微软亚洲研究院院刊

Matrix

NO.66
2023年 7 - 9 月

微软亚洲研究院推出AI编译器界“工业重金属四部曲”

科学匠人 | 边江：在研究院的七年“技痒”，探寻大模型助力AI与产业融合之道

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！

01 焦点

02 前沿求索

科研第一线

03 文化故事

04 媒体报道

微软亚洲研究院推出 AI 编译器界“工业重金属四部曲”

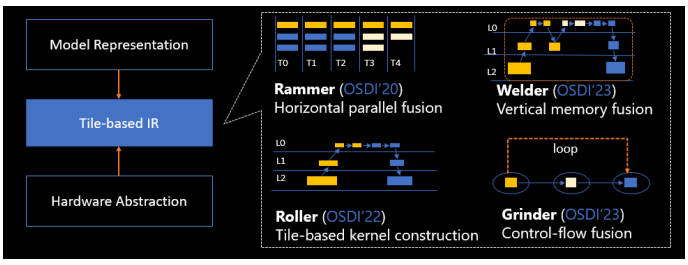
编译器在传统计算科学中一直是一个重要的研究课题。在人工智能技术快速发展和广泛应用的今天，人工智能模型需要部署在多样化的计算机硬件架构上。同时，训练和部署大型人工智能模型时又对硬件性能有着更高的要求，有时还需根据硬件定制化代码。这些都对人工智能时代的编译器提出了新的更高的要求。

为了适应迅速发展的人工智能模型和加速硬件的需求，微软亚洲研究院以设计和构建具有高度灵活性、高效性、可扩展的 AI 编译器架构为目标，与海内外合作者展开研究并提出了一套包含 Rammer、Roller、Welder、Grinder 四款 AI 编译器的系统性解决方案，将提升硬件并行利用率、提高编译效率、优化全局访存效率、优化控制流的高效执行等几大难题通通搞定。四篇相关论文已先后被 2020 年、2022 年、2023 年的 OSDI 大会接收。

编译是程序开发的一个重要步骤 —— 把用高级语言书写的源代码翻译成在计算机硬件可执行的机器码，而编译器就是实现这一功能的特殊应用程序。如今，人工智能技术和大模型无疑是当今计算机领域的 C 位担当，其自身的特性也对编译器提出了新的挑战。

从最初的 RNN、CNN 到 Transformer，人工智能的主流模型架构在不断变化，这意味着上层的应用程序也在随之改变。同时，底层加速器硬件如 GPU、NPU 等，也在快速迭代更新，有些新的硬件设计甚至颠覆了之前的架构。那么，要想让新的人工智能模型更好地运行在新的芯片等计算机硬件上，就要全新的 AI 编译器。

对此，微软亚洲研究院的研究员们和国内外合作者围绕着 AI 编译器的核心问题展开了一系列研究工作，并陆续推出了 AI 编译界的“工业重金属四部曲”：Rammer、Roller、Welder、Grinder，为当前主流的人工智能模型和硬件编译提供了系统性的创新解决方法。



基于统一切块（tile）抽象的四个核心 AI 编译技术

AI 编译“夯土机”Rammer：提升硬件并行利用率

深度神经网络（DNN）是当前图像分类、自然语言处理和许多其他人工智能任务中广泛采用的方法。由于其重要性，许多计算设备，如 CPU、GPU、FPGA 和专门设计的 DNN 加速器被用来执行 DNN 计算。其中影响 DNN 计算效率的关键因素之一是调度，即决定在目标硬件上执行各种计算任务的顺序。现有的 DNN 框架和编译器通常将数据流图（data flow graph）中的 DNN 算子视为不透明的库函数，并将它们调度到加速器上单独执行。同时，这一过程还依赖于另一层调度器（通常在硬件中实现）来利用算子中可用的并行特性。这样的两层方法就导致了显著的调度开销，并且通常不能充分利用可用的硬件资源。

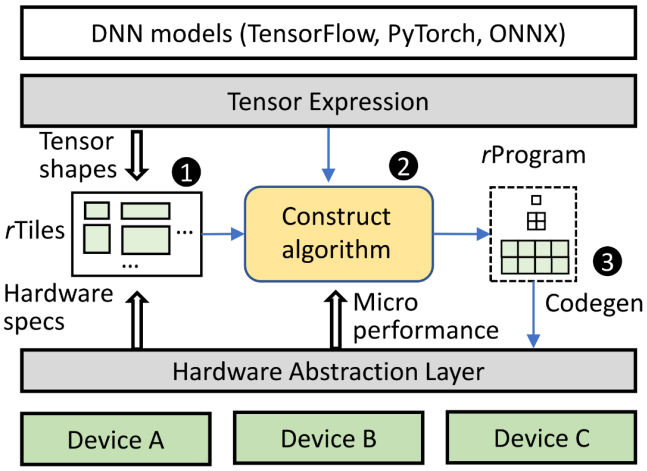
为此，研究员们提出了一种新的 DNN 编译器 Rammer，它可以优化 DNN 工作负载在大规模并行加速器上的执行。事实上，我们可以将进行 AI 编译时的调度空间想象成一个二维空间，并将计算任务看作是可以被拆分成不同大小和形状的“砖块”，调度的目的就是在二维空间的计算单元上将这些“砖块”像垒墙一样紧密排列起来，最大程度地利用计算单元不留空隙，因为一旦出现空隙，不仅已有空间得不到有效利用，而且还会影响“垒墙”的速度。而 Rammer 就是这个二维空间中的一台“夯土机”，在将 DNN 程序翻译成“砖块”后，可放置在芯片的不同计算单元上，将其压实。

换言之，Rammer 在编译时为 DNN 生成了有效的静态时空调度，最大限度地减少了调度开销。同时，通过为计算任务和硬件加速器提出的几个新的、与硬件无关的抽象，使 Rammer 获得了更丰富的调度空间，实现了算子间和算子内的协同调度，从而可以全面利用并行性。这些新颖且具有启发式的方法，让 Rammer 可以更好地探索空间并找到有效的调度，大幅提高硬件利用率。

研究员们在 NVIDIA GPU、AMD GPU 和 Graphcore IPU 等多个硬件后端对 Rammer 进行了测试。实验表明，Rammer 在 NVIDIA 和 AMD GPU 上的性能显著优于 XLA 和 TVM 等最先进的编译器，加速比高达 20.1 倍。与 NVIDIA 的专有 DNN 推理库 TensorRT 相比，加速比达 3.1 倍。

AI 编译“压路机”Roller：提高编译效率

在计算机芯片上不仅有并行计算单元，还有多层内存，一个大的计算任务需要一层层地向上传递，并在这个过程中将任务逐层切分成更小的“砖块”，最终交给最上层的处理器进行计算。这其中的难点在于如何把大的“砖块”铺满内存空间，进而更好地利用内存并提升效率。目前已有的方法是通过机器学习进行搜索，寻找更好的“砖块”切分策略，但这通常需要数千个搜索步骤，每个步骤都要在加速器中进行评估，以找到合理的解决方案，所以这会花费大量的时间，如编译一个完整模型甚至需要几天或几周。



Roller 技术框架

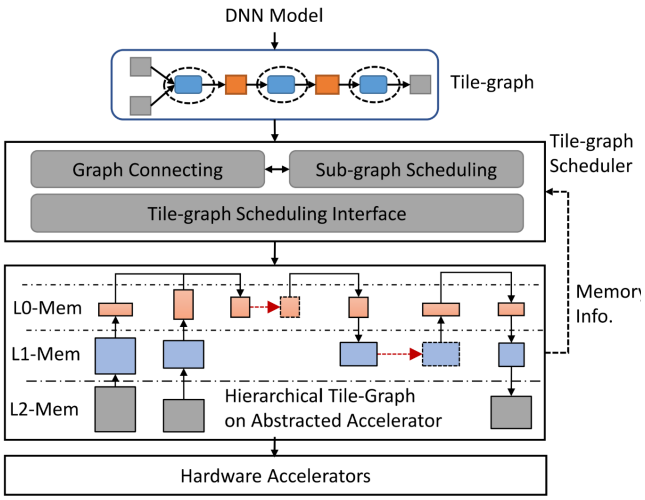
研究员们认为，在了解了计算逻辑和各内存层的参数，也就是在已知软件和硬件信息的情况下，其实完全可以估算出“砖块”切割的最佳方法和大小，从而实现更快的编译。这也是 Roller 的设计思路，它相当于一台压路机，在考虑内存特性的前提下，像铺地板一样把高维的张量数据平铺到二维的内存中，找到最优的切块（tile）大小。同时，它还封装了与底层加速器的硬件特性一致的张量形状，通过限制形状选择来实现高效编译。

通过对 6 种主流 DNN 模型和 119 种流行的 DNN 算子的评估表明，Roller 可以在几秒内生成高度优化的内核，尤其是对于大型昂贵的自定义算子。最终，Roller 在编译时间上比现有的编译器实现了三个数量级的改进。Roller 生成的内核性能与包括 DNN 库在内的最先进的张量编译器的性能相当，有些算子甚至表现更好。与此同时，Roller 也已被用于微软内部开发的自定义 DNN 内核上，在实际开发中验证了 Roller 可以显著加快开发周期的优越性能。

AI 编译“电焊机”Welder：降低访存量，提升计算效率

现代 DNN 模型对高速内存的要求变得越来越高，在分析了一些最新的 DNN 模型后，研究员们发现当前大部分 DNN 计算的瓶颈主要在于 GPU 的访存，如这些模型对内存带宽利用率高达 96.7%，但计算核的平均利用率只有 51.6%，而且随着硬件与 DNN 模型的不断发展，这两者之间的差距还会持续增大。尤其是当前的人工智能模型需要处理高保真度的数据，如更大的图像、更长的句子、更高清的图形，这些数据在计算中都占用了更多的内存带宽。同时，更高效的专有计算核（如 TensorCore）也进一步加大了内存压力。

为了解决内存问题，研究员们提出了 Welder 深度学习编译器，全面优化由通用算子组成的端到端 DNN 模型的内存访问效率。其实，DNN 模型可以看作是由多个算子连成的一张图，整个计算过程涉及多个阶段，即数据需要流过不同的算子，在每个阶段都需要将张量切分成块，先搬运到处理器上进行计算，然后再搬运回内存，这就会造成很大的搬运开销。由于整个计算过程包含多个流程，所以还可以将这一过程想象成逐层向上搬运“砖块”的场景，其中第一个“工人”将“砖块”拿上去加工然后再放回去，第二个“工人”再拿上来雕刻一下再放回去，然后是第三个、第四个……，反复搬运，可以预想其中的开销不言而喻。那么是否可以让第一个“工人”在顶层完成一部分子任务后直接交给下一个“工人”继续处理，然后再将多项任务“焊接”起来，实现流水化作业呢？ Welder 正是扮演了电焊机这个角色，通过链接不同的算子，它可以让数据块以流水线的方式处理，大大降低了访存量，在近几年人工智能模型对访存效率要求越来越高的情况下，可以大幅提升计算效率。



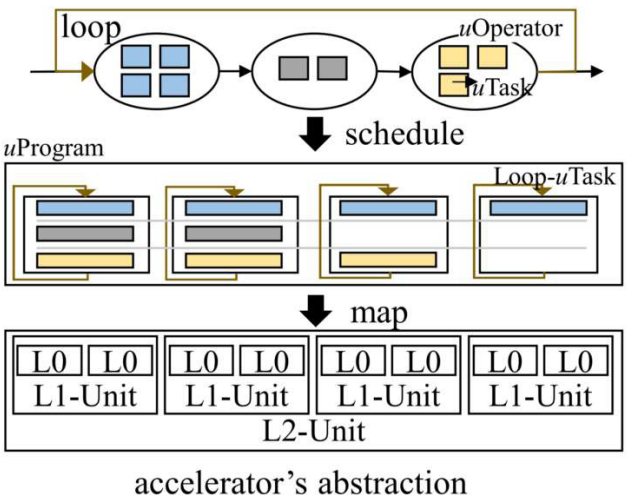
Welder 技术框架

在对 10 个主流的 DNN 模型（包括用于各种任务的经典和最新模型结构，如视觉、自然语言处理、3D 图形等）进行评估后可以表明，Welder 在 NVIDIA 和 AMD 的 GPU 上都显著超过了现有的主流框架和编译器，如 PyTorch、ONNXRuntime 和 Ansor，速度提升分别到达 21.4 倍、8.7 倍和 2.8 倍。Welder 的自动优化甚至超过了 TensorRT 和 Faster Transformer，最高可实现 3.0 倍和 1.7 倍的加速。此外，当在 TensorCore 等具有更快计算核心的硬件上

运行这些模型时，其性能有了更大的提高，突显了内存优化对未来人工智能加速器的重要性。

AI 编译“研磨机”Grinder：让控制流也能在加速器上高效执行

在计算程序中，对数据块的搬运过程有时候需要引入一些更复杂的控制逻辑，这就是人工智能程序数据流之外的控制流，如循环地遍历一个句子中的每个单词，或者根据输入动态决定执行哪一部分程序。当前的编译器大多都是在解决数据流问题，对控制流的支持并不高效，因此导致了控制流较多的模型无法高效利用加速器性能。研究员们认为，可以将控制流和数据流切分重组以此来进行更高效的优化，并推出了 Grinder（Grinder 为项目名称，论文中系统名称为 Cocktailer）。Grinder 好像一个便携的研磨切割机，它在把数据流切分成不同规模的并行计算块后，会再把控制流融入数据流，让控制流也能在加速器上高效执行。



Grinder 技术框架

Grinder 可以在硬件加速器上共同优化控制流和数据流的执行，并通过一种新的抽象来统一包括控制流和数据流的人工智能模型的表示，这就允许 Grinder 向较低级别的硬件并行性暴露用于重新调度控制流的整体调度空间。Grinder 使用启发式策略找到了有效的调度方案，且能够自动将控制流移动到设备内核中，进而实现了跨控制流边界的优化。实验表明，Grinder 可以对控制流密集的 DNN 模型加速 8.2 倍，是目前针对控制流的 DNN 框架和编译器中速度最快的一个。

基于同一套抽象和统一的中间表示层（Intermediate Representation，IR），这四款 AI 编译器解决了当前 AI 编译器中的不同问题 —— 并行、编译效率、内存、控制流，构成了一套完整的编译解决方案。在推进研究的进程中，微软亚洲研究院的编译原型系统已经为 Office、Bing、Xbox 等微软产品的部署和模型优化提供了帮助，同时也在微软新型算子的定制和优化中发挥了作用。

“在大模型成为主流的今天，人工智能模型对效率、算力有了更高的要求。一方面，AI 编译器需要针对硬件资源做出极致的算子融合、定制和优化；另一方面，也需要对新型大规模硬件架构进行系统编译支持，如片上网络互联（NoC）的芯片、混合内存架构等，甚至通过白盒编译方法指导硬件定制。我们提出的这套 AI 编译器已被证明能够大幅提升 AI 编译的效率，可以更好地助力人工智能模型的训练和部署。同时，大模型的发展也为 AI 编译带来了机遇，未来大模型本身或许就可以帮助我们实现优化和编译。”微软亚洲研究院首席研究员薛继龙表示。

相关论文：

Rammer: Enabling Holistic Deep Learning Compiler Optimizations with rTasks
https://www.microsoft.com/en-us/research/publication/rammer-enabling-holistic-deep-learning-compiler-optimizations-with-rtasks/

ROLLER: Fast and Efficient Tensor Compilation for Deep Learning
https://www.microsoft.com/en-us/research/publication/roller-fast-and-efficient-tensor-compilation-for-deep-learning/

WELDER: Scheduling Deep Learning Memory Access via Tile-graph
https://www.microsoft.com/en-us/research/publication/welder-scheduling-deep-learning-memory-access-via-tile-graph/

Cocktailer: Analyzing and Optimizing Dynamic Control Flow in Deep Learning
https://www.microsoft.com/en-us/research/publication/cocktailer-analyzing-and-optimizing-dynamic-control-flow-in-deep-learning/

Qlib 全新升级：强化学习能否重塑金融决策模式?

2020 年，微软亚洲研究院开源了金融 AI 通用技术平台 Qlib。Qlib 以金融 AI 研究者和金融行业 IT 从业者为用户，针对金融场景研发了一个适应人工智能算法的高性能基础设施和数据、模型管理平台。一经开源，Qlib 便掀起了一阵热潮，相关开源项目在 GitHub 上已收获了 11.4k 颗星。作为一个通用技术平台，Qlib 不仅大大降低了行业从业者使用 AI 算法的技术门槛，还为金融 AI 研究者提供了一个相对完整的研究框架，让他们可以基于专业知识探索更广泛的金融 AI 场景。

微软亚洲研究院对 Qlib 的研究并未止步于此，经过两年多的深入探索，Qlib 迎来了重大更新，在原有的 AI 量化金融框架基础上，又引入了基于强化学习和元学习的新范式以及订单执行优化和市场动态性建模的新场景，帮助相关从业者使用更先进和多样的人工智能技术来应对更复杂的金融挑战。

金融业务的目标复杂性和顺序决策流程的特殊本质，让构建有效的金融决策模型成为一项十分困难的任务。一方面，金融市场的交易规则及其相互作用十分复杂，给金融策略的模拟和评估带来了巨大挑战。并且，策略模型优化往往涉及收益最大化和风险最小化，是一个多目标优化问题，这进一步增加了获得监督信号的难度。

另一方面，金融市场一系列决策之间相互依赖，这些决策共同决定了最终的策略表现，这使得一系列机器学习算法的独立同分步（Independent and Identically Distributed，IID）假设无效，导致传统的监督学习、半监督学习、无监督学习方法很难适用于这些金融场景的决策。

而基于强化学习（Reinforcement Learning, RL）的学习范式不需依赖标注样本，可通过智能体与环境的交互来收集相应的样本（如状态、动作、奖励）进行试错学习，从而不断改善自身策略来获取最大的累积奖励。这种通过不断试错和探索环境来进行学习，以寻找更好策略的学习范式更有利于满足上述金融决策的需求。

“ 在应用强化学习时，应用环境需要具有一定的沙盒属性。因为强化学习是通过反复试错的机制进行学习的，如果结果正确它就会得到强化。在现实世界中，游戏和金融领域的回测都是典型的沙盒场景。所以，我们希望能够利用强化学习来帮助解决金融决策的问题。” 微软亚洲研究院高级研究员任侃说。

基于这一认知，Qlib 团队的研究员们针对交易决策和投资组合管理策略展开了研究，并在全新升级的 Qlib 中增加了基于强化学习的单智能体订单执行优化和多智能体批量订单联合优化的示例算法及其相应的平台支撑功能。

OPD 先知策略提取：更好的订单优化策略

订单执行优化是算法交易中的一个基本问题，目的是通过一系列交易决策完成预设的元交易订单（meta-order），如平仓、建仓及仓位调整。从本质上讲，订单执行的目标是双重的，不仅要求完成整个订单，而且追求更经济的执行策略，实现收益最大化或资本损失最小化。针对订单执行的顺序决策特点，强化学习方法可以发挥优势捕捉市场的微观结构，从而更好地执行订单。

但简单、直接地使用强化学习会遇到一个问题 —— 原始的订单及市场数据中存在大量噪声和不完美的信息。噪声数据可能导致强化学习的样本效率低下，使学习订单执行策略的有效性降低。更重要的是，在采取行动时，可以利用的信息只有历史市场信息，缺少明显的线索来对市场价格或交易活动的未来趋势做准确预测。

为此，Qlib 团队提出了一个通用的订单执行策略优化框架，引入了全新的策略提取方法 OPD（Oracle Policy Distillation，先知策略提取），来弥合噪声和不完美市场信息与最优订单执行策略之间的差距。该方法是一种 “ 教师 - 学生 ”（teacher-student）的学习范式，“ 教师 ” 在获得完美信息的情况下，会先被训练成一个可以找出最佳交易策略的 “ 先知 ”，然后 “ 学生 ” 通过模仿 “ 教师 ” 的最佳行为模式来进行学习。而当模型训练阶段结束进入到实际使用阶段时，OPD 会在没有 “ 教师 ” 或未来信息的情况下，使用 “ 学生 ” 策略进行订单执行的规划。而且，与传统强化学习方法只为单一股票训练单一模型的思路不同，Qlib 团队提出的这一强化学习算法可以利用所有股票的数据做联合训练，从而极大缓解学

习过程中的过拟合问题。

实验结果显示，OPD 的性能显著优于其它方法，证明了 OPD 的有效性，也证实了传统基于金融市场假设的方法在真实场景中并不适用。此外，其它基于训练的数据驱动的方法因为未能很好地捕捉到市场的微观结构，所以也无法相应地调整策略，导

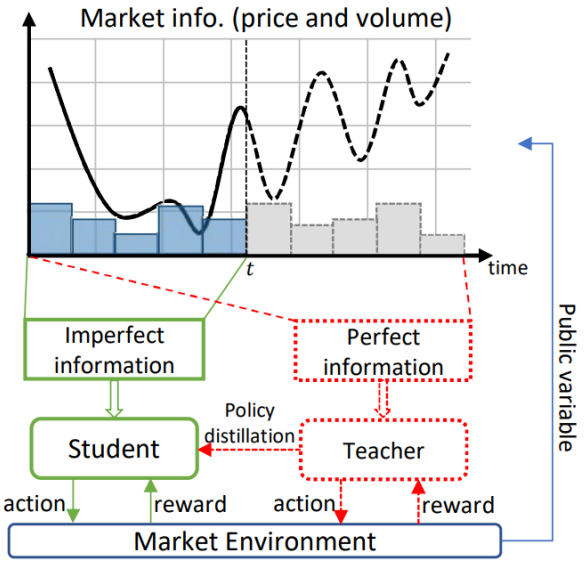


图 1：OPD 先知策略提取示意图

Category	Strategy	Reward($\times 10^{-2}$)	PA	GLR
financial model-based	TWAP (Bertsimas et al. 1998)	-0.42	0	0
	AC (Almgren et al. 2001)	-1.45	2.33	0.89
	VWAP (Kakade et al. 2004)	-0.30	0.32	0.88
learning-based	DDQN (Ning et al. 2018)	2.91	4.13	1.07
	PPO (Lin et al. 2020)	1.32	2.52	0.62
	OPD^S (pure student)	3.24	5.19	1.19
	OPD (our proposed)	**3.36***	**6.17***	**1.35**

表 1：OPD 方法实验结果（数值越高性能越好）

多智能体协作方案 MARL：显著提高批量订单的执行性能

在量化金融中，对资产管理的一类主要目标是通过在市场上连续交易多种资产来最大化长期价值。所以，除了订单执行外，投资组合管理也是量化金融中一个基础的场景，其目标是在一定的时间范围内，完成投资组合管理策略指定的大量订单，从而实现本轮的投资组合持仓调整，并尽可能降低换仓的成本甚至通过订单执行提高整体收益。

在多订单执行的联合优化中存在三个问题。首先，订单数量及交易金额每天都会根据投资组合的分配而变化，这要求订单执行策略具有可扩展性和灵活性，以支持多种不同的订单情况。其次，现金余额有限，所有的买入资产操作都会消耗交易者有限的现金供应，而出现的现金缺口只能通过卖出资产操作来进行补充。另外，

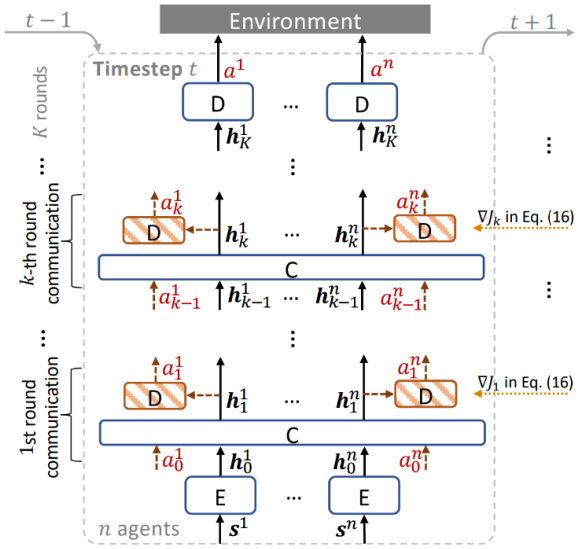


图 2：多智能体强化学习算法中多轮意图感知机制示意图

现金不足可能会使得投资者错过更好的交易执行机会，所以投资者要在买入及卖出之间实现平衡，避免交易决策因为现金短缺导致交易执行业绩不佳。

尽管市场上存在许多用于订单执行的工具，但这些工具很少能够同时解决上述三个问题。为了解决这些挑战，Qlib 团队推出了多智能体协作强化学习（Multi-Agent Reinforcement Learning, MARL）方法，让每个智能体执行一个单独的订单，再以分解联合行动空间（joint action space）扩展到多个不同的订单，并且所有智能体协作可以在较少的决策冲突情况下实现更高的总利润。为了加强各智能体之间的协作，研究员们还提出了一种新的多轮意图感知通信机制，以了解每个协作阶段智能体的行动意图，该机制使用了新的行动价值归因（value attribution）方法，可以直接优化和细化每一轮智能体的预期行动。

实验表明，在 A 股及美股数据上共 6 个不同测试时间窗口里，MARL 相比于单智能体强化学习、简单的多智能体强化学习及传统金融模型方法都在各项指标上有显著提升。此外，意图感知通信机制大大降低了 TOC 度量（用于衡量多订单执行中买、卖操作不均衡带来的现金短缺情况），这表明采用通信共享意图行动的方法比以前的 MARL 方法提供了更好的协作性能。并且研究员们提出的 IaC 方法的效果，远远超出了此前一些利用通信共享智能体意图的方法，这表明在单个时间段内细化多轮的行动意图对于智能体在复杂环境中实现良好协作来说至关重要。

强化学习在金融领域的研究离不开专用框架的支撑

研究新算法通常需要快速地进行反复迭代，而迭代效率则在很大程度上取决于研究框架的完善程度。为了更好地推进强化学习在金融领域的前沿研究，Qlib 针对金融领域的特性，提供了全面的框架支持。

Qlib 新发布的金融领域强化学习框架提供了三个关键特性，以解决强化学习在金融领域应用的常见问题。

1. 在金融领域使用强化学习时，用户往往需要对接金融强化学习环境，通过设计马尔可夫决策过程（Markov Decision Process, MDP），集成强化学习策略算法。整个过程需要大量的工程工作，同时也需要大量的金融专业知识和实战经验，非常费时费力，导致研究人员无法专心于研究问题本身。Qlib 直接提供了涵盖上述问题的完整技术栈，免去了研究人员大量繁琐的重复工作。

2. 强化学习是通过与环境交互试错来优化策略的。但模拟环境与实际市场环境之间往往存在较大差异，这种差异可能导致模拟环境的最优解与真实环境的最优解存在很大的差距，这是强化学习研究落地的难点之一。这种差距一方面来自于真实交易包含了大量繁琐的规则，而一般用于学术研究的交易框架常常会忽略这些规则；另一方面，真实交易中通常是不同层次的交易互相结合使用（如日频交易和高频交易），忽视这部分交互影响也会对模拟产生偏差。Qlib 在设计时尽可能考虑到了各种规则，并将嵌套决策框架（nested decision making）用于模拟真实交易时不同层次交易策略的互相影响，从而最大限度地减少模拟误差。

3. 强化学习需要大量计算资源，涉及与环境的交互和试错，可能需要多次迭代才能达到最优策略。特别是在金融市场的复杂规则下，这些交互可能非常耗时，需要大量内存和计算。为了加速强化学习的研究迭代，优化训练和测试流程至关重要。Qlib 提供了不同仿真程度的模拟器，用户可以在训练时在不同的阶段使用不同仿真程度的模拟器（例如，在训练早期使用低仿真度但运行效率极高的模拟器，在训练后期使用高仿真同时资源开销较大的模拟器），从而实现在获得高仿真环境下的最优策略的同时，节约计算资源并加快训练速度。在测试环节，通过 Qlib 可灵活调度强化学习智能体的训练及测试环境的这一功能，实现提高回测并行度以加速策略的评估。

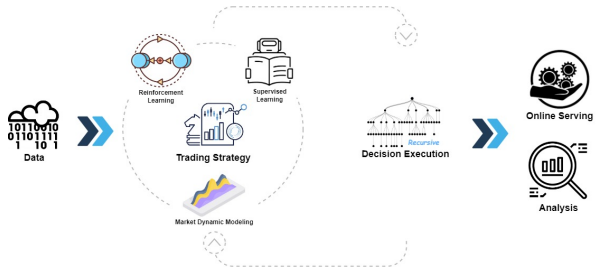


图 3：全新升级的 Qlib 框架示意图

实时市场动态建模：更有效地预测未来数据分布

在现实世界的真实场景中，人们处理的数据往往是随时间顺序收集的流式数据，但机器学习算法能够被广泛应用于现实世界一般依赖数据独立同分布的假设。然而，金融领域的数据是非独立同分布的，它的规律会随着时间产生变化，这就导致传统的依赖独立同分布假设的机器学习模型难以在不同时间上同时进行有效预测。这种流数据分布以不易预测的方式发生变化的现象被称为概念漂移。

为了处理概念漂移，此前的方法是先检测概念漂移发生的时间，再调整模型以适应最新数据的分布，但是这类方法无法应对数据分布在下一个时刻继续发生变化的问题。Qlib 团队的研究员们发现，除了一些极端难以预料的分布突变，概念漂移常常以渐进地非随机方式演变，且这种渐进的概念漂移在某种程度上是可预测的，即概念漂移本身就存在一定的趋势和规律。而实际上这种场景在流数据中十分常见，但大多数现有研究都较少关注这一方向。因此，Qlib 团队通过预测未来的数据分布来关注可预测的概念漂移，并提出了新的方法 DDG-DA（Data Distribution Generation for Predictable Concept Drift Adaptation），来有效地预测数据分布的演变，并提高模型的性能。其具体的思路是，首先训练预测器来估计未来的数据分布，然后利用它生成训练样本，最后在生成的数据上训练模型。

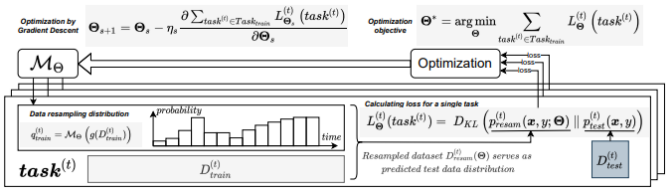


图 4：DDG-DA 算法示意图

DDG-DA 方法已经在三个实际任务中进行了实验：股票价格趋势、电力负荷和太阳辐照度的时序预测，并在多个广泛使用的模型上获得了显著的性能提升。

微软亚洲研究院高级研究员杨骁表示，“如果用户在使用工具时没有考虑到时间上数据分布的动态变化，那么最终的建模将是不完善的。我们的动态市场建模方法可以动态调整数据分布，让模型更好地学习和适应当前市场的规律。相比于传统使用历史数据构建模型进行预测的方法，DDG-DA 能够根据实时的市场规律变化，使用与未来分布更相似的数据建模，从而可以更准确地预测未来。”

元学习框架助力市场动态性建模

在市场动态性研究中，DDG-DA 通过调整数据分布间接地影

响预测模型的训练过程，从而影响最终的预测结果。这种训练模式本质上是在学习如何训练一个模型，可以归结到元学习（Meta Learning）范畴。Qlib 提供了一套元学习框架，定义了元学习中任务、数据、模型的接口规范。

使用这套框架，研究者和从业人员不仅可以训练模型，还可以设计元模型（meta model）来自动地学习如何更好地训练模型，这为开展 DDG-DA 类似的研究工作提供了极大的便利。未来，Qlib 团队希望这个框架能够为更多的元学习算法提供支持，从市场动态性研究开始，扩展到更多的场景和问题。

更新版 Qlib 已开源，全新功能等你来探索

集成了最新功能的多范式 Qlib 现已在 GitHub 上发布。其新增的框架和组件能更好地支持强化学习这一学习范式在金融领域中进行智能决策相关的研究和应用。同时 Qlib 团队还发布了基于 Qlib 框架在订单执行这一典型场景下，基于强化学习的先知策略提取 OPD 及多智能体协作 MARL 的两个示例算法。而对元学习范式的支持也使得类似于市场动态性建模这类依赖元学习范式场景上的相关研究得以高效地开展并且更方便于实际应用，为智能金融决策又增加了一个成功的砝码。

“从数据处理到算法支撑，再到模型的训练与验证，此前的 Qlib 在纵向深度上为金融 AI 研究者和金融行业从业者提供了一个全方位面向 AI 量化投资的研究框架，而升级后的 Qlib 则在横向广度上为智能金融决策提供了更多新的学习范式，能够帮助使用者更精准地匹配金融业务及相关研究的需求。全新升级的 Qlib 将更多的 AI 算法、学习范式与更广阔的金融任务、场景相连接，提供了一个更易用、更高效的量化金融研究平台。”微软亚洲研究院首席研究员刘炜清表示。

如何在微软 Edge 浏览器上一键观看高清视频?

视频是当下最流行的媒体形式之一。但由于视频压缩、网络不稳定等原因，我们常常可以看到互联网上的很多视频其画面质量并不理想，尤其是在浏览器端，这极大地影响了观看体验。不过，近期微软 Edge 浏览器推出了一项新功能，一键就可以让浏览器中的视频变为高清版。这项神奇功能背后的技术秘诀是什么？今天，让我们一起来了解一下微软 Edge 视频超分辨率功能的“秘密武器”——来自微软亚洲研究院的智能视频增强工具集 DaVinci 2.0。

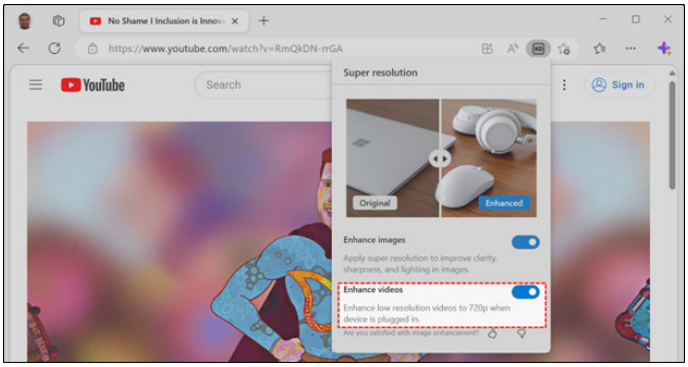
近期，微软 Edge 浏览器推出了一项新功能——视频超分辨率（ VSR）。用户只需在 Edge 浏览器中开启 VSR 功能，就能够在浏览器端观看高清视频。即使是几十年前的 360P、480P 老电影，或者在网络不稳定被迫降低视频画质的情况下，用户也可以时刻享受高清体验。

VSR 功能的背后是来自微软亚洲研究院的智能化视频增强工具集“达芬奇（DaVinci）”。该功能在不占用网络带宽的情况下即可在用户端实时消除视频压缩的伪影，提高视频分辨率，从而整体提升用户浏览视频的视觉体验。

现在，就跟着微软 Edge 的节奏，让视频高清起来吧！

第一步，打开微软 Edge 浏览器；第二步，单击 Edge 地址栏中的高清图标并选择增强视频的切换开关；第三步，播放视频，享受高清体验。（注：视频超分辨率由 Edge 自动启用，用户可自行决定启用或禁用该功能）

* 值得注意的是，受限于模型计算代价较高的限制，该功能目前仅针对具有相对高端显卡的台式机用户开放测试（需要 Edge Stable 版本不低于 117， Edge Canary 版本不低于 119）。同时，微软 Edge 团队也在不断努力，希望可以将该功能逐步开放给所有具有独立显卡、集成显卡的用户。



Long Description
启用或禁用视频超分辨率功能的流程

从特定视频域到开放域的挑战

据微软 Edge 团队调查，近四成用户曾表示在 Edge 浏览器观看视频时，网页上的视频质量较低，通常为 360P 或 480P，非常影响用户体验。为此，微软 Edge 团队希望与微软亚洲研究院开展合作，借助创新技术来提升 Edge 网页端所有低清视频的质量，给用户以高清体验。2022 年微软亚洲研究院推出的智能视频增强工具集“达芬奇（DaVinci）”，能够实现视频超分辨率、视频插帧、压缩视频超分辨率等功能，很好的满足了微软 Edge 团队的需求。(DaVinci 项目链接：https://github.com/microsoft/DaVinci)

然而，在将 DaVinci 算法模型应用到产品的过程中存在着不小的挑战。DaVinci1.0 主要是针对特定领域进行的训练，有明确的训练目标；特定领域的数据分布一致，所以模型的优化过程更加容易，优化的上限也更高；而且，高质量的垂直领域的数据更易于收集，可以获得大量公开的训练数据。但进入到 Edge 应用场景下的开放域（open domain），技术难度呈指数级增加。在开放域中，视频类别众多，视觉差异较大，比如用户在 Edge 中打开的可能是包含动物、植物、建筑、车辆等众多元素在内的影视、动画、视频会议等各种不确定类型的视频。要让一个模型补充不同类别视频的细节，是 DaVinci 首先要面对的难题。

与此同时，模型的容量是否足够大，可以支撑真实场景下的大量数据，并捕捉到不同的数据模式？如何定义开放域？开放域需要包含哪些特定领域的数据？评估指标是什么？这些都是 DaVinci 模型需要克服的问题。

更适合开放域视频的超分辨率算法

DaVinci 1.0 视频超分辨率模型的目的是在从低质量（LQ）或低分辨率的对应帧预测的高质量（HQ）帧的过程中来学习映射函数。然而，为了从高质量的训练数据集生成对应的低质量 / 低分辨率的视频帧，现有方法大多是使用预定义的算子（如，双三次下采样，bicubic down-sampling）来模拟退化过程，得到 LQ 输入。这就限制了模型在真实视频场景上的通用性，特别是对于具有高压缩率的视频流数据。所以在 DaVinci 2.0 的视频超分辨率技术中，微软亚洲研究院的研究员们将视频压缩也纳入到模型中，并通过运行具有不同压缩策略的几个流行视频编解码器来合成 LQ-HQ 视频对，以训练模型。

同时，受到大语言模型的启发，研究员还利用自监督的 LQ-HQ 复原范式（restoration paradigm），使用来自不同类别的 15 万个视频片段对模型进行了预训练。通过进一步考虑来自不同编码器的视频压缩伪影类型，使得 DaVinci 模型可以显著恢复具有大范围低质量的不同视频内容。

为了进一步提高模型的视觉质量，研究员们采用两阶段训练策略。其中，第一阶段旨在恢复结构信息（如，对象的边缘和边界），第二阶段则针对高频纹理（如，树叶和毛发），使用视觉感知和生成对抗性目标进行优化。

由于当前该领域中的现有指标，如 LPIPS（Learned Perceptual Image Patch Similarity，学习感知图像块相似度）和 FVD（Fréchet Video Distance，弗雷歇视频距离）不能完全反映人类的视觉偏好，因此研究员们构建了一个端到端流水线（pipeline），用于视频增强任务的主观评估，以便更好地了解改进后的 DaVinci 模型性能，评估它在开放域视频场景中所发挥的作用。

具体而言，就是让参与者在十个类别中标注出他们对真实场景视频数据不同方法的偏好。参与者不仅要考虑每个视频帧的静态质量，还要考虑动态质量，这对于改善用户体验尤为重要。该流水线评估方法表明，相比于浏览器中默认的双线性放大，超过 90% 的用户更喜欢使用 DaVinci 2.0 模型来提升视频质量。

在微软亚洲研究院与微软 Edge 团队的通力合作下，Edge 浏览器的 VSR 功能基于 DaVinci 2.0 超分辨率模型，可以提升所有不确定内容类型的视频质量，并且在不产生伪影的情况下，提高视频清晰度，为 Edge 用户提供丝滑、高清的视频体验。

从视频超分辨率到视频生成

尽管 DaVinci 1.0 并没有涉及到如此大规模的低质量数据预训练，但 DaVinci 2.0 在 Edge 浏览器中的成功应用，证明了模型具有从低质量预训练到大规模高清数据应用的高泛化能力。这也进一步促进了微软亚洲研究院研究员们将创新技术应用到更多开放域场景的探索。

“DaVinci 2.0 对视频增强功能的创新，实现了对开放域视频图像

细节的补充。基于视频帧间具有本质关联的特性，DaVinci 最终实现了高清结果。接下来，我们希望对技术进行更深入的探索，最终达到从 0 到 1 的创造，”微软亚洲研究院高级研究员傅建龙表示。

在以视频为主流媒介的大趋势下，微软亚洲研究院希望未来还可以给用户提供自动生成视频、创建个性化视频内容的工具。在全方位为用户提供极致的视频观看体验的同时，也帮助用户从事更复杂、更具创造力的内容创作工作。

相关链接：

相关论文：

Universal Trading for Order Execution with Oracle Policy Distillation
https://arxiv.org/abs/2103.10860

DDG-DA: Data Distribution Generation for Predictable Concept Drift Adaptation
https://arxiv.org/abs/2201.04038

GitHub 链接：
https://github.com/microsoft/qlib

扫描二维码查看相关视频

MABIM：多智能体强化学习算法的“炼丹炉”

现实世界中，许多问题和任务都是由多个参与者交互进行的，所以要想使用人工智能技术解决真实世界的问题，就需要更好地模拟这种复杂的环境，而这正是多智能体强化学习（MARL）的强项。早在 2020 年，微软亚洲研究院基于多智能体强化学习，推出了面向多行业横截面上的多智能体资源调度平台 MARO。

随着研究的深入，研究员们发现互动式的学习环境和测试平台对多智能体强化学习的发展至关重要。为此，近期微软亚洲研究院在 GitHub 开源了一个能够灵活适应多智能体强化学习各种挑战的学习测试平台——MABIM，从而可以更好地测试 MARL 算法，让其更容易迁移到真实的应用场景中。

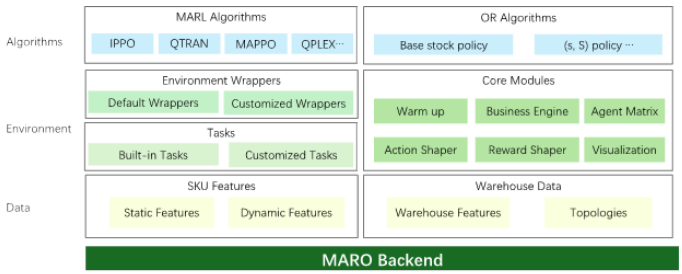
多智能体强化学习 (Multi-Agent Reinforcement Learning，MARL) 是强化学习研究的一个重要分支，旨在让多个智能体在特定环境中通过合作与竞争的方法来实现共同目标。与传统的单智能体强化学习相比，MARL 具有多项优势：能够更好地模拟现实世界的复杂环境，解决涉及多个参与者的问题，并提高系统的鲁棒性、学习效率、自适应与可扩展性。正是这些优势让 MARL 成为了解决实际问题的有力工具，在机器人协同控制、自动驾驶、游戏、经济学、金融、医疗等领域具有广泛的应用前景。

MABIM 基准测试平台：助力训练最具实用价值的 MARL 算法

强化学习算法的发展与进步离不开互动式学习环境和测试平台。这些环境为强化学习提供了丰富的学习空间，使智能体得以在实践中不断优化决策策略，从而在各种复杂应用场景中取得成功。近年来，MARL 领域涌现出许多不同类型的学习环境，对 MARL 算法的发展产生了积极的影响。然而，目前还没有学习环

境既能充分考虑到 MARL 领域的众多挑战，又能提供灵活的定制和扩展。

库存管理作为供应链领域最关键的场景之一，在企业运营中具有非常重要的地位。通过合理的库存管理，企业可以降低成本、提高客户满意度、保障生产稳定、提高资金周转速度，进而实现企业经济效益的最大化。因此，微软亚洲研究院的研究员们以运筹学领域的库存管理问题为基础，设计了一个具有高自由度、支持多级多商品库存网络的 MARL 基准测评框架——MABIM (Multi-Agent Benchmark for Inventory Management)，并已在 GitHub 上开源。



MABIM 框架图

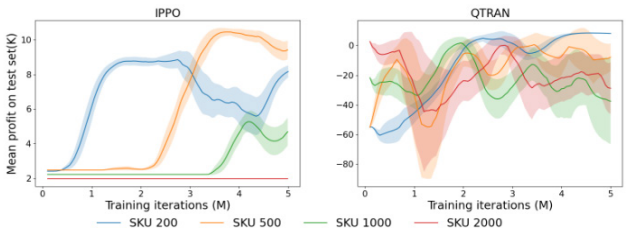
MABIM GitHub 链接：https://github.com/victoryxl/replenishmentenv

MABIM 平台可以灵活适应 MARL 的各种挑战。通过配置参数，MABIM 能轻松定制不同环境，模拟各种具有挑战性的场景。例如，可以通过设置不同层级的库存网络和不同数量的商品来模拟众多智能体之间的协作、通过设置不同的仓库空间来模拟智能体之间不同程度的竞争与合作、通过设置不同的客户需求来模拟非平稳的环境等等。

MABIM 共内置了 51 个具有挑战性的任务，涉及 MARL 领域多种不同挑战的组合，可以用来测试 MARL 算法在复杂场景下的适配能力和运行效果。比如，针对解决复杂合作和竞争关系的 MARL 算法，可以使用多个层级库存网络加上受限的仓库库容测试；对于着重解决可扩展性的 MARL 算法，可以使用含有更多个商品(>=1000)的任务进行测试。此外，MABIM 还具有高运行效率、基于 GYM 标准接口、完整的策略可视化工具和基于真实数据等特点，使其能更好地支持 MARL 的研究。

MARL 挑战犹在，MABIM 的研究还将继续

研究员们利用 MABIM 测试了多种经典的运筹学和多智能体强化学习算法，发现了一些有趣的结论，如 IPPO 算法在智能体数量增多时训练将变得困难，QTRAN 算法会变得不稳定；在资源紧张的竞争环境中，IPPO 表现出短视行为，为了避免短期的损失而采取长期不盈利的策略；在需要上下游合作的环境中，纯 MARL 算法难以学习到有效的上下游策略；在非平稳环境中，MARL 策略优于普通运筹学算法等。这说明，虽然 MARL 算法在业界有很大的应用潜力，但也面临着更大的挑战，如计算复杂度会随智能体数量指数级增加、智能体之间的合作与竞争、不稳定的环境等。



IPPO 和 QTRAN 算法的训练随着智能体数量的增加变得不稳定

计算复杂度：随着智能体数量的增加，MARL 的计算复杂度会呈指数级增加。这是因为每个智能体都需要考虑其他智能体的策略，从而导致状态空间和动作空间迅速增大。这给学习和优化过程带来了巨大的挑战，尤其是在大规模多智能体系统中，如在库存管理领域，当有大量成千上万的商品需要做决策时，每个商品都可能需要考虑其他商品的决策。这使得计算复杂度迅速增加，让实时决策和控制变得困难。

合作与竞争：智能体之间的合作和竞争关系是 MARL 的核心挑战之一。合作关系需要智能体之间共享信息和协调行动，而竞争关系需要智能体在有限资源下优化自身目标。这些关系的建立和维护对于学习有效策略至关重要，但在实际应用中可能非常困难，比如在库存管理场景中，多个商品需要在有限的资源下竞争（预算、仓库货架空间等），同时也需要与其他商品合作以维持整体效益最大化。在这种情况下，设计既能合作又能竞争的强化学习算法是一项巨大的挑战。

不稳定的环境：在 MARL 中，智能体的行为会影响环境，从而影响其他智能体的学习过程，这使环境变得非平稳和不确定，给学习和优化带来了额外的困难。比如在库存管理领域，每个商品的未来需求是不确定的，导致了整个环境有很大的不确定性。

虽然 MABIM 是基于库存管理任务的学习环境，但其涉及的众多问题在业界具备一定的普遍性，经过 MABIM 测试的 MARL 算法将更容易迁移到业界的其它应用中。未来，微软亚洲研究院还将继续完善 MABIM，包括将库存管理模型扩展到树形或网络结构，以评估智能体之间的通信能力；隐藏部分商品特征，以评估算法在部分观测情况下的表现。通过这些扩展，研究员们希望 MABIM 能够更接近真实场景，进一步降低算法从实验室到真实系统迁移的代价，助力业界解决真实场景中的难题。

机器学习开源工具 BatteryML，一站式分析与预测电池性能

天下苦锂电池寿命久矣，时闻“开车出，推车回”，又闻“充电两小时，待机两分钟”，亦闻“气温骤降，请注意电池保暖”……随着以锂离子电池为动力源的产品，如手机、电脑、新能源汽车等，逐步成为人们生活的必需品，关于电池寿命的碎碎念也越来越多。电池性能预测也成为了产业人工智能研究的重要课题之一。

为了更好地分析电池性能，预测电池使用寿命，微软亚洲研究院开发并开源了一站式机器学习工具 BatteryML，希望可以集结更多的专业力量，共同推动电池领域的研究。

近年来，锂离子电池由于其高能量密度、长循环寿命和相对较低的自放电，已成为储能解决方案的基石，也被广泛应用于各种商业场景中，包括新能源汽车、消费电子和储能设施等。尽管锂电池带来了诸多优势，但它仍面临着容量衰减和性能优化等挑战。

在不断循环使用的过程中，锂电池因固有的电化学特性不可避免地导致了其性能的衰退，具体表现为充放电容量下降。这种不受控的性能衰退会对下游的商业场景造成极大的影响，比如导致新能源汽车用户的“里程焦虑”、影响储能系统的供电稳定性等等。而且，过快的锂电池容量衰减也会给可持续发展带来很大的挑战，包括设备维护成本增加、造成稀缺资源消耗、加剧环境污染和影响产业经济效益等。因此，有效地分析与预测锂电池的性能衰退，进而为提前预防和干预提供指导，成为了一个非常重要的产业人工智能研究课题。

克服建模挑战，实现电池性能的分析和预测

锂电池的性能衰退是一个复杂的电化学过程，其涉及到固体电解质膜的增长、锂析出、活性材料损失等等。这个过程会受到电极材料、环境温度、充放电条件和速率等多种因素的复合影响，从而导致基于有限电化学规律的物理模型很难有效建模实际条件下的电池性能衰退过程。不过，如今的机器学习方法则能够自动从数据中归纳复杂的规律，并在近些年成为了建模电池性能衰退的重要工具，引起了学术界和工业界的广泛关注。

然而，在电池性能建模场景中研究和应用机器学习模型也极具挑战。一方面，对于电池行业的领域专家来说，尽管他们对于电池衰退背后的机制和原理有着深刻的认识，但却不擅长打造有效的机器学习模型。因为构建机器学习模型往往需要特殊的数据处理、有效的特征建设、精细的模型构建与调优等准备工作。另一方面，对于计算机行业的数据科学家来说，电池领域存在数据异构化严重、领域知识要求高、任务定义多样化等特点，极大地阻碍了他们研究及应用最先进的机器学习方法。

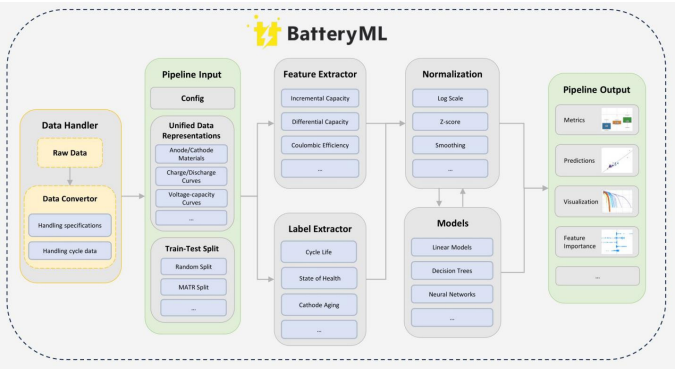
为了应对这些挑战，微软亚洲研究院的研究员们开发了 BatteryML 工具，以用于电池性能的分析与预测。同时，微软亚洲研究院也希望 BatteryML 既可以成为电池行业领域专家的一站式解决方案，也能够作为计算机科学家研究电池性能预测的高效开发平台。

获取 BatteryML 详细的文档、示例及源代码，请访问 BatteryML GitHub 链接：

https://github.com/microsoft/BatteryML

打造社区驱动的开源平台

作为一个一站式解决方案，BatteryML 简化了电池数据的研究和实验过程，涵盖了锂电池研究领域的各种经典模型。



BatteryML 架构示意图

BatteryML 主要有六大特点：

开源和社区驱动：BatteryML 旨在打造社区驱动的开源平台，从而集电池领域专家和数据科学家的合力，共同推动电池性能建模领域的进步。

统一的数据表征：BatteryML 构建了一套统一的数据表征来应对电池数据异构化问题，并提供了全面的处理脚本来汇总目前几乎所有的公开电池数据集。

预处理和特征工程：BatteryML 默认提供了一套标准的数据预处理流程，并内置了基础的特征工程来应对领域知识的高要求挑战。

丰富的模型：BatteryML 涵盖了电池性能预测领域已有的经典模型和基准评测。

明确的任务体系：BatteryML 包含了当前最受关注的电池性能预测任务，既与经典研究论文中的任务相对齐，又包含产业中更加关注的一些变形任务。

可扩展和可定制：BatteryML 预留了全面的接口，让研究员和开发者们可以根据自己的需求，高效地定制独有的数据集、开发新的数据处理及特征工程、研究更先进的机器学习模型等。

未来，微软亚洲研究院关于电池性能预测的研究进展将陆续在 BatteryML 平台开源，包括此前提出的 “多面深度对比回归（multi-faceted deep contrastive regression）” 方法、大模型在电池领域的应用等。

通过开源 BatteryML 解决方案，微软亚洲研究院期待加速电池性能预测领域的产学研融合，并希望有更多关注电池性能建模的伙伴加入，贡献新功能、新模块代码，向社区分享新的开源数据，共同推动电池领域的研究进步。

Distributional Graphormer：从分子结构预测到平衡分布预测

近年来，深度学习技术在分子微观结构预测中取得了巨大的进展。然而，分子的宏观属性和功能往往取决于分子结构在平衡态下的分布，仅了解分子的微观结构还远远不够。在传统的统计力学中，分子动力学模拟或增强采样等是获得平衡分布中采样的常用方法，但这些方法昂贵又耗时。

针对这个长期且艰巨的挑战，微软研究院发布了可用于预测分子结构平衡分布的深度学习框架 Distributional Graphormer (DiG)。DiG 可以快速生成真实多样的构象，进而为实现从单一结构预测到平衡分布预测的突破奠定基础。实验表明，DiG 在蛋白质、蛋白质 - 配体复合物和催化剂 - 吸附质系统等采样任务中，展现出了优异的性能和潜力，为分子科学研究打开了新的图景，并为药物设计、材料科学等领域带来新的可能。

结构预测是分子科学中的一个根本课题，因为分子的三维结构决定了分子的特性和功能。近年来，深度学习方法在分子结构预测方面取得了显著进展，并产生了重大影响。例如，深度学习模型 AlphaFold 和 RoseTTAFold 在从氨基酸序列中预测最有可能的蛋白质结构方面达到了前所未有的准确度；由微软研究院研发的 Graphormer 模型可以精准预测催化剂表面分子的吸附构象，并在全球首届公开催化剂挑战赛中夺冠。尽管深度学习方法改变了分子科学的游戏规则，但为分子的静态结构提供单一快照，仅揭开了复杂分子系统的冰山一角。

以蛋白质分子为例，蛋白质并不是刚性物体，它们是动态的分子，在平衡状态下可以呈现不同的结构，每种结构都有特定的出现概率。平衡分布下的结构及其出现的概率决定了分子的宏观属性和功能，从而才能揭示其生物学原理并对现实应用产生影响。而获得这些平衡分布的传统方法，如分子动力学模拟或蒙特卡洛采样都是从分布中顺序采样，由于其计算成本高，并且采样样本之间统计不独立，所以导致该类方法难以轻易用于复杂的实际应用场景中。因此，分子科学领域迫切需要找到全新方法，可以从分子结构预测问题迈进到分子的平衡分布预测。

DiG：预测平衡态下分子结构的分布

微软研究院发布的全新深度学习框架 Distributional Graphormer (DiG)，可以用于预测平衡态下分子结构的分布，旨在攻克平衡分布预测这一基础性难题，为分子科学研究创造了新的机遇。DiG 实现了从单一结构预测扩展到对平衡分布的整体预测的重要突破。平衡分布预测弥合了由统计力学和热力学控制的分子系统微观结构和宏观特性之间的差距。这是一项非常具有挑战性的任务，因为它需要对高维空间中的复杂分布进行建模，以捕捉不同分子状态的概率。

通过对此前研究工作 Graphormer 的扩展，DiG 实现了分布预测的全新解决方案。Graphormer 是一种通用的图 (Graph) Transformer，可以有效地对分子结构进行理解和建模，在分子科学中表现出了优异的性能，在量子化学或分子动力学模拟中也得到了应用。现在，DiG 具有更新、更强大的功能——通过深度神经网络直接预测平衡分布。

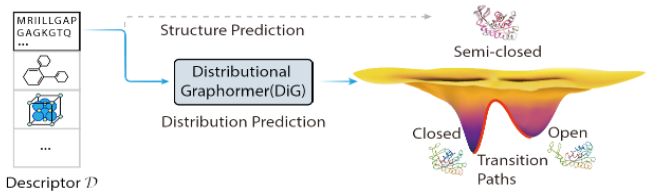


图 1：DiG 的目标是以分子系统的基本描述符（例如氨基酸序列或分子化学式）作为输入，预测符合平衡分布的结构及其概率。

DiG 受到热力学和优化的经典方法——模拟退火算法（simulated annealing）启发，通过模拟一个随机过程，将一个简单分布逐渐完善，从而产生一个复杂分布。此随机过程的预测在深度学习框架中完成。这也是最近将生成式人工智能推向火热的扩散模型（diffusion models）的模式。DiG 将这一思想又带回了热力学研究，形成了一个灵感和创新的闭环。可以想象在不久的未来，科学家们将可以像使用 AI 作画一样来使用 DiG 生成分子结构：通过输入简单的描述，例如氨基酸序列，DiG 就可以快速生成符合平衡分布的、真实多样的分子结构。这将大大提高科学家的生产力和创造力，使其能够在药物设计、材料科学和催化等领域获得新的发现与应用。

在多种分子体系采样任务中，DiG 颠覆传统

DiG 框架已在多个分子采样任务上展现出优异的性能和潜力，这些任务涵盖了广泛的分子系统，如蛋白质、蛋白质 - 配体复合物和催化剂 - 吸附质系统等。研究结果显示，DiG 不仅能够以高效率和低计算成本生成真实、多样的分子结构，还可以提供状态密度的估计，这对于使用统计力学计算宏观性质至关重要。DiG 在从统计学角度理解微观分子并预测其宏观特性方面取得了重大进展，为分子科学创造了更多令人兴奋的研究机会。

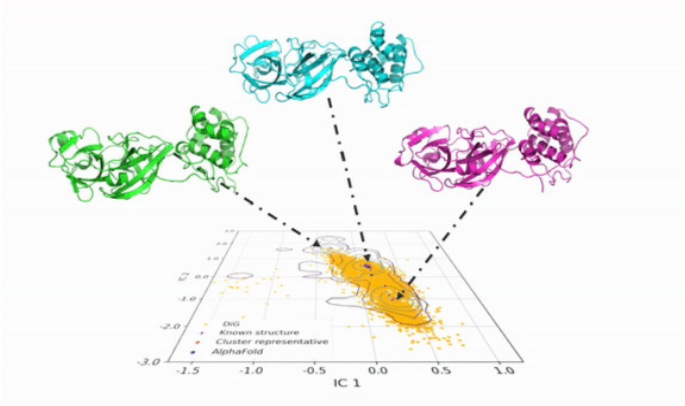


图 2：DiG 生成的结构覆盖了新冠病毒主蛋白酶长时间平稳态动力学模拟在 2 维投影构象空间中分布的主要区域

DiG 的重要应用之一是对蛋白质构象进行采样，这对于理解蛋白质性质和功能是必不可少的。蛋白质是动态分子，在平衡状态下会形成不同的结构且形成的概率各不相同，而这些结构通常又与其生物功能和与其他分子的相互作用有关。但是预测蛋白质构象的平衡分布是一个长期存在且具有挑战性的问题，原因在于构象空间中的概率分布取决于复杂和高维的能量景观图（Energy Landscape）。与昂贵且低效的分子动力学模拟或蒙特卡洛采样方法相比，DiG 可以从氨基酸序列中生成多样化并与功能相关的蛋白质结构，不仅速度快，而且成本显著降低。

DiG 可以从相同的蛋白质序列中产生多种构象。如图 2 所示，DiG 生成了 SARS-CoV-2 病毒主蛋白酶的结构，并与分子动力学模拟和 AlphaFold2 的预测结果进行了比较。在二维空间中，等高线图（以线条表示）显示了由大规模分子动力学模拟采样的三个簇，DiG 在三个簇中均生成了高度相似的结构。

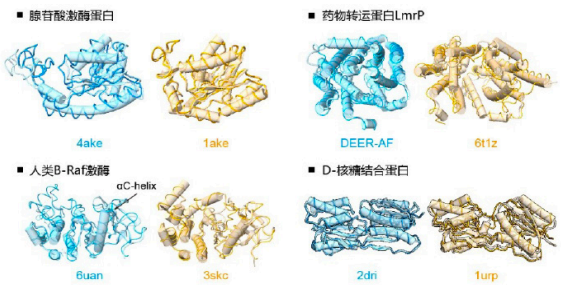


图 3：DiG 在产生蛋白质多种构象方面的性能。在 4 种不同蛋白，DiG（薄带状）产生的结构与实验确定的结构（圆柱）高度一致。

图 3 是将 DiG 在四种蛋白质上产生的结构与实验结构进行对比，每种蛋白质都有两种可区分的构象，对应着独特的功能状态。对于左上的腺苷酸激酶蛋白（Adenylate kinase）有开放和闭合状

态，两者都被 DiG 很好地采样。类似地，对于右上的药物转运蛋白 LmrP，DiG 也生成了对应两个功能状态的结构。值得注意的是，闭合状态是通过实验确定的（第二列下方的棕色示例，PDB ID 6t1z），而另一种状态则是与实验数据一致的 AlphaFold2 预测的模型。对于图 3 左下的人类 B-Raf 激酶而言，主要的结构差异位于 A 环 (A-loop) 区和附近的螺旋，也被 DiG 很好地捕获到了。另一个有趣的例子是具有两个分离结构域的 D- 核糖结合蛋白（右下），可以被包装成两种不同的构象。虽然 DiG 完美地生成了垂直构象，但未能预测扭曲 / 倾斜构象。尽管如此，DiG 还是生成了似乎是中间态的构象。总之，DiG 展示了生成与功能相关状态对应的多样化结构的能力，这在此前专注于结构预测的方法中尚未实现。

DiG 的另一个应用是对催化剂 - 吸附质系统进行采样，这是多相催化的核心。识别活性吸附位点和稳定的吸附质构型是理解和设计催化剂的关键，但由于复杂的表面分子相互作用，这项工作也非常具有挑战性。密度泛函理论（DFT）计算和分子动力学模拟等传统方法往往非常耗时且成本高昂，特别是对于大型的复杂表面。DiG 提供了快速、准确的解决方案，可以根据基质和吸附质描述符，预测吸附位点和构型及其相应的概率。DiG 还可以处理不同类型的吸附质，如单原子或分子，以及金属或合金等不同类型的基质。

通过 DiG，研究员们预测了各种催化剂 - 吸附质系统的吸附位点，并将预测结果与 DFT 计算得到的能量进行了比较。如图 4 所示，DiG 可以找到所有稳定的吸附位点，并产生类似于 DFT 结果的吸附质构型，效率高且成本低。DiG 还可以估算不同吸附构型的形成概率，这与 DFT 能量非常一致。

DiG 还在蛋白 - 配体采样，逆设计等任务中展现了前所未有的能力。具体内容请参考论文原文。

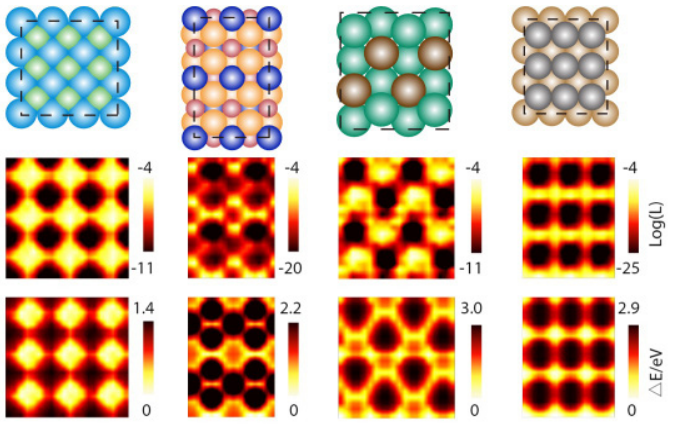


图 4：单个 N 原子和 O 原子在催化剂表面的吸附预测结果。模型预测的催化剂表面吸附质吸附概率分布与量子化学计算得到的相互作用能分部对比图。

DiG 是如何工作的?

类似于模拟退火过程的模式，DiG 通过使用 Graphormer 模型预测一个扩散过程，将简单分布转换为复杂分布。简单分布通常是标准高斯分布，复杂分布则是分子结构的平衡分布。转换是一步一步进行的，如此建模复杂分布的难度便被拆解到每一步成为较为简单的问题。

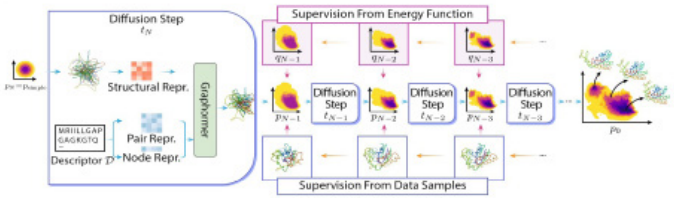


图 5：DiG 的设计和骨干架构

DiG 可以使用不同类型的数据或信息来进行训练。DiG 首先可以使用模拟数据，例如分子动力学轨迹，来学习分布。DiG 也可以直接使用分子系统的能量函数来训练，因为平衡分布可通过统计力学理论直接由能量函数给出。由于分子体系平衡分布预测不同于传统 AI 任务，其数据生成需要耗费长时间的模拟计算因而难以大规模得到，直接从能量函数学习便是一个缓解对数据严格依赖的手段。

DiG 在许多分子系统上都显示出与基于深度学习的结构预测方法相似的良好泛化能力。这是因为 DiG 继承了先进的深度学习架构，如 Graphormer 的优势，并将其应用于一个新的、具有挑战性的分布预测任务。训练好后，DiG 可以通过反转转换过程来生成分子结构，从一个简单的分布开始，并以相反的顺序调用深度学习模型。DiG 还可以通过计算转换过程中概率的变化来提供每个生成结构的概率估计。可以看到，DiG 是一个灵活而通用的框架，可以处理不同类型的分子系统和描述符。

未来，为分子科学研究开辟更多新机遇

DiG 是从单一结构预测到对平衡分布整体建模的重大进展，为在深度学习框架下连接微观结构和宏观属性奠定了基石。DiG 使用生成式 AI 技术，可以在多种分子系统中对符合平衡分布的分子结构进行采样。研究员们在包括蛋白质在内的不同类别的分子上展示了 DiG 的灵活性，同时也证明了以这种方式生成的单一结构是符合物理化学相互作用规律的。

然而，要获得对任意分子系统平衡分布更精准的预测，仍需要进行更多的研究。微软研究院希望 DiG 能够沿着这一方向激发更多的研究与创新，期待未来能够看到 DiG 和其他方法在分子平衡分布预测问题上带来更多令人兴奋的成果和影响。

相关链接：

DiG 论 文：Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning
https://www.microsoft.com/en-us/research/publication/towards-predicting-equilibrium-distributions-for-molecular-systems-with-deep-learning/

Demo 页面
https://distributionalgraphormer.github.io

KDD Cup 2021 | 微软亚洲研究院 Graphormer 模型荣登 OGB-LSC 图预测赛道榜首
https://www.msra.cn/zh-cn/news/features/ogb-lsc

公开催化剂挑战赛冠军、通用 AI 分子模拟库 Graphormer 开源！
https://www.msra.cn/zh-cn/news/features/graphormer

微软研究院团队获得首届 AI 药物研发算法大赛总冠军

AI 药物研发是人工智能未来应用的重要方向之一。自新冠病毒（SARS-CoV-2）首次爆发以来，新冠病毒的小分子药物研发备受关注，于近期举行的首届 AI 药物研发算法大赛便聚焦于此。在比赛中，来自微软研究院科学智能中心的团队，凭借创新的 AI 模型系统 AI2BMD 和 ViSNet 取得了绝佳的成绩，斩获桂冠。

近日，由清华大学药学院、百度飞桨、百度智能云和临港实验室联袂主办的首届 AI 药物研发算法大赛公布了比赛结果，来自微软研究院科学智能中心的团队，利用研发的量子精度动力学模拟系统 AI2BMD 和通用分子三维结构网络 ViSNet 在初赛、复赛、决赛中均位列第一，并获得大赛的总冠军，展现了 AI 在促进药物研发方面的应用潜力。

本次大赛由中国药学会等业内权威机构鼎力支持，共有来自全球的 878 支团队参赛。作为一场全球性的技术创新活动，此次大赛聚焦于新冠病毒（SARS-CoV-2）小分子药物研发。事实上，自新冠病毒首次爆发以来，新冠病毒的小分子药物研发就备受关注。若要抵抗新冠病毒肆虐，深入了解病毒复制与感染机制至关重要。其中，新冠病毒主蛋白酶（Mpro）作为关键酶，负责感染过程中剪切病毒产生的蛋白质前体，促进病毒复制，所以主蛋白酶是一个潜在的治疗靶点，抑制其活性可有效干扰病毒的复制过程，为治疗方法提供突破口。因此，本次比赛的初赛阶段，参赛者需要使用深度学习、分子对接等方法进行建模，预测小分子抑制主蛋白酶活性的概率，复赛则重点关注小分子在 Caco 细胞上抑制新冠病毒复制的概率。

在初赛对新冠病毒主蛋白酶的药物预测中，面对常用分子对接软件无法有效区分正负样本与靶点蛋白结合自由能的问题，微软研究院科学智能中心团队利用了最新开发的 AI2BMD 模拟系统，将药物预测精度显著提升。AI2BMD 模拟系统实现了对超 10000 原子的各种蛋白质能量和力的精确计算，并具有广泛的适用性。相较于密度泛函理论（DFT），AI2BMD 模拟系统的计算时间缩短了数个数量级。凭借几百纳秒的动力学模拟，AI2BMD 展现了在探索蛋白质构象空间、预测核磁共振实验数据以及模拟蛋白质折叠过程等方面的卓越能力。与传统分子对接、经典动力学模拟方法相比，AI2BMD 在计算结合自由能方面也有明显优势。

微软研究院科学智能中心团队获得首届 AI 药物研发算法大赛冠军

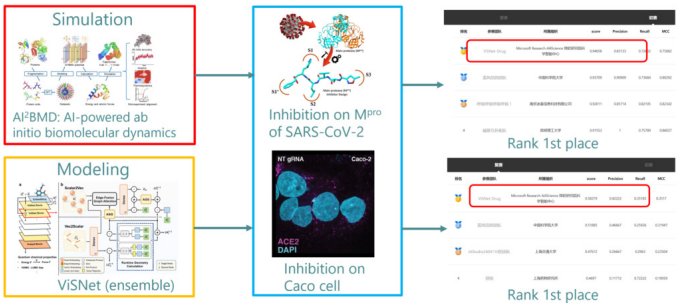
AI2BMD 模拟系统论文链接：https://www.biorxiv.org/content/10.1101/2023.07.12.548519v1

复赛中，团队运用自主开发的分子建模几何深度学习模型 ViSNet 化合物分子进行了表征学习。ViSNet 是 AI2BMD 模拟系统中的机器学习势能函数。作为一种等变的几何增强图神经网络，ViSNet 能在线性计算的复杂度下提取几何特征（距离、角度、二面角等）。在多个分子动力学基准（包括 MD17、rMD17 和 MD22）上，ViSNet 表现均优于其他先进方法，同时也在 QM9 和 Molecule3D 数据集上实现了卓越的量子化学性质预测。

分子建模几何深度学习模型 ViSNet 论文链接：https://arxiv.org/abs/2210.16518v1

团队在复赛阶段，还利用自主研发的首个蛋白大分子全构象空间数据集 AIMD-Chig 和小分子公开数据集 OGB 分别对蛋白和小分子的三维结构表征进行了预训练，然后通过多任务学习对模型进行微调。该方法不仅取得了最佳的预测精度，而且以大比分领先比赛的第二名团队。在最终的决赛答辩中，微软研究院科学智能中心团队的新冠药物预测算法方案取得了总分 99.60 分的绝佳成绩，相较比赛亚军 90.76 分、季军 85.31 分的最终成绩具有显著优势。

蛋白大分子全构象空间数据集 AIMD-Chig 论文链接：https://www.nature.com/articles/s41597-023-02465-9



微软研究院科学智能中心团队提出的新冠药物预测算法方案

通过此次药物研发大赛，微软研究院科学智能中心开发的量子精度动力学模拟系统 AI2BMD 展现了出色的实际应用潜力。未来，AI2BMD 有望在生命活动的分子机理解释、药物设计、酶催化等方面进行更广泛的探索，助力 AI 药物研发的加速发展。

知识产权、隐私和技术滥用：如何面对大模型时代的法律与伦理挑战？

如今，随着 GPT 技术的发展和创新，信息化世界的每一个角落都在以自身高速的变化印证着大模型时代的到来。想象中未来人工智能（AI）改变世界的场景，正逐渐从缥缈的远方走近我们身边。然而，迈入现实的人工智能也同时落入了纷繁复杂的人类社会，它不仅是技术工具，也将作为一个社会对象影响着你我。如何与这一当今世界最具革命性但也蕴含着最多挑战的工具相处，怎样科学地看待、解决人工智能在社会维度上的挑战，是摆在全人类面前的重要课题。

为了洞悉人工智能发展所带来的新问题、新挑战，更好地为世界打造负责任的人工智能，微软亚洲研究院在 2023 年特别组织了“社会责任人工智能（Societal AI）”系列研讨会，让计算机领域的科研人员与国内外高校及研究机构的社会科学领域的专家学者，共同深入探讨人工智能在开发、部署和应用中产生的，包括法学、心理学、社会学在内的跨学科前沿问题。

法律，既与传统社会伦理有着密切渊源，也体现了政策制定者对新兴技术发展持有的态度。在中国人民大学法学院副教授郭锐的大力支持与协助下，微软亚洲研究院举办了“社会责任人工智能”系列研讨会的法律与伦理专题讨论。研讨会上，来自法律和计算机领域的顶尖专家们聚焦探讨了大模型与知识产权、大模型与隐私、大模型的技术滥用问题等人工智能发展所带来的与法律规范和社会伦理相关的问题，以期在这个最为紧迫且关键的话题上引发更多深入思考与探索。

大模型与知识产权：生成内容的保护

从 2023 年年初开始，大模型的重大技术突破吸引了全世界的目光；但与此同时，与大模型有关的知识产权纠纷也开始走进公众的视线。大模型对于现有知识产权法律的挑战，是技术快速发展和应用所带来的最直接的影响之一。

在微软亚洲研究院“社会责任人工智能”系列研讨会的法律与伦理专题讨论上，来自日内瓦大学数字法学中心的 Jacques de Werra 教授指出，目前透明度在版权生态系统中正变得愈发重要。由于目前的知识产权只保护人类作者创作的作品，披露创作中非人类作者来源的部分是必要的。为了应对这一问题，法律和技术两方面的解决方案都应被考虑在内。香港大学的孙皓琛教授认为确定 AI 生成内容的独创性门槛对于讨论 AI 生成内容是否需要被版权法保护是至关重要的。

这也就要求人们进一步区分辨识 AI 生成的内容和 AI 辅助产生的内容，尤其是在二者之间的界限日益模糊的今天。而白盒方法是针对这一问题一个具有潜力的解决方案。因此，接下来应当予以关注的关键问题便是：有哪些白盒方法能够用可解释的方式实现内容生成过程的全透明和披露？

显然，大模型在知识产权上陷入的纠纷已经提示人们考虑如何保障用于大模型开发的作品的人类创作者的权利。清华大学人工智能研究院常务副院长孙茂松教授和莱斯大学的胡侠副教授认为，大家需要通过全球对话与合作的方式来找到更有效的解决方案，来自动识别和解释内容中是否包含有人类创造力。若要达成大模型相关的知识产权问题的共识，有必要制定国际公认的规则，力求在尊重知识产权持有者的权利、公共利益和合理使用例外情况之间达到平衡。

大模型和隐私：尊重隐私，保障安全，促进开放

让一个大模型运行起来，需要使用海量的文本语料进行学习，而在这个过程中大模型使用的是无监督学习方式对大量的文本数据进行预训练。仅 GPT-3 的参数量就达到了 1750 亿，其训练语料达到了 45 TB（文本）。用于大模型训练的这些文本数据来自于互联网的各个角落，包括但不限于书籍、文章、百科、新闻网站、论坛、博客等等，凡是互联网上可以找到的信息，几乎都在其学习之列。即便科研人员会对语料进行数据清洗，但其中仍有可能包含个人的隐私信息。

不论是大型语言模型（Large language models, LLMs）还是图像生成模型，大模型都会记住训练所使用的样本，可能会在无意中泄露敏感信息。因此，苏黎世联邦理工学院的 Florian Tramèr 教授认为，当前的隐私保护技术方法，如数据去重和差分隐私，可能与人们对隐私的普遍理解并不完全一致。所以，应该在微调阶段纳入更严格的保障措施，以加强对于数据隐私的保护。

研讨会上，各位专家明确了大模型存在隐私风险的三个方面：互联网数据训练、用户数据收集和生成内容中的无意泄露。这其中首先需要确保公共数据是不具有个人可识别性的，并与私人或敏感数据明确区分开来。未来应重点关注算法的透明度和对个人信息主体的潜在伤害问题。

其实，对于隐私的保护和大模型的效率之间存在着一个两难的矛盾 —— 既要最大限度地保护数据隐私，又要最大限度地发挥模型的功效。微众银行人工智能首席科学家范力欣博士和微软亚洲研究院高级研究员张辉帅一致认为，人们需要通过协作开发一个统一、可信的框架，从而在隐私保护、模型效用和训练效率之间取得一种平衡。

美国科文顿 • 柏灵律师事务所（Covington & Burling LLP）的罗嫣和微软公司法律顾问丁倩强调，在大模型开发过程中面临的数据隐私问题上，要确保遵守现行法律法规的规定，并充分评估隐私数据的使用对个人信息主体的影响，采取有效措施防止可能带来负面影响。另外，在确保透明性的基础上，鼓励个人信息主体同意分享隐私数据，以解决我们共同面对全球重大问题。这样才可以确保负责任地开发和安全地利用人工智能，进而带来更加广泛的社会效益。

大模型和技术滥用问题：边缘群体的数字平等

当大模型在技术和社会中扮演起越来越关键的角色时，它能否承担起相应的责任？如何促进负责任的人工智能进步并确保其在价值观上与人类价值观相一致？这些宏观的问题十分棘手，但也十分迫切，因为大模型一旦遭到滥用，其强大的效用和能力有可能反过来损害社会的利益。

微软亚洲研究院资深首席研究员谢幸认为，负责任的人工智能需要技术和社会学两方面的策略双管齐下，而且有必要将大模型与多样化、个性化以及特定文化的人类价值观结合起来，达到一致。

这其中对于边缘群体（尤其是残障人士）的数字平等问题需要更加关切。AI 技术可能产生错误陈述和歧视，使得对残障人士

的歧视被制度化。因此，AI 开发者必须注意不要让残障人士与 AI 产生角色和利益上的冲突，开发者有责任去主动对抗那些有偏见的态度，倡导平等参与，提高平等意识。

哈佛大学的崔凤鸣博士和一加一残障公益集团的蔡聪在研讨会上强调了数字平等问题所包含的两个关键维度：其一是“赋能”，要让 AI 设备的价格可以被边缘群体所承受，并为他们提供适当的培训；其二是“包容”，要将对于边缘群体的关注整合到人工智能从模型设计到数据创建的整个开发过程中，这样才能打破壁垒，消除歧视。

欲了解本次法律与伦理专题研讨会的更多详细信息，请查看链接：The Workshop on Legal and Ethical Governance Challenges Faced by Big Models

https://www.microsoft.com/en-us/research/event/2023-legal-and-ethical-governance-challenges-faced-by-big-models-workshop/

从大模型对知识产权和隐私保护产生的冲击，到可能的技术滥用风险，伴随技术快速发展所带来的诸多挑战，一种更负责任、更为健全的人工智能治理规范也在成长之中。人工智能的法律问题本身并不是孤立存在的，它涉及到复杂的传统社会伦理观念，也涉及到政策可能给与人工智能技术的发展空间。为了让讨论和思考真正有益于 AI 与社会的和谐相处，在此次针对法律和伦理的充分交流之后，微软亚洲研究院还将深入法律、伦理和政策制定的更深层次，并拓展心理学和社会学等领域的探索。

大模型时代，如何评估人工智能与人类智能?

随着人工智能（AI）应用的不断落地，AI 之于人类的角色也在悄然改变。人们对于人工智能的期待和看法，从完成特定任务的机器转向了真正的智能伙伴。然而，这些新伙伴所具有的复杂性和未知性却是人们前所未见的，因此大模型的测评工作成为了当下亟待解决的关键问题。人类若要深入地理解这些复杂且高度智能的模型，就需要心理学及教育学等涉及认知能力内在的研究领域与计算机科学合力探索。

在心理测量领域，研究者们已将对人类能力的深刻理解和洞察进行了汇集，并提供了丰富的理论模型以及对其进行有效测评的方法，这些都能够为人工智能的评估和进一步发展提供启示。近期，在北京师范大学心理学部骆方教授的大力支持与协助下，微软亚洲研究院举办了“社会责任人工智能（Societal AI）”系列研讨会的心理与教育专题讨论。研讨会上，来自心理测量领域、教育领域以及计算机领域的顶尖专家们共同探讨了心理测量技术应用与人工智能测评的可行性、大模型如何赋能心理测评，并展望了人工智能辅助下的未来教育。

人类及大模型的能力评估：汇聚与整合

目前，人工智能领域的传统评估对象是为特定任务设计和构建的 AI 模型，如机器翻译模型等，模型评估即评估这些 AI 模型在这一特定任务上的表现。然而，新一代的人工智能并不是为了执行某一个特定任务而设计的，它们能够广泛地模拟人类智慧，胜任多样化的任务，比如 ChatGPT。因此，基于单一任务表现的传统评估方式不再适用于新一代人工智能模型。

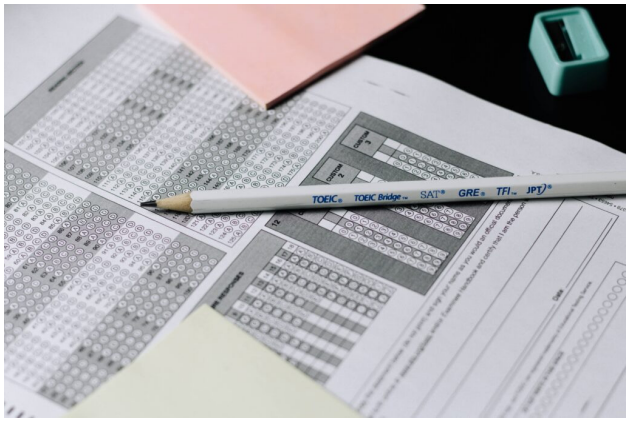
面对这样的现状，来自瓦伦西亚理工大学的 José Hernández-Orallo 教授判断，已有的评估方法“任务导向的评估”（task-oriented evaluation），现在需要转向“能力导向的评估”（capability-oriented evaluation），把评估的重点从衡量某项特定任务的表现，转向这些表现背后的思维能力。

尽管对于大模型而言，能力测评是一个全新的议题，但对于人类的能力测评却有着悠久的历史。心理测量学认为人类所具有的某种潜在特质会导致一系列典型行为，测量这些典型行为的目的便是获取对能力的估计，而能力又可以进一步预测在更多相关行为上的表现。因此，心理测量领域的研究者们多年来以人类为研究对象，研究如何量化人类的不可见心理特质，如思维能力，并将它们与可观察的行为表现进行关联。José Hernández-Orallo 教授认为，为了实现对大模型能力的准确评估，人工智能领域的科研人员需要做好进行“范式转变”的准备，把心理测量方法融入到 AI 评估中。

然而，在将基于人类智能所构建的心理测评方法应用于大模型的能力时，必须要十分审慎。José Hernández-Orallo 教授指出，“心理测量学的方法和技术需要根据 AI 的特点进行有针对性的调整，相关结果的意义和解读可能也需要重新思考”。来自卑尔根大学的 Marija Slavkovik 教授也认为，人工智能并不是完全模仿人类智能，所以计算机完成任务所采用的是与人类不同的方式。

奥本大学的范津砚教授提到，相比于能否将心理测验应用于测评大模型这一问题本身，观念和思维方式的转变更为重要，通过心理测评的视角去探索全新的人工智能测评道路是十分有意义的。而这要求计算机科学和心理学的科研人员要共同探索适用于 AI 能力测评的新范式，不断地去发现和解决那些前所未有的新问题。

此外，圣加仑大学的 Clemens Stachl 教授还指出，之前已有的测验很可能已经出现在了大模型的训练数据中，因此研究者应该关注大模型在那些全新测验上的表现，即考察大模型是否具备应对和解决新问题的能力。



大模型辅助下的未来教育及测评

在心理测量领域既有的知识经验不断为人工智能领域带来深远影响的同时，新兴的人工智能技术也给测量领域带来了新的思路和启示。

传统心理测量学更多采用的是自上而下的思路 —— 基于理论构建某种比较简单的统计模型，然后获取结构化的数据，以验证模型的有效性。而在人工智能技术充分发展的当下，各种在线学习平台提供了丰富的学习资料、学习场景以及交互形式。牛津大学的 Alina A von Davier 教授提出，未来的学习和测评系统将会具有数字化、自适应、个性化以及沉浸感等特点，且人们可获取的学习及测验数据也愈发丰富，包含了语音、视频等，因此研究者需要思考如何建模这些多模态数据。在这一背景下，测量领域也需要转变范式，探索人工智能辅助下的测评模式。

Alina A von Davier 教授提到，人工智能在测验的编制、施测、评分以及结果报告整个过程中的每一个环节都能够发挥重要作用，但各个环节中仍需要人类专家进行监控和决策，每个环节都应该是人工智能和人类智能协作的结果。针对编制环节，来自剑桥大学的 David Stillwell 教授分享了尝试采用大模型自动编制测验题目的经验，他认为大模型能够帮助研究者想出更多元、更丰富的题目情境，从而提高测验的编制效率。然而目前大模型生成的题目质量还够不理想，需要人类专家进行细致的筛查。Clemens Stachl 教授则表示，大模型在实现自动化测量上具有情境，但其可信度和有效性以及透明度等问题则会构成挑战。

Alina A von Davier 教授的团队目前已经尝试将 AI 技术融入心理测量中，并提出了 Digital-First Assessment 这一新型的测验方法。Digital-First Assessment 基于数字化环境设计，可提供交互式的操作和功能，利用人工智能算法辅助进行测验的生成和分发，在自动采集被试多模态的过程性数据后，再结合心理测量理论进行分析和解读，从而保证了测评结果的有效性和可靠性。Alina A von Davier 教授认为这种融合有望成为人工智能时代下，心理及教育测评的主要形式。

当大模型应用进入教育领域，学校、家长以及社会都在担忧由其引发的一些有损教育初衷和公平性的情况，如学生使用大模型完成作业和考试等。来自中国科学技术大学的研究员朱孟潇认为科研人员需要思考如何识别和避免这些情况的发生，包括对异常作答的检测，但更为重要的是思考如何重新设计作业和评估的形式。对此，范津砚教授提供了一个思路 —— 过程性评估，即不像以往那样完全关注结果，而是更关注产出结果的过程，根据过程来反映被试的能力。

大模型的出现和应用直接推动了教育观念的改变，促使人们重新思考未来教学和评估的焦点。Marija Slavkovik 教授认为大模型的出现让大家开始反思如今的教育是否在培养和评估学生的能力而非特定知识，但这实际上是教育本就需要思考的问题，是大模型的出现增加了这个问题的紧迫性。来自北京师范大学的卢宇教授强调，如今我们比以往更需要强调高阶思维能力的培养和测评。José Hernández-Orallo 教授则指出了一个更具有前瞻性的方向：评估人类与人工智能的共同体（the hybrid of human and AI system），即评估个体能否利用人工智能工具来更好地解决问题。孟菲斯大学的胡祥恩教授认为，新一代 AI 代表了数字化的文明，人们需要具备与它们合作的能力。面对一个 AI 无处不在的未来世界，社会各界必须帮助下一代在这个世界的生存和发展做好准备。

欲了解本次研讨会的更多详细信息，请查看链接：The Workshop on Understanding and Evaluating Big Models for Human Intelligence and Learning

https://www.microsoft.com/en-us/research/event/the-workshop-on-understanding-and-evaluating-big-models-for-human-intelligence-and-learning/

大模型与心理测量的结合预示着一场划时代的变革。心理测量学可以帮助人们深刻、透彻地理解和挖掘大模型的真实能力。与此同时，大模型也将成为心理及教育测评研究者深度合作的伙伴，通过将 AI 技术融入心理及教育测量的全过程之中，心理及教育测评领域将能够实现个性化、自动化且沉浸式的评估。可以预见的是，一旦大型语言模型与心理测量技术结合的巨大潜力被激发，一个更为智能、开放和人性化的教育新纪元将会成为现实！

AI 将怎样影响人类社会?

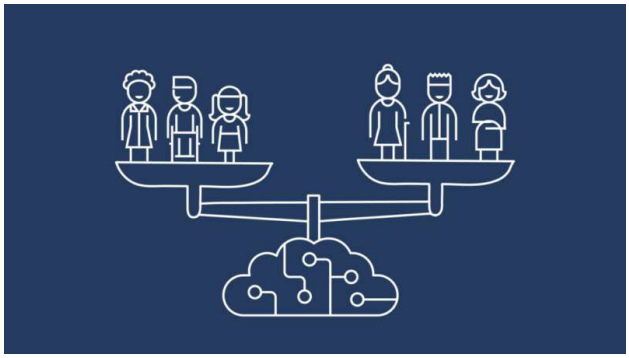
人工智能（AI）将会对人类未来产生重大的影响，这一点现在已经变得越来越清晰和明确。但对于这种影响到底涉及哪些方面，其中有哪些值得期待和着手准备的，又有哪些潜在的风险需要担忧，却是一个尚未被充分重视的问题。如同人类技术史上那些划时代的变革一样，AI 一旦真正走入世界发展的进程，对于社会的影响将是全方面的：从社会分工，到经济结构、产业形态，以及科学技术研究和青少年教育 ……

为了洞悉人工智能发展所带来的新问题、新挑战，更好地为世界打造负责任的人工智能，微软亚洲研究院在 2023 年特别组织了“社会责任人工智能（Societal AI）”系列研讨会，让计算机领域的科研人员与国内外高校及研究机构的社会科学领域的专家学者，共同深入探讨人工智能在开发、部署和应用中产生的，包括法学、心理学、社会学在内的跨学科前沿问题。

人类社会是一个诸多领域环环相扣的精微整体，任何一个部分的变化都可能影响你我每一个人。在普林斯顿大学黄俊铭博士的大力支持和协助下，微软亚洲研究院举办了“社会责任人工智能”系列研讨会的社会与科技专题讨论。研讨会上，来自社会科学和计算机领域的专家们聚焦探讨了人工智能影响下的数字平等、社会公平、经济结构变化等面向未来发展趋势的重大问题，努力为这个尚在探索阶段的议题提供严肃的思考。

AI 引发的社会公平问题

在 AI 兴起所引发的种种担忧当中，最为广泛的一种便是对于 AI 大规模应用加剧社会不平等现象的隐忧。我们都能预见到 AI 应用所带来的“马太效应”：AI 竞争当中的强者不断累积资源，优势也就不断扩大；而落后者似乎会面临越来越被动的窘境。



普林斯顿大学的谢宇教授指出，在人工智能发展的起步阶段，不同国家的规模和体量可能会构成一种原始优势。例如，网民群体庞大的国家，就具有很大的数据优势。来自北京交通大学的桑基韬教授则表示，越早采用新兴技术的一方会更有机会启动正反馈循环。

相比其他科技领域，AI 领域的资源更为集中，目前主导整个行业的企业和国家仍属于少数。针对这一情况，卡内基·梅隆大学的 Patrick Park 教授认为，未来需要深入思考人们是应该追求技术的普及，将技术置于大众的手中，还是应该加强管制以防范技术滥用带来的风险？这两条道路各有优劣，也都会为社会公平议题带来新的挑战，因此需要各界共同探索不同的应对方式。

AIAI

AI 的全面应用将很有可能在未来彻底重塑我们的工作方式。2023 年微软推出的“智能副驾”Microsoft 365 Copilot 就直接带来了办公软件领域的新变革。来自纽约大学的程思薇教授强调，人工智能对于人们工作内容的直接影响将不仅限于科技领域，而是涉及各行各业。

桑基韬教授对此进行了补充，表示人工智能的应用会使得那些例行性的任务变得更加次要，一些工作将会由 AI 来完成，而这时，人们就可以更加专注于更高层次的规划和分析工作。由此可以推测，未来的职场将会更加重视发现和解决问题的能力、创造力以及批判性思维，还有主动学习和获取新技能的能力。

谢宇教授的观点更加强调人类的多样性。他表示，那些需要个体独特性的服务在未来将变得更受欢迎，比如，尽管老师上课讲授内容的工作可能会被自动化，但师生关系却是不能被取代的。

信息安全和社会观念

如今人们生活在一个被各类信息包围的社会当中，这也意味着 AI 强大的信息处理能力将会对信息安全构成潜在威胁。AI 可以通过内容生成直接改变我们所看到的内容，影响我们所接触信息的真实性，乃至通过信息手段影响每个人的社会观念。不可否认，像其他信息处理方式一样，AI 不仅会产生错误信息，而且也会受到其他错误信息的影响。

桑基韬教授将展望的视野推至了更远的方向：在未来，随着

AI 的发展，互联网上的内容会有越来越多是由 AI 生成的，模型训练的数据中 AI 生成的数据比例也将不断增加。也就是说，AI 生成的信息被用来训练 AI 的情况将普遍存在。在这种情况下，如果 AI 使用了以往模型生成的错误信息来进行训练，那么后续的模型可能会不断地放大这些错误信息。面对这样棘手的情况，学术界和工业界都需要思考如何用更加多样的数据来训练模型、在推理过程中增加验证手段，以及增加对于信息可信度的评估来协助人们进行判断。

不论如何，广泛地使用 AI 生成的内容将极有可能潜移默化地影响社会的全体成员。程思薇教授认为，随着人工智能进入到社会观念的建构过程当中，理解人们的观点将会变得更加复杂。来自长江商学院的张维宁则提醒人们注意在这种情况下，内容筛选的重要性，那些可验证的信息将会在公共空间内发挥举住轻重的作用。Patrick Park 表示，在我们用于训练的数据中建立起基础的可信信息将变得越来越重要。谢宇教授则认为，人们更应该在教育过程中强调自主思考的重要性，并充分重视起人文主义方法。

科研与教育会被自动化吗?

由人工智能带来的数字化浪潮的影响范围包含方方面面，其中当然也包括人类探索知识、传递知识的领域。毋庸置疑，人工智能将成为人类科学技术研究过程中的有力工具 —— 促进科研的数字化进程、扩大实验范围、产生新的想法，从而达到推动科学探索的目的。来自哥本哈根信息技术大学的 Anna Rogers 提到，AI 模型还可以被用于更有效地表达科研结果，有助于非英语母语的科研人员更好地参与研究工作。

加州大学的 Bernard Koch 描述了近年来机器学习研究社区的变化，科研人员们正倾向于采用更少的基准来衡量人工智能的进展，并逐渐缩小了数据集的规模。而这可能会引发一系列问题，包括模型在特定数据集上过拟合、提出的认知问题变得更狭窄、数据集不代表完整分布时会出现的伦理风险，以及选择基准问题的参与者变得更少，等等。对于科学研究来说，提出与众不同的问题是十分重要的，但人工智能的介入可能会带来研究领域变得认知范围过窄的风险。与会者们认为，人们应该更专注于开发那些人工智能与人类互补的能力。

与此同时，AI 也可能会改变人们对教育内容、教师角色以及教育本身价值的看法，因此将对教育系统产生深远的影响。来自芝加哥大学的 James A. Evans 观察到，其实教育的很多内容是保持不变的，尽管顶尖大学的研究多种多样，但讲座和研讨会的内容通常可以追溯到课程首次教授时。大模型的到来将使得教育者可以更好地认识到此前的忽视，并充分利用技术来创建更具吸引力的教育过程。

欲了解本次研讨会的更多详细信息，请查看链接：The Workshop on AI's impact on Society and Advancements in Technology

https://www.microsoft.com/en-us/research/event/the-workshop-on-ais-impact-on-society-and-advancements-in-technology/

人工智能对于这个世界的深远影响才刚刚开始显露，它到底会走向何方还需要我们耐心的观察并持续关注。为了迎接具有全新可能性的智能未来，社会各界都需要做好各方面的准备：不论是对社会公平的守护，还是对经济秩序的维护以及对科教人文的保护，还有许多工作有待完善。微软亚洲研究院将继续以最前沿的视野和跨学科的广度，建构“社会责任人工智能”，为快速变革的世界注入微软力量。

如何评测一个大语言模型?

大型语言模型（Large language models, LLMs）因其在学术界和工业界展现出前所未有的性能而备受青睐。随着 LLMs 在研究和实际应用中被广泛使用，对其进行有效评测变得愈发重要。近期已有多篇论文围绕大模型的评测进行研究，但尚未有文章对评测的方法、数据、挑战等进行完整的梳理。日前，微软亚洲研究院的研究员们参与完成了介绍大模型评测领域的第一篇综述文章《A Survey on Evaluation of Large Language Models》。该论文一共调研了 219 篇文献，以评测对象 (what to evaluate)、评测领域 (where to evaluate)、评测方法 (How to evaluate) 和目前的评测挑战等几大方面对大模型的评测进行了详细的梳理和总结。研究员们也将持续维护大模型评测的开源项目以促进此领域的发展。

通俗来讲，大模型是一个能力很强的函数 f，与之前的机器学习模型并无本质不同。那么，为什么要研究大模型的评测？大模型评测跟以前的机器学习模型评测有何不同？

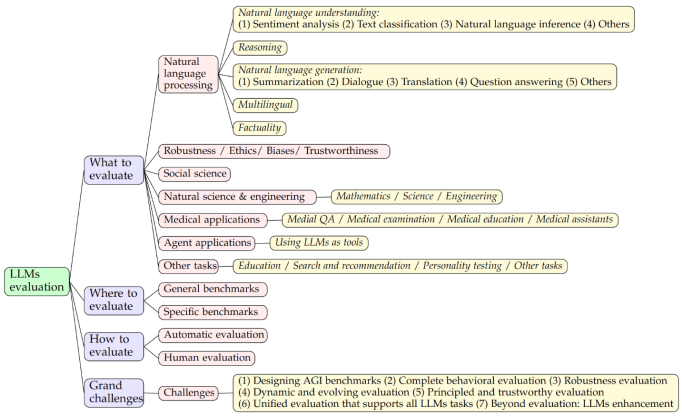
首先，研究评测可以帮助我们更好地理解大模型的长处和短处。尽管多数研究表明大模型在诸多通用任务上已达到类人或超过人的水平，但仍然有很多研究在质疑其能力来源是否为对训练数据集的记忆。如，人们发现，当只给大模型输入 LeetCode 题目编号而不给任何信息的时候，大模型居然也能够正确输出答案，这显然是训练数据被污染了。

其次，研究评测可以更好地为人与大模型的协同交互提供指导和帮助。大模型的服务对象终究是人，那么为了更好地进行人机交互新范式的设计，我们便有必要对其各方面能力进行全面了解和评测。如，我们最近的研究工作 PromptBench：首个大语言模型提示鲁棒性的评测基准，便详细地评测了大模型在“指令理解”方面的鲁棒性，结论是其普遍容易受到干扰、不够稳定，这便启发了我们从 prompt 层面来加强系统的容错能力。

最后，研究评测可以更好地统筹和规划大模型未来的发展的演变、防范未知和可能的风险。大模型一直在不断进化，其能力也越来越强。那么，通过合理、科学的评测机制的设计，我们能否用演化的角度来评测其能力？如何提前预知其可能的风险？这都是重要的研究内容。

因此，研究大模型的评测具有十分重要的意义。

我们于近期完成了介绍大模型评测领域的第一篇综述文章《A Survey on Evaluation of Large Language Models》。该论文一共调研了 219 篇文献，以评测对象 (what to evaluate)、评测领域 (where to evaluate)、评测方法 (How to evaluate) 和目前的评测挑战等几大方面对大模型的评测进行了详细的梳理和总结。其研究目标是增强对大模型当前状态的理解，阐明它们的优势和局限性，并为其未来发展提供见解。同时，我们也将该项工作进行了开源，希望有更多同行参与，共同促进该领域的发展。



研究框架

相关链接：

论文链接：https://arxiv.2307.03109

GitHub 链接：
https://github.com/MLGroupJLU/LLM-eval-survey
https://github.com/microsoft/promptbench

大模型评测相关研究：https://llm-eval.github.io/

扫描二维码查看完整内容

科研第一线

科研上新 | 强可控视频生成；定制化样本检索器；用脑电重建视觉感知；大模型鲁棒性评测

“科研上新”汇聚了微软亚洲研究院最新的创新成果与科研动态。在这里，你可以快速浏览研究院的亮点资讯，保持对前沿领域的敏锐嗅觉，同时也能找到先进实用的开源工具。

本期内容包括

01. 强可控视频生成模型 DragNUWA

02. LLM Retriever：通过定制化样本检索器来提高大语言模型的上下文学习效果

03. 解码大脑信号重建视觉感知图像

04. PromptBench：首个大语言模型提示鲁棒性的评测基准

扫描二维码了解更多信息

ICML 2023 | 拓展机器学习的边界

ICML
International Conference
On Machine Learning

如今，机器学习已成为人类未来发展的焦点领域，如何进一步拓展机器学习技术和理论的边界，是一个极富挑战性的重要话题。7 月 23 日至 29 日，第四十届国际机器学习大会 ICML 2023 在美国夏威夷举行。该大会是由国际机器学习学会（IMLS）主办的年度机器学习国际顶级学术会议，旨在推动机器学习领域的学术进步。在 ICML 2023 上，微软亚洲研究院的研究员们有多篇论文入选，本文为大家简要介绍了其中的 5 篇。

扫描二维码了解更多信息

ACL 2023 | 持续进化中的语言基础模型

尽管如今的 AI 模型已经具备了理解自然语言的能力，但科研人员并没有停止对模型的不断改善和理论探索。自然语言处理（NLP）领域的技术始终在快速变化和发展当中，酝酿着新的潮流和突破。

NLP 领域的顶级学术会议国际计算语言学年会（简称 ACL）是关心 NLP 领域的研究者们观察动向的关键窗口。在 2023 年 7 月 9-14 日于加拿大多伦多举行的 ACL 2023 上，微软亚洲研究院有多篇针对基础模型的论文被 ACL 2023 收录，本文介绍了相关的 4 篇前沿工作。

扫描二维码了解更多信息

ACL 2023 | ｜大模型时代，自然语言领域还有什么学术增长点?

随着人工智能技术的快速发展，确保相关技术能被人们信赖是一个需要攻坚的问题。微软亚洲研究院也在不断推进负责任的人工智能的探索发现与应用实践。本文为大家带来了 3 篇微软亚洲研究院以负责任的人工智能为主题入选 ACL 2023 的论文。

扫描二维码了解更多信息

科学匠人｜边江：在微软亚洲研究院的七年“技痒”，探寻大模型助力 AI 与产业融合之道

基础科研的创新为技术落地应用提供了动力，而来自真实世界的业务需求则为基础科研提供了灵感和方向。当人工智能进入大模型时代，什么样的技术创新才能更好地落地于产业？对此，微软亚洲研究院资深首席研究员边江有着切身的体会和独特的见解。自 2016 年重回微软亚洲研究院后，边江一直专注于人工智能的快速开发和应用，助力行业企业提高生产效率，加速智能转型。在加入微软亚洲研究院的第七个年头，边江将面临哪些新“技痒”？

身处第四次工业革命的进程中，机器学习、深度学习等人工智能技术已成为公认的核心驱动力。近几年，在各大研究机构的努力下，人工智能技术已取得众多突破性成果，并在产业中落地生花。

多年来，微软亚洲研究院在探索计算机基础科研创新的同时，也在持续推动人工智能技术与现实产业场景的融合，通过与各行业合作伙伴的联合研究，目前已经让 AI 在智慧金融、物流运输、医疗健康、能源与可持续性发展等行业场景中得到应用。这些项目的成功，离不开微软亚洲研究院资深首席研究员边江和同事们多年的努力。从出走创业，到回归七年，边江一直奔走在创新技术与产业融合的第一线。

带着创业视角重回微软亚洲研究院

学术界是边江职业生涯的起点。在美国佐治亚理工学院完成计算机科学博士学业后，边江加入了美国雅虎研究院，负责雅虎首页内容推荐和垂直搜索模块的研究和优化工作。但异国的漂泊终究难抵对故土的深情，没多久，他选择了回国发展，加入微软亚洲研究院继续从事推荐算法相关的研究。

2013 年，各行各业掀起了一轮创业热潮。怀揣着“科技改变世界”的梦想，边江短暂地离开了熟悉的科研圈，加入了一家新成立的算法推荐新闻资讯公司，成为其初创团队的一员，将一腔热血投入到产业界中，期望可以将实验室的创新技术应用到现实场景中。

“相比纯粹的学术研究，我更喜欢将技术应用到实际产业的过程。除了发表论文和在公共数据上做实验之外，产业界还要考虑更多的实际问题。做产品需要系统性的思维，包括对数据的考量、KPI 的设定等，比如如何弥补模型优化和产品实际上线之间的差距，这其中存在许多新的挑战。”边江说。

创业的几年间，边江在将机器学习技术应用于产品实践方面积累了丰富的经验。而扎实的学术功底加上创业实战经历，让边江得到了众多科技公司的青睐。但在比较了若干机会之后，边江还是选择了回归微软亚洲研究院。“无论是产业界，还是在学术界，我一直都对前沿技术很感兴趣，微软亚洲研究院一直走在新技术研究与探索的前沿，并已经发展成为了一个横跨产、学、研的机构，给科研人员提供了更广阔的舞台。重新回到这里，让我可以探索新技术从想法到研发，再到产业落地的全过程。而且研究院拥有众多顶尖的人才，我也能够与这些优秀的同事和实习生合作，一起做出更有影响力的研究成果。”边江说。

身处大模型时代浪尖，科研方向也需应势而变

重返微软亚洲研究院后，边江主要聚焦于机器学习、深度学习技术的研究，他和团队陆续将这些技术成功应用到了物流、金融、医疗等行业。例如，在物流领域，他们与东方海外航运（OOCL）合作，通过深度学习和强化学习技术大幅优化了航运网络运营；在金融领域，他们探索出了金融 AI 通用技术平台 Qlib，帮助金融 AI 研究者和从业者使用更先进和多样的人工智能技术来应对更复杂的金融挑战；在医疗领域，他们训练的新冠疫情（COVID-19）预测模型被美国疾控中心采用，其表现优于全球其它四十多家科研机构的预测模型。边江和团队还与医疗健康服务提供商美国哈门那公司（Humana）合作，基于深度学习技术构建了健康预测模型，帮助用户及时了解自身的健康状况，获得个性化的医疗服务。

2023 年是边江回到微软亚洲研究院任职的第七年。在这期间，边江见证了人工智能技术发展与产业化的起伏变化。如今，随着基础大模型的成熟，人工智能赋能产业发展又迎来了新的机遇。面对新一轮的人工智能应用浪潮，边江与团队成员也在重新思考

产业界的新需求，定位新的科研方向，并确定了从智能决策、工业数据智能、差异化隐私安全和智能认知学习四个方向着手，深化人工智能技术的应用研究。

边江（第二排右三）与团队部分成员合影

在之前与东方海外航运的合作研究中，边江和团队使用了多智能体强化学习来优化航运网络，但在实际工业界的智能决策中，外部环境时刻变化，智能体要如何适应这种变化是一个问题。对此，边江团队积极探索环境强化学习（Situational RL），从而让强化学习模型可以更好地适应外界环境的变化。

与此同时，边江与团队还在积极探索将大语言模型（LLM）和强化学习结合，以构建更强大的智能决策能力。一方面，他们在探索通过基于 LLM 的预训练方法构建可以从交互环境中获取常识知识的世界模型（World Model），凭借预训练世界模型的能力，使决策智能体适应不同的工业场景。另一方面，团队也在研究如何集成 LLM 与在线学习的能力，在无需承担大量训练或微调成本的前提下，实现决策智能在工业控制场景中更强的泛化能力。

与纯学术研究相比，将工业数据应用到深度学习时，数据质量和特性等方面都存在较大的差异。比如 COVID-19 和 Humana 预测模型都有独特的数据特性，其中 COVID-19 模型预测的美国各大州数据具有空间关联性，即一个州的新冠确诊和死亡趋势与另一个州有相关性，而在 Humana 模型的数据中，既有病人定期检查信息这样的有规则数据，还有用户吃饭、运动等不规则数据。另外，在工业环境下，还会遇到数据缺失的情况。面对这些挑战，如何利用深度学习挖掘数据中的规律，是边江与团队的重要研究方向之一。

随着基础大模型的成熟，边江与团队正在进一步探索构建具有跨领域数据知识的工业基础大模型，使其能更有效处理工业数据中的数值和结构化信息。借助新的工业基础大模型，边江和团队期望创建工业数据智能与解决方案的新范式。

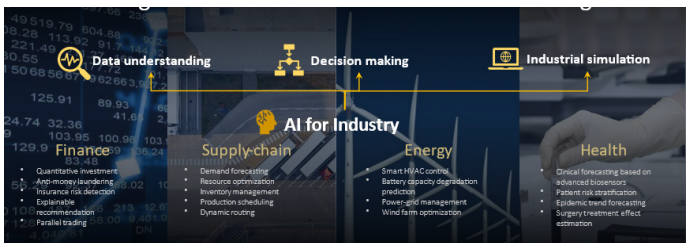
而在与工业界合作伙伴的交流中，边江也发现企业对隐私安全的需求越来越高。隐私安全一直是学术界重点关注的问题，也是微软亚洲研究院持续探索的一个方向。如何在利用好个体数据的同时保护个人隐私也是边江和团队未来关注的重点之一。

另外，微软亚洲研究院机器学习组在智能认知的研究上也取得了诸多成果，包括机器翻译、语音识别、自然语言处理等。边江希望将这些技术融入数字虚拟人（Digital Human）中，创造更逼真的虚拟人。目前，边江团队已在以文本或声音驱动的虚拟人面部口型（talking face）和音乐生成（music generation）方面取得了进展。

以产业创新中心为平台，加速技术与产业融合

除了带领机器学习组推进相关科学研究，边江还是微软亚洲研究院产业创新中心（Industry Innovation Center）的负责人。作为微软亚洲研究院产、学、研融合创新的平台，产业创新中心以研究院前沿研究与技术转化的丰硕成果与丰富经验为基石，聚集了一批既擅长技术创新又具备行业知识的“接地气”的计算机科学家和工程师，致力于发展与企业的战略合作，面向真实世界的关键场景，通过联合技术创新，实现共同发展，引领产业未来。

边江表示，产业创新中心希望从三个方面推动人工智能技术与产业的快速融合：一是在工业应用场景基础上探索最先进的人工智能技术，并将人工智能技术真正应用于工业界，赋能工业界；二是将创新的技术转化成云服务和开源项目，赋能更多的产业客户；三是为微软研究院搭建一个与业界沟通交流的平台，帮助研究人员从产业界中发现新的研究课题，也让产业界了解最新的技术发展动态。



微软亚洲研究院产业创新中心所涵盖的研究领域

具体而言，产业创新中心计划聚焦四个领域。在供应链领域，从物流、零售和制造业切入，利用智能决策进行供需匹配预测，帮助企业实现从生产、仓储，到运输物流的全产业链资源优化。在能源领域，从生产端助力风能和太阳能等清洁能源的生产趋势预测，最大化能源网络效率；从消费端帮助企业减少资源消耗，构建智能楼宇、绿色数据中心等，实现节能减排；并在两者之间，通过电池性能和生命周期预测以及充放电策略等研究，优化新能源储存方式。在金融领域，利用深度学习识别洗钱和欺诈行为，帮助金融机构更好地识别诈骗或洗钱团伙；同时模拟金融投资市场，助力金融从业人员预测极端的市场动荡，做好风险控制。而在医疗健康领域，边江和团队将推动机器学习在慢性病方面的应

用，帮助进行病程预测，例如根据糖尿病病人的饮食、运动行为，预测血糖变化，采用更好的胰岛素治疗方案等。此外，他们还将提升模型的泛化性，将模型推广到更多慢性病的诊断与辅助治疗场景中。

当然，将前沿技术应用到行业中并不是一件容易的工作，对此边江深有体会。基于已有的与行业企业合作的成功经验，边江认为将创新技术与产业更好地融合有两个关键因素：一是需要有大量掌握产业知识的科研人员参与其中，这些科研人员要真正深入到实际的业务场景中，下沉到真实的数据里，了解行业面临的真正问题。同时，因为基础科研和技术落地应用之间的关系并不是简单的线性模式，而是一个复杂的双向生态系统，这就需要行业企业从上至下全面接受智能化转型，支持创新技术与业务的融合，真正做到学术界和产业界的共同创新。

“AI 在学术研究领域已经进入文本、图像、语音等多模态数据的大一统和大模型时代，大模型背后所蕴藏的数据表示能力、知识能力、逻辑推理能力为更强大的数据理解、优化决策、场景生成与模拟提供了新的可能性。而在产业界，经过此前的摸索，很多行业和企业已经建立了自己的数字化平台，并积累了大量的行业数据，为 AI 落地应用提供了基础。这些学术与产业的最新发展趋势预示着 AI 驱动的产业数字化转型将迎来新的爆发阶段。”边江说。

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）。

“在成为微软亚洲研究院的实习生之前，我对研究院的了解就是 ResNet（Residual Network，残差网络）的论文。在这篇论文中，微软亚洲研究院的研究员们首次引入了残差学习的思想，让 ResNet 成为了计算机视觉技术发展的一个里程碑。“黄昶互（Changho Hwang）说。

“前沿的技术研究，顶尖的创新人才”，是黄昶互对微软亚洲研究院的第一印象。

加入微软亚洲研究院，以正确的方式做正确的事

博士学业的第二年，黄昶互在韩国科学技术院（KAIST）导师的推荐下，成为了微软亚洲研究院的一名实习生。在 2018 年寒假与 2019 年暑期两个阶段的实习之后，黄昶互对微软亚洲研究院有了全新的认识，并且确定了自己博士毕业后的职业目标 —— 加入微软亚洲研究院，从事更具前瞻性的技术研究。“当时有同学和同事给我介绍过其他实验室和公司，但是在微软亚洲研究院的实习经历让我非常明确，我更喜欢这里的工作环境和研究氛围，它能够让我专注于自己感兴趣的研究领域。“黄昶互说。

黄昶互表示，微软亚洲研究院最吸引他的一点是始终在以正确的方式做正确的事。微软亚洲研究院不会随波逐流地追逐技术的风口，而是有着独到的战略和研究方向，并一直将目光放在更大的蓝图之上，专注于探索前沿的技术研究。

此外，共事的同事以及研究院多元的研究方向，也是黄昶互选择加入微软亚洲研究院的重要原因。研究院有一群可爱且技术实力深厚的研究员，黄昶互实习期间的导师平易近人，在研究中给予了他自由的研究空间和极大的学术支持，在工作和生活中，同事们也都热情相助，这让身处异国他乡的黄昶互倍感温暖。同时，

在微软亚洲研究院所进行的前沿探索中，不仅有与黄昶互的电气工程专业高度匹配的研究领域和项目，还有不少横跨领域 / 行业的科研方向，让研究员们有机会拓展研究的广度与深度。因此，2022 年博士毕业后，黄昶互毫不犹豫地加入了微软亚洲研究院，成为网络与基础设施组（Networking Infrastructure Group）的一员。

聚焦提升 AI 系统性能：层层递进研究，持续打磨成果

实习期间，黄昶互所在团队的重点研究课题是优化支持人工智能模型运行的 GPU 的性能，当时黄昶互的工作比较明确，主要是探索如何通过软硬件协同设计来提高人工智能系统的吞吐量和利用率。然而，科学研究是一项长期工作，有些研究并不会在短期内就显现成果。作为一名坚信长期主义理念的科研人员，黄昶互在这项研究中并没有把自己当作匆匆过客，相反，在实习结束回到学校继续攻读博士学位的两年里，他仍然与微软亚洲研究院的团队保持合作，持续跟进这一课题。最终，他和研究院团队的研究成果获得了 2022 年 MLArchSys 大会的最佳论文奖。

论文标题：Towards GPU-driven Code Execution for Distributed Deep Learning
https://chhwang.github.io/pubs/mlarchsys22_hwang.pdf

随着大模型的发展，GPU 愈发成为训练和部署人工智能模型的关键硬件，GPU 的性能和利用效率直接影响着人工智能的发展。因此，在成为微软亚洲研究院正式的研究员之后，黄昶互依然致力于这一方向的研究，而他的角色从曾经的项目参与者，转变为了项目的主导者。

黄昶互认为，如今最先进的深度学习应用需要大量并行的 GPU 提供充足的算力，但 GPU 和 CPU 之间的通信效率却制约着人工智能模型的性能。具体来说，在当前主要依靠 GPU 驱动的人工智能系统通信模式下，CPU 却扮演着总指挥的角色，CPU 负责给多个 GPU 布置任务，而 CPU 与 GPU 之间的消息传递存在可观的延迟，这就导致了任务执行效率的低下，造成了 GPU 资源的浪费。

黄昶互的研究目标与思路是希望 GPU 可以自己指挥自己，从而提升通信效率。为此，他和组里的同事们设计了一种由 GPU 驱动的代码执行系统，并开发了一种能够被 GPU 直接驱动的 DMA 引擎，让 GPU 能够自己解决原本需要 CPU 指挥的通信问题，降低了人工智能系统的通信延迟，提高了 GPU 计算资源的利用率。这种方法释放了之前通信模式下被占用的 CPU 资源，让 CPU 专注于自己的工作，也让 GPU 实现自主调度，做它最擅长的工作——给人工智能模型提供更高的算力性能。这项研究工作首次证明了基于分布式 GPU 的人工智能系统可以由 GPU 自己完成任务调度，相关论文已被 2023 年 NSDI 大会接收。

论文标题：ARK: GPU-driven Code Execution for Distributed Deep Learning
https://www.usenix.org/system/files/nsdi23-hwang.pdf

“ 系统性能优化是一个永恒的话题。在过去的十几年中，我们见证了人工智能的快速发展，其中一个主要的驱动因素就是不断增强的算力支持。充足的算力让系统性能持续提升，也使得人工智能模型变得越来越大，功能越来越强。当前，提升系统性能的研究方向主要有两个切入点，一是提升 GPU 等硬件的性能，二是提出新的人工智能算法，但这两种方法都相当困难，并且硬件的设计和制造成本高昂。” 黄昶互说道。

在这样的背景下，黄昶互和同事们提出了硬件与算法协同设计的方法，这或将是另一种提升人工智能系统性能的有效解决方案。因此，在证明了 GPU 可以自主调度，实现性能提升后，黄昶互将继续探索 GPU 的调度算法，避免调度冲突，进一步提升 GPU 之间的通信效率。黄昶互表示，“ 希望未来 GPU 不再需要额外的 DMA 引擎就能实现自主调度，从而推动人工智能系统性能再上一个台阶。”

黄昶互（中）在微软亚洲研究院第二次实习后与组内小伙伴合影留念

“在微软亚洲研究院，我可以自由选择研究方向”

微软亚洲研究院一直以来所拥有的开放、包容、多元的研究文化，也对黄昶互有着巨大的吸引力。在研究院工作一年有余的黄昶互对这里有了更深刻的认识，“ 微软亚洲研究院更像是一个实验室，一个真正的研究机构，在这里，所有人都是平等的，所做的工作都是透明的，大家了解彼此的想法，思想上也能够保持同步。在研究院，我们有更大的自由度来选择自己的研究方向。”

除了在内部营造自由的学术氛围，微软亚洲研究院还在学术交流和人才培养方面，与包括韩国在内的全球学术界持续保持着紧密的合作。例如，微软亚洲研究院联合清华大学、北京大

学、新加坡国立大学、首尔国立大学等多所亚洲地区高校成立了 OpenNetLab 开放网络平台联盟，以推动人工智能在网络研究中的应用与发展，黄昶互在 KAIST 求学时的导师就参与其中。再如，持续了十多年的面向韩国高校人才培养和学术研究的 MSIT 项目，为微软亚洲研究院与韩国学术界搭建了学术交流的桥梁，通过合作项目，学者们开展了深入的科研合作，并丰富了全球计算机领域的人才储备。黄昶互在微软亚洲研究院实习后也参与了一个学术合作项目，相关论文还获得了 2021 年 APSys 大会的最佳论文奖。

论文标题：Accelerating GNN Training with Locality-Aware Partial Execution
论文链接：https://dl.acm.org/doi/10.1145/3476886.3477515

作为微软亚洲研究院乃至整个计算机学术生态体系的一部分，这些多样的交流与合作项目不仅产出了众多前沿的科研成果，也成为了众多学者和学生与微软亚洲研究院结缘的起点。以韩国为例，截至目前已有超过 150 多名来自韩国的跨学科人才在微软亚洲研究院进行过实习，也吸引了像黄昶互这样的优秀人才，成为了微软亚洲研究院的正式员工。

坚持长期主义研究的心得

科学研究之路道阻且长，坚持长期主义研究、行而不辍并非一件易事。除了本身执著的性格之外，黄昶互也有自己的方法和心得。

黄昶互认为从事科研工作，首先要对研究事业保持高度的热情，比如他自身就十分享受科学研究中发现问题、解决问题的整个过程。“有些工作的目标是找到避开问题的最佳方式；而科学研究的目标是找到问题、直面问题、解决问题。我更享受从发现问题到解决问题的科研过程。”黄昶互说。

而在长期研究中难免会遇到阻碍，或者结果达不到预期，例如黄昶互的研究论文也曾被所投大会一次次拒之门外。面对这种情况，“不要气馁或是怨天尤人，而是要反思自己，复盘已有工作，找出其中的问题，再投入新的研究。”黄昶互认为，“这是一个说服自己的过程，要让自己看到研究的价值。”

当面对研究困境时，黄昶互表示，不能画地为牢，将自己困于当前的问题中，而是要学会放松自己，例如他会弹弹钢琴，或者与他人谈心交流，以此来摆脱桎梏，转换思路也许问题就能迎刃而解。

科学匠人｜罗琳：用真诚赢得信任，用 AI 助力无障碍沟通

你知道吗？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。世界聋人联合会希望通过设立这两个国际日，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。

在微软亚洲研究院也有一群人在为此努力，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，致力于利用人工智能技术为聋听之间搭建沟通的桥梁，让聋人朋友可以使用他们自己的语言——手语，与听人无障碍交流。

如果你用一个人听得懂的语言与他交流，他会记在脑子里；如果你用他自己的语言与他交流，他会记在心里。——纳尔逊 · 曼德拉（前南非总统）

这句话深深地打动了微软亚洲研究院视觉计算组的研究工程师罗琳。在参与研究院手语识别与翻译研究的过程中，罗琳了解了聋人朋友的心声，也因此把这句话作为了她与人交往、合作的一条准则，尤其是和聋人朋友的交流。

罗琳与微软亚洲研究院的缘份始于微软亚洲研究院与北京大学联合开设的软件实现技术系列课程。当时还是北京大学软件与微电子学院计算机技术专业硕士一年级学生的罗琳，参与了整个课程的学习，并凭借优异的表现，获得了研究院实习岗位的面试机会，最终通过面试成为实习生。研究院在人工智能领域所进行

的前沿探索，以及将人工智能技术应用于真实世界的落地成果，给实习中的罗琳留下了极其深刻的印象。

毕业后，罗琳毫不犹豫地选择加入了微软亚洲研究院。从实习生到成为正式员工，罗琳感受最深的，也是她加入微软亚洲研究院的直接原因是，“这里是一个可以让人自由生长的地方，每个人都能获得所需的养分，充分发挥自己的主观能动性。”

“不会手语怎么能做好手语的识别与翻译呢？”

一直以来，微软亚洲研究院都致力于开展有温度且面向未来的科学研究，并通过提供具有“包容性设计”的技术创新，来满足不同人群的实际需求。罗琳在加入研究院后参与的手语识别与翻译研究就是这样的项目。

此前，微软亚洲研究院曾与星巴克中国手语门店合作，通过语音识别技术，将顾客的需求转换成文字，帮助减少顾客与聋人店员之间的沟通障碍。对聋人来说，语音识别是让聋人看到听人在说什么，但这项技术仍有其挑战。一方面，在嘈杂的环境中语音识别的准确率会受到影响。另一方面，“对于大部分聋人来说汉语并不是母语，手语（视觉语言）与汉语（有声语言）在语法、词汇、表达法等各方面都有很大的差异，对于聋人来说，学好汉语是需要付出很大努力的。”罗琳介绍道。聋人用手语沟通才是他们直接、准确和更高效的沟通方式。让聋人使用手语直接和听人交流，通过 AI 识别翻译手语，从而让听人理解聋人的表达，是微软亚洲研究院进行手语识别与翻译研究的目标。

手语是一门视觉语言，它的识别与翻译是一个多模态问题，研究员们希望利用前沿的计算机视觉、自然语言处理等技术来尝试解决手语识别与翻译中的问题。微软亚洲研究院手语识别与翻译研究项目的目标是希望建立一个实时的手语与文字的双向翻译系统，将连续手语视频转换为文本或语音，也能将文本转换为连续手语视频，以实现聋人和听人之间的高效沟通。然而，微软亚洲研究院首席开发经理陈刚在项目启动时提出了一个问题——“如果我们都不会手语，怎么能做好手语识别与翻译呢？”

幸运的是，当时微软中国的一个团队招聘了一名手语非常优秀的聋人实习生。无独有偶，罗琳所在的研究组也招聘了一名来自天津理工大学聋人工学院的实习生。此后，罗琳和同事们迅速在组内建立起了手语课堂，由两位聋人实习生担任手语老师，开展了长达半年时间的手语学习。

正是在这个学习的过程中，罗琳发现手语是一门很有趣的语言，并对其产生了浓厚的兴趣。为了更深入地了解手语，在组内学习的同时，罗琳还报名了网上的手语课程，利用业余时间，一周三次跟随聋人老师学习手语。如今，罗琳已经可以和聋人朋友进行比较顺畅的沟通，但她依然通过多种渠道继续学习手语，并积极参与聋人社区的活动。“希望这些学习与交流，不仅能提升我的手语水平，也能让我深入聋人群体了解他们真正的需求，更好地利用前沿科技帮助到他们。”罗琳说。

罗琳（左四）与同事们的手语课堂开班合影

手语数据集是手语识别与翻译研究的一座大山

可能在外行人看来，手语识别与翻译就是将手语的各种手势、姿态汇总，再与汉字序列一一对应，通过语法规则匹配来达到两种语言的转换，但事实上手语翻译远比想象的要复杂得多。

首先，手语手势与汉字词语并不是一一对应的关系，存在一势多义，如“今天“和“现在”是同一个手势。而有些手语手势在不同上下文中也会表达不同的意思。

手语表达还存在很强的空间性，手形的运动、位置、朝向都会影响到意思的表达。例如，手语中“借”是一个方向动词，在“他从我这借钱”和“我从他那借钱”的手语表达中，手形的朝向和运动方向表达了不同的主谓关系。

再例如，“起风了，树被吹的摇摇晃晃，砸到了车”，在这个句子中，车和树存在相对的位置关系，而且“风吹”与“树晃”是同时发生的，这就需要计算机在识别时，可以理解物品在空间中的位置关系和交互关系。

另外，手语不仅包括手部的动作，还涉及面部表情、口部动作以及身体动作等非手控信息。例如，“吃完了吗？”，这句话除了手部动作之外，还需要配有“疑问”的表情来表达这是疑问句。手语的这些特点都对计算机视觉技术提出了更高的要求。

建立一个日常覆盖面广、质量高的手语数据集对解决手语识别与翻译难题至关重要，这也是罗琳的主要任务。目前，市面上虽然有一些公开的学术数据集，但还没有一个统一的标准，且数据质量仍有提升的空间。同时，模型训练需要大规模的数据量，相比语音数据上百万的量级，手语数据量仅在几万级别，远远达不到模型的需求。

相比其他数据集的建立，手语数据的采集和标注也有着更高的难度。罗琳和同事们需要找到高水平的手语使用者，就像听人的普通话水平不同一样，手语使用者的水平和习惯也大相径庭，

不同地区的手语表达也不尽相同。这就需要手语使用者拥有广泛的手语知识，熟知不同地区的多种手语打法，才能让数据集的词汇更丰富多样，手语识别准确率才会更高。同时，研究员自己也要懂手语，能够与手语使用者深入沟通，才能设置更好的采集任务，并在标注时兼顾计算机视觉与手语语言学的需求，提升数据集的质量。

“我们希望采集到的手语是聋人最日常的真实表达。这里有两个概念——手势汉语和中国手语，也就是人们常说的自然手语。手势汉语以汉语为基准，与汉语一一对应，如‘他坐在门口’，手势按顺序逐词对应，但这并不是聋人的日常表达；自然手语则是用最自然的手语语法来描述，同样的这句话，在视觉上‘门’与‘坐着的人’也是有位置关系的。我们希望引导手语老师打出最地道的手语词汇和句子。”罗琳说，“我们邀请到的手语老师就像一本手语大词典，每一位老师都拥有深厚的手语经验，无论南方手语还是北方手语，他们都能打出多种常用的手语表达，以此来扩充词汇量，让数据集更多样化。”

在手语识别与翻译研究项目中，微软亚洲研究院将手部动作、面部表情、口部动作等作为一个整体来进行识别，进一步提升了识别与翻译的准确率。目前，该项目在算法方面由微软亚洲研究院资深研究工程师魏芳芸带领，已在手语识别与翻译学术领域的多个子任务上处于领先地位。项目团队期待高质量的手语数据集在投入使用后，能更好地助力于相关研究的发展。

由于手语识别与翻译的研究尚处于早期阶段，并且是一个需要长期投入的领域。因此，在未来的一段时间，罗琳和同事们将继续推进手语识别与翻译的研究，帮助聋人与听人实现无障碍沟通的目标。

语言就像一把钥匙，帮你打开一个新的世界

要让有声世界与无声世界交融，并不是一件容易的事。罗琳认为，建立信任，展现真诚与尊重是第一要素。“就像曼德拉那句话说的，当你使用手语与聋人交流时，他们会觉得你更亲近，也更愿意与你交流，你也就能走进更多聋人朋友的世界，了解他们的生活。语言就像一把钥匙，可以帮你打开一个新世界，只要你展现出真诚、尊重与理解，他们会非常包容你、欢迎你。”

此外，在与聋人朋友们进行交流时，还要留意作为听人的固有表达沟通习惯，可能只是听人世界的习惯，不要简单的想当然，以己度人。因为，听人常常会从自己的角度考虑如何给聋人群体提供帮助，而缺少了设身处地的思考。罗琳曾作为 mentor 参加了第一届微软 Engage 残障大学生培养计划。该计划通过向听障、视障、肢障、自闭症等残障大学生提供为期六周的线上培训和项目指导，来提高大家编程能力，同时也帮助同学们成为未来优秀的“职场人”。

罗琳（右五）与热爱手语的聋人和听人朋友在手语星巴克

在项目筹备期，为确保不同残障类别的学生能在课堂上有效获取信息，筹备组决定提前录制课程视频并加入字幕，再在线上课堂中进行实时播放。然而，罗琳在一次指导学生修改答辩视频时，电脑没有开启声音，仅播放了带有字幕的 PPT 展示视频，她发现当自己在看字幕时完全无暇去关注 PPT 的内容，更不用说判断字幕与 PPT 内容的对应关系了。这个经历让罗琳深深感受到站在聋人群体之外，听人的设想很多都是不实际的。“只有进入到无声世界中，你才能真正了解他们所面临的问题，同时，与聋人相关的项目，也必须要倾听聋人的声音。”罗琳说。

正如全球残障人士社区经常使用的口号“没有我们的参与，请不要做与我们有关的决定（Nothing about us without us）。”罗琳和团队在进行手语识别与翻译研究时，也始终坚持这一理念。在近期举行的 2023 年微软全球骇客松大会上，微软亚洲研究院的手语识别与翻译项目获得了国际团队的关注。罗琳也期望在微软亚洲研究院的支持下，能够与更多的团队和聋人朋友深入合作，加速推动有声世界与无声世界的无障碍交流、沟通。

扫描二维码查看相关视频

量子位｜微软亚洲研究院提出 RetNet，或将成为 Transformer 有力继承者！

近期，微软亚洲研究院上线了大模型新架构的论文“Retentive Network: A Successor to Transformer for Large Language Models”，该基础架构采用了新的 Retention 机制来代替 Attention，向 Transformer 发起挑战！

Computer Science > Computation and Language

[Submitted on 17 Jul 2023]

Retentive Network: A Successor to Transformer for Large Language Models

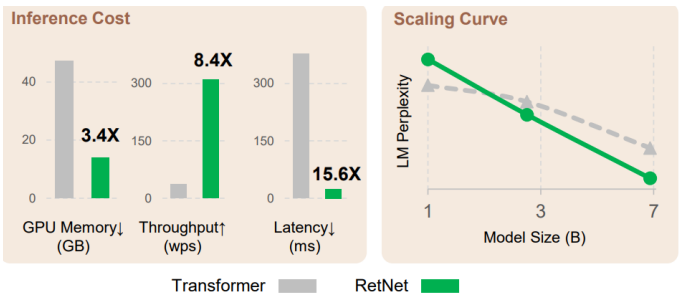
Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, Furu Wei

In this work, we propose Retentive Network (RetNet) as a foundation architecture for large language models, simultaneously achieving training parallelism, low-cost inference, and good performance. We theoretically derive the connection between recurrence and attention. Then we propose the retention mechanism for sequence modeling, which supports three computation paradigms, i.e., parallel, recurrent, and chunkwise recurrent. Specifically, the parallel representation allows for training parallelism. The recurrent representation enables low-cost $O(1)$ inference, which improves decoding throughput, latency, and GPU memory without sacrificing performance. The chunkwise recurrent representation facilitates efficient long-sequence modeling with linear complexity, where each chunk is encoded parallelly while recurrently summarizing the chunks. Experimental results on language modeling show that RetNet achieves favorable scaling results, parallel training, low-cost deployment, and efficient inference. The intriguing properties make RetNet a strong successor to Transformer for large language models. Code will be available at this https URL.

实验数据也显示，在语言建模任务上：

- RetNet 可以达到与 Transformer 相当的困惑度（perplexity）
- 推理速度达 8.4 倍
- 内存占用减少 70%
- 具有良好的扩展性

并且当模型大小大于一定规模时，RetNet 的表现会优于 Transformer。



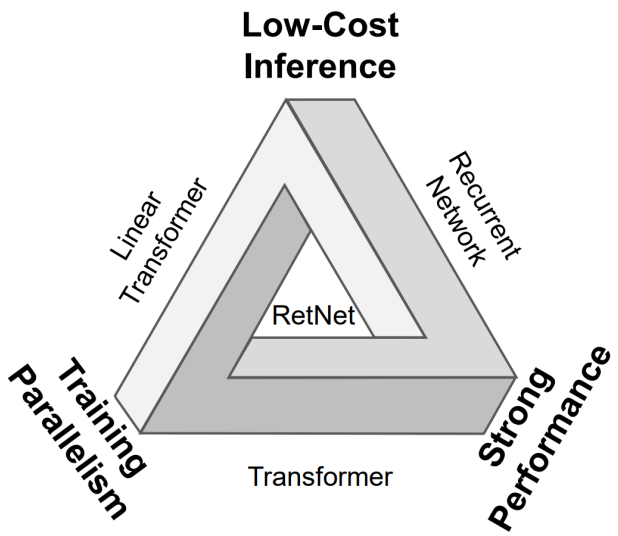
解决“不可能三角”

Transformer 在大语言模型中的重要性毋庸置疑。无论是 OpenAI 的 GPT 系列，还是谷歌的 PaLM、Meta 的 LLaMA，都是基于 Transformer 打造。

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，每个步骤的复杂度为 O(N)；Transformer 是内存密集型模型，序列越长，占用的内存越多。

在此之前，大家也不是没想过继续改进 Transformer。但主要的几种研究方向都有些顾此失彼：线性 Attention 可以降低推理成本，但性能较差；循环神经网络则无法进行并行训练。

也就是说，这些神经网络架构面前摆着一个“不可能三角”，三个角代表的分别是：并行训练、低成本推理和良好的扩展性能。



RetNet 的研究人员想做的，就是化不可能为可能。

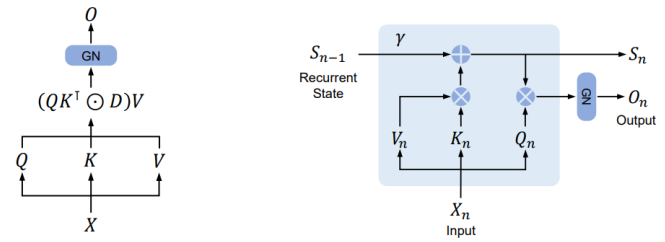
具体而言，RetNet 在 Transformer 的基础上，使用多尺度保持（Retention）机制替代了标准的自注意力机制。

与标准自注意力机制相比，保持机制有几大特点：

引入位置相关的指数衰减项取代 softmax，简化了计算，同时使前步的信息以衰减的形式保留下来。

引入复数空间表达位置信息，取代绝对或相对位置编码，容易转换为递归形式。

另外，保持机制使用多尺度的衰减率，增加了模型的表达能力，并利用 GroupNorm 的缩放不变性来提高 Retention 层的数值精度。



(a) Parallel representation. (b) Recurrent representation.

Figure 3: Dual form of RetNet. "GN" is short for GroupNorm.

RetNet 的双重表示

每个 RetNet 块包含两个模块：多尺度保持（MSR）模块和前馈网络（FFN）模块。

保持机制支持以三种形式表示序列：

- 并行
- 递归
- 分块递归，即并行表示和递归表示的混合形式，将输入序列划分为块，在块内按照并行表示进行计算，在块间遵循递归表示。

其中，并行表示使 RetNet 可以像 Transformer 一样高效地利用 GPU 进行并行训练。

递归表示实现了 O(1) 的推理复杂度，降低了内存占用和延迟。分块递归则可以更高效地处理长序列。这样一来，RetNet 就使得 " 不可能三角 " 成为可能。以下为 RetNet 与其他基础架构的对比结果：

Architectures	Training Parallelization	Inference Cost	Long-Sequence Memory Complexity	Performance
Transformer	✔	$O(N)$	$O(N^2)$	✔✔
Linear Transformer	✔	$O(1)$	$O(N)$	✘
Recurrent NN	✘	$O(1)$	$O(N)$	✘
RWKV	✘	$O(1)$	$O(N)$	✔
H3/S4	✔	$O(1)$	$O(N \log N)$	✔
Hyena	✔	$O(N)$	$O(N \log N)$	✔
RetNet	✔	$O(1)$	$O(N)$	✔✔

Table 1: Model comparison from various perspectives. RetNet achieves training parallelization, constant inference cost, linear long-sequence memory complexity, and good performance.

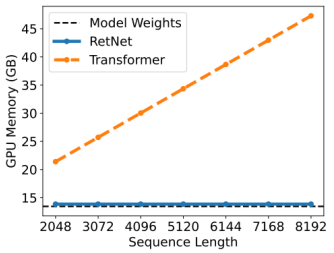
在语言建模任务上的实验结果，进一步证明了 RetNet 的有效性。

结果显示，RetNet 可以达到与 Transformer 相似的困惑度（PPL，评价语言模型好坏的指标，越小越好）。

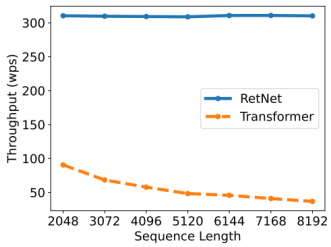
同时，在模型参数为 70 亿、输入序列长度为 8k 的情况下，RetNet 的推理速度能达到 Transformer 的 8.4 倍，内存占用减少 70%。

在训练过程中，RetNet 在内存节省和加速效果方面，也比标准 Transformer+FlashAttention 表现更好，分别达到 25-50% 和 7 倍。

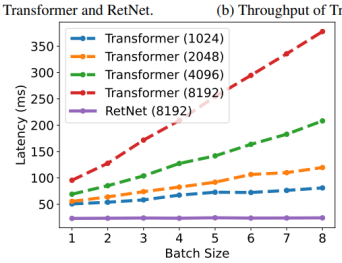
值得一提的是，RetNet 的推理成本与序列长度无关，推理延迟对批量大小不敏感，允许高吞吐量。



(a) GPU memory cost of Transformer and RetNet.



(b) Throughput of Transformer and RetNet.



(c) Inference latency with different batch sizes.

另外，当模型参数规模大于 20 亿时，RetNet 的表现会优于 Transformer。

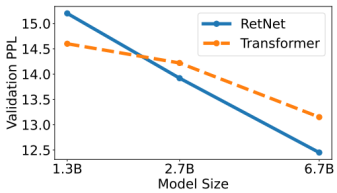


Figure 5: Perplexity decreases along with scaling up the model size. We empirically observe that RetNet tends to outperform Transformer when the model size is larger than 2B.

论文名称：Retentive Network: A Successor to Transformer for Large Language Models

论文地址：https://arxiv.org/abs/2307.08621

深科技 | 谢幸：做经得起时间检验的研究，打造负责任的人工智能

最近几年，人工智能（AI）技术发展迅猛，特别是在视觉、语音识别、阅读理解、机器翻译等单个任务上逐渐达到、甚至超过人类的水平。

2022 年 11 月底，OpenAI 发布 ChatGPT，再次掀起了人们对大语言模型关注的热潮。一直以来，大语言模型被视为通往通用人工智能的必然发展阶段，这让人们再次看到通用人工智能的曙光。

技术发展的同时，我们也必须看到其所带来问题，正如微软总裁布拉德·史密斯（Brad Smith）在《工具，还是武器？》一书中提到，当一个技术或工具能力非常强大时，它所带来的帮助和危害同时是巨大的。

目前，ChatGPT 运转的 GPT 模型的参数已经逾 1 万亿。当模型变得越来越大以后，随之带来的是计算的急剧增加。在传统的隐私保护模式下，其内存、计算和通信开销都不可接受。

此外，模型的复杂度也逐年增加，保障模型的鲁棒性、正确性和可解释性显得愈发困难。因此，如何应对和解决大语言模型可能带来的危害成为研究者们的重点研究方向之一。

谢幸是微软亚洲研究院的资深首席研究员、空间数据挖掘领域的先驱之一。为确保以负责任和符合人类价值观的方式使用人工智能，其团队致力于“社会责任人工智能（Societal AI）”的研究方向，以保证技术和大模型应用的公平性、可靠和安全、隐私和保障、包容性、透明度和责任。

他们通过设计有效的隐私保护方法，来保护用于大模型训练或微调的隐私数据，以及通俗易懂的模型解释方法，来帮助用户理解大模型的工作原理。并且，着力消除或减轻用于训练大模型的数据中固有的社会偏见和仇恨，从而让大模型在不断变化的环境中更加安全和鲁棒。

谢幸的研究在全球产生了深远的影响，截至目前，他共发表 400 余篇学术论文，h-index 为 103，共被引用 40000 余次，其中 8 篇论文被引次数超过 1000。凭借在深入理解人类自身行为规律的基础上，建立人与机器之间的信任关系，致力于人工智能技术规范发展、使计算技术变得更友好和负责任，谢幸成为 DeepTech 2022 年“中国智能计算科技创新人物”入选者之一。为 AI 与人类和谐共处奠定基础。

为 AI 与人类和谐共存奠定基础

当大模型的尺寸呈指数级上升时，我们如何确保 AI 能够在未来与人类是和谐共存的关系呢？谢幸表示，其团队将从五个方面来建设社会责任人工智能，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、跨学科合作。

第一，保证 AI 价值观与人类价值观对齐，也就是需要确保 AI 在与人和社会合作时，秉承和人类相同的价值观。

第二，保证训练 AI 的数据以及人工智能模型的安全，包含隐私、版权、知识产权等。大语言模型或 AI 很容易受“越狱”等欺骗，去完成未被设计的任务，因此需要防范安全风险。

最近，该团队与国内外多所知名高校的法学专家团队围绕版权进行了深入探讨。随着 AI 能力的提升，它能够生成图像、视频，或者将文稿润色成为一篇论文或文章。因此，当 AI 和人类共同完成内容时，如何来定义版权以及如何划分 AI 和人类的贡献成为关键的难题。

谢幸表示：“我们与很多法学专家交流，他们对于版权的定

义也有很多不同的意见。但我相信通过版权的定义来保护训练数据和作者们的版权，对未来创造性的工作肯定会有巨大的影响。”

第三，正确性或可验证性。很多时候 AI 会生成一些错误的信息，也就是人们常说的“一本正经地胡说八道”，如何保证 AI 的输出是可以验证的、是正确的也是谢幸与团队关注的研究方向之一。

第四，模型评测。由于现在大模型的能力越来越强，传统的 AI 的模型评测并不适用于现在的 AI 模型。为此，该团队正在试图提出一个全新的评测方式框架，以评测 AI 真正的能力。

第五，与社会科学的跨学科合作，包括法学、社会学、心理学、传媒、经济、教育等，共同研究 AI 未来的发展最终目标。据介绍，该团队最近与心理学相关专家及团队正在进行深入的研究探索，通过借鉴心理评测领域积累的经验，更好地评测 AI 未来的能力。

“在过去，心理学领域开发了非常专业的评测方法，来评测人类的核心能力，包括创造力、辩证思维、解决复杂问题的能力等。我们讨论后发现，这种方式和理念也适用于 AI。”谢幸表示。

时间是最好的试金石

谢幸博士于 2001 年 7 月加入微软亚洲研究院，现任微软亚洲研究院资深首席研究员，中国科学技术大学兼职博士生导师，微软 - 中国科大联合实验室主任。

微软亚洲研究院一直鼓励跨学科、国际化的交流以及前瞻性方向的合作。回顾过去二十几年的研究生涯，谢幸坦言，团队的研究方向随时代的趋势经过了几次调整和转变。“因此当我们在看到一些技术的发展趋势后，会马上尝试在这个趋势到来之前进行深入探索。”

2003 年 2 月，谢幸在微软研究院技术节上展示移动内容浏览的技术

他回忆道：“我记得我加入微软时用的还是诺基亚的非智能手机，但很快就出现了各种智能设备，大家逐渐意识到手机不仅是打电话的工具，还能做很多其他的事情。”基于此，他在微软亚洲研究院做的第一个项目是如何让用户在移动设备上更好地浏览内容，包括视频、图像和网页。

随着技术的发展，越来越多的传感器被加入到移动设备中，能力也越来越强，例如 GPS。因此从 2006 年开始，谢幸带领团队围绕着用户的位置数据做了一系列研究，例如通过这些数据理解用户的兴趣爱好、出行规划，然后从用户扩展到群体，再扩展到城市。他解释说道：“我们去理解一个城市的人群是怎么样移动、怎么样出行的，他们的节奏是什么样的，当时我们在时空数据挖掘方向展开了一系列的研究。”

有一句话叫“时间是最好的试金石”。谢幸团队的多项研究于十几年前发表，并在最近陆续获得了多项时间检验奖（Test of Time Award）。“这几个研究实际上都围绕着更好地去理解用户位置数据的方向展开，回头来看，这些研究都经受住了时间的检验，这也更坚定了我们对前沿技术探索的信心。”他说。

2021 年 7 月，《在物理世界知识指导下驾驶》（Driving with Knowledge from the Physical World）获得了 ACM 数据挖掘中国分会（SIGKDD China Chapter）颁发的时间检验奖。该团队通过对出租车轨迹数据的深入研究，进而摸索出交通模式和驾驶员的行为模式，设计了一种为用户提供定制化导航路线的服务。

发表于 KDD 2012 的论文《基于人类移动规律和兴趣点的城市功能区域发现》（Discovering Regions of Different Functions in a City Using Human Mobility and POIs）获得了 KDD 2022 时间检验奖。KDD 会议被认为是反映业内“最前沿数据领域研究风向”，被誉为“全球数据挖掘最高级别的学术会议”之一。在这篇论文中，作者们基于大规模人群移动数据以及地图兴趣点数据，提出了一种基于数据的城市功能分区自动发现方法，从而帮助人们更深入地理解不断演化的大城市 [2]。

谢幸团队获得的时间检验奖证书

在移动数据领域的国际学术会议 IEEE MDM 2023 上，谢幸团队于 2010 年发表的论文《一种基于交互式投票的地图匹配算法》（An interactive-voting based map matching algorithm）获得了时间检验奖 [3]。在该论文中，为解决用户 GPS 轨迹语义理解中面临的低采样率挑战，研究人员研发了一种地图匹配算法。该算法是在交互投票基础上实现的，根据真实轨迹数据集上的实验，算法大幅地提高了匹配性能。

后来，该团队重点研究的方向为社交网络数据和个性化推荐，更加深入理解人类用户。最近，他们聚焦于社会责任人工智能方向，从对人类自身的深入理解方向调整为对人和 AI 同时进行深入理解，从而让 AI 在未来与人和谐相处。

谈及 AI 发展的未来，谢幸表示，未来是人和 AI 共生的社会，AI 会完成很多从前人类做的事情，它会辅助人类完成具有创造力或挑战性的工作，例如现在科学家已经开始借助 AI 辅助进行科学研究等。“ 我们希望未来 AI 可以发展为一个安全的工具和助手，它的价值观与人类是对齐的，而且它的产出是可以信任的，是人类可控的助手。”

参考资料：

1. Jing Yuan et al. Driving with Knowledge from the Physical World. ACM SIGKDD international conference on Knowledge discovery and data mining, 2011s, 316–324
https://doi.org/10.1145/2020408.2020462

2. Jing Yuan et al. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, 186–194
https://doi.org/10.1145/2339530.2339561

3. Jing Yuan et al. An interactive-voting based map matching algorithm. IEEE MDM 2023.
https://doi.org/10.1109/MDM.2010.14

周礼栋

微软亚洲研究院院长

“二十多年来，微软亚洲研究院始终秉承开放、积极的心态，致力于打造自由、平等、可持续的科研协作环境，让分工、协调、合作链环上的每个人都成为新的发现与贡献的核心主体，为各种创造性想法的星星之火提供形成燎原之势的催化剂。

一个创新型组织的成长是不断拓展视野并承担更大社会责任的过程。微软亚洲研究院从创立伊始就持续与国内外计算机科研机构展开深度合作，携手进步，共同发展。在面对当下可持续发展、碳中和、医疗健康等人类社会亟待解决的关键问题时，微软亚洲研究院将守正创新，践行所有有利于激发创新力的原则，大胆接受和改造各种新的范式，与各界伙伴共同推动计算技术的跨界融合发展。”

关于微软亚洲研究院

微软亚洲研究院成立于 1998 年，在北京和上海拥有 300 多位科学家和工程师，是微软公司在亚太地区设立的、美国本土以外最大的研究机构。通过来自世界各地不同学科和背景的专家学者们的鼎力合作，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，致力于推动整个计算机科学领域的前沿技术发展，将最新研究成果快速转化到微软的关键产品中，并且着眼于下一代革命性技术的研究，助力公司实现长远发展战略和对未来计算的美好构想。

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，同时扎根中国，辐射亚洲，通过融合东西方创新文化的精髓，以高度的社会责任感，持续开展有影响力、有温度、面向未来的基础科学研究和技术创新。微软亚洲研究院始终秉持相互信赖、相互尊重以及开放合作的理念，承诺与高校和科研机构开展持久而有效的合作，激发创新潜力、推进行业发展。

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：提高人类的知识水平，推动基础研究的发展；增强人类的创造力和成就；培育有韧性、可持续的社会；支持健康的全球社会；确保技术值得信赖，让每个人都可以受益。

扫描二维码观看视频介绍

微软研究院全球布局

微信

知乎

电话：86-10-59178888

网址：http://www.msra.cn/

微博：http://t.sina.com.cn/msra