ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 3,May 2025,pp. 585－602
DOI:10. 16337/j. 1004-9037. 2025. 03. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

通信网络与AI大模型融合发展研究综述

瞿崇晓[1]，唐宇波[2]，吴高洁[2]，范长军[1]，张永晋[1]，刘　硕[1]

(1. 中国电子科技集团公司第五十二研究所,杭州 310012;2. 智能博弈重点实验室,北京 100091)

摘　要: 随着生成式AI技术的快速发展,特别是大模型领域的突破,学术界和产业界正积极寻求AI大模型与通信网络的深度融合。本文致力于深入探索这一新兴领域,通过梳理最新的相关研究进展,全面剖析如何通过AI大模型提升通信网络的智能化水平,以及如何利用通信网络增强AI大模型的效能。首先,介绍了基于Transformer的大模型主流架构,阐述了大模型的训练过程和智能涌现机制。随后,分析了AI大模型在网络设计、诊断、配置、安全和网络语言理解以及规范分析方面的智能化应用,并讨论了相关的技术实现手段。此外,探讨了通信网络在支撑AI大模型训练、推理和部署中的关键作用,重点关注基于云边协同的分布式大模型构建技术和多智能体大模型网络构建方案。最后,提出了若干亟待解决的关键研究议题,并对未来研究方向进行了展望。

关键词: AI大模型;通信网络;融合与协同;Transformer;云边协同

中图分类号: TP393;TP181　　**文献标志码:** A

Review on Integrated Development of Communication Networks and Large-Scale AI Models

QU Chongxiao[1], TANG Yubo[2], WU Gaojie[2], FAN Changjun[1], ZHANG Yongjin[1], LIU Shuo[1]

(1. The 52nd Research Institute, China Electronics Technology Group Corporation, Hangzhou 310012, China; 2. State Key Laboratory of Intelligent Game, Beijing 100091, China)

Abstract: With the rapid development of generative AI technologies, especially breakthroughs in the field of large language models (LLMs), both academia and industry are actively seeking deeper integration between these large-scale AI models and communication networks. This paper aims to explore this emerging field in depth by reviewing the latest research advancements. It provides a comprehensive analysis of how LLMs can enhance the intelligence of communication networks and how communication networks can improve the performance of LLMs. First, the paper introduces the mainstream Transformer-based architectures of LLMs, elaborating on their training processes and the mechanism of intelligent emergence. It then analyzes the intelligent applications of LLMs in network design, diagnostics, configuration, security, network language understanding, and specification analysis, and discusses the corresponding technical implementation methods. Furthermore, the paper explores the crucial role of communication networks in supporting the training, inference, and deployment of LLMs, with a focus on distributed LLM construction technologies based on cloud-edge collaboration and multi-agent LLM network construction solutions. Finally, the paper identifies several key research challenges that remain to be addressed and provides insights into future research directions.

收稿日期:2024-07-29;**修订日期:**2024-11-06

Key words: large-scale AI models; communication networks; integration and collaboration; Transformer; cloud-edge collaboration

引　言

在当今移动通信网络日益复杂、通信业务生态不断多元化的背景下，网络通信正面临着前所未有的挑战。在民用领域，随着物联网和5G应用的蓬勃发展，从简单的文本消息到高清视频流等数据流量呈现出爆炸性增长，对网络带宽、连接数密度以及低延迟传输提出了更高要求。用户对通信质量、传输效率、网络安全和个性化服务体验的期望也在不断提高[1]。在军事领域，现代战争已进入信息化时代，联合作战和全域作战成为主要作战形态，亟需构建完善的网络信息体系作为支撑[2]。这一转变要求网络通信具备卓越的稳定性和安全性，以建立高度可靠且隐蔽的通信系统，从而有力支撑战术部署和战略决策。通信系统不仅要在极端环境下保持稳定运行并有效抵御外界干扰，还需实现跨战区的无缝通信连接，以及与无人机、卫星系统和其他高科技武器系统的高效数据交互，以确保指挥控制与战术操作的实时性和精确性[3]。这对通信网络的设计、规划、运营与维护提出了更加严苛且细致的要求。

在这一背景下，生成式人工智能(Artificial intelligence, AI)技术，特别是以ChatGPT为代表的大型语言模型(Large language models, LLMs)，为通信网络的创新与发展带来了新的思路，开辟了应对上述挑战的新路径[4]。生成式AI技术能够从大规模数据集中学习，并自动生成新的内容，包括文本、图像、视频等，展现出强大的创新能力。作为生成式AI的一个重要应用分支，LLMs在语义理解、情感分析和文本生成等多个自然语言处理任务中表现出色。多模态大模型更是将这种能力从文本扩展到2D/3D视觉、音频信号和X射线等多个维度，为包括通信网络在内的多个领域提供了潜在的提升方案[6]。目前，关于大型语言模型的技术原理及其在教育、工业、医疗等领域应用的研究已经比较丰富，这些研究主要关注文本数据，也有一些涉及图像数据[7]。相较之下，现有研究对AI大模型在通信领域应用的关注仍然有限，大型语言模型与网络数据相结合的潜力尚未得到充分挖掘。

此外，作为信息服务的基础设施，通信网络在大型语言模型的训练、部署、推理与使用过程中发挥着至关重要的作用[8]。一方面，通信网络需要为大型语言模型的训练和推理过程提供高效的数据传输服务；另一方面，网络系统应充分利用其云计算和边缘计算能力，为大型语言模型协调和分配所需的计算资源。而且，为了支持各类智能应用的开发与运行，需要在网络各节点上部署大型语言模型，通过集群智能为用户提供灵活且多样的智能服务。因此，利用通信网络来赋能AI大模型的研究逐渐崭露头角，并吸引了越来越多研究人员的关注。

本文旨在探索通信网络与AI大模型的交叉融合与协同创新，力求通过两者的结合实现互补与增益。首先，简要介绍了大模型技术的基本原理，包括其基本架构、训练过程及智能涌现机制。然后，探讨了大模型技术在通信网络领域的潜在应用，并总结了相关的技术实现手段。同时，分析了通信网络如何为大模型提供支撑，以更好地为用户提供智能服务。在此基础上，深入分析和讨论了大模型技术与通信网络相结合过程中面临的挑战、不足以及未来的发展方向，为后续研究提供参考和启示。

1　大模型技术与理论

大型语言模型，通常简称为大模型，是指在海量文本语料上训练得到的、参数规模通常超过数十亿的大型神经网络模型。这些AI大模型通过自监督或半监督学习方法在大量未标记文本上进行训练，具备通用的智能，能够执行各种自然语言处理任务[9]。多模态大模型是对大型语言模型的一种重要扩展，它通过跨模态的学习和融合，能够处理图像、音频和视频等多种信息，从而更全面地理解和表达世界。

1.1 大模型基础架构

当前主流的大模型结构主要是在Transformer架构的基础上进行的持续改进。与传统的循环神经网络和卷积神经网络不同,Transformer采用自注意力机制来处理数据序列,在自然语言处理及其他序列到序列的多模态任务中展现出卓越的性能[10]。

Transformer由Vaswani等[11]于2017年提出,是一种由编码器和解码器组成的神经网络架构。其中,编码器负责理解输入文本并生成其语义表示,而解码器则利用这些语义表示生成目标序列。编码器和解码器均由若干相同结构的自注意力层和前馈神经网络层堆叠而成,如图1所示。自注意力层能够为序列中各个位置的词分配不同的注意力权重,通过考虑一个词与其他词之间的位置关系,确定与之最相关的信息。这使得模型能够并行处理整个序列,而无需像循环神经网络那样依赖前一步的计算结果,从而提高了计算效率。在其之后的前馈神经网络层则将每个位置的特征独立地映射到更高维度的空间。为了引入序列中各元素的相对位置信息,Transformer还采用了位置编码[12]。此外,通过残差连接和层归一化技术,Transformer有效解决了训练过程中的梯度消失问题,显著提升了表示能力和训练稳定性,增强了模型的整体性能。

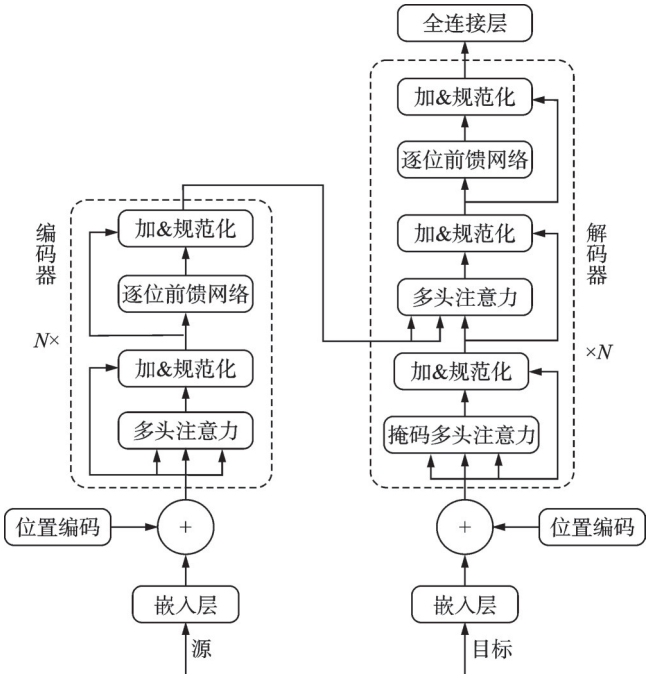

图1 Transformer网络架构

Fig.1 Transformer architecture

Transformer模型的发展衍生出了多种不同架构的大型语言模型,主要包括BERT、GPT系列、T5、XLNet和BART等[13]。这些模型可大体分为3类:仅编码器(Encoder-only)、仅解码器(Decoder-only)以及编码器-解码器(Encoder-decoder)架构。BERT是由谷歌开发的预训练模型,它采用双向Transformer编码器,在理解上下文关系方面表现出色。目前,BERT已被谷歌纳入其搜索算法中,用来提升搜索结果的精准度和上下文关联性,为用户提供更优质的查询体验[14]。在BERT的基础上,百度推出了ERNIE大模型,该模型整合了结构化知识,如实体概念,在语言推断和命名实体识别方面展现出卓越性能,增强了对中文语言的理解[15]。GPT系列是由OpenAI推出的一系列基于Transformer解码器的生成式预训练模型,涵盖了GPT-1至GPT-4多个版本[16]。这些模型通过大规模无监督学习进行预训练,利用自回归技术逐步预测序列中的下一个单词,并通过微调或迁移学习来适应特定领域的任务。尤其是GPT-3及其后续版本,展现出卓越的文本生成和自然语言对话能力,在与人类意图、价值观对齐方面具有显著优势。T5模型采用了Transformer的编码器-解码器架构,是一种由谷歌提出的统一预训练模型框架,它将所有的语言处理任务统一为文本到文本的格式[17]。这种方法简化了模型的训练和应用过程,提升了模型在摘要、翻译和分类等任务中的表现。XLNet由谷歌和卡内基梅隆大学联合研发,采用了Transformer的编码器。它引入了基于排列的训练方法,通过预测序列中的所有标记(而不是一次预测一个标记),更好地捕捉双向上下文。在需要深刻理解上下文顺序的任务上,如问答和文档排序,XLNet的表现优于BERT[18]。此外,由Facebook提出的BART结合了自编码器和自回归方法,在序列到

序列任务和文本生成方面非常有效。它能够在不丢失关键信息的情况下生成长文档的简明摘要，广泛应用于文本摘要等领域[19]。上述这些模型代表了Transformer架构在自然语言处理领域的多样化发展方向和应用场景，在其基础上演化出一系列功能强大的AI大模型产品[20]，具体如图2所示。

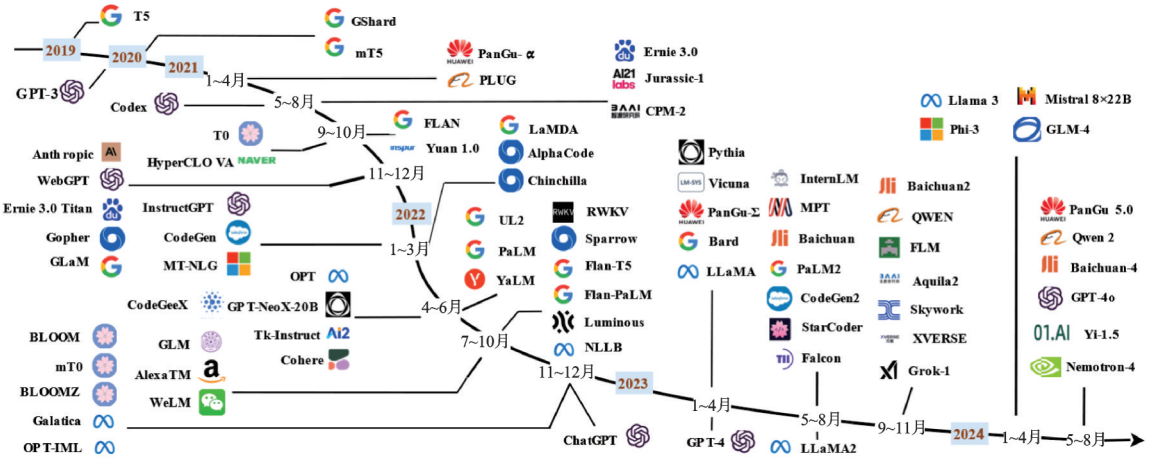

图2 近年来的AI大模型发展时间线

Fig.2 Development timeline of existing LLMs in recent years

1.2 大模型训练与智能涌现

相较于其他深度神经网络，大模型在性能上表现出显著优势，这是由扩展定律决定的[21]。研究表明，通过大幅度扩展语言模型的参数规模、数据规模和训练计算量，模型的能力和性能会大幅提升。当参数规模超过某一临界点时，这些更大规模的语言模型不仅在语义理解、逻辑推理等复杂任务上的表现得到显著提升，还展示出一些小规模模型所不具备的特殊能力。这种引起“质变”的能力被称为涌现能力，是大模型迈向通用人工智能的关键。正因为涌现能力的存在，大模型在广泛的自然语言处理任务中展现出独特的优势。

为了实现大模型的智能涌现，需要依托海量文本数据对基于Transformer架构的模型进行深入的预训练。然后，根据特定领域的应用需求，对模型进行精细的微调，以确保其适应性和实用性。此外，还可以引入基于人类反馈的强化学习等技术[22]，以确保模型输出更符合人类的思维和表达方式，从而提升其在实际应用中的效果。

在大模型的预训练过程中，自监督学习起到了至关重要的作用。这种方法无需大量人工标注数据，只需在原始未标注数据的基础上生成伪标签即可进行训练，从而显著降低了数据收集和处理的成本。通过构建诸如掩码语言模型或自回归预测等任务[23]，自监督学习使模型能够从海量未标注文本中学习并掌握丰富的语言结构和语义信息，这使得模型具有了准确、灵活地处理自然语言的通用能力。

为了使大模型适应特定任务或领域的需求，在预训练完成后，需要对模型进行微调，以进一步优化其性能。微调通常在特定任务的带标签数据集上进行，使用监督学习方法对整个模型进行训练。这需要大量高质量的特定领域数据，并对领域知识有一定要求。经过微调，大模型在特定领域的表现能够显著提升。其中，指令微调是一种通过提供明确的任务指令或要求来引导模型进行微调的方法。这种方法可以使模型更好地理解和执行复杂任务，增强其在执行不同任务时的灵活性和适应性。

预训练利用自监督学习赋予模型广泛的语言理解能力，而微调则在此基础上针对特定任务进行优化。这种互补策略显著增强了模型在各类自然语言处理任务中的泛化性能。然而，由于人类价值观的

多样性和偏好的复杂性,模型的输出可能会存在潜在的危害、误导或偏见。为了确保大模型的行为符合人类的期望,需要引入一系列对齐技术[24]。例如,InstructGPT采用了一种创新的方法——基于人类反馈的强化学习,使模型能够响应并适应人类的反馈,从而打造出与人类期望高度契合的大模型。

2 AI大模型赋能通信网络

目前,通信网络的智能化应用正在迅速发展,涵盖了监督学习、无监督学习以及强化学习等多种机器学习技术。尽管生成式AI在通信网络中的应用尚处于起步阶段,但其代表性技术——大模型在网络通信赋能方面展现出巨大的潜力,正受到越来越多的关注。以6G网络为代表,未来的网络将实现架构级的智慧内生[25],能够利用AI大模型优化网络性能,提升用户体验,实现自动化网络运营,从而极大地提高其智能化水平。

2.1 AI大模型在通信网络中的应用

开发通信系统和管理网络基础设施是一项知识密集型和劳动密集型工作,需要丰富的专家经验和大量的手动操作。以往,网络智能化主要依赖于分散部署的小模型,这些模型在特定环境中独立运行[26],例如用于客户服务的智能助手、提高服务质量的自适应路由算法或减少手动错误的确定性配置生成。大模型通过自然语言提供了统一的网络智能接口,使网络智能得以通用化,能够理解并传播知识。大模型在通信网络领域的潜在应用主要包括以下几类。

(1) 网络设计:通过处理包含网络性能指标、设备规格和历史设计模式的大量数据集,大模型能够在设备选择、网络规划、协议制定等多个网络设计领域帮助工程师做出更为明智的决策。在设备选择方面,大模型可以通过分析兼容性要求、性能基准和成本因素,提供与特定网络目标相匹配的推荐。在网络规划方面,大模型能够模拟不同的网络方案,预测潜在的瓶颈,并给出最优布局建议,以平衡效率、可扩展性和弹性。在协议制定方面,大模型可以通过分析历史协议的成功与失败案例,结合当前的网络需求和技术发展,提出创新的协议方案。Bariah等[27]探讨了生成式AI大模型在无线网络设计中的广泛应用,涵盖了网络感知与成像、定位、信道估计、波束成形、联合源信道编码以及多智能体决策等关键领域。Taki等[28]提出了一种新型的网络设计,使用户通过大模型进行通信,而不是直接通过互联网,从而减少了数据传输量,降低了对互联网的依赖。

(2) 网络诊断:对于网络运营商而言,故障排除是一项繁琐且艰巨的任务,尤其是在大规模广域网络环境中。这不仅需要在多个地域间实现跨部门的精密协作,还常常面临难以解释的网络故障和性能下降问题,从而导致难以估量的经济损失。通过将AI大模型集成到网络诊断系统中,运营商可以显著提升故障排除的效率和准确性。大模型能够根据网络状态信息生成故障报告,加速故障定位,并基于报告分析和历史操作数据提供合理的处理建议。Huang等[29]分析了生成式AI大模型在网络问题诊断中的作用及其实现途径,指出大模型能够通过生成故障报告、实时异常检测和优化网络配置,显著提升网络问题诊断的效率和效果。Long等[30]探讨了大模型通过处理和实时分析多模态数据,快速识别和预警网络故障,提供智能化处理建议,并预测未来故障,从而显著提升网络问题诊断的效率和准确性。Maatouk等[31]将大模型视为电信专业人员的潜在辅助工具,以帮助他们进行网络问题的诊断并提供相应的解决方案。

(3) 网络配置:网络中存在大量异构设备,例如交换机、路由器和中间件。鉴于不同供应商产品之间的差异,网络专家需要投入大量精力学习用户手册、收集合适的命令、验证配置模板,并将模板参数映射到控制器数据库。在此过程中,即便是单个访问控制列表配置错误也可能导致网络中断。考虑到需要管理的大量异构云网络及其计算和存储设备,统一的自然语言配置界面对简化配置过程和实现自配置网络至关重要。Ahmed等[32]指出,大模型可以提供统一的自然语言接口,简化网络配置流程,并有

助于管理各种网络设备。Nguyen等[33]研究了如何通过采用意图驱动网络利用AI大模型实现网络配置自动化,从而使网络管理员能够基于业务意图高效管理复杂的网络。Wang等[34]开发了用户友好的网络配置工具Netbuddy,利用大模型将自然语言形式的高层网络策略翻译为低层设备配置。

(4) 网络安全:网络经常面临各种潜在的安全问题,如分布式拒绝服务攻击、地址欺骗和数据泄露。为了保护网络免受恶意攻击,需要实施一系列措施,包括安全评估、漏洞扫描、入侵检测和防御等。大模型作为强大的交互平台,能够访问和使用各种安全工具和系统。例如,通过逻辑严密的提示引导,大模型可以调用Wireshark的解析工具,并更新防火墙策略,以完成异常流量拒绝任务[29]。Zhou等[35]探讨了网络安全相关语言在结构和语义上与通用语言的不同,并提出了大模型技术用于攻击检测的两个主要研究方向。Nguyen等[33]深入分析了将AI大模型集成到6G网络中的安全挑战,提出了一个全面的威胁分类法,并详细讨论了将大模型应用于网络安全的多种方法。Alwahedi等[36]概述了在物联网环境中应用机器学习进行网络威胁检测的当前趋势、方法和挑战,并提出了利用AI大模型来增强物联网安全的未来愿景。

(5) 网络语言理解:与自然语言的纯文本不同,网络语言包含更多非标准化的格式和符号,涉及从高层管理策略到低层访问控制列表、命令行界面和数据建模语言(如XML和JSON)等方面。传统的映射方法通常仅限于形式化翻译,例如实体抽象和模板填充。相比之下,大模型通过为特定的人类自然语言输入提供定制化响应,能够显著提高网络服务的用户体验质量(Quality of experience, QoE)。然而,网络语言具有特定领域的术语、协议、规则以及数学约束,这使得通用大模型容易因概念模糊而产生“幻觉”,或者因丧失上下文相关性而“胡言乱语”。为了解决这些问题,Huang等[29]提出了一种名为ChatNet的网络大模型架构,通过大量网络知识对大模型进行微调,以适应网络通信领域的需求,并通过访问外部文档的检索增强来促进自然语言到网络语言的映射。此外,这些网络大模型还可以访问外部工具,如搜索引擎、数据分析器、数学求解器和网络工具,以自动化执行复杂任务。

(6) 网络规范分析:第3代合作伙伴计划(3rd generation partnership project, 3GPP)文档制定了移动通信技术的标准,涵盖无线接入技术、核心网络架构以及服务和系统的各个方面,这些文档以其详尽的细节和严格的规范著称。然而,由于文档篇幅庞大,尤其是在新版本发布时,跟踪所有细节成为一项艰巨且耗时的任务。对于试图在产品中实现其中的技术和功能的工程师来说,这一挑战尤为明显,因为他们必须在众多文档中花费大量时间寻找相关信息。AI大模型为解决这一问题提供了有前途的解决方案。Bariah等[37]通过微调大模型来识别3GPP技术文档中的网络规范类别,证实了大模型在电信领域的应用价值。Karapantelakis等[38]评估了现有主流模型在回答3GPP标准文档相关问题时的表现,并提出了用于评估大模型表现的基准和测量方法。此外,他们通过数据预处理和微调显著提高了大模型在回答相关问题时的准确性,并引入了一种名为TeleRoBERTa的模型,该模型在参数数量显著减少情况下的表现与主流大模型相当。

此外,随着技术的不断发展,新兴的大模型逐步出现,为解决通信网络中的各类应用难题提供了新的视角。通信网络通过地理上分布的异构无线节点生成并收集多模态数据,这些数据反映了网络在不同时间和环境条件下的动态运行状态,涵盖了从网络设计、部署、维护到故障修复的各个阶段[39]。由于这些数据具有显著的时间和空间特性,且以多模态、多粒度和异步报告的形式呈现,通信网络数据显示出复杂的时序特征[40]。近年来,时序大模型的研究取得了显著进展,展现出在处理此类数据方面的巨大潜力和应用价值。时序大模型能够有效捕捉通信网络数据中的内在时序依赖关系,在多种网络应用中实现情境感知,从而支持更为高效的决策[35]。例如,通过时序大模型分析交通设备收集的历史和实时网络性能数据,可以为自动驾驶车辆提供优化的路线选择和交通流量预测。

2.2 AI大模型赋能通信网络的途径

在通信网络领域应用AI大模型的关键技术手段主要包括预训练与微调、提示工程以及检索增强生成等。预训练是基于海量数据对大模型进行初步训练的过程,旨在为模型构建广泛且深厚的知识框架。微调是在预训练模型的基础上,利用标注数据针对通信网络领域的特定任务进行深入学习,通过调整模型参数,提升其在特定任务中的表现,增强模型的针对性和专业性。提示工程通过精心设计查询输入,引导模型生成预期的响应。该过程需要根据用户意图设计合适的提示,从而从模型中获取所需的信息或行为[41]。检索增强生成(Retrieval-augmented generation, RAG)结合了大模型与信息检索技术,特别是向量检索方法,提升模型在记忆与事实准确性方面的表现[42]。这一技术使大模型能够更准确地调用和利用大量通信网络相关知识,从而生成更可靠和精确的输出。综合应用上述技术手段,能够显著提升大模型在通信网络领域的应用质量、效率和精准度。

2.2.1 预训练与微调

AI大模型的训练过程依赖于大规模的无监督学习。这些模型在海量的文本语料库(约PB级)上进行预训练,旨在学习预测句子中的下一个词。该训练过程极为耗时,通常需要消耗约10^5~10^6 GPU小时。图1显示了Transformer架构模型的预训练数据流。在训练过程中,词嵌入向量与位置编码相加后被输入模型,并逐层传递下去。编码器和解码器是模型的基本组件,它们均由多头注意力模块和逐位前馈网络构成。多头注意力模块配备多个注意力头,它们并行工作,评估句子中每个词与其他词的相关性。该机制利用查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)和值($\boldsymbol{V}$)向量,通过特定计算捕捉词与词之间的复杂依赖关系。经过注意力模块处理后,各层的输出结果被整合,并通过线性变换进一步计算,以生成输入序列的语义表示。通过堆叠多个编码器和解码器,模型采用自回归方法逐层计算并迭代调整参数,逐步学习得到丰富的参数。据估计,1次完整的训练过程耗资可达数百万美元,但每次API调用的成本较为经济,每千次调用的费用通常不超过1美元[22]。在预训练通信网络领域的大模型时,数据集至关重要。目前,业界正积极收集、处理并推出专门针对电信领域的数据集。例如,Maatouk等[43]收集了来自电信领域的10 000个问题及其答案,推出了TeleQnA数据集,旨在评估大模型在电信领域的知识水平。据称,这一数据集是专为电信行业量身定制的首个开源基准数据集。类似地,Nikbakht等[44]推出了一个涵盖从Release 8到Release 19所有3GPP文档的开源综合数据集TSpec-LLM。基于这些电信领域的数据集,Meng等[45]推出了一个生成式预训练模型NetGPT,用于准确建模网络流量,以帮助提高网络服务质量并保护数据隐私。

典型的大模型训练过程通常在预训练之后使用特定领域的数据进一步调整预训练模型的部分参数,以使其包含特定领域的知识[46]。常见的微调策略包括指令微调和对齐微调。指令微调通过使用自然语言格式的实例来微调预训练模型,从而增强其对未见任务的泛化能力。这一过程涉及收集或构建指令格式的实例,并以监督方式进行训练[47]。这些实例通常包括任务描述、可选输入、对应输出以及示范例子,主要来源于传统自然语言处理任务数据集、日常聊天数据和合成数据。研究表明,指令实例的设计和质量对模型性能有显著影响,高质量且多样化的指令实例能够显著改善模型的任务泛化能力。对齐微调的目标是确保模型输出符合人类价值观,避免生成有害、偏见或误导性的内容。该方法通过收集不同标注者的反馈,指导模型生成符合预定义标准的输出。对齐微调的关键技术之一是基于人类反馈的强化学习,其过程包括监督微调、训练反映人类偏好的奖励模型,并最终使用该奖励模型对原大模型进行微调[48]。此外,直接优化方法通过监督学习构建高质量对齐数据集,可以直接对模型进行微调。在通信网络领域,这些不同的微调方法各有其优缺点和适用场景。例如,Bariah等[37]采用常规方法微调了包括RoBERTa和GPT-2在内的多个大模型,以适应电信领域的应用,并将其用于识别3GPP标准规范。Kan等[49]提出了一种专门用于5G网络分析的开源大模型Mobile-LLaMA。该模型通过使用

公开的真实5G网络数据集进行指令微调,增强了LLaMA 2 13B模型,使其在数据包分析、IP路由分析和性能分析等任务中表现出色。

2.2.2 提示工程

有效的提示工程对于最大化AI大模型在各种应用中的潜力至关重要。通过精心设计问题或陈述,提示工程能够引导大模型生成所需的输出。特别是对于复杂的网络通信任务,提示应当包含明确且充足的上下文信息,例如特定的约束条件和预期的响应格式。以下是几种常用的提示方法。

零样本提示不依赖任何示例,只需提供一个清晰的任务描述或相关问题,即可引导模型生成尽可能准确的响应。这种方法特别适用于简单的网络任务。例如,在询问大模型关于3GPP技术规范中的概念时,可以直接提出问题,而无需提供示例。少样本提示稍显复杂,它包含1个小型示例集(通常为1~3个),用于向模型展示所需任务或答案的格式。这些示例相当于一个小型训练集,引导模型理解上下文和预期响应的具体性质。在网络通信领域,这些示例可能包括网络配置的简要描述、故障排除场景或协议交互,紧随其后的是需要类似响应的查询。Shao等[50]利用少样本提示技术将领域知识嵌入大模型中进行频谱感知,通过提供少量训练示例,使模型在不同信噪比下的性能接近最佳能量检测器。

上下文学习(In-context learning, ICL)最早在GPT-3中引入,它通过格式化的自然语言提示,结合任务描述和示例来引导大模型执行任务[51]。这种方法使模型能够充分利用上下文信息,从而识别并完成新的任务。示例的设计在ICL中至关重要,包括示例的选择、格式化和先后顺序。示例的格式化涉及将选定的示例转换为结构化的提示,整合任务特定的信息,并可能结合推理过程以增强效果。示例的顺序则可根据与查询的相似性进行编排,或采用信息论方法来优化其排列组合,以减少大模型的偏差。ICL的基本机制包括任务识别和任务学习,使得大模型能够通过预训练的知识和结构化提示推断并解决新任务。任务识别是指大模型从提供的示例中识别任务类型,并利用预训练数据中的相关知识。任务学习则是指大模型通过给定的示例获取新的任务解决策略。随着模型规模的不断增大,这种能力变得愈发显著。Zhang等[52]设计了3种上下文学习方法,显著增强了大模型在全自动网络入侵检测任务中的处理效率。Zecchin等[53]通过上下文学习,利用预训练的大模型将输入信号和导频符号直接映射到输出数据符号,从而有效提升了多输入多输出(Multiple-input multiple-output, MIMO)均衡的效果。

思维链(Chain-of-thought, CoT)提示使得大模型能够处理更为复杂的算术、常识性问题及符号推理任务。该方法鼓励大模型遵循逐步推理的过程,类似于人类在解决问题时将复杂任务拆解的思维方式[54]。借助CoT提示,大模型能够系统地分析和应对网络通信中的复杂情境。例如,在诊断网络问题时,可以设计流程化的提示,引导模型逐步考虑各类因素,包括网络拓扑、硬件状态以及软件配置等。此外,CoT提示还可以促进即插即用网络工具的级联使用,将大模型从单纯的答案生成器转变为能够使用工具、进行逻辑推理和解决问题的网络专家。CoT方法的核心是依赖一系列推理步骤(即“思维链”),将生成提示或回答过程中的中间认知环节具体化,从而增强模型的推理能力,并提高结果的可解释性与调试性。在此基础上,自一致性方法通过采样多个独立的CoT,并选择最频繁出现的结果作为最终输出,进一步提升了模型在算术推理任务中的表现。此外,近期提出的树状思维(Tree-of-thoughts, ToT)和图状思维(Graph-of-thoughts, GoT)方法,通过模拟人类层次化的推理过程和构建灵活的依赖关系图,进一步提升了大模型在复杂推理任务中的表现[55]。这些CoT及其衍生方法在数学、常识推理和符号推理等任务中展示了广泛的有效性和卓越的性能。文献[56]提出了基于双通道理论的通信认知增强技术框架,探讨了包括CoT、ToT和GoT等思维推理技术在内的认知增强技术,并讨论了它们在通信领域中的应用。

2.2.3 检索增强生成

凭借上述技术手段,AI大模型在思维推理方面表现出色,但在获取即时、专业及私有信息时仍存在一定局限。为了解决这一问题,RAG技术应时而生。RAG技术将大模型与信息检索技术,特别是与向量检索技术结合起来,通过向大模型提供外部知识,实现对外部信息源的有效整合,从而增强模型的决策和生成能力,为构建更加智能和专用的通信领域大模型提供了新的研究思路和方法[57]。

RAG过程始于从各种来源(如数据库和文档库)收集并创建外部数据。这些数据通常通过嵌入语言模型转换为便于高效检索的格式,并存储在向量数据库中,形成一个全面的知识库。在接收到查询时,大模型首先通过语义搜索在庞大的数据库中寻找与查询语义高度相关的文本片段。这一过程依赖于语义索引,它通过分析数据的意义,而不仅仅依靠关键词匹配,来组织和检索信息。随后,大模型将这些检索到的信息作为提示的一部分,连同原始用户输入一起输入到生成过程中,从而为模型提供更为丰富的背景信息,帮助其生成更加精准且富有见地的响应。

通过将最新的数据集(如迭代中的网络标准草案或更新的网络维护日志)动态引入大模型,RAG技术有效解决了大模型无法及时获取最新知识的问题,同时避免了频繁重新训练的繁琐工作。它通过将与通信网络相关的外部数据与模型对接,使模型能够在丰富的网络通信上下文中进行学习和生成响应,从而显著提升了大模型在通信网络智能化应用中的灵活性、准确性和实用性。Piovesan等[58]通过检索增强生成方法提升了Phi-2语言模型的能力,整合了专门针对电信标准规范的大规模知识库。增强后的Phi-2模型在回答关于电信标准的问题时,准确性明显提升,其精度接近资源需求更高的GPT-3.5。Bornea等[59]提出了一种开源的检索增强生成框架——Telco-RAG,用于处理电信标准文档,特别是3GPP文档,为大模型在电信领域的应用铺平了道路。

3 通信网络赋能AI大模型

以6G为代表的未来通信网络将实现算网一体化。网络和计算将能够相互感知、协同工作并深度融合,从而实现实时、准确的算力发现以及灵活、动态的计算和连接服务调度。这一进程将提供无处不在的计算和服务,推动云、边、网、算的高效协同[60]。在此背景下,通信系统在AI大模型的训练、推理、部署和使用过程中将发挥越来越重要的作用,为AI大模型的广泛应用提供坚实支持。

3.1 基于云边协同的大模型

传统上,AI大模型通常依赖集中式云服务器进行训练和托管,以满足其巨大的计算需求。尽管这种集中式架构能够提供强大的计算能力,但也面临一系列不可忽视的挑战,包括高网络延迟、高能耗、带宽受限以及隐私泄露等问题。集中式云计算架构的高延迟问题主要源于数据传输路径的长距离和处理队列的拥堵。在需要即时反馈的应用场景中,如自动驾驶、智能制造或智能家居等,高延迟会导致严重的性能瓶颈,影响用户体验和系统安全性。Dong等[61]指出,大模型的请求响应时间通常在几秒到几十秒之间,而自动驾驶决策处理对性能的最低要求则在几十毫秒级别,显然,集中式架构无法满足这一需求。此外,集中式云计算的巨大能源消耗不仅增加了运营成本,还导致了大量碳排放,给环境的可持续性带来了压力。同时,随着用户终端设备规模的不断扩大,这些设备不断产生大量数据,并涉及个人隐私,对需要频繁上传和下载数据的集中式计算模式提出了更大的挑战。为了应对这些问题,学术界和工业界已开始探索通过边缘设备来实现AI大模型的训练与推理。

边缘训练通过软硬件协同设计,利用数据压缩、稀疏性和流水线等技术手段,减少片外内存访问和计算量,从而优化训练过程。边缘推理则采用低秩分解、参数量化、网络剪枝、知识蒸馏等模型压缩技术,并结合软硬件优化,适配并提升大模型在边缘端的推理效率[62]。Frantar等[63]提出了一种名为SparseGPT的大模型剪枝方法,在几乎不损失精度的情况下,将GPT模型剪枝至至少60%的稀疏性。Ma

等[64]提出了基于梯度信息的结构剪枝方法LLM-Pruner,能够选择性地移除模型中非关键的耦合结构。这些研究为减少大模型部署时对内存和计算力的需求,提升其在边缘设备上的运行可行性提供了有力支持。然而,这些技术在资源极其有限的设备上的应用仍然存在难点,并且难以满足大规模、高精度模型部署时对时延和能耗等方面的要求。

边缘云计算通过网络连接多个节点(如终端设备、边缘服务器或云计算中心)协同工作,扩展计算能力并实现并行化,从而促进数据资源的共享,以满足计算密集型大模型的应用需求。这是一种高效的分布式计算策略,它将计算功能从传统的云数据中心转移至网络边缘,使得模型训练或推理所需的计算资源更接近数据源或边端设备。这样的协作机制使得数据可以在更靠近用户的位置上进行处理,从而显著降低数据处理延迟,提升响应速度,减少带宽占用,同时保护用户隐私。此外,这种模式还能够缓解中央服务器的负荷,降低整体系统的能源消耗。Wang等[65]对生成式AI大模型与边缘云计算的最新发展进行了调研,分析了当前在大规模生成式AI大模型训练和部署中所面临的技术挑战,并讨论了边缘云计算如何通过分布式计算、数据压缩、个性化和隐私保护等技术手段,提升大模型的计算效率和响应速度。Xu等[66]研究了在移动网络中通过边缘云协同部署生成式AI大模型的应用、挑战及未来研究方向,并指出:“为了支持AI大模型的预训练、微调和推理,需要大量的计算、通信和存储资源。因此,为了提供低延迟和个性化的智能服务,需要构建云边协同的智能计算框架,以实现异构资源共享者之间的广泛合作。”

边缘云协同计算系统通常由云计算中心与网络边缘设备相互连接构成,后者可进一步划分为边缘服务器和终端设备,如图3所示。将计算任务拆分至网络边缘的关键在于合理减轻云端的计算负担,同时确保边缘节点与云端之间的高效协同和紧密互联互通。为了研究无线通信如何支持大模型,Xue等[67]提出了一种基于专家混合(Mixture of experts, MoE)的无线分布式大模型范式WDMoE,将LLM的计算任务分散到基站的边缘服务器和无线通信系统中的移动设备上。具体而言,该方法通过解构MoE层,将门控网络和前置神经网络层部署于基站,而专家网络则分布到各个设备上,以充分发挥分布式设备上专家网络的并行处理能力。文献[68]提出了一种面向生成式AI的云边端协同智能网络框架GainNet。通过深度融合生成式AI与6G网络设计,GainNet实现了正向闭环的知识流动,并促进了生成式AI大模型的可持续优化演进。在此基础上,提出了一种集成感知、通信与计算的通用资源编排机制,以确保GainNet的高效运作。更进一步地,Chen等[69]提出了一种智慧内生的网络架构NetGPT。该架构通过创新的云边协同部署策略,巧妙地在云端和边缘端配置不同规模的AI大模型,实现了通信与计算资源的深度集成以及AI逻辑工作流的灵活设计。

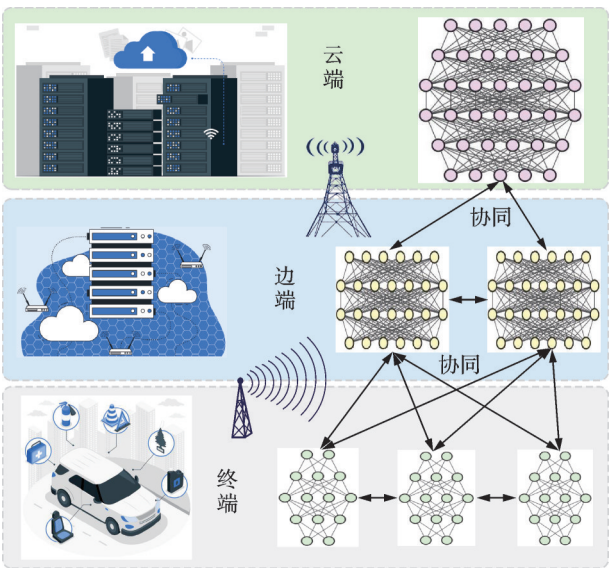

图3 基于云边协同的分布式大模型方案

Fig.3 Distributed LLM solution based on cloud-edge collaboration

在具体实施中,在边缘设备上部署轻量级大模型,以适应这些设备有限的计算能力;而在云计算中心运行规模更大、功能更全面的模型,以充分利用其强大的计算能力和充足的存储资

源[70]。边缘端的轻量级模型能够利用地理位置等信息,提升服务的个性化程度,满足特定区域用户的需求,但存在泛化性不足的问题。云端的大规模LLMs则能够提供通用智能服务,但存在时延大和个性化不足等问题。通过云端与边缘端的协同,靠近用户的边缘端轻量级模型负责获取并初步处理本地区域内的请求,处理后的请求再汇总传输至云端的大规模LLMs,由其凭借强大的算力生成高质量的反馈信息。在此过程中,通信网络作为连接云端与边缘端AI大模型的桥梁,促成两者之间的分工合作,优化了资源分配和优势互补,既提高了内容生成的个性化水平,又保证了其品质。此外,通过边缘端轻量级LLMs的高效预处理,减少了不必要的数据传输,有效降低了通信成本,并优化了网络响应速度,从整体上提升了AI大模型的智能服务效率和质量。

3.2 多智能体大模型网络

尽管AI大模型具有强大的能力,但单一大模型在理解现实世界方面仍有不足,且在维持长链推理过程中的一致性上也面临困难。这些缺陷阻碍了它们进行深入、迭代的思维过程,即所谓的“慢思考”[71]。随着AI大模型逐步赋能更多的边缘设备,通过多智能体网络的互联与协同,能够使大量无线设备展现出集体智能,从而有效应对这些挑战。为实现这一目标,通信系统需从数据和模型驱动的范式转变为知识和推理驱动的范式。传统通信网络主要依据关键性能指标在网络节点之间传输数据,但这种方式对于由AI大模型驱动的大规模设备网络来说效率较低。未来的网络系统应开发一种能够在网络内部传递知识的计算架构,将当前在云端的基于海量知识训练的AI大模型推向分布式集体智能。为此,通信领域的AI大模型应扎根于现实世界情境,通过知识驱动的通信来执行多智能体的规划、决策和推理等复杂任务。

Bariah等[27]提出了支撑多智能体LLMs无线网络实现的3个关键研究视角,分别是语义通信、新型协议学习和多智能体强化学习。具体地,网络系统应具备从原始数据中提取有效语义抽象的能力,这对于在资源受限的无线设备上部署和连接AI大模型至关重要。通过重新设计无线物理层,采用语义化高效通信机制,可以大大减少计算开销和通信冗余。借助分布式规划和语义通信,具备感知、规划、批判和推理能力的多个智能体可以协同工作。它们不仅能够利用AI大模型进行推理,还能感知环境、规划任务、记忆经验并评估行动,从而提高执行效率和决策质量。在此过程中,为了使基于大模型的多智能体网络能够高效地协同工作,合理、有效的智能体间通信协议是必要的。通信协议为智能体的协作提供了结构化的框架和方法,通过在多智能体网络中建立消息交换的规则和机制,指导具体动作的执行[71]。Cheng等[72]指出,实施定义明确的通信协议可以确保智能体的交互遵循一致的结构和语义,从而减少模糊和误解,提高通信效率。Pham等[73]提出了一种名为CIPHER的新型通信协议,它基于嵌入表示实现模型间的通信。这一协议颠覆了传统方法,从大模型中移除了token采样步骤,直接通过Transformer输出的嵌入特征来实现模型间的交互。在多类推理规划任务中,该方法展现出了卓越的性能,为智能体之间的高效通信提供了一种有效的途径。

为应对边缘节点有限的计算和存储资源,无线设备上运行的大模型应专注于理解特定领域的知识,并与环境有效互动,以执行特定任务。构建这类模型的关键在于将其置于一个现实世界的框架中,该框架包含模块化结构和因果关系的知识,确保模型既可扩展又保持轻量化,以适应边缘设备的运行条件。借助无线网络连接,这些封装了领域专业知识的大模型能够实现知识的传输与编码,进而依据接收到的信息做出逻辑决策。面对复杂任务时,通常需要多个无线设备上的多模态传感器和多任务执行器的协同工作,而单个设备端大模型由于资源限制,往往只能专注于特定领域的知识。因此,需要将高层目标分解为更具操作性的低层任务,并通过先验知识和信念体系进行复杂问题的推理。这种分而治之的策略,结合设备的交互、规划及推理能力,赋予了它们智能体的特性,使其能够在各自的领域中

自主运行。无线网络作为连接纽带，提供了必要的通信和传输基础设施，支持智能体之间的信息交换与协作。因此，通过无线网络协同多个大模型智能体的集体智慧，成为高效完成复杂任务的可行且有前景的方案。

为了实现上述构想，Zou等[74]提出了一种创新的多智能体大模型网络架构，如图4所示。该架构通过在大量边缘设备上分散部署大模型，促进了各智能体之间的协作规划和任务决策。这一方法巧妙地解决了由于大规模LLMs巨大的计算和存储需求而难以直接集成到边缘计算环境中的问题。在该框架中，大模型智能体既作为感知器来观察环境，也作为执行器来执行决策，实现了生成式大模型、边缘网络与多智能体系统三者之间的融合与协同。无线生成式智能体利用大模型开展闭环任务规划、执行和优化的具体步骤包括：首先，智能体将高层意图或目标进行解构，并随着时间的推移规划一系列低层任务。随后，智能体感知环境，做出决策并相应地采取行动，然后根据奖励优化决策策略。根据每一步的结果，智能体创建新的任务，直到目标达成，并生成最终响应。此外，文献[75]提出了一种名为GenAINet的分布式多智能体框架，用于管理网络协议与应用程序，旨在通过生成式AI与通信网络的融合，充分发挥集体智能的力量。在该框架中，生成式大模型智能体从多模态原始数据中提取语义概念，并构建表示其语义关系的知识库，供生成式AI大模型检索，以进行规划与推理，如图5所示。在这一范式下，智能体可以快速从其他智能体的经验中学习，并通过高效的通信做出更优决策。该研究还开展了无线设备查询和无线功率控制两个案例研究。前者表明，通过从教师智能体到学生智能体的知识传递，可以提高查询的准确性与通信效率；后者展示了分布式生成式大模型智能体通过协同推理，能够为目标数据速率提供有效的功率分配解决方案。

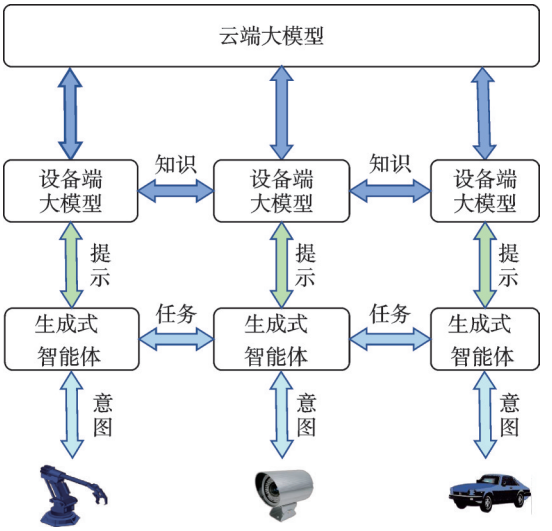

图4 多智能体大模型网络架构

Fig.4 Multi-agent LLM network architecture

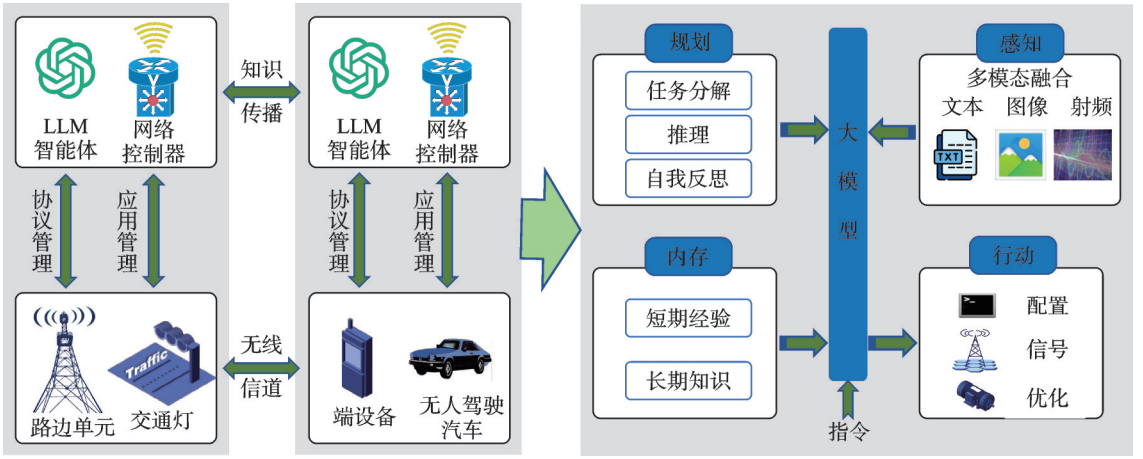

图5 无线生成式LLM智能体网络框架[75]

Fig.5 Wireless generative LLM network and agent architecture

4 挑战、机遇与未来展望

在AI大模型和通信网络迅速发展的背景下,AI大模型在通信系统中的应用已渗透到多个层面,包括网络设计、规划、配置、安全和运营等领域,相关研究也取得了显著进展。同时,随着新兴网络技术对AI大模型训练、部署、推理和使用的支持不断加强,通用智能技术得到了进一步发展,大模型技术也得以持续演进。可以预见,AI大模型与通信网络的深度融合和协同发展将带来更大的收益。然而,该领域的研究尚处于起步阶段,仍面临诸多问题和挑战。本文将逐一探讨这些问题和挑战,并展望未来的发展方向。

(1) 大模型的可解释性:在通信网络的整个生命周期——从设计、规划到管理与运营,深入理解网络大模型的决策机制和内在逻辑,是确保系统高效运行的关键。这些模型的可解释性对于网络管理人员至关重要,因为它直接影响到他们是否能够清晰地洞察AI智能体的决策过程,从而进行有针对性的优化和增强,并有效应对过程中出现的不明问题,实现更加可信的系统行为。因此,未来的研究应加强对生成式AI大模型可解释性的探索,使其在通信网络领域的应用更加透明和可靠。

(2) 通信数据与大模型架构的适配:现有的生成式AI大模型架构虽然能够处理多模态数据,但要将这些模型的范式有效适配到通信网络领域,还需对架构进行重新设计和训练。通信网络领域中的某些数据具有独特的特征,例如射频数据,现有的大模型架构难以有效处理这些数据,也难以将其与其他数据模态进行整合。未来的研究应加强针对通信网络领域的大模型设计与训练,以确保这些模型能够高效处理并整合射频数据等多种不同类型的数据。

(3) 设备端大模型的部署:考虑到边端设备的个体差异,尽管将AI大模型部署到这些设备上能够显著降低响应延迟并支持本地操作,但一些设备的计算能力和能源供应无法满足某些模型的运行需求,使这一方案面临严峻挑战。为了解决这一问题,需要研发新颖的模型压缩、量化和知识蒸馏等技术,以优化AI大模型。未来的研究应着重探索如何进一步轻量化AI大模型,确保其在适应边缘设备有限计算资源和能源消耗的同时,依然能够维持卓越的性能,从而实现效能与可行性的最佳平衡。

(4) 设备端大模型的资源管理:在无线网络,尤其是物联网系统中,固有的资源约束对多智能体系统的高效运行构成了挑战,可能严重制约其性能。尽管无线多智能体大模型技术展现出前所未有的潜力,为智能化应用开辟了新的前景,但无线节点的异构性和资源的有限性要求设备端大模型在计算资源、能源效率、内存占用、推理稳定性,以及满足终端用户期望方面达到精妙的平衡。因此,设计和优化资源管理方案变得尤为关键。为实现这一目标,首先需要对每个智能体的资源消耗模式进行精细化分析,包括但不限于CPU利用率、内存使用情况以及能量消耗。基于此,智能体可以根据特定的通信场景和任务需求进行自适应调整。未来的智能资源调度不仅应关注单个智能体的即时需求,还应具备前瞻性的视野,预测整个系统的动态变化,以期在全局范围内实现资源的最优分配。

(5) 设备端多模态大模型:在无线网络环境中,智能体需要处理多样化的数据形态,包括文本、图像、地理位置信息以及射频信号等多模态数据。然而,通信和计算资源的限制对设备端大模型的设计提出了很高的要求。目前,用于编码原始数据的多模态大模型通常体积庞大,难以在移动设备上部署。为克服这一难题,应采用基于知识的大模型,从概念空间出发对多模态数据进行编码,通过捕捉数据的最小必要结构来缩小模型规模并降低推理成本。未来的研究应着眼于构建更加高效的多模态数据概念空间模型,这将有助于开发既轻量又高效的模型,使其能够在移动设备上流畅运行,同时保持对多模态数据的深度理解和处理能力。

(6) 网络延迟与实时处理:为了在物联网环境中实现完全沉浸式的智能服务体验,大模型应能够实时处理和解析大量的本地感知数据,将响应延迟降至人类或敏感系统无法察觉的程度,从而促进无缝

的人机交互。未来的研究不仅应将数据处理部署在靠近数据生成源的用户端,还需进一步优化边缘计算及云边协同计算架构,确保网络基础设施、边缘计算能力和AI大模型3个领域同步取得显著进展,使三者能够协调发展,共同推动实时智能处理技术的创新与进步。

(7) 分布式大模型的协调与数据同步:分布式大模型通过在网络中不同节点生成的大量数据上进行训练,能够支持更多用户参与,从而降低延迟并提高动态适应性。然而,分布式大模型在多个模型之间的协调、聚合和数据同步方面依然面临诸多挑战,尤其是在不同规模和部署位置的大模型之间,实现高效的协同存在显著困难。未来的研究应着重探索如何实现分布式大模型的有效协调与数据同步,以应对分布式节点间的通信瓶颈和资源差异,从而提升整体系统的效率与性能。

(8) 分布式多智能体收敛:在分布式系统中,实现多智能体在无线网络中的收敛是一项极具挑战性的任务,这主要源于无线网络的动态特性。面对不断变化的网络条件,各智能体的大模型必须持续进行自我调整和优化。在多个大模型各自追求不同网络目标的情况下,这一挑战尤为突出。如果无法顺利收敛,可能会导致无法找到最优解,从而影响整个系统的性能。为了应对这一问题,多个大模型之间需要保持高度的协调与配合。然而,整体协调性与大模型的自我治理及个体优化目标之间存在一定的矛盾。因此,未来的研究应当加强对集体智能方法的探索,使多个智能体能够在快速变化的网络状态下,实时达成共识,并实现高效的协作与响应。

5 结束语

本文深入探讨了通信网络与AI大模型的交叉融合及其协同创新发展的现状与未来趋势。首先,详细阐述了大模型技术的基本原理,并探讨了其在通信网络领域的潜在应用场景及实现路径。接着,分析了通信网络如何有效支撑基于云边协同的分布式大模型和多智能体大模型网络的构建,从而提供更加高效、智能的用户服务。最后,深入剖析了大模型技术与通信网络融合过程中面临的主要问题与挑战,并展望了未来的发展方向。通过积极应对这些挑战,通信网络与AI大模型的深度融合将为两者的协同发展开辟全新路径,推动更加智能化的通信网络和更加网络化的通用智能服务的实现。

参考文献:

[1] 尹浩,黄宇红,韩林丛,等.6G通信-感知-计算融合网络的思考[J].中国科学: 信息科学,2023,53(9):1838-1842.
YIN Hao, HUANG Yuhong, HAN Lincong, et al. Thoughts on 6G integrated communication, sensing and computing networks[J]. Science China Information Sciences, 2023, 53(9): 1838-1842.

[2] 刘超,曹晨,段俊奇.现代空中战场通信网络化发展研究[J].战术导弹技术,2024(3):32-40.
LIU Chao, CAO Chen, DUAN Junqi. Research on the development of modern air battlefield communication networking[J]. Tactical Missile Technology, 2024(3): 32-40.

[3] 向夏雨,顾钊铨,曾丽仪.网络威胁情报共享与融合技术综述[J].网络空间安全科学学报,2024,2(2): 2-17.
XIANG Xiayu, GU Zhaoquan, ZENG Liyi. A survey of cyber threat intelligence sharing and fusion technologies[J]. Journal of Cybersecurity, 2024, 2(2): 2-17.

[4] WANG J Y, ZHANG L, YANG Y R, et al. Network meets ChatGPT: Intent autonomous management, control and operation[J]. Journal of Communications and Information Networks, 2023, 8(3): 239-255.

[5] HAO J G, VON DAVIER A A, YANEVA V, et al. Transforming assessment: The impacts and implications of large language models and generative AI[J]. Educational Measurement: Issues and Practice, 2024, 43(2): 16-29.

[6] WANG X, CHEN G, QIAN G, et al. Large-scale multi-modal pre-trained models: A comprehensive survey[J]. Machine Intelligence Research, 2023, 20(4): 447-482.

[7] RAIAAN M A K, MUKTA M S H, FATEMA K, et al. A review on large language models: Architectures, applications, taxonomies, open issues and challenges[J]. IEEE Access, 2024, 12: 26839-26874.

[8] LONG S F, TANG F X, LI Y F, et al. 6G comprehensive intelligence: Network operations and optimization based on large language models[EB/OL]. (2024-04-28). https://arxiv.org/abs/2404.18373.

[9] XU H. Natural language processing in biomedicine: A practical guide[M]. Cham: Springer International Publishing, 2024: 265-297.

[10] HOFFMANN J, BORGEAUD S, MENSCH A, et al. An empirical analysis of compute-optimal large language model training[J]. Advances in Neural Information Processing Systems, 2022, 35: 30016-30030.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2017: 5998-6008.

[12] CHU X X, TIAN Z, ZHANG B, et al. Conditional positional encodings for vision Transformers[EB/OL]. (2023-02-12). https://arxiv.org/abs/2102.10882.

[13] CHOUDHARY A, ALUGUBELLY M, BHARGAVA R. A comparative study on transformer-based news summarization[C]//Proceedings of 15th International Conference on Developments in eSystems Engineering. [S.l.]: IEEE, 2023: 256-261.

[14] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24). https://arxiv.org/ abs/1810.04805.

[15] LIU Y D, LU W X, CHENG S Q, et al. Pre-trained language model for web-scale retrieval in Baidu search[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. [S.l.]: ACM, 2021: 3365-3375.

[16] KALYAN K S. A survey of GPT-3 family large language models including ChatGPT and GPT-4[J]. Natural Language Processing Journal, 2024,6: 100048.

[17] GROVER K, KAUR K, TIWARI K, et al. Deep learning based question generation using t5 transformer[C]//Proceedings of Advanced Computing: 10th International Conference. Singapore: Springer: 2021: 243-255.

[18] YANG Z, DAI Z, YANG Y, et al. XLNet: Generalized autoregressive pretraining for language understanding[C]//Proceedings of Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc, 2019, 32: 1-11.

[19] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 7871-7880.

[20] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[EB/OL]. (2023-11-24). https://arxiv.org/ abs/2303.18223.

[21] MUENNIGHOFF N, RUSH A, BARAK B, et al. Scaling data-constrained language models[C]//Proceedings of Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, Inc, 2024, 36: 1-19.

[22] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural in Formation Processing Systems, 2022, 35: 27730-27744.

[23] SINHA K, JIA R, HUPKES D, et al. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little[EB/OL]. (2021-09-09). https://arxiv.org/abs/2104.06644.

[24] SHEN T H, JIN R R, HUANG Y F, et al. Large language model alignment: A survey[EB/OL]. (2023-09-26). https: //arxiv.org/abs/2309.15025.

[25] SINGH R, KAUSHIK A, SHIN W, et al. Towards 6G evolution: Three enhancements, three innovations, and three major challenges[EB/OL]. (2024-02-16). https://arxiv.org/abs/2402.10781.

[26] 董天健. 知识定义网络的资源调度关键技术研究[D]. 北京: 北京邮电大学, 2023.
DONG Tianjian. Research on key technologies of resource allocation in knowledge defined networking[D]. Beijing: Beijing University of Posts and Telecommunications, 2023.

[27] BARIAH L, ZHAO Q Y, ZOU H, et al. Large generative ai models for telecom: The next big thing?[EB/OL]. (2023-11-23). https://arxiv.org/abs/2306.10249.

[28] TAKI S U, MASTORAKIS S. Rethinking Internet communication through LLMs: How close are we?[EB/OL]. (2023-09-25). https://arxiv.org/abs/2309. 14247.

[29] HUANG Y D, DU H Y, ZHANG X Y, et al. Large language models for networking: Applications, enabling techniques, and challenges[EB/OL]. (2023-11-29). https:// arxiv.org/abs/2311.17474.

[30] LONG S F, TANG F X, LI Y F, et al. 6G comprehensive intelligence: Network operations and optimization based on large language models[EB/OL]. (2024-04-28). https:// arxiv.org/abs/2404.18373.

[31] MAATOUK A, PIOVESAN N, AYED F, et al. Large language models for telecom: Forthcoming impact on the industry [EB/OL]. (2024-02-25). https://arxiv.org/abs/2308. 06013.

[32] AHMED T, PIOVESAN N, DE DOMENICO A, et al. Linguistic intelligence in large language models for telecommunications[EB/OL]. (2024-02-24). https://arxiv.org/abs/2402.15818.

[33] NGUYEN T, NGUYEN H, IJAZ A, et al. Large language models in 6G security: Challenges and opportunities[EB/OL]. (2024-03-18). https://arxiv.org/abs/ 2403.12239.

[34] WANG C J, SCAZZARIELLO M, FARSHIN A, et al. Making network configuration human friendly[EB/OL]. (2023-09-12). https://arxiv.org/abs/2309.06342.

[35] ZHOU H, HU C M, YUAN Y, et al. Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities[EB/OL]. (2024-05-17). https://arxiv.org/abs/2405.10825.

[36] ALWAHEDI F, ALDHAHERI A, FERRAG M A, et al. Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models[J]. Internet of Things and Cyber-Physical Systems, 2024, 4: 167-185.

[37] BARIAH L, ZOU H, ZHAO Q, et al. Understanding telecom language through large language models[C]//Proceedings of GLOBECOM 2023 IEEE Global Communications Conference. [S.l.]: IEEE, 2023: 6542-6547.

[38] KARAPANTELAKIS A, THAKUR M, NIKOU A, et al. Using large language models to understand telecom standards[EB/OL]. (2024-04-12). https://arxiv.org/abs/ 2404.02929.

[39] ZANOUDA T, MASOUDI M, GEBRE F G, et al. Telecom foundation models: Applications, challenges, and future trends [EB/OL]. (2024-08-02). https://arxiv.org/abs/2408. 03964.

[40] NABEEL M, NIMARA D D, ZANOUDA T. Test code generation for telecom software systems using two-stage generative model[C]//Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops). [S.l.]: IEEE, 2024.

[41] MARVIN G, HELLEN N, JJINGO D, et al. Prompt engineering in large language models[C]//Proceedings of International Conference on Data Intelligence and Cognitive Informatics. Singapore: Springer Nature, 2023: 387-402.

[42] CHEN J, LIN H, HAN X, et al. Benchmarking large language models in retrieval-augmented generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2024: 17754-17762.

[43] MAATOUK A, AYED F, PIOVESAN N, et al. TeleQnA: A benchmark dataset to assess large language models telecommunications knowledge[EB/OL]. (2023-10-23). https://arxiv.org/abs/2310.15051.

[44] NIKBAKHT R, BENZAGHTA M, GERACI G. TSpec-LLM: An open-source dataset for LLM understanding of 3GPP specifications[EB/OL]. (2024-08-07). https://arxiv. org/abs/2406.01768.

[45] MENG X Y, LIN C G, WANG Y Q, et al. NetGPT: Generative pretrained transformer for network traffic[EB/OL]. (2024-05-17). https://arxiv.org/abs/2304. 09513.

[46] 陈浩泷,陈罕之,韩凯峰,等. 垂直领域大模型的定制化: 理论基础与关键技术[J]. 数据采集与处理, 2024, 39(3): 524-546.
CHEN Haolong, CHEN Hanzhi, HAN Kaifeng, et al. Domain-specific foundation-model customization: Theoretical foundation and key technology[J]. Journal of Data Acquisition and Processing, 2024, 39(3): 524-546.

[47] 夏润泽,李丕绩. ChatGPT大模型技术发展与应用[J]. 数据采集与处理, 2023, 38(5): 1017-1034.
XIA Runze, LI Piji. Large language model ChatGPT: Evolution and Application[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1017-1034.

[48] 瞿崇晓,郑寄平,张永晋,等. GPT技术原理及其潜在军事应用研究[J]. 中国电子科学研究院学报, 2023, 18(7): 624-633.
QU Congxiao, ZHENG Jiping, ZHANG Yongjin, et al. Research on the principles and potential millitary applications of GPT technology[J]. Journal of CAEIT, 2023, 18(7): 624-633.

[49] KAN K B, MUN H, CAO G, et al. Mobile-LLaMA: Instruction fine-tuning open-source LLM for network analysis in 5G networks[J]. IEEE Network, 2021, 14(8): 1-8.

[50] SHAO J W, TONG J W, WU Q, et al. WirelessLLM: Empowering large language models towards wireless intelligence[EB/OL]. (2024-06-15). https://arxiv.org/abs/ 2405.17053.

[51] DONG Q X, LI L, DAI D M, et al. A survey on in-context learning[EB/OL]. (2024-06-18). https://arxiv.org/abs/2301.00234.

[52] ZHANG H, SEDIQ A B, AFANA A, et al. Large language models in wireless application design: In-context learning-enhanced automatic network intrusion detection[EB/OL]. (2024-05-16). https://arxiv.org/abs/ 2405.11002.

[53] ZECCHIN M, YU K, SIMEONE O. In-context learning for MIMO equalization using transformer-based sequence models [EB/OL]. (2024-01-22). https://arxiv.org/ abs/2311. 06101.

[54] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.

[55] BESTA M, MEMEDI F, ZHANG Z Y, et al. Demystifying chains, trees, and graphs of thoughts[EB/OL]. (2024-04-05). https://arxiv.org/abs/2401.14295.

[56] 廉霄兴,宋勇,朱军,等.基于双通道理论的通信认知增强技术研究[J].电信科学,2024,40(1): 123-135.
LIAN Xiaoxing, SONG Yong, ZHU Jun, et al. Research on dual process theory based telecom augmented cognition[J]. Telecommunications Science, 2024, 40(1): 123-135.

[57] SIRIWARDHANA S, WEERASEKERA R, WEN E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023(11): 1-17.

[58] PIOVESAN N, DE DOMENICO A, AYED F. Telecom language models: Must they be large?[EB/OL]. (2024-06-25). https://arxiv.org/abs/2403.04666.

[59] BORNEA A L, AYED F, DE DOMENICO A, et al. Telco-RAG: Navigating the challenges of retrieval-augmented language models for telecommunications[EB/OL]. (2024-08-07). https://arxiv.org/abs/2404.15939.

[60] 柴蓉,邹飞,刘莎,等. 6G移动通信: 愿景、关键技术和系统架构[J].重庆邮电大学学报(自然科学版), 2021, 33(3): 337-347.
CHAI Rong, ZOU Fei, LIU Sha, et al. 6G mobile communication: Vision, key technologies and system architecture[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2021, 33(3): 337-347.

[61] DONG Q, CHEN X, SATYANARAYANAN M. Creating edge AI from cloud-based LLMS[C]//Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications. New York: ACM, 2024: 8-13.

[62] 侯祥鹏,兰兰,陶长乐,等. 边缘智能与协同计算:前沿与进展[J]. 控制与决策, 2024, 39(7): 2385-2404.
HOU Xiangpeng, LAN Lan, TAO Changle, et al. Edge intelligence and collaborative computing: Frontiers and advances[J]. Control and Decision, 2024, 39(7): 2385-2404.

[63] FRANTAR E, ALISTARH D. SparseGPT: Massive language models can be accurately pruned in one-shot[C]//Proceedings of the 40th International Conference on Machine Learning. Hawaii: [s.n.], 2023: 10323-10337.

[64] MA X Y, FANG G F, WANG X C. LLM-Pruner: On the structural pruning of large language models[C]//Proceedings of the Advances in Neural Information Processing Systems. New Orleans: [s.n.], 2021: 2533-2552.

[65] WANG Y C, XUE J T, WEI C W, et al. An overview on generative AI at scale with edge-cloud computing[J]. IEEE Open Journal of the Communications Society, 2023(4): 2952-2971.

[66] XU M, DU H, NIYATO D, et al. Unleashing the power of edge-cloud generative ai in mobile networks: A survey of AIGC services[J]. IEEE Communications Surveys & Tutorials, 2024, 26(2): 1127-1170.

[67] XUE N, SUN Y P, CHEN Z Y, et al. WDMoE: Wireless distributed large language models with mixture of experts[EB/OL]. (2024-05-05). https://arxiv.org/abs/2405. 03131.

[68] CHEN N, YANG J, CHENG Z P, et al. GainNet: Coordinates the odd couple of generative AI and 6G networks[EB/OL]. (2024-01-05). https://arxiv.org/abs/ 2401.02662.

[69] CHEN Y X, LI R P, ZHAO Z F, et al. NetGPT: An AI-native network architecture for provisioning beyond personalized generative services[J]. IEEE Network, 2024, 38(6): 404-413.

[70] 王永威,沈弢,张圣宇,等. 大小模型端云协同进化技术进展[J]. 中国图象图形学报, 2024, 29(6): 1510-1534.
WANG Yongwei, SHEN Tao, ZHANG Shengyu, et al. Advances in edge-cloud collaboration and evolution for large-small models[J]. Journal of Image and Graphics, 2024, 29(6): 1510-1534.

[71] HÄNDLER T. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous LLM-powered multi-agent architectures[EB/OL]. (2023-10-05). https://arxiv.org/abs/2310.03659.

[72] CHENG Y H, ZHANG C Y, ZHANG Z W, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects[EB/OL]. (2024-01-07). https:// arxiv.org/abs/2401.03428.

[73] PHAM C, LIU B, YANG Y, et al. Let models speak ciphers: Multiagent debate through embeddings[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: [s.n.], 2024.

[74] ZOU H, ZHAO Q Y, BARIAH L, et al. Wireless multi-agent generative AI: From connected intelligence to collective intelligence[EB/OL]. (2023-07-05). https://arxiv. org/abs/2307.02757.

[75] ZOU H, ZHAO Q Y, BARIAH L, et al. GenAINet: Enabling wireless collective intelligence via knowledge transfer and reasoning[EB/OL]. (2024-02-28). https://arxiv.org/abs /2402.16631.

作者简介:

瞿崇晓(1985-),男,正高级工程师,研究方向:智能物联技术、信息系统集成,E-mail: quchongxiao@cetc.com.cn。

唐宇波(1974-),男,正高级工程师,研究方向:作战模拟与仿真、计算机兵棋推演。

吴高洁(1981-),男,博士,研究方向:军事运筹、作战试验与作战仿真、兵棋系统建模。

范长军(1983-),**通信作者**,男,博士,研究方向:机器学习、大型语言模型,E-mail: fcjun @126.com。

张永晋(1982-),男,高级工程师,研究方向:无线通信技术、多模态大模型。

刘硕(1989-),男,高级工程师,研究方向:无线通信技术、云边协同计算。

(编辑:刘彦东)