于数据质量(如图像分辨率)。常用的改变输入数据的技术有数据压缩(如JPEG)和下采样(如区域裁剪和像素级下采样),对推理质量产生影响。

2 边缘智能驱动的沉浸式元宇宙核心技术

2.1 面向元宇宙场景的内容自动生成

2.1.1 基于虚实结合的内容生成技术

近年来,深度学习用于游戏等的程序性内容生成(PCG)相较传统方法有很多突破^[21],但多数工作是关于二维场景的内容生成且不支持实时交互。在处理三维场景内容生成方面,一种方法是采用生成对抗网络(GAN)结构^[22],通过如自然语言处理Word2Vec中的密集表示,使其生成三维结构,并形成更大规模的开放世界。在支持用户风格偏好、个性需求实时生成内容方面,可利用基于不同目标的适应度函数,离线训练多种对抗生成网络模型,根据用户移动的历史轨迹和摄像头捕捉的物理环境,生成新的虚拟场景。

2.1.2 基于神经渲染的三维重建技术

神经渲染(NeRF)是近年来提出的一种三维建模方式,简单来说是通过神经网络来隐式地表示三维模型^[23]。使用者可

以输入一个五维坐标（如物体位置坐标与观察位置）来获得对应点的RGB值与密度值，即可以任意视角查看三维模型，并通过体素渲染技术生成需要视角的图片。从渲染的角度，神经渲染能进行无限分辨率的渲染。尽管神经渲染三维重建的效果十分优秀，但由于网络参数量较大导致训练速度较慢，2021年微软团队将神经渲染分解为两个神经网络，并通过缓存架构来提高测试时间效率，实现了以200 Hz以上的速度渲染高分辨率逼真图像；而英伟达团队利用多分辨率哈希编码技术，通过完全融合的CUDA内核来实现神经渲染，大幅降低了时间开销。

2.2 多元复杂场景的内容自动生成

2.2.1 面向多元场景的多模态大模型内容生成

基于多模态大模型的生成技术已展示了其在图像生成、图像合成、图像编辑等场景的巨大潜力。现有的基于多模态大模型的生成技术主要在模型参数量和数据规模上进行扩展，但在多模态特征融合、模型结构轻量化方面缺乏深入探索，相关算法和模型落地仍有广阔的前景。如图2所示，可从多模态生成大模型预训练、多模

态生成大模型轻量化以及多模态生成大模型应用3个方面进行相关的理论和算法研究，以实现基于多模态大模型的生成技术在多模态场景生成上的高效应用与部署。

2.2.2 预训练与轻量化

（1）多模态生成大模型的预训练

基于多模态大模型的生成技术通过在海量的文本-图像对上进行模型的预训练来获得一个具有良好泛化性能的特征表示。在多模态相关技术中，关键问题是对不同模态的特征进行特征融合，并将其映射到同一个特征表示空间，以实现不同模态数据语义信息的统一。在文本和视觉特征融合方面，目前的方法主要是将文本和视觉建模为token序列，然后采用基于Transformer的架构进行特征融合。然而这样的方法没有对输入特征进行对齐，在Transformer中直接进行特征融合具有较高的训练难度，因此需要大量的数据来迫使模型学习其中不同模态的联系。因而模态特征的对齐和更好的特征融合将使多模态模型训练更加容易和高效，能提升模型特征表示的能力。在多模态生成大模型的预训练阶段，可以探究如何将不同模态的数据进行更好的融合，从而获得更好的多模态语义表示。

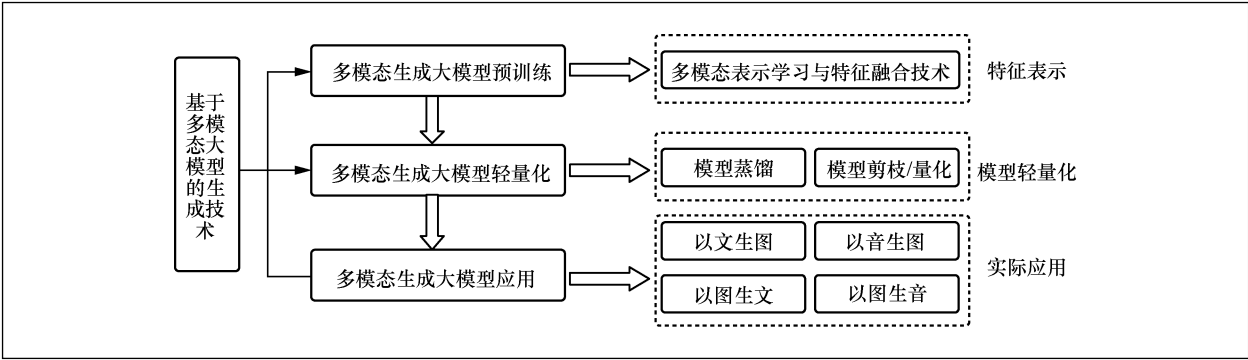


图2 基于多模态大模型的内容生成技术方案

(2) 多模态生成大模型的轻量化

多模态生成大模型目前强大的能力基于模型巨大的参数量,如DALL-E模型的参数量高达120亿,参数量如此巨大的模型在实际应用中的部署中是存在困难的,探究多模态大模型的轻量化对于模型在实际应用中的部署具有重要意义^[24]。大模型的轻量化通过模型蒸馏、模型剪枝和模型量化等方式,实现模型规模和参数量减少,并保留模型的特征提取能力,尽可能不降低其在下游任务中的性能^[25]。多模态大模型具有更强的特征表达能力并且参数规模更大,因此在模型轻量化方面存在一定的工程和技术挑战。

2.3 延迟与精度联合感知的边缘推理

2.3.1 数据自适应的高效边侧部署模型技术

为了将模型部署到资源受限的设备上,现有部署方法一般需进行部署策略搜索,而这一过程往往需要消耗大量时间才能完成。以模型量化为例,策略搜索算法需要为深度模型的每一层确定在待部署设备上的最优运行位宽^[26]。现代深度模型结构复杂,尽管可以应用强化学习、进化算法等智能搜索算法减少搜索规模,但过程仍耗时较长。同时,已有的部署策略搜索算法还面临着如下两个问题。第一,搜索策略与耗时设备无法迁移,即在一类设备上搜索得到的策略与这类设备高度相关,无法迁移到其他设备。当设备数量少、种类单一时,传统的模型部署方法产生的时间开销尚在可接受范围之内;而在元宇宙应用中设备数量大、种类多,继续使用传统方法为每个设备进行独立部署策略搜索,时间开销会成为整个部署过程的短板。第二,搜索策略与模型绑定无法更改。在真实部署场景中,边侧设备由于没有云

侧设备那样稳定的运行环境,常常需要模型及时进行调整。如当设备过热时,模型应调整其部署策略以减少计算资源的占用,而传统的策略搜索方法无法使模型在运行时感知设备的变化。

为了解决上述问题,可以通过学习层次重要性,实现混合精度神经网络量化。在混合精度量化中,指数级大的离散搜索空间使得确定每一层的最优位宽变得困难。参考文献[27]发现量化中的一些独特的可学习参数,即量化器中的缩放因子,可以作为一个层的重要性指标,反映该层在某些位宽下对最终精度的贡献。这些重要性指标自然地感知了在量化感知训练过程中的数值变换,能够准确且正确地提供层的量化敏感性度量。参考文献[28]提出了一种联合训练方案,一次性获得所有指标。它通过并行化原始序列训练过程,显著加速了指标训练过程。有了这些学到的重要性指标,可以将搜索问题表述为一次性整数线性规划问题,形成多模型版本用于边缘部署。

2.3.2 边缘推理服务联合模型与数据适配策略技术

多模型版本和多数据版本联合推理可以带来更大的推理服务优化空间。如图3所示,参考文献[29]差异化的边缘侧节点可部署不同版本的轻量化模型,并相应地调度不同压缩率数据输入。首先,对大模型进行轻量化,形成不同参数规模、算力需求的轻量化模型版本(图3中右侧模型轻量化部分);其次,通过精度延迟联合感知模型部署,选择性地将深度学习模型的不同“版本”放置在不同的云边端位置(图3中中部模型部署);最后,根据差异化的用户请求精度和延迟需求,规划数据版本,并发送到已部署的模型版本上进行

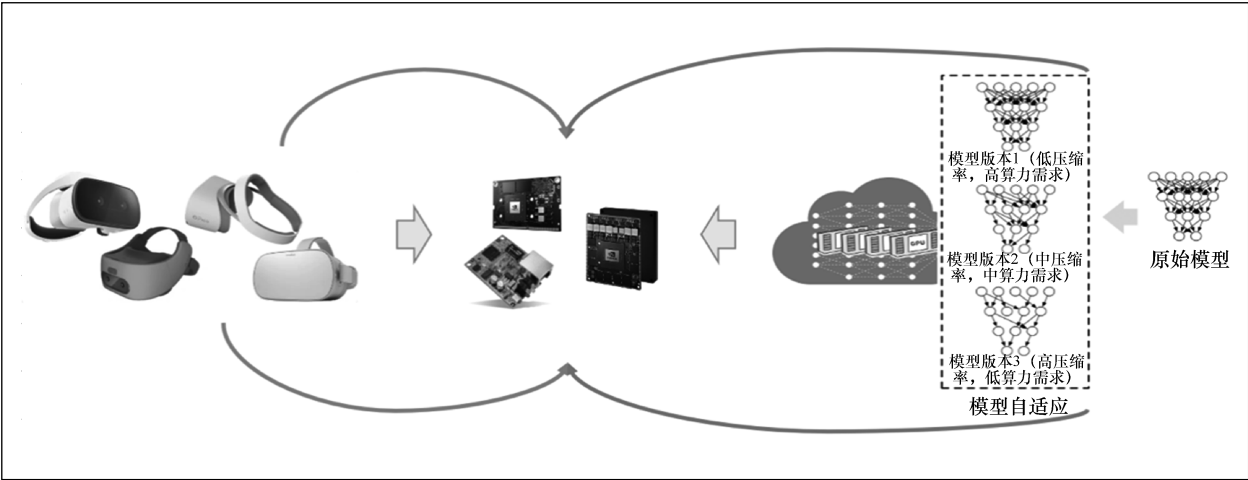


图3 算力受限边缘设备模型切分与部署

推理服务的完成(图3中左侧请求调度)。在以上问题中,可以将模型部署和推理查询服务问题表述为一个规划问题,在满足查询延迟和准确性约束的基础上最小化系统成本,通过联合调整模型复杂度和数据压缩率的决策方法,实现满足不同的延迟和准确性的要求。

2.3.3 元宇宙服务中的内容部署

元宇宙基建随着5G网络、云服务、边缘计算等的逐渐完善愈发蓬勃发展。随着边缘网络技术的采用,元宇宙场景中的内容交付基础设施正在经历从传统的集中式模式到基于边缘的模式巨大变化。通过允许用户从附近的缓存点(如基站)下载内容,边缘内容交付显著减轻了网络主干的负载。由于成本效益和访问延迟的减少,如今越来越多的内容分发网络(CDN)提供商使用边缘缓存来满足不断增长的内容量需求。传统解决方案无法充分处理边缘内容复制。一方面,传统的“被动”缓存替换策略,包括经典的LRU和LFU,其有效性基于用户请求和流行模式的稳定性假设^[30],由于边缘节点的本地流行度的剧烈

变化而无法很好地执行。一个更基本的解释是,只知道本地请求的边缘缓存没有足够的信息来生成对未来请求具有足够适应性的复制策略,尤其是当本地内容请求模式严重违反时。另一方面,使用从所有边缘缓存收集的内容请求信息生成内容复制策略的集中式机制更能够“预测”未来的请求^[31]。

针对越来越频繁使用的边缘设备内容分发基础设施,不仅缓解了集中式服务器的压力,也让用户更接近访问内容。本地节点与邻居节点进行缓存协作的方式,打通了元宇宙交互的地域壁垒,允许边缘缓存参考其他节点的内容复制策略^[32]。首先,通过协同边缘缓存使用深度强化学习处理本地请求,并交换不同节点的流行分布来学习策略。其次,通过设计一种邻居维护机制迭代选择少量邻居节点协作。最后,可以研究深度强化学习通过大量可以迁移的邻居节点知识,帮助本地节点更好地适应动态请求模式^[33]。

综上,可以在需要更高速、大带宽的元宇宙场景中尽可能缩短内容分发网络的用户延迟,达到高效进行高通量低延迟的元宇宙通信传输要求。

3 结束语

传统的众包直播和视频流媒体,以及扩展和虚拟现实应用的发展,使沉浸式多媒体服务和应用兴起。边缘智能可以满足这些应用的要求,比如更低的成本、更低的延迟,以及更好的隐私保护。为了发挥边缘计算的力量,本文综述了边缘智能与沉浸式元宇宙服务的关系、影响面向沉浸式元宇宙服务的边缘智能性能关键因素,以及边缘智能驱动的沉浸式元宇宙核心技术,包括面向元宇宙场景的内容自动生成、多元复杂场景的内容自动生成研究,以及延迟与精度联合感知的边缘推理。本文提供了一种联合模型压缩和数据调度框架,实现自适应深度学习模型部署和数据流,支持边缘内容复制和基于学习的速率自适应^[34]。本文展示了边缘智能和沉浸式多媒体令人兴奋的技术演进:边缘智能技术栈的发展为新型元宇宙服务和应用提供了可能性,包括支持360度和虚拟现实流媒体的新型多功能视频编码、6G网络等,将物理、网络和人类世界融合,有望实现高保真、沉浸式交互、社交和个性化的沉浸式元宇宙新世界。

参考文献:

- [1] JENSEN J F, KJEMS E, LEHMANN N, et al. Virtual space: spatiality in virtual inhabited 3D worlds[M]. [S.l.]: Springer Science & Business Media, 2002.
- [2] ZAHARIADIS T, DARAS P, BALLESTEROS L B. Towards future 3D media internet[EB]. NEM Summit, 2008.
- [3] LAEEQ K. Metaverse: why, how and what[Z]. 2022.
- [4] NING H, WANG H, LIN Y, et al. A survey on metaverse: the state-of-the-art, technologies, applications, and challenges[EB]. arXiv preprint, 2021, arXiv: 2111.09673.
- [5] BALLARDINI R M, HE K, ROOS T. AI-generated content: authorship and inventorship in the age of artificial intelligence[M]//Online distribution of content in the EU. [S.l.]: Edward Elgar Publishing, 2019: 117-135.
- [6] LEE L H, BRAUD T, ZHOU P, et al. All one needs to know about metaverse: a complete survey on technological singularity, virtual ecosystem, and research agenda[EB]. arXiv preprint, 2021, arXiv: 2110.05352.
- [7] FANG Z X, CAI L B, WANG G. MetaHuman creator the starting point of the metaverse[C]//Proceedings of 2021 International Symposium on Computer Technology and Information Science. Piscataway: IEEE Press, 2021: 154-157.
- [8] PAWAR C S, GANATRA A, NAYAK A, et al. Use of machine learning services in cloud[M]//Computer networks, big data and IoT. Singapore: Springer, 2021: 43-52.
- [9] HU C, JIANG J, WANG Z. Decentralized federated learning: a segmented gossip approach[EB]. arXiv preprint, 2019, arXiv: 1908.07782.
- [10] DHELMIM S, KECHADI T, CHEN L, et al. Edge-enabled metaverse: the convergence of metaverse and mobile edge computing[EB]. arXiv preprint, 2022, arXiv: 2205.02764.
- [11] WU C L, WANG Z, SUN L F. PAAS: a preference-aware deep reinforcement learning approach for 360° video streaming[C]//Proceedings of the 31st ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. New York: ACM, 2021: 34-41.

- [12] WANG Y T, SU Z, ZHANG N, et al. A survey on metaverse: fundamentals, security, and privacy[EB]. arXiv preprint, 2022, arXiv:2203.02662.
- [13] XU M, NG W C, LIM W Y B, et al. A full dive into realizing the edge-enabled metaverse: visions, enabling technologies, and challenges[EB]. arXiv preprint, 2022, arXiv: 2203.05471.
- [14] MACH P, BECVAR Z. Mobile edge computing: a survey on architecture and computation offloading[J]. IEEE Communications Surveys & Tutorials, 2017, 19(3): 1628–1656.
- [15] ZHANG J L, CHEN B, ZHAO Y C, et al. Data security and privacy-preserving in edge computing paradigm: survey and open issues[J]. IEEE Access, 2018, 6: 18209–18237.
- [16] 朱锐, 王宏志, 崔双双, 等. 面向元宇宙的云边端协同大数据管理[J]. 大数据, 2023, 9(1): 63–77.
- ZHU R, WANG H Z, CUI S S, et al. Cloud-edge-end collaborative big data management for metaverse[J]. Big Data Research, 2023, 9(1): 63–77.
- [17] 周知, 于帅, 陈旭. 边缘智能: 边缘计算与人工智能融合的新范式[J]. 大数据, 2019, 5(2): 53–63.
- ZHO U Z, YU S, CHEN X. Edge intelligence: a new nexus of edge computing and artificial intelligence[J]. Big Data Research, 2019, 5(2): 53–63.
- [18] NOMA T, ZHAO L, BADLER N I. Design of a virtual human presenter[J]. IEEE Computer Graphics and Applications, 2000, 20(4): 79–85.
- [19] MA G, WANG Z, ZHANG M, et al. Understanding performance of edge content caching for mobile video streaming[J]. IEEE Journal on Selected Areas in Communications, 2017, 35(5): 1076–1089.
- [20] JIANG J Y, LUO Z Y, HU C H, et al. Joint model and data adaptation for cloud inference serving[C]//Proceedings of 2021 IEEE Real-Time Systems Symposium. Piscataway: IEEE Press, 2021: 279–289.
- [21] SHAKER N, TOGELIUS J, NELSON M J. Procedural content generation in games[M]. Cham: Springer, 2016.
- [22] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: an overview[J]. IEEE Signal Processing Magazine, 2018, 35(1): 53–65.
- [23] WANG Z, WU S, XIE W, et al. NeRF: neural radiance fields without known camera parameters[EB]. arXiv preprint, 2021, arXiv: 2102.07064.
- [24] CHOUDHARY T, MISHRA V, GOSWAMI A, et al. A comprehensive survey on model compression and acceleration[J]. Artificial Intelligence Review, 2020, 53(7): 5113–5155.
- [25] CHEN C, YIN Y C, SHANG L F, et al. bert2BERT: towards reusable pretrained language models[EB]. arXiv preprint, 2021, arXiv:2110.07143.
- [26] POLINO A, PASCANU R, ALISTARH D. Model compression via distillation and quantization[EB]. arXiv preprint, 2018, arXiv: 1802.05668.
- [27] TANG C, OUYANG K, WANG Z, et al. Mixed-precision neural network quantization via learned layer-wise importance[C]//Proceedings of European Conference on Computer Vision. [S.l.:s.n.], 2022.
- [28] HE Z L, LI H S, WANG Z, et al. Adaptive compression for online computer vision: an edge reinforcement learning approach[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(4): 1–23.
- [29] LI H S, HU C H, JIANG J Y, et al. JALAD: joint accuracy-and latency-aware deep structure decoupling for edge-cloud execution[C]//Proceedings of 2018 IEEE 24th International Conference on Parallel and Distributed Systems. Piscataway: IEEE Press, 2019: 671–678.

- [30] GOIAN H S, AL-JARRAH O Y, MU-HAIDAT S, et al. Popularity-based video caching techniques for cache-enabled networks: a survey[J]. IEEE Access, 2019, 7: 27699-27719.
- [31] NARAYANAN A, VERMA S, RAMADAN E, et al. DeepCache: a deep learning based framework for content caching[C]// Proceedings of the 2018 Workshop on Network Meets AI & ML. New York: ACM, 2018: 48-53.
- [32] YE J H, LI Z C, WANG Z, et al. Joint cache size scaling and replacement adaptation for small content providers[C]// Proceedings of IEEE INFOCOM 2021 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2021: 1-10.
- [33] ZHOU S J, WANG Z, HU C H, et al. Caching in dynamic environments: a near-optimal online learning approach[J]. IEEE Transactions on Multimedia, 2023, 25: 792-804.
- [34] 朱文武, 王智. 数据驱动的网络多媒体边缘内容分发[J]. 中国科学(信息科学), 2021, 51(3): 468-504.
- ZHU W W, WANG Z. Data-driven multimedia edge network and content delivery[J]. Scientia Sinica (Informationis), 2021, 51(3): 468-504.

作者简介



王智 (1985-), 男, 清华大学深圳国际研究生院副教授, 从事多媒体网络和系统方向研究, 包括高性能数据处理系统优化、分布式机器学习系统、边缘计算与网络等。先后获得2017年教育部自然科学奖一等奖、2018年国家自然科学奖二等奖、2019年深圳市青年科技奖、2020年中国电子学会技术发明一等奖, 入选中组部万人计划。研究成果获得多媒体领域高水平会议ACM Multimedia最佳论文奖、2014年计算机学会优秀博士论文奖。担任*IEEE TMM*副编辑和*ACM TIST*客座编辑。关键技术获得2014年移动互联网创业大赛第一名。研究成果获得工业届广泛应用和关注, 相关成果被知名技术媒体MIT Technology Review、Synced Review等报道。



夏树涛 (1972-), 男, 清华大学深圳国际研究生院教授、博士生导师, 计算机科学与技术研究所所长, 深圳智能语义挖掘技术工程实验室主任。1997年1月博士毕业于南开大学数学系, 2004年1月进入清华大学深圳研究生院, 主要研究方向为编码/量化/压缩、机器学习/计算机视觉、AI安全等。在*IEEE TIT*、*TPAMI*、*TSP*、*ToN*、*TDSC*等国际权威期刊以及NIPS、ICML、ICLR、CVPR、ICCV、ECCV、ISIT等领域高端会议上发表学术论文100余篇。近年来主持完成国家973项目1项、国家重点研发计划项目1项、国家自然科学基金项目6项, 以及深圳市重点科技项目多项。



毛睿 (1975-), 男, 长江特岗学者, 深圳大学特聘教授、博士生导师, 主要研究方向为通用大数据处理。现任深圳大学大数据系统计算技术国家工程实验室副主任, 深圳计算科学研究院执行院长。提出了应对多样性挑战的大数据泛构模式, 建立了基于度量空间的通用大数据管理分析理论框架, 获6项国家、省部级或重要行业组织教学/科研奖励。

收稿日期: 2023-05-30

基金项目: 深圳市科技计划资助项目 (No.JCYJ20220818101014030)

Foundation Item: Science and Technology Plan Project of Shenzhen (No.JCYJ20220818101014030))