微软亚洲研究院院刊

Matrix

NO.68
2024年 1- 3月

01 焦点

02 前沿求索

科研第一线

03 文化故事

04 观点

05 媒体报道

大模型时代的计算机系统革新：更大规模、更分布式、更智能化

“大模型的不断涌现和下一代人工智能需求的迅速增长，促使我们加速对传统计算机系统的革新。同时，构建于大规模高性能计算机系统之上的现代人工智能技术也为未来计算机系统的研究带来了无限的机遇。创新超级计算机系统、重塑云计算、重构分布式系统，将是实现计算机系统自我革新的三个重要方向。”

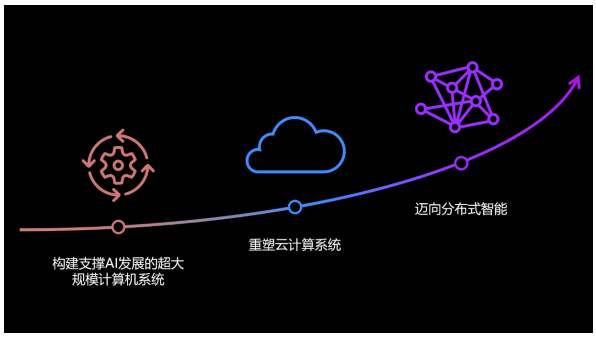
——杨懋，微软亚洲研究院副院长

杨懋，微软亚洲研究院副院长

在计算机科学的诸多细分研究领域之中，计算机系统研究可能是最兼具“古典”与“摩登”特质的研究方向。说它古典，是因为计算机系统的雏形可以追溯到古代的算盘、算筹、数据表等计算工具，其发展远远早于软硬件、云计算、人工智能等技术的研究；至于摩登的一面，大数据、云计算等现代技术又促进了计算机系统的不断进化。传统计算机系统研究领域，如分布式系统理论和实践、编译优化、异构计算等成果，已在当今的大模型时代大放异彩。同时，以大规模 GPU 集群为代表的高性能计算机系统也推动人工智能实现了质的飞跃。

然而，随着人工智能技术更新迭代速度的加快，我们也愈发清晰地看到传统计算机系统面临着新的挑战：当前的 GPU 集群在规模和效率上，已经难以满足新一代人工智能模型的训练和服务的需求，而现有的云计算和移动计算系统平台，也需要从服务传统的计算任务向服务智能应用转变。

面对这一系列挑战，我们意识到构建于大规模高性能计算机系统之上的现代人工智能技术，将为计算机系统的研究带来无限的机遇。因此，计算机系统的革新也势必要从这三个方向展开：创新超大规模计算机系统以支持未来人工智能的发展；重构云计算这一重要的 IT 基础平台；设计前沿的分布式系统，以适应更广泛的分布式智能需求。



大规模和更高效的计算机系统是下一代人工智能发展的基石

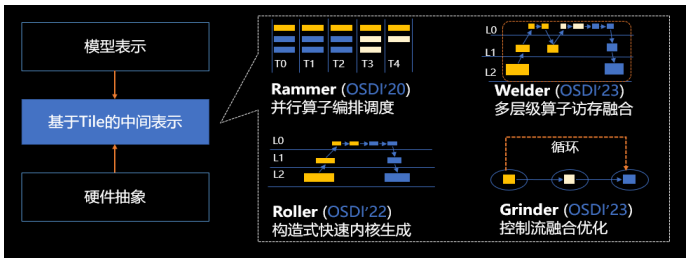
强化学习领域的创始人之一 Rich Sutton 曾说过，“从70年的人工智能研究中可以总结出的最重要的经验是，最大化利用计算能力是最有效，也是最有优势的方法。从长远来看，唯一重要的事情就是利用好算力。”

超级计算机系统作为当前最有效的计算力“源力”，是现代人工智能成功的重要基石。然而，在基于超级计算机系统构建大规模 GPU 集群的过程中，系统的可靠性、通信效率和总体性能优化成为制约大模型训练性能上限的关键问题。因此，我们需要创造一个更高性能、更高效率的基础架构和系统，以推动下一代人工智能的发展。

过去五年中，我们从体系结构、网络通信、编译优化和上层系统软件等多个角度，开展了计算机系统的创新研究，为人工智能基础架构的演化提供了有力支持。例如，我们推出了能够跨多个加速器执行集体通信算法的微软集体通信库 MSCCL[1]，以及有助于开发大规模深度神经网络模型的高性能 MoE（Mixture of Experts，混合专家）库 Tutel[2]。这些研究成果为包括大语言模型训练及推理在内的各种人工智能任务提供了高效的支持。

超级计算机系统不能仅依靠传统系统方法来实现革新，而是要利用人工智能实现创新和演进。这也是微软亚洲研究院正在探索的研究方向，我们认为人工智能的新能力将为解决传统计算机系统问题提供新视角，包括更智能和高效地优化复杂系统的性能，更快速和智能的问题诊断，以及更便捷的部署和管理。人工智能与系统结合将为计算系统设计带来新的范式。从芯片设计、体系结构创新、编译优化到分布式系统设计，人工智能可以成为

系统研究者的智能助手，甚至承担大部分工作。在人工智能的协助下，系统研究者可以将更多精力用于更大规模系统的整体设计，关键模块和接口的抽象，以及系统整体的演进路线。比如，对于人工智能编译系统的设计，我们推出了 Welder、Grinder 等编译器[3]，可以更专注于模型结构、编译系统和底层硬件之间的关系和抽象，而更多具体的编译优化搜索算法和实现可以由人工智能辅助完成。这些新的系统研究范式将成为构建更大规模和更高效的人工智能基础架构的真正基石。



基于统一切块（tile）抽象的四个核心 AI 编译技术

以智能化为内核，重塑云计算系统

“操作系统管理着计算机的资源和进程以及所有的硬件和软件。计算机的操作系统让用户在不需要了解计算机语言的情况下与计算机进行交互。”这是我们对计算机系统的最初理解。

但是，随着以 GPU、HBM、高速互联网络为代表的分离式（Disaggregation）服务器架构逐渐取代传统以 CPU 为中心的服务器，人工智能智能体（AI Agent）和大模型成为云计算平台的主流服务，深度学习算法逐渐替代传统服务核心算法，云计算这个始于本世纪初的最重要的 IT 基础系统也需要重塑自身。

传统云计算领域的研究方向，如虚拟机（VM）、微服务（Micro-services）、计算存储分离、弹性计算等，在人工智能时代下需要被重新定义和发展。虚拟化技术需要在分离式架构的背景下进行重新设计；微服务及其相关云计算模块需要为 AI Agent 和大语言模型构建高效且可靠的服务平台；数据隐私和安全需要成为云计算系统创新的核心要素。所有这些变革创新都要服务于云计算系统的智能化（Cloud + AI）。

在过去几年中，我们在体系结构方面围绕分离式架构展开研究，在系统软件上以大语言模型和 AI Agent 为核心，提出了诸多构想，推出了多项创新技术。这些技术将在未来的云计算平台中发挥重要作用。

云计算自身的变革也为云计算平台上的传统服务，如数据库系统、大数据系统、搜索和广告系统、科学计算等大规模系统，带来了新的进化机遇。一方面，大规模异构计算系统在云端的普及为传统大规模系统提供了新的计算平台；另一方面，深度学习特别是大模型的发展为传统大规模系统的内在算法设计和实现提供了崭新的思路。以搜索系统为例，我们基于异构计算系统和深度学习方法对搜索系统进行了创新，从 Web Scale 的矢量搜索系统 SPANN[4]到最新的 Neural index 索引系统 MEVI[5]的设计，这些创新不仅极大提升了搜索和广告系统的性能，也为未来信息检索系统提供了新的范式。类似的创新也发生在数据库系统、科学计算系统等领域。

云计算系统不仅为人工智能的发展提供了保障，其自身和构建其上的大规模系统服务也将受益于人工智能技术，从而实现持续演进。未来的云计算平台也将成为新一代人工智能基础架构的关键组成部分。

分布式系统将是分布式智能的关键基础设施

“人类的智能不单存在于人类的头脑中，还广泛分布在整个物理世界、社会活动和符号体系中——这就是‘分布式智能’。”美国认知科学家 Roy Pea 在 1993 年发表的一篇论文《Distributed Cognition: Toward a New Foundation for the Study of Learning》中提出了分布式智能（distributed cognition）的概念，为我们提供了一种新的视角来理解人工智能系统与社会以及环境之间的相互作用。

目前，大模型的技术链条，从训练到推理都依赖于云计算中心。但我们相信，智能广泛存在于分布式环境中，未来的智能计算也必然存在于任意的分布式环境中。

人类和物理世界的交互、基于符号系统的交流，都是智力活动的体现。在未来，这些智力活动应该能被大模型更好地感知和学习，人们也可以在任意终端更实时地获取人工智能模型的能力。这种泛在的相互感知和不断演进的能力，将是未来分布式系统研究的重点之一。

那么，如何支持智能技术在更分布式的场景下发展？我们需要考虑在由云端、边缘端和设备组成的广泛计算平台中，如何更好地进行人工智能计算。除了传统的模型稀疏化、压缩等优化模型推理性能的技术外，更为关键的是要克服大模型等算法在边缘端运行时遇到的挑战，如实时性和可靠性等基础问题。为此，我们推出了PIT[6]、MoFQ[7]等多种移动端模型量化、稀疏化以及运行时优化的技术。

另外，对于边缘计算平台和设备，硬件和推理算法的创新也至关重要，这将从根本上革新端侧的推理方式，比如利用基于查找表（Lookup Table）等全新的计算范式来提升端侧推理效率，包括 LUT-NN[8]等技术。

我们还与多个不同的机器学习团队紧密合作，使学习算法可以更好地从任意信号（Signals）中捕捉智能。除了传统的多模态模型，我们也在寻找更简洁和内在一致的模型结构和学习算法，

可以从任意信号中进行学习。我们也在探索更优的模型结构和算法，这些模型应当更稀疏、更高效，且具有良好的可扩展性，能够有效地支持自学习和实时更新。

未来，智能将融入广泛的分布式环境中，而创新的分布式系统将是分布式智能的关键基础设施，也是人类社会获得更实时、更可靠的人工智能交互能力的前提。

未来的计算机系统将自我进化

未来的计算机系统研究将是一个持续自我革新的过程。这不仅意味着计算机系统需要不断进化来满足未来人工智能发展的需求，也意味着计算机系统本身将更加智能化，并具备自我演化的能力。

过去几年的变革创新让我们窥见了些许未来的样貌。然而，从基础架构、云计算平台到分布式智能化，人工智能时代的计算机系统研究领域，还有很多新的可能性等待我们去探索。当然，我坚信那些更加智能、更强大的助手和工具，一定会在未来的研究道路上给我们带来尚未被发现，但又足以令人兴奋的惊喜。

相关链接：

[1] Microsoft Collective Communication Library (MSCCL)
https://github.com/Azure/msccl

[2]Tutel MoE: An Optimized Mixture-of-Experts Implementation
https://github.com/microsoft/tutel

[3] 微软亚洲研究院推出AI编译器界“工业重金属四部曲”
https://www.msra.cn/zh-cn/news/features/ai-compilor

[4] SPANN: Highly-efficient Billion-scale Approximate Nearest Neighbor Search
https://www.microsoft.com/en-us/research/publication/spann-highly-efficient-billion-scale-approximate-nearest-neighbor-search/

[5] Model-enhanced Vector Index
https://www.microsoft.com/en-us/research/publication/model-enhanced-vector-index/

[6] PIT：通过排列不变性优化动态稀疏深度学习模型
https://www.msra.cn/zh-cn/news/features/new-arrival-in-research-3#research3

[7] Integer or Floating Point? New Outlooks for Low-Bit Quantization on Large Language Models
https://arxiv.org/abs/2305.12356

[8] LUT-NN: Empower Efficient Neural Network Inference with Centroid Learning and Table Lookup
https://www.microsoft.com/en-us/research/publication/lut-nn-empower-efficient-neural-network-inference-with-

本文作者：

杨懋博士现任微软亚洲研究院副院长，领导微软亚洲研究院在计算机系统和网络领域的研究工作。

杨懋博士于2006年加入微软亚洲研究院，主要从事分布式系统、搜索引擎系统和深度学习系统的研究、设计与实现。同时领导团队在计算机系统、计算机安全、计算机网络、异构计算、边缘计算和系统算法等方向进行关键技术研究。团队及个人在 OSDI、SOSP、NSDI、SIGCOMM、ATC 等计算机系统和网络的顶级会议上持续发表多篇论文。团队在研究的同时还注重与实际计算机和网络系统的演进结合，与 Azure 云计算、Bing 搜索引擎系统、Windows 操作系统、SQL Server 数据库系统以及多个开源社区密切合作。杨博士同时还是中国科学技术大学博士生导师。

杨懋博士拥有北京大学计算机体系结构专业博士学位以及哈尔滨工业大学硕士和学士学位。

人工智能≠机器“人”：激活基础模型在产业中的巨大应用潜力和商业价值

“人工智能基础模型（Foundation Models），也就是人们常说的大语言模型（Large Language Models）或大模型，在不同产业具有巨大的应用潜力，但一些企业和机构对于人工智能大模型的应用方式，还局限在智能客服、对话机器人，或者文字、图片生成等方面。事实上，基础模型拥有强大的推理、生成和泛化能力，适用于产业界中最具商业价值的任务，如精准预测和控制、高效优化决策，以及智能化、可交互的工业模拟。”

——边江，微软亚洲研究院资深首席研究员

边江，微软亚洲研究院资深首席研究员

随着人工智能大模型（基础模型）的发展，很多企业和机构都对其在生产力场景中的应用表现出极大的热情。不过，我们也观察到这样一个现象：很多产业从业者似乎更关注人工智能接近“人”的一面——像人类一样对话、写作、创作，以及拥有近似于人类的感知能力。比如，很多企业在引入人工智能大模型时，首选场景都倾向于智能客服，对话机器人等“类人”岗位。毫无疑问，这种倾向存在着对大模型理解和应用上的局限，并不能让它在产业界发挥出应有的潜力。

然而，这些局限有其必然性。因为基础模型与生产场景的融合还缺乏成熟且普遍的先例。如果把人工智能看作一种“生产工具”，那么它的应用就类似于“先有工具再发掘用途”，而且人类历史上可能从未有过这种不针对特定需求，而是有着广泛用途但又存在不确定性的工具。

此外，由于不同产业存在更加丰富、复杂的场景，适用于产业界相应场景的基础模型，与通常意义上的基础模型也不尽相同。这就需要对产业大模型进行同步创新，在更多产业场景中充分发挥基础模型的能力，实现人工智能与应用场景的匹配。对于各个产业来说，我们不妨从摆脱思维局限开始，不要将人工智能等同于机器“人”。然后，重新审视和改变现有的业务流程和业务架构，梳理出适应人工智能时代的人与基础模型的合作模式。

基础模型在产业界潜力无限

基础模型是具有通用的数据表示能力、知识理解能力和推理能力的人工智能模型，可以在不同的领域和场景中自然迁移，并快速适应新的环境。与此同时，产业界的数字化平台在经过多年的发展后，已经积累了大量的行业数据，为基础模型提供了更丰富和更适用于特定场景的知识和信息——这让基础模型有了融入产业场景的基础。

在应用价值方面，基础模型强大的推理能力，能够帮助使用者更好地理解数据，从海量的数据中提取有价值的信息，发现数据之间的关联和规律，从而提供更深刻的洞察和更有效的建议。这一优势可以在产业领域的预测、决策、模拟等场景中发挥关键作用。

基础模型的另一个优势是泛化能力。在基础模型出现之前，每个行业场景都需要使用特定数据来训练一个专属人工智能模型，这难以大规模复用，限制了人工智能的商业价值，而基于全世界通用知识训练的基础模型，极大地提升了模型的泛化能力，让产业界不再需要像传统人工智能解决方案那样，为每个场景训练专属模型。

基础模型还可以和生成式人工智能结合，提升工业仿真和智能模拟的准确性、真实性与可交互性，促进数字孪生的实现。工业仿真和模拟都是对真实世界的还原和测试，涉及众多复杂的角色和环境。传统人工智能模型难以支撑大规模仿真，模拟时往往会简化真实情况，或忽略重要的极端事件，影响了模拟和仿真的质量和真实性。生成式人工智能大模型能够支持更广泛的场景，在深入学习特定领域专业知识的基础上，建立特定数据维度分布与真实事件的映射，实现接近现实世界的模拟，更好地辅助工业预测与决策任务，达到工业应用标准。

基础模型在产业界落地需要克服四个难题

在产业界，最重要也最有商业价值的任务包括精准预测和控制，高效优化决策，以及智能化、可交互的工业模拟等复杂任务。这些领域也是传统行业企业应该重点关注的应用方向。然而，通过对现有的 GPT 等基础模型的评测，并结合产业领域的实际情况，我们发现基础模型与真实产业需求之间还存在明显差距，需要克服若干难题，才能使其在产业界发挥更大的作用。

首先，我们缺乏一个能够从纷繁的领域数据中理解复杂领域知识，且可以基于领域知识来构建智能体的通用框架。不同的领域具有各自丰富且复杂的数据，例如物流企业中的海关信息、跨国政策等相关信息；医药行业中 FDA（食品药品监督管理局）药物审查文档；法律行业中的各类法规文档等。构建基于领域知识的智能体需要更通用的框架，从这些数据中提炼出重要的领域知识，发现数据和知识之间的隐含关联，并对它们进行有效的组织和管理。

其次，在文本数据之外，基础模型对结构化数据的处理和理解能力较弱。目前的基础模型最擅长的还是纯文本内容的生成和创作，部分模型也能处理图像、语音等数据。但是，工业场景中的数据往往是数值型、结构化的，如健康监测指标、电池充放电信号、金融信用行为等时序数据或表格数据。现有的大模型还没有针对这类数据进行特定的设计和优化，不能充分理解并处理这些数据，因此很难精准的完成基于这类数据的预测和分类任务。

第三，从应用层面来看，基础模型的决策能力不够稳定和可靠。能源、物流、金融、健康等关键产业场景中最重要的往往是决策类任务，包括物流路径优化、能耗设备控制、投资策略制定、医疗资源调度等，这些任务往往涉及多个变量和多个约束，特别是当面对动态变化的环境时，基础模型还没有完全适应这些复杂的任务，无法直接在产业领域应用。

最后，我们还缺乏对一些特定领域的基础数据的洞察，以及构建特定领域基础模型的方法和经验。很多特定领域的核心信息并不是单纯的文本，因此他们的基础数据也不再是文本中的字和词，而是包含独特的语义结构和关系的新型基础数据，例如金融投资行业中的交易订单信息；生物医药行业中的分子结构信息等，相关领域的核心知识往往隐含在这一类基础数据中，需要更深入和更细致的分析。只有在此基础上构建特定领域的基础模型，才能更有效地挖掘和释放数据的潜力。

构建产业基础模型：融合通用知识与领域专业知识

为了推动基础模型在产业界更快地落地和应用，我们可以着重从以下几个方面入手：

首先，我们可以利用丰富和复杂的产业领域数据，构建更通用、高效和实用的检索增强生成（RAG）框架，可以适配各个垂直领域，帮助提炼出重要的领域知识，发现数据和知识之间的隐含关联，并对它们进行有效的组织和管理。

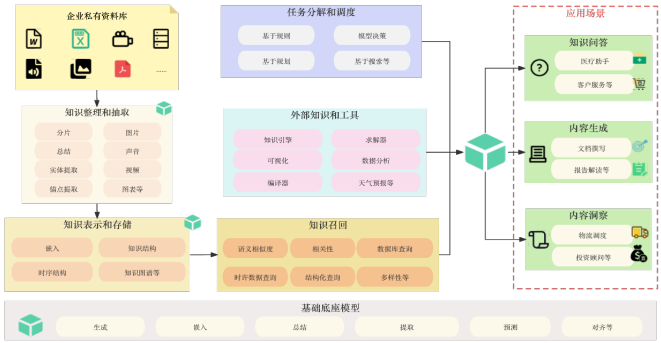


图 1：基于基础模型的更通用、高效和实用的检索增强生成（RAG）框架

其次，基于工业场景中重要的数值数据和相应的结构化依赖，构建适合产业化的基础模型，通过有效融合通用知识和时序数据或表格数据中的领域知识，更有效地解决产业中的预测、分类等任务。

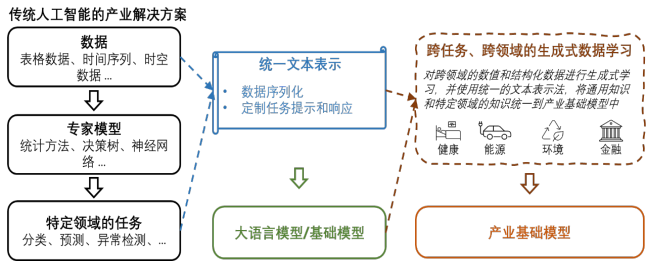


图 2：从传统人工智能的产业解决方案到融合通用与领域知识的产业基础模型

另一个我们目前正在着重探索的方向：利用基础模型已具备的强大的生成、泛化和迁移能力，提高产业决策的质量和效率。在这方面，我们在探索两种路径，一是将基础模型作为一个智能体，二是让基础模型辅助强化学习智能体。

将基础模型作为一个智能体：我们可以利用基础模型的先验知识，结合离线强化学习（Offline Reinforcement Learning），通过持续收集新的领域知识并不断微调，促进智能体的进化，提高作为智能体的基础模型的优化决策能力，使其能更专注于处理产业领域内的任务。

经过优化的基础模型可以在多种产业场景中发挥作用。例如，在方程式赛车中，该基础模型能够优化赛车的轮胎维修策略，根据赛车轮胎的损耗和维修成本，找到最佳的进站维修时间，以缩短赛程、提高赛车排名；在化工企业的产品调度中，利用这一基础模型可以大幅提高产品存储与生产过程中管线协同的效率，从而提升生产执行效率；另外，基于基础模型的泛化能力与鲁棒性，还可以将其快速迁移至空调控制优化的场景中，在保

证舒适温度的同时实现能耗最小化。

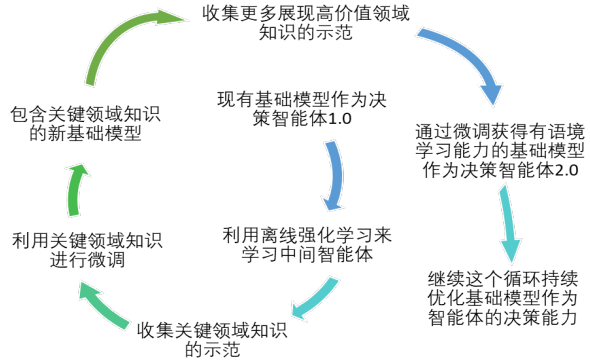


图 3：协同基础模型与离线强化学习构建决策智能体

使用基础模型辅助强化学习智能体：我们可以让模型学习通用表示，在不同的环境和任务中快速适应，从而提升泛化能力。在这一方法中，我们引入了预训练世界模型（Pretrained World Model），它可以模拟人类的学习和决策过程，增强产业决策的效果。通过利用具有广泛知识的预训练世界模型，并采用两阶段预训练框架，开发者能够更全面和灵活地训练基础模型进行产业决策，并将其扩展到任何特定的决策场景。

我们与微软 Xbox 团队合作，在游戏测试的场景中验证了这个框架的有效性。我们利用该框架针对游戏地图预训练了世界模型，解决了在新游戏场景中利用地标观察进行长期空间推理或导航的问题。该预训练模型明显优于没有世界模型或使用传统学习方法的模型，极大地提高了游戏探索的效率。

此外，我们还可以使用领域内专有的基础数据和所蕴含的特定语义信息，打造领域内的基础模型，为智能可交互的决策和模拟开拓新的可能性。比如，我们可以基于金融市场交易订单数据构建金融投资基础模型，这些基础数据是包含丰富语义结构和信息的交易订单，而不是纯文本字符。基于这一金融基础模型，我们可以实现针对不同市场风格的订单流生成，模拟不同市场环境下的大规模订单交易，实现对金融投资市场的可控模拟，从而更好地理解市场变化的规律，探索应对极端场景的策略。

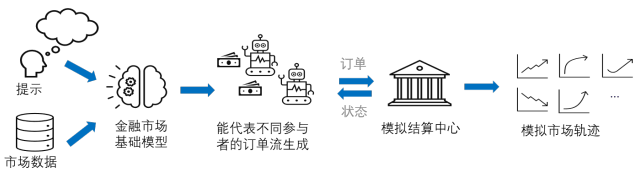


图 4：基于金融基础模型实现针对不同市场风格的订单流生成，从而模拟多样的市场环境

基础模型引领产业数字化转型的下一波浪潮

很早之前，微软亚洲研究院就已经意识到人工智能在产业界的广泛应用需要新的技术探索、尝试和突破，通过跟来自不同产业的合作伙伴合作，我们陆续研发出 Qlib 人工智能量化投资平台、MARO 多智能体资源优化平台、FOST 时空预测工具、BatteryML 电池性能分析与预测平台等开源模型。这些面向产业的人工智能平台、工具和模型，不仅在工业界发挥了重要作用，也为目前基础模型的产业落地提供了重要的数据和工具基础。

借鉴成功的人工智能产业化经验，我们已经开始从前文介绍的几个维度深入探索面向工业领域的人工智能基础模型及其应用。我们发现，在这些突破传统大模型认知的维度上，基础模型拥有巨大潜力，能够深刻促进产业变革。

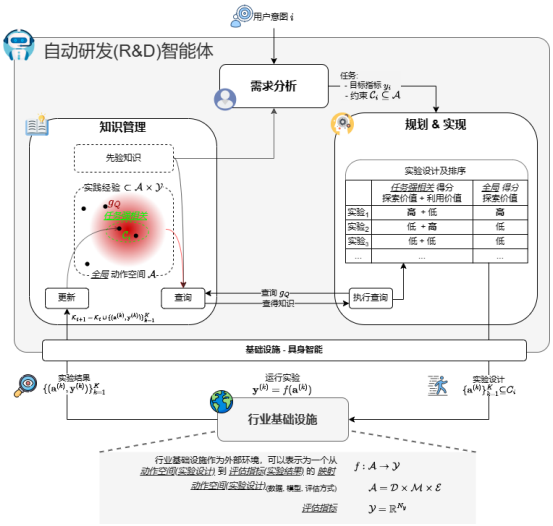


图 5：研发智能体：自动演进以工业数据为中心的研发周期

可以想象，未来基础模型将能够帮助产业界实现行业内的知识自动管理、自动提取、自动迭代。在此之外，我们也在探索基础模型帮助企业实现自动研发，包括研发方向的自动发掘、算法研究方案的自动生成、研发过程和科学实验的自动生成和执行，以及研究思路的自动迭代。换言之，人工智能将能够自主进行数据驱动的产业化研发，这将深刻改变产业界的运作模式。

基础模型将是继互联网和云计算之后，加速产业数字化转型的新动力，并将带来新一波的产业创新爆发。我们期待与更多产业界的合作伙伴一起，深入真实场景，探索基础模型在产业领域应用的更多可能性，充分释放基础模型的商业价值。

相关链接：

Qlib 人工智能量化投资平台
https://www.msra.cn/zh-cn/news/features/qlib-2

MARO 多智能体资源优化平台
https://github.com/microsoft/maro

FOST 时空预测工具
https://github.com/microsoft/FOST

BatteryML 电池性能分析与预测平台
https://github.com/microsoft/BatteryML

本文作者：

边江博士，现任微软亚洲研究院资深首席研究员、微软亚洲研究院机器学习组和产业创新中心负责人，所带领的团队研究领域涉及深度学习、强化学习、隐私计算等，以及人工智能在金融、能源、物流、制造、医疗健康、可持续发展等垂直领域的前沿性研究和应用。

边江博士曾在国际顶级学术会议和期刊上发表过上百篇学术论文，并获得数项美国专利。他曾是多个国际顶级学术会议程序委员会成员，并担任多个国际顶级期刊审稿人。过去几年，他的团队成功将基于人工智能的预测和优化技术应用到金融、物流、医疗等领域的重要场景中，并将相关技术和框架发布到开源社区。

边江博士本科毕业于北京大学，获计算机科学学士学位，之后在美国佐治亚理工学院深造，获计算机科学博士学位。

2024 年值得关注的三大人工智能趋势

对生成式人工智能而言，2023年是具有重要意义的一年。在这一年里，数百万人通过使用 ChatGPT 和 Microsoft Copilot 等工具，让人工智能从实验室走进了我们的现实生活。展望2024年，人工智能将会变得更加便利和实用、更加深入细微，并将融入到那些可以提升我们日常工作和协助解决全球性问题的技术中。以下是微软研究院认为在2024年值得关注的三个重要的人工智能发展趋势。

小语言模型

如果你曾使用 Microsoft Copilot 来回答复杂的问题，那么相信你已经感受到了大语言模型（LLMs）的强大能力。这些模型非常巨大，需要占用大量的计算资源来运行，所以小语言模型（SLMs）的发展就变得十分重要。

参数是影响模型行为的变量或可调节元素。尽管与大语言模型的数千亿参数相比，小语言模型所拥有的数十亿参数要少得多，可这仍然相当庞大，但已足够在手机端离线运行。

微软研究院机器学习基础组负责人 Sebastien Bubeck 表示："小语言模型在规模和成本上的优势，将会让人工智能更加普及。与此同时，我们也在探索新的方法，让它们能够达到和大语言模型一样强大的水平。"

目前，微软研究院已经开发并发布了两个小语言模型——Phi 和 Orca，它们在一些领域展现出了与大语言模型同样甚至更好的性能。而这对"规模是性能的保证"观点也提出了质疑。

不同于使用海量互联网数据进行训练的大语言模型，小语言模型使用的都是筛选的高质量训练数据，研究员们在模型的规模和性能之间取得了平衡。2024年，我们有望看到更多优化过的小语言模型，它们将会推动研究和创新的进一步发展。

多模态人工智能

大多数大语言模型只能处理文本数据，但多模态模型则可以理解来自文本、图像、音频和视频等不同类型的数据。多模态模

型强大的综合能力，可以让技术在搜索工具和创意应用等领域发挥更好的作用，提升它们的丰富性、准确性和连贯性。

Microsoft Copilot 的多模态模型能够处理图像、自然语言和必应的搜索数据，因此用户可以使用 Microsoft Copilot 来获得上传图像中的相关信息。例如，借助 Microsoft Copilot 用户可以了解某张照片中的纪念碑的历史故事。

多模态人工智能也为 Microsoft Designer 提供了支持。Microsoft Designer 是一款可以根据用户描述智能生成视觉图像的图形设计应用。除此之外，多模态人工智能还可以实现微软 Azure AI 服务中的自定义神经语音，或为文本阅读器和需要语音支持的残障人士提供自然的语音输出。

微软首席技术官办公室首席工程师 Jennifer Marsman 表示："多模态能够创造更加人性化的体验，更好地利用人类的各种感官，如视觉、语言和听觉。"

科学探索中的人工智能

专家预测，人工智能工具有潜力在促进科学发现方面实现突破，从而帮助应对如气候变化、能源危机和疾病等全球性问题。

为了缓解气候变化并帮助农民更高效地工作，微软的研究员们正在利用人工智能构建更精确的天气预测工具、碳排放估算工具以及其他有利于农业可持续发展的工具。不仅如此，研究员们还在探索能够帮助农民进行田间作业的人工智能技术。比如，使用聊天机器人帮助农民识别陌生的杂草、基于农场特定数据比较不同灌溉方法的效率等。

在生命科学领域，研究员们正在合作构建全球最大的基于图像的人工智能模型以抗击癌症，并利用先进的人工智能技术来寻找治疗传染性疾病的新药物和突破性药物的新分子。这类技术的研发通常需要经过数年的科学实验和试错，现在借助人工智能，这个过程将缩短至几个月甚至几周。

材料科学是一个致力于开发具有特定性能的新材料的广泛领域。人工智能在这个领域发挥着创新作用。最近一项研究突破展示了人工智能和高性能计算加速寻找低毒性电池材料的能力。

微软研究院科学智能中心负责人 Chris Bishop 表示："人工智能正在推动科学发现的革命。这可能是人工智能最令人兴奋的应用领域，也是最重要的应用领域。"

MicroCinema 与 CCEdit：让文生视频兼具创造性与可控性

随着视频生成技术的飞速进步，我们见证了人工智能技术在视频清晰度、长视频连贯性以及对物理变化理解和镜头转换处理能力方面的显著提升。不过，这些高质量的生成结果是否完全符合我们的需求呢？显然，并非总是如此。由于生成模型的不可预测性，其生成结果常常与用户预期偏离。

在微软亚洲研究院的研究员们看来，创造性与可控性兼备的视频生成模型才是人工智能技术落地应用的关键。基于这一理念，研究员们研发了两项创新的视频生成技术——文生视频模型 MicroCinema 和视频编辑框架 CCEdit，旨在让视频生成更加贴合用户的切身需求。相关论文已被 CVPR 2024 接收。

在这个个性化表达的时代，每个人都是社交媒体内容的创作者。想象一下，拥有一套自己情有独钟的动物表情包——无论是快乐奔跑、好奇探头、悲伤低头，还是悠闲躺卧，每个动作都栩栩如生，同时还能适应各种场景，在古代与现代、虚拟与现实中自由穿梭——这将让你的社交互动更加生动、有趣，也更能展现自我。微软亚洲研究院的 MicroCinema 和 CCEdit 让这一创意想法成为了可能。

为什么由文生视频模型 MicroCinema 和视频编辑框架 CCEdit 生成的视频，能够更精准地匹配文字描述，尤其在视频运动效果上满足用户的具体需求？关键在于它们的"可控性"。

微软亚洲研究院首席研究员罗翀表示："尽管现有的生成技术极大地拓宽了创造的边界，但如果天马行空的创意不能准确地满足用户的实际需求，那么技术的潜力就无法被完全发挥出来并实现落地应用。理想的生成模型应当能够在精确理解用户指令和适应多变场景的基础上，根据用户的特定需求进行实时调整，确保创造力的可控性。"

MicroCinema：分而治之，实现视频按需生成

MicroCinema 之所以能够生成与文本描述高度匹配且流畅连贯的高质量视频，在于其采取了"分而治之"的策略。当前文生视频扩散模型（diffusion model）的主流方法是采用级联时空扩散模型，在文本-视频对之间进行学习，并通过在文本到图像生成模型中加入时间维度，再对文本和视频数据进行微调以创建文本到视频的模型。但是这种方法生成的视频常常会出现外观与时间不一致、不连贯的问题。

MicroCinema 通过将文本到视频的生成过程分为两个阶段来解决这一挑战：首先是文本到图像的生成，其次是图像加文本到视频的生成。在第一阶段，用户可以灵活利用先进的文本到图像模型（如 Stable Diffusion、Midjourney 和 DALL-E）来生成逼真且细节丰富的图像，这些图像作为视频的关键帧，为之后的视频片段生成提供了基础。在第二阶段，通过将这些生成的图像与初始文本一同作为输入，模型便可以减少对细节外观的关注，更专注于学习动态变化。

为了有效实施这一策略，研究员们引入了两项核心技术：利用外观注入网络（Appearance Injection Network）来增强保持给定图像外观的能力；通过外观噪声先验机制（Appearance Noise Prior）保持预训练的 2D 扩散模型的能力。

两段式的设计策略不仅使 MicroCinema 能够生成根据文本提示精确控制动作的高质量视频，而且显著降低了模型从头训练的成本。

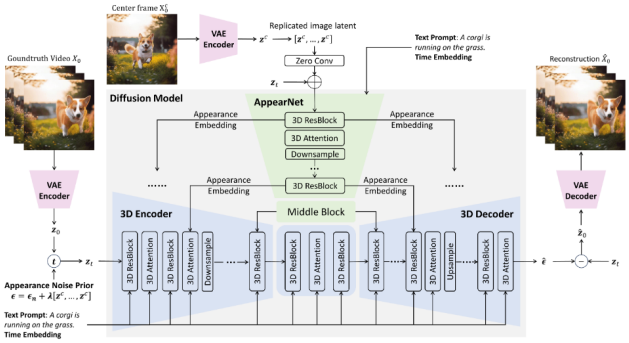


图 1：MicroCinema 架构示意图

CCEdit：三叉戟网络，让视频编辑更可控

尽管现有视频生成模型拥有强大的创造力，但其生成的视频结果并不是总能精确符合用户的编辑意图或艺术构想，特别是无

法对结果进行二次编辑。CCEdit 通过其创新的三叉戟网络结构，有效分离结构控制和外观控制，为用户提供了一系列广泛的编辑功能，包括前景替换、背景修改、风格转换和特效添加等。

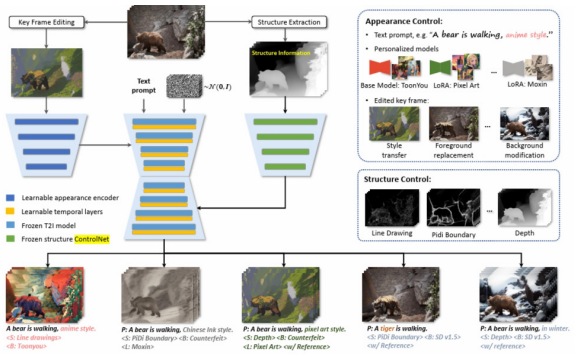


图 2：CCEdit 三叉戟网络示意图

CCEdit 网络由三个关键组件构成：负责文本到视频生成的主分支，以及两个专门用于结构和外观控制的辅助分支。

主分支利用预训练的文本到图像扩散模型，通过插入时间序列层将其转换为文本到视频模型。除了使用文本提示进行外观图像控制，CCEdit 还允许用户调用 Stable Diffusion 社区的个性化文本到图像模型（如 ToonYou、LoRA、Rev Animated）来增强内容的创造性与灵活性。结构分支采用多样的 ControlNet 架构，从输入视频中提取每帧的结构信息和动作轨迹，以实现无缝的结构继承和不同粒度上的结构控制。外观分支则会把通过特征提取处理编辑后的关键帧信息，有效地融入到主分支，支持将用户自己设计的图像作为关键帧，进行细粒度的外观调整。而所有这些控制选项都是在同一框架内无缝集成的。

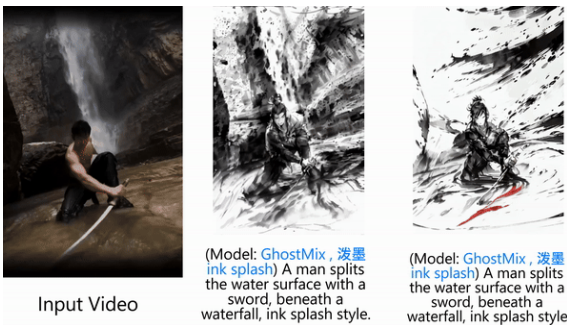
CCEdit 网络的设计策略为用户提供了广泛的控制权，使视频生成不仅拥有无限的创造潜力，同时也具备高度的可控性，满足了用户从细节到整体的个性化编辑需求。

以下是使用 CCEdit 进行视频编辑的几个示例，展现了其在不同场景下的应用能力：

背景编辑，把一名女性在春日的田野里享受美酒的背景改为绿色的田野

全局修改，将一个年轻女孩对着镜头微笑的场景转变成漫画风格

前景编辑，把一只行走的老虎转换成 2D 动画风格



通过插入关键帧，将视频修改为自定义风格

多轮互动修改将是未来视频生成模型必不可少的功能

在这个技术革新的时代，文生视频模型凭借其独有的创造力和强大的性能，“创作”出了许多令人瞩目的作品，同时也引发了大众的广泛关注。然而，在对技术进行前沿探索的同时，微软亚洲研究院智能多媒体组的研究员们深入思考了一个关键问题：如果无法对生成结果进行按需求的精确编辑和调整，那么这些先进工具在日常生活中的实际应用价值又有多大？

正是基于这样的思考，研究员们在开发 MicroCinema 和 CCEdit 的过程中，特别强调了“可控性”这一核心原则。这不仅体现了对技术性能的追求，也反映了他们对人工智能技术未来发展方向的期待。“未来，视频编辑和生成技术将不仅限于单次指令的响应。相反，它们需要能够通过多轮对话与用户进行互动，准确地理解并实施用户的具体需求，从而逐步精细化并满足用户的个性化编辑意图，创造出更加匹配需求的内容。”罗翀说。

MicroCinema 和 CCEdit 的开发标志着研究员们在可控视频生成技术领域的初步探索。未来，研究员们将继续沿着这一思路，进一步推动人工智能视频生成技术的发展，使其更贴近用户的实际需求和创意愿景。

相关链接：

MicroCinema: A Divide-and-Conquer Approach for Text-to-Video Generation
论文链接：https://arxiv.org/abs/2311.18829
GitHub链接：https://wangyanhui666.github.io/MicroCinema.github.io/

CCEdit: Creative and Controllable Video Editing via Diffusion Models
论文链接：https://arxiv.org/abs/2309.16496
GitHub链接：https://ruoyufeng.github.io/CCEdit.github.io/

扫描二维码查看 MicroCinema 和 CCEdit 的视频生成与编辑效果

ViSNet：用于分子性质预测和动力学模拟的通用分子结构建模网络

尽管几何深度学习已经彻底颠覆了分子建模领域，但最先进的算法在实际应用中仍然面临着几何信息利用不足和高昂计算成本的阻碍。为此，微软研究院科学智能中心（Microsoft Research AI4Science）的研究员们提出了通用分子结构建模网络 ViSNet。在多个分子动力学基准测试中，ViSNet 均表现优异。

分子几何建模在理解生物活性机制、化学性质预测、药物设计和蛋白质工程方面发挥着关键作用。然而，虽然几何深度学习（geometric deep learning）是一种低成本、高精度且可以被广泛使用的计算方法，在过去十年取得了巨大进展，但这种技术仍然存在一些有待解决的问题和局限性：

- 分子可解释性不足：深层神经网络尽管可以进行预测，但缺乏对分子的深入洞察；
- 随着分子尺寸的增加，计算成本迅速增加：一些目前最先进的方法中采用的高阶 Clebsch-Gordan 系数计算是计算密集型的，因此阻碍了其在大分子中的应用；
- 需要实际应用中的盲目测试和评估：模型总是在基准测试上进行测试，同时也需要仔细评估在实际应用中的有效性。

为了解决这些难题，微软研究院科学智能中心的研究员们将研究重点聚焦在了如何提高分子可解释性、降低计算成本以及评估实际应用几个方面，并创新性地提出了通用分子结构建模网络 ViSNet (Vector-Scalar interactive graph neural Network)。相关文章"Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing"已发表在《自然-通讯》（Nature Communications）杂志上，并同时入选了"AI and machine learning"和"Biotechnology and method"两个领域的编辑精选文章。

有效提升分子几何表示

研究员们最初计划通过有效且充分地利用分子结构的领域知识来设计模型。由于经典分子动力学（molecular dynamics, MD）通过明确描述势能函数中的键长、键角和二面角来模拟分子运动，所以受经典 MD 模拟的启发，研究员们将这些项目转换并扩展，从而构建了 ViSNet 独特的模型设计。

与通过简单的特征工程过程直接采用角度或二面体信息不同，研究员们提出了"方向单元"这个概念作为节点的向量化表示，即从中心节点到其任何第一个相邻节点的所有归一化向量的总和，作为中心节点的矢量化表示。再以此将键长、键角和二面角计算扩展到二体、三体和四体相互作用。然后，通过设计运行时几何计算（runtime geometry calculation, RGC）模块来描述模型操作等多体交互。

更重要的是，三体和四体相互作用的 RGC 计算都只有线性时间复杂度。因此，研究员们又进一步提出了向量标量交互式消息传递机制（ViS-MP），其中方向单元会通过构建块由节点和边的标量表示迭代更新，反过来，标量表示由方向单元同时更新 RGC 模块。RGC 和 ViS-MP 的独特设计显著增强了几何编码能力并加速了分子图神经网络中的消息传递过程。

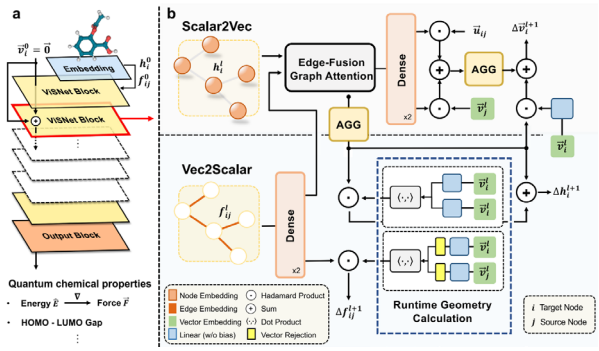


图 1：ViSNet 网络结构示意图

图 2：ViSNet 在 Pytorch Geometry 中作为基础模型

ViSNet在分子建模和性质预测实际应用中的表现

研究员们首先将 ViSNet 在广泛使用的分子化学性质预测基准上进行了评估。在 MD17、修订版 MD17 、 MD22、QM9 以及 Molecule3D 数据集上显示出卓越的性能，证明了分子几何表示的强大能力。然后，研究员们还在自己开发的 DFT（密度函数理论）精度的蛋白质数据集 AIMD-Chig 数据集上训练了 ViSNet，并对蛋白质 Chignolin 进行了 MD 模拟。

ViSNet 取得了比经验力场和现代机器学习力场更好的性能及令人满意的结果。ViSNet 的模拟结果与在 DFT 水平上获得的结果非常接近，这表明 ViSNet 在数据效率和模拟保真度方面具有潜力。

研究员们用 ViSNet 参加了全球首届 AI 药物研发算法大赛。该大赛旨在根据小分子的序列信息（即SMILES）预测针对新冠病毒 SARS-CoV-2 主要蛋白酶的抑制剂。共有来自全球878支团队的1105名参赛者参与了此次比赛。最终，研究员们凭借 ViSNet 获得了比赛的总冠军，也展现了 ViSNet 优异的预测准确性。

如何获取ViSNet模型？

为了促进更广泛的应用和便捷的使用，ViSNet 已被微软纳入 Pytorch Geometry 库，作为分子建模和属性预测领域的基本模型。ViSNet 的定期维护和更新版本也可在 GitHub上 获取。

此外，考虑到图神经网络随着模型变得越来越大、越来越深，可能会遇到"过度平滑"的风险，研究员们还进一步设计了 ViSNet 的 Transformer 版本，可以将 RGC 模块转移到 Transformer 注意力计算中，并提出了一种新颖的原子间位置编码（IPE），命名为 Geoformer（Geometric Transformer的缩写）。作为 ViSNet 的 Transformer 变体，Geoformer 可通过堆叠数百个注意力块来进行大模型训练。相关研究论文发表于 NeuralPS 2023。

分子动力学模拟的未来：兼具人工智能与从头计算精度的能力

作为人工智能（AI）驱动的从头算分子动力学（AI2BMD）项目的重要组成部分，ViSNet 致力于实现加速分子动力学模拟的目标，使大型分子系统的模拟精度接近从头算法。

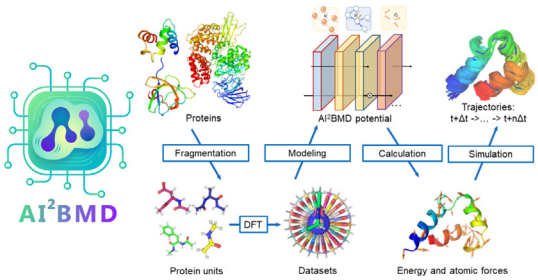


图 3：AI2BMD 整体流程

ViSNet 可以让 AI2BMD 实现对包含超过10,000个原子的蛋白质的能量和力计算达到接近从头算法的精度。利用 ViSNet 进行蛋白质动力学模拟还可提高自由能估计的准确性，提供有关蛋白质折叠热力学的深入预测，并有助于表征蛋白质的特性，从而潜在地增强实验研究。

相关链接：

ViSNet论文
https://www.nature.com/articles/s41467-023-43720-2

AIMD-Chig 数据集
https://www.nature.com/articles/s41597-023-02465-9

首届AI药物研发算法大赛官方网页
https://aistudio.baidu.com/competition/detail/1012/0/leaderboard

ViSNet-Pytorch Geometry

https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.nn.models.ViSNet.html

ViSNet-GitHub
https://github.com/microsoft/AI2BMD/tree/ViSNet

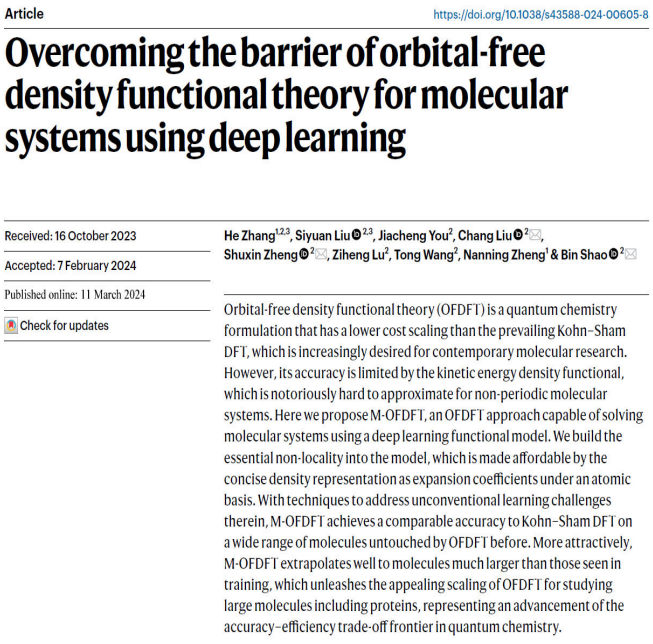
Geoformer
https://github.com/microsoft/AI2BMD/blob/Geoformer/Geoformer.pdf/

AI2BMD
https://microsoft.github.io/AI2BMD/index.html

AI 助力 M-OFDFT 实现兼具精度与效率的电子结构方法

为了使电子结构方法突破当前广泛应用的密度泛函理论（KSDFT）所能求解的分子体系规模，微软研究院科学智能中心的研究员们基于人工智能技术和无轨道密度泛函理论（OFDFT）开发了一种新的电子结构计算框架 M-OFDFT。这一框架不仅保持了与 KSDFT 相当的计算精度，而且在计算效率上实现了显著提升，并展现了优异的外推性能，为分子科学研究中诸多计算方法的基础——电子结构方法开辟了新的思路。相关研究成果已在国际知名学术期刊《自然-计算科学》（Nature Computational Science）上发表。

近几十年来，理论与计算化学领域取得的一大成就是能够通过计算手段得到分子体系的物理化学性质。这为药物发现和材料设计等诸多工业界问题带来了全新的研究手段，有望缩短开发流程并降低开发成本。这些计算方法的基础步骤是使用电子结构方法求解给定分子体系的电子状态，进而得到该体系的各种性质。



Article https://doi.org/10.1038/s43588-024-00605-8

Overcoming the barrier of orbital-free density functional theory for molecular systems using deep learning

Received: 16 October 2023
Accepted: 7 February 2024
Published online: 11 March 2024
Check for updates

He Zhang[1,2,3], Siyuan Liu[2,3], Jiacheng You[2], Chang Liu[2], Shuxin Zheng[2], Ziheng Lu[2], Tong Wang[2], Nanning Zheng[1] & Bin Shao[2]

Orbital-free density functional theory (OFDFT) is a quantum chemistry formulation that has a lower cost scaling than the prevailing Kohn–Sham DFT, which is increasingly desired for contemporary molecular research. However, its accuracy is limited by the kinetic energy density functional, which is notoriously hard to approximate for non-periodic molecular systems. Here we propose M-OFDFT, an OFDFT approach capable of solving molecular systems using a deep learning functional model. We build the essential non-locality into the model, which is made affordable by the concise density representation as expansion coefficients under an atomic basis. With techniques to address unconventional learning challenges therein, M-OFDFT achieves a comparable accuracy to Kohn–Sham DFT on a wide range of molecules untouched by OFDFT before. More attractively, M-OFDFT extrapolates well to molecules much larger than those seen in training, which unleashes the appealing scaling of OFDFT for studying large molecules including proteins, representing an advancement of the accuracy–efficiency trade-off frontier in quantum chemistry.

M-OFDFT 相关研究已发表在《自然 - 计算科学》（Nature Computational Science）上

然而，各种电子结构方法的求解精度和计算效率往往无法兼得。当前，取得相对合理的"精度-效率"权衡而被广泛应用的方法是 Kohn-Sham 形式的密度泛函理论（Kohn-Sham density functional theory, KSDFT）。但 KSDFT 具有较高的计算复杂度，不能很好地满足日益增长的求解大规模分子体系的需求。为此，微软研究院科学智能中心的研究员们提出了一种基于深度学习和无轨道密度泛函理论（OFDFT）的电子结构计算框架 M-OFDFT，其不仅显著超越了 KSDFT 的计算效率，还能保有其求解精度。这一成果展示了人工智能在提升电子结构计算中"精度-效率"权衡方面的卓越能力，并将助力加速相关业界问题的研究与开发。M-OFDFT 的相关研究成果已在国际知名学术期刊《自然-计算科学》（Nature Computational Science）上发表。

人工智能给电子结构方法带来新机会

电子结构方法是求解分子体系各种物理化学性质的基础工具。由于多电子体系本身具有一定的求解难度，所以高精度电子结构方法因其较高的计算代价很难应用到工业界所关注的分子体系中，而可计算较大分子的方法则会因引入一些近似而损失精度。目前 KSDFT 因其相对合适的精度与效率权衡得到了广泛应用。

不过，近期人工智能技术的喜人进展也为其他电子结构计算框架带来了新的机会。为了使电子结构方法突破 KSDFT 所能求解的分子体系规模，微软研究院的研究员们利用人工智能技术，

开发了 M-OFDFT，该方法比 KSDFT 效率更高，同时又能保有其精度。基于 OFDFT 的开发，让 M-OFDFT 成为了一种比 KSDFT 理论复杂度更低的电子结构计算框架，因为它只需优化电子密度函数 ρ(r) 这一个函数来求解电子状态即可，KSDFT 则需要优化与电子数相同的多个函数。

不过，OFDFT 面临着一个巨大的挑战——需要电子动能关于密度函数的泛函 T_S [ρ]，但它的形式未知，并且难以构造适用于分子体系的高精度近似。

针对这一难题，M-OFDFT 使用一个深度学习模型 T_(S,θ) 来近似动能泛函。借助深度学习模型的强大拟合能力，M-OFDFT 可实现比基于近似物理模型设计的经典动能泛函更高的准确度。对于一个待求解的分子体系结构，M-OFDFT 会使用动能泛函模型 T_(S,θ) 以及其他可直接计算的能量项构造出一个电子密度的优化目标，然后通过优化过程求解最优（基态）电子密度（图1），进而可计算能量、力、电荷分布等分子属性。

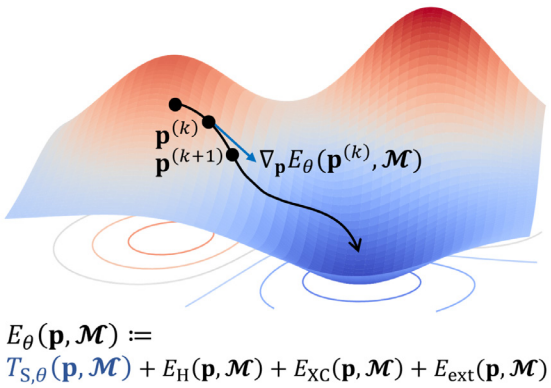


图 1：对于待求解的分子体系结构 M，M-OFDFT 通过最小化电子能量 E_θ 来求解电子密度（以其向量化系数 p 表示），其中难以近似的动能部分由深度学习模型 T_(S,θ) 来近似

M-OFDFT实现兼具精度与效率的电子结构方法

研究员们对 M-OFDFT 进行了一系列的实验验证。首先考察的是 M-OFDFT 在常见小分子体系上的求解精度。结果显示，M-OFDFT 在乙醇分子构象以及 QM9 数据集的分子上可以达到与KSDFT相当的精度（能量达到化学精度1 kcal/mol）。相较于经典 OFDFT 方法，精度提高了两个数量级（图2-a）。M-OFDFT 解得的电子密度也与 KSDFT 的结果重合（图2-b），特别是得到了电子壳层结构，而经典 OFDFT 的结果则有明显偏差。由 M-OFDFT 解得的乙醇构象空间上的势能面（每个点都是通过密度优化得到的，并不是直接预测）也与 KSDFT 的结果一致（图2-c）。

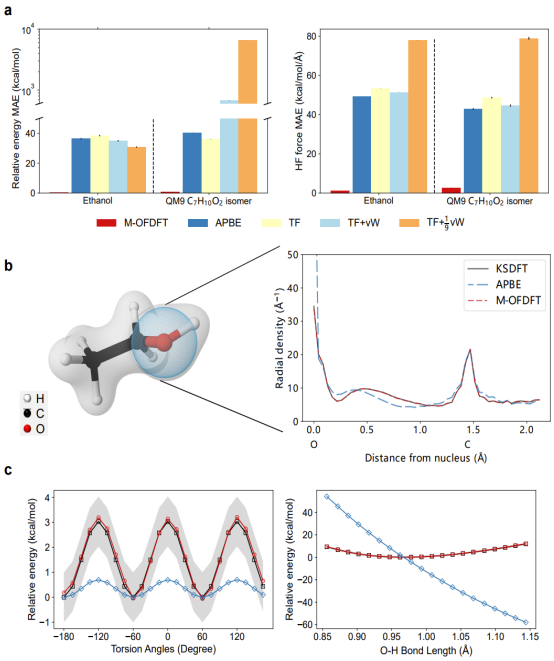


图 2：M-OFDFT 和一些经典 OFDFT 在分子体系上与 KSDFT 的比较

之后，研究员们又验证了 M-OFDFT 不仅保有 KSDFT 级别的精度，更低的理论计算复杂度还使其在效率上也超越了KSDFT。在实际计算中 M-OFDFT 取得了 O(N^1.46) 的复杂度（如图3所示），比 KSDFT 的实际复杂度 O(N^2.49) 低了一阶，且其所需绝对时间也明显少于 KSDFT。在两个更大的蛋白质体系上（包含2676和2750个电子），M-OFDFT 实现了25.6倍和27.4倍的加速。

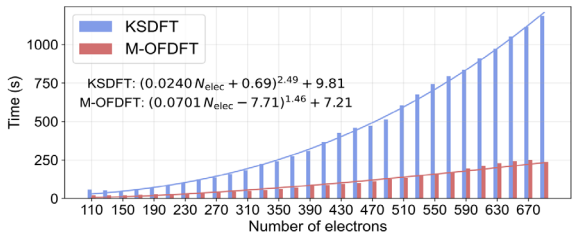


图 3：M-OFDFT 和 KSDFT 的实际计算时间及复杂度

M-OFDFT具有更强的泛化能力

深度学习模型在科学任务中的应用面临一大挑战是，在具有与训练数据不同特点的数据上的泛化问题。但采用了 OFDFT 框架后，动能泛函模型遇到的泛化问题就会减轻，从而使 M-OFDFT 可以在比训练集分子规模更大的体系上展现出良好的外推能力。

实验结果表明，M-OFDFT 的能量预测误差显著低于基于深度学习的端到端能量预测模型（图4-a）。此外，研究员们还利用在多肽片段上训练的 M-OFDFT 模型求解完整蛋白结构，并取得了超越端到端模型和经典 OFDFT 的泛化性能（图4-c）。不仅如此，相较端到端模型，M-OFDFT 还可以用更少的大分子体系

训练数据取得更好的泛化表现（图4-b与图4-d）。

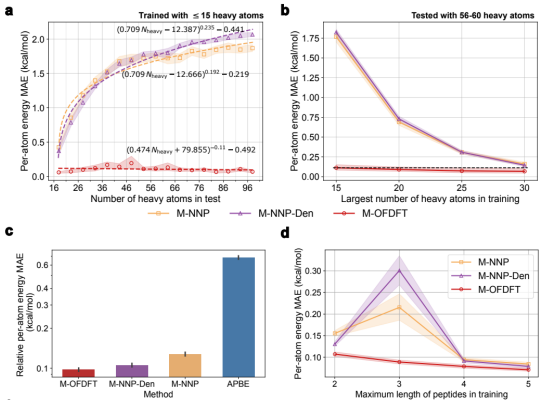


图 4：M-OFDFT 和其他深度学习方法的泛化性能比较

M-OFDFT的工作原理:“神龙见首又见尾”:高效捕获非局域效应的动能泛函模型

动能密度泛函具有明显的非局域效应，而用经典的基于格点（grid）的方式表征电子密度则会带来高昂的非局域计算代价。为此，M-OFDFT 将电子密度在一组原子基组函数上展开，并使用展开系数 p 作为电子密度表征。由于基函数叠加的形状与电子分布接近，所以其数量可远小于格点数，使得非局域计算代价大大降低，并有助于刻画电子密度中的壳层结构。

M-OFDFT 将每个原子上的电子密度系数 p 和类型 Z 与坐标 x 作为节点特征，并基于 Graphormer 模型[1]预测电子动能 T_(S,θ)（图5），其自注意力机制显式刻画了荷载在每两个原子上的电子密度特征之间的相互作用，从而可捕捉非局域性质。此外，为了保证动能的旋转不变性，M-OFDFT 使用了以各个原子为中心、基于其相邻原子的局部坐标系，将电子密度系数转换为旋转不变的特征。

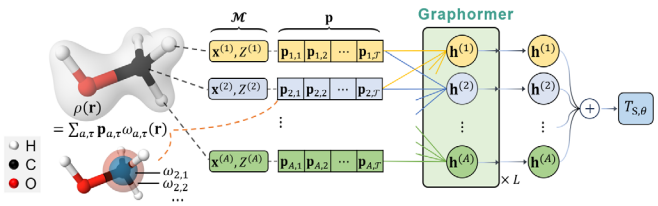


图 5：基于非局域图神经网络的动能密度泛函模型

“横看成岭侧成峰，远近高低各不同”：高效学习电子能量曲面的训练策略

与传统机器学习任务不同，动能泛函模型是被当作其输入变量的优化目标使用的，而非用于在一些单点上做预测，这对模型的学习提出了更高的要求：模型必须捕捉到每个分子结构上电子能量曲面的轮廓。

为此，研究员们深入分析了用来生成数据的电子结构方法，发现它其实可以为每个分子结构生成多个数据点，而且还能提供梯度标注，从而让模型可以拥有更丰富的曲面轮廓特征。然而梯度的巨大范围也使神经网络难以优化。对此，研究员们还提出了一系列增强模块，让模型能够更容易地表达巨大的梯度。

开启未来电子结构方法的新篇章

M-OFDFT 成功突破了无轨道密度泛函框架在分子体系中的瓶颈，将其求解精度提升到了常用的 KSDFT 的水平，同时保有了其更低的计算代价，推进了电子结构方法在“精度-效率”方面的权衡，为分子科学研究提供了一种更有潜力的研究工具。

尽管 M-OFDFT 已经在某些分子体系上展现了出色的泛化性能，但在更大的分子体系上实现长时间且稳定的高精度模拟仍是一个巨大的挑战。微软研究院期待 M-OFDFT 可以沿着这一方向激发更多研究与创新，并在未来和其他方法一起为电子结构计算带来更多突破性的成果和影响。

相关链接：

Do Transformers really perform badly for graph representation? Advances in Neural Information Processing Systems 34 (NeurIPS 2021)
https://proceedings.neurips.cc/paper/2021/hash/f1c1592588411002af340cbaedd6fc33-Abstract.html

Overcoming the Barrier of Orbital-Free Density Functional Theory for Molecular Systems Using Deep Learning
《自然-计算科学》文章链接：https://www.nature.com/articles/s43588-024-00605-8
SharedIt链接：https://rdcu.be/dANtS
论文链接：https://arxiv.org/abs/2309.16578

优化 LLM 数学推理；深度学习建模基因表达调控；基于深度学习的近实时海洋碳汇估算

本期内容速览:

01. CoT-Influx：专注剪枝冗余的 CoT 样本和 token，突破少样本学习极限，显著提高各类大语言模型解决数学推理任务的能力。

02. CREaTor：通过自注意力机制建模细胞特异的基因表达调控，解决数据缺失与调控机制表征的复杂问题，同时具有优秀的泛化能力和零样本学习能力。

03. CMO-NRT：利用深度学习和半监督方法，提高全球海洋碳汇的及时性和精确性，助力气候变化管理。

扫描二维码了解更多信息

高性能计算与深度学习结合；提升云人工智能基础设施可靠性；基于心理测量学的通用型人工智能评估；模仿人脑思维模式的视觉语言规划框架

本期内容速览:

01. PPoPP 2024 最佳论文ConvStencil：突破高性能计算与人工智能的“软硬”边界

通过高效将stencil计算转换为张量核心单元上的矩阵乘法，实现了传统高性能计算利用深度学习硬件进行加速，有望提高各种科学和工程应用的性能。

02. Anubis：通过主动验证提升云人工智能基础设施的可靠性

针对云人工智能基础设施进行主动验证，减少硬件冗余引起的性能下降，提升整体可靠性，已成功部署在微软 Azure 云平台的云人工智能基础设施中。

03. 基于心理测量学的通用型人工智能评估方法

通过识别关键心理构念、开发测验、引入信度和效度概念，实现对人工智能进行评估的准确性及可靠性。

04. VLP：类似于人类左右脑思维模式视觉语言规划框架

结合语言规划分支和视觉规划分支，实现多模态大模型的视觉推理能力。

扫描二维码了解更多信息

提示词专场：从调整提示改善与 LLMs 的沟通，到利用 LLMs 优化提示效果

本期内容速览：

01. EvoPrompt：连接大语言模型与进化算法，打造高效的提示词优化器

通过将大语言模型卓越的语言处理能力与进化算法卓越的优化能力相结合，在多个自然语言理解和生成任务中获得出色的提示词。

02. 具有结构化语言知识的分层视觉语言模型提示学习

分层提示微调（HPT）方法可将大语言模型中的结构化语言知识和传统语言知识进行结合，以分层学习提示的方式提高提示效果。

03. LLMLingua系列研究：通过提示压缩重排优化人与大语言模型的沟通

受信息论和通信系统的启发，采用压缩与重排来浓缩内容并确保关键信息获得更多关注，进而显著提高"多文档问答"等任务的效果。

04. 基于大语言模型的工业控制优化

通过优化提示词中的示例部分，无需微调模型即可优化工业控制任务，展现了基于 LLMs 的方法强大的泛化性和鲁棒性。

扫描二维码了解更多信息

AAAI 上新：从生成检索、跨语言迁移，到负责任的视觉合成

本期内容速览:

01. HORIZON: 高分辨率语义可控的全景图像合成

高分辨率全景图生成框架HORIZON利用球面建模技术，有效解决了球面扭曲和边缘不连续的问题，同时通过并行解码技术显著提高了全景图生成的效率。

02. 排序学习用于生成式检索

通过直接学习序列排序目标，实现生成模型的优化目标和排序结果的统一。

03. 使用机器创建的通用语言实现更好的跨语言迁移

将不同语言中的共享概念映射到相同的通用词汇，有效提高跨语言转移的效果。

04. ORES: 开放词表下的可靠视觉合成

开放词汇表下的负责任的视觉合成以及对应的解决方案 TIN 框架，实现负责任的视觉输出，降低图像生成风险。

05. 基于强化条件的文本扩散

通过奖励信号优化扩散模型的去噪过程，提升文本生成质量及效率，显著降低文本扩散模型对大量去噪步数的依赖。

扫描二维码了解更多信息

科学匠人 | 韩东起：从理论物理到神经科学和人工智能，迈向学科交叉的新方向

从对物理现象的探索到对生物本质的研究，从托卡马克可控核聚变到认知神经机器人，微软亚洲研究院研究员韩东起为什么会做出如此大跨度的科研转向？从冲绳科学技术大学院大学（OIST）到微软亚洲研究院，从实习生到研究员，是什么原因推动他坚持长期主义研究？从人工智能驱动的神经科学研究，到脑启发的神经网络和具身智能机器人，他所从事的研究对全球社会又有着怎样的意义和价值？让我们通过科学匠人韩东起的故事一起来感受跨学科研究的魅力。

做基础研究，还是做应用研究？或许是许多科研人员都曾面临的选择。然而，对即将毕业的韩东起来说，他却"贪心"地希望自己以后能够兼顾基础研究和应用研究。

通过多方面了解，韩东起选择了微软亚洲研究院作为实现自己愿望的初步尝试，并在博士毕业前成为了微软亚洲研究院（上海）的一名实习生。半年的实习，让韩东起亲身感受到了研究院不仅致力于各种基础研究创新，而且还积极推动与产业界的合作，并为创新成果的应用创造一切可能的机会。这为他后来正式加入研究院埋下了一颗种子。

"微软亚洲研究院为研究人员搭建了一个能够结合基础研究和工业应用实践的平台。在这里，我既可以和世界一流的研究人员合作，探究智能的原理和方法，也可以将人工智能技术应用于医疗健康等领域。这也是我博士毕业后选择加入微软亚洲研究院的主要原因之一。这里可以说是我科研职业生涯的最佳起点。"韩东起说。

韩东起，微软亚洲研究院研究员

从人工智能与脑科学交叉研究中，探索智能的本质

近年来，人工智能等计算机技术与其他学科的交叉融合日益紧密，微软亚洲研究院也一直积极布局相关研究，并不断加大投入力度。而上海这座海纳百川的城市，拥有众多知名学府和顶尖医疗机构，为人工智能在医疗健康领域的跨界探索提供了良好的环境。因此，人工智能与脑科学等医疗健康领域的交叉研究也成为了微软亚洲研究院（上海）的重要研究方向之一。

韩东起博士毕业于日本冲绳科学技术大学院大学（OIST），主要从事认知神经机器人，即机器人与脑神经科学方面的研究。作为神经科学家，韩东起的加入为微软亚洲研究院的人工智能与脑科学交叉研究团队带来了全新的专业力量。

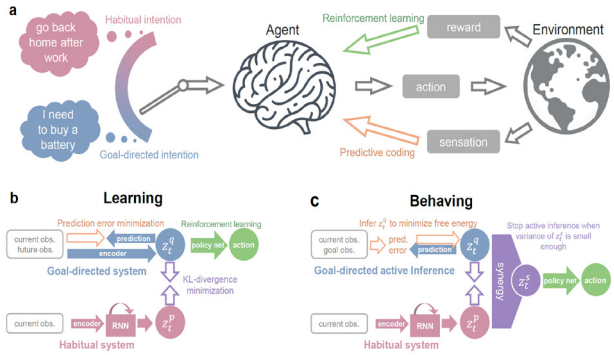
韩东起（第二排左二）与团队同事合影

韩东起当前的研究主要集中在两个方向：人工智能与脑科学，以及人工智能在医疗健康领域的应用。在他看来，人工智能与神经科学的交叉研究不仅有着重要的理论意义，也有着广泛的应用价值。"无论是脑机接口，还是人机交互，都需要我们对人类的认知和感受机制有更深入的了解，这样才能设计出更自然、高效的人机交互系统。"韩东起表示。而在医疗健康领域，目前全球有约10亿人受神经系统疾病的影响。人工智能与大脑的研究，可以为医生和患者提供更多信息，从而更好地理解、预防和治疗这些神经性疾病。"人工智能与脑科学之间有着千丝万缕的联系，它们都是在探索智能的本质和运行机制，两者有许多共同的问题和挑战，同时也可以相互借鉴、启发。"

事实上，人工智能的很多技术都受到了人脑神经网络的启发，比如多层感知器、长短时记忆网络等。这意味着我们可以通过研究人类或动物的认知功能，如学习、记忆、决策，来增强人工智能的能力。当前人工智能面临的一个严峻挑战是" 灾难性遗忘 "现象，即模型在学习新数据时会忘记之前学到的信息，但人类大脑却不会出现这样的问题。韩东起和同事们希望通过研究人脑来帮助人工智能克服这类难题。

另一方面，人工智能强大的数据处理和建模能力，可以推动神经科学的研究和应用。人脑大约有1011个神经元和1015个神经连接，要想有效地处理如此庞大的数据量，可能就需要用AI来建模人脑的工作原理和计算方法，验证神经科学理论的有效性。

目前，韩东起和同事们已经在相关研究上取得了一些成果。在一项研究中，他们主要关注两种不同的人类行为：习惯性行为（habitual behavior）和目标导向性行为（goal-directed behavior）[1]。以人类行走为例，习惯性行为是无需刻意思考的自动化行为，比如下班回家的路线选择；而目标导向性行为则是需要考虑目的和结果的有意识行为，比如领结婚证时要如何到达民政局。这两种行为模式可以解释很多生物行为特征，然而人们目前仍不清楚人脑是如何在这两种行为中做出选择，以及两种行为是如何相互影响的。

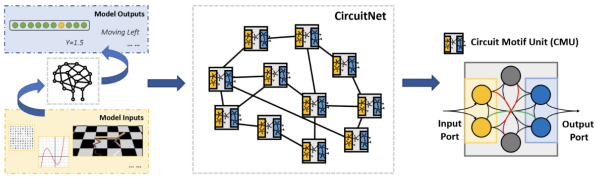


习惯性和目标导向性行为的计算建模

韩东起说："我们团队用深度学习和机器学习理论进行建模，来探究两种行为的特性和神经机制，为认知科学和心理学研究提供了支持。同时，通过对这两种行为的研究，我们也能够从中汲取灵感，进而设计新的人工智能算法。"

韩东起和同事们的另一项研究是基于对大脑神经回路结构的模拟，提出了一种新型神经网络 CircuitNet[2]。人脑神经元连接具有局部密集和全局稀疏的特点，即相邻的神经元在同一个神经核团内紧密连接，但在不同脑区之间稀疏连接，微软亚洲研究院的研究员们希望探究这种连接模式的原理和优势。韩东起从实习期间就参与到了这个研究项目中，最终在全组同事的共同努力下 CircuitNet 破壳而出，相关论文也成功入选了 ICML 2023。

作为一种更加节能、高效的神经网络架构，CircuitNet 能够以更少的参数取得更好的性能。据估算，人脑的平均功耗仅不到20瓦，而人工智能大模型如 GPT-4 的功耗却高达数百上千瓦。未来，韩东起和同事们将在 CircuitNet 的基础上，继续探索人脑节能的奥秘。



CircuitNet 的模型结构

此外，韩东起还致力于深度强化学习和具身智能（Embodied AI）的研究，希望让人工智能可以更好地学习和决策，实现智能机器人与真实世界的交互。"现在的人工智能大模型主要用于内容生成，比如文字、图像，生成的结果就是最终的状态。而具身智能机器人输出的则是动作，在与真实世界交互的过程中会存在很多不确定性，例如机器人画一幅画可能会出现各种状况——画错了、笔断了、笔被拿走了，这些都会影响最终的输出结果。这些动作的选择对于机器人来说是非常复杂的，而我们可以借鉴人脑的决策原理，加速具身智能的实现。"

跨学科的思想碰撞为跨领域研究带来新灵感

韩东起一直对世界充满了好奇心，对科学研究有着广泛且浓厚的兴趣。他的本科专业是理论物理，在他看来，物理是一门极具挑战性的学科，它不仅要求学习者具备严谨的逻辑思维，还要有扎实的数学功底、实验设计和数据分析的能力，这些都能有效地提升一个人的综合素质。此外，物理科学也是很多现代科技的基石，应用范围十分广泛。

在物理领域，韩东起本科期间主要感兴趣的研究方向是托卡马克（一种可控核聚变装置）中物理参数的探测[3]。"核聚变是一种充足、高效、清洁的能源，一旦我们掌握了它，将给人类带来不可估量的收益。"实现可控核聚变是他最初的科研理想。但是，经过一段时间的学习，他发现，制约可控核聚变落地的并不是基础物理理论的突破，更多的是工程实践上的问题。这意味着，他所学习的理论物理无法直接推动理想的实现。而在托卡马克的研究中，韩东起接触到了机器学习技术，这让他萌生了转换专业的想法。于是在攻读博士学位期间，他选择了认知神经机器人专业。

跨学科学习意味着很多工作都要从零开始。博士一年级时，韩东起就学习了很多新的基础课程，如基础神经科学、机器学习、机器人控制自动化技术，以及认知科学、心理学等跨学科知识。虽然学习过程十分艰辛，但他也从不同学科知识的碰撞中激

发出了很多新的想法，为人工智能与脑科学的研究带来了优势。

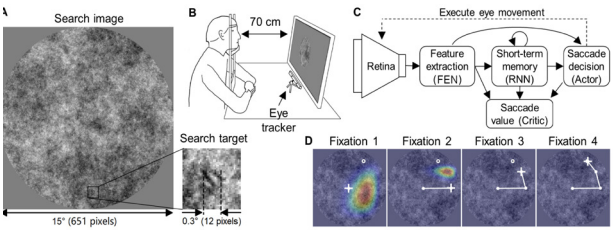
在强化学习和具身智能的研究中，韩东起会代入物理学的思维方式，把物理实验中的严谨思维（例如常用的测量统计方法）和逻辑推理训练的方法引入到人工智能的研究中。

“跨学科学习的好处，不仅仅体现在知识的跨领域应用，更体现在逻辑思维的借鉴和启发上。比如，在解决神经网络机器学习的问题时，传统的机器学习思维，往往首先考虑的是数据量和模型规模。神经科学的训练则可以让我们从人类或动物大脑的角度出发，以更具拓展性和灵活性的方式来思考问题。”韩东起说。

跨领域合作：用前沿技术解决实际问题

除了跨学科的基础研究之外，韩东起和同事们还与国内外高校、医疗机构展开了跨领域的合作，将前沿技术应用于解决实际问题，寻求更好的解决方案。

韩东起所在团队与复旦大学联合开发了一种仿人眼视觉的机器视觉人工智能模型[4]，目标是提高计算机视觉系统的节能性。“在合作中，我们发现人类视觉和计算机视觉有很大的差异，人眼视觉中心的分辨率远高于边缘分辨率，而且人脑是用脉冲（spike）来传输信号的，脉冲神经网络也是大脑的一个特征。”为了探究人类视觉的优势，联合研究团队使用脉冲神经网络建模了一个视觉系统，可模拟不均匀的视觉分辨率和大脑脉冲信号传输。这个模型在完成视觉任务的同时，在理论上可以实现百倍的能效优化，能够以更少的能耗处理相同的视觉任务。



仿人眼脉冲神经网络进行视觉搜索任务示意图

在与韩国科学技术院合作的机器人项目中，双方的研究员们尝试使用可穿戴的、非侵入式设备读取人的脑电波信号，让机器人能够更准确地理解人的意图。“比如，当一位老人伸手指向厨房，机器人就要判断老人是不是饿了或者想要拿什么东西，又比如，指向桌子，桌子上有一瓶水和一盒纸巾，他到底需要哪一个？”韩东起说，“韩国科学技术院在神经科学领域积淀深厚，而我们则能够提供更优秀的人工智能算法。通过将脑科学的模型与AI技术结合，我们可以从多角度的研究和实践来帮助机器人更好地完成任务。这种跨领域的合作研究为双方带来了新的研究思路。”

从对游戏的钻研到对科研的坚持

工作之余，韩东起有着丰富多样的兴趣爱好。他喜欢徒步旅行以及羽毛球等体育运动，对各类电子游戏也情有独钟。“我从小学就开始玩电子游戏，后来一段时间，我迷上了一款需要很大耐心且挑战无上限的游戏。只有通过屡次失败和反复练习，我才能有一点进步。这种持续的自我提升给了我很强的成就感，也磨练了我的心态，使我更具耐心和毅力去面对生活中的各种困难。”他说，“这也影响了我的研究风格——进行长期持续性的研究。现在，研究院又给我提供了一个多元、包容的研究环境，让我能够与不同背景、不同专业和不同性格的同事一起，热情、高效的合作和交流，这使得我在长期主义研究的道路上得到了更多支持，也有了更多志同道合的伙伴。”

相关链接：

[1] Synergizing Habits and Goals with Variational Bayes
https://osf.io/preprints/psyarxiv/v63yj

[2] CircuitNet：A Generic Neural Network to Realize Universal Circuit Motif Modeling
https://proceedings.mlr.press/v202/wang23k/wang23k.pdf

[3] Energy-Efficient Visual Search by Eye Movement and Low-Latency Spiking Neural Network
https://arxiv.org/abs/2310.06578

[4] In situ relative self-dependent calibration of electron cyclotron emission imaging via shape matching
https://doi.org/10.1063/1.5038866

科学匠人|李潇：Aspire的力量——年轻科研人员的成长加速器与沟通桥梁

*他曾是微软亚洲研究院的“首席实习生”，也是2019届微软亚洲研究院与中国科学技术大学的联合培养博士生。经历了互联网行业的实践之后，他选择重回微软亚洲研究院，成为了一名多媒体计算组的研究员，同时也是微软亚洲研究院 Aspire 社团的核心力量之一。他就是江湖人称“潇帝”的李潇。*让我们通过李潇的视角，了解一下微软亚洲研究院的 Aspire 究竟是一个怎样的组织，以及从“首席实习生”到高级研究员，李潇的心态和研究之路又发生了怎样的变化。

心理学研究表明，在陌生的环境中，人们常因不熟悉当地文化、语言和规则而感到迷茫、焦虑和不适。这种不适应会增加心理和情绪压力，进而影响个人的行为和表现。然而，对于像李潇等新入职的员工来说，他们在加入微软亚洲研究院后却并没有这样的困扰。

“尽管我在正式加入微软亚洲研究院之前有过一段较长的实习经历，但在从事产品开发一段时间后重新回归学术研究，在过渡的过程中，我仍面临一些挑战。借助 Aspire，我‘无痛’地渡过了适应期，重新融入了研究院这个大家庭。”李潇说。

李潇，微软亚洲研究院高级研究员

回归微软亚洲研究院，Aspire让我“无痛”渡过适应期

李潇提到的 Aspire 是微软亚洲研究院为入职三至四年的年轻员工创建的一个内部社群。这个社群为年轻人提供了一个可以相互交流与合作的平台，帮助他们快速适应研究院的科研工作和文化，建立多元化的跨组合作氛围，并通过组织各种项目和活动促进个人成长。

在过去的两年中，李潇作为 Aspire 组织委员会的核心成员之一，与团队的伙伴们一起组织了一系列既富有趣味性又具有挑战性的社群活动，比如研究院年会上的智力游戏、运动会和创意工作坊等等。这些活动既为参与者提供了互动及展现个人才华的机会，也激发了大家的创造力，促进科研想法的交流。

而且，还可以帮助新入职的员工们快速结识志同道合的伙伴。“我们曾组织过面向 Aspire 的冰壶体验和比赛，希望帮助新同事们提升团队合作能力。因为冰壶是一项既有趣，又能够锻炼身体，同时最重要的是对团队协作和战略思维要求极高的运动，所以在比赛中，队员们需要共同制定战术策略，相互支持与沟通。”李潇说。

Aspire 还鼓励新员工大胆表达自己的想法和建议，为微软亚洲研究院带来新鲜的视角和活力，促进不同代际科研人员之间的思想碰撞。在微软亚洲研究院成立25周年的庆祝活动中，李潇和其他 Aspire 组织委员会的成员们共同策划了一场名为“共筑未来城”的活动——借助研究院内部团队建设的机会，让所有员工对未来城市进行一次大胆预想。活动中，每位参与者都为这座微软亚洲研究院的“未来城”贡献了自己的一砖一瓦，共同构建了一个充满创新与梦想的科技乌托邦。

李潇参与组织的 Aspire 社群冰壶活动

此外，Aspire 还为年轻员工搭建了与研究院管理层沟通的桥梁，帮助管理层更快、更好地理解年轻员工的需求和想法，为研究院的发展建言献策。Aspire 会定期举办线上线下沟通会，建立年轻人自己的交流沟通渠道。在 Aspire 组织里，没有新老员工或是上下级的单向教育，年轻的研究员们可以畅所欲言，内容涵盖

年轻群体的兴趣点、当下科研的新趋势、对研究院研究氛围和变化的看法、对学术领域和科研环境的观察等多个方面。同时，为了更好地营造开放、包容的文化，Aspire 还会不定期开展问卷调查，让每一位年轻的 Aspire 成员都能有机会提出自己关注的问题，发出自己的声音。通过 Aspire 这一平台，研究院的管理层能够及时了解年轻员工最真实的需求和遇到的挑战，确保相关问题能够得到快速的响应和解决。

“在我看来，研究院希望通过 Aspire 这个平台，鼓励、激发我们这些年轻人在工作和生活中保持热情，发掘科研工作与日常生活的美好，以更舒适的方式投身于所热爱的事业中。事实上，Aspire 确实拉近了我们与不同领域同事之间的距离，使我们能够自由提出创新的想法和研究需求，也可以直接向管理层反馈建议。参与各种社群活动的过程不仅减少了我们在陌生环境中的迷茫和焦虑，还增强了我们对研究院的归属感与责任感。”李潇说。

从“首席实习生”到联培博士，探索AI在图形学中的应用

虽然2020年才正式成为微软亚洲研究院的研究员，但李潇可是研究院的一位“老人”了。李潇与研究院的缘份始于十几年前，当时还在中国科学技术大学读大四的他，加入了研究院的网络图形组，开始了他的实习之旅。这段经历让李潇有机会与敬仰的学者们近距离交流，也为他积累了宝贵的知识和经验。本科毕业后，李潇凭借出众的能力入选了中国科学技术大学与微软亚洲研究院的联合培养博士生项目，这一决定改变了他原本计划出国深造的人生轨迹。

“追求高水平的技术研究才是我的最终目标，微软亚洲研究院作为一个国际化的顶尖研究机构，为我提供了与世界级科研人才共事的机会以及理想的学习和研究环境。”从那以后，李潇就一直在深度学习、计算机图形学等领域的最前沿进行探索。

2015年 ResNet 的提出，标志着深度学习在计算机视觉领域的重大突破，随后深度学习技术很快便受到了学术界的广泛关注，并逐渐在多个领域开始普及应用。李潇所在的微软亚洲研究院网络图形组便是最早一批将深度学习技术应用于图形学研究的团队。2017年，李潇作为第一作者在图形学领域的全球顶级会议 SIGGRAPH 上发表了论文。该论文是早期探索在图形学领域直接利用深度学习技术识别图像材质的方法之一，给后续研究带来了全新的视角和思路。2019年，李潇和 Mentor 一起就三维生成技术进行了更深入的研究，其中关于使用多投影生成对抗网络、利用无标注图片生成三维物体的论文，被计算机视觉领域的全球顶会 CVPR 接收。

2019年6月，在完成了近六年的学习并从中国科学技术大学博士毕业后，李潇也从微软亚洲研究院“毕业”了。面对人生的新选择，李潇决定先去更接近生产实践的互联网公司，更紧密地接触用户，了解用户的需求，研发更多符合用户需求的产品和技术。

重回微软亚洲研究院，为多模态大模型构建媒体基础

经过在互联网企业的历练，李潇深刻体会到用户需求的多样性和变化的迅速。他意识到，在快速变化的市场环境中，单纯依赖人力和功能堆砌的方式难以持续且高效地满足用户需求。他认为，更通用且普适的创新方法——尤其是更加通用的人工智能模型——是应对这一挑战的关键。这种洞察驱使他回到科研的最前线，重新加入微软亚洲研究院。

“微软亚洲研究院清楚地知道哪些研究领域是有价值的，而且这里的研究不仅仅是为了发表论文，更是为了让前沿技术真正应用到不同的领域，对现实世界产生积极的影响。”李潇相信，通用型人工智能技术将极大提高多媒体内容创作的生产效率，并能激发人们的创作灵感。

回到研究院后，李潇加入了多媒体计算组，将自己在短视频平台研发工作中积累的经验运用到音视频理解与生成的前沿研究中，与组里的同事们共同探索构建多模态模型。他们的研究不仅限于理论层面，更着眼于将研究成果应用到实际产品中。比如，李潇和同事们开发的音频驱动口型重建技术，能够允许用户在视频会议中创建自己的虚拟形象，并通过语音驱动虚拟形象的口型，显著减少交流时的违和感，提升了用户体验。

在微软亚洲研究院全球研究合伙人吕岩的带领下，李潇和同事们在人工智能与多媒体研究的交叉领域也取得了突破，他们利用神经编解码器将视觉、语言和声音等多种信息类型转化为隐空间的神经表达。这种方法能够将复杂而含有噪音的现实世界，转化为能够捕获世界本质信息和动态变化的抽象表示，与纯粹依赖自然语言处理的大模型相比，它提供了一种更为多样和全面的信息理解方式。基于这一思路，李潇和同事们通力合作，验证了利用视频和音频数据构建全新的媒体基础（Media Foundation）的可行性，为构建人工智能的多模态模型开辟了新的研究方向。

李潇（第二排右二）与多媒体计算组的同事合影

“媒体基础与大语言模型一样，将成为构建通用型人工智能的关键组成部分。基于媒体基础的多模态模型能实现对文本、图像、语音、视频的深层次理解，并促进这些不同模态之间的流畅转换与互动。”李潇介绍道。

在多元包容的平台上，创造影响深远的科研成果

自2012年首次以实习生身份加入微软亚洲研究院，结识众多学术界大牛，到经历四年联合培养博士项目与顶尖学者深入交流，再到成为正式研究员参与多媒体计算技术的前沿探索，李潇的每一步成长也伴随着他对微软亚洲研究院认知上的刷新。

如今，微软亚洲研究院汇聚了全球顶尖人才，不仅包括计算机科学领域的专家，还有来自不同领域和学科的多元化人才。“在这里，尽管大家背景各异，但我们一直都处于一个平等的讨论环境。无论是实习生、研究员，还是 Mentor 或者管理者，每个人都可以自由地提出问题、分享观点，不存在因等级或资历而产生的沟通壁垒。Aspire 则通过组织结构有效促进了这种文化的实践，它在帮助新员工快速融入研究院的同时，也促进了不同代际同事间的理解与合作，共同迈向我们的愿景。”李潇说。

相关阅读：

吕岩：媒体基础，打开多模态大模型的新思路

随着AI的快速发展，人们希望AI能够从现实世界的数据中进行学习和迭代。然而如何在复杂且充满噪声的真实世界和AI所处在的抽象语义世界之间架起桥梁？是否可以为不同类型媒体信息构建与自然语言平行的，另一种可被人工智能学习理解的语言？微软亚洲研究院全球研究合伙人吕岩在其署名文章中介绍了，微软亚洲研究院在多媒体与AI融合发展方面的最新进展，及相关领域的未来展望。

扫描二维码查看文章

科学匠人｜罗琳：用真诚赢得信任，用AI助力无障碍沟通

每年的9月23日是国际手语日，每年9月第四个星期日是国际聋人日，其设立目的是为了提高社会对聋人群体的关注与支持以及对手语的认识，保护聋人权利。在微软亚洲研究院，一群致力于用科技创造更包容世界的工程师们正在为此努力，来自视觉计算组的研究工程师罗琳就是其中一员。她和同事们合作开发的针对手语识别与翻译的研究项目，致力于利用AI技术为聋听之间搭建沟通的桥梁。

扫描二维码查看文章

科学匠人｜黄昶互：坚持长期主义研究，是一个不断说服自己的过程

他，刚入职微软亚洲研究院一年，却有着丰富的学术合作经验；刚刚博士毕业，就有多篇论文被业内顶会收录并获奖；能够长期投入到一项研究课题中，并持续跟进三、四年；选定一个研究领域，层层递进，展开了多方面的研究；热衷于科学创新，坚持长期主义的研究理念。他是来自韩国的微软亚洲研究院研究员黄昶互（Changho Hwang）。让我们一同走进他的故事，感受他对研究的热忱与专注。

扫描二维码查看文章

对话 | ACM、IEEE 双 Fellow 谢幸：坚持长期主义研究的秘诀是什么？

日前，全球计算机领域影响力最大的专业学术组织国际计算机学会（Association for Computing Machinery，简称 ACM）公布了2023年度 ACM Fellow 名单。微软亚洲研究院资深首席研究员谢幸博士因其在空间数据挖掘和推荐系统方面的卓越贡献获选其中。学术研究是一个厚积薄发的过程，那么，谢幸博士是如何在漫长的科研道路上保持数十年如一日的研究热情和持续动力的？在科研工作中，他如何能够不断产出具有深远影响的成果？随着人工智能大模型的快速发展对社会造成深刻影响，我们要如何确保这些技术安全、可信、可靠，并且符合人类的价值观？对于希望在学术研究领域深耕的年轻学者们，谢幸博士有哪些经验和建议？为了寻找这些问题的答案，我们与谢幸博士进行了一次深入的对话。

谢幸博士于2001年7月加入微软亚洲研究院，二十多年来，他一直潜心钻研，取得了一系列令人瞩目的成果，并屡获国内外奖项和荣誉。他的学术成就包括但不限于2019年获得 ACM SIGSPATIAL 十年影响力论文奖和中国计算机学会青竹奖，2020年获得 ACM SIGSPATIAL 十年影响力论文荣誉奖，2021年获得 ACM SIGKDD China 时间检验论文奖（Test of Time Award），2022年获得 ACM SIGKDD 时间检验论文奖，2023年获得 IEEE MDM 时间检验论文奖和中国计算机学会自然科学一等奖。

微软亚洲研究院资深首席研究员谢幸

Q1：恭喜您获选 2023 ACM Fellow，对此次获选您有什么感受？这对您的科研工作和职业生涯有怎样的意义？

谢幸：ACM Fellow 代表着全球计算机科学领域内极高的荣誉，能与众多著名和有影响力的前辈学者一同入选，我感到非常的荣幸！回顾20多年前，我作为一名刚获得博士学位的研究员加入微软亚洲研究院，开启了我的学术生涯，至今的旅程充满了令人激动的经历。我感谢在这段旅程中给予我支持的每一个人，也希望未来能帮助更多青年学者实现他们的梦想，正如我在这条路上所获得的帮助一样。期待和大家共同努力，继续推动科学的边界！

Q2：您是从何时开始从事空间数据挖掘和推荐系统方面的研究的？这些研究成果有怎样的科研价值和应用价值？

谢幸：我的研究开始于大约20年前，也就是2005年左右。那时传感器，尤其是 GPS，开始被越来越多地集成到移动设备中，其能力也愈发强大。在注意到这一发展趋势之后，我和团队展开了在时空数据挖掘方向上的研究。通过分析脱敏后的用户位置数据，来理解城市人群的移动、出行习惯和生活节奏等，并利用社交媒体或其他来源的数据分析用户行为，以开发个性化的应用，如推荐系统，从而为人们的工作和生活提供更多便利。

这些研究本质上都是对人类行为的探索，属于"社会计算"范畴，即利用计算技术和方法解决社会学问题。这一领域的研究范围广泛，涉及社会学、心理学、脑科学等多个跨学科方向。最初，我们的研究主要是围绕空间数据来展开，选择将人类行为预测作为研究重点。现在，我们又进一步把研究领域扩展到了社会责任人工智能（Societal AI），确保人工智能技术更安全、更负责任、更符合人类价值观。

Q3：您和您所在团队的研究曾获得2019年 ACM SIGSPATIAL "十年影响力论文奖"，并在2021、2022年、2023年分别获得 ACM SIGKDD China、ACM SIGKDD、IEEE MDM 的"时间检验奖"。能否介绍一下这些研究的内容？

谢幸：2019年获奖的论文 "Map-Matching for Low-Sampling-Rate GPS Trajectories" 提出了以地图匹配的方式，把不连续的 GPS 点映射到具有明确语义的地图位置上，并结合人类运动轨迹的特性，提高用户定位的精度。这是一个计算机科学和空间信息

科学的交叉问题，也是一项由数据和应用驱动的研究。早在2009年，我们就已经开始了这一跨学科研究，算是这个领域的先行者，这也是我们获奖的主要原因。

2021年，我们的论文 "Driving with Knowledge from the Physical World"（《在物理世界知识指导下驾驶》）获得了 ACM SIGKDD China Chapter 时间检验奖。

这篇论文通过对出租车轨迹数据的深入研究，发现了交通模式和驾驶行为的规律，并设计了一种定制化的导航服务，可以根据用户的需求和环境的变化，提供最优的路线建议。

获得 KDD 2022 时间检验奖的论文是发表在 KDD 2012 的 "Discovering Regions of Different Functions in a City Using Human Mobility and POIs"（《基于人类移动规律和兴趣点的城市功能区域发现》）。这篇论文中，我们基于大规模人群移动数据和地图兴趣点数据，提出了一种数据驱动的城市功能分区自动发现方法，可以帮助人们更深入地理解城市的动态变化。

2023年获得时间检验奖的论文是我们2010年发表的 "An interactive-voting based map matching algorithm"（《一种基于交互式投票的地图匹配算法》）。当时为了解决用户 GPS 轨迹语义理解中的低采样率问题，我们研发了一种基于交互式投票的地图匹配算法，大幅提高了地图匹配性能。

这些研究都是以用户轨迹数据社交网络数据和个性化推荐为重点的研究，目的是深入理解人的行为，提升人们的生活质量和满意度。

谢幸和团队获得的时间检验奖证书

Q4：您的很多研究都经受住了时间的考验。您是如何保持对研究的长期热情的？在坚持长期主义研究方面，您有哪些经验和心得？

谢幸：对于如何进行长期主义研究，没有一个确定的答案。对我来说，深入探索一个问题并彻底理解它，带来的满足感是我研究中最大的乐趣。我追求的不仅仅是发表一两篇论文，更重要的是与人交流和讨论的过程。比如，我们会探讨推荐系统的本质是什么，用户行为是否可以被准确预测，如何将推荐系统与知识图谱结合，以及将 Transformer 模型和领域知识与推荐系统结合的可能性。这些讨论虽然不一定直接反映在论文中，但是这种频繁的交流，让我了解了不同人对于领域的理解，这是一个很有趣的过程，也为我提供了持续前进的动力。

另一方面，解决问题需要广泛的知识储备，这不仅要求我们深入理解自己的研究领域，还要紧跟最新的前沿动态。持续的研究还需要投入大量的时间和耐心，不能急于求成，这就需要我们反复实验和实践，验证和改进自己的方法和模型，才能取得好的成果。我想再次强调的是，时间是必不可少的，必须要深入挖掘，这是一个厚积薄发的过程，不能半途而废。

Q5：在选择研究方向时，您有哪些经验可以分享，特别是如何挑选出有价值且能产生更大影响力的研究课题？

谢幸：衡量研究影响力的维度很多，我认为可以从以下几个角度来考虑。首先是定义一个新颖或重要的问题。如果一项研究能够发现一个以前被忽视或者有巨大潜力的问题，就会引起他人的关注，然后再提出一个行之有效的解决方法，这样的研究往往能产生比较大的影响力。这类问题通常存在于不同领域的交叉点，或者是新兴应用场景的早期阶段，这需要我们有敏锐的洞察力和创造力。

另外，着手解决长期存在但尚未解决的难题也是一个好的选择。如果论文能够在一个已有的或者经典的问题上，提出一个比以前更优、更完善的方案，也会引起很大的反响，比如之前长期存在的"量子霸权"问题。但是这类问题的研究往往会有很多的参与者，竞争激烈，很有挑战性，需要研究者具备扎实的理论基础和过硬的技术实力。

再者，紧跟科研热点趋势也是选择研究方向的一个策略。热点就意味着新兴、前沿和快速发展，创新和突破的机会相对较多。比如，六年前我们将热门的推荐系统与知识图谱结合，由此在2018年发表的研究论文《将知识图谱特征学习引入推荐系统的思路与实现方法》也获得了很大的影响。

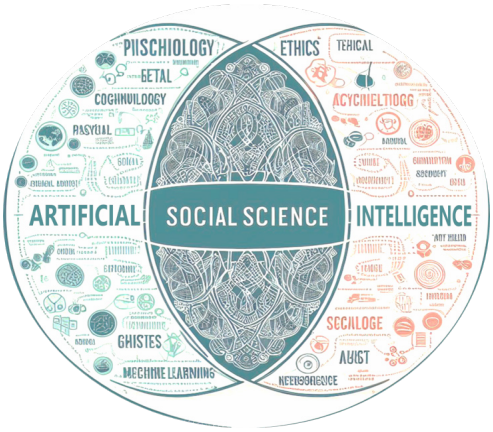
此外，一些基础性研究也很有价值，例如发布一个领域的基础数据集，这类研究虽然没有提出新问题，但却为该领域提供了标准化的问题框架。还有一些实用性的研究，比如计算机视觉领域的图像生成模型的研究，有助于提高人们的创造力。总之，选择有影响力的研究方向没有固定的套路，需要根据自己的兴趣和能力来决定。

Q6：您的研究成果和经验都显示了跨领域研究的重要性和价值。您是如何看待跨领域研究与合作的？

谢幸：微软亚洲研究院成立之初的一个宗旨就是与各大高校合作，推动计算机科学的前沿研究与发展。我们通过各种形式的项目，比如合作研究、访问学者、联合课程以及"明日之星"实习生项目等等，来开展跨领域和跨学科的研究。此外，我们还积极寻求与产业界的合作，与跨领域的专家探讨他们面临的挑战，从而更有的放矢地解决行业痛点。

我目前重点关注的研究方向——社会责任人工智能，更加需要跨学科的支持，特别是与社会科学的融合。社会科学可以为计算机科学提供新的视角和工具，但这对我们而言是一个全新的领域，需要从零开始搭建理论框架与方法。我们也需要引入"双料人才"，因为目前在社会学、法学等社会学科中，能够掌握跨学科研究所需知识的人才还比较匮乏。

当然，如何平衡并有机结合计算机科学与社会科学这两种截然不同的学科思维和研究方式，我们也还在探索中。我们期待与各行各业、各个学科、各个领域的伙伴们携手合作，共同推进跨学科领域的前沿技术创新与应用。



Q7：当前您重点关注的是社会责任人工智能研究，请介绍一下该领域的研究方向及研究重点？

谢幸：当前人工智能的发展一次次地超出人们的预期，尤其是大模型在不断突破以往的限制，它在促进行业发展和提高效率方面的潜力已经得到了广泛认可。然而，在技术发展给人们带来便利的同时，技术本身是否价值观中立以及技术安全性问题也愈发引人关注。为了确保人工智能谨守造福人类的原则，我们近年来一直致力于社会责任人工智能的研究。

这并不是一个全新的方向，微软多年前在进行人工智能研究时，就制定了"负责任的人工智能（Responsible AI）"准则。过去一年中人工智能的爆发式成长，使得社会责任人工智能 成为了面向人工智能未来的、同样具有前瞻性的研究方向。

我们认为构建社会责任人工智能应该包含五个方面，包括价值观对齐、数据及模型安全、正确性或可验证性、模型评测，以及前面提到的跨学科合作。通过与社会学、心理学等领域的交流与合作，我们已经在这些方向上取得了初步进展，并发表了相关的研究成果，例如《大模型道德价值观对齐问题剖析》等。
让人工智能的创造和决策符合全人类的福祉，使其道德和价值观与人类相一致，这是一个长期而艰巨的任务，还需要我们持续探索研究，也需要所有人共同努力，才能让人工智能技术更好地造福人类社会。

关于社会责任人工智能（Societal AI）研究的更多前瞻观点，请点击阅读谢幸博士的署名文章《跨学科合作构建具有社会责任的人工智能》。

Q8：对于有志于科学研究的青年学子，您认为他们应该具备哪些品质？

谢幸：我觉得科学研究需要有三个重要的品质：兴趣、能力和自驱力。兴趣是科学研究的动力源泉，如果研究者对自己所研究的领域或者问题没有兴趣，就很难坚持长期探索，也就无法做出深入的研究。能力是科学研究的基础，这需要长期培养自己的基础研究能力和积累研究领域的专业知识，因为只有兴趣而没有能力，可能只能做出非常肤浅的研究。

另外一个非常重要的品质是自我驱动力。在博士阶段，我们经常会遇到寻找科题、发表论文等方面的困难，这时候自我驱动力，保持科研热情，坚持不懈的精神就会发挥非常重要的作用，这也是我们考察博士生时最看重的品质。

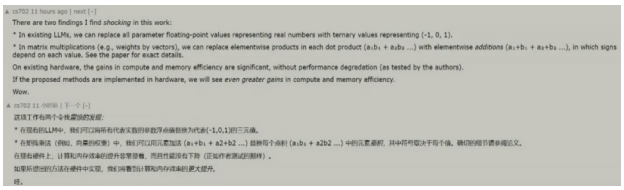
Q9：您在微软亚洲研究院工作了20多年，研究院给您的科研生涯带来了哪些帮助和影响？

谢幸：微软亚洲研究院是一个国际化的研究平台。在这里，我们不仅有机会进行前沿的基础研究，还能与各领域的杰出科学家合作。这种开放、自由和包容的研究氛围极大地拓宽了我的视野，锻炼了我的学术研究和沟通合作能力。研究院丰富的资源和机遇为我们的成长和进步提供了助力。

此外，研究院还是一个连接学术研究与工业应用的平台。我们与工业界、学术界，以及微软产品团队都有着紧密的合作，这让我们可以更加便捷地参与到既有实际需求又富有应用价值的研究中。紧跟工业界的发展趋势，也使我们能够进行更具创新性和挑战性的研究。

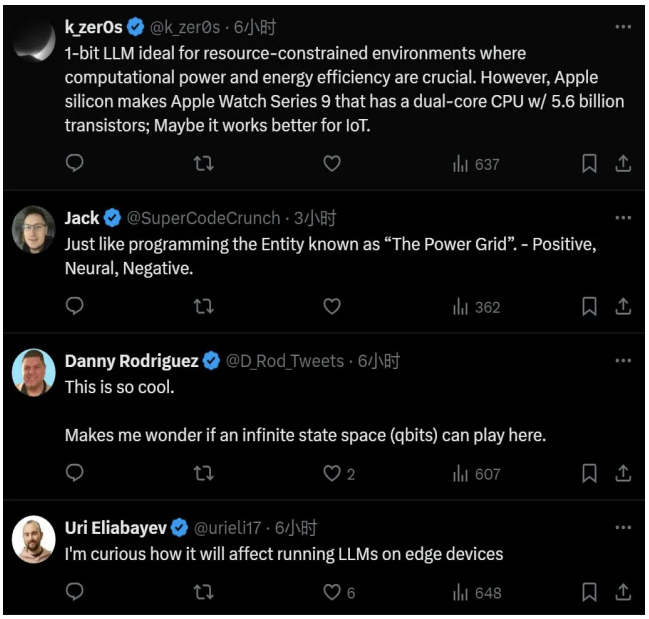
机器之心 | BitNet b1.58：开启 1-bit 大语言模型时代

近期，由微软亚洲研究院、中国科学院大学（国科大）等机构提交的一篇论文在 AI 圈里被人们争相转阅。该研究提出了一种 1-bit 大模型，实现效果让人只想说两个字：震惊。



如果该论文的方法可以广泛使用，这可能是生成式 AI 的新时代。

对此，已经有人在畅想 1-bit 大模型的适用场景，看起来很适合物联网，这在以前是不可想象的。



人们还发现，这个提升速度不是线性的 —— 而是，模型越大，这么做带来的提升就越大。

近年来，大语言模型（LLM）的参数规模和能力快速增长，既在广泛的自然语言处理任务中表现出了卓越的性能，也为部署带来了挑战，并引发人们担忧高能耗会对环境和经济造成影响。

因此，使用后训练（post-training）量化技术来创建低 bit 推理模型成为上述问题的解决方案。这类技术可以降低权重和激活函数的精度，显著降低 LLM 的内存和计算需求。目前的发展趋势是从 16 bits 转向更低的 bit，比如 4 bits。然而，虽然这类量化技术在 LLM 中广泛使用，但并不是最优的。

最近的工作提出了 1-bit 模型架构，比如 2023 年 10 月微软研究院、国科大和清华大学的研究者推出了 BitNet，在降低 LLM 成本的同时为保持模型性能提供了一个很有希望的技术方向。

BitNet 是第一个支持训练 1-bit 大语言模型的新型网络结构，具有强大的可扩展性和稳定性，能够显著减少大语言模型的训练和推理成本。与最先进的 8-bit 量化方法和全精度 Transformer 基线相比，BitNet 在大幅降低内存占用和计算能耗的同时，表现出了极具竞争力的性能。

此外，BitNet 拥有与全精度 Transformer 相似的扩展法则（Scaling Law），在保持效率和性能优势的同时，还可以更加高效地将其能力扩展到更大的语言模型上， 从而让 1 比特大语言模型（1-bit LLM）成为可能。

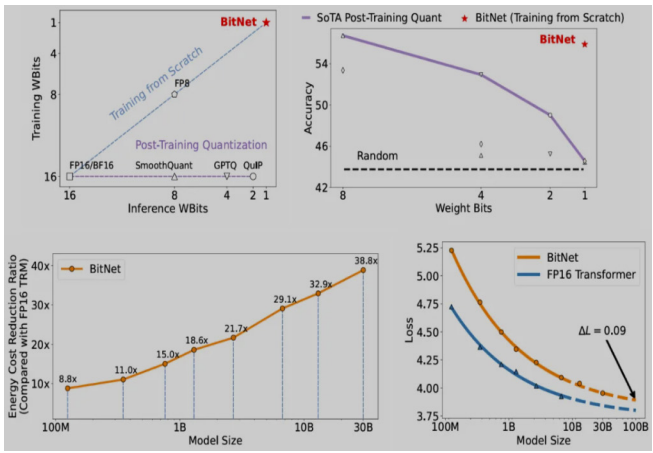


图 1：BitNet 从头训练的 1-bit Transformers 在能效方面取得了有竞争力的结果

如今，微软亚洲研究院、国科大同一团队（作者部分变化）的研究者推出了 BitNet 的重要 1-bit 变体，即 BitNet b1.58，其中每个参数都是三元并取值为 {-1, 0, 1}。他们在原来的 1-bit 上添加了

一个附加值 0，得到二进制系统中的 1.58 bits。

BitNet b1.58 继承了原始 1-bit BitNet 的所有优点，包括新的计算范式，使得矩阵乘法几乎不需要乘法运算，并可进行高度优化。同时，BitNet b1.58 具有与原始 1-bit BitNet 相同的能耗，相较于 FP16 LLM 基线在内存消耗、吞吐量和延迟方面更加高效。

此外，BitNet b1.58 还具有两个额外优势。其一是建模能力更强，这是由于它明确支持了特征过滤，在模型权重中包含了 0 值，显著提升了 1-bit LLM 的性能。其二实验结果表明，当使用相同配置（比如模型大小、训练 token 数）时，从 3B 参数规模开始， BitNet b1.58 在困惑度和最终任务的性能方面媲美全精度（FP16）基线方法。

如下图所示，BitNet b1.58 为降低 LLM 推理成本（延迟、吞吐量和能耗）并保持模型性能提供了一个帕累托（Pareto）解决方案。

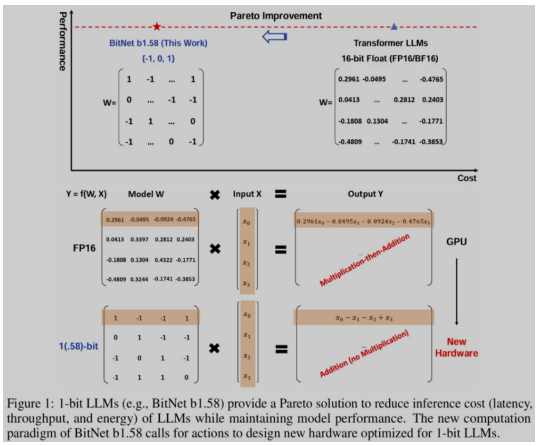


Figure 1: 1-bit LLMs (e.g., BitNet b1.58) provide a Pareto solution to reduce inference cost (latency, throughput, and energy) of LLMs while maintaining model performance. The new computation paradigm of BitNet b1.58 calls for actions to design new hardware optimized for 1-bit LLMs.

BitNet b1.58介绍

BitNet b1.58 基于 BitNet 架构，并且用 BitLinear 替代 nn.Linear 的 Transformer。BitNet b1.58 是从头开始训练的，具有 1.58 bit 权重和 8 bit 激活。与原始 BitNet 架构相比，它引入了一些修改，总结为如下：

用于激活的量化函数与 BitNet 中的实现相同，只是该研究没有将非线性函数之前的激活缩放到 [0, Q_b] 范围。相反，每个 token 的激活范围为 [−Q_b, Q_b]，从而消除零点量化。这样做对于实现和系统级优化更加方便和简单，同时对实验中的性能产生的影响可以忽略不计。

与 LLaMA 类似的组件。LLaMA 架构已成为开源大语言模型的基本标准。为了拥抱开源社区，该研究设计的 BitNet b1.58 采用了类似 LLaMA 的组件。具体来说，它使用了 RMSNorm、SwiGLU、旋转嵌入，并且移除了所有偏置。通过这种方式，BitNet b1.58 可以很容易的集成到流行的开源软件中（例如，Huggingface、vLLM 和 llama.cpp2）。

$$\widetilde{W} = \text{RoundClip}(\frac{W}{\gamma+\epsilon}, -1, 1), \tag{1}$$

$$\text{RoundClip}(x, a, b) = \max(a, \min(b, \text{round}(x))), \tag{2}$$

$$\gamma = \frac{1}{nm}\sum_{ij}|W_{ij}|. \tag{3}$$

实验及结果

该研究将 BitNet b1.58 与此前该研究重现的各种大小的 FP16 LLaMA LLM 进行了比较，并评估了模型在一系列语言任务上的零样本性能。除此之外，实验还比较了 LLaMA LLM 和 BitNet b1.58 运行时的 GPU 内存消耗和延迟。

下表总结了 BitNet b1.58 和 LLaMA LLM 的困惑度和成本：在困惑度方面，当模型大小为 3B 时，BitNet b1.58 开始与全精度 LLaMA LLM 匹配，同时速度提高了 2.71 倍，使用的 GPU 内存减少了 3.55 倍。特别是，当模型大小为 3.9B 时，BitNet b1.58 的速度是 LLaMA LLM 3B 的 2.4 倍，消耗的内存减少了 3.32 倍，但性能显著优于 LLaMA LLM 3B。

Models	Size	Memory (GB)↓	Latency (ms)↓	PPL↓
LLaMA LLM	700M	2.08 (1.00x)	1.18 (1.00x)	12.33
BitNet b1.58	700M	0.80 (2.60x)	0.96 (1.23x)	12.87
LLaMA LLM	1.3B	3.34 (1.00x)	1.62 (1.00x)	11.25
BitNet b1.58	1.3B	1.14 (2.93x)	0.97 (1.67x)	11.29
LLaMA LLM	3B	7.89 (1.00x)	5.07 (1.00x)	10.04
BitNet b1.58	3B	**2.22 (3.55x)**	**1.87 (2.71x)**	**9.91**
BitNet b1.58	3.9B	**2.38 (3.32x)**	**2.11 (2.40x)**	**9.62**

Table 1: Perplexity as well as the cost of BitNet b1.58 and LLaMA LLM.

下表结果表明，随着模型尺寸的增加，BitNet b1.58 和 LLaMA LLM 之间的性能差距缩小。更重要的是，BitNet b1.58 可以匹配从 3B 大小开始的全精度基线的性能。与困惑度观察类似，最终任务（end-task）结果表明 BitNet b1.58 3.9B 优于 LLaMA LLM 3B，具有更低的内存和延迟成本。

Models	Size	ARCe	ARCc	HS	BQ	OQ	PQ	WGe	Avg.
LLaMA LLM	700M	54.7	23.0	37.0	60.0	20.2	68.9	54.8	45.5
BitNet b1.58	700M	51.8	21.4	35.1	58.2	20.0	68.1	55.2	44.3
LLaMA LLM	1.3B	56.9	23.5	38.5	59.1	21.6	70.0	53.9	46.2
BitNet b1.58	1.3B	54.9	24.2	37.7	56.7	19.6	68.8	55.8	45.4
LLaMA LLM	3B	62.1	25.6	43.3	61.8	24.6	72.1	58.2	49.7
BitNet b1.58	3B	**61.4**	**28.3**	**42.9**	**61.5**	**26.6**	**71.5**	**59.3**	**50.2**
BitNet b1.58	3.9B	**64.2**	**28.7**	**44.2**	**63.5**	**24.2**	**73.2**	**60.5**	**51.2**

Table 2: Zero-shot accuracy of BitNet b1.58 and LLaMA LLM on the end tasks.

内存和延迟：该研究进一步将模型大小扩展到 7B、13B 和 70B 并评估成本。图 2 显示了延迟和内存的趋势，随着模型大小的增加，增长速度（speed-up）也在增加。特别是，BitNet b1.58 70B 比 LLaMA LLM 基线快 4.1 倍。这是因为 nn.Linear 的时间成本随着模型大小的增加而增加，内存消耗同样遵循类似的趋势。延迟和内存都是用 2 位核测量的，因此仍有优化空间以进一步降低成本。

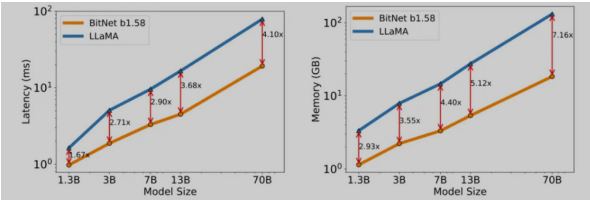


Figure 2: Decoding latency (Left) and memory consumption (Right) of BitNet b1.58 varying the model size.

能耗。该研究还对 BitNet b1.58 和 LLaMA LLM 的算术运算能耗进行了评估，主要关注矩阵乘法。图 3 说明了能耗成本的构成。BitNet b1.58 的大部分是 INT8 加法计算，而 LLaMA LLM 则由 FP16 加法和 FP16 乘法组成。根据 [Hor14，ZZL22] 中的能量模型，BitNet b1.58 在 7nm 芯片上的矩阵乘法运算能耗节省了 71.4 倍。

该研究进一步报告了能够处理 512 个 token 模型的端到端能耗成本。结果表明，随着模型规模的扩大，与 FP16 LLaMA LLM 基线相比，BitNet b1.58 在能耗方面变得越来越高效。这是因为 nn.Linear 的百分比随着模型大小的增加而增长，而对于较大的模型，其他组件的成本较小。

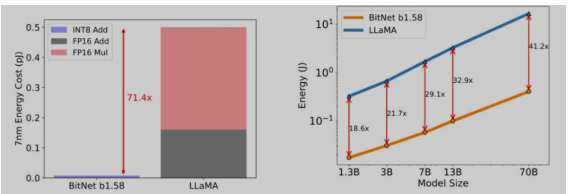


Figure 3: Energy consumption of BitNet b1.58 compared to LLaMA LLM at 7nm process nodes. On the left is the components of arithmetic operations energy. On the right is the end-to-end energy cost across different model sizes.

吞吐量。该研究比较了 BitNet b1.58 和 LLaMA LLM 在 70B 参数体量上在两个 80GB A100 卡上的吞吐量，使用 pipeline 并行性 [HCB+19]，以便 LLaMA LLM 70B 可以在设备上运行。实验增加了 batch size，直到达到 GPU 内存限制，序列长度为 512。表 3 显示 BitNet b1.58 70B 最多可以支持 LLaMA LLM batch size 的 11 倍，从而将吞吐量提高 8.9 倍。

Models	Tokens	Winogrande	PIQA	SciQ	LAMBADA	ARC-easy	Avg.
StableLM-3B	2T	64.56	76.93	90.75	66.09	67.78	73.22
BitNet b1.58 3B	2T	**66.37**	**78.40**	**91.20**	**67.63**	**68.12**	**74.34**

Table 4: Comparison of BitNet b1.58 with StableLM-3B with 2T tokens.

相关链接：

The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits
https://arxiv.org/pdf/2402.17764.pdf

相关阅读：

韦福如：人工智能基础创新的第二增长曲线

在人工智能快速发展的今天，如何突破现有的技术瓶颈，实现跨越式增长？微软亚洲研究院全球研究合伙人韦福如在其署名文章中，分享了微软亚洲研究院在推进人工智能基础创新"第二增长曲线"方面所作出的努力。他表示，"我们希望直击人工智能第一性原理，通过革新基础网络架构和学习范式，构建能够实现效率与性能十倍、百倍提升，且具备更强涌现能力的人工智能基础模型，为人工智能的未来发展奠定基础。"

扫描二维码查看文章

微软亚洲研究院提出全新大模型基础架构RetNet，或将成为Transformer有力继承者！

Transformer 在大语言模型中的重要性毋庸置疑。但 Transformer 也并非完美无缺，其并行处理机制是以低效推理为代价的，且序列越长占用的内存越多。而微软亚洲研究院提出的新型神经网络架构——RetNet 在 Transformer 的基础上，使用多尺度保持（retention）机制替代了标准的自注意力机制，让 RetNet 可以同时实现良好的扩展结果、并行训练、低成本部署和高效推理。这些特性将使 RetNet 有可能成为继 Transformer 之后大语言模型基础网络架构的有力继承者。

扫描二维码查看文章

络绎科学 | 微软亚洲研究院开展视觉内容生成研究，助力解决多模态生成式 AI 核心难题

智能计算是一个非常广阔的概念，涵盖基础理论、软硬件、通用人工智能模型、脑科学、人机协作，以及 AI 在科学、文学、社会学、经济学中应用等一系列问题。

今天，AI 技术的发展已经显露出人类历史上的巨大机遇。通过不断设计、整合与优化软硬件架构和性能、不断获取高质量多模态异构大数据，AI 基础模型的表示、理解和生成能力、小样本学习能力、终身学习能力和推理能力也在不断变强。

"智能计算有望在未来几年取得更加突破性的进展，在诸多领域开创全新的应用场景，并从根本上改变人们的工作和生活。"微软亚洲研究院自然语言计算组资深首席研究员段楠说。

他的主要研究方向为多语言多模态预训练基础模型、多模态生成式人工智能、代码智能和机器推理等。

多年来，他带领团队与微软内部多个产品部门进行长期深入的产研结合合作，所开发技术成功转化到必应搜索/广告/新闻、微软小娜、Visual Studio/VSCode、Azure 云服务等产品，为全球用户提供多样化 AI 服务。

段楠实现的主要成果包括：主导研发了多语言预训练语言模型 Unicoder，实现单一预训练语言模型对 100 种人类语言的覆盖；多模态预训练模型 Unicoder-VL，以及全球首个多语言多模态预训练模型 M3P；代码预训练模型 CodeBERT 及其后续版本 GraphCodeBERT 和 UniXcoder，构建代码智能领域基准测试集 CodeXGLUE，引领预训练技术在软件工程领域的快速发展等。

截止目前，他已发表学术论文 100 余篇，Google Scholar 被引用次数超过 10000 次，持有专利 20 余项。凭借构建多语言多模态预训练基础模型，探索基于基础模型的复杂任务推理和任务完成机制，推动通用型人工智能技术的发展，段楠成为 DeepTech 2022 年"中国智能计算科技创新人物"入选者之一。

从算法层面助力解决多模态生成式AI核心难题

最近三年，段楠带领团队开展视觉内容生成研究，从算法的层面解决了多模态生成式人工智能中的一些核心问题。

他和团队主导实现了业界首个开放域视觉内容生成预训练模型 NUWA（女娲）[1]及其后续版本NUWA-Infinity（任意分辨率图像和视频生成）[2]、NUWA-XL（超长视频生成）[3]和DragNUWA（可控视频生成）[4]，引领了人工智能在高清、超长和可控视觉内容生成场景下的创新和落地。

NUWA-Infinity可生成任意大小的高清图片或长时间视频。该技术将自然语言处理中的"变长"生成算法与视觉内容生成场景进行结合，从而能够生成用户指定大小的高清图片和视频内容。NUWA-XL 是针对超长视频生成的核心问题提出的解决方案。该算法可以在保障关键帧场景切换明显的同时，又能够让整个视频内容保持丰富、连贯，并且可以生成较长时间的视频内容。

在以往的视频生成系统或模型中，一般仅可生成几秒钟或十几秒种以内的视频，而 NUWA-XL 实现了用 16 句话的简短描述生成 11 分钟动画片，充分展示了该模型对超长视频生成的能力。

最近，段楠与团队提出了新型视频生成模型 DragNUWA，从可控性角度提出视频生成新方案。实际上，从相对比较离散的文字空间到相对比较具体的视觉内容，映射的不是"一对一"，而是"一对多"的关系。也就是说，一句话可以用成千上万种不同的视觉内容来表达。因此，如何让视频生成的内容更加可控是该研究的重点。

与以往研究相比，该研究引入了用户输入的轨迹信息。用户不仅可以输入一句话来生成视频，还可以输入一张静态图片，并在图片上标记希望运动的目标方向和力度、运动角度和轨迹等。

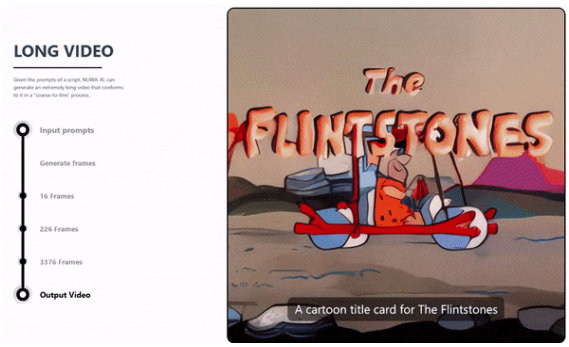


图 1：NUWA-XL 长视频生成

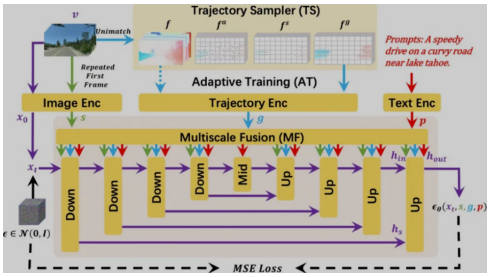


图 2：DragNUWA 的训练过程概述

段楠表示："通过从语义、空间和时间这三个角度提供约束信息，DragNUWA 极大地增强了视频生成模型的可控性。"

多模态生成式人工智能不仅具备巨大的商业价值，还是构建通用型人工智能不可或缺的重要组成部分。当前的人工智能系统主要利用大语言模型进行规划和推理，但人类在做决策时不仅仅会思考一段文字，还会去想象相关的视觉场景。因此，段楠团队对生成算法模型研究的最主要目标，是希望通过设计多模态生成模型帮助人们做更好的规划和推理，以在未来辅助 AI 智能做决策。

"未来，我们希望基于 NUWA 系列工作中沉淀出来的视频生成技术构建通用世界模型（General World Model）。这样就从多模态角度将大语言模型和视频生成模型相结合，从而可以更好地完成任务规划、推理和决策。"段楠说。

从首名联合培养博士到微软亚洲研究院资深首席研究员

2007 年夏天是段楠的硕士毕业季，他从硕士期间就开始在前微软亚洲研究院副院长周明（现澜舟科技创始人兼 CEO）负责的自然语言计算组实习。也正是那时，周明开始在天津大学招收博士研究生。

基于对研究方向的热爱和实习期间优异的工作表现，段楠成为了微软亚洲研究院-天津大学的首名联合培养博士，以统计机器翻译为主要研究方向。彼时，与计算机视觉、计算机图形学等方向的发展相比，自然语言处理还处于相对发展平缓的阶段。

2012 年，段楠在博士毕业后正式加入了微软亚洲研究院，研究方向以问答和自然语言理解为主。段楠与团队早期在多语言模型方面进行系列研究，并开发了微软内部首个多语言预训练模型 Unicoder[5]，该技术此后落地到微软必应搜索引擎，支持全世界 100 种语言和 200 个地区的搜索和问答。2021 年 4 月，Unicoder 模型登上 XTREME 多语言自然语言理解基准测试排行榜。

当时，人们总是想找到一种无歧义的中间形式化语言来标注数据，这样机器可以非常方便地理解自然语言。但这充满挑战，怎么样才能够很好地获得大量的结构化数据，来更好地理解自然语言呢？

图 3：段楠团队部分成员合影

在 Visual Studio/VSCode 等微软产品陆续开发后，2019 年左右，他们意识到，编程语言就是一种无歧义的语言，用它作为语言机器将非常容易理解。

随着语料的累积，大语言模型（Large Language Model，LLM）对自然语言具有非常丰富的理解能力。段楠与团队带着对语义分析挑战的问题，直接开展了后续一系列代码智能的研究，包括：业界首个代码预训练模型 CodeBERT[6]及其后续版本 GraphCodeBERT[7]和 UniXcoder[8]，构建代码智能领域基准测试集 CodeXGLUE[9]等。后续 OpenAI 的 CodeX、ChatGPT 和 GPT-4 更是直接验证了代码预训练对构建 LLM 的重要性。

回顾多年来的研究方向和研究课题，段楠团队好像总是在特殊的时间节点上"提早准备"，对此段楠表示："作为团队负责人，必须对领域发展有系统的、明确的认识，再根据技术的发展，抓住时机及时调整研究方向。"

他同时指出，现在随着 ChatGPT 的出现，领域的选题也越来越难。研究者不仅要考虑当下能解决的问题，也要更长远地去解决那些根本性的问题。

将持续推动通用型人工智能技术的发展

当前，AI 技术大致分为四个主要研究方向，即 LLM、多模态 LLM、多模态生成式人工智能以及 AI 智能体。

简单来理解这四个方向，LLM 构造类似于大脑的功能；多模态 LLM 是对声音、图像等外界环境具有感知能力；生成式人工智能是让 AI 模型具备听、说、读、写的能力；AI 智能体则是让整个 AI 系统能够调用甚至制造工具，并利用这些工具完成数字世界或物理世界各种类型的任务。

段楠认为，未来与 AI 本身发展同样重要的是，需要构建负责任的 AI。也就是说，要保证 AI 模型与人类共生、共创，让 AI 更好地辅助人类去提升生产力及创造力。

他指出，目前 AI 模型在没有见过的、复杂的问题上做得还不够好。以最新的一些 AI 模型为例，一旦它们遇到新出现的、相对复杂的数学类或算法类问题，相关生成质量明显下降。

因此，增强模型的推理能力是非常重要的方向之一，将让它们在未遇到过的问题上具备泛化能力。这样，即便在没有数据驱动的研究方向，也能够通过 AI 模型进行推理。

段楠表示，在感知层构建多语言、多模态、多领域统一预训练模型，在认知层探索具备小样本学习和终身学习能力的可解释复杂推理机制，这两方面研究能够从总体上推动通用型人工智能技术的发展，并驱动其在创新性应用产品中的落地。

相关链接：

[1] NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion
https://arxiv.org/abs/2111.12417

[2] NUWA-Infinity: Autoregressive over Autoregressive Generation for InÞnite Visual Synthesis
https://arxiv.org/abs/2207.09814

[3] NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation
https://arxiv.org/abs/2303.12346

[4] DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory
https://arxiv.org/abs/2308.08089

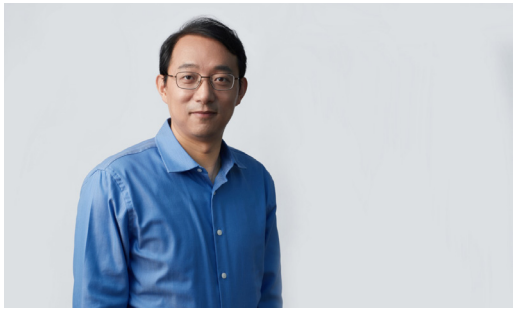
[5] Unicoder: A Universal Language Encoder by Pre-training with Multiple Cross-lingual Tasks
https://arxiv.org/abs/1909.00964

[6] CodeBERT: A Pre-Trained Model for Programming and Natural Languages
https://arxiv.org/abs/2002.08155

[7] GraphCodeBERT: Pre-training Code Representations with Data Flow
https://arxiv.org/abs/2009.08366

[8] UniXcoder: Unified Cross-Modal Pre-training for Code Representation
https://arxiv.org/abs/2203.03850

[9] CodeXGLUE: A Machine Learning Benchmark Dataset for Code Understanding and Generation
https://arxiv.org/abs/2102.04664



周礼栋

微软全球资深副总裁
微软亚太研发集团首席科学家
微软亚洲研究院院长

如今，我们正处在孕育新一代计算范式的关键节点。在不久的将来，虚拟世界和现实世界的边界会不断消弭，计算会像电力一样无处不在。新的计算范式将赋能人类生活和工作的方方面面，给各行各业带来颠覆性的变革，也将催生众多新的机遇。

面对科技发展的新浪潮，微软亚洲研究院将践行所有利于激发创新力的原则，持续致力于营造多元、包容、自由、平等、开放、可持续的研究氛围和科研协作环境，让各种具有创造性的想法、观点和创意，在微软亚洲研究院这个"化学反应池"中交流、碰撞、提炼和升华，使创新的星星之火形成燎原之势。同时，我们也将保持积极开放的姿态，与国内外各界伙伴携手，共同推动技术进步，实现人类社会的可持续发展。

关于微软亚洲研究院

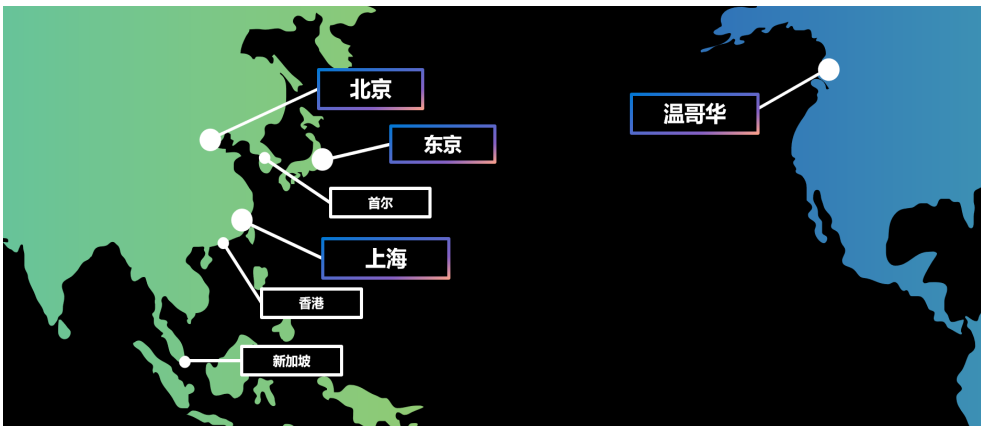
微软亚洲研究院成立于1998年，是微软公司在亚太地区设立的研究机构，在北京、上海、温哥华、东京、首尔、新加坡和香港设有实验室及研究岗位，研究方向涵盖计算基础创新、下一代智能交互、多维感知与通信、人工智能与社会福祉、科学发现与行业赋能等。通过来自世界各地不同学科和背景的多元人才的鼎力合作，微软亚洲研究院已经发展成为世界一流的计算机基础及应用研究机构。多年来，从微软亚洲研究院诞生的新技术层出不穷，对微软公司的产品创新以及全球范围的科技发展产生了深远的影响。

作为微软研究院全球体系的一员，微软亚洲研究院拥有广阔的国际视野，同时融合了东西方创新文化的精髓。秉持开放合作的理念，微软亚洲研究院始终与高校和科研机构开展持久而有效的合作，推动跨地区、跨文化和跨学科的交流，激发创新潜力，促进行业发展。

微软亚洲研究院倡导对技术进步怀有远大抱负，推崇富于冒险的极客创新精神，鼓励研究人员拓展研究的深度与广度，跨越计算机领域的界限，把视野拓展到解决具有广泛社会意义的问题上，为未来的计算新范式奠定基础，并为AI和人类发展创造更美好的未来。

扫描二维码观看视频介绍

微软亚洲研究院实验室分布



微信

知乎

电话：86-10-59178888
网址：http://www.msra.cn/
微博：http://t.sina.com.cn/msra