美国国家工程院外籍院士沈向洋：

合成数据成未来大模型训练关键

本报记者 宋婧

近日，美国国家工程院外籍院士、粤港澳大湾区数字经济研究院创院理事长沈向洋在2024 IDEA大会上分享了其对人工智能“三件套”（算力、算法、数据）的最新思考。他表示，在技术大爆发时期开展创新，对技术的深度理解尤为重要。站在商业的视角，新技术快速冲入市场，则意味着技术需要理解需求。技术要在持续不断的反馈和创新中与市场完成匹配。

算力需求持续增长

大模型参数规模会越来越大，训练这样的模型，对算力的需求会呈现出“平方级”的增长。

整个计算行业在过去四五十年的发展中，最重要的一件事情是算力的不断提升。根据英特尔创始人之一戈登·摩尔提出的摩尔定律，当价格不变时，集成电路上可容纳的元器件的数目，每隔约18～24个月便会增加一倍，性能也将提升一倍。或者换句话说，性能每2年翻一倍，价格下降一半。

然而过去这十几年随着人工智能的蓬勃发展，特别是深度学习的发展，对算力的需求显著提高。英伟达提出了以其创始人黄仁勋的名字命名的“黄氏定律”，认为GPU（图形处理器）将推动人工智能性能实现逐年翻倍。

“从以前的百亿级，到现在的千亿级、万亿级，大模型参数规模会越来越大，对训练的要求越来越高。要训练这样的模型，数据量要增长，性能要随之提升，对算力的需求也会呈现出‘平方级’的增长。”沈向洋感慨道。

同时，沈向洋评论称，英伟达是过去十几年中，信息技术和人工智能行业最了不起、最成功的一家公司之一，它硬生生把自己从一家做硬件的乙方公司做成了甲方。“大家对英伟达的未来充满信心，其中最重要的是看到了行业对算力的需求。这也解释了为什么过去十年英伟达的市值涨了300倍。”沈向洋说道。

人工智能发生范式转移

接下来这几年，算法沿着增强学习这条道路走下去，一定会有令人惊艳的全新突破。

沈向洋表示，从2017年Transformer架构（一种基于注意力机制的深度学习模型架构）出来开始，人工智能、深度学习、大模型基本上是沿着该架构这条线“堆”数据、“堆”算力。OpenAI推出GPT4之后，一些新的突破性技术陆续推出，包括多模态GPT-4o。最新的o1（OpenAI最新发布的大语言模型）推理学习能力展现出的人工智能的范式转移非常值得我们认真思考。

GPT系列做的事情是通过预训练来预测“下一个Token（吞吐量）”。技术背景是把所有的数据聪明地压缩，并能很快给出答案，只要问一句话就能得到一个结果。而新的范式变革是增强学习，即可以自行改善的学习，在给出答案之前还有一个后训练、后推理的过程。

实际上，增强学习并不是一个新鲜事物。AlphaGo下围棋就是用这种增强学习的方法打败人类世界冠军的。不过新的增强学习“打法”更为通用。以前做一个系统只能解决一个问题，比如下围棋或者做其他工作。今天o1不仅可以做数据、做编程，还可以做物理、做化学等。

“我觉得接下来这几年，算法沿着增强学习这条道路走下去，一定会有令人惊艳的全新突破。”沈向洋表示。

未来大模型训练需要合成数据

未来大模型的训练需要强逻辑性的数据，很多数据是网上没有的，需要进一步优化、合成。

公开数据显示，GPT3的训练用了2T（即2万亿Token）的数据。GPT4大概用了20T（即20万亿Token）的数据，相当于今天能找到的几乎所有清洗过的互联网数据。而GPT5预计要比GPT4有长足的进步，预测数据量大概会达到200T（即200万亿Token）的规模。

为了进一步阐释大模型训练所需的数据量规模，沈向洋列举了几个例子：1万亿数据相当于500万本书，或者20万张高清照片，抑或是500万篇论文。一个人从小学、中学、大学到念完大学的时候，真正学到的东西相当于1000本，也仅仅是0.00018T的数据。人类历史上所有的书加起来大概也只有21亿Token的数据。

“现在互联网上已经找不到那么多高质量的数据了，人工智能向前发展要造数据、合成数据，这有可能带来大模型创业的下一个‘百亿美金’问题，就是怎么来合成数据。”沈向洋表示。

GPT系列模型的训练依靠的是互联网语料数据，比如文本、图片、音频、视频等多模态数据，o1的训练则需要强逻辑性的数据，很多数据是网上没有的，需要进一步优化的合成数据。

沈向洋表示，我们不能盲目造数据，而是需要有理有据、有逻辑关系的数据。要先采集真实数据，建一个语境图谱，然后再合成数据，把这些合成数据放进大模型里继续做预训练和推理。

沈向洋透露，IDEA Data-Maker（数据合成平台）知识驱动大模型数据合成技术已经可以将模型推理准确率提升25.4%以上，平均节约成本达85.7%。同时，IDEA大模型合成数据加密训练技术可以打破数据孤岛，助力私域数据的安全流通。相较于基础模型，该技术可将大模型专业推理能力提升12.8%~24.1%。

中国“天河”新一代超算夺得世界图计算领域桂冠

本报讯 记者**张琪玮**报道：近日，记者从国家超级计算天津中心获悉，中国“天河”新一代超级计算机系统，在最新公布的国际Graph500排名中，以6320.24 MTEPS/W的性能夺得Big Data Green Graph500（大数据图计算能效）榜单世界第一。

据了解，Graph500排行榜于2010年首次发布，是国际上评价超级计算机图计算性能的最权威榜单。该榜单主要针对当前热门的数据密集型应用，如人工智能、大数据处理等实施评测，可充分体现超级计算机的访存和通信性能，直接反映超级计算机的数据处理能力。其中，由“天河”新一代超级计算机系统摘得头名的Big Data Green Graph500主要被用于评估超算在图计算中的能耗水平。

当前，大规模数据分析需求日益增长，图计算正成为大数据和人工智能的重要支柱。记者了解到，图计算是一种以图结构为核心的数据处理与分析方法，是研究复杂网络、关联模式和结构化数据的重要工具。

国家超算天津中心党组书记、首席科学家孟祥飞表示，此次中国“天河”新一代超级计算机系统得以摘得世界桂冠，不仅标志着“天河”超算处理复杂数据分析任务的能力取得了国际性领先突破，还为推动新一代智能化技术发展提供了重要支撑。

记者从国家超级计算天津中心官方网站了解到，当前“天河”新一代超级计算机系统已经为天文数据、材料物理、基础物理等学科的发展提供了有效技术助力。

2024年第三季度全球云计算支出同比增长21%

本报讯 近日，全球知名研究机构Canalys发布报告显示，2024年第三季度，全球云基础设施服务支出同比增长21%，达到820亿美元。客户对顶级云厂商人工智能产品的投资成为增长的主要推动力，这也促使主要云厂商加大了在人工智能领域的投入。

全球前三大云厂商——AWS、微软Azure和谷歌云的排名与上季度保持不变，三者合计占全球云支出的64%。这三家厂商的总支出同比增长26%，均实现了环比增长。市场领导者AWS的年增长率为19%，与上季度持平，但增速低于微软（33%）和谷歌云（36%）。然而，从实际金额来看，AWS的收入增长（同比增加约44亿美元）超过了微软和谷歌云。

2024年第三季度，云服务市场延续了强劲的增长态势。三大云计算巨头都表示，从人工智能投资中获得了积极回报，人工智能的应用开始对其整体云业务表现产生影响。这种回报反映了人工智能作为云计算创新和竞争优势的关键驱动因素的日益重要性。

人工智能技术的广泛应用，对高性能计算和存储的需求持续上升，这对云厂商的基础设施扩张提出了更高要求。为应对这一挑战，领先的云厂商正优先大规模投资下一代人工智能基础设施。为了规避投资不足的风险——如无法满足未来需求或错失关键机会。这些厂商采取了超额投资的策略，确保其服务能力能够满足人工智能客户不断增长的需求。因此，这些厂商普遍表示，其资本支出将继续支持快速增长的势头，并预计这一趋势将延续至2025年。

（云 编）

5G打造沉浸式“宋韵文化”新体验

（上接第1版）此外，他们还依赖5G网络的低时延和云上渲染的高效率，打造了5G+AR/VR沉浸式宋韵文化新体验。目前，景区日均接待游客突破10万人次。

高品质的游客体验之下，是坚实的基础设施保障与高效的运营管理在支撑。据了解，中国移动通信集团浙江有限公司杭州分公司通过5G室内定位的辅助支撑，克服古迹内部难布线的问题，精准监测园区巡检覆盖率，有的放矢进行调度优化；此外，借助5G网络的泛链接优势，为南宋皇城小镇打造5G+全域景区智慧化管理平台，大幅提高了运营管理水平。在该平台支持下，遇到人流饱和状况时，单兵监控视频回传延迟能控制在毫秒级水平，并利用5G双域专网的特性保障了数据安全。“在5G网络和智慧管理平台的加持下，景区真正实现了降本增效，运营成本下降了200万元。”中国移动通信集团浙江有限公司杭州分公司相关负责人告诉记者。

南宋皇城小镇实现智慧化转型，正是智慧文旅发展大势的最佳例证之一。中国文化产业协会副会长、北京京和文旅发展研究院院长范周表示：“随着新时代文旅产业升级，‘Z世代’已经成为文旅消费主力，文旅产业的发展正从资源要素驱动向数据要素驱动转变。对此，要更加深度地应用智能技术，挖掘消费需求，进行精准营销，赋能文旅场景，培育壮大新业态，激发消费活力。”

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号