科技新闻

·深度报道·

[编者按] 随着大模型和生成式人工智能(AI)的快速发展,AI彰显出了更强的创造力,在众多领域展现出了广阔的应用潜力与价值。中国科学技术协会主席万钢非常关注AI的发展与应用,于2024年11月11—12日深入调研了北京零一万物科技有限公司和北京深度求索人工智能基础技术研究有限公司,实地考察了AI领域的最新进展。结合调研情况与相关资料,本文剖析了2024年AI技术的突破与未来发展趋势。

“花”开有声:2024年人工智能技术突破与发展趋势

文/刘志远[1],孙鸿航[2]

2024年,人工智能(AI)大模型技术在多个领域取得显著进展,其应用场景不断拓展,尤其在金融、医疗和制造业等领域初步落地,显著提升了行业效率。技术上,大模型正朝着多功能化、小型化方向发展,通过增加参数、国产AI芯片自研和模型小型化等策略,提升处理复杂问题的能力。同时,大模型的参数规模不断扩大,从单模态向多模态演变,能够处理来自不同数据模态的信息。在商业化方面,大模型的商业模式主要包括应用开发、应用程序编程接口(API)服务和模型平台服务,其中API服务占据核心地位。此外,大模型在算力底座、推理分析、创意生成、情绪智能等方面展现出强大潜力,预示着AI正进入“机器外脑”时代。

1 人工智能技术的全面“开花”

在调研中,北京零一万物科技有限公司CEO李开复指出,大模型多模态能力将极大拓展AI原生应用边界。多模态大模型在2024年实现了快速提升,能够高效处理文本、图像、音频、视频等多种类型的数据,并实现跨模态推理。2024年12月10日,OpenAI公司正式发布了Sora模型,这一模型采用最新的文生视频技术,可根据用户输入的文本描述或图像生成长达20 s的高质量视频。同时,谷歌公司发布的Gemini 2.0和DeepMind公司推出的Genie 2也展示了多模态大模型在虚拟世界生成领域的突破,Genie 2甚至允许用户仅通过一张图片和简单文本描述生成“无限”种类的可玩3D世界,标志着AI在虚拟世界生成领域的重要进展。

在复杂推理能力方面,大模型技术更加注重推理性能的优化和工程化改进。斯坦福大学提出的Quiet-STaR模型利用强化学习优化中间推理过程,显著提升了零样本准确率和困惑度。北京深度求索人工智能基础技术研究有限公司的DeepSeek模型通过多头潜在注意力(MLA)和DeepSeek混合专家(MoE)算法创新,大幅降低了推理成本。2024年9月13日,OpenAI发布了最强推理模型o1,标志着AI行业开启了一个全新的纪元。

长上下文处理能力也得到显著增强。大模型通过改进Transformer架构、优化注意力机制、实现长期记忆、改进位置编码和上下文预后处理等方法来支持更长的序列长度。国际上,Anthropic的Claude 3和谷歌的Gemini 1.5 Pro已经支持长达1 M的上下文长度,OpenAI的GPT-4 Turbo支持128 K的上下文长度。在国内,Kimi智能助手的上下文无损输入长度达200万字,阿里通义千问和百度文心一言等相继宣布支持更长的长文本处理。

自监督学习与强化学习的结合被视为AI领域的一项重要进展。这种结合不仅提高了强化学习算法的策略和价值函数的优化效率,还提升了学习效率和性能。自监督学习通过从大量未标注数据中自动学习模式,提高生成精度;强化学习则通过调整生成内容的质量,使其更接近用户意图和期望,特别是在对话型AI中,引入了反映用户反馈以提高响应精度的机制。2024年,Meta公司推出了Llama系列的大语言模型,特别是Llama-3,它采用了优化的Transformer架构,并结合了自监

1. 科技导报社《前瞻科技》编辑部
2. 中国科学技术协会办公厅

督学习、有监督微调和基于人类反馈的强化学习(RLHF),这种结合方式不仅提高了模型的实用性和安全性,还在多个基准测试中超越了同规模模型。

模型架构的创新与优化也在不断推进。例如,谷歌公司的Gemini 1.5 Pro和阿里云计算有限公司的Qwen等都采用了MoE架构,以提高模型效率和性能。2024年12月26日,DeepSeek发布DeepSeek-V3,其利用MoE模型,实现了与世界顶尖的闭源模型GPT-4o及Claude-3.5-Sonnet相媲美的性能,且其训练成本仅为558万美元,不足GPT-4的1/20。此外,苹果公司的Ferret系列专注于端侧设备上的模型部署,通过ReALM项目提高了智能设备对屏幕的理解和用户响应速度。

在数据层面,合成数据的重要性日益凸显。合成数据可以由计算方法和模拟来创建,包括文本、数字、表格、图像、视频等,具有低边际成本、隐私保护、减少偏见等优点。例如,微软Phi系列强调使用高质量的小规模合成数据来训练更有效的模型,而不是单纯追求参数量的增长。从Phi-1到Phi-3,数据质量不断提高,同时模型尺寸保持克制增长,体现了"小而美"的设计理念。

在编码能力方面,随着大模型技术的发展,未来智能编码将实现低代码和无代码开发。DeepSeek-Code-V2在此方面表现极佳,不仅增强了编码和数学推理能力,还扩展了对编程语言的支持范围,从86种增加到了338种,并将上下文长度从16 K扩展到了128 K。据DeepSeek技术负责人介绍,该模型是全球首个在代码、数学能力上超越GPT-4-Turbo、Claude3-Opus、Gemini-1.5Pro等的开源代码大模型。

检索增强生成(RAG)技术在2024年也经历了架构上的优化和模块化发展。例如,Advanced RAG成为主流,效果优于Naive RAG且易于实现。RAG框架包括检索模块和生成模块,通过索引、检索和生成3个基本步骤来提升模型的性能。在检索方法上,Hybird检索(BM25+语义相似检索)和RRF Fusion(多种检索方式融合)成为主流,同时Rerank技术用于优化检索结果。

在安全与隐私保护方面,大模型的安全框架得到了进一步完善,涵盖了数据合规获取、数据标注安全、数据集安全检测、数据增广与合成、内生安全评测、鲁棒性增强、"幻觉"缓解、偏见缓解、可解释性提升、系统安全加固、插件安全保护、输入输出安全保护等多个方面。蚂蚁数科旗下的摩斯团队在NeurIPS 2024隐私挑战赛中展示了他们先进的隐私保护技术,确保了大语言模型训练数据的安全性和私密性,这对于推动AI技术的健康发展至关重要。

值得注意的是,虽然大模型依然主导市场,但"小而精"的模型正在崛起,特别是在边缘计算和移动设备上的应用。通过蒸馏(distillation)和剪枝(pruning)技术,大模型被压缩为轻量化版本,同时保留核心功能。小型生成式AI模型被广泛部署在边缘设备上,实现了离线运行和更低的延迟。如在智能工厂设备中部署轻量化AI模型,用于实时监控和优化生产线。通过云端大模型与本地小模型协同工作,既能提供强大计算能力,又能保证隐私和实时性。例如,设备在本地使用小模型处理敏感数据,必要时通过云端调用大模型进行复杂推理。

2 人工智能应用的多领域落地"生花"

2024年,"新质生产力"成为年度热词。李开复认为,AI不仅是"新质生产力"这一概念的良好落地范例,甚至是迄今为止最佳的落地范例,行业先行者将获得前所未有的市场机会。AI已经成为发展新质生产力的最强大技术之一,2024年其在多个领域取得了显著的应用与商业化进展,推动了各行业的智能化升级。

新一代AI助手变得更加智能,能够理解上下文并完成复杂任务。这些助手不仅能同时处理多个任务,还能在任务间共享信息,提供情感化的反馈。国内外硬件厂商纷纷推出搭载端侧AI模型的终端产品,如三星Galaxy Z系列、谷歌Pixel 9系列和小米汽车SU7等,未来将为消费者带来更便捷的智能体验。

在能源管理方面,大模型被用于能耗优化、气候建模和能源分配。例如,百度智能云与南方电网、国家电网山东电力等合作,通过AI技术实现了智能创作、设备巡检和电力调度等应用场景的优化。此外,Lepton AI的能耗优化技术也在数据中心、云计算、边缘计算等领域得到广泛应用,显

著降低了系统能耗。

医疗领域也受益于AI技术的进步。大模型被用于医学影像分析,帮助医生更快、更准确地诊断疾病。例如,Geneformer模型在有限数据条件下成功预测了基因网络变化,为精准医疗注入新动力。ShukunGPT大模型在医疗领域的突破性进展也得到了认可,其诊断准确率高达98%,显著提升了诊断效率。华为云的盘古大模型通过增加AI制药核心场景,药物设计效率提升了33%,优化后的分子结合能力提升了40%以上。

金融服务领域也迎来了AI技术的赋能。多家科技公司与金融机构紧密合作,将生成式AI能力应用在投资研究、投资顾问、风险控制、客户服务等场景。如蚂蚁金融大模型聚焦于真实金融场景需求,涵盖理财选品、产品评测、行情解读、资产配置等服务。

游戏娱乐领域也因AI技术焕发出新的活力。生成式AI贯穿从游戏开发、游戏发行到用户体验的全过程。例如,巨人网络公司发布的"千影QianYing"有声游戏生成大模型,包括游戏视频生成大模型YingGame和视频配音大模型YingSound,为玩家带来了全新的游戏体验。

法律服务领域也迎来了AI技术的应用。生成式AI技术可用于自动化合同起草、法律文件摘要、法律案例分析等任务,提高效率并减少人为错误。国外的CS Disco和中国的法大慧云等初创企业已推出法律相关的AI智能化解决方案。

在教育教学领域,生成式AI模型为学生和教师提供了强大的辅助工具。对于学生而言,生成式AI可提供自适应学习、口语陪练、智能排课、在线答疑等学习支持。对于教师而言,生成式AI可优化教学设计,创新规划单元教学设计、一键生成互动教学课件、智能教学评价,提升备课、教学和评价的效率和质量。讯飞星火认知大模型被广泛应用于智慧教育硬件产品及服务,如AI听说课堂、AI教研平台等。

3 人工智能未来绽放"绮花"

展望2025年,AI技术将持续受到高度关注,其发展前景令人充满期待。

据量子位智库发布的《2024年度AI十大趋势报告》显示,未来大模型将更加注重多模态数据融合。这种多模态数据融合能够利用不同模态之间的关联和互补,提高模型的表达、理解、创造和推理能力,从而在自动驾驶、艺术创作等领域带来实际应用的突破。然而,多模态数据融合也面临格式、特征和语义等方面的挑战,需要深入研究和优化。

同时,大模型将提升自适应和迁移学习能力,以满足多应用场景下的通用性、灵活性和效率需求。自适应能力使模型能够根据不同的应用场景自动调整参数和结构,而迁移学习能力则能让模型在一个任务上学到的知识应用到另一个相关任务上,从而加速学习过程并提高性能。

此外,模型的可解释性也将成为未来发展的关键。采用可解释性算法等技术手段将变得至关重要,这些算法能够解释模型预测结果,帮助人们理解模型的内部逻辑和决策过程,增加对模型的信任,推动AI技术的更广泛应用。与此同时,AI的发展还将向"负责任"的方向迈进,强调以合乎伦理、安全、透明、可靠和尊重知识产权的方式开发和部署。

垂直领域大模型的研发也是未来不可忽视的一个重要方向。通过聚焦具有深厚知识背景、高质量数据、稳定的数据供给、清晰规则和明确需求的行业领域,开展专用大模型的设计和开发,能够更有效地满足行业实际需求,提升行业效率,优化工作流程。

在关注大模型的发展中,也需要高度重视隐私保护与数据安全等问题。数据加密技术、匿名化处理、完善的访问控制机制以及合规与审计等措施,将共同确保用户数据的安全性和隐私性。与此同时,大模型的发展还需更加注重能效比与绿色计算。随着模型规模的不断扩大,能效优化和绿色计算技术的发展与应用将成为关键,包括改进模型架构和算法设计、采用高效环保的计算设备和能源利用方式,以及建立绿色计算标准和评估体系。

研判和前瞻技术、产业发展趋势,虽然量子AI初现端倪,处于起步阶段,但其巨大的发展潜力已经开始受到关注。2024年备受关注的具身智能也将成为驱动AI发展的新型驱动力。人形机器人作为具身智能的主要应用场景,将通过与环境的交互实现感知、认知、决策和行动等能力,推动AI在工业制造、物流运输、医疗护理、家庭服务、体育训练、自动驾驶等领域的落地。