秀湖会议集萃

科学智算 (AI4S) : 交叉与赋能

——第六期CCF秀湖会议报告

关键词 : 科学智算 AI4S 基础理论 大语言模型

编者按 : 2023年10月20~22日, 第六期CCF秀湖会议在苏州CCF业务总部 & 学术交流中心举办, 就“科学智算 (AI for Science, AI4S) : 交叉与赋能”进行深入交流和研讨。来自学术界与工业界的20多位专家全程参与, 围绕AI4S基础理论与模型算法、AI4S在基础科学和工程技术领域突破、大语言模型与AI4S结合、AI4S计算支撑平台等方面进行探讨, 并形成如下报告。

背景与意义

科学研究的两大基本范式分别是以数据观测为核心的实验科学和以发现物理世界基本原理为核心的理论科学。但是, 对于成本高昂的科学实验 (如可控核聚变的核心是让“托卡马克”环中等离子体进行可控核聚变等) 或者以假设为前提的科学研究 (如数学猜想证明和未知物质合成等), 传统科学研究范式往往难以满足个性化或复杂繁多的要求。

人工智能与基础科学和工程技术等领域研究相互结合, 推动科学智算的蓬勃兴起, 帮助科学家生成假设、设计实验、计算结果、解释机理, 特别是辅助科学家在不同的假设条件下进行大量重复的验证和试错, 让人工智能利用自身强大的数据归纳和分析能力学习科学规律和原理, 从而得到科学模型, 解决实际的科研问题, 大大加速了科研探索的进程, 并获得仅使用传统科学方法可能无法获得的见解。

本次秀湖会议的主题是“科学智算 : 交叉与赋

能”, 一方面探讨人工智能这一通用目的使能技术所具有的学科交叉内禀促进基础科学研究, 另一方面剖析人工智能在工程技术领域的赋能途径, 推动基础科学和工程技术领域学科范式变革。

观点集萃

人工智能赋能生命科学

构建高质量生命科学智算数据集。“数据是燃料、模型是引擎、算力是加速器”, 数据、模型和算力是这一轮人工智能受到关注的三驾马车。深度学习迅速崛起一定程度上是因其其在ImageNet这类大规模标注图像数据集上取得了令人耳目一新的性能, 但是在生命科学领域尚无高质量科学智算数据集, 一定程度上限制了人工智能在生命科学领域的发展。尽管生物领域正在布局高通量的生物自动化大设施平台, 以促进高质量标准化数据的获取, 但是生命科学领域数据产出的成本高, 部分数据具有敏感性或较高经济价值, 使大规模数据的构建和共

享变得困难。

生物学问题与可计算问题。当前生命科学研究往往基于少量数据，其一直沿用的思维方式与当前人工智能技术依赖于海量数据不同。这导致当前大量生物学问题尚未被抽象为数据驱动模式下的可计算问题，难以利用人工智能开展辅助分析，当前生命科学领域的人工智能研究集中在少量环节，没有与特定领域深入结合。

人工智能与科学的双向奔赴。由于知识体系、思维习惯等方面的差异，人工智能与科学领域的科研工作者面临合作壁垒，限制了人工智能在生命科学领域的赋能。人工智能和科学双方都应进一步努力推进在 AI4S 方向的合作。双方合作时，人工智能方可以提供简明人工智能教程或培训，使科学方充分了解当前人工智能的优势与不足。同时科学方应提供一定的领域知识，如物理化学特征、形态学特征和逻辑知识等，以辅助人工智能方明晰特定领域知识，使模型尽量学习到更本质的特征。

总之，AI4S 归根到底是要解决科学问题，因此需要更加深入地认识问题、理解问题。只有对问题理解得够深入，才能做到“解决真问题，真解决问题”。为此，需要鼓励人工智能专家更加勇敢地走出自己的领域，拥抱更广的科学领域问题，并扎根进去；同时，鼓励科学领域专家多多总结和抽象问题，并且对新思路和新方法抱有更加开放的态度，大胆尝试。

人工智能赋能数学与量子计算

明确人工智能适用边界。对于实际具体场景的图论与组合等 NP 难问题的求解，机器学习将带来新的技术路线，降低对专家的依赖，以及更好地利用 AI 算力和处理实际不确定性。但是在很多实际问题中，人工智能模型能获得和处理的信息是有限的，如何甄别哪些问题可以利用 AI 技术进行有效求解值得更深入的探索。

加强数理基础科学研究。从数学或物理的角度有可能揭示 AI 模型中的本质问题。充分思考和结合科学中的第一性原理，从数学优化手段和物理定

律约束等角度出发，寻找人工智能模型提取通用特征的最佳途径，使所提取的区别性特征能够反映所求解对象的第一性原理，这样才能有助于提升人工智能模型的专业科学性。

加强内嵌机理驱动下人工智能科学范式研究。目前，国际上形成了将物理机理嵌入深度神经网络的研究趋势，例如布朗大学提出的内嵌物理规律的神经网络 (Physics-Informed Neural Networks, PINNs)、麻省理工学院发表的通用微分方程 (Universal Differential Equations, UDE) 范式、加州理工学院发表的傅里叶神经算子 (Fourier Neural Operator, FNO) 等，这些研究已成为解决部分复杂科学问题的有效思路，但如何利用人工智能统一近似理论和发挥数据拟合能力，让人工智能既有效利用数据又从容引入知识，仍有待进一步探索。

加强计算平台支撑。AI 的成功离不开大规模算力的支撑，在目前 GPU 算力受限的环境下应当加强计算平台的支撑。除了经典的 GPU 算力外，量子计算也为科学智算提供了新的可能。不过目前量子计算还处于发展阶段，只被少数机构拥有，一定程度上限制了 AI 和量子结合的研究。因此，一方面需要开发更高效的量子计算经典模拟器，以支撑用传统的计算设备开展小规模量子算法的研究；另一方面，发展量子计算云平台，以支撑远程调用研究。

大模型赋能工程应用

大模型学习方法论。实现机器学习的自动化是通向自主智能的关键，大模型本质是学习方法论。任何学习任务是对现实世界规律的总结和刻画，通过观察数据学习得到逼近函数。这一学习过程由数据方案、机器架构、损失函数和优化方法构成，本质要求是对数据、网络、损失、算法、任务等要素做一体化的设计和调控。因此，实现这一目标就是在任务和方法之间建立映射，即掌握学习方法论。

加强大模型可解释性探索。大模型已在许多应用领域展现出强大的性能，然而它们通常被认为是

黑盒模型,难以理解和解释。在一些关键应用领域,如医疗问答、药物发现等场景,模型的可解释性对可信性和可靠性至关重要。

提升大模型在垂直领域的泛化性。大语言模型有潜力推动人工智能科学计算,当前大语言模型在科学领域的泛化性仍需进一步提升。需要关注利用大语言模型进行多模态、多源、多尺度的知识融合,实现对复杂系统的全面表征;构建涵盖不同领域和层次的专业化语料库,提高大语言模型在科学领域的适应能力和泛化能力;引入下游行业多任务微调,提升模型在医疗行业、地球科学、可对话交互的生物计算、新型电力系统场景的落地效果。

提升大模型结构化知识能力。尽管大语言模型的文本理解、内容生成创作有能力有巨大提升,但在数学逻辑推理、发现和定义新问题等任务上仍有不足。目前训练的语料数据主要围绕互联网网页信息、电子书、维基百科等非结构化语料知识,需要重新定义和思考训练语料来提升模型在理科推理、发现和定义新问题上的能力。从搜集各个领域的语料库到构建结构化-知识性质的数据,让大模型“吃得好”,科学思维能力也得到提升。

计算平台

加强多领域算子库设计与加速。科学智算离不开能够支撑不同领域的统一算子库,然而目前此类通用算子库依然非常有限。因此首先应依据共性算子规范,研制复杂算子,支持国产芯片适配。其次,根据跨领域共性基础模型,研制模块化算子库。最后,覆盖科学领域标准应用套件,支持管理工具、优化策略等服务,提供数据、模型、调试等多种开发接口。

构建高性能计算平台。大模型时代对算力有极高的需求,高性能计算平台也成为一种新型的基础设施。该平台应该具备高吞吐量、高通信速率、高并行浮点计算、高存储容量的特点,以支撑在人工智能算法、大数据算法框架、开源科学计算库、商用计算软件的软硬件集成等方面的需求。

加强国产框架开发。国外已建立成熟的人工智

能框架 PyTorch 和 TensorFlow,亦有 JAX 等新技术代表进行了人工智能求解科学计算的相关优化与适配。然而,现有框架不能完全满足科学智算需求:动态编程的 PyTorch 运行效率较低,且难以满足工业级部署需求;TensorFlow 缺少基础算子体系,科学计算的高阶自动微分功能可扩展性不足;JAX 则尚未形成成熟生态。因此应当在共性范式的引导下,基于国产硬件平台,搭建异构适配的科学智算共性框架,以无服务计算(serverless)等跨学科友好的部署方式,形成人工智能科学计算共性平台,以有效服务多学科的科学研究问题求解,支撑各种复杂科学问题研究。

AI4S面临的挑战

AI赋能科学计算的局限性

模型设计缺乏理论指导。目前,科学智算主要依赖深度神经网络等通用的机器学习算法或模型,然而这些算法或模型可能不适用于所有的科学问题。首先,如何将特定学科中待解决的某一个或某一类科学任务抽象为计算机可计算任务是一个重要挑战;其次,如何设计更适合特定领域和问题的人工智能算法和模型,将该科学领域中的先验知识和物理规律转化为 AI 模型约束和指导的过程更是困难重重,需要专家同时具备两方面专业知识。

AI 模型缺乏可解释性。科学的目标是理解物理世界的规律,很多时候就是发现和解释科学现象,将“黑盒”变成“白盒”。然而目前的研究尚未完全理解模型的内部运作方式,无法系统地解释模型的决策过程、判断标准和推理结果,给 AI 技术在科学领域的深入应用带来了很大的挑战。

AI 模型泛化能力有限。训练数据和测试数据分布的偏差,会带来模型的泛化问题,尤其在真实应用落地中,往往会降低 AI 模型的有效性。而在科学智算场景下,模型的错误输出造成的后果可能是致命的。这要求 AI 模型必须对输入的微小变化

和噪声，以及数据分布偏移具有强大的鲁棒性，才能使科学智算的结果可信可靠。

数据安全与隐私问题。这个问题广泛地伴随着现有几乎所有的人工智能方法，在科学智算领域可能更为凸显，尤其涉及个人敏感信息或关键基础设施这两类场景。虽然联邦学习等致力于解决数据安全隐私隐患的人工智能方法受到越来越多的关注，有了一定的发展，但是“道高一尺、魔高一丈”，还是存在推理攻击等手段威胁数据隐私的安全。因此我们要不断创新 AI 模型，以确保科学智算场景下个人敏感信息不被泄露以及关键基础设施的安全。

科学智算的基础理论和方法体系

科学领域具体问题的抽象。科学智算的初期发展面临一个重要挑战：许多科学问题尚未被抽象为可计算问题。以数字孪生人类任务为例，如何整合和映射多源异构数据，解决人类认知阻碍建模的问题，完成从微观到宏观的多尺度综合建模，以实现人体机能的全面理解，是一个重要的挑战。科学领域具体问题的抽象，要求超越传统的学科边界，培养将实际问题转化为可计算问题的能力，以跨学科的方式处理复杂的数据和模型，以便利用计算机的处理能力加速科学研究的进程。这涉及跨学科的合作和新的方法论，是一个严峻挑战。

AI 模型和科学领域问题的融合。AI 模型与科学领域问题的融合一直是科学智算领域亟须解决的问题。目前，该领域仍处于早期发展阶段，普遍存在缺乏共性范式引导、科学计算编程框架不统一、领域套件难复用等问题，导致各个学科之间难以形成合力，进一步阻碍了规模放大效应的产生。解决这一问题，需要建立一套适用于各种科学问题的通用框架和原则。从数理逻辑、计算复杂性、信息论等角度出发，分析科学智算的本质和特征，揭示人工智能科学计算与科学之间的内在联系和规律是至关重要的。

人才培育和评价体制。科学智算的基础理论和方法体系的发展还面临着人才培育和评价体制方面

的重大挑战。首先，科学智算的交叉学科性质需要培养具备深厚科学素养和人工智能技能的综合型人才。而在国内现有的教育体系中，学科之间是独立的，交叉学科不够深入，也较难吸引到特别优秀的人才，亟须改革以提供更优质的资源。其次，科学智算领域需要综合考量科学贡献和人工智能技术创新，然而传统的学术评价标准主要注重科研论文和学术成果，往往无法全面评估人工智能在科学研究中的实际应用价值。

科学智算的数据处理与计算平台

高质量标注数据获取困难。AI 模型的训练离不开高质量的数据。不同于当前计算机视觉和自然语言处理等领域中的大量高质量数据，许多科学领域中的数据标注高度依赖于实验测量，这使高质量的标注数据采集变得非常困难，同时由于科学数据具有高维、稀疏和有噪声等特点，基于这些数据训练的模型难以有效地帮助相关学科的研究。虽然一些科学领域已经开始布局自动化大设施平台来获取高质量的标注数据，但是如何提升数据标注的准确性和效率是科学智算面临的重大挑战。

跨域数据理解与融合困难。科学研究的跨学科特性要求 AI 模型能够理解并融合不同领域的数据和知识。然而对科学文献中的专业术语的理解依赖大量的专业数据和知识，导致许多科学知识无法形式化表示，进而影响科学智算的效果。同时，许多科学领域的知识以图表、知识图谱等形式存在，如何充分利用这些多模态的数据并与人类知识融合成为科学智算面临的新挑战。

科学智算的统一计算框架。人工智能科学计算需要大量的计算资源和数据资源，因此需要高效、稳定、安全的科学智算平台，实现数据的标准化、共享和保护，提供易用、灵活、可扩展的人工智能科学计算工具和服务等。然而目前已有的科学智算框架存在运行效率低、缺少基础算子体系、可扩展性不足等问题。如何设计高吞吐量、高通信速率、高并行浮点计算的统一计算框架，是科学智算面临的基础挑战。

科学智算的算力平台。科学智算的发展非常依赖高性能的算力平台，然而很多科学领域的专家缺乏构建大规模计算平台的能力，因此很难有机会直接使用类似于人工智能领域的大规模计算资源。虽然量子计算等领域的发展为降低计算资源门槛提供了可能，但是目前仍不具有普适性。同时，大规模算力给生态环境与可持续发展带来了巨大的挑战。如何克服底层技术、成本分摊等问题，建设公共算力服务平台和算力网，是科学智算当前面临的巨大挑战。

AI4S发展建议

科学智算的“范式化”和“平台化”

推动科学智算理论“范式化”。在推动科学智算的“范式化”方面，我们迫切需要发展科学智算的理论方法。随着人工智能引发科学发现模式的变迁，我们应该鼓励学者深入挖掘多学科智算融合模式，共同探索人工智能技术在科学计算中的应用边界与方法论，包括对统一近似理论和数据拟合能力在科学计算中更深层次的探讨，以及人工智能如何更好地为传统科学计算服务流程赋能。构建这一基础理论体系将有助于完善科学计算问题的解决方案，促进科学智算的范式化进程。

推动科学智算软硬件“平台化”。为了实现人工智能科学计算的“平台化”，我们必须努力以共性范式为引导，基于国产硬件平台，建立异构适配的科学智算共性框架。通过采用无服务计算等跨学科友好的部署方式，我们可以形成科学智算共性平台，有效服务于多学科的科学研究问题求解。这样的平台将为各种复杂科学问题的研究提供有力支撑，使不同学科领域的研究者能够更便捷地利用人工智能技术，推动科学智算的平台化取得更为显著的成果。

重视科学智算领域智能体研究。大模型在知识获取、指令理解、泛化、规划和推理方面都表现出了强大的能力。智能体(agent)是能够感知自身

所处环境、自我决策并采取行动以完成特定任务的人工智能模型。因此，研究大语言模型(LLM)支撑下的智能体技术，完成科学智算中不同种类任务，形成端云协同的大小模型计算模式，在云上LLM支撑下完成端侧不同智能体任务，构建科学实验机器人。

构建AI4S交流平台

与会专家共同提出交流平台构想。多位专家学者提出建立科学智算的跨学科交流和合作平台，建立全面促进人工智能与科学深度合作与交流的创新生态系统，包括搭建跨学科的合作平台，加强人才培养和引进，以及投入科学计算基础和应用研究。科学智算交流平台旨在建立一个全面促进人工智能和科学深度合作与交流的创新生态系统，这需要更多的交流将科学问题和人工智能方法链接起来，也需要针对特定重大科学问题开发新的人工智能方法。

交流平台促进科学研究。在研究层面，平台的存在将鼓励领域学者探索前沿理论和实际问题，推动该领域的理论创新和技术突破。针对这一点，可以尝试创立AI4S的学术会议和期刊，更好地促进交流与合作以及研发解决科学问题的新方法。鼓励跨学科、跨组织、跨行业组队申报国家重大科研项目，在安全界限内，打破学科壁垒，实现信息、数据、知识的互通。对跨学科组队各方的贡献，不能单纯以投入的人力和硬性实验成本计算，也要考虑智力投入和产出贡献，客观评价。在政策和制度层面，引导和促进跨学科的融合互通。同时，与会专家建议，想要从事AI4S的研究者应该有决心，放下以前的研究路径，拥抱新的科研体系。人工智能和科学领域的学者都需要更主动一些，努力做好这个方向，互相往对方多走几步，人工智能方向的学者应该更加主动地了解科学方面的科学问题和挑战，科学方向的学者需要更加主动地了解人工智能方法的机制、优势和局限性。

交流平台整合社会资源。在数据、政策、人才培养和资源层面，平台将推动建立数据开放共

享机制，鼓励跨学科组队参与国家科研项目，并在政策和制度上促进跨学科的融合。平台将关注有效的资源共享机制，确保 AI4S 领域的计算资源、数据资源得到充分利用。虽然国家自然科学基金委于 2020 年底正式成立了交叉科学部，但学科保护主义具有天然的惯性，交叉学科的理念和文化还需要更多的鼓励和支持。建议相关政府部门、互联网企业以及垂直行业企业能够加大对 AI4S 领域资助的广度和力度，并对交叉学科的科研成果形成更加合理的评价体系，一起更好地推动 AI4S 的发展。

数据共享和开源

建立科学智算数据开放共享机制。高质量数据是 AI 发挥作用的基础，然而科学智算涉及的数据来自不同行业和机构。当前我国没有统一的科学智算数据管理机制，导致数据留存散落在各个研究机构的实验室中，要么无从考证，要么尘封为历史，要么几经转手丢失严重，无法发挥出数据的更多价值。以生物领域数据为例，可以参考美国布鲁克海文国家实验室的模式，由国家级的高校、机构牵头，负责建设具备一定规模且质量较高的生物领域权威数据，定期维护更新实验数据，组织人员对数据进行标签化整理和清洗，在合法合规范畴内有计划地面向国内科研工作者、教育工作者和企业开放，搭建人工智能科学计算的共享数据和模型库，推动人工智能科学计算的开源和开放。

提升科学数据高质量标注的能力和效率。相比经典的图像和自然语言，科学智算数据的标注更为困难。一方面，科学智算数据标注需要深入理解科学智算的机理，因此需要科学专家和人工智能专家进行深入的探讨与合作，而不是仅由科学领域专家进行手工标注或者由人工智能专家在不理解科学原理的情况下进行标注。另一方面，为提升科学数据的标注效率，需要在考虑数据隐私和数据安全的前提下建立开放共享的标注平台，让更多科学领域的从业人员可以合作共建高质量的科学标注数据。

加强学科建设和人才培养

加强科学智算学科体系建设。学科建设是推动科学智算长期持续发展的必要保障。建议在智能科学与技术一级学科设立的背景下，构建 AI4S 面向不同学科的完善的课程体系，以推动 AI4S 人才的培育。而为了培养具有交叉学科背景的未来科学家，建议加强人工智能与科学的融合，鼓励学生探索领域间的交叉点，培养创新思维和实践能力。

建立完善的人才培育体系。目前我国在人工智能领域已经建立起相对完善的人才培养体系，但是在 AI4S 领域，尚无完善的人才培养机制。建议加大人才培养和引进力度，通过建设专业课程体系和引进国内外专家，培养具有双重背景的复合型人才，为领域的可持续发展提供支持。为具有交叉属性的研究设立专项基金和人才奖励，支持相关人员获得经费的支持，并鼓励更多研究人员参与，持续推进 AI4S 的研究。

整理

吴飞 周晟 王永威 卜佳俊

会议发起人

吴飞 卜佳俊

特邀嘉宾（按姓名拼音排序）

鄂维南 漆远 宋乐 徐宗本 杨焕明
曾震宇 张林峰 张渝

参会嘉宾（按姓名拼音排序）

白磊 陈阳 董伟 洪亮 黄刚
黄合良 李斐然 李建欣 李子青 刘祎
林洲汉 刘扶芮 石荣晔 王太峰 王永威
王钰 汪志华 忻琦 薛贵荣 严骏驰
叶杰平 叶凯 张铂 张贵军 张肖男
周号益 周晟 朱霖潮

（本文责任编辑：杨易）