



“人工智能+”行动正当时

DeepSeek 的一小步 国产AI芯片生态突围的一大步

中经记者 李玉洋 上海报道

在人工智能训练和推理加速的竞赛中,浮点数(Floating Point)的表示方式正成为关键突破口。

随着 DeepSeek-V3.1 的发布,算力“破壁人”——深度求索(DeepSeek)在这方面更进一步:继成功训练出世界首个使用 FP8(8 位浮点数)精度的开源大模型 DeepSeek-V3 后,在 DeepSeek-V3.1 中又使用了 UE8M0 FP8 Scale 的参数精度,并置顶留言“UE8M0 FP8 是针

实现两倍的浮点算力提升

DeepSeek 通过 DeepGEMM 开源库实现了工程化突破,为整个国产生态在大模型中的落地提供了可能。

计算机里的小数是用科学计数法表示,只不过不是10的次方,而是2的次方。浮点数则是计算机用于表示小数的核心手段,由符号位(Sign)、指数(Exponent)和尾数(Mantissa)三部分构成。其中,符号位决定正负,指数决定小数点的位置,尾数影响精度。

在 AI 计算领域,数据精度格式长期被国际巨头所主导。比如眼下被业内推崇的 FP8(FP 代表浮点数,8 代表数据用 8bit 即 8 位 0、1),是一种较新的数据格式,能显著降低显存占用和计算资源需求,同时保持较高的计算精度。

DeepSeek 在开源社区 Huggingface 放出信息:DeepSeek-V3.1 使用 UE8M0 FP8 缩放数据格式进行训练,以确保与微缩放数据格式兼容。这里所谓“微缩放数据格式”,即业界的 Microscaling FP8 (MXFP8) 标准。这是 Open

对即将发布的下一代国产芯片设计的”。

UE8M0 FP8 是什么? 下一代国产芯片又是什么? 这个充满谜团的声明带火了 UE8M0 FP8,也迅速传导至 A 股资本市场,国产芯片及半导体上市公司股价应声大涨,其中寒武纪(688256.SH)一路高歌猛进,总市值已跃居科创板头名。

《中国经营报》记者注意到,DeepSeek V3/R1、Kimi-K2 等主流开源大模型均原生生态支持 FP8 的低精度格式,这代表了低精度

Compute Project 在 2023 年发布的 Microscaling (MX) Formats Specification v1.0 里定义的 8bit 微缩块格式。

UE8M0 是 FP8 的一种特殊的数字表示格式。U 表示没有符号,E8 表示 8 位指数,M0 表示没有尾数。DeepSeek 在 V3.1 模型中提出的 UE8M0 FP8,相当于一条别人没走过的路。有观点认为,DeepSeek-V3.1 把原本只在硬件/内核实现层面存在的 scale 表达方式,提升到模型部署和训练标准,并适配国产下一代芯片,需要说明的是,UE8M0 FP8 并非 DeepSeek 独创,但 DeepSeek 通过 DeepGEMM 开源库实现了工程化突破,为整个国产生态在大模型中的落地提供了可能。

摩尔线程方面表示,UE8M0 是指无符号 8 位指数的 FP8 格式,用于对 FP8 张量数据进行分块缩

量化在未来 AI 大模型领域中的重要性。

“DeepSeek 距离上次的版本发布已经有段时间了,所以这一轮的发布特别受到关注。UE8M0 FP8 指的是为中国市场而特别定制的模式格式,此格式对中国的芯片厂商比较友好,基于它们的芯片规格而设置。”行业研究机构 Omdia 人工智能首席分析师苏廉节告诉记者。

事实上,DeepSeek 已经成为一个游戏规则的改变者,改变了大模型竞赛“谁拥有算力谁才能



DeepSeek-V3.1 针对下一代国产芯片设计。

本报资料室/图

放,提升表达精度;通过提升 FP8 张量表达精度,来提升大模型训练推理的精度。

“摩尔线程旗舰 AI 训推产品 MTT S5000 是国内首批原生支持 FP8 并已大规模量产的 GPU。”摩尔线程方面表示,其 MUSA 架构原生支持硬件 FP8 张量加速计算,能完美支持 UE8M0 FP8 Scale,利用硬件原生 FP8,相对于传统的 FP16 计算能够实现两倍的浮点算

胜出”的游戏规则。

对于 DeepSeek 针对下一代国产芯片引入 UE8M0 FP8 Scale 参数精度的意义,摩尔线程方面表示,意义在于“能够充分发挥已量产的国产芯片架构特性,通过国产大模型和国产芯片协同设计优化,实现 1+1>2”。

东莞证券在研报中指出,该格式通过更高的灵活度支持复杂模型推理,提升芯片的解码效率与运算能力,为国产芯片适配更大模型提供技术路径,国产 AI 算力生态有望加速成型。

力提升、访存和通信带宽效率提升和存储容量利用率提升,同时最优化张量表达精度。

苏廉节指出,UE8M0 FP8 是特别为中国芯片厂商准备的,尽管它的精度没常规的 FP8 高,也更依赖于训练、量化、校准等算法补偿和硬件支持。

截至目前,记者了解到,寒武纪、华为昇腾、海光、沐曦、摩尔线程等芯片公司支持 FP8。

国产算力生态协同的积极进展

随着中国人工智能行业的不断发展,国产芯片+国产引擎+国产模型的生态协同将不断深化。

曾几何时,“大模型需大力出奇迹”“精度不能降低,否则模型会崩”的论调左右着行业认知,彼时业内多采用如 FP32、FP16 或 BF16 这样的传统浮点格式。

然而,在 AI 训练和推理方面,高精度格式向低精度格式过渡的趋势已经出现。比如,英伟达开发者论坛上发布的一篇技术博客展示了 FP8 格式的高效性,它能够在不牺牲性能的情况下将显存占用率减半。

值得一提的是,英伟达多年来也一直用低精度数字表示法提升推理和训练效率,并具有强大的生态推动力,包括 DeepSeek 使用的 FP8 和微软使用的 FP4,都是由英伟达率先产品化并推向市场。

据壁仞研究院研究人员观察,自 OpenAI 开源首个原生态支持 FP4 格式的 GPT-OSS 系列大模型,AI 领域实际上已进入低精度计算时代。

“在 AI 领域中,特别是大模型场景下,模型普遍对数值精度不敏感,不像 HPC 领域中需要 FP64 甚至更高的精度格式。从实践来看,模型的参数量规模远比高精度带来的收益要大,这一点在如今混合专家大模型的时代得到很好的验证,包括目前大部分的主流开源大模型均采用 MoE 的结构。”壁仞研究院研究人员这样总结。

DeepSeek 对 UE8M0 的调整,其实提供了 FP8 格式的变体,可以进一步降低对计算能力、存储和带宽的需求。有观察人士表示,中国

的人工智能产业正在进入一个软硬件紧密合作的新时期。中国芯片公司可以逐步建立自己的 FP8 生态系统。UE8M0 FP8 的设计能“最大限度地利用硬件计算能力”。

“这是国产算力生态协同的积极进展的体现。”清程极智 CEO 汤雄超表示。

汤雄超还指出,过去,企业大模型落地的一大痛点是“最先进的模型与最难获取的硬件绑定”,目前这个情况已经获得一定程度的改善。

其中一点得益于国产算力芯片、国产开源模型以及国产推理引擎的生态协同。他还提到,中国工程院院士、清华大学教授郑纬民在今年 WAIC(世界人工智能大会)上提到,大模型推理引擎是人工智能产业竞争的关键,因为推理引擎是最终决定什么模型能够部署什么芯片的关键一环。

今年年初,清程极智推出的赤兔推理引擎率先实现了国产算力运行 FP8 原生精度 Deepseek-V3 模型;而本次最新发布的 Deepseek-V3.1 模型,赤兔推理引擎也在发布当天实现了昇腾、沐曦、海光等国产算力芯片的适配。

“本次 DeepSeek 新模型使用 UE8M0 数据格式有利于充分发挥下一代国产算力芯片的潜能。我们相信随着中国人工智能行业的不断发展,国产芯片+国产引擎+国产模型的生态协同将不断深化,最终让用户第一时间用上部署在国产算力上的国产大模型。”汤雄超表示。

AI+产业发展提速:智算规模提升 应用多点开花

中经记者 谭伦 北京报道

随着新一轮科技革命与产业变革的加速演进,我国人工智能产业正走向产业应用阶段,成为

推动经济社会高质量发展的新质生产力。

8 月 26 日,国务院印发《关于深入实施“人工智能+”行动的意见》(以下简称《意见》),其中,不仅

提出了涵盖“人工智能+”科学技术、产业发展、消费提质、民生福祉、治理能力、全球合作在内的六大重点行动方向,还提出要强化提升模型基础能力、强化智能算力统

筹在内的八项基础支撑能力。

《中国经营报》记者了解到,最近两年,在供给端与需求端双轮驱动下,中国人工智能产业保持高速扩张。中国互联网络信息中心统

计数据显示,2024 年我国人工智能产业规模已突破 7000 亿元,连续多年保持 20% 以上的增速,生成式 AI、行业模型与垂直场景化服务则成为增长最活跃的产业细分领域。

筑牢 AI+底座 算力基础设施加速完善

“人工智能+”行动的顺利推进,离不开强大的算力支撑。

工业和信息化部最新数据显示,截至 2025 年 6 月底,我国在用算力中心标准机架达 1085 万架,智能算力规模达 788EFLOPS(每秒百亿亿次浮点运算),干线 400G 端口数量大幅增加至 14060 个,存力总规模超过 1680EB(艾字节,存储容量的度量单位),全国算力中心平均电能利用效率(PUE)降至

1.42,算力基础设施规模和水平不断提升。

而智算基础设施正是这场“算力基建”浪潮的核心。Omdia 电信战略分析师杨光向记者表示,近年来,我国在智算基础设施建设上取得了长足进展,构建起了覆盖全国、协同发展的算力网络。这一布局的核心,在于打破传统数据中心的局限,建立集人工智能芯片、智能计算集群、高速网络为一体的新

型基础设施。

其中,过去两年,以电信运营商为代表的 AI 基础设施建设厂商,在算力与网络两条底座线上同时发力,形成“网络+算力+云边缘”协同的基础能力体系。从国家级超算中心、区域智算中心到企业自建机房、云厂商的专用智算集群,呈现出“多主体并进、分层布局”的格局。

以中国移动为例,记者从

“2025 中国算力大会”上获悉,其最新推出的算网大脑 3.0,深度融合了 AI 升级,全量纳管现网超 80EFLOPS 多样化算力,以及超 25 万个网络链路资源。

中国移动副总经理李慧镪认为,我国算力发展正呈现结构、形态、效能三个“转变”。其中,算力结构由以“训练为主”向以“推理为主、训推一体”转变,算力形态由“集中化”向“分布式”转变,算力效

能释放由“硬件横向扩展”向“软硬垂直优化”转变。

据其预计,随着大模型轻量化、开放化发展激发推理应用爆发,未来推理算力有望达到训练算力的 10 倍规模。在此背景下,他表示,当前我国算力网络正加速迈向以智为核心、算网智一体化的算力网络 3.0 新阶段,并最终成为支撑我国数字经济发展和人类社会智能化转型的新质生产力。

智能新业态剧增 AI 应用场景多点开花

如果说智算基础设施是“人工智能+”的坚实底座,那么在产业领域的深度应用,则是其价值的最终体现。当前,我国“人工智能+”正从平台能力向行业深度渗透,医疗、教育、制造、金融和消费服务等领域涌现出一批可复制的典型案例。“人工智能+”产业应用正呈现出从试点走向规模、从通用走向行业的发展走势。

在制造与矿业场景中,行业大模型正与多模态感知结合,把“经验密集型”作业转为“数据驱

动型”流程。以“盘古矿山”为代表的行业模型为例,记者从华为了解到,其围绕带式输送、井下巡检、选煤智能控制等环节进行“感知—决策—执行”的闭环重构,目标直指安全生产、良率与能耗等核心指标的提升;同时,在电力系统,国家电网与南方电网相继发布了行业大模型,聚焦设备状态评估、隐患识别、负荷预测、调度辅助与应急处理等,推进了数字化资产全生命周期管理。

在通信网络与城市基础设施

领域,“自智网络”则在进一步走向高阶智能化阶段。如中国移动提出了“3+13”大模型智能体与 L4 高价值场景解决方案,围绕装维、优化、开通、运营支撑等环节落地。记者从中国移动方面获悉,其相关方案已在国际电信联盟与 TM Forum 体系内获得多项荣誉,显示出网络从“可观测、可自动化”向“可感知、可决策、自优化”的能力加速进阶。

在医疗健康领域,AI 应用正在显著改善医疗服务的可及性和

效率。AI 辅助诊断系统,能够通过深度学习分析医学影像(如 CT、MRI),为医生提供更精准、更快速的诊断建议。

在药物研发领域,望石智慧创始人兼 CEO 周杰龙向记者表示,传统药物研发流程中,研发人员主要依赖于药物化学专家的灵感和经验,缺乏高效的工具来辅助药物设计。但依托 AI 大模型,融合几何深度学习等算法,药物研发企业能够精准地生成与靶点口袋结构契合的分子,实现更高

效、更准确的蛋白质比对和合成。据其透露,目前公司在多模态 AI 小分子生成大模型已经取得了显著成果,其中部分项目更是进入了临床阶段。

杨光认为,这将为我国 AI+行动的顺利实施奠定良好基础。记者了解到,此次《意见》提出的六大重点行动中,着重强调了将培育智能原生新模式新业态、推进工业全要素智能化发展、加快农业数智化转型升级、创新服务业发展新模式等举措。

应用早期阶段,还远远没有到成熟定型的阶段,因此,在兼顾应用的同时,的确应继续加强发展基础大模型,并加强专业大模型研发。

虽然挑战重重,但总体来看,对于《意见》的出台,产业链普遍认为,随着此次政策路线的发布,我国人工智能产业自身的技术演进、应用落地和生态完善,有望在未来几年迎来加速发展,并成为我国经济高质量增长的重要引擎。

上接 C1

在 AI 手机与 AI PC 蓬勃发展的同时,人形机器人这些具身智能终端也在智能浪潮中崭露头角,从市场数据来看,多家企业已在量产方面取得显著进展。自今年 6 月份启动量产以来,松延动力仅在 6 月和 7 月两个月内,就实现了人形机器人交付量破百。

袁帅表示,目前,新一代智能终端和智能体的应用普及尚处于初级阶段,虽然在一些领域和场景中得到了应用,但整体普及率仍有待提升,用户对于这类新型智能产品的认知度和接受度还需进一步提高。

谈及《意见》政策出台的背景,国家发展改革委有关负责同志在接受媒体采访时表示,智能化新产品、新场景竞相涌现,智能网联汽车、智能穿戴、智能家居等新一代智能终端不断推陈出新,智能机器人进工厂、进家庭渐成趋势,各类智能体已在生产生活领域发挥重要作用。与此同时,仍存在对人工智能作用认识不一致、供需对接不畅、应用落地存在“最后一公里”障碍等突出问题,亟须加强战略引导和统筹指导,推动全社会深刻认识人工智能对生产力的革命性影响,全面推进人工智能科技创新、产业发展和赋能应用,为加快培育发展新质生产力、推动高质量发展提供新动能。

《意见》的发布,也将给智能终端产业链带来深刻变革。周迪分析,上游层面,芯片、传感器、操作系统等核心技术研发将加速推进;中游层面,设备制造环节的智能化水平将提升;下游层面,教育、医疗、交通等服务领域将与 AI 深度融合。

“一方面将促进芯片、传感器、操作系统等核心技术的研发和创新,提升产业链的整体竞争力;另一方面,也将带动智能终端产品的升级和迭代,满足市场多样化的需求。”袁帅指出。