通用大模型演进路线

任福继[1,2],张彦如[1,2*]

1. 电子科技大学计算机科学与工程学院,成都 611731
2. 电子科技大学(深圳)高等研究院,深圳 518110

摘要 随着人工智能技术的飞速发展,通用大模型(GLMs)已经成为人工智能领域的重要研究方向。通用大模型拥有超大规模参数,通过大规模数据进行训练,具备强大的学习和推理能力。这些模型在自然语言处理、图像识别、代码生成等多种任务中展现出卓越的能力。回顾了通用大模型的发展历程,梳理关键技术节点,从早期基于规则的系统和传统机器学习模型,到深度学习的崛起,再到Transformer架构,以及GPT系列及国内外通用大模型的进展。尽管GLMs在多个领域取得了显著进展,但其发展也面临诸多挑战,包括计算资源需求、数据偏见与伦理问题及模型的解释性与透明性。分析了这些挑战,并探讨了GLMs未来发展的5个关键方向:模型优化、多模态学习、具情感大模型、数据与知识双驱动以及伦理与社会影响。通过这些策略,通用大模型有望在未来实现更广泛和深入的应用,推动人工智能技术的持续进步。

关键词 通用大模型;人工智能;深度学习;Transformer架构;GPT系列

随着人工智能技术的飞速发展,通用大模型(general large models,GLMs)已经成为人工智能领域的重要研究方向,通常具备以下特点。

1)大规模。通用大模型通常拥有大量的参数,从几十亿至上千亿参数不等,通过大规模数据进行训练,从而具备强大的学习和推理能力。

2)预训练—微调。通用大模型通常采用预训练和微调的策略。首先在大规模未标注数据上进行无监督或自监督预训练,然后通过有监督的微调适应特定任务。

3)通用性。通用大模型具备广泛的适用性,可以处理不同类型的数据和任务,如文本、图像、音频等。

4)多模态。一些通用大模型能够处理多种模态的数据,如文本与图像结合,体现了广泛的应用潜力(图1)。

5)高度复杂。由于拥有大量参数和复杂的架构,通用大模型具备强大的表现力和学习能力,但是,同时也面临着计算资源需求高、模型解释性差等挑战。

收稿日期:2024-05-14;修回日期:2024-05-28

作者简介:任福继,教授,日本工程院院士、欧盟科学院院士、俄罗斯工程院外籍院士,研究方向为先进智能、情感计算、智能机器人等,电子信箱:renfuji@uestc.edu.cn;张彦如(通信作者),教授,研究方向为智能博弈与决策,电子信箱:yanruzhang@uestc.edu.cn

引用格式:任福继, 张彦如. 通用大模型演进路线[J]. 科技导报, 2024, 42(12): 44-50; doi:10.3981/j.issn.1000-7857.2024.05.00531

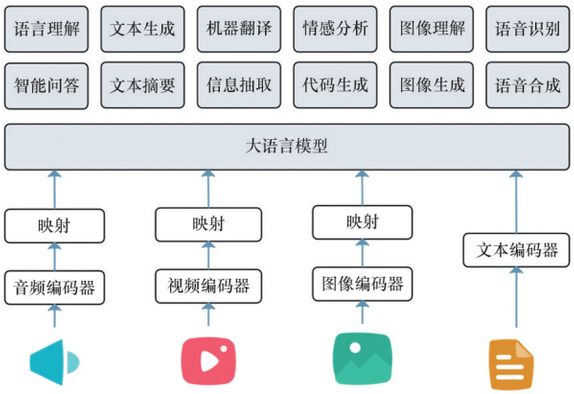

图1　多模态通用大模型

通用大模型为实现更高级的理解、交互和生成任务提供了可能,被广泛认为是推动人工智能技术向通用智能发展的关键因素[1]。自生成式预训练变换器(generative pre-trained transformer,GPT)系列模型问世以来,这一领域取得了长足的进步。随着以GPT为代表的大模型不断涌现,研究人员已深刻认识到通用大模型不仅代表着当今人工智能技术的前沿,更预示着未来智能系统的发展方向。

通用大模型的发展得益于深度学习的进步以及计算能力的提升。Transformer架构的引入,打破了传统循环神经网络在处理长序列任务时的瓶颈,开启了大规模预训练模型的时代。GPT系列模型进一步展现了通过大规模预训练来学习通用知识的潜力,为实现通用人工智能(artificial general intelligence,AGI)奠定了基础。

本文探讨通用大模型的演进路线,分析其发展历程、面临的挑战及未来可能的方向。

1　通用大模型的发展

1.1　早期模型

在通用大模型崭露头角之前,人工智能领域主要依赖于基于规则的系统和早期的机器学习模型。这些模型包括决策树、支持向量机(SVM)及朴素贝叶斯分类器等。虽然这些方法在特定任务上取得了一定的成功,但它们在处理复杂语言任务和大规模数据时显得力不从心。随着数据量的增大,简单的机器学习模型逐渐难以应对。

深度学习的崛起为通用大模型的发展奠定了基础。循环神经网络(RNN)和卷积神经网络(CNN)是深度学习的两大支柱。RNN擅长处理序列数据,被广泛应用于语言建模和语音识别等任务,而CNN则在图像处理方面表现出色。然而,这两类模型都存在固有的局限性:RNN难以处理长序列数据,存在梯度消失和梯度爆炸的问题;CNN在捕捉全局特征时效率较低。

1.2　Transformer架构的出现

Transformer架构的出现彻底改变了这一领域。Vaswani等[2]在2017年提出的Transformer架构,通过自注意力机制解决了RNN在处理长序列任务时的瓶颈问题。自注意力机制使得模型可以关注输入序列中的不同部分,从而有效地捕捉全局信息。Transformer的另一大优势在于并行计算能力。传统的RNN需要逐步处理序列数据,而Transformer则能同时处理整个序列,极大提升了计算效率。此后,Transformer架构成为众多通用大模型的基础,并广泛应用于自然语言处理、图像处理等领域。

1.3　GPT系列的发展

OpenAI公司的GPT系列模型是通用大模型的典范。GPT-1[3]于2018年发布,参数量达到15亿,引入了预训练和微调的框架,通过在大规模语料上进行无监督预训练,再通过监督学习进行微调,实现了出色的性能。GPT-2[4]在GPT-1的基础上进一步扩展了模型规模,展示了惊人的文本生成能力。GPT-3[5]于2020年发布,拥有1750亿参数,约是GPT-2的10倍。GPT-3在更广泛的数据集上进行了预训练,展现了强大的通用能力。除了出色的文本生成能力,GPT-3还能执行各种任务,包括翻译、问答、代码生成等。GPT-3的发布标志着通用大模型进入一个新的阶段。GPT-4[6]于2023年推出,OpenAI并未公开其确切的参数数量。根据行业内的推测和相关报道,GPT-4的参数量可能在数百亿到数万亿之间。它进一步提升了模型的规模和能力,并首次引入多模态功能。GPT-4能够处理文本、图像等多种输入形式,使其在广泛的任务中表现出色。GPT-4o在GPT-4基础上进行了优化,提

高了模型的处理速度和效率。相比于GPT-4,GPT-4o引入了改进的架构和训练方法,是OpenAI首个端到端训练的跨越文本、视觉和音频的新模型。截至目前(2024年),GPT-5尚未发布,但可以预见,未来的GPT模型将进一步提升模型规模和能力,并在多模态、持续学习等方面取得新的进展。GPT-5可能会进一步优化模型效率、增强多模态学习能力、提升模型的可解释性和公平性。随着计算能力和数据规模的进一步提升,GPT系列模型将继续引领通用大模型的发展方向(图2)。

图2 GPT系列模型的发展

1.4 国内外通用大模型的发展

全球不同国家和地区在通用大模型研究和发展上呈现出多样化的特点。美国作为该领域的先行者,相关研究机构和企业在开源社区中非常活跃,OpenAI、Google、Microsoft和Facebook等公司在大模型研究方面处于领先地位,发布了许多重要的通用大模型。例如,Google开发了多款大模型,其中T5[7](text-to-text transfer transformer)和PaLM[8](pathways language model)在设计和应用上都展示了高度的通用性,能够在多种自然语言处理任务中表现优异,其最新发布的Gemini 1.5[9]系列在各项性能评估中更是直追GPT-4。由前OpenAI研究人员于2021年创立Anthropic公司开发的Claude[10]系列模型,在多模态和语言能力上取得了显著成就。目前开源的大模型中,参数最大的是Meta(Facebook AI)发布的Llama 3[11]模型,其参数规模超过4000亿,是迄今为止开源的最大参数规模的大模型之一。这些模型在不同评测体系中表现优异,展现了强大的语言理解、知识推理、数学计算和多任务处理等综合能力。总的来说,美国在通用大模型领域的研究处于全球领先地位,并在模型规模和多任务处理能力方面树立了全球行业标杆。美国拥有世界上最强大的计算资源和基础设施,包括超级计算机和大规模数据中心,为训练超大规模模型提供了必要的支持,预计美国在该领域仍能保持较长一段时间的全球领先地位。

在通用大模型领域,中国也展现出了迅速的发展势头。国内的领军企业,如百度、阿里巴巴、华为、字节跳动、腾讯等纷纷推出了各自的大模型。由清华大学研发的ChatGLM[12]系列模型在多个评测中表现亮眼,与国际一流模型水平接近,且差距逐渐缩小。由百度开发的文心一言在中文语言理解、中文知识和中文创作上表现优秀。由字节跳动研发的豆包大模型在企业市场的定价极具竞争力,主力模型的价格为每1000个标码(tokens)0.0008元人民币,比行业价格便宜99.3%。大幅降低了模型推理的单位成本,有助于企业以更低成本加速业务创新。中国在大模型的研究和发展上,注重结合本土文化和市场需求,在大模型的应用落地上展现出巨大活力,在医疗、教育等行业涌现出了一大批

垂直应用。同时在政策支持和投资驱动下,中国在通用大模型研究和发展上正展现出强劲的追赶势头,在技术创新、人才培养、政策支持等方面不断加强,正快速缩短与领先国家的差距。同时也在积极探索国际合作和开源生态建设,力求在全球人工智能领域占据重要地位。

除了美国和中国,其他国家和组织也在开发通用大模型。在欧洲,德国初创公司Aleph Alpha发布了拥有700亿参数的预训练模型Luminous[13],英国的DeepMind(属于Google)开发了多款大模型,如Gopher[14]、Chinchilla[15]等,是欧洲在通用大模型领域的重要代表。在亚洲,日本东京工业大学正在基于自主研发的超级计算机"富岳"开发大模型,韩国的互联网巨头Naver和Kakao,移动运营商巨头KT、SKT,以及通信巨头LG都在开发大模型;阿联酋的技术创新研究所也在进行1800亿参数模型的研发。多个国家正在逐步构建自己的大模型研究和应用生态,其中也不乏通过全球研究人员合作研发的通用大模型。例如,由BigScience发布的BLOOM[16]系列模型,就是由Hugging Face协调,联合法国国家大型计算中心(GENCI)和高密度科学计算发展与资源研究所(IDRIS)组织共同参与的国际合作项目。随着全球科技创新的加速,各国在通用大模型领域的竞争和合作将进一步加强,推动整个行业的发展。

通用大模型的发展不仅改变了人工智能领域的格局,也为实现通用人工智能提供了新的思路。不同国家和地区在通用大模型的研究和发展上各具特色,正通过各自的优势和战略,共同推动这一前沿科技领域的进步。未来,随着模型规模和数据规模的进一步扩大,通用大模型有望在更多领域展现出其强大的通用能力;同时随着技术的不断成熟和应用场景的不断拓展,可以预见全球在这一领域将会出现更多合作与竞争并存的局面。

2　通用大模型的挑战

2.1　计算资源需求

通用大模型的性能与其规模密切相关。随着模型规模的扩大,通用大模型需要大量的计算资源进行训练和推理[17],包括图形处理单元(graphics processing unit,GPU)资源和电力。如此庞大的计算资源需求不仅提高了训练成本,而且限制了中小型研究机构进入该领域的机会,导致了资源分配不平等。此外,通用大模型的推理阶段也需要大量计算资源,特别是在需要实时响应的应用场景中。为了应对这些挑战,研究人员正在探索模型压缩[18-19]、知识蒸馏[20-21]等方法,以提高计算效率。然而,尽管这些技术可以降低计算资源需求,但模型的规模增长仍然远超技术的优化速度。

2.2　数据偏见与伦理问题

通用大模型通常在大规模的文本或多模态数据集上进行训练,这些数据集可能包含各种形式的偏见,如性别、种族、政治等方面的偏见。如果模型不加甄别地学习这些偏见,它们可能在生成内容时反映,甚至放大这些问题。这不仅可能导致不公平的决策,还可能引发一系列社会伦理问题。为了应对数据偏见与伦理问题,研究人员提出了多种方法,如通过构建更加平衡和多样化的数据集来减少偏见[22],通过公平性约束和损失函数调整来限制模型输出的不公平行为[23-24]。此外,提高透明度和设计问责制也是解决这一问题的重要途径,开发者需要对模型的潜在风险和影响进行评估,并采取措施防止滥用。

2.3　模型解释性与透明性

通用大模型的复杂性和庞大规模使得理解其内部工作机制变得非常困难[25]。由于其内部决策过程不透明,通用大模型的输出可能难以解释,这对某些应用场景尤其不利。例如,在医疗、法律等高风险领域,用户需要了解模型的决策依据,以建立信任并确保决策的可靠性。为了提高通用大模型的解释性,研究人员提出了多种解释技术,如对注意力机制的分析[26]、特征重要性分析[27]等。然而,这些方法的有效性在一定程度上受限于模型的复杂性。此外,透明性还涉及算法公平性、数据隐私等问题。为了增强通用大模型的透明性,研究人员需要对模型的设计、训练和部署过程进行全面的分析和记录,并确保用户可以获取相关信息。

3 通用大模型的未来发展方向

3.1 模型优化

通用大模型的规模和复杂性不断增加,因此,模型优化是一个关键的发展方向。人脑在处理效率和能耗方面具有显著优势,能够在极低能耗下迅速做出反应。相比之下,GPT模型虽然在参数规模上略有优势,但在能效和反应时间方面还有很大改进空间,因此,模型优化是一个关键的发展方向[28]。模型优化旨在减少计算资源需求,提高性能和节省成本。现有的优化方法包括模型压缩、知识蒸馏和高效架构设计等。通过模型压缩可以去除不必要的参数和结构,知识蒸馏则通过训练一个小模型来模仿大模型的行为,高效架构设计则通过简化网络结构或改进计算方式来提高效率[29-30]。模型优化不仅有助于降低训练和推理的成本,也为在资源受限的设备上部署通用大模型提供了可能(图3)。

图3 迈向通用人工智能

3.2 多模态学习

多模态学习是通用大模型的另一个重要发展方向。多模态学习旨在整合和处理多种类型的数据,如文本、图像、音频等。通过多模态学习,模型可以更全面地理解信息,并在更广泛的应用场景中表现出色。多模态预训练、多模态检索及多模态生成是该领域的主要研究方向。通过多模态预训练,模型可以在不同模态之间建立联系;多模态检索允许通过一种模态查询另一种模态的数据;多模态生成则可以根据一种模态的输入生成另一种模态的输出。多模态学习不仅提高了通用大模型的实用性,也拓宽了其应用领域。

3.3 具情感大模型

开发具有情感理解和表达能力的大模型被视为通用大模型未来发展的一个重要方向。大模型的推理能力在过去几年不断提升,虽然在许多任务上其智商表现出色,但在需要理解和表达情感的人机交互中,它在情商上的局限性可能影响大模型应用的广泛性和深入性。引入情感元素的大模型将把研发重点由大模型本身转移到对人和大模型的交互上[31],也就是先进智能推动的大模型,是通用大模型未来发展的趋势。

3.4 数据与知识双驱动

数据与知识双驱动为通用大模型的未来发展开辟了新的可能,通过这种复合策略,模型不仅能从大数据中学习,还能利用人类的知识体系进行更深层次的推理和决策,发挥两者的优势,弥补各自的不足。大模型的记忆机制一直是研究人员关注的问题,当前也有方案为大模型提供了处理无限长度序列的能力。但是否能够处理无限长度序列就能解决像人类智能那样的记忆问题?记得越多就越好吗?数据与知识双驱动是通用大模型发展的一个重要方向,这里的知识是指静态知识和动态知识[32],有望给大模型的记忆和忘却提供解决方案。

3.5 伦理与社会影响

通用大模型在取得技术进步的同时,也带来了伦理与社会方面的挑战。由于通用大模型在处理数据时可能引入偏见、不公平和隐私泄露等问题,伦理和社会影响成为重要的关注点。解决这些问题需要在数据采集、模型训练、应用部署等各个环节采取措施。确保公平性、保护隐私以及制定责任归属和问责机制是其中的重要任务。只有通过建立透明、可控和负责任的开发与应用框架,通用大模型才能实现真正的可持续发展,并为社会带来积极的影响。

4 结论

通用大模型代表了人工智能领域的一次重大进步。通用大模型的发展不仅体现了人工智能技术的

进步,也展现出了深度学习在多样化任务中的巨大潜力。

通用大模型随着规模和复杂性的增加,面临的挑战也愈发明显。通用大模型在计算资源需求、数据偏见与伦理问题、模型解释性与透明性等方面的挑战,对通用大模型的广泛应用提出了严峻的考验,也为研究人员提供了丰富的研究课题。为了应对这些挑战,通用大模型需要在未来继续优化和发展。通用大模型未来发展有5个关键方向:模型优化、多模态学习、数据与知识双驱动、具情感大模型以及伦理与社会影响。通过模型优化,可以提高通用大模型的效率和性能;通过多模态学习,可以扩展通用大模型的应用范围;通过赋予大模型情感特质,可以让大模型更擅长与人交互;通过数据与知识双驱动,可以赋予大模型更深层次的推理和决策能力;通过注重伦理与社会影响,可以确保通用大模型的负责任开发与应用。

总的来说,通用大模型代表了人工智能的未来发展方向。虽然面临诸多挑战,但随着技术的不断进步和人们对人工智能系统开发与应用的持续关注,通用大模型有望在未来取得更加卓越的成就。

参考文献(References)

[1] 陶建华, 聂帅, 车飞虎. 语言大模型的演进与启示[J]. 中国科学基金, 2023, 37(5): 767-775.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2017.

[3] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-06-11) [2024-05-14]. https://openai. com/index/language-unsupervised.

[4] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL]. (2019-02-14) [2024-05-14]. https://openai. com/index/better-language-models.

[5] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 33th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2020: 1877-1901.

[6] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J/OL]. [2024-05-04]. https://doi. org/10.48550/arXiv.2303.08774.

[7] Colin R, Noam S, Adam R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of machine learning research, 2020, 21(140): 1-67.

[8] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[9] Reid M, Savinov N, Teplyashin D, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[J/OL]. [2024-04-25]. https://doi. org/10.48550/arXiv.2403.05530.

[10] Anthropic. The claude 3 model family: Opus, sonnet, haiku[EB/OL]. (2024-03-04) [2024-05-28]. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc-618857627/Model_Card_Claude_3.pdf.

[11] Meta. Introducing Meta Llama 3: The most capable openly available LLM to date[EB/OL]. (2024-04-18)[2024-05-28]. https://ai.meta.com/blog/meta-llama-3.

[12] Du Z X, Qian Y J, Liu X, et al. GLM: General language model pretraining with autoregressive blank infilling[EB/OL]. [2024-05-28]. http://arxiv.org/abs/2103.10360.

[13] Aleph Alpha. Luminous performance benchmarks[EB/OL]. [2024-05-28]. https://aleph-alpha. com/luminous-performance-benchmarks.

[14] Rae J W, Borgeaud S, Cai T, et al. Scaling language models: Methods, analysis & insights from training gopher[EB/OL]. [2024-05-28]. http://arxiv.org/abs/2112.11446.

[15] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[EB/OL]. [2024-05-28]. http://arxiv.org/abs/2203.15556.

[16] Le Scao T, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model[EB/OL]. [2024-05-28]. http://arxiv.org/abs/2211.05100.

[17] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge, USA: MIT Press, 2022: 30016-30030.

[18] Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2015.

[19] Dettmers T, Pagnoni A, Holtzman A, et al. Qlora: Efficient finetuning of quantized llms[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge: MIT Press, 2023: 10088-10115.

[20] Wang Z Y, Huang S H, Liu Y X, et al. Democratizing reasoning ability: Tailored learning from large language model[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 1948-1966.

[21] 邵仁荣, 刘宇昂, 张伟, 等. 深度学习中知识蒸馏研究

综述[J]. 计算机学报, 2022, 45(8): 1638-1673.
[22] Qi J R, Fernández R, Bisazza A. Cross-lingual consistency of factual knowledge in multilingual language models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2023: 10650-10666.
[23] Shi Y X, Liu Z L, Shi Z, et al. Fairness-aware client selection for federated learning[C]//Proceedings of IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ: IEEE, 2023: 324-329.
[24] 古天龙, 李龙, 常亮, 等. 公平联邦学习及其设计研究综述[J]. 计算机学报, 2023, 46(9): 1991-2024.
[25] Shanahan M, McDonell K, Reynolds L. Role play with large language models[J]. Nature, 2023, 623(7987): 493-498.
[26] Ethayarajh K, Jurafsky D. Attention flows are shapley value explanations[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 49-54.
[27] 鞠天杰, 刘功申, 张倬胜, 等. 自然语言处理中的探针可解释方法综述[J]. 计算机学报, 2024, 47(4): 733-758.
[28] 刘学博, 户保田, 陈科海, 等. 大模型关键技术与未来发展方向:从ChatGPT谈起[J]. 中国科学基金, 2023, 37(5): 758-766.
[29] Du N, Huang Y, Dai A M, et al. Glam: Efficient scaling of language models with mixture-of-experts[C]//Proceedings of the 39th International Conference on Machine Learning(ICML). NY: PMLR, 2022: 5547-5569.
[30] Rajbhandari S, Li C L, Yao Z, et al. DeepSpeed-MoE: Advancing mixture-of-experts inference and training to power next-generation AI scale[C]//Proceedings of the 39th International conference on machine learning(ICML). NY: PMLR, 2022: 18332-18346.
[31] Ren F J. Construction of metaverse center based on advanced intelligence, 2022-Annual-Report-EU-Academy-of-Sciences[R]. Salzburg: European Academy of Sciences, 2023: 106-116.
[32] Ren F J, Shi H. A general ontology based multi-lingual multi-function multimedia intelligent system[C]//Proceedings of SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and their Complex Interactions' (Cat. No. 00CH37166). Piscataway, NJ: IEEE, 2000: 2362-2368.

Evolution of general large models

REN Fuji[1,2], ZHANG Yanru[1,2*]

1. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
2. Shenzhen Institute for Advanced study, UESTC, Shenzhen 518110, China

Abstract With the rapid development of artificial intelligence (AI) technology, general large models (GLMs) have become a significant research focus in the AI field. GLMs typically possess an extensive number of parameters, are trained on massive datasets and exhibit robust learning and reasoning capabilities. These models demonstrate outstanding performance in various tasks, including natural language processing, image recognition, and code generation.

This paper reviews the evolution of GLMs and the key technology nodes, from the early rule-based systems and traditional machine learning models to the rise of deep learning, the introduction of the Transformer architecture, and the advancements in the GPT series and other GLMS over the world. Despite the significant progress, GLMs face numerous challenges, such as high computational resource demands, data bias, ethical issues, and model interpretability and transparency. This paper analyzes these challenges and explores five key future development directions for GLMs: model optimization, multimodal learning, emotionally intelligent models, data and knowledge dual-driven models, and ethical and societal impacts. By adopting these strategies, GLMs are expected to achieve broader and deeper applications, driving continuous progress in AI technology.

Keywords general large models; artificial intelligence; deep learning; transformer architecture; GPT series

（责任编辑　王微）