大数据 大算力 大模型

——2023 数博会人工智能大模型高端对话嘉宾观点荟萃

5月25日，2023中国国际大数据产业博览会人工智能大模型高端对话在贵阳国际生态会议中心举行。活动以“大数据、大算力、大模型”为主题，汇聚了中外院士及顶尖专家、领军企业家代表，共同探讨“数据、算力、模型”的技术发展趋势，以及数字经济产业发展趋势。

在主旨演讲环节，中国工程院院士、清华大学计算机系教授郑纬民，美国国家工程院院士、东方理工高等研究院常务副院长兼教务长张东晓，清华大学人工智能研究院常务副院长孙茂松，上海交通大学人工智能研究院常务副院长杨小康结合大模型时代背景下新一代信息技术的新趋势与新发展，从算力、模型等角度阐述了自己的观点及建议。

在圆桌对话环节，分别以“数实相融，创新智算”和“大模型时代的数据智能产业机遇与挑战”为题举行了两场圆桌对话，参与嘉宾深入探讨了大模型时代算力发展、高质量数据、数据激活与价值释放、数据治理等领域的诸多关键议题。

本次活动由中国国际大数据产业博览会组委会主办，上海张江（集团）有限公司、贵州贵安发展集团有限公司承办。

2023中国国际大数据产业博览会人工智能大模型高端对话现场。 贵阳日报融媒体记者 石照昌 摄

中国工程院院士、清华大学计算机系教授郑纬民：

软硬件协同优化助力大模型应用

“ChatGPT出来后，技术层次有了三方面进展，即数据清洗、人工标注反馈和整体系统工程化。”5月25日，在人工智能大模型高端对话上进行《整体系统工程化——在大模型系统中的应用》主旨演讲时，中国工程院院士、清华大学计算机系教授郑纬民说。

在郑纬民看来，整体系统工程化实际上是软硬件协同的系统设计与优化。近年来，新型硬件层出不穷，做人工智能大模型训练的机器类型也有很多。新型硬件的使用，对软件系统的设计提出了更大的挑战。因此，大模型必须和硬件匹配。新的应用软件出来了，也需要对新型硬件系统进行设计和优化。

郑纬民认为，当前整体系统工程化的主要挑战主要是两方面。一是硬件层面，新型异构高性能计算机的体系结构在计算、网络、存储等方面存在硬件限制；二是软件层面，不规则应用程序导致节点间负载不均衡、并行扩展难。

演讲中，郑纬民以大模型“八卦炉”为例进行了讲解，认为对新一代神威超级计算机来说，大规模算力给了扩展预训练模型的绝佳机会。

郑纬民介绍了三种典型的大模型并行训练方式，即数据并行、模型并行和专家并行，并认为分布式训练可通过不同并行模式扩展模型规模与吞吐量。

随后，郑纬民介绍了三种优化模式——

拓扑感知的混合并行模式方面，他认为混合数据与专家的并行模式，相比其他策略性能提升可达1.6倍。

体系结构感知的访存性能优化方面，针对环网可能带来的性能问题，利用核间通信辅助，使用对角线上的核心合并各行的访存请求，可有效降低环网上的请求数量。针对存控可能带来的性能问题，通过排布核组访存模式，可避免同时间负载不均问题。两种优化方法可以带来5.3倍性能提升。

大规模检查点存储性能优化方面，他认为硬件可靠性与检查点的优化重点是存储效率。在新一代神威平台，存储性能同样受网络拓扑影响。而最大化存储带宽，需要满足足够的进程数和进程在超节点间均匀分布两个条件。

“我们从网络拓扑与并行优化、体系结构与访存策略、存储结构与检查点策略三个角度，通过软硬件结合的方式进行软件优化，最终我们将训练系统扩展到新一代神威超级计算机的全机规模，可以支持最高174万亿参数量的模型训练任务，是人类首次达到人脑神经元突触规模（百万亿）的预训练模型。”郑纬民说。

美国国家工程院院士、东方理工高等研究院常务副院长兼教务长张东晓：

知识嵌入和知识发现在科学机器学习中同等重要

在题为《科学机器学习中的知识嵌入与知识发现》的主旨演讲中，美国国家工程院院士、东方理工高等研究院常务副院长兼教务长张东晓从数据驱动模型、理论指导的数据驱动模型（知识嵌入）、数据驱动的模型挖掘（知识发现）三个维度，进行了分享。

张东晓介绍，机器学习模型预测原理是利用大量历史数据，寻找并确定输入多元变量与目标变量的复杂映射关系，构造模型并基于该模型对未来的目标变量进行预测。常用机器学习模型有人工神经网络、支持向量机、卷积神经网络、循环神经网络等。

数据驱动存在一定局限性，如现实场景数据极度稀缺、MSE等指标的局限性、易被攻击与误导（没有常识，缺少知识）、知识工程难以处理复杂现实问题等。

知识嵌入是将领域知识整合到数据驱动模型中的过程，目的是创建具有物理常识的模型，构建物理上合理、数学上准确、计算上稳定高效的智慧能源模型。其核心问题包括复杂形式控制方程的嵌入方法，控制方程以外的通用知识的嵌入方法，不规则物理场的知识嵌入方法，损失函数中正则项权重的自动调整策略。

数据驱动的模型挖掘（知识发现）是从数据到机器学习再到模型的过程。知识发现的目标是从数据中提取未发现的知识，并拓展人类认知的边界。发挥机器学习算法描述高维非线性映射的优势，从实验数据中直接挖掘新的知识。

通过研究，张东晓认为，机器学习算法可以有效解决具有复杂非线性映射关系的问题，数据是基础，通过信息化、物联网，实现从数据到大数据的转变。同时，通过引入行业知识，可以有效提升机器学习模型的效果，即可以在数据预处理、机器学习模型结构以及模型效果评估环节嵌入领域知识，提升精度和鲁棒性，在一定程度上降低数据需求。总而言之，知识的嵌入和知识的发现，在科学机器学习、人工智能技术发展当中同样重要。

清华大学人工智能研究院常务副院长孙茂松：

大模型可以重塑一个产业

5月25日，清华大学人工智能研究院常务副院长孙茂松以《ChatGPT和大模型：开启人类通用的人工智能之旅》为题，在人工智能大模型高端对话作了主旨演讲。

他说，ChatGPT给他的个人感受就像是通用人工智能之“幽灵”。他举例说，把杜牧的《阿房宫赋》提供给ChatGPT，请其找出描写阿房宫的句子，结果“六王毕，四海一，蜀山兀，阿房出”“覆压三百余里，隔离天日”等相关的句子全部被选了出来，一句都没有落下。接着要求ChatGPT根据这些挑出的句子分场景描绘出阿房宫的图像，然后其分五个场景生成了结果。“这说明什么呢？最直接的体会是机器好像有了举一反三的能力，但这件事后面有大模型，叫语言大模型。”

孙茂松梳理了人工智能半个多世纪的发展历程。他认为人工智能能做的事情中，最重要的是语言的处理。

“毕达哥拉斯有一句名言叫万物皆数，我套用他的话，万物皆向量。”孙茂松说，词向量可以得到句子向量、篇章向量，把词、句子、篇章所有的单位打通，这件事是人类历史从来没有做到过的事，ChatGPT完成了。它取得的效果不是平白无故的，是有深度原因的，那就是大模型。

孙茂松说，恩格斯在谈到事物普遍联系的“辨证图景”时指出：“当我们深思熟虑地考察自然界或人类历史或我们自己的精神活动的时候，首先呈现在我们眼前的，是一幅由种种联系和相互作用无穷无尽地交织起来的画面。”而大模型在语义空间里面做到了这一点。

“大模型对经济社会的作用是非常巨大的。它可以重塑一个产业，也可以重塑产业生态。”孙茂松说。

上海交通大学人工智能研究院常务副院长杨小康：

生成式人工智能是新型生产力新型创造力

5月25日，上海交通大学人工智能研究院常务副院长杨小康在《生成式人工智能》主旨演讲中表示，人工智能发展60多年后，有了ChatGPT，某种程度上已经是通用人工智能（生成式人工智能）了。这种通用语言大模型跟人类的大脑可以高度类比，是新型的生产力。

杨小康认为，通用语言大模型也是新型创造力，例如，对很多顶尖设计师来说都很难设计出的一张图，利用通用语言大模型结合图像处理，基本上一秒钟就可以设计出来。在2021年ChatGPT流行之前，就有研究机构预测，到2025年生成式人工智能所产生的数据将占据人类全部数据的10%。当生成式人工智能所产生的数据超过80%的时候，人类是不是会进入元宇宙？

“在元宇宙当中，我觉得两个东西非常重要，一个是人的虚拟化，一个是物的虚拟化，生成式人工智能可以比较好地构建虚拟数字人和虚拟世界。”杨小康说，生成式人工智能在做开创时代的事，是当代的艺术。

杨小康说，生成式人工智能从视觉角度来讲有以下趋势：大模型要多，模态要更通用，物理世界的模拟要更逼真，数字人要更丰富更立体，虚拟人跟虚拟世界可以交互。世界模型是未来的方向，这是Yann LeCun提出AI新架构，让AI像人类（直觉+自监督）一样，对物理世界进行学习和推理。有了大模型之后，这个是可能的。

杨小康认为，生成式人工智能是新型的生产力也是新型的创造力，当然，生成式人工智能不可解释不可控，未来要发展基础理论，再把生成式人工智能用好，加速元宇宙的构建，促进行业的数字化虚实结合。

“数实相融，创新智算”圆桌对话举行

企业家代表共话大模型时代的算力之路

本报讯 5月25日，2023数博会人工智能大模型高端对话活动中，举行了“数实相融，创新智算”圆桌对话。

对话由上海市人工智能行业协会秘书长、上海市人工智能标准化技术委员会秘书长钟俊浩主持，贵安新区科创产业发展公司常务副总经理邓周灰，优刻得科技股份有限公司副总裁刘杰，IBM大中华区混合云及人工智能专家实验室总经理魏永明，沐曦联合创始人、CTO兼首席软件架构师杨建四位企业家代表，共同探讨大模型时代算力发展的关键议题。

大模型出来后，算力产业面临哪些机遇和挑战？

“我觉得现在应该有一个‘铁三角’，即算力、模型和数据。”邓周灰说，大模型训练的时候，大部分来自互联网。现在算力多样化，大家都在做大模型，需要不断重复投入，这会不会造成大量投入浪费，是需要讨论的问题。

刘杰认为，在算力方面，目前高端算力还达不到要求，而国产化的算力还需要时间。在数据方面，在做大模型时发现存在数据总量不够、质量不高等问题。在算法方面，大模型要基于上千亿参数进行计算，计算的模型很多，调试很困难。

在商业和技术深度结合方面，结合“东数西算”“东数西训”，魏永明认为，目前大家更多关注算力层面，就是基础设施层面。但分布式结构最重要的是软件技术。在训练模型方面，要把计算需求通过软件虚拟化，从而实现从东部、西部、南部等任意大空间的跨度计算。“我们的研发不能光在基础架构上，在应用层也需要关注，这对于国内人工智能大数据或者数字化应用，是非常重要的。”

针对超算中心在未来低成本服务于大模型时代，杨建说，全国已有11家超算中心，应用于科学计算、AI等方面。在科学计算上，国内和国外所有AI的序列框架完全可以在生态上运行，比如大数据处理、GPU数据库等。

嘉宾表达了对算力发展的期待。

“都说算力是数字经济的生产力，数据是数字经济的生产要素，大模型的出现让我看到了数据作为生产要素推动数字经济发展的可能。”刘杰说。

“算力最终是通过人工智能在产业或者社会上产生一系列价值，可以预见这个价值在未来会有颠覆性效果。我们要具备这种应用人工智能、利用数字化的技术，重新设计自己的职业，重新设计自己将来工作的能力。”魏永明说。

“大模型时代的数据智能产业机遇与挑战”圆桌对话举行

嘉宾探讨以大模型推动生产力发展

本报讯 5月25日，2023数博会人工智能大模型高端对话活动中，举行了“大模型时代的数据智能产业机遇与挑战”圆桌对话。

达而观信息科技（上海）有限公司首席战略官刘江贤主持了该场圆桌对话，香港理工大学先进制造研究院副院长、讲座教授黄国全，北京柏睿数据技术股份有限公司董事长兼首席科学家刘睿民，中国人工智能开源软件发展联盟副理事长王健宗，晶泰智药技术（上海）有限公司副总裁王明泰四位嘉宾共同探讨了高质量数据的定义与标准、数据激活与价值释放、数据治理等领域的关键议题。

“大模型的三大促进元素中，除了算力和算法，就是数据。”刘江贤说，大模型需要大数据，而且数据的质量是关键。但如同淘金一样，需要有针对性地选择“矿藏”挖掘数据，用科学的角度和方法推动大模型发展。

用好大模型要有高质量的大数据支撑，要采集好的大数据就要有大规模精细化的布局。黄国全表示，在智能制造方面，要用好大数据的万有价值定律、数据的共享定律、采来的数据要形成碰撞这三大定律，解决智能制造大数据面临的“成本高、风险高、技术门槛高”三高问题，推动生产力发展。

王明泰认为，国内外的大型制药企业都在关注大模型，但不是关注大模型本身，而是关注大模型背后开发的范式、计算的范式。谁掌握新的高质量数据和新的模型，就有可能集中产出大量新药。不过，强调数据的高质量，不能忘记安全性，数据应该被负责任地存储，然后负责任地使用。

“所有的模型到最后都是向量或者是张量的计算，离不开矩阵计算。”刘睿民说，大模型将带来更多挑战和机遇。工业领域的规上企业实现自动化后要进入数字化，提升过程其实就是把所有机器运行的状况全部进行数字化，反映在一个虚拟的世界里，某种意义上可以认为这就是一个模型。

王健宗说，数据在大模型时代有三个明显的变化，一是数据的获取难度变大，二是对于数据多维标签的要求变高，三是万物皆向量。对于整个大模型的生产来说，最核心的就是数据和算力。现在，中国电信、中国联通、中国移动、华为、苹果、腾讯等纷纷将数据中心落地贵安，上百家大数据企业在此聚集，未来，贵安可能成为一个大数据的工厂、大模型的工厂，专注于大模型的企业可以来贵安安家。

贵阳日报融媒体记者 谢孟航 张鹏 / 文