超智融合为突破算力瓶颈提供有效路径

AI世界

◎本报记者 崔 爽

“人工智能大模型是新质生产力的代表，大模型和超级计算融合发展十分重要，我国需认真布局、考虑。”近日，中国科学院院士陈润生在2024中国算力发展专家研讨会上表示。

在这场由中国智能计算产业联盟与全国信标委算力标准工作组共同主办的研讨会上，超智融合技术路线的趋势与发展方向引发专家学者热议。

在数字化、智能化浪潮中，算力正成为经济社会高质量发展的重要驱动力。但千行百业的计算场景纷繁复杂，单一计算架构无法应对。与会专家学者认为，超智融合兼具超算的强大处理能力与智算的算法优化能力，二者融合发展已成大势所趋。

图为国家超级计算成都中心主机房“硅立方”。 新华社记者 刘坤摄

迈出探索性步伐

资料显示，当前流行的预训练大模型具有数十亿乃至上万亿参数，训练时用到数万亿个Token（大模型用来表示自然语言文本的单位），训练的过程就是海量数据处理的过程，这消耗了巨大算力。

据美国开放人工智能研究中心（Open AI）测算，2012年开始，全球大模型训练所用的计算量呈指数级增长，平均每3.43个月便会翻一番。目前，计算量已远超算力增长速度。

“大模型的迅猛发展体现出新质生产力的特质，但目前遇到了算力瓶颈。”中国科学院计算技术研究所研究员张云泉说。

北京应用物理与计算数学研究所研究员袁国兴说：“现在的应用越来越复杂，不同应用需要不同算法，对计算机也有不同要求。”

张云泉认为，中国在超算领域拥有深厚技术积累，超智融合有望化解这些挑战。

国家信息中心信息化和产业发展部主任单志广说，超智融合随着基础算力、智算算力、超算算力等应用多元化发展而诞生。这一技术通过混合型算力资源或融合型算力体系，可同时满足多种不同算力的应用需求。

实际上，超智融合技术正成为近年来全球计算领域热点话题。今年5月，在以“重塑超算”为主题的国际超算大会上，超智融合相关方案遍地开花。

在我国，超智融合技术已被应用于超算互联网建设。今年4月，国家超算互联网平台上线，标志着我国在超智融合领域迈出探索性步伐。平台依托一体化算力调度、数据传输、生态协作体系，实现算力供给、软件开发、数据交易、模型服务等产业链相关各方紧密链接，构建市场化、互联网化、标准化的先进计算服务环境。

数据显示，平台上线以来，已有超200家应用、数据、模型等服务商入驻，并提供超3200款商品。这些商品覆盖科学计算、工业仿真、人工智能模型训练等领域，可满足全社会对先进计算服务的需求。

增强软硬件协同

不过，要更好实现超智融合，仍需大量创新探索。

陈润生认为，发展大模型与智算，不仅要改进应用层面的模型和算法，还要在基础理论层面有所突破。在他看来，随着模型规模扩张，一味“堆芯片”并不可取。根本上还要向人脑学习，把空间复杂度、时间复杂度压缩得更小，以更低能耗实现更高性能。

此外，软硬件协同创新程度有待进一步提升。

中国科学院院士钱德沛认为，在硬件方面，要尽量以最低能耗实现最高性能。未来不一定要做面面俱到的硬件，可重构或柔性或许是主要发展路径。而在软件方面，要从基本大模型理论出发，形成完整支撑人工智能的软件栈。

“我国一些超算中心已能为大模型训练提供支撑，未来还应重点围绕国产算力芯片发展关键软件，进一步实现软硬件协同优化。”中国工程院院士郑纬民说。

中国信通院云计算与大数据研究所所长何宝宏认为，传统超算和智算训练，对底层基础设施的要求各不相同。“这需要判断在什么场景下实现兼容统一，在哪些场景下凸显各自独特性。”何宝宏说。

呈现三阶段演进

在通用性与专用性之间，应如何选择超智融合的技术路线？与会专家学者普遍认为，应保持一定通用性。尤其在技术和方法论持续发展的背景下，应保持芯片、系统与软件的普适性，为研究提供广阔空间，深化底层理论与方法探索。

对此，单志广提醒，未来一体化算力体系的构建，要做好算力资源和业务应用的统筹衔接。须避免有效应用需求不足、缺乏网络服务质量保证、没有成熟调度体系的普遍性算力互联，不能脱离实际应用需求进行异地计算和远地计算算力设施布局，要从算力资源供给侧和业务应用需求侧两个维度进一步深化研究。

未来，超智融合具体将以何种路径演进？钱德沛认为，其将沿着超算支撑人工智能应用、用人工智能技术改进超算、超智实现内生融合这三个阶段清晰演进。

他进一步解释，在第一阶段，对现有计算机系统进行改造与升级。要发展专用硬件，确保可高效支持和执行人工智能任务，为人工智能研究提供坚实基础设施。在第二阶段，用人工智能改造传统计算。一方面要用人工智能的方法求解传统超算问题，另一方面人工智能也将影响传统计算机的结构，这个趋势会逐渐明显。在最终阶段，计算机系统将呈现内在的智能特性。人工智能不再是一种外加能力，而将成为计算机的核心属性和基本组成，可能计算能力或智能化水平会远超今天的超算或智算。

新增功能支持动漫风格

我国自研视频大模型全球上线

科技日报讯（记者**崔爽**）记者8月3日获悉，亮相2024中关村论坛年会的人工智能视频大模型Vidu日前宣布在全球正式上线。Vidu开放文生视频、图生视频两大核心功能，提供4秒和8秒两种时长选择，分辨率最高达1080P。

Vidu是北京生数科技有限公司（以下简称“生数科技”）联合清华大学发布的自研长时长、高一致性、高动态性视频大模型。据介绍，此次面向全球上线，Vidu在基础功能外新增动漫风格、角色一致性等功能。生数科技有关负责人表示，Vidu实现了业界最快实测推理速度，仅需30秒就能生成一段4秒片段。目前Vidu无需申请，用户直接使用邮箱注册即可上手体验。

今年初，文生视频大模型Sora在全球引发广泛关注。目前业界对视频模型的评价主要围绕三大核心维度：语义理解准确性、画面美观性、主体动态一致性。Vidu较好平衡了这三方面的表现。它能准确理解并生成提示词中的文字，包括字母、数字等，并能生成文字特效。对第一人称、延时摄影等镜头语言，Vidu也能精准表达，用户只需细化提示词，即可大幅提升视频可控性。同时，Vidu支持大幅度、精准的动作生成，保持高流畅、高动态的画面效果。

此外，Vidu在构图、叙事和光影等方面，能达到接近电影级效果。Vidu还能生成影视级特效画面，如烟雾、炫光效果、CG（计算机图形学）特效等。

Vidu生成的视频截图。 生数科技供图

朱松纯在研讨会上表示——

通用人工智能关键在立“心”

◎本报记者 何 亮 实习生 胡轶慧

“很多人将规模定律（Scaling Law）奉为圭臬，认为只要数据越多、算力越强、模型参数越大就行了。但我认为，要实现通用人工智能，仅靠数据是不够的，我们需要探索另外一种路径——为人工智能立‘心’。”近日，在湖北鄂州召开的莲花山研究院二十周年学术思想研讨会上，北京通用人工智能研究院院长朱松纯分享了他的思考。他认为，通用人工智能已成为全球科技竞争制高点。要在科技竞争中取得突破，关键是厘清大数据源头，定位好人工智能发展方向。

“如果无法处理视觉数据，人工智能系统就只剩一个空架子。”在朱松纯看来，数据标注就像为计算机戴上一副特殊的“眼镜”，让其具备识别并理解图像、文本和其他数据细节的能力。

1997年，斯科特·科尼什（Scott Konish）完成了世界上第一个数据集的标注——图像边界，用来训练分类器。也正是看到了统计对图像理解的可能性，2004年朱松纯开启大规模高颗粒度数据标注工作。

“2008年，我和团队成员在数据标注上遇到两个瓶颈。”朱松纯告诉记者，其一，价值、因果、意图等要素潜藏于感知数据表象之下，无法被传感器直接探测，更难以标注；其二，数据标注的过程与特定任务高度相关，不同任务要求不同的标注方法，继续扩大数据或模型规模，仍然无法提升泛化能力。这让朱松纯对通用人工智能有了更深入思考。

在朱松纯看来，通用人工智能是由计算机视觉、自然语言处理等核心领域构成的复杂巨系统，其研发道路之艰难好比“登月”；而大数据路线就好比“攀登珠峰”，两者目标相差甚远。

那么如何探索通用人工智能这条道路呢？朱松纯认为，人工智能研究需要由“理”向“心”转变。“理”是数理模型，“心”是认知架构、价值对齐。

“经过近30年发展，人工智能多个核心领域已然呈现对内融合、对外交叉的发展态势，朝着通用人工智能方向推进。”朱松纯说，在融合过程中，必定会形成统一的人工智能架构，以实现从解决单一任务为主的专项人工智能向解决大量任务、自主定义任务的通用人工智能转变。

在朱松纯看来，为机器立“心”，实现由“理”到“心”的过渡以及从大数据到大任务、从感知到认知的飞跃，是未来10—20年的学术前沿，也是智能学科需要承担的核心使命。

2023年市场规模同比增长72.5%

我国智算服务释放巨大潜力

IT之窗

◎王 鹏

国际数据公司（IDC）近日发布的《中国智算服务市场（2023下半年）跟踪》报告显示，2023年下半年，中国智算服务市场整体规模达114.1亿元，同比增长85.8%。全年来看，这一市场规模达194.2亿元，同比增长72.5%，动力强劲。

智算服务是一种专门提供智能计算能力的服务，广泛应用于机器学习、深度学习、自然语言处理等人工智能（AI）领域，并助力生命科学、工业制造、自动驾驶等行业智能化转型升级。智算服务“算”出AI落地加速度，让更多创新企业迎来发展新机遇。

双轮驱动增长

我国智算服务市场由智算集成服务、智算基础设施即服务（AI IaaS）两大板块构成，后者进一步细分为面向生成式人工智能的GenAI IaaS和面向非生成式人工智能的Non-GenAI IaaS。从2023年智算服务市场表现看，智算集成服务市场贡献了主要增量，去年下半年规模达36亿元，同比增速高达129.4%。GenAI IaaS市场在2023年实现“从0到1”的爆发式增长，规模达32.2亿元。Non-GenAI IaaS市场保持稳健增长态势，规模达45.9亿元。

我国智算服务市场快速发展，得益于智算服务需求与宏观政策引领双轮驱动。

一方面，我国智算服务需求保持高速增长。从“大炼模型”到“炼大模型”，千亿、万亿参数大模型的孵化，推动智算基础设施加快建设，智能算力需求持续爆发，各方纷纷加大对AI算力的布局。各地政府前瞻规划，现已投产上线的智算中心数量接近百个，可用算力已接近万PFlops（1PFlops=1千万亿次浮点运算/秒）。同时，人工智能技术公司、信息与通信技术服务商、数据中心服务商、实体企业等多方主体纷纷重金投入智算市场。

另一方面，宏观政策出台发挥引领作用，营商环境和创新生态持续改善。国家发展改革委等部门联合发布的《全国一体化大数据中心协同创新体系算力枢纽实施方案》提出，在京津冀、长三角、粤港澳大湾区、成渝，以及贵州、内蒙古、甘肃、宁夏等地布局建设全国一体化算力网络国家枢纽节点。北京、浙江、广东等地也纷纷发布相关政策规划，央地协同推动智算市场发展。政府加快实施“东数西算”工程、发放“算力券”，提升跨区域算力调度水平，降低企业使用算力设施和服务的门槛，并对示范效果突出的智算中心给予资金奖补，使营商环境进一步改善。随着云计算、大数据、深度学习、自然语言处理和图像识别等技术发展，智算能力大幅提升，创新生态不断优化。

图为位于浙江嘉兴的智能计算产业园。 邱道岑/视觉中国

机遇前所未有

智能算力作为新型基础设施，是万千行业进行智能化变革的核心驱动力，我国智算服务市场正迎来前所未有的发展机遇。智算服务市场蓬勃发展为行业带来新机遇。

受益于AI领域先发优势和资源技术积淀，目前字节跳动、阿里巴巴、百度、腾讯市场份额位居我国智算服务市场前列。同时，一些智算服务企业也纷纷紧抓智算市场发展窗口期，快速获取GenAI IaaS市场份额。商汤科技凭借资源储备、“大模型+大装置”全方位技术能力成功跨界，表现抢眼；并行科技凭借在超算领域的积累赢得一席之地；首都在线以互联网数据中心资源为基础整合产业生态，分得一杯羹……未来，随着技术快速发展，新的行业巨头有望产生。

随着智算市场快速增长，智算技术逐步渗透到农业、金融、医疗等领域，AI与各行业融合程度日益加深。AI大模型通过优化生产流程、提升管理效率等方式助力传统产业数字化转型升级，催生新模式和新业态。例如，智能网联汽车行业就在智算支撑下驶上了“快车道”。

合力共促发展

还要看到，智算服务市场快速增长的背后也伴随着挑战。

首先，数据安全与隐私保护问题亟待解决。智算服务涉及大量用户数据和企业机密信息，这些数据可能被恶意利用。其次，智算行业面临前沿技术待突破、高端人才短缺等问题，可能难以适应智算应用场景复杂化的发展趋势。此外，人工智能训练和推理需要大量算力，高性能计算设备运行过程中也会产生大量热量，能耗问题不容忽视。最后，智算持续发展需要跨领域合作和资源共享，而这往往面临交互标准与互操作性、供应链管理、盈利模式与利益分配等问题。

面对智算市场发展的一系列挑战，政府和企业应形成合力，共同推动行业良性发展。

第一，加强智算技术自主研发。可设立智算产业投资基金，吸引金融机构、企业等多元资本，对开展高端智算自主研发的企业进行投资。推动企业加大对智算前沿技术的研发投入，搭建智算技术的适配、验证与调优平台，推动产学研用深度融合。

第二，建立健全数据安全和隐私保护机制，实现数据态势可知、威胁可现、风险可控。要重视数据质量核验任务，完成数据质量规范性、一致性、准确性和完整性检查。

第三，积极培养高水平数据科学家和分析师人才。建立以应用能力为基础、以工作过程为导向的课程体系，构建集生产实训、虚拟仿真为一体的实习实训基地，促进高水平人才供给。

第四，最大限度节省能源消耗。算力基础设施建设应优先考虑算力中心地理位置，充分利用自然条件散热降温。合理安排调度业务进程，优先处理用户驱动型业务，在计算资源闲时处理结果驱动型业务，提高资源利用效率。

（作者系北京市社会科学院副研究员）

图说智能

人机对弈

在第八届中国—南亚博览会上，数字经济、人工智能、绿色能源、低空经济等领域科技范十足的展品吸引众多观众。图为小朋友在博览会上和智能机器人下棋。 新华社记者 胡超摄