基于生成式人工智能的算力网络自智优化研究综述

Self-Intelligent Optimization of Computing Power Networks Based on Generative Artificial Intelligence: A Review

崔佳怡/CUI Jiayi[1]，谢人超/XIE Renchao[1,2]，
唐琴琴/TANG Qinqin[1]
（1. 北京邮电大学网络与交换全国重点实验室，中国 北京 100876；
2. 紫金山实验室，中国 南京 211111）
(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Purple Mountain Laboratories, Nanjing 211111, China)

DOI：10.12142/ZTETJ.202406009
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250106.1155.002.html
网络出版日期：2025-01-07
收稿日期：2024-10-25

摘要：生成式人工智能（GAI）技术可以在多样化业务处理过程中赋予算力网络（CPN）精准的意图分析能力，通过算网专家库的构建进而辅助算力网络实现高效的自适应智能决策，通过模型微调技术使资源配置决策适应突发网络变化，为用户提供精准且稳定的服务。基于上述目标，首先介绍生成式人工智能和算力网络概述，然后讨论了基于生成式人工智能的网络自智优化相关研究进展，创新性提出生成式算力网络的架构，对其核心流程和所需关键技术进行讨论，并对所提架构的优越性进行仿真验证和分析，最后对生成式算力网络应用场景进行分析，期望对该领域的后续研究提出可供借鉴的新思路。

关键词：生成式人工智能；算力网络；意图分析；模型微调

Abstract: Generative artificial intelligence (GAI) technologies can endow the computing power networks (CPNs) with precise intent analysis capabilities in diverse business processing scenarios. By constructing an expert database within the CPNs, it assists in achieving efficient adaptive intelligent decision-making. Through model fine-tuning techniques, the resource allocation decisions can adapt to sudden network changes, providing users with accurate and stable services. Based on these objectives, this paper firstly introduces an overview of GAI and CPNs, then discusses the research progress on network self intelligence optimization based on GAI. A novel architecture for generative computing power networks is proposed, along with discussions on its core processes and necessary key technologies. Furthermore, the superiority of the proposed architecture is validated and analyzed through simulation. Finally, an analysis of the application scenarios of generative computing power networks is provided, aiming to propose new perspectives for subsequent research in this field.

Keywords: generative artificial intelligence; computing power network; intentional analysis; model fine-tuning

引用格式：崔佳怡, 谢人超, 唐琴琴. 基于生成式人工智能的算力网络自智优化研究综述 [J]. 中兴通讯技术, 2024, 30(6): 54-62. DOI: 10.12142/ZTETJ.202406009
Citation：CUI J Y, XIE R C, TANG Q Q. Self-intelligent optimization of computing power networks based on generative artificial intelligence: a review [J]. ZTE technology journal, 2024, 30(6): 54-62. DOI: 10.12142/ZTETJ.202406009

随着人工智能的快速发展，越来越多的领域开始应用人工智能这一技术，如自然语言处理、计算机视觉等领域。在众多新兴人工智能应用中，生成式人工智能（GAI）作为其中的一个重要分支，在近年来取得了迅猛发展，它能够在几秒钟内生成高质量的内容，并根据用户的需求提供个性化的内容[1]。在新型人工智能技术的支持下，算力网络多种基础功能如任务分配、数据存储、计算处理等方面得到进一步优化，算力网络的应用场景也得到了不断拓宽。当前对于算力网络的研究正处于与新兴技术广泛融合的关键时期，生成式人工智能将促进算力网络的进一步发展，该技术可以自动化部署和管理算力网络中的异构资源，例如根据不同的任务特性和资源状态，动态地分配计算资源，优化计算路径，提高算力网络的运行效率和性能[2]。凭借对意图的精确感知和对海量数据的分析能力，生成式人工智

基金项目：国家自然科学基金项目（92367104）

能可应对多元化的算力场景和复杂化的业务需求，在匹配用户多维度需求方面为算力网络提供了更加智能高效的网络服务策略定制方案。

当前中国正在积极推动生成式人工智能和算力网络相关建设。2023年2月，中共中央、国务院发布《数字中国建设整体布局规划》，系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局。2023年7月，中央网信办等七部门发布《生成式人工智能服务管理暂行办法》，首次明确了各方面的法定责任及法律依据，确立了人工智能产品的安全评估规定及管理办法。总之，数字经济已成为驱动中国经济发展不可或缺的力量，“网络赋能智能，智能使能网络”的创新发展已经逐渐成为推动国家数字化转型的重要力量。

本文通过对生成式人工智能和算力网络的系统调研，提出一种面向用户意图的生成式算力网络架构，并对其核心流程和关键技术进行探讨，随后进行仿真验证分析，最后对应用场景和未来发展方向进行分析展望，以响应生成式人工智能服务快速增长的算力需求，优化算力网络的整体机制，推动智能和网络的高效融合。

1 生成式人工智能与算力网络研究现状

在研究面向意图的生成式算力网络进行前需要明确相关概念，因此本节对生成式人工智能和算力网络相关概念和研究现状进行简要介绍。

1.1 生成式人工智能与算力网络概述

1）生成式人工智能

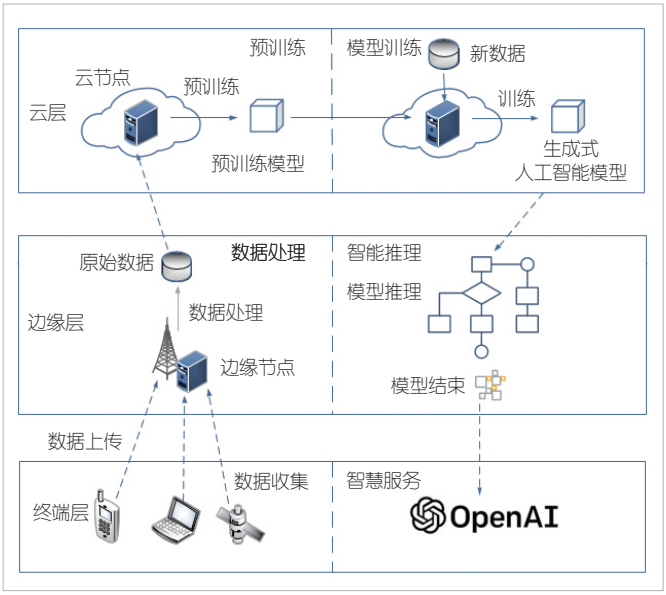
生成式人工智能是一种利用人工智能算法创造性地生成、操纵和修改有价值及多样化个性化数据的自动化方法[3]。生成式人工智能提供信息的过程不需要用户参与。在AI模型训练完成后，用户只需提供任务描述等输入，即可高效获取生成的内容。因为超高的生产内容效率，生成式人工智能逐渐成为新型网络的重要支撑工具。本节将介绍生成式人工智能最典型的服务架构和关键技术。

生成式人工智能架构以云–边–端三层网络架构为主。提供生成式人工智能服务的流程主要分为6个部分，包括数据收集、数据处理、模型预训练、模型训练、智能推理、智慧服务，如图1所示。该架构的核心在于根据任务的复杂性和数据规模，选择合适的模型架构来进行预训练，随后用特殊场景下的数据的分布规律对预训练模型进行再次更新，这个过程称为“微调”。生成式人工智能使用的预训练模型主要包括变分自编码器、生成对抗网络、扩散模型、Transformer等以及在这些基本模型上进行改进的架构。

生成式人工智能最初的应用面向人类与计算机的交互，针对人类需求自动化生成多媒体应用服务，如文章、音乐和图片等内容。随着研究范围的进一步扩大，多模态生成式人工智能技术赋予了基础大模型针对不同学科的强大知识和理解能力。当前全球针对生成式人工智能在网络当中应用的研究也更加深入，利用基础大模型能够生成各类面向系统的内容，辅助网络进行规划设计或资源管理等，对网络管理模块进行智能升级。例如：在生成式人工智能驱动网络架构的研究中，文献[4]介绍了生成式人工智能利用提示词工程将下游任务和生成式人工智能知识相关联，通过不同的提示策略优化生成质量，并使用离散提示和连续提示实现系统自动化。文献[5]主要考虑使用生成式人工智能作为网络组件或者增强网络功能，利用网内多模态数据，通过大语言模型（LLM）训练实现面向网络和任务的专业生成能力，例如网络协议、网络配置的设计和资源分配、链路管理策略的设计等。在感知通信数据处理方面，文献[6]提出生成式人工智能也不断增强数据集的多样性，利用有限的真实数据提高人类活动检测的精度，且可以解决相位模糊、信号到达方向估算等复杂问题。在利用信号频谱数据的训练中，生成式人工智能方法的性能明显优于深度强化学习方法。

2）算力网络

算力网络的核心思想是将分布的计算节点连接起来，动态实时感知计算资源和网络资源状态，进而统筹分配和调度计算任务，形成一张计算资源可感知、可分配、可调度的网



▲图1　生成式人工智能服务架构

络，满足新业务新应用对算力的要求。凭借其泛在算力按需分配的特点，算力网络已经成为驱动各行各业变革的重要解决方案。

当前全球针对算力网络的学术研究主要围绕算力资源建模、感知和调度三大类问题。对于算力资源建模，算力网络需要解决的最基本问题是如何衡量底层异构的算力资源质量和大小，并对其进行合理的表征和度量。近年来算网资源度量领域的研究者们对网络各种资源进行了全面度量分析[7-8]，兼顾算力资源的基础性能以及算力节点的工作状态，但随着智能的发展，需要引入更加自动化、更加智能的方法来提升度量的精确性。对于算力资源感知来说，基础设施层的算力资源庞大且分散，算力资源需要对计算任务进行按需匹配，感知机制的发展使得广泛的算力能够得到充分的调配协同，但目前的感知机制仍然无法解决跨域跨层级异构算力的全面感知[9-10]。对于算力网络协同调度来说，网络需要合理地对任务进行分配以及动态地检测和平衡运行中的节点，根据计算任务的要求，结合实时的计算负载和网络状态条件，动态地将计算任务调度到最匹配的边缘计算节点，实现对算力资源的协同利用和调度。尽管当前的算网调度机制已经非常完善，但在面对算网突发情况时，例如大规模节点环境的变化，仍难以做出即时决策[11]。

1.2 生成式人工智能与网络自智优化

生成式人工智能在网络中有显著优化能力，体现在自动学习、生成和改进网络相关的组件，以实现更加精准的网络服务响应。全球的相关研究主要集中在通过大模型训练实现面向网络和任务的专业能力。本文主要介绍以下几个方面：

1）网络模型应用

基于大模型的微调能力可以面向多种任务场景训练出针对特定场景问题的生成式解决方案。近年来，大语言模型如GPT-4和Llama-3等逐渐应用到文本解析、对话生成等多种自然语言处理任务中。在生成式人工智能对于网络自智优化的场景中，基于大语言模型的语义分析、内容生成、上下文学习等能力，通过少量数据微调训练能够捕捉专业化的语义关系和复杂的数据模式，实现面向网络和任务的专业生成能力。基于大语言模型设计的无线网络大语言模型能够解决正交频分复用（OFDM）系统的功率分配问题、无线网络的频谱感知问题、网络协议理解问题等，不仅可以突破大语言模型在无线通信中遇到的固有局限，还显著提升了其处理无线通信问题的能力[12]。除无线网络大语言模型之外，越来越多的定制化大模型逐渐支持6G客户端业务，通过数据、知识驱动的分布式协同部署和微调适配，在边缘侧充分发挥基站、边缘云的潜力，实现定制化大模型支撑多样个性化6G客户端业务。

2）网络自主设计

随着大规模网络的发展，第3代合作伙伴计划（3GPP）、美国电气电子工程师学会（IEEE）和国际电信联盟（ITU）等组织发布各种协议、标准和规范来确保设备之间的可靠高效通信。然而，协议的多样化和复杂性增加了网络对协议应用的困难性，并且制订的网络协议往往缺少自适应匹配网络环境的能力。生成式人工智能模型凭借突出的数据理解能力，可以快速获取与无线网络协议相关的信息。因此，利用生成式人工智能模型可以设计出更加智能的路由协议，根据网络流量的变化自动调整路由路径，以优化带宽使用和减少延迟，增强系统整体的链路性能[13]。除此之外，在大型网络环境中手动配置设备是非常耗时且容易出错的工作，借助生成式人工智能模型可以将这一过程变得自动化、智能化。系统可通过学习最佳实践和历史配置数据来生成合适的配置文件。

3）网络自智操作

生成式人工智能模型能够借助相关技术分辨和模拟复杂数据模式，更加智能地感知预测网络状态。该技术广泛应用于网络资源分配、链路管理策略生成等网络自智操作场景。因此，可以利用生成式人工智能模型学习网络历史流量模式，并预测未来流量，从而提前做出路由决策，自适应地分配带宽生成转发策略，甚至在实际数据缺乏时填补数据空白[14]。在某些应用场景下，比如视频流媒体或在线游戏，网络延迟和丢包率需要严格控制。对此，生成式人工智能模型可以用来优化服务质量（QoS）参数，确保优先级较高的流量获得更好的服务保障。除了这些资源分配策略生成的应用外，生成式人工智能还能够根据不断变化的网络条件和用户行为调整激励机制。例如，在混合现实（MR）场景中，生成式人工智能与契约理论等技术的结合能有效激励全双工设备对设备语义信息共享，避免重复的计算任务，以解决计算能力受限的问题。

2 生成式算力网络方案设计

为了应对算力网络多种应用场景下用户意图的差异化和个性化带来的新挑战，基于生成式人工智能技术和算力网络的研究现状，本节中我们提出面向用户意图的生成式算力网络，并给出生成式算力网络的定义。生成式算力网络在算网深度融合的基础上，更加关注用户意图的适配和算网环境的敏感变化，增强算网决策的智能生成能力和灵活适应能力，旨在精准满足用户意图的同时保障服务过程的稳定性，包括决策生成、模型更新、环境突变响应等功能。本节将介绍生

成式算力网络的设计动机，以及生成式算力网络的基础架构、核心流程和关键技术。

2.1 网络架构

生成式算力网络旨在根据生成式人工智能的经验、对算力网络中的当前基础设施资源状态和用户意图的感知，动态生成算力网络中网算存资源分配的最佳策略，实现“意图–策略”的高效匹配，同时实现网络内资源的灵活调度和协同，保障其全生命周期的安全性和可靠性，以提供高质量高可靠的生成式人工智能应用服务。生成式人工智能融合算力网络架构设计为3个层面，如图2所示，包括基础设施层、感知决策层、应用服务层，每一层实现功能具体如下：

1） 基础设施层

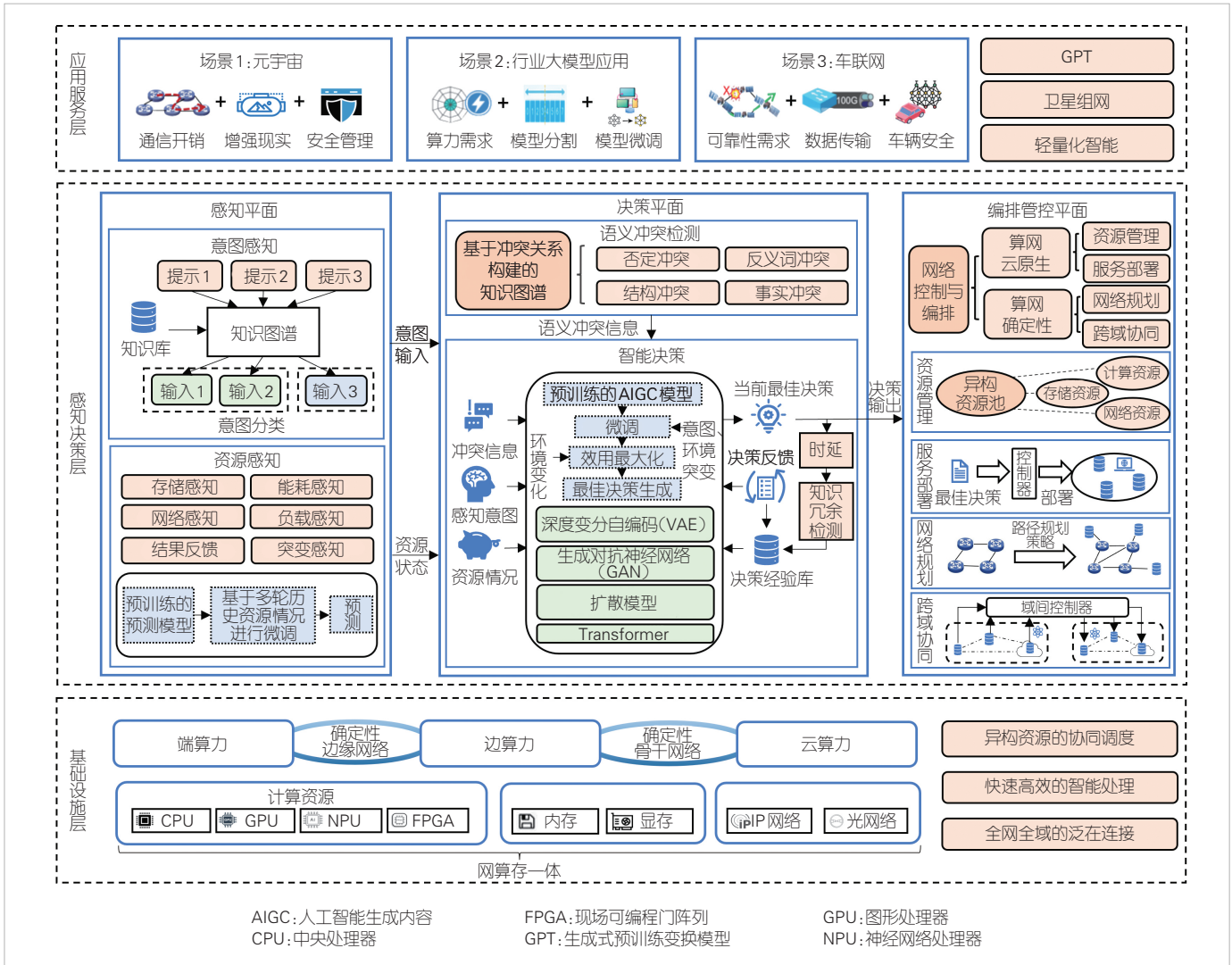
基础设施层为生成式算力网络提供了全网广泛的网算存资源。基础设施层中包含分布式算力节点和广泛的计算资源，算力节点包括端算力节点、边算力节点和云算力节点，这些算力节点部署基础算力、智能算力、超级算力等多种计算资源。确定性网络连接算力节点，实现全网全域的泛在连接。生成式算力网络将网络内的网算存资源整合统一，协同调度异构资源来高效地协同处理不同的任务，更好地响应用户需求，具备快速高效的智能处理能力。

2）感知决策层

感知决策层是生成式算力网络架构的核心。根据该层的功能，该层可分为3个平面：感知平面、决策平面、编排管控平面。

（1）感知平面

感知平面部分接收来自基础设施层和应用服务层的信息。感知平面从应用服务层获取用户的意图，用户意图往往



AIGC：人工智能生成内容　　FPGA：现场可编程门阵列　　GPU：图形处理器
CPU：中央处理器　　GPT：生成式预训练变换模型　　NPU：神经网络处理器

▲图2　面向用户意图的生成式算力网络架构

包含于语义信息中，感知平面主要应用基于知识图谱的意图转译方法。除此之外，感知平面实时接收基础设施层上传的资源状态，包括“网、算、存”资源的使用情况和负载均衡情况，以便后续根据用户意图和资源情况进行协同考虑，在完成最优决策部署的同时最大化资源的利用率。为了灵活应对环境资源产生的多种敏感变化，提高生成决策性能的稳定性，感知平面对资源的感知不能只停留在当前时间状态，而是需要根据历史资源状态的存储记录对未来某一时刻资源状态进行预测，以适应决策生成过程中时延带来的环境改变。同时，感知平面也感知任务的执行状态，检查决策是否执行以及其性能优劣。此外，感知平面还具备监测环境突变（如基础设施节点大范围的减少或增加）的能力，及时发现对用户服务不利的情况以做出灵活应对，保持决策的可用性，提供服务的持久稳定性。

（2）决策平面

决策平面是整个生成式算力网络工作的关键，其核心能力是使用生成式人工智能相关技术，基于感知平面输入的用户意图、接收到的当前和未来预测资源生成最优策略，为用户目标定制个性化资源调度策略。决策平面引入“以意图为中心”的算网专家库，它相当于一个历史记录存储库，记录生成式算力网络中所有的用户意图部分信息和与之匹配的最终决策模型的参数。利用用户意图之间的相似性，算网专家库为新用户意图提供有效的预训练生成式人工智能模型选择空间，减少重新训练模型的时间。算网专家库辅助选择的预训练生成式人工智能模型，不能够完全满足用户个性化的意图需求，因此需要根据用户意图对预训练模型进行适应性微调，使模型的性能和损失逐渐符合用户的期望，同时通过不断与资源环境反馈状态进行实时交互，从而不断更新模型参数，更好地适应环境。除此之外，预训练生成式人工智能模型的微调不止发生在用户意图输入的时刻，还发生在环境大规模突变的时刻，以适应环境输入维度的变化，改变模型使其能够生成更加合理的决策，保障用户服务执行过程中不受环境突变的影响。

（3）编排管控平面

编排控制平面是决策的执行者，它根据最优决策对网算存资源进行一体化管理，部署最佳决策内容到基础设施层的节点执行，调度分配、管理算力资源和规划网络，同时控制跨域协作，从而可以实现多域算力共享和域间服务协作。

3）应用服务层

生成式算力网络的应用服务层不仅能够满足更多智能应用的需求，如元宇宙、GPT、车联网等，还能提供生成式人工智能服务，如文本、图像、视频等内容，并且支持内容根据用户意图自适应调整等功能，以不断满足用户新需求。

2.2 核心流程

针对生成式算力网络的核心部分，即决策平面，分析决策产生的主要工作流程，包括决策生成、模型更新和环境突变响应。

1）决策生成

决策生成流程根据当前资源状态生成最优策略。基础设施层进行语义分割，度量每个节点的资源状态，包括存储、能耗、网络、负载和历史策略执行性能，然后将输入传输到感知决策层。在感知平面上，基于知识图谱将语义分割生成的文本划分为不同的意图组，同时感知资源状态和决策反馈。接下来，决策平面根据知识图中不同单词之间的关系，检测意图组之间的语义冲突。同时，利用预测模型，根据之前的输入，预测未来短时间内资源的状态。将用户意图在经验库中进行搜索和近似匹配，调取预训练生成式人工智能模型。使用用户意图对预训练生成式人工智能模型进行小样本微调，然后将资源预测结果、环境冲突信息、用户意图、当前资源状态等信息反馈给核心预训练生成式人工智能模型生成最优决策，同时进行模型更新和环境突变响应以实现决策效用最大化。收到决策平面响应后，系统将决策发送到编排控制平面，控制器根据该决策管理部署节点，执行任务调度。最后，基础设施层中的节点接收部署的决策并执行，完成用户服务后反馈结果信息。

2）模型更新

模型更新过程包括两部分，这两部分分别根据来自应用服务层的用户意图和来自基础设施层的资源反馈进行模型参数的更新。用户意图对从算网专家库中选取的预训练生成式人工智能模型进行微调，使其更满足当前用户的需求，以获得用户意图的最佳匹配决策。这比直接使用预训练生成式人工智能模型更准确。同时，模型在决策生成与执行完成的过程中，不断与基础设施层反馈的资源环境以及执行过程中反馈的结果进行交互，并不断更新模型参数，使模型适应环境敏感变化，保障生成内容的有效性。执行完决策之后，算网专家库确认最终决策，将最终模型添加到专家库中，同时将时间较远的模型参数删除，以减少专家库的数据冗余。

3）环境突变响应

环境突变响应是一种发生于环境急剧变化的决策模型适应过程。发生环境突变主要有以下几种情况：（1）基础设施层中负载不均衡，造成大量节点过载无法使用，空闲节点没能得到充分利用；（2）网络规模的扩大和缩小，例如算力供应商的加入和退出造成大量节点的增删情况，资源环境维度

发生变化。针对第一种情况在基础设施层设置负载阈值，以判断整体网络环境内所有决策的综合性能。如果过载节点数大于阈值，感知决策层将接收到相关信息，将目前正在执行的决策终止，根据当前信息立即制定新的策略。针对第二种情况，需要对当前执行模型和算网专家库中的模型全部进行微调，使其迅速适应大范围的网络变化。

2.3 关键技术

本小节将从生成式算力网络架构和流程中的几种关键支撑技术展开探讨，包括意图感知和转译技术、预测和决策生成模型、微调技术等。

1）意图感知和转译技术

在基础设施层，使用包括解码器网络的语义分割将用户的需求划分为文本提示。然后，意图感知部分利用知识图谱对文本提示进行分类。知识图谱用于表示实体之间的关系和属性，以及其语义信息。首先基于开放关系抽取[15]等方法可以通过分析用户输入的语句，从非结构化文本中提取开放领域的关键实体、特征和关系信息，并将这些信息映射到知识图谱中相应的节点和边。目标本身固有的属性信息是用于目标意图分析的参数，如网算存资源参数和网算存节点等。以各种特征为基本节点的节点信息主要包括距离、能耗、资源利用率等特征参数，而进行意图分析的节点关系主要包括目标实体意图间的包含关系、关系值域、关系约束等。借助一系列的知识抽象、知识推理、构建上述知识图谱，并通过反馈实时更新数据库，系统可实现更加精确的目标意图知识图谱。利用知识图谱即可进行意图的分析和转译，获得意图的各种属性、特征和关系解析结果。

2）预测和策略生成模型

在生成决策前，系统可通过构建神经网络对未来资源状态进行预测，输入资源的当前状态并估计资源的后续状态，以减少时间延迟对策略性能的影响。决策生成模型主要基于生成式人工智能模型。生成式人工智能模型包括VAE、GAN、扩散模型、Transformer等经典内容生成模型，提供动态自适应调度方案[16]，提升系统的智能水平。除此之外，也可采用强化学习算法对流程进行优化，例如Actor-Critic结构和经验缓冲区。

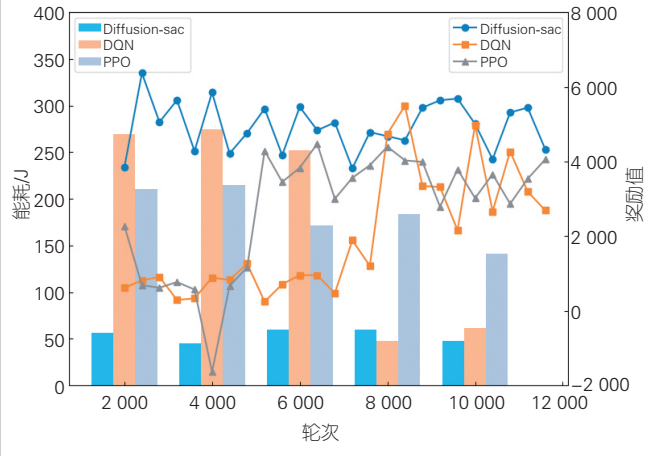
3）微调技术

参数高效微调是指微调少量或额外的模型参数，固定大部分预训练模型参数，从而大大降低计算和存储成本，同时也能实现与全量参数微调相当的性能。参数高效微调方法甚至在某些情况下比全量微调效果更好，可以更好地泛化到域外场景。高效微调技术可以粗略分为以下三大类：增加额外参数（例如AttentionFusion）、选取一部分参数更新（例如BitFit）、引入重参数化（例如LoRa）。其中，增加额外参数这类方法又主要分为类适配器（Adapter-like）方法和软提示（Soft prompts）两个小类[17]。如果环境中的突变对决策的性能产生较大影响，就需要使用微调技术来调整模型。此外，为了满足用户输入意图，对从算网专家库中复制的预训练模型也要进行微调，基于新的数据集更新预训练模型部分参数以适应新的任务。

2.4 仿真验证

本文中，我们通过仿真实验对生成式算力网络架构以及扩散模型集成的强化学习算法进行简单验证，并对3种算法的奖励值进行仿真：集成扩散模型的强化学习算法（Diffusion-sac）、深度Q网络（DQN）和基于策略梯度的强化学习算法（PPO）。对生成式算力网络的智能生成策略机制进行仿真验证主要分为两部分：1）选择每轮训练的奖励值和网络能耗作为系统性能指标，验证生成式算力网络和普通算力网络的策略生成性能；2）对不同用户算力需求下系统的表现性能和稳定性进行仿真时延，以证明其在不同场景下的适应能力。我们使用扩散模型作为Actor网络核心算法，对于实时变化的网络环境和用户需求做出算力的分配策略，关注3种学习方法的奖励值曲线和网络整体的能耗指标。

在策略生成性能方面，如图3所示，Diffusion-sac算法学习奖励值的曲线平稳且高于其他算法，在任务处理能耗方面也使得网络始终具备最低能耗。这些结果证明了Diffusion-sac算法在高性能、低能耗和快速收敛方面的优越特性。因此，Diffusion-sac算法是解决算力网络场景策略生

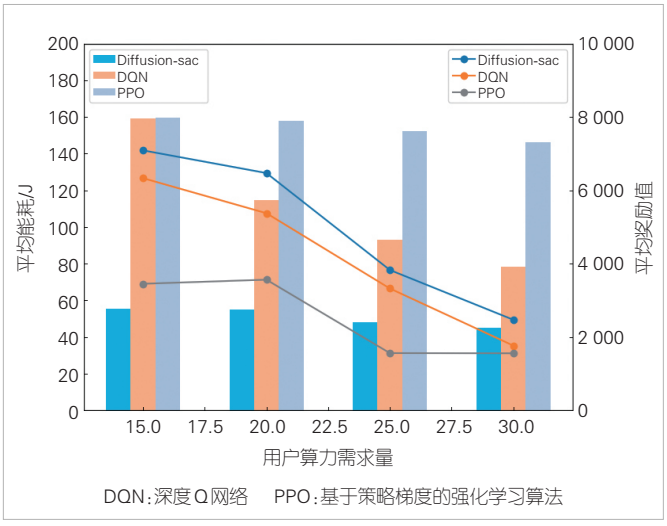


DQN：深度Q网络　PPO：基于策略梯度的强化学习算法

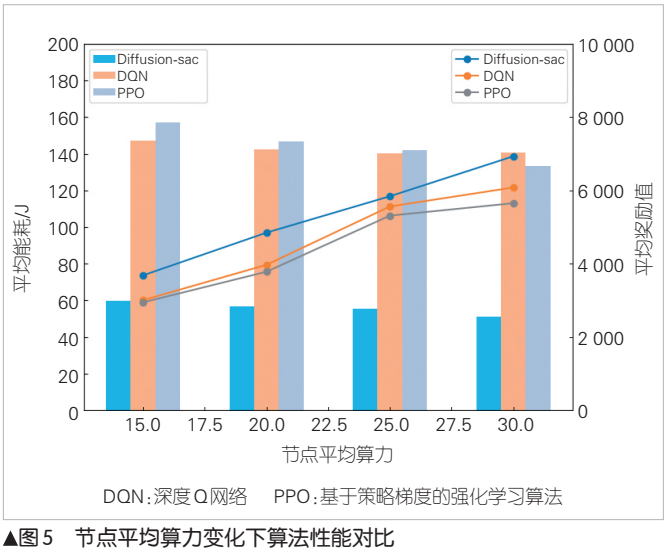
▲图3　3种算法奖励值和能耗对比

成问题的首选。

在系统多场景下的稳定性方面，图4和图5显示了在不同的用户算力需求以及不同节点算力容量情况下Diffusion-sac算法的稳定性。可以看出，随着用户算力需求量的增加，3种算法的平均奖励值有所下降，这是因为节点资源承受的资源供给压力较大，但是Diffusion-sac算法始终保持着最高的平均奖励值，并且能耗值也远小于其他算法。由节点平均算力的增加可以看出，3种算法的奖励值都有所提升。与此同时，系统的整体平均能耗有所下降，这表明系统具备负载均衡的能力，不会出现能耗过高情况，而且Diffusion-sac算法仍然保持着最优效果。实验证明，Diffusion-sac算法不仅在单一场景下具备优秀的训练能力，并且在变化的负载条件下也能够适应场景的变化进行最优决策，具备强大的泛化能力和稳定性能。



▲图4　用户算力需求量变化下算法性能对比



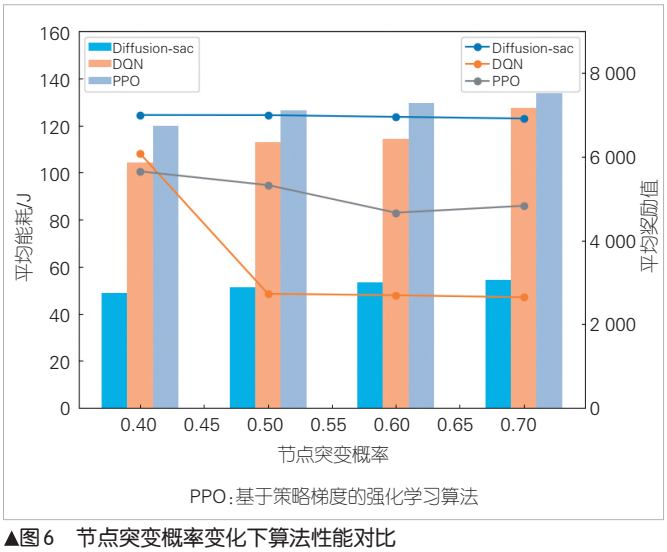
▲图5　节点平均算力变化下算法性能对比

除此之外，我们还进行了针对模型泛化能力的仿真验证。图6展示了改变环境节点突变概率时3种算法的表现情况。我们设置节点突变概率变量，用于控制算网环境中每个节点的崩溃概率。若节点发生崩溃，则任务调度过程中需要考虑其他节点情况。其中，3种算法的奖励值整体趋势都随着节点突变概率的提升而下降。这是因为可用节点数量越少，任务处理和调度决策困难就越会导致节点负载不均匀，奖励值就越下降。由图6可知，Diffusion-sac算法的奖励值变化幅度较小，且始终高于其他两种传统算法，这显示了Diffusion-sac算法的稳定性能。3种算法产生的能耗整体趋势都随着节点突变概率的提升而有所提升，这是因为节点负载不均衡导致部分节点使用过载，能耗极具增大。其中，能耗最小的算法为Diffusion-sac，在每种情况下都表现为最小，这说明该算法能够尽可能在环境突变的情况下保持负载均衡的决策。

3 生成式算力网络应用场景

3.1 生成式算力网络服务于垂直行业

大模型的应用对于各专业场景都是新兴的发展趋势，但训练大模型对大多数企业来说不现实，因为这些企业往往只专注于一类业务，更希望利用大模型在行业场景的个性化智能能力。企业的需求是支付尽可能低廉的成本，而不是付出巨大算力直接进行大模型训练。总结来说，当前垂直行业应用生成式人工智能技术存在以下问题：1）由于大模型训练需要大量的算力资源，目前只有云端能够支撑大模型训练方案，边缘算力仅依靠协同计算无法支撑大模型训练，绝大多



▲图6　节点突变概率变化下算法性能对比

数用户和企业没有能力训练自己的大模型；2）垂直行业设备大量分布在边缘端，因此生成式人工智能的服务交互需要向云端请求并获得结果反馈，通信开销巨大；3）目前现有的大模型大多数是通用模型，不符合垂直行业的专业需求。

生成式算力网络为垂直行业训练属于自己的大模型以提供算力条件支撑。网络基础设施层结合多种分布式智能技术，使得预训练大模型在边缘为各专业领域提供个性化服务成为可能，大大减少了实现个性化智能能力的复杂性。文献[18]中基于给定主题的几张图像，将这几张图片的特征嵌入到模型的输出域中，并对预训练的基于扩散的文本到图像框架进行微调。这样的解决方案不仅能够满足企业训练模型需求，也保证了智能资源的充分使用，解决高昂的成本问题。

3.2 生成式算力网络实现服务个性化定制

传统算力网络根据业务需求，通过移动边缘计算等技术将计算资源、存储资源以及网络资源进行集中和灵活调度，以实现按需的算力分配和灵活调度。由于不同的业务需求需要不同的计算、存储和网络资源，而且同一业务的不同阶段也可能需要不同的资源分配，因此，算力网络需要提供个性化的服务，以满足不同业务需求和用户需求，同时对资源进行高效利用，避免资源的浪费。但随着智能业务的海量涌现和发展，传统的静态资源管控调度方案面临灵活性差、适配度低等问题。

在生成式算力网络中，生成式人工智能算法会根据用户的行为和偏好，生成个性化的服务内容和推荐。例如，网络可以根据用户的浏览历史、购买记录、搜索关键词等信息，生成个性化的推荐列表，帮助用户更快速地找到自己感兴趣的内容。同时，算力网络还可以根据用户的反馈和评价，不断优化和改进个性化服务。例如，如果用户对某个推荐的服务或商品不满意，融合网络可以根据用户的反馈信息，重新调整算法模型，提高个性化推荐的准确性和质量[19]。

3.3 生成式算力网络驱动虚拟世界扩展

元宇宙作为未来虚拟世界的起点已经获得了极大的关注，然而数据的映射是在虚拟世界和现实世界之间建立共生互联网的先决条件。生成式人工智能技术通过利用人工智能的力量来自动化信息创建过程，为快速创建数字内容提供了技术支持。但随着虚拟世界的逐渐扩展和演进，虚拟世界用户数据量的增多带来大模型广泛访问的问题。生成式算力网络的出现能够为大模型提供缓存空间，解决访问频繁问题。

通过生成式算力网络可以实现跨平台一体化的元宇宙体验。无论是对于虚拟现实设备、增强现实设备，还是针对智能手机等移动终端，算力网络承载的生成式人工智能都可以提供一致且无缝的用户体验，使用户可在不同设备上连续获得元宇宙服务。算力网络不仅能为用户提供娱乐环境，还能帮助用户实现元宇宙的任务管理、资源调配、时间规划等，进而提高用户的工作效率[20]。

4 生成式算力网络未来发展

4.1 激励机制

在生成式算力网络中，生成式人工智能服务的整个生命周期需要对参与者进行适当的激励。由于网络中参与生成式人工智能服务的节点提供异构资源，在提供服务的过程中，数据收集、预训练、调优和推理都需要大量异构资源参与。这就需要对服务的各方参与者按照其贡献设计合理的激励机制，例如引入区块链技术，实现网络对生成式人工智能服务全生命周期的去中心化管理。用户可以根据服务提供商的交易历史来评估其声誉，从而促进服务的优化和改进。

4.2 服务定制

在生成式算力网络中，生成式人工智能模型的预训练、微调和推理通常会消耗大量的计算和网络资源。因此，我们可以更加关注生成式算力网络的绿色运营，以最小的能耗和碳排放提供生成式人工智能服务。此外，还可以提出智能资源管理和调度技术来平衡服务质量和资源消耗。

4.3 模型压缩

随着生成式人工智能大模型变得越来越复杂庞大，在提供生成式人工智能服务时，模型压缩技术对于减少服务延迟和资源消耗变得越来越重要。目前已经研究出的模型压缩技术包括：修剪、量化和知识蒸馏技术。其中，修剪技术的目的是去除不重要的权重，而量化则降低了权重的精度，知识蒸馏目的是训练一个较小的模型来模仿较大模型的行为。未来对于模型技术的研究可以基于上述技术进行完善，以平衡模型的大小和精度。

4.4 服务安全隐私

为了提供保护隐私的生成式人工智能服务，在模型训练和推理中都需要考虑隐私计算技术。差分隐私、安全多方计算和同态加密等技术可用于保护敏感数据并防止未经授权的访问。其中，差分隐私涉及在数据中添加噪声以保护个人隐私，安全多方计算允许多方在不向彼此透露其输入的情况下计算一个函数，同态加密允许在不解密的情况下对加密数据

执行计算。

5 结束语

由于生成式人工智能对用户需求高效快速响应，在任务处理方面带来显著的性能提升，同时算力网络能够为生成式人工智能提供非常充足的计算条件，因此在算力网络中引入生成式人工智能的新型网络架构具备充分的条件。本文对算力网络的概念进行详细介绍，并系统梳理了生成式人工智能的发展现状，重点对生成式人工智能融合算力网络的系统架构、通信流程及关键技术进行概述，最后通过仿真实验验证了生成式算力网络架构的合理性和有效性。在未来的工作中，我们可以对生成式算力网络架构每一层进行详细的研究，以期提供更加完善高效的智能算力服务。

参考文献

[1] LEE L H, LIN Z J, HU R, et al. When creators meet the metaverse: a survey on computational arts [EB/OL]. [2024-10-25]. http://arxiv.org/abs/2111.13486

[2] 彭开来, 王旭, 唐琴琴. 算力网络资源协同调度探索与应用 [J]. 中兴通讯技术, 2023, 29(4): 26-31. DOI: 10.12142/ZTETJ.202304006

[3] HARSHVARDHAN G M, GOURISARIA M K, PANDEY M, et al. A comprehensive survey and analysis of generative models in machine learning [J]. Computer science review, 2020, 38: 100285. DOI: 10.1016/j.cosrev.2020.100285

[4] LIU Y Q, DU H Y, NIYATO D, et al. Optimizing mobile-edge AI-generated everything (AIGX) services by prompt engineering: fundamental, framework, and case study [J]. IEEE network, 2024, 38(5): 220-228. DOI: 10.1109/MNET.2023.3335255

[5] BARIAH L, ZHAO Q Y, ZOU H, et al. Large generative AI models for telecom: the next big thing? [J]. IEEE communications magazine, 2024, 62(11): 84-90. DOI: 10.1109/MCOM.001.2300364

[6] ZHENG J K, ZHANG J Y, DU H Y, et al. Flexible-position MIMO for wireless communications: fundamentals, challenges, and future directions [J]. IEEE wireless communications, 2024, 31(5): 18-26. DOI: 10.1109/MWC.011.2300428

[7] 杜宗鹏, 李志强, 陆璐. 算力网络四面三级算力度量技术体系 [J]. 中兴通讯技术, 2023, 29(4): 8-13. DOI: 10.12142/ZTETJ.202304003

[8] 李重严, 毕成, 张晟. 面向信息能源融合的低碳算力网络架构研究 [J]. 电信工程技术与标准化, 2022, 35(11): 1-6. DOI: 10.3969/j.issn.1008-5599.2022.11.001

[9] 闫实, 彭木根, 王文博. 通信-感知-计算融合: 6G愿景与关键技术 [J]. 北京邮电大学学报, 2021, 44(4): 1-11. DOI: 10.13190/j.jbupt.2021-081

[10] 许胜, 许方敏, 赵成林. 基于数字孪生的算力网络自优化技术研究 [J]. 中兴通讯技术, 2023, 29(3): 46-50. DOI: 10.12142/ZTETJ.202303009

[11] 衷璐洁, 王目. 区块链赋能的算力网络协同资源调度方法 [J]. 计算机研究与发展, 2023, 60(4): 750-762

[12] SHAO J W, TONG J W, WU Q, et al. WirelessLLM: empowering large language models towards wireless intelligence [J]. Journal of communications and information networks, 2024, 9(2): 99-112. DOI: 10.23919/JCIN.2024.10582827

[13] 任天骐, 李荣鹏, 张宏纲. 通信网络与大模型的融合与协同 [J]. 中兴通讯技术, 2024, 30(2): 29-36. DOI: 10.12142/ZTETJ.202402005

[14] LIU Y Q, DU H Y, NIYATO D, et al. Deep generative model and its applications in efficient wireless network management: a tutorial and case study [J]. IEEE wireless communications, 2024, 31(4): 199-207. DOI: 10.1109/MWC.009.2300165

[15] DAI H, DU D, LI X, et al. Improving fine-grained entity typing with entity linking [J/OL]. [2019-09-26]. https://arxiv. org/abs/1909.12079

[16] SU N. Research on multiparty participation collaborative supervision strategy of AIGC [C]//Proceedings of IEEE 13th International Conference on Electronics Information and Emergency Communication (ICEIEC). IEEE, 2023: 268-272. DOI: 10.1109/ICEIEC58029.2023.10200392

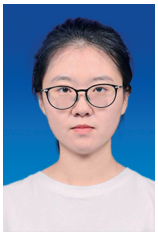
[17] 丁鑫, 邹荣金, 潘志庚. 基于高效参数微调的生成式大模型领域适配技术 [J]. 人工智能, 2023(4): 1-9

[18] RUIZ N, LI Y Z, JAMPANI V, et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 22500-22510. DOI: 10.1109/CVPR52729.2023.02155

[19] DU H Y, LI Z H, NIYATO D, et al. Enabling AI-generated content services in wireless edge networks [J]. IEEE wireless communications, 2024, 31(3): 226-234. DOI: 10.1109/MWC.004.2300015

[20] DU B X, DU H Y, LIU H F, et al. YOLO-based semantic communication with generative AI-aided resource allocation for digital twins construction [J]. IEEE Internet of things journal, 2024, 11(5): 7664-7678. DOI: 10.1109/JIOT.2023.3317629

作 者 简 介



崔佳怡，北京邮电大学在读博士研究生；主要研究领域为算力网络、工业互联网等。

谢人超，北京邮电大学教授；主要研究领域为未来网络体系架构、算力网络、云网融合、工业互联网、信息中心网络等；作为项目负责人主持或参与国家重点研发计划、国家自然科学基金、北京市自然科学基金、工信部重大专项、华为企业合作基金等项目20余项；发表论文70余篇。

唐琴琴，北京邮电大学博士后；主要从事边缘计算、算力网络、卫星互联网、网络人工智能相关研究工作；参与多个国家重点研发计划、国家自然科学基金等项目；发表论文20余篇，申请国家发明专利10余项。