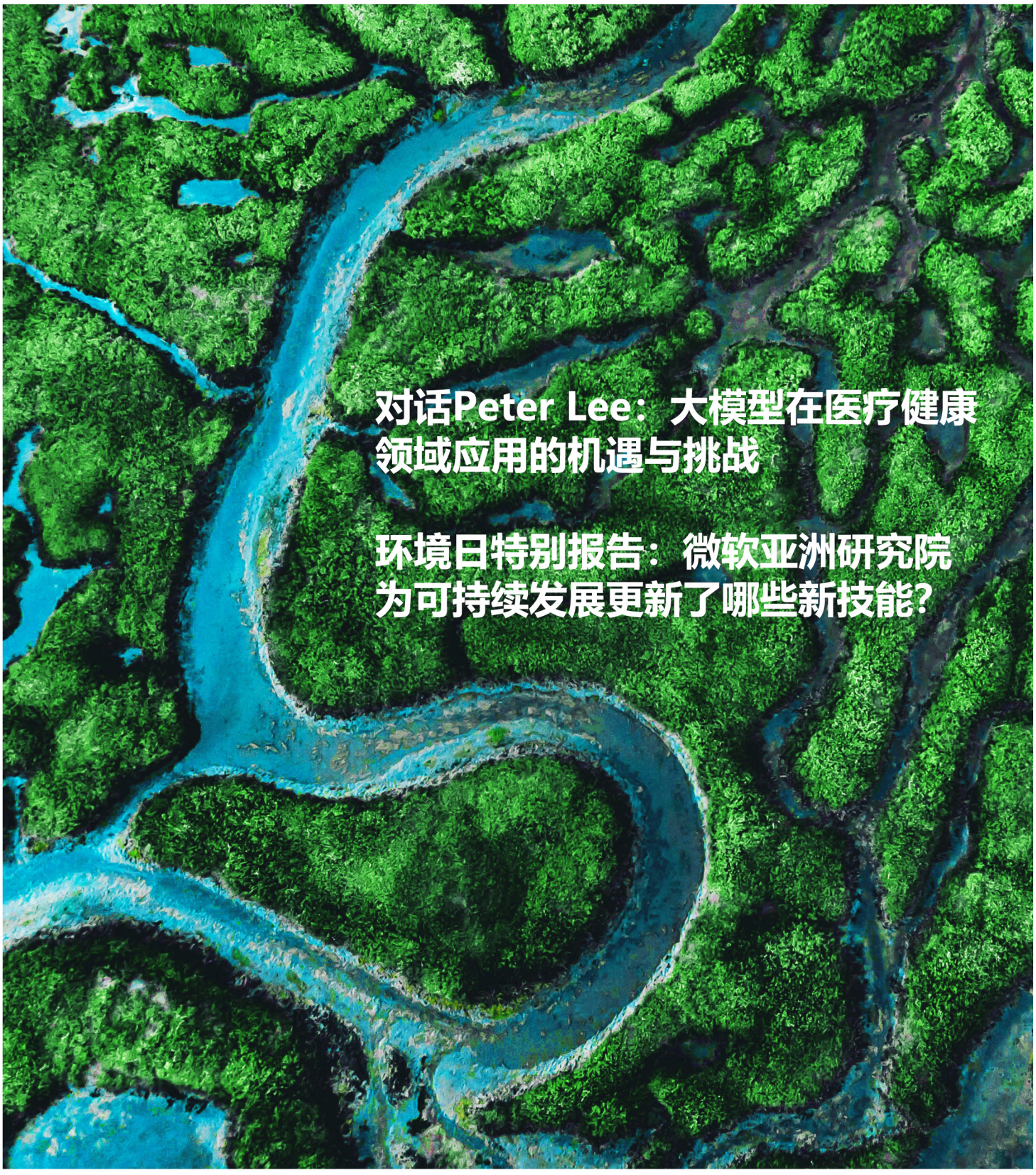
微软亚洲研究院院刊

Matrix

NO.65
2023年 4 - 6 月



对话Peter Lee：大模型在医疗健康领域应用的机遇与挑战

环境日特别报告：微软亚洲研究院为可持续发展更新了哪些新技能？

01 焦点

02 前沿求索

科研第一线

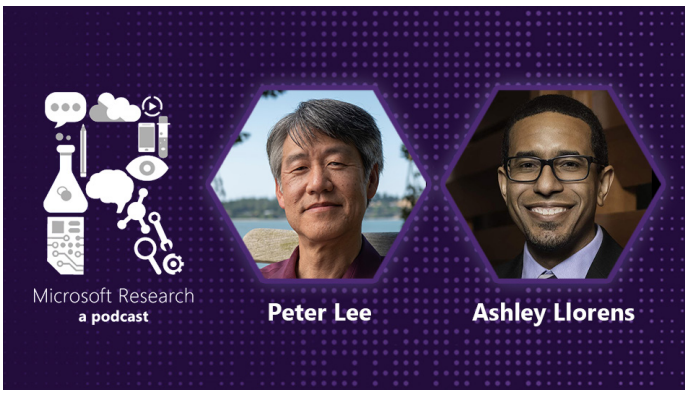
03 文化故事

04 观点

05 媒体报道

对话 Peter Lee：大模型在医疗健康领域应用的机遇与挑战

2023 年 3 月，OpenAI 推出了大语言人工智能模型 GPT-4，其在推理、解决问题和语言等方面的能力都有了显著提高，使得发展长达数十年的人工智能进入了一个新阶段。微软全球资深副总裁、微软研究院负责人 Peter Lee 是微软内部最早使用 GPT-4 进行评估和实验的成员之一。在微软研究院 AI 前沿系列播客节目中，Peter Lee 与微软研究院副总裁、微软杰出首席科学家 Ashley Llorens 进行了一次深度对话，表达了他对于大模型在医疗健康领域应用潜力和挑战的看法，以及在大模型潮流的引领下，微软研究院对未来计算的研究规划。本文节选了对话中的部分内容，完整版请扫描文末二维码收听博客。



Ashley Llorens：通过科学研究为社会创造更多机遇和价值，带给整个社会更有意义的影响是我们共同的目标。你一直关注情境研究（research in context），在 GPT-4 等大模型引领人工智能潮流的当下，你有何新的看法？

Peter Lee：情境研究是一个非常重要的课题。试想，你知道未来某个时刻世界的样子，然后再倒推回今天的工作会是怎样的？举个例子，科学家们相信 10 年之后我们将在很大程度上解决癌症问题，但随着人口老龄化加剧，未来与年龄相关的神经系统疾病将大大增加。如果我们现在就可以意识到神经系统疾病在未来的重要性并增加投入，这将使未来的世界与我们今天的处境截然不同。但如今的医学研究更聚焦于癌症研究，而非神经系统疾病。

这种变化意味着什么，是否在指导我们的科研方向？当然科学研究仍是未来导向的，但它既要展望未来十年，也要着眼现实世界，也就是情境研究。现在看来，通用人工智能超越人类智能可能是不可避免的，甚至在未来 5-10 年就会发生。那这对科研会有什么影响？它可能比癌症和神经系统疾病更具颠覆性和挑战性，影响也更深远。

之前我已经经历过五次类似的技术变革。第一次是上世纪 80 年代后期，我在卡内基梅隆大学担任助理教授，当时许多顶尖大学计算机科学系都在 3D 计算机图形学领域做出了优秀的研究成果，像光线追踪（ray tracing）、辐射度（radiosity）、硅结构（silicon architectures）这些想法都是在那时提出的。还有 SIGGRAPH 大会，当时每年都会吸引全球数百名科研人员展示各自的成果。到了 90 年代初，有些初创公司开始采用这些创新想法，试图将 3D 计算机图形变为现实，这其中的一家就是英伟达（NVIDIA）。最终，3D 计算机图形学成为了人们生活的基础设施，这是基础计算机科学研究取得的一次巨大成功，以至于今天你的口袋中如果没带"GPU"，没带手机，整个人都会感到不舒服。这种变革，对研究产生积极影响的同时，也具有颠覆性。

所以，即使某些领域取得了成功，但谈到扩展为人类社会的基础设施时，就脱离了基础研究的范畴，同样的情况还包括编译器设计（这是我自己的研究领域）、无线网络、超文本和超链接文档、以及操作系统等研究。现在它们已经成为我们生活中不可或缺的东西，都代表着计算机科学的伟大成就。而今天，我们正处于向大语言模型的过渡阶段中。

Ashley Llorens：你认为这次技术过渡是否在本质上与其他后台（background）技术有所不同？你提到我们每天出门时口袋里都装着"GPU"，但我不是这样想的，或许我对我的手机有某种拟人化的想法。但可以肯定的是，语言模型是一种具有前台效应（foreground effect）的技术，我想知道，你在其中是否看到了不同之处？

Peter Lee：我认为，对研究机构、学术界、领域内的研究人员来说没有什么不同，但对于技术的消费者和使用者，感受却有很大不同。相比同样从学术研究走入现实的触控可扩缩放的用户体验，大语言模型的影响可能会更深远。

这又带来一个大问题，当我们与大语言模型交互时，即使知道它不是有感情、有情绪、有知觉的生物，但又不由自主地这么想，这是进化中的固有思想。就像我们产生视觉幻像时，理智上深知这是幻觉，但大脑却无法克服，这种硬性连接引导我们将系统拟人化，也因此让它们走到了前台。

Ashley Llorens: 接下来我们把话题转向目前你正在努力的医疗健康领域以及在微软的历程。你曾说过把前沿的人工智能技术引入医疗健康系统面临诸多挑战，在 GPT-4 和大规模人工智能模型发展的背景下，人工智能与医疗健康结合时是否会有不同？

Peter Lee: GPT-4 是否会给医疗健康领域带来不同还需要检验。因为我们也曾对计算机技术帮助医疗健康领域或促进医学进步持乐观态度，但却一次次失望。这些挑战可能源于过度乐观。

作为计算机科研人员，我们看到了医疗领域的一些问题，例如对读取放射图像和测量肿瘤生长的研究，或对鉴别诊断选项或治疗选项排序问题的研究，我们认为自己知道如何用计算机科学解决这些问题。而医学界也在关注着计算机科学研究和技术的发展，他们对人工智能、机器学习和云计算印象深刻。因此，来自两个领域的这种难以置信的乐观情绪，最终变成了过度乐观。因为将计算机技术整合到医疗健康和医学工作流程中的实际挑战，是要确保它的安全性，并且真正发挥计算机技术的最大能力，但这是非常困难的。

另外，在医学实际应用中，诊断和治疗过程都发生在不稳定的环境中，这就导致在机器学习的环境中涉及很多混杂因素。由于医学是建立在对因果关系的精确理解和推理之上的，所以这些混杂因素至关重要，但现在机器学习里最好的工具本质上是相关性的机器（correlation machines）。相关性和因果关系是不同的，例如，吸烟是否会致癌，考虑到混杂因素的影响并了解其中存在的因果关系是非常重要的。

我第一次见到 GPT-4，是 OpenAI 的人员演示代号为 Davinci 3 的 GPT-4 早期版本，并让它回答 AP Biology（大学进阶生物学）的问题。在这次考试中，我认为它得了最高分 5 分。AP Biology 的试题通常是选择题，但该系统却能够使用自然语言对其选择的答案做出解释，让我吃惊的是，它在解释中使用了 "因为" 这个词。

例如，它会说"我认为答案是 C。因为当你从这个角度看问题时，会引发其他生物学问题，因此我们可以排除答案 A、B 和 E，然后又因为其他因素，排除答案 D，所有的原因和结果都是一致的。"我们都不清楚为什么一个大语言模型会具有因果分析能力。

这只是 GPT-4 百分之一的能力，它似乎克服了一些阻碍机器智能融入医疗健康和医学中的因素，例如推理、解释能力。再加上 GPT-4 的泛化能力，这似乎让我们对其在医学领域的作用更乐观，认为它有可能带来不同的未来。

另一方面，我们不必完全专注于临床应用。GPT-4 很擅长填写表格，减轻文本工作的负担，它知道如何申请医保报销的事先授权，这是医生目前主要的行政和文本负担。相关工作并没有真正影响到攸关生死的诊断或治疗的决定，但这些后台功能同样也是微软的重要业务。有很多理由可以让我们相信，与 OpenAI 的合作能够带来颠覆性的改变。

Ashley Llorens: 每一项新技术的出现都会伴随着相关的机遇和风险。这种新型的人工智能模型和系统有着根本的不同，因为它们不是学习特定功能的映射。但在各种各样的应用中，即使是这样的机器学习也有很多悬而未决的问题。你如何看待这种通用技术在医疗健康等领域所带来的机遇和风险？

Peter Lee: 我认为有一件事引起了大量社交媒体和公共媒体不必要的关注，那就是系统出现幻觉（hallucination）或者脱轨的时候。这是 GPT-4 和其他类似系统有时会遇到的问题，比如它们会编造一些信息。过去几个月，随着 GPT-4 的稳步发展，它产生的幻觉越来越少。我们也了解到，这种倾向似乎与 GPT-4 的创造力有关，它能做出明智的、有根据的猜测，能进行智能的推测。

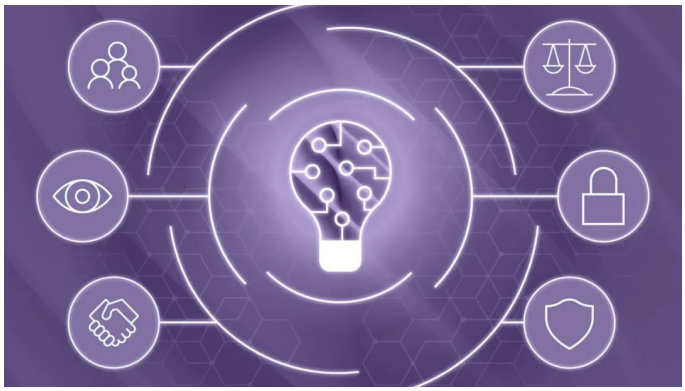
这是第一个你可以问它没有任何已知答案的问题的人工智能系统。而问题是，我们能完全相信它所给出的答案吗？ GPT-4 具有局限性，尤其在数学问题中。它很擅长解基本的微分方程和微积分，但在统计中却会犯基础性错误。我在哈佛医学院的同事就遇到过一个问题，在一个标准皮尔逊相关的数学问题上，它似乎总忘记对一个数据项进行平方。有趣的是，当你向 GPT-4 指出错误时，它的第一回答是，"不，我没犯错，是你错了。"随着系统的改进，现在这种指责用户犯错误的行为不会再发生了。

另外一个更大的问题与"负责任的人工智能"有关，这一直是整个计算机科学领域的重要研究课题，但我想这个词现在有可能不再合适了，我们可以称之为"社会责任人工智能（societal AI）"或其他的术语。它不是正确与错误的问题，也不仅仅是它会被误用而产生有害的信息，而是在监管层面的更大的问题，还有在社会层面的工作流失，新的数字鸿沟，以及富人和穷人获得这些工具的权利。这些问题也会直接影响着它在医疗健康领域的应用。

Ashley Llorens: 信任问题是多方面的，既包括在机构层面，也包括做出决策的个人。他们需要作出艰难的抉择，比如，在工作流程中，何时何地以及是否使用人工智能技术。你如何看待医疗健康专业人员做出此类决定？在将这些决策付诸应用时，存在哪些障碍？努力的方向又是什么？

Peter Lee： 关于 GPT-4 及同类技术应该在多大程度上应用，如何监管，有着很多讨论。美国有一个监管机构是食品和药物管理局（FDA），他们有权监管医疗设备。有一类医疗设备叫做软

件即医疗设备（software as a medical device，SaMD），在过去四五年中大家讨论最多的是如何监管基于机器学习或人工智能的SaMD。FDA 越来越多地批准使用机器学习的医疗设备。在我看来，FDA 和美国已经趋近于拥有真正公平的基础框架，来验证基于机器学习的医疗设备在临床的用途。但这些新兴框架不适用于 GPT-4，也就意味着用这些方法对 GPT-4 进行临床验证没有意义。



你的第一个问题是，这件事应该被监管吗？如果要监管，应该怎么做？这相当于把医生的大脑放在一个盒子里。假设，有一位伟大的脊柱外科医生，如果把他的大脑放在一个盒子里，请你验证这个东西，你会怎么想？什么样的框架适用于它？监管机构可能会做出反应并实施一些规则，但我认为这将是错误的，至少在目前，实施的规则应该是针对人的，而不是机器。

现在的问题是医生和护士、接待员和保险理赔员，以及所有相关人员，他们的指导方针是什么。这些决定不是监管机构的事情，而是医学界本身应该对这些指导方针和规则的制定负责，甚至通过医疗许可和其他认证来强制执行。这就是我们今天所处的位置，人类要自我负责，自我监管和规范自己的行为。

Ashley Llorens： 围绕测试和评估，以及相关的许可问题进行研究，也和创建模型本身一样有意思。

Peter Lee : 在这里，我想借机赞扬一下 OpenAI 团队的成员。我们在微软研究院的同事非常幸运，可以提前了解新技术对人类发展关键领域的影响，如健康和医学、教育等。OpenAI 团队看到了这样做的必要性，他们与微软研究院进行了深入地探讨，给了我们很大的自由度，让我们尽可能诚实且不加修饰地深入探索 GPT-4。这很重要，当我们与世界分享这些探索时，就能对它更加了解，能辩证地讨论。我们需要研究、考虑，以辩证的思想去看待它，而不是过度反应。

Ashley Llorens： 就你的观点而言，所有围绕各种社会重要性框架的思考都在试图追赶上一代技术，还没有完全瞄准这些新技术。在这种情况下，你认为计算机研究的下一步是什么？

Peter Lee： 我们是让技术从研究到成为生活中真正的基础设施这之间的纽带。微软研究院处于一个非常有趣的位置，既是研究的贡献者，让 OpenAI 正在做的事情成为可能，也是微软公司的一部分，希望与 OpenAI 一起让技术成为每一个人生活中的基础设施。作为变革的一部分，微软研究院已经确定了五个人工智能研究方向。

第一个是我们讨论的人工智能在社会中的作用和影响，包括负责任的人工智能等。其次，微软研究院的同事一直在推动 AGI（通用人工智能）的物理学概念。计算机科学理论一直是机器学习中重要的主线。这种研究风格越来越适用于理解大语言模型的基本功能、边界和趋势。即便你不再需要获得那些理解困难的数学定理，但它仍然是数学导向的，就像宇宙和大爆炸的物理学一样，AGI 的物理学也是如此。

第三方面是应用层面的。在微软研究院内部，我们称它为副驾驶（copilot）。我们期望让它成为你的伙伴，辅助你高效、高质地完成任务。

再有就是 AI4Science，我们在这方面做了很多工作，同时越来越多的证据表明，大型人工智能系统可以提供更新的方法，促进物理学、天文学、化学、生物学等方面的科学发现。

最后是核心的基础，我们称之为模型创新。不久前我们发布了新的模型 KOSMOS，用于进行多模态机器学习以及分类和识别交互。我们还创新提出了 VALL-E，基于三秒钟的语音样本就能够确定你的语音模型并复刻语音。这些模型创新还将继续发生。

从长远来看，如果微软、OpenAI 等公司获得成功，那么大模型将会真正成为生活基础设施工业化的一部分。我预计，大语言模型的研究将在未来十年开始消退，但是，全新的视野将会开启，这是在我们在网络安全、隐私和安全、物理科学等方面所做的所有其他事情之上的。可以肯定的是，现在人工智能正处于一个特殊时期，尤其是在以上这五个维度上。

扫描二维码收听完整播客

环境日特别报告：微软亚洲研究院为可持续发展更新了哪些新技能？

2023 年 3 月，联合国政府间气候变化专门委员会发布的一份报告[1]表明："人类活动主要通过排放温室气体，已毋庸置疑引起了全球变暖。不利的气候影响已经比预期的更加深远和极端。"这些巨大的排放量来自全球各行各业，与整个人类的活动都密切相关，随着人类文明的发展，地球也正在遭受破坏。保护环境，建设和谐的生态家园，是人类共同的责任。

一直以来，微软在以数字技术创新，实现绿色低碳的可持续发展道路上不断投入与探索。作为微软的创新前沿，微软亚洲研究院也希望运用计算机科学技术帮助解决碳减排、绿色能源等问题，同时通过创新生态平台与各界合作伙伴共同努力，以创新科研成果助力个人和组织实现可持续发展目标。

在 2023 年第 50 个世界环境日之际，让我们一起来了解一下微软亚洲研究院在可持续发展方面的科研和应用成果。

可持续发展已经成为当今全球社会的一个重要话题，为了确保全球的可持续发展，"双碳"（碳达峰、碳中和）策略、节能减排、空气污染治理、维护生态平衡等也成为人类面临的关键问题。

近年来，微软亚洲研究院将可持续发展作为重要的研究主题之一，并致力于以计算机科学推动气候、能源等领域的前沿研究和变革性创新，帮助全球应对气候变化，实现碳中和目标。这些研究既包括与学术界专家学者推进的跨学科基础研究，帮助人们更好地了解地球生态系统的碳汇过程，从而有效地指导碳中和工作；也有聚焦于计算机本身，推动绿色计算，减少碳排放的工作；同时，微软亚洲研究院还与产业界的伙伴合作，深入各种真实场景，在业务流程和供应链中最大程度地实现节能减排。

小贴士："碳达峰"很简单，就是指在某一个时点，二氧化碳的排放不再增长达到峰值，之后逐步回落。"碳中和"指的是，在一定时间内，通过植树造林、节能减排等途径，抵消自身所产生的二氧化碳排放量，实现二氧化碳"零排放"。

跨学科基础创新，助力实现碳中和目标

2020 年 1 月，微软宣布了一项重大承诺：在 2030 年实现碳负排放，并到 2050 年消除自 1975 年公司成立以来的碳排放量总和，包括直接排放或因用电产生的碳排放。为了助力公司实现这一目标，微软亚洲研究院开始从计算机基础科研发力。

要想制定出更好的碳减排策略和路线，首先要在宏观上掌握地球各个生态系统中碳排放和吸收的进程与机制。具体来说，由于二氧化碳在大气中的过量排放会导致温室效应，因此防止全球变暖就要控制排放到大气中的二氧化碳浓度，这需要准确预估二氧化碳的排放以及海洋和陆地生态系统在不同时间、不同地点对二氧化碳的吸收情况。

为此，微软亚洲研究院与清华大学地球系统科学系刘竹副教授的团队合作开发了全球首个近实时负碳数据库。该数据库基于清华大学提供的数据，利用微软亚洲研究院的先进人工智能技术，分别对海洋、陆地以及人工碳汇进行了高精度和快速的统计、反演、预测与补全。相比于当前时间滞后 1-2 个月的方法，该数据库实现了近实时的海洋和陆地碳汇数据预测。下一步，微软亚洲研究院将会合并海洋与陆地模型，构建统一的碳中和平台来显示当前碳中和进程，以便为政策的制定和调整提供有力依据，机构和组织也可以基于该数据库开展后续工作。

小贴士：碳汇是指通过植树造林、植被恢复等措施，吸收大气中的二氧化碳，从而减少温室气体在大气中浓度的过程、活动或机制。

如今，在碳减排的大趋势下，新能源行业快速发展，储能、电动汽车等行业对电池的需求越来越大，提高电池利用效率，减少废旧电池污染也成为一项重要课题。对于电池问题，微软亚洲研究院从电池寿命预测入手，见微知著减少碳排放。

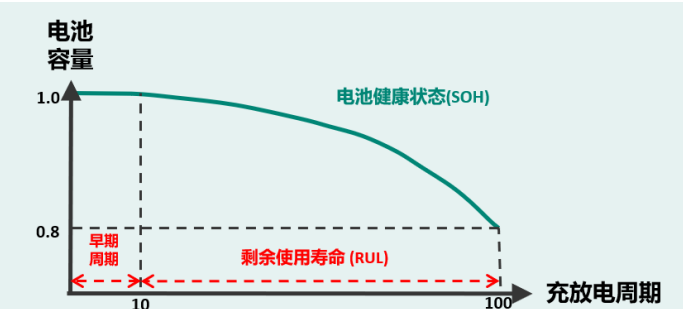


图 1：电池使用生命周期

微软亚洲研究院提出了一种多面深度对比回归(multi-faceted deep contrastive regression）方法，通过构建多面特征图来编码多维度不规则长度的时序电信号，并利用深度对比回归模型大幅提升了在数据匮乏情况下的模型的稳定性和泛化能力。该方法与当前最先进的方法相比，误差减少了 15%，针对电池容量退化的预测偏差可以控制在 1% 之内。通过更准确地预测锂离子电池的寿命（循环寿命）、剩余使用寿命和健康状态，这一技术可以进一步促进电池材料（阴极、阳极、电解质等）的快速筛选和设计，支持快速充电协议的高通量优化，设计更好的快速充电策略，改进电池管理系统，并通过提供电动汽车电池寿命终止预警来促进储能的二次利用。

解决真实场景中的真问题

可持续发展是一项长期且艰巨的任务。在多年的探索中，微软亚洲研究院已有不少研究成果应用在了实际的场景中，并在煤电厂低碳转型、空气污染治理、楼宇智能节能降耗等方面取得了显著实效。

煤电厂低碳转型。作为电力行业的支柱产业，煤电厂所排放的二氧化碳在全部的碳排放中占比极重。为了尽快解决煤电厂的排放问题，煤电行业一直在寻找低碳转型的新出路，其中基于生物质资源的负排放技术就是最值得挖掘的技术之一。但是，煤电厂低碳路径改造需要考虑对整个产业链的影响，比如，生物质资源分布不均衡，就需要考虑生物质资源的跨地区采集，而在资源运输过程中也会产生污染和相应的成本。

对此，微软亚洲研究院与清华大学地球系统科学系蔡闻佳副教授的团队合作提出了一种机器学习 + 运筹优化的解决方案（论文链接：https://pubs.acs.org/doi/10.1021/acs.est.2c06004）。通过将煤电厂自身情况与未来用电需求相结合，综合考虑资源分布、运输过程、煤电厂规模、技术改造之间的复杂关系，从而优化求解，最终制定出成本最低的最优碳中和策略。该方案已在全国 137 家省级煤电厂低碳转型实践中得到应用，与基准实验相比，改造策略在仅需要额外花费 10% 的电力成本时，就达到了 100% 碳排放减少的目标。

> 小贴士：生物质是指利用大气、水、土地等通过光合作用而产生的各种有机体，即一切有生命的、可以生长的有机物质，这些都通称为生物质，包括植物、动物和微生物。生物质的特点是具有可再生性、低污染性、广泛分布性、碳中性，且资源丰富。

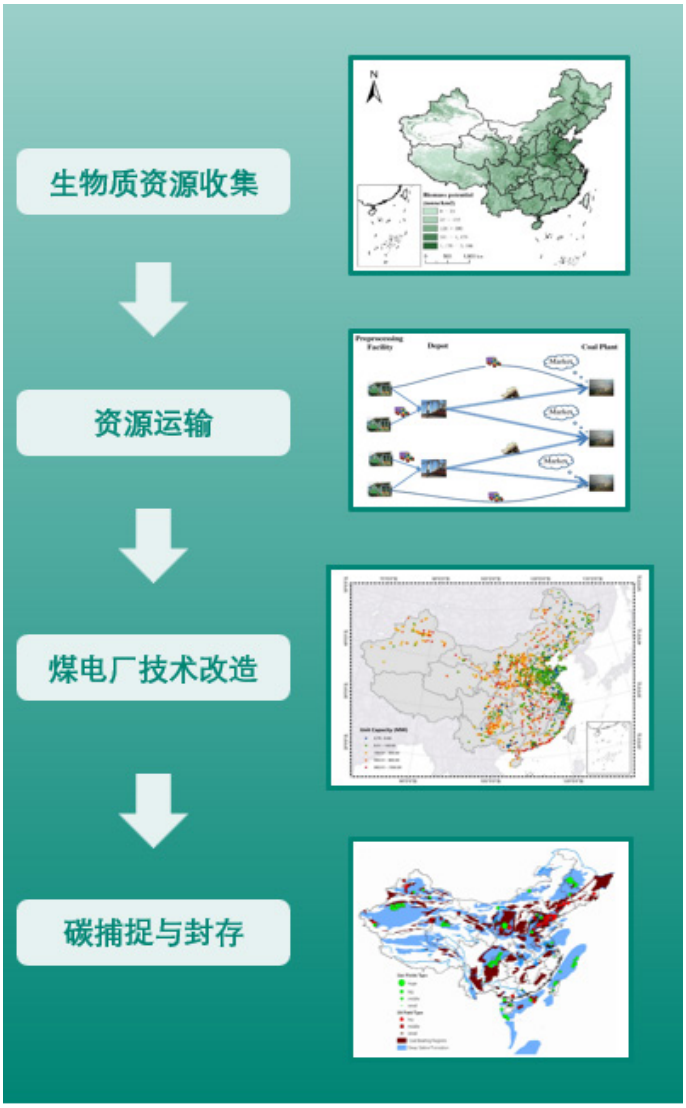


图 2：生物质资源

空气污染治理。空气污染已经成为“人类健康的最大环境威胁”之一[2]，空气污染治理一直是环境科学中的重要课题。控制和减少污染物排放是治理空气污染的重要手段，然而传统排放 - 污染物浓度响应曲面需要依赖大气化学传输模型（CTM）来进行模拟，其中存在计算量庞大、运行耗时耗力、时效性低等问题。

针对这些问题，微软亚洲研究院与清华大学地球环境学院邢佳副教授的团队共同研发了 DeepRSM 大气响应模型，能够更精细地刻画空气污染物浓度，可将传统大气响应数值模型的运算速度提升近百倍，模型错误率下降一半，从而帮助决策者快速找到最佳的减排方案。该工作已被中国“十四五”规划采纳为评估治理大气污染的重要模型，并被美国环境保护署（USEPA）用于预测美国本土的空气质量。

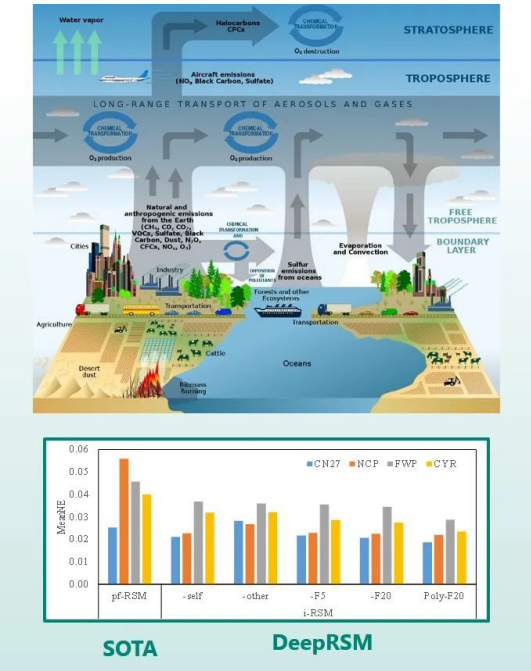


图 3：DeepRSM 大气响应模型

节能降耗。节能降耗是从源头减少碳排放的重要方式，它可以用较低的代价实现"双碳"目标。在日常生活中，供暖和空调对能源的消耗相对较高。而且目前大部分的 HVAC（供暖通风与空气调节）设备的控制通常还是依靠人工或者是简单的规则，这使得控制本身无法快速适应建筑内外部因素的变化，无法有效地达到节能效果。

通过收集历史上积累的相关数据，微软亚洲研究院的研究员们利用离线强化学习算法（offline reinforcement learning）对 HVAC 的控制策略进行了优化。该算法可根据外部天气情况和内部人流变化，实时调整空调设备的送风温度和送风压力，在保证系统安全和人体舒适度的情况下，使能源消耗减少高达 30%（线上测试），实现能源消耗最小化。该方案已经在多个楼宇中部署应用，效果显而易见。

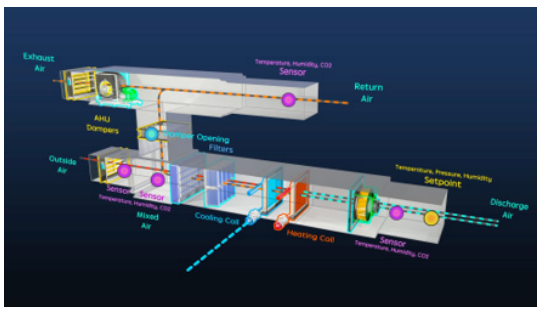


图 4：DeepRSM 大气响应模型

节能减排下一步：多技术角度打造绿色低碳数据中心

随着人工智能、云计算技术的发展，数据中心的部署越来越多，规模也愈发庞大，数据中心已然成为能耗大户之一。据中国信息通信研究院预测 [3]，2030 年，全国数据中心耗电量将超过 3800 亿度，如果不采用可再生能源，碳排放量将超过 2 亿吨。然而，可再生能源的使用存在间歇性和波动性影响。

为此，微软亚洲研究院联合清华大学碳中和研究院院长助理，环境学院长聘教授鲁玺教授的团队计划展开研究，开发动态风电能源评估研究框架、数据中心终端用户模型和集成风力发电系统的数据中心协调优化模型，从而为数据中心提供最优的运行方案，最大限度地节约能源。

此外，针对微软自己的数据中心，微软研究院也在探索节能减排方案。为了确保 IT 设备正常运行，数据中心会使用大量的水资源用于设备冷却。微软研究院希望能够利用前沿的机器学习方法，在满足温度标准的同时优化用水量，进而促进微软承诺到 2030 年实现"水资源正效益（water positive）"目标的达成。

共建创新合作平台，共营碳中和发展生态

人类只有一个地球，各国共处一个世界，保护生态环境、应对气候变化，实现碳中和目标至关重要，这是全人类面临的共同挑战，需要包括政府、高校与企业等不同主体多方联合，共建创新合作平台。

2021 年，清华大学与多家知名跨国企业共同发起"气候变化与碳中和国际合作联合行动"，微软亚洲研究院也是其中一员。联合行动成员倡议"共同传播碳中和绿色理念、共同加强碳中和人才培养、共同引领碳中和技术创新、共同开展碳中和集成示范、共同推动碳中和产业转型"。以此来加强绿色理念的传播、碳中和人才培养和技术创新，在促进社会可持续发展中提供切实可行的方案，为实现全球碳中和愿景作出贡献。

创新技术在实现"双碳"目标中正发挥着越来越重要的作用。当前，碳中和技术主要体现在低碳利用及能效提升技术、可再生能源技术、碳捕获和储存等负碳技术这几个重点领域。未来，基于人工智能、云计算等技术发展的数字能源将成为能源发展的新形态，也将是推动实现碳中和的重要力量。微软亚洲研究院将持续在全球可持续发展的生态中发挥自身技术优势，与各界伙伴一道共享碳中和的治理举措和经验，共建碳中和领域前沿基础研究交流平台，共营碳中和理念传播及发展生态。

图 5：2023 年 3 月 14 日，"气候变化与碳中和国际合作联合行动"启动仪式在清华大学顺利举行，微软亚洲研究院学术合作总监马歆（右三）与其它十一家创始理事单位代表共同见证了"联合行动"的开启

相关链接：

[1] 联合国政府间气候变化专门委员会 (IPCC)，AR6 Synthesis Report: Climate Change 2023
https://www.ipcc.ch/report/ar6/syr/

[2] 世卫组织全球空气质量指南
https://www.who.int/zh/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution

[3] 中国信息通信研究院云计算与大数据研究所，算力基础设施的现状、趋势和对策建议
http://ictp.caict.ac.cn/article/2022/2096-5931/2096-5931-48-3-2.shtml

多位微软亚洲研究院研究员斩获殊荣

2022 年底到 2023 年 4 月，微软亚洲研究院的科研人员获得了国内外多个表彰技术创新及其社会影响力的荣誉奖项：

2022 年底，微软亚洲研究院副院长邱锂力获选美国国家发明家科学院院士（NAI Fellow）。作为无线及移动网络领域的国际顶级专家，邱锂力因其多年来在相关研究领域所做出的卓越贡献而获此殊荣。NAI Fellow 是授予学术创新发明家的最高荣誉，旨在表彰对人类社会福祉产生切实影响的发明者。

2023 年 3 月，微软亚洲研究院高级研究员胡瀚作为"远见者"入选了 2022 年度《麻省理工科技评论》"35 岁以下科技创新 35 人"中国榜单。该榜单专注于挖掘新兴科技创新领域的中国青年力量。胡瀚的入选理由为：其所提出的 Swin Transformer 促进了视觉 Transformer 取代经典的卷积神经网络，让计算机能够像理解语言一样看周围世界。

2023 年 4 月，微软亚洲研究院首席研究员谢幸和段楠一同获选了由 DeepTech 发起的"2022 年中国智能计算科技创新人物"，该奖项旨在让更多人了解在智能计算领域极具才华与创新精神的年轻中坚力量，展现智能计算领域最新的学术研究成果和技术突破，推动智能计算科学与技术的进步。两位科学家的获奖理由分别是：谢幸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；段楠构建了多语言多模态预训练基础模型，积极探索基于基础模型的复杂任务推理和任务完成机制，推动了通用人工智能技术的发展。

再次祝贺各位同事！

NUWA 系列再添新成员——超长视频生成模型 NUWA-XL

最近，大型语言模型展现出的强大能力引发了新一轮的 AIGC （人工智能生成内容）研究和应用热潮。人工智能的创作能力边界已经从文字问答、编程逐渐扩展到了绘画、音频等多模态领域。但在视频领域，尤其是超长视频内容的生成上，目前大多数模型的效果还不尽如人意。近期，微软亚洲研究院 NUWA 多模态生成模型家族迎来了新成员——NUWA-XL，其以创新的 Diffusion over Diffusion 架构，首次实现了高质量超长视频的并行生成，为多模态大模型提供了新的解题思路。

输入 16 句简单描述就能生成一段长达 11 分钟的动画片？

没错！微软亚洲研究院提出的超长视频生成模型 NUWA-XL 可以根据文字自动生成高质量动画作品。扫描文末二维码查看由 NUWA-XL 生成的动画片。

早在多年前，微软亚洲研究院就开始了包括图像和视频在内的视觉生成方面的研究，并于 2021 年推出了多模态生成模型 NUWA。NUWA 可以通过自然语言指令实现文本、图像、视频之间的生成、转换和编辑，为视觉内容创作提供灵感。随后推出的 NUWA 升级版——无限视觉生成模型 NUWA-Infinity，则可以支持更高分辨率的图像和短视频生成任务，让视觉艺术创作趋于"无限流"（还记得那个无限延展的 Windows 桌面吗？）。

随着视频行业需求的增长和技术的发展，近两年人工智能在视频生成方面取得了一定的进展，然而，大多数模型还仅能够生成 3 到 5 秒左右的短视频。但在实际应用中，人们所需的视频通常要比 5 秒长得多，例如，一部电影通常持续在 90 分钟以上，一集动画片往往也超过 20 分钟，即使是常见的短视频时长也多在 30 秒以上。因此，超长视频的快速生成对于人工智能来说仍然是一个巨大的挑战。

"视频生成任务和语言、图像的生成类似，但图片是静止的，只包含了空间信息，而视频还需要考虑时间等因素。我们认为视频生成模型是可以对标语言生成模型的，并且拥有更大的应用潜力和更多的应用场景。所以在基于大模型的 AIGC 发展初期，我们就已经将视觉生成列为研究对象 (https://arxiv.org/abs/2104.14806)，并放在了与文本生成同等重要的位置上。"微软亚洲研究院首席研究员段楠表示。

当前，长视频生成的多数方法是采用"Autoregressive over X"架构，"X"表示任何能够生成短视频片段的生成模型，包括 Phenaki、TATS、NUWA-Infinity 使用的自回归模型（Autoregressive Models），或者 MCVD、FDM、LVDM 使用的扩散模型（Diffusion Models）。这些方法的主要思想是在短视频片段上训练模型，再通过推理，像滑动窗口一样自回归的自左向右生成长视频。

由于在训练时只需要短视频数据，"Autoregressive over X"架构在一定程度上降低了对长视频数据的要求，但微软亚洲研究院的研究员们发现了这种方法存在的问题：

首先，在短视频上进行训练再推理出长视频，会导致巨大的训练 - 推理差距（Train-Inference Gap）。也就是说，这种方法只知道所生成长视频的开始和结尾的故事信息，视频中间的情节则完全依赖前一段小视频的再推理，这种状态不断叠加之后就会导致不真实的、扭曲的镜头变化。缺乏长视频数据的训练，还会让模型生成的视频存在帧与帧之间不连贯以及故事情节无法逻辑自洽等问题。

其次，由于滑动窗口的依赖性限制，模型只能顺序自左向右生成视频，无法并行推理，因此需要花费更长的时间。例如，TATS 需要 7.5 分钟才能生成 1024 帧，而 Phenaki 需要 4.1 分钟。

全新 Diffusion over Diffusion 架构，"从粗到细"的生成过程

为了解决这些问题，微软亚洲研究院提出了 NUWA-XL（eXtremely Long），它采用 Diffusion over Diffusion 架构，通过 "从粗到细"的生成过程，以相同的粒度并行生成视频，并应用全局扩散模型（Global Diffusion）来生成整个时间范围内的关键帧，然后通过局部扩散模型（Local Diffusion）递归地填充附近帧之间的内容，既提升了生成效率，也确保了视频的质量和连续性。具体而言，如图 1 所示，NUWA-XL 中的全局扩散模型首先会基于 L 个文本提示生成 L 个视频关键帧，形成视频的"粗略"故事情节。然后将第一个局部扩散模型应用于 L 个提示和相邻的关键帧，将其视为第一帧和最后一帧，以完成中间的 L – 2 帧，从而总共产生 L+（L – 1）×（L – 2）≈ L^2 个"精细"帧。通过迭代应用局部扩散来生成中间帧，视频的长度将以指数级增加，进而生成非常长的视频。例如，具有 m 深度和 L 局部扩散长度的 NUWA-XL 能够生成具有 O（L^m） 大小的长视频。

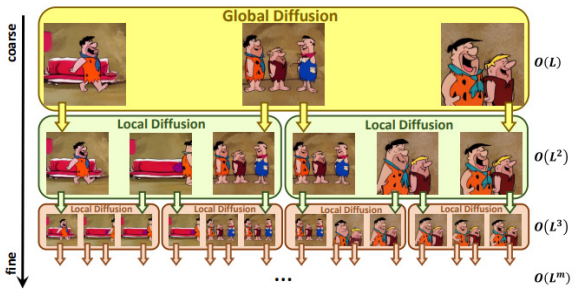


图 1：“粗略到精细”——NUWA-XL 超长视频生成概述

NUWA-XL “从粗到细”的生成方法具有三个优势：

1. 分层结构使模型能够直接在长视频上进行训练，从而消除了训练和推理之间的差距。NUWA-XL 会先生成类似于连环画的关键帧，既加强了场景切换又保证全局内容的统一，再在关键帧之间生成更多帧画面。同时，模型从 L 到 L^m 的每一层关键帧还支持文字提示生成关键帧，极大地确保了视频情节的连续性。

2. 模型包含多个局部扩散模型，自然支持并行推理，可以显著提高生成长视频时的推理速度。例如在相同的硬件设置下，当生成 1024 帧时，NUWA-XL 使平均推理时间从 7.55 分钟减少到 26 秒，速度提升了 94.26%。

3. 由于视频的长度可以相对于深度 m 呈指数级扩展，因此模型可以很容易地扩展出更长的视频。

长视频生成的时间、质量和连续性均获最优性能

除了生成时间的大幅缩短外，NUWA-XL 在 Avg FID 和 Block FVD（B-FVD）两个指标上也优于其它模型。Avg FID 起始距离（FID）是一种评估图像生成的度量，用于计算生成帧的平均质量，数值越低越好。B-FVD 视频距离（FVD）则被广泛用于评估生成视频的质量。

如表 1 所示，对于 “X over AR” 架构，由于误差累积，生成帧的 Avg FID 随着视频长度的增加而下降，例如 Phenaki，生成 16 帧的数值是 40.14，生成 1024 帧时是 48.56。与之相比，NUWA-XL 不是按顺序生成帧，所以质量不会随着视频长度的增长而下降，Avg FID 始终保持在 35 左右。

同时，与仅在短视频上训练的 “AR over X” 相比，NUWA-XL 能够生成更高质量的长视频，而且随着视频长度的增长，NUWA-XL 的生成片段（B-FVD-16）质量下降得更慢，因为 NUWA-XL 已经学习了长视频的模式。此外，可并行执行这一特性使得 NUWA-XL 在生成 256 帧时，推理速度提高了 85.09%；生成 1024 帧时，推理速度提高了 94.26%。

	Method	Phenaki [23]	FDM* [5]	NUWA-XL	NUWA-XL
	Resolution	128	128	128	256
	Arch	AR over AR	AR over Diff	Diff over Diff	Diff over Diff
16f	Avg-FID↓	40.14	34.47	35.95	32.66
	B-FVD-16↓	544.72	532.94	520.19	580.21
	Time↓	4s	7s	7s	15s
256f	Avg-FID↓	43.13	38.28	35.68	32.05
	B-FVD-16↓	573.55	561.75	542.26	609.32
	Time↓	65s	114s	17s (85.09%↓)	32s
1024f	Avg-FID↓	48.56	43.24	35.79	32.07
	B-FVD-16↓	622.06	618.42	572.86	642.87
	Time↓	259s	453s	26s (94.26%↓)	51s

表 1：NUWA-XL 长视频生成与最先进模型的定量比较（其中 Avg FID 数值越小代表性能越好）

NUWA-XL 为人工智能视频生成提供新思路

“在长视频生成的研究过程上，我们也咨询了专业的动画制作人员，了解了真正的动画制作流程，即先画出故事中的几个关键画面（即关键帧），再在关键帧之间不断添加更多的画面，来丰富故事情节确保连续性。正是受到真实动画创作流程的启发，我们在 NUWA-XL 工作中采取了 Diffusion over Diffusion 这样一种由粗到细的设计。相较传统从左至右的生成方法，NUWA-XL 由粗到细的生成方法从根本上改变了人工智能生成长视频的方式。”微软亚洲研究院主管研究员吴晨飞说。

NUWA-XL 以动画片为例验证了 Diffusion over Diffusion 架构的有效性，为超长视频的人工智能生成研究打开了新的思路。未来，通过在电影、电视等更多的视频数据上的训练，以及更强大的算力，NUWA-XL 或将进一步帮助动画、电影、电视、广告等视觉制作领域提高生产力。

对于人工智能多模态大模型的发展，段楠认为，“现在的大模型还停留在文字生成阶段，尽管 GPT-4 在理解端加入了视觉信息，但也仅限于图片，输出端还是文字或代码。因此，当前及未来的研究路线非常清晰，就是将语言与视觉的理解和生成融入到一个基础大模型中，在输出端加强图像、视频、音频的生成。我们希望未来可以用一套架构来融合支持语言、视觉的生成算法，让人工智能模型更加通用。”

相关链接：

NUWA-XL 项目页面：
https://msra-nuwa-dev.azurewebsites.net/#/

论文链接：https://arxiv.org/abs/2303.12346

扫描二维码查看视频

隐藏在 Microsoft Designer 背后的新科技，让人人都是设计师

在视觉图像设计中，用户的需求与最终的设计成品往往是"想象很美好，现实很骨感"。这通常是因为用户在与设计师沟通时，双方理解不一致，导致最终设计结果不尽如人意。但是，如果能够"自给自足"，借助人工智能技术为每个人赋予设计能力，是否会更容易让自己脑海中的画面变为现实？智能化设计工具 Microsoft Designer 就是一个能辅助用户成为设计师的好帮手。

2022 年 10 月，微软在 Ignite 大会上发布了 Microsoft Designer 内测版，为 Microsoft 365 家族再添一个视觉生产力工具。2023 年 4 月 27 日，经过半年的迭代和改进，微软宣布推出 Microsoft Designer 公开预览版。利用人工智能技术"猜想"用户的想法，智能辅助生成文字提示和视觉图像，Microsoft Designer 大大降低了设计难度，让人人都能成为视觉设计师。

如今，市场上充斥着各种各样视觉的设计工具，然而这些专业软件有很高的技术门槛且操作复杂，非专业人员难以熟练使用。也有一些工具提供了海量的模板库，用户可以基于模板进行修改，虽然这简化了部分操作，但其呈现效果与用户的设想仍有不小的差距。Microsoft Designer 则能够智能理解用户的需求，自动生成文字表述，实现从文字到视觉图像的自动化创造，并将这些素材用于设计项目。

作为智能化的设计工具，Microsoft Designer 将先进的科研成果快速吸纳并转化为生产力，其中包括来自微软亚洲研究院视觉计算组的 Provence、Swin Transformer 模型，自然语言计算组与微软图灵团队合作的图灵通用语言表示模型，系统研究组的 SPANN（存储器 - 磁盘混合索引和搜索系统）算法等众多前沿技术。

厚积薄发：Provence 多模态内容推荐模型助力实现“一键式”设计配图

微软亚洲研究院很早就开始研究通过自然语言生成图像或视频的技术。2018 年，正值短视频发展的上升时期，研究员们意识到视频化的传播形态将成为未来互联网主要的沟通交流方式。然而视频内容的制作流程繁杂，高质量视频的拍摄更需要专业人员的参与，那么是否可以通过技术创新创造出一个简化视频制作和生成的工具？在这一目标的驱动下，视觉计算组开始了文字到图像和视觉的生成技术的研究。

经过一年多的潜心钻研，2020 年视觉计算组推出了第一代基于检索的文字到视频的生成模型 Provence（Retrieval-based text-to-video generation）。Provence 模型能够根据文本描述搜索相匹配的视频或图像，同时确保跨模态对应具有较高的准确率，达到了"一键式（即检索到的第一个图像推荐就是用户所需）"的水平。

Provence 模型的潜力很快就被微软 Microsoft 365 产品部门发现，并将其引入到了 PowerPoint Design Ideas（PowerPoint 设计器）功能中。为了更好地满足产品端的工程化需求，微软多个研究组的研究员们将 Provence 与 Swin Transformer、图灵通用语言表示模型和 SPANN 算法结合，在 Design Ideas 功能的底层构建了零样本多模态的内容检索引擎，让用户在几秒钟内就能通过文字自动检索出最适合于当前幻灯片的配图，并给出布局设计建议，良好的使用体验让 Design Ideas 功能的用户使用率提升了 20% 以上。

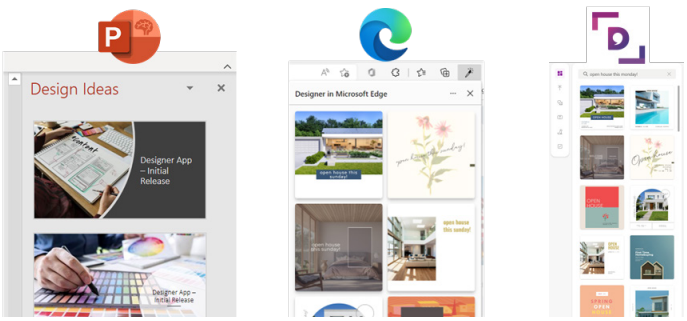


图 1：Provence 模型分别应用于 PowerPoint Design Ideas，Microsoft Designer 及其生态系统中

2021 年 3 月，在微软内部的骇客松（Hackathon）活动期间，Microsoft 365 产品团队通过与研究员们的思想碰撞，更加深入地了解了机器学习领域的前沿技术趋势，不仅看到了走向成熟的 Provence 检索技术，也看到了更多创新的机器学习算法的应用潜力。由此，双方共同发起了 Designer in Edge 的 Hackathon 项目，此后这一项目进入产品化迭代过程，也形成了如今的 Microsoft Designer 和 Designer Platform 这两款由人工智能驱动的产品。

微软 Microsoft 365 产品部副总裁张大川表示，“在与微软亚洲研究院多个研究组的交流中，我们看到了 Provence、Swin Transformer、图灵通用语言表示模型等诸多超前的 AI 技术理念，这些前沿技术完全满足 Microsoft Designer 及其生态系统的需求。

双方的紧密合作，不仅大大加速了产品的创新周期，而且还革新了传统设计的流程。下一步，我们将共同致力于创新技术的落地应用，为全球用户提供更加便捷易用的视觉设计工具，更好地激发人们的创造力和创新力。”

“很高兴看到微软亚洲研究院越来越多的创新研究成果走向了实际应用，成为支持产品开发的核心技术。以 Microsoft Designer 为例，它的关键技术始于研究院五年前的创新突破，正是因为微软亚洲研究院持续致力于探索计算机领域前瞻性的基础研究，才使得这种拿来即用的技术转化成为可能。未来，微软亚洲研究院将一如既往地着眼于下一代革命性技术的研究，并将科研成果快速转化到微软的产品中，赋能更多用户。”微软亚洲研究院常务副院长郭百宁表示。

Microsoft Designer：从多模态推荐走向具有“创作”能力的 AI

生成式视觉设计的一个关键环节是用语言或者文字将用户脑海中想象的画面清晰地表达出来。因此，微软亚洲研究院视觉计算组的研究员们进一步对 Provence 模型进行了升级，让 Microsoft Designer 在从文字描述中精准检索出用户所需图像的基础上，又实现了根据用户意图智能输出文字提示的功能。

其核心思想是基于学习到的自动模板为不同的输入文本创建不同的提示，具体包括三个步骤：

首先，将用户原始输入的文本与一组字符（token）结合，这些字符是对用户所需要的设计图像的视角、样式、氛围、用途等的描述。然后，根据美术设计的评分，使用学习到的自动模板找到与不同字符匹配的最佳组合。最后，将输入文本和自动模板提示的组合返回给用户，并使用评分指标对结果进行排序，再从中检索出最佳的图像。智能输出提示文字，为用户原始的输入文本添加了更多的描述和细节，从而激励视觉模型“创作”出更符合用户需求的结果。如图 2 所示：用户输入“a cat hacker wearing a VR headset”后，Microsoft Designer 自动输出了相关的提示与图像。另外，研究员们还提出了一种检索增强提示的推荐方法，通过使用提示数据库来增强自动提示的结果。随着用户对 Microsoft Designer 的频繁使用，模型会学习到更多的提示，而这些数据将

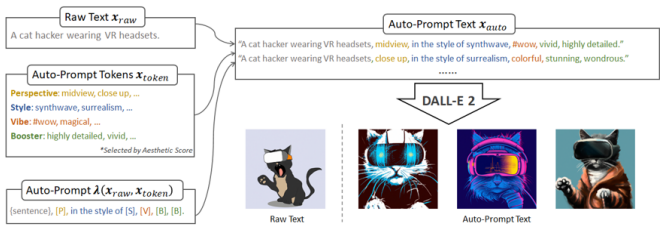


图 2：智能输出提示文字的流程

能进一步提高提示质量。如图 3 所示，对于用户输入，Microsoft Designer 会先使用语言模型从提示数据库中检索最相似的提示文本，然后通过评估分数对检索结果排序，再将自动提示与排序检索提示结合，以获得更好的结果。由于模型具有持续学习的特性，最终将有越来越多的用户数据纳入到提示数据库中来增强提示。

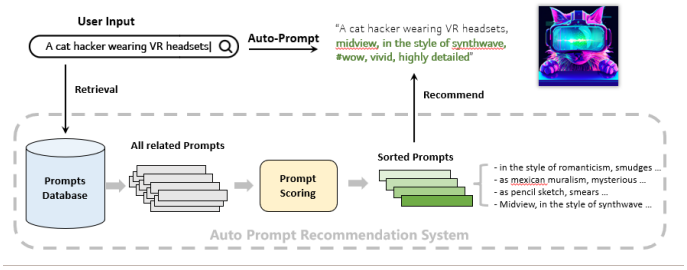


图 3：检索增强提示

借助智能输出文字提示和智能图像生成的双重加持，用户可以更好地用文字描述出脑海中的画面，让 Microsoft Designer “创作”更符合需求的视觉图像，并从推荐的图像中选择出最匹配需求的用于后续的定制化设计。由人工智能技术驱动的 Microsoft Designer 极大地降低了设计工作的专业门槛，让设计更加大众化，人人都能成为设计师。未来，Microsoft Designer 还将引入更多的人工智能算法，比如个性化的智能修图、借助大模型实现平面布局等等，以此丰富 Microsoft Designer 的功能，为更多用户带来更高水平的创意和创造力生产工具。

现在就来试用 Microsoft Designer，开启 AI 设计之旅吧！（产品页面：https://designer.microsoft.com）随着人工智能技术的发展，确保相关技术能被人们信赖是一个需要攻坚的问题。微软采取了一系列措施来预判和降低人工智能技术所带来的风险。微软致力于依照以人为本的伦理原则推进人工智能的发展，早在 2018 年就发布了“公平、包容、可靠与安全、透明、隐私与保障、负责”六个负责任的人工智能原则（Responsible AI Principles），随后又发布了负责任的人工智能标准（Responsible AI Standards）将各项原则实施落地，并设置了治理架构确保各团队把各项原则和标准落实到日常工作中。微软也持续与全球的研究人员和学术机构合作，不断推进负责任的人工智能的实践和技术。

相关论文：

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows
https://arxiv.org/abs/2103.14030

SPANN: Highly-efficient Billion-scale Approximate Nearest Neighbor Search
https://arxiv.org/abs/2111.08566

BEiT: BERT Pre-Training of Image Transformers
https://openreview.net/forum?id=p-BhZSz59o4

多项创新技术加持，实现零 COGS 的 Microsoft Editor 语法检查器

Microsoft Editor 是一款人工智能写作辅助工具，其中的语法检查器（grammar checker）功能不仅可以帮助不同水平、领域的用户在写作过程中检查语法错误，还可以对错误进行解释并给出正确的修改建议。神经语法检查器模型是这款提供了强大拼写检查和语法纠正服务的 Microsoft Editor 背后的关键技术，该模型采用了微软亚洲研究院创新的 Aggressive Decoding 算法，并借助高性能 ONNX Runtime（ORT）进行加速，使服务器端的模型推理速度提升了 200%，在不损失模型预测质量的情况下，节省了三分之二的成本。神经语法检查器模型还使用了微软亚洲研究院前沿的客户端 seq2seq 建模技术 EdgeFormer，构建了性能优异的轻量级生成语言模型，结合部署中的模型和系统优化，该技术可赋能用户在设备上的部署，从而实现零销货成本（zero-COGS, zero-cost-of-goods-sold）的目标。本文编译自微软研究院博客 "Achieving Zero-COGS with Microsoft Editor Neural Grammar Checker"。

自上世纪 70 年代以来，语法检查器（grammar checker）所依赖的技术已经取得了显著的发展，最初的第一代工具只是基于简单的模式匹配（pattern matching）。1997 年，一个标志性的事件发生了，当时 Microsoft Word 97 引入了一个基于成熟的自然语言处理系统（Heidorn, 2000）的语法检查器，以支持更复杂的错误检测和修改，并提高了准确率。2020 年，语法检查器再次实现关键性突破，微软推出了神经语法检查器（neural grammar checker），通过利用深度神经网络和全新的流畅度提升学习和推理机制，神经语法检查器在 CoNLL-2014 和 JFLEG 基准数据集上均取得了 SOTA 结果 [1, 2]。2022 年，微软发布了高度优化后的 Microsoft Editor 神经语法检查器，并将其集成到 Word Win32、Word Online、Outlook Online 和 Editor Browser Extension 中。

如今 Microsoft Editor 版本中的神经语法检查器模型主要采用了微软亚洲研究院创新的 Aggressive Decoding 算法，而且借助高性能 ONNX Runtime（ORT）进行加速，可以使服务器端模型的推理速度提升 200%，在不损失模型预测质量的情况下，节省了三分之二的成本。此外，该神经语法检查器模型还使用微软亚洲研究院前沿的客户端 seq2seq 建模技术 EdgeFormer，构建了性能优异的轻量级生成语言模型，结合部署过程中设备开销感知的模型和系统优化，该技术满足交付要求，赋能用户设备上的部署，最终实现了零销货成本（zero-COGS, zero-cost-of-goods-sold）的目标。

不仅如此，Microsoft Editor 中的神经语法检查器模型在转换为客户端模型后，还有三个优势：

1. 提升隐私性。客户端模型在用户设备本地运行，无需向远程服务器发送任何个人数据。

2. 增强可用性。客户端模型可以离线运行，不受网络连接、带宽或服务器容量的限制。

3. 降低成本、提高可扩展性。客户端模型运行在用户设备上，省去了服务器执行所需的所有计算，从而可以服务更多客户。

> "Aggressive Decoding 算法具有巨大价值，它不仅适用于 Microsoft Editor 这样对响应时间、请求频率和准确度都有很高要求的应用场景，还可以拓展到更多功能模块，如文本重写、文本摘要等。Aggressive Decoding 算法能够在保证模型预测质量不受损的同时更快地服务更多的客户，降低服务成本并提高产品的竞争力和影响力，这一创新技术也将在未来的客户端模型研发中发挥重要作用。"
>
> —— 陈思清，微软首席应用科学家

Aggressive Decoding：首个在 seq2seq 任务上无损加速的高效解码算法

Microsoft Editor 中的人工智能语法检查器主要基于 Transformer 模型，并采用了微软亚洲研究院在语法纠错方面的创新技术 [1, 2, 3]。与大多数 seq2seq 任务一样，Microsoft Editor 此前的模型使用了自回归解码来进行高质量的语法校正。然而，传统的自回归解码效率很低，尤其是由于低计算并行性，导致模型无法充分利用现代计算设备（CPU、GPU），从而使得模型服务成本过高，并且难以快速扩展到更多终端（Web/ 桌面）。

为了降低服务成本，微软亚洲研究院的研究员们提出了创新的解码算法 Aggressive Decoding[3]。与之前以牺牲预测质量为代价来加速推理的方法不同，Aggressive Decoding 是首个应用在 seq2seq 任务（如语法检查和句子重写）上达到无损加速的高效

解码算法。它直接将输入作为目标输出，并且并行验证它们，而不是像传统的自回归解码那样逐个顺序解码。因此，这一算法可以充分发挥现代计算设备（如带有 GPU 的 PC）强大的并行计算能力，极大地提升解码速度，能够在不牺牲质量的前提下以低廉的成本处理来自全球用户（每年）数万亿次的请求。

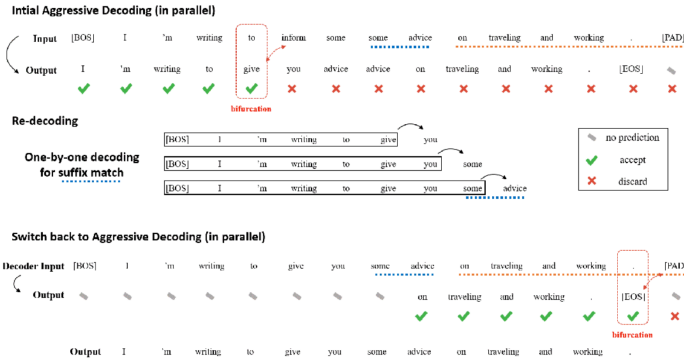


图 1：Aggressive Decoding 的工作原理

如图 1 所示，如果模型在 Aggressive Decoding 过程中发现了一个分歧点，那么算法将舍弃分歧点后的所有预测，并使用传统的逐个自回归解码重新解码。如果在逐个重新解码时发现了输出和输入之间存在唯一的后缀匹配（图 1 中蓝色点线突出显示的建议），那算法会通过把输入的匹配字符（token）之后的字符（图 1 中用橙色虚线突出显示的部分）复制到解码器的输入中并假设它们是相同的，从而切换回 Aggressive Decoding。通过这种方式，Aggressive Decoding 可以确保生成的字符与自回归贪婪解码一致，但解码步骤大幅减少，显著提高了解码效率。

> " 我们在做模型推理加速算法研究时最重要的考虑就是无损，因为在实际应用中，模型生成质量是排在第一位的，以损失质量来换取更小的开销会严重影响用户体验。为此，我们提出了 Aggressive Decoding 算法，它利用了语法纠错任务的一个重要特性，即输入与输出高度相似，将整个计算过程（pipeline）高度并行化，充分利用 GPU 在并行计算上的优势，在生成质量无损的前提下实现大幅加速的效果。"
>
> —— 葛涛，微软亚洲研究院高级研究员

离线 + 在线评估结果：Aggressive Decoding 可显著降低 COGS

离线评估：研究员们在语法校正和其他文本重写任务如文本简化中，采用了一个 6+6 标准的 Transformer 及深度编码器和浅层解码器的 Transformer 测试 Aggressive Decoding。结果表明 Aggressive Decoding 可在没有质量损失的情况下大幅提升速度。

	CoNLL14		NLCC-18		Wikilarge		
	F0.5	speedup	F0.5	speedup	SARI	BLEU	speedup
6+6 Transformer (beam=1)	61.3	1	29.4	1	36.1	90.7	1
6+6 Transformer (AD)	61.3	6.8	29.4	7.7	36.1	90.7	8

表 1：6+6 标准 Transformer 测试结果

	CoNLL14	
	F0.5	speedup
12+2 Transformer (beam=1)	66.4	1
12+2 Transformer (AD)	66.4	4.2

表 2：深度编码器和浅层解码器的 Transformer 的测试结果

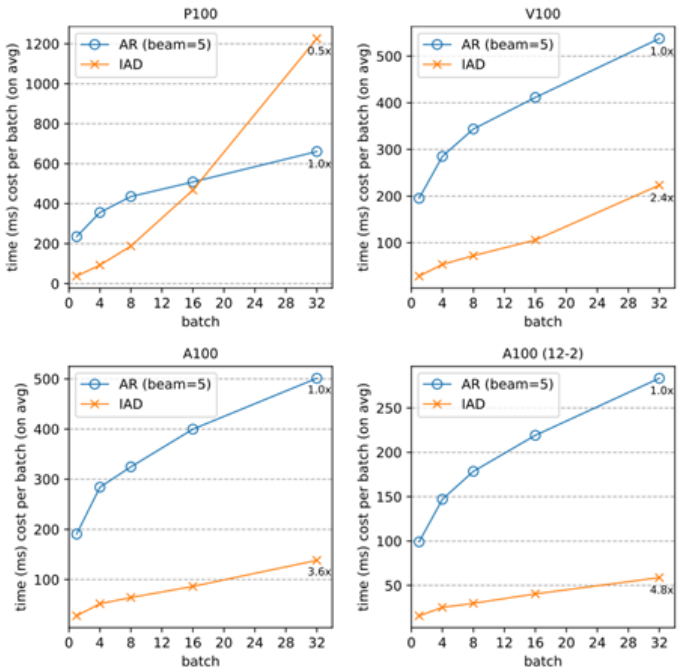


图 2：Aggressive Decoding 算法在更强大的并行计算设备上的运行效果更好

在线评估：研究员们还在 Marian 服务器模型和使用 ONNX Runtime 的 Aggressive Decoding 的同等服务器模型之间进行了 A/B 实验。结果如图 3 所示，与在 CPU 中使用传统自回归解码的 Marian 运行时相比，后者在 p50 延迟上有超过 2 倍的提升，在 p95 和 p99 延迟上有超过 3 倍的提升。此外，与之前的自回归解码相比，后者提供了更高的效率稳定性。这种显著的推理时间加速，将服务器端的 COGS 降低了三分之二。

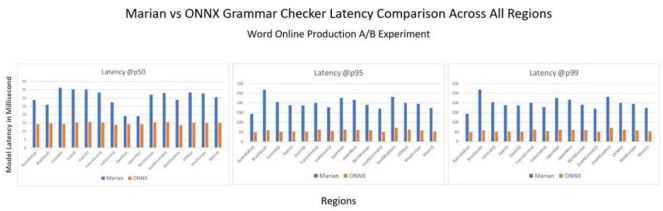


图 3：所有区域 Marian 和 ONNX 语法检查器延迟对比

离线和在线评估都验证了 Aggressive Decoding 能够在不降低模型预测质量的情况下显著减少 COGS。基于此，研究员们将 Aggressive Decoding 也应用到了更通用的 seq2seq 任务中 [4]。Aggressive Decoding 的高效率和无损质量特性，或将使其成为 seq2seq 任务高效解码的标准范式，在降低 seq2seq 模型部署成本中起到重要作用。

ONNX Runtime 加速语法检查器

ONNX Runtime 是微软开发的高性能引擎，它可在各种硬件平台上加速人工智能模型。许多基于机器学习的微软产品都利用 ONNX Runtime 来加速推理性能。为了进一步降低推理延迟，ORT 团队的研发人员们首先将 PyTorch 版的 Aggressive Decoding 语法检查器，通过 PyTorch-ONNX 导出器导出为 ONNX 格式，再使用 ONNX Runtime 进行推理。ONNX Runtime 支持 Transformer 的特定优化以及 INT8 量化，这不仅实现了 Transformer 在 CPU 上的性能加速，同时还可以缩减模型大小。该端到端解决方案使用了多项前沿技术，以实现高效地运行这个先进的语法检查器模型。

> "ONNX Runtime 是一个具有很好延展性的跨硬件模型加速引擎，可以支持不同的应用场景。为了最高效运行 Aggressive Decoding 这一创新解码算法，我们对 PyTorch 导出器和 ONNX Runtime 做了一系列提升，最终让这一先进的语法检查器模型以最高性能运行。"
>
> —— 宁琼，微软首席产品主管

PyTorch 提供了一个内置函数，可以轻松地将 PyTorch 模型导出为 ONNX 格式。为了支持语法检查模型的独特架构，研发人员们在导出器里实现了复杂嵌套控制流导出到 ONNX，并扩展了官方 ONNX 规范来支持序列数据类型和运算符，以表示更复杂的场景，例如自回归搜索算法。这样就不需要单独导出模型编码器和解码器组件，再使用序列生成逻辑将它们串联在一起。由于 PyTorch-ONNX 导出器和 ONNX Runtime 支持序列数据类型和运算符，所以原模型可以导出成单一的一个包括编码器、解码器和序列生成的 ONNX 模型，这既带来了高效的计算，又简化了推理逻辑。此外，PyTorch ONNX 导出器的 shape type inference 组件也得到了增强，从而可以得到符合更严格的 ONNX shape type 约束下的有效的 ONNX 模型。

在语法检查器模型中引入的 Aggressive Decoding 算法最初是在 Fairseq 中实现的。为了使其与 ONNX 兼容以便于导出，研发人员们在 HuggingFace 中重新实现了 Aggressive Decoding 算法。在深入实施时，研发人员们发现 ONNX 标准运算符集不直接支持某些组件（例如分叉检测器）。目前有两种方法可以将不支持的运算符导出到 ONNX 并在 ONNX Runtime 中运行：1. 利用 ONNX 已有的基本运算符组建一个具有等效语义的图；2. 在 ONNX Runtime 中实现一个更高效的自定义运算符。ONNX Runtime 自定义运算符功能允许用户实现自己的运算符，以便灵活地在 ONNX Runtime 中运行。用户可以权衡实现成本和推理性能来选择合适的方法。考虑到本模型组件的复杂性，标准 ONNX 运算符的组合可能会带来性能瓶颈。因此，研发人员们选择在 ONNX Runtime 中实现自定义运算符。

ONNX Runtime 支持 Transformer 的优化和量化，这在 CPU 和 GPU 上都能提升性能。此外，ONNX Runtime 针对语法检查器模型进一步增强了编码器 attention 以及解码器 reshape 图算融合。支持该模型的另一大挑战是多个模型子图，而 ONNX Runtime Transformer 优化器和量化工具对此也实现了子图融合。ONNX Runtime 量化压缩已被应用于整个模型，进一步改善了吞吐量和延迟。

GPT-3.5 助力模型实现质的飞跃

为了进一步提高生产中模型的精度和召回率，研究员们使用了强大的 GPT-3.5 作为教师模型。具体而言，GPT-3.5 模型通过以下两种方式来帮助提高结果：

训练数据增强：通过对 GPT-3.5 模型进行微调，使其为大量未标注的文本生成标签。所获得的高质量标注，可以用作增强训练数据来提高模型性能。

训练数据清理：利用 GPT-3.5 强大的零样本和少样本学习能力来区分高质量和低质量的训练示例。然后，通过 GPT-3.5 模型重新对已识别的低质量示例生成标注，从而产生更干净、更高质量的训练集，直接增强模型性能。

EdgeFormer：用于客户端 seq2seq 建模的成本效益参数化

近年来，客户端设备的计算能力大大增加，使得利用深度神经网络来实现最终的零销货成本成为可能。然而，在这些设备上运行生成式语言模型仍然是一个很大的挑战，因为这些模型的内存效率必须受到严格的控制。在涉及生成式语言模型时，自然语言理解中用于神经网络的传统压缩方法往往不适用。

图 4：使用深度神经网络来实现零销货成本（zero-COGS）

运行在客户端的语法模型应该具有很高的效率（例如延迟在 100ms 内），这个问题已经由 Aggressive Decoding 解决了。此外，客户端模型还必须具有高效的内存（例如占用的空间在 50MB 以内），这是强大的 Transformer 模型（通常超过 5000 万个参数）在客户端设备上运行的主要瓶颈。

为了应对这一挑战，微软亚洲研究院的研究员们引入了前沿的客户端 seq2seq 建模技术 EdgeFormer[6]，用于构建性能优异的轻量级生成语言模型，让模型可以在用户的计算机上轻松运行。

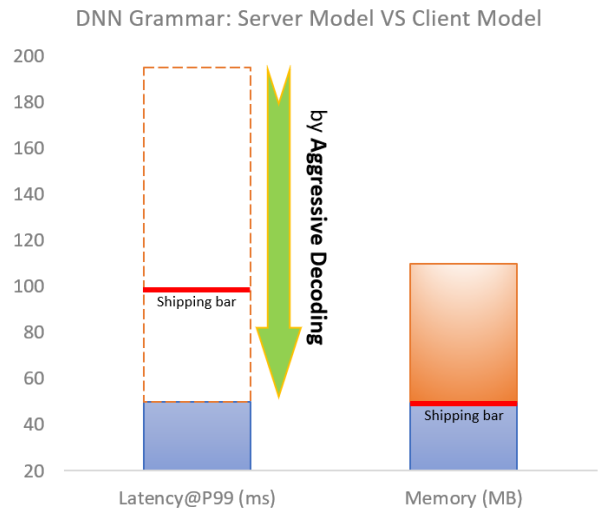


图 5：DNN 语法：服务器模型 VS 客户端模型

EdgeFormer 有两个原则，主要是为了参数化的成本效益：有利于编码器的参数化和负载均衡参数化。

遵循上述具有成本效益参数化原则而设计的 EdgeFormer，使得每个参数都能发挥最大潜力，即使客户端设备存在严格的计算和内存限制，也能获得有竞争力的结果。在 EdgeFormer 的基础上，研究员们进一步提出了 EdgeLM——EdgeFormer 的预训练版本，这是第一个在设备上公开可用的预训练 seq2seq 模型，可以让 seq2seq 任务的微调变得更容易，进而获得好的结果。EdgeLM 作为语法客户端模型的基础模型，实现了零销货成本，与服务器端模型相比，该模型以最小的质量损失实现了超过 5 倍的模型压缩。

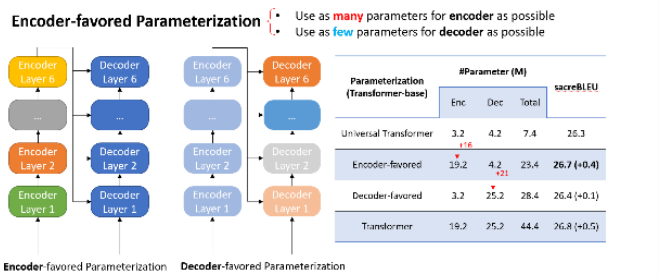


图 6：有利于编码器的参数化

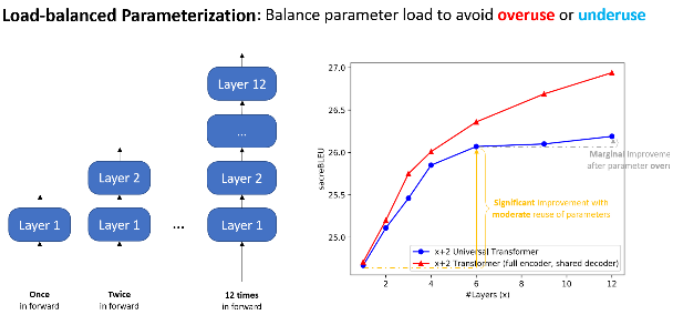


图 7：负载均衡参数化

> “微软亚洲研究院异构计算组致力于以全栈协同设计的思想，构建深度学习模型到实际设备部署之间的桥梁。以 Microsoft Editor 为例，我们与算法、产品和 AI 框架团队深度合作，通过系统和硬件感知的模型优化和压缩，以及针对不同硬件的推理系统和运算符优化等，使模型开销能够满足实际设备运行的要求，为未来将更多微软产品的 AI 服务部署到设备端铺平了道路。”
>
> —— 曹婷，微软亚洲研究院高级研究员

降低推理成本，赋能客户端部署

客户端设备的模型部署对硬件使用有严格的要求，如内存和磁盘使用量等，以避免干扰其他的应用程序。由于 ONNX Runtime 是一个轻量级的引擎并提供全面客户端推理解决方案（如 ONNX Runtime 量化和 ONNX Runtime 扩展），所以其在设备部署方面也具有明显的优势。此外，为了在保持服务质量的前提下满足交付要求，微软亚洲研究院引入了一系列优化技术，包括系统感知的模型优化、模型元数据简化、延迟参数加载以及定制量化策略。基于 EdgeFormer 建模，这些系统优化可以进一步将内存成本降低 2.7 倍，而不会降低模型性能，最终赋能模型在客户端设备的部署。

系统感知的模型优化。由于模型在推理系统中被表示为数据流图，因此该模型的主要内存成本来自于生成的许多子图。如图 8 所示，PyTorch 代码中的每个分支被映射为一个子图。所以，需要通过优化模型实现来减少分支指令的使用率。这其中尤为重要的是，因为波束搜索包含更多的分支指令，研究员们利用了贪婪搜索作为解码器搜索算法，从而将内存成本降低了 38%。

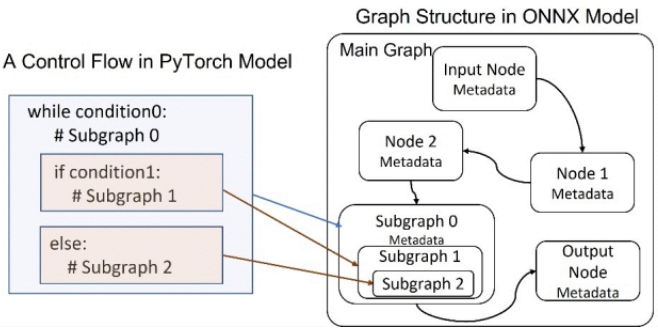


图 8：PyTorch 模型和 ONNX 模型图的映射

模型元数据简化。如图 8 所示，模型包含大量消耗内存的元数据，如节点名称和类型、输入和输出以及参数等。为了降低成本，研究员们需要简化元数据，只保留推理所需的基本信息，例如，节点名称从一个长字符串简化为一个索引。此外，研究员们也优化了 ONNX Runtime 模型图的实现，对所有子图只保留一个元数据副本，而不是在每次生成子图时复制所有可用的元数据。

延迟模型权重加载。当前的模型文件包含模型图和权重，并在模型初始化期间将它们一起加载到内存中。这会增加内存使用量，如图 9 所示，这是因为在模型图解析和转换过程中会重复复制权重。为了避免这种情况，研究员们提出将模型图和权重分别保存成独立的文件，并将该方法在 ONNX Runtime 加以实现。通过该方法，在初始化期间，只有模型图被加载到内存中进行实际解析和转换，而权重仍然留在磁盘上，通过文件映射只把权重文件指针（pointer）保留在内存中，实际的权重到内存的加载将被推迟到模型推理之时。该技术可将峰值内存成本降低 50%。

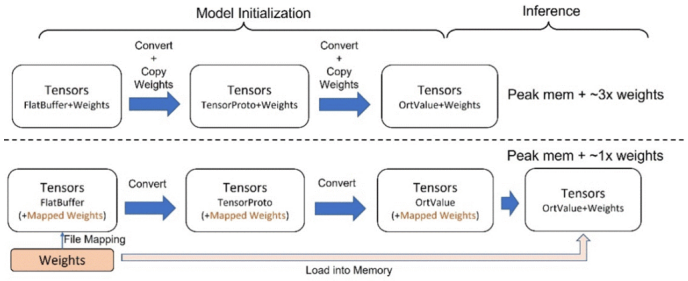


图 9：对比现有的模型图和权重同时加载（虚线上），以及模型初始化期间通过文件映射实现的延迟权重加载（虚线下）

ONNX Runtime 量化和扩展。量化是众所周知的模型压缩技术，它在牺牲模型精度的同时，带来了性能加速和模型缩减。ONNX Runtime 量化提供了多种微调选择，使其能够应用定制的量化策略。研发人员们为 EdgeFormer 模型定制了最优量化策略，以减少量化对精度的影响，具体包括训练后、动态和 UINT8 量化，以及 per-channel 和既有所有运算符量化策略。Onnxruntime-extensions 提供了一组 ONNX Runtime 定制运算符，以支持视觉、文本和自然语言处理模型的常见预处理和后处理运算符。利用这一工具，研发人员们将模型的预处理和后处理，例如标记化（tokenization）、字符串操作等，都集成到一个独立的 ONNX 模型文件中，从而提高性能、简化部署、减少内存使用率并提供更好的可移植性。

这些创新成果只是微软亚洲研究院为降低生成式语言模型的销货成本而做出的长期努力中的第一个里程碑。这些方法并不局限于加速神经语法检查器，它可以很容易地应用在抽象摘要、翻译或搜索引擎等广泛的场景中，从而加速降低大语言模型的销货成本 [5, 8]。在人工智能的未来发展中，这些创新对微软乃至对整个行业都将至关重要。

相关链接：

ONNX Runtime
https://onnxruntime.ai

EdgeFormer
https://www.microsoft.com/en-us/research/publication/edgeformer-a-parameter-efficient-transformer-for-on-device-seq2seq-generation/

EdgeLM
https://github.com/microsoft/unilm/tree/master/edgelm

ONNX Runtime 量化
https://onnxruntime.ai/docs/performance/model-optimizations/quantization.html

Onnxruntime-extensions
https://github.com/microsoft/onnxruntime-extensions

参考文献

[1] Tao Ge, Furu Wei, Ming Zhou: Fluency Boost Learning and Inference for Neural Grammatical Error Correction. In ACL 2018.

[2] Tao Ge, Furu Wei, Ming Zhou: Reaching Human-level Performance in Automatic Grammatical Error Correction: An Empirical Study.
https://arxiv.org/abs/1807.01270

[3] Xin Sun, Tao Ge, Shuming Ma, Jingjing Li, Furu Wei, Houfeng Wang: A Unified Strategy for Multilingual Grammatical Error Correction with Pre-trained Cross-lingual Language Model. In IJCAI 2022.

[4] Xin Sun, Tao Ge, Furu Wei, Houfeng Wang: Instantaneous Grammatical Error Correction with Shallow Aggressive Decoding. In ACL 2021.

[5] Tao Ge, Heming Xia, Xin Sun, Si-Qing Chen, Furu Wei: Lossless Acceleration for Seq2seq Generation with Aggressive Decoding.
https://arxiv.org/pdf/2205.10350.pdf

[6] Tao Ge, Si-Qing Chen, Furu Wei: EdgeFormer: A Parameter-efficient Transformer for On-device Seq2seq Generation. In EMNLP 2022.

[7] Heidorn, George. "Intelligent Writing Assistance." Handbook of Natural Language Processing. Robert Dale, Hermann L. Moisl, and H. L. Somers, editors. New York: Marcel Dekker, 2000: 181-207.

[8] Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, Furu Wei: Inference with Reference: Lossless Acceleration of Large Language Models.
https://arxiv.org/abs/2304.04487

ICLR 2023 杰出论文奖得主独家分享：适配任意密集预测任务的通用小样本学习器

国际学习表征会议 ICLR（International Conference on Learning Representations），被公认为当前最具影响力的机器学习国际学术会议之一。在今年的 ICLR 2023 大会上，微软亚洲研究院发表了在机器学习鲁棒性、负责任的人工智能等领域的最新研究成果。其中，微软亚洲研究院与韩国科学技术院（KAIST）在双方学术合作框架下的科研合作成果，因出色的清晰性、洞察力、创造力和潜在的持久影响获评 ICLR 2023 杰出论文奖。研究员们提出了首个适配所有密集预测任务的小样本学习器 VTM，以轻量化的迁移成本，赋予了计算机视觉模型预测新任务标签的能力，为计算机视觉中密集预测任务的处理以及小样本学习方法打开了全新思路。

密集预测任务是计算机视觉领域的一类重要任务，如语义分割、深度估计、边缘检测和关键点检测等。对于这类任务，手动标注像素级标签面临着难以承受的巨额成本。如何从少量的标注数据中学习并作出准确预测，即小样本学习，是该领域备受关注的课题。近年来，关于小样本学习的研究不断取得突破，尤其是一些基于元学习和对抗学习的方法，深受学术界的关注和欢迎。

然而，现有的计算机视觉小样本学习方法一般针对特定的某类任务，如分类任务或语义分割任务。它们通常在设计模型架构和训练过程中利用特定于这些任务的先验知识和假设，因此不适合推广到任意的密集预测任务。微软亚洲研究院的研究员们希望探究一个核心问题：是否存在一种通用的小样本学习器，可以从少量标记图像中学习任意段未见过的密集预测任务。

一个密集预测任务的目标是学习从输入图像到以像素为单位注释的标签的映射，它可以被定义为：

$$\mathcal{T}: \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times C_{\mathcal{T}}}, \quad C_{\mathcal{T}} \in \mathbb{N}.$$

其中 H 和 W 分别是图像的高与宽，输入图像一般包含 RGB 三个通道，C_T 表示输出通道的数目。不同的密集预测任务可能涉及不同的输出通道数目和通道属性，如语义分割任务的输出是多通道二值的，而深度估计任务的输出是单通道连续值的。一个通用的小样本学习器 F，对于任何这样的任务 T，在给定少量标记样本支持集 S_T（包含了 N 组样本 X^i 和标注 Y^i）的情况下，可以为未见过的查询图像 X^q 产生预测，即：

$$\hat{Y}^q = \mathcal{F}(X^q; \mathcal{S}_{\mathcal{T}}), \quad \mathcal{S}_{\mathcal{T}} = \{(X^i, Y^i)\}_{i\leq N}.$$

如果存在适配任意密集预测任务的通用小样本学习器，那么必须满足以下期望：

首先，它必须具备 e 统一的体系结构。该结构能够处理任意密集预测任务，并共享大多数任务所需的参数，以便获取可泛化的知识，从而能以小量样本学习任意未见过的任务。

其次，学习器应该灵活地调整其预测机制，以解决具有各种语义的未见过的任务，同时足够高效，以防止过度拟合。

因此，微软亚洲研究院的研究员们设计并实现了小样本学习器视觉令牌匹配 VTM（Visual Token Matching），其可用于任意的密集预测任务。这是首个适配所有密集预测任务的小样本学习器，VTM 为计算机视觉中密集预测任务的处理以及小样本学习方法打开了全新的思路。该工作获得了 ICLR 2023 杰出论文奖。

VTM 的设计灵感源于类比人类的思维过程：给定一个新任务的少量示例，人类可以根据示例之间的相似性快速将类似的输出分配给类似的输入，同时也可以根据给定的上下文灵活变通输入和输出之间在哪些层面相似。研究员们使用基于图像块（patch）级别的非参数匹配实现了密集预测的类比过程。通过训练，模型被启发出了捕捉图像块中相似性的能力。

给定一个新任务的少量标记示例，VTM 首先会根据给定的示例以及示例的标签调整其对相似性的理解，从示例图像块中锁定与待预测图像块相似的图像块，通过组合它们的标签来预测未见过的图像块的标签。

VTM 采用分层的编码器 - 解码器架构，在多个层次上实现了基于图像块的非参数匹配。它主要由四个模块组成，分别为图像编码器 f_T、标签编码器 g、匹配模块和标签解码器 h。给定查询图像和支持集，图像编码器首先会独立地提取每个查询和支持图像的图像块级表达。标签编码器也会类似地提取每个支持标签的标记。在每个层次的标记给定后，匹配模块会执行非参数匹配，最终由标签解码器推断出查询图像的标签。

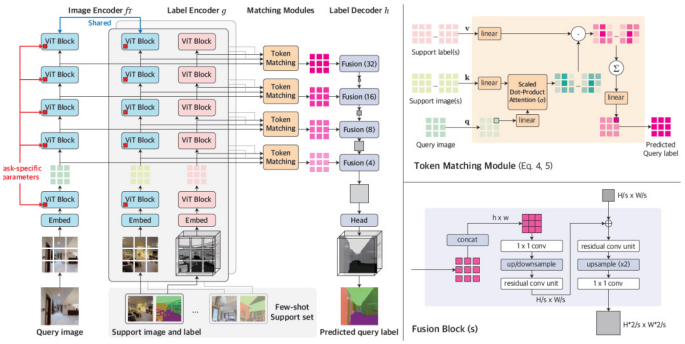


图 1：VTM 的整体架构

VTM 的本质是一个元学习方法。其训练由多个 episode 组成，每个 episode 模拟一个小样本学习问题。VTM 训练运用到了元训练数据集 D_train，其中包含多种有标签的密集预测任务示例。每个训练 episode 都会模拟数据集中特定任务 T_train 的小样本学习场景，目标是在给定支持集的条件下，为查询图像产生正确的标签。通过多个小样本学习的经验，模型能够学习到通用的知识，以便快速、灵活地适应新的任务。在测试时，模型需要在训练数据集 D_train 中未包含的任意任务 T_test 上进行小样本学习。

在处理任意任务时，由于元训练和测试中的每个任务的输出维度 C_T 不同，因此使得为所有任务设计统一的通用模型参数成为了巨大挑战。为了提供一个简单而普适的解决方案，研究员们将任务转换为 C_T 个单通道子任务，分别学习每个通道，并使用共享的模型 F 独立地对每个子任务进行建模。

为了测试 VTM，研究员们还特别构建了 Taskonomy 数据集的一个变种，从而模拟未见过的密集预测任务的小样本学习。Taskonomy 包含各种标注过的室内图像，研究员们从中选择了十个具有不同语义和输出维度的密集预测任务，将其分为五部分用于交叉验证。在每个拆分方式中，两个任务用于小样本评估（T_test），其余八个任务用于训练（T_train）。研究员们仔细构造了分区，使得训练和测试任务彼此有足够的差异，例如将边缘任务（TE，OE）分组为测试任务，以便对新语义的任务进行评估。

表 1 和图 2 分别定量与定性地展示了 VTM 和两类基线模型在十个密集预测任务上的小样本学习性能。其中，DPT 和 InvPT 是两种最先进的监督学习方法，DPT 可独立地针对每个单一任务进行训练，而 InvPT 则可以联合训练所有任务。由于在 VTM 之前还没有针对通用密集预测任务开发的专用小样本方法，因此研究员们将 VTM 与三种最先进的小样本分割方法，即 DGPNet、HSNet 和 VAT，进行对比，并把它们拓展到处理密集预测任务的一般标签空间。VTM 在训练期间没有访问测试任务 T_test，并且

仅在测试时使用了少量（10 张）的标记图像，但它却在所有小样本基线模型中表现得最好，并且在许多任务中的表现都具备与全监督基线模型比较的竞争力。

Supervision	Model	Tasks									
		Fold 1		Fold 2		Fold 3		Fold 4		Fold 5	
		SS	SN	ED	ZD	TE	OE	K2	K3	RS	PC
		mIoU ↑	mErr ↓	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓
Full	DPT	**0.4449**	**6.4414**	**0.0534**	**0.0268**	**0.0188**	**0.0689**	**0.0358**	**0.0357**	**0.0860**	**0.0347**
	InvPT	0.3900	12.9249	0.0589	0.0298	0.0517	0.0788	0.0456	0.0384	0.0949	0.0370
10-Shot (< 0.004%)	HSNet	0.1069	24.9120	0.2375	0.0748	0.1746	0.1643	0.1056	0.0651	0.2627	0.0610
	VAT	0.0353	25.8134	0.2718	0.0779	0.1719	0.1655	0.1450	0.0678	0.2709	0.0796
	DGPNet	0.0261	29.1668	0.4579	0.2846	0.1881	0.2130	0.1104	0.1308	0.3680	0.3574
	Ours	**0.4097**	**11.4391**	**0.0741**	**0.0316**	**0.0791**	**0.0912**	**0.0639**	**0.0519**	**0.1089**	**0.0420**

表 1：在 Taskonomy 数据集上的定量比较（ Few-shot 基线在训练了来自其他分区的任务后，在需测试的分区任务上进行了 10-shot 学习，其中完全监督的基线在每个 fold(DPT)或所有 fold(InvPT)上训练和评估了任务）

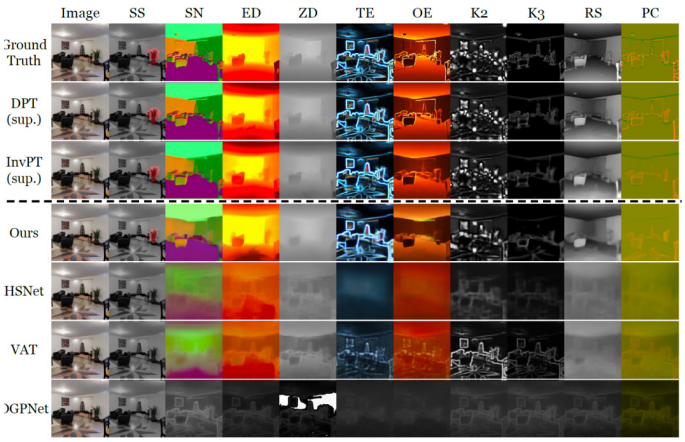


图 2：在 Taskonomy 的十个密集预测任务中，在新任务上仅提供十张标记图像的小样本学习方法的定性比较。在其他方法失败的情况下， VTM 成功地学习了所有具有不同语义和不同标签表示的新任务。

在图 2 中，虚线上方的分别是真实标签和两种监督学习方法 DPT 和 InvPT。虚线下方的是小样本学习方法。值得注意的是，其他小样本基线在新任务上出现了灾难性的欠拟合，而 VTM 成功地学习了所有任务。实验说明，VTM 可以在极少量的标记示例（<0.004%的完全监督）上现表现出与完全监督基线类似的竞争力，并能够在相对较少的附加数据（0.1%的完全监督）下缩小与监督方法的差距，甚至实现反超。

总结来说，尽管 VTM 的底层思路非常简单，但它具有统一的体系结构，可用于任意密集预测任务，因为匹配算法本质上包含所有任务和标签结构（例如，连续或离散）。此外，VTM 仅引入了少量的任务特定参数，就能具备抗过拟合性与灵活性。未来研究员们希望进一步探究预训练过程中的任务类型、数据量、以及数据分布对模型泛化性能的影响，从而帮助我们构建一个真正普适的小样本学习器。

相关论文：

Universal Few-shot Learning of Dense Prediction Tasks with Visual Token Matching
https://arxiv.org/abs/2303.14969

相关阅读

扫描二维码查看文章

ICLR 2023 | 更适合研究员体质的机器学习鲁棒性论文合集

本文将带来微软亚洲研究院在机器学习鲁棒性方面的四篇研究成果，其研究主题分别为领域泛化问题、分布外泛化、自适应阈值法与半监督学习中伪标签质量数量的权衡。

ICLR 2023 | 负责任的人工智能，守护机器学习的进阶思考

本文将与大家分享在微软亚洲研究院负责任的人工智能方向的三篇研究工作，它们分别拓展了差分隐私深度学习效率的边界、时序图的可解释性研究以及预训练语言模型在文本生成中的安全性。

LLM Accelerator：使用参考文本无损加速大语言模型推理

如今，基础大模型正在诸多应用中发挥着日益重要的作用。大多数大语言模型的训练都是采取自回归的方式进行生成，虽然自回归模型生成的文本质量有所保证，但却导致了高昂的推理成本和长时间的延迟。由于大模型的参数量巨大、推理成本高，因此如何在大规模部署大模型的过程中降低成本、减小延迟是一个关键课题。针对此问题，微软亚洲研究院的研究员们提出了一种使用参考文本无损加速大语言模型推理的方法 LLM Accelerator，在大模型典型的应用场景中可以取得两到三倍的加速。

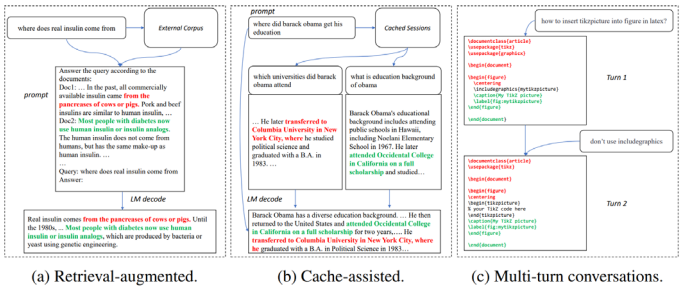
随着人工智能技术的快速发展，ChatGPT、New Bing、GPT-4 等新产品和新技术陆续发布，基础大模型在诸多应用中将发挥日益重要的作用。目前的大语言模型大多是自回归模型。自回归是指模型在输出时往往采用逐词输出的方式，即在输出每个词时，模型需要将之前输出的词作为输入。而这种自回归模式通常在输出时制约着并行加速器的充分利用。

在许多应用场景中，大模型的输出常常与一些参考文本有很大的相似性，例如在以下三个常见的场景中：

1. 检索增强的生成。New Bing 等检索应用在响应用户输入的内容时，会先返回一些与用户输入相关的信息，然后用语言模型总结检索出的信息，再回答用户输入的内容。在这种场景中，模型的输出往往包含大量检索结果中的文本片段。

2. 使用缓存的生成。大规模部署语言模型的过程中，历史的输入输出会被缓存。在处理新的输入时，检索应用会在缓存中寻找相似的输入。因此，模型的输出往往和缓存中对应的输出有很大的相似性。

3. 多轮对话中的生成。在使用 ChatGPT 等应用时，用户往往会根据模型的输出反复提出修改要求。在这种多轮对话的场景下，模型的多次输出往往只有少量的变化，重复度较高。



(a) Retrieval-augmented. (b) Cache-assisted. (c) Multi-turn conversations.

图 1：大模型的输出与参考文本存在相似性的常见场景

基于以上观察，研究员们以参考文本与模型输出的重复性作为突破自回归瓶颈的着力点，希望可以提高并行加速器利用率，加速大语言模型推理，进而提出了一种利用输出与参考文本的重复性来实现一步输出多个词的方法 LLM Accelerator。

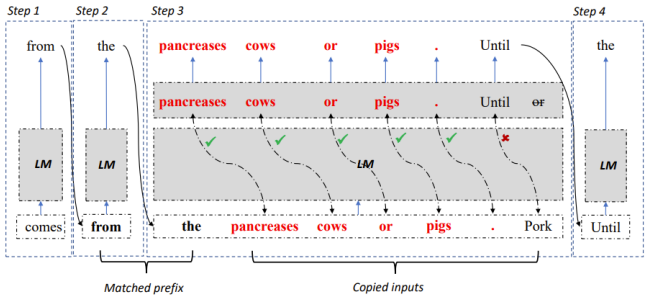


图 2：LLM Accelerator 解码算法

具体来说，在每一步解码时，让模型先匹配已有的输出结果与参考文本，如果发现某个参考文本与已有的输出相符，那么模型很可能顺延已有的参考文本继续输出。因此，研究员们将参考文本的后续词也作为输入加入到模型中，从而使得一个解码步骤可以输出多个词。

为了保证输入输出准确，研究员们进一步对比了模型输出的词与从参考文档输入的词。如果两者不一致，那么不正确的输入输出结果将被舍弃。以上方法能够保证解码结果与基准方法完全一致，并可以提高每个解码步骤的输出词数，从而实现大模型推理的无损加速。LLM Accelerator 无需额外辅助模型，简单易用，可以方便地部署到各种应用场景中。

使用 LLM Accelerator，有两个超参数需要调整。一是触发匹配机制所需的输出与参考文本的匹配词数：匹配词数越长往往越准确，可以更好地保证从参考文本拷贝的词是正确的输出，减少不必要的触发和计算；更短的匹配，解码步骤更少，潜在加速更快。二是每次拷贝词的数量：拷贝词数越多，加速潜力越大，但也可能造成更多不正确的输出被舍弃，浪费计算资源。研究员们通过实验发现，更加激进的策略（匹配单个词触发，一次拷贝 15 到 20 个词）往往能够取得更好的加速比。

为了验证 LLM Accelerator 的有效性，研究员们在检索增强和缓存辅助生成方面进行了实验，利用 MS-MARCO 段落检索数据集构造了实验样本。在检索增强实验中，研究员们使用检索模型对每个查询返回 10 个最相关的文档，然后拼接到查询后作为模型输入，将这 10 个文档作为参考文本。在缓存辅助生成实验中，每个查询生成四个相似的查询，然后用模型输出对应的查询作为参考文本。

Model	Tokens/sec ↑		Time (sec) ↓		Speed-up ↑
	baseline	LLMA	baseline	LLMA	
7B	23.9	59.2	511.2	206.0	2.48x
13B	18.5	41.1	658.4	296.8	2.22x
30B	4.9	12.1	2503.2	1005.8	2.49x

表 1：检索增强的生成场景下的时间对比

Model	Tokens/sec ↑		Time (sec) ↓		Speed-up ↑
	baseline	LLMA	baseline	LLMA	
7B	24.3	53.8	730.8	329.8	2.22x
13B	19.3	42.3	918.4	419.3	2.19x
30B	5.1	15.6	3467.7	1133.0	3.06x

表 2：使用缓存的生成场景下的时间对比

研究员们使用通过 OpenAI 接口得到的 Davinci-003 模型的输出结果作为目标输出，以获得高质量的输出。得到所需输入、输出和参考文本后，研究员们在开源的 LLaMA 语言模型上进行了实验。由于 LLaMA 模型的输出与 Davinci-003 输出不一致，所以研究员们采用了目标导向的解码方法来测试理想输出（Davinci-003 模型结果）结果下的加速比。研究员们利用算法 2 得到了贪婪解码时生成目标输出所需的解码步骤，并强制 LLaMA 模型按照得到的解码步骤进行解码。

对于参数量为 7B 和 13B 的模型，研究员们在单个 32G NVIDIA V100 GPU 上进行实验；对于参数量为 30B 的模型，在四块同样的 GPU 上进行实验。所有的实验均采用了半精度浮点数，解码均为贪婪解码，且批量大小为 1。实验结果表明，LLM Accelerator 在不同模型大小（7B，13B，30B）与不同的应用场景中（检索增强、缓存辅助）都取得了两到三倍的加速比。

进一步实验分析发现，LLM Accelertator 能显著减少所需的解码步骤，并且加速比与解码步骤的减少比例呈正相关。更少的解码步骤一方面意味着每个解码步骤生成的输出词数更多，可以提高 GPU 计算的计算效率；另一方面，对于需要多卡并行的 30B 模型，这意味着更少的多卡同步，从而达到更快的速度提升。在消融实验中，在开发集上对 LLM Accelertator 的超参数进行分析的结果显示，匹配单个单词（即触发拷贝机制）时，一次拷贝 15 到 20 个单词时的加速比可达到最大 （图 4 所示）。在图 5 中我们可以看出，匹配词数为 1 能更多地触发拷贝机制，并且随着拷贝长度的增加，每个解码步骤接受的输出词增加，解码步骤减少，从而达到更高的加速比。

Algorithm 2 Infer decoding sequence from target sequence and reference documents.

Input: $\boldsymbol{y}, \boldsymbol{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n), n, k$;
Output: $\boldsymbol{s} = (i_1, o_1), \ldots, (i_m, o_m)$;

$step \leftarrow 0$
$\boldsymbol{s} \leftarrow []$
while $step <$ LEN$(\boldsymbol{y})$ **do**
  $matched, d, pos \leftarrow$ MATCH_NGRAMS$(\boldsymbol{y}, step, D, n)$
  if $\neg matched$ **then**
    $step \leftarrow step + 1$
    APPEND$(\boldsymbol{s}, (1, 1))$
    continue
  end if
  $num\_valid \leftarrow$ GET_MATCHED_TOKENS$(d, pos, \boldsymbol{y}, step)$
  $num\_valid \leftarrow$ MIN(k, num_valid)
  $num\_output\_tokens \leftarrow num\_valid + 1$
  $step \leftarrow step + num\_output\_tokens$
  APPEND$(\boldsymbol{s}, (1 + k, num_output_tokens))$
end while

图 3：利用算法 2 得到了贪婪解码时生成目标输出所需的解码步骤

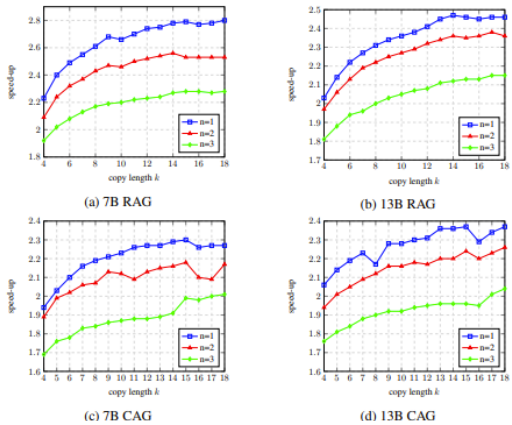


图 4：消融实验中，开发集上对 LLM Accelertator 的超参数的分析结果

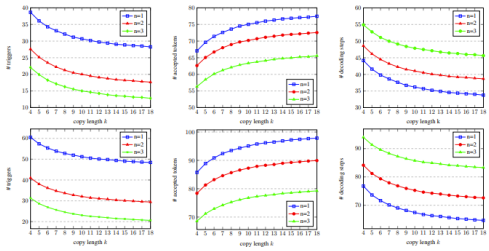


图 5：在开发集上，具有不同匹配词数 n 和拷贝词数 k 的解码步骤统计数据

LLM Accelertator 是微软亚洲研究院自然语言计算组在大语言模型加速系列工作的一部分，未来，研究员们将持续对相关问题进行更加深入的探索。

相关链接：

论文链接：
https://arxiv.org/pdf/2304.04487.pdf

项目链接：
https://github.com/microsoft/LMOps

LLM 时代，探索式数据分析的升级之路有哪些新助攻?

信息爆炸的时代，数据已经成为我们生活、工作中不可或缺的重要资源。大量的数据犹如一座座金矿，蕴藏着无尽的价值。但如果无法从数据中提取出知识和信息并加以有效利用，那么数据本身并不能驱动和引领技术应用取得成功。如何让数据发挥它最大的价值是数据、知识与智能（DKI）组持续探索的方向，为了实现这一目标，研究员们在探索式数据分析（Exploratory Data Analysis，EDA）领域进行了一系列的研究工作，相关成果已发表在 SIGMOD 2017、SIGMOD 2019、SIGMOD 2021、KDD 2022 和 SIGMOD 2023 等全球顶会上。该系列工作也已应用于 Microsoft Excel，Microsoft Power BI 和 Microsoft Forms 等微软产品中。

本文为大家介绍 DKI 组与香港科技大学合作完成的在探索式数据分析领域的成果。让我们一起探讨数据分析的外延和内涵在如何演变，一起了解在大语言模型 LLM（Large Language Model）的强力助攻下，运用数据洞察与因果信息来促进探索数据分析的潜力!

什么是数据洞察（Data Insight）?

"Insight" 在中文中可以翻译为"洞察"。而在这数据分析中，"insight" 是指从多维数据中发现的 interesting data pattern（有趣的数据模式），它反映了数据在某种特定视角下的有趣特征。例如:

在销售数据中，发现安装安卓系统的平板里，某种连接配置的销售额比别的配置的销售额高出很多。

在医疗数据中，发现某种疾病的新增患者数目随年份增加。

在社交媒体数据中，发现某个话题的讨论量存在周期性规律。

一个具体的洞察反应了人们对数据原始信号的某种特定规律的总结。它有如下特点：洞察是服务于数据分析任务的，往往是对原始信号的某种符号化抽象；相比于原始信号，洞察具有更高的分析语义。基于此，形式化的定义洞察以及如何从数据中挖掘潜在的洞察是一个首要任务。此外，可以预见到，由于洞察具有符号化的属性，基于这些属性的代数操作（当然需要有分析语义）可以进一步扩展分析的广阔空间。

QuickInsights 是一种能够快速自动发现多维数据中洞察的技术。它提出了一种统一定义洞察的抽象方法，将之前提出的不同类型的数据模式归到统一架构下，而且 QuickInsights 的挖掘框架的目标是自动发现高质量、高效率的数据洞察。来自 DKI 组的 QuickInsights 技术已经应用于 Microsoft Excel 和 Microsoft Power BI 等微软产品中。

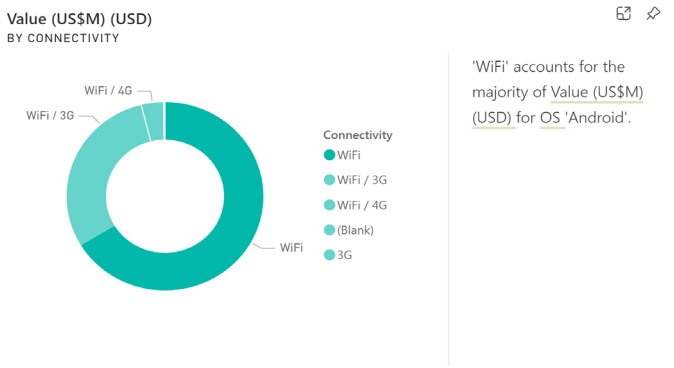


图 1：Microsoft Power BI 中的 QuickInsight

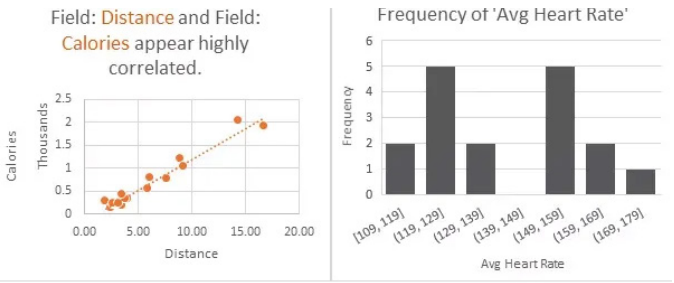


图 2：Microsoft Excel 中的 QuickInsight

QuickInsight 在以 Analysis Entity（AE，即由 <Subspace, Breakdown, Measure> 三元组构成的 data cube）为核心的传统数据分析范式的基础上进行了拓展。以 Analysis Entity 为基点，QuickInsight 将 <AE, Type, Property> 三元组作为数据分析的基本单元，其中 Type 表示该 AE 原始数据分布下的特征类型，如趋势（trend）、变更点（change point）等特征；Property 则对该特征的属性进行编码（即符号化），如趋势型特征的方向、变更点的具体位置等。Type 和 Property 的加入进一步扩展了数据分析的操作空间，打开了通往数据分析新视界的大门。

MetaInsight：数据洞察的内涵挖掘之道

在 QuickInsight 的框架下，数据洞察被表征为 <AE, Type, Property> 三元组。实际分析场景中，研究员们发现这样的洞察主要适用于服务基础的分析任务。一些复杂的分析任务，则需要对若干具有特定关系的洞察进行复合才能完成。这种复合操作赋

予了这些洞察作为一个整体，通过结构化关系而呈现出来的新的语义信息。基于这样的观察与思考，研究员们从数据分析过程中典型的意义构建机制（sensemaking mechanism）获得启发，提出了 MetaInsight。MetaInsight 是在基本数据洞察的基础上，利用多个 AE 内部的同质关系，形成更高层面的数据洞察，从而可更有效地推动探索性数据分析。

City	House Style	Month	Sales	…
Los Angeles	2Story	Jan	208,500	…
…	…	…	…	…
Los Angeles	1.5Fin	Dec	163,200	…
…	…	…	…	…
Yuba	1.5Unf	Dec	118,000	…

House Sales in California

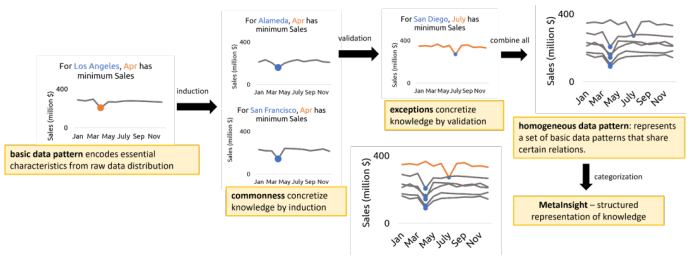


图 3：一个典型的探索式数据分析的过程

以一个典型的探索式数据分析的过程为例。如图 3 所示，假设有一个关于 2019 年美国加州房屋销售的多维数据。房地产经纪人 Bob 正在分析这个数据集。他首先注意到，与其他月份相比，洛杉矶在 4 月份的房屋销售量最低。然后，Bob 提出了一个归纳假设（inductive hypothesis）："其他城市的销售情况也同样糟糕。"通过进一步的数据探索，他了解到这是一个普遍现象，因为还有相当多的其他城市在 4 月份的销售表现也不佳（commonness，共性）。基于这种共性，Bob 现在开始关注异常情况，他提出了一个验证性质询（validity inquiry）："是不是所有城市在 4 月份的销售都很差，还是有例外？"经过对加州所有城市的仔细检查，Bob 发现圣地亚哥在 7 月份的销售情况不佳（exception，异常）。此时，Bob 已经完成了一个探索式数据分析流程，他从数据中得到了以下洞察：首先他提取了一条知识（洛杉矶在 4 月份的销售表现不佳），并进一步从数据中获得了结构化的知识（除圣地亚哥外，大多数城市在 4 月份的销售表现不佳）。因此，他希望将这个异常情况（圣地亚哥）作为进一步分析的新起点。

受人类数据分析过程中意义构建机制的启发，MetaInsight 可抽象出"归纳假设"和"验证性质询"两个关键机制，并由此在 QuickInsight 的基础上，获得更具结构化语义的数据洞察。依靠基本的数据洞察机制，MetaInsight 通过捕捉某个特定数据区域（data scope）原始数据分布的关键特征，可实现该数据区域的知识提取。在此基础上，MetaInsight 对其进行扩展，从而获得一系列语义上有归纳潜力的同质数据区域（homogeneous data scope）。例如，针对上述案例洛杉矶房价的这个数据区域，可以将其扩展为不同城市的房价同质数据区域。基于 QuickInsight 在初始数据区域的发现，进而评估同质数据区域内不同数据区域是否呈现出相似的特征，最终可获得具有结构化语义的 MetaInsight。

在 MetaInsight 中，研究员们还针对 Insight Scoring、Mining 以及 Ranking 等具体问题进行了研究，设计出了一套高效、可行的解决方案。

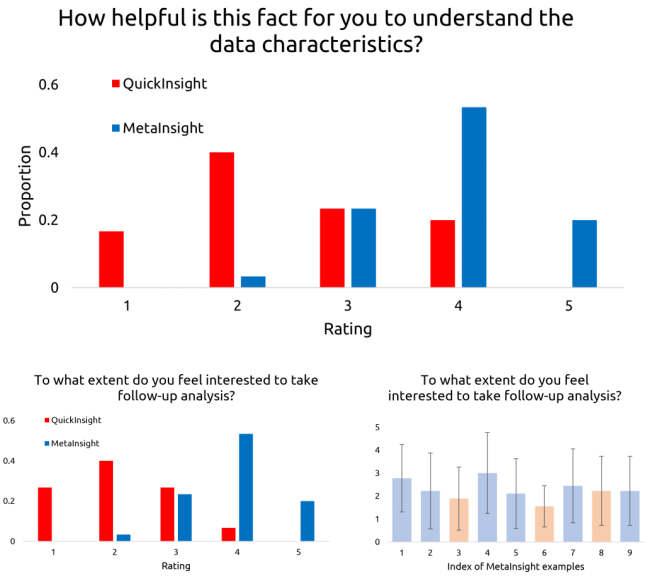


图 4：QuickInsight 和 MetaInsight 的用户使用反馈

一系列的用户研究表明，MetaInsight 能够更好地帮助用户理解数据、启发更深层次的数据探索。

XInsight：利用因果知识扩展数据洞察的外延

在数据分析场景中，可解释性是固有的一个需求。人们对各种数据中的现象进行观察，总结出适用概念，提出假设性的机制来解释观测，再提供预测，等等。这些分析过程中可解释性是普遍存在的。

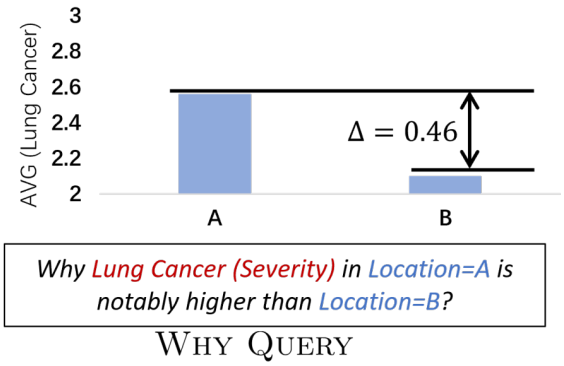
一个典型的场景是对异常数据进行解释。对异常的解释不仅在 MetaInsight 的使用中较为常见，而且更广泛地存在于日常数据分析的场景中。例如在销量数据中，用户希望探知某一年销量下降的原因；在游客数据中，用户希望解析出不同地域之间游客的差异。这种对解释的需求促使研究员们在传统探索式数据分析的基础上，提出了新的数据分析范式——可解释数据分析（Explainable Data Analysis, XDA）。

处理可解释数据分析的第一个要素是利用变量之间的因果关系。现有的工具往往基于相关性来生成数据的解释，但相关性并不意味着因果关系。基于相关性的数据解释往往会让用户产生因果幻觉（illusion of causality），而导致错误的数据理解和决策。

解释可以被分为因果（causal explanation）和非因果（non-causal explanation）两类。因果解释通过寻找因果因素来解释结果。例如图 5，在一个模拟的肺癌数据集中，患者的地理位置（反映地区控烟政策）和压力程度会影响他们是否吸烟，吸烟会进一步影响肺癌的严重程度，而肺癌严重程度又会进一步影响患者是否接受手术和五年存活率。

Location	Stress	Smoking	Lung Cancer	Surgery	5Y Survival
A	High (3)	Yes (1)	Severe (3)	Yes (1)	No (0)
…	…	…	…	…	…
B	Low (1)	No (0)	Mild (1)	No (0)	Yes (1)

(a) raw data



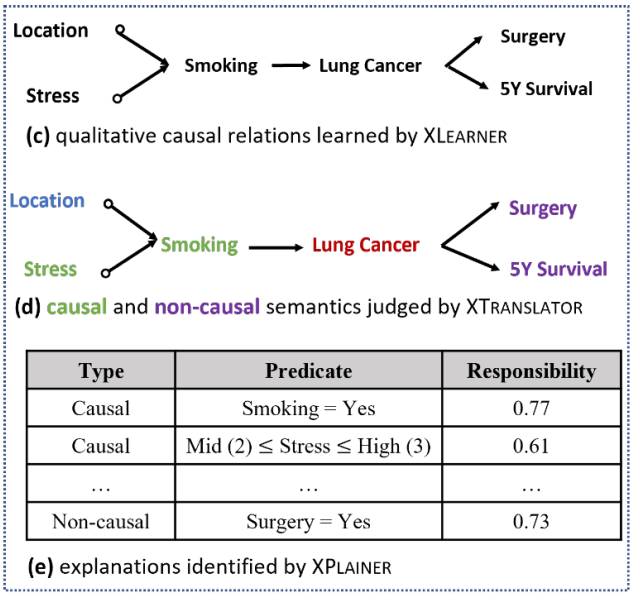
(b) typical EDA output and the derived WHYQUERY

图 5：肺癌数据集、不同地区的肺癌差异以及对应产生的 Why Query

吸烟为患者为什么患有不同严重程度的肺癌提供了解释。这是一个因果解释。相比之下，非因果的解释仅通过统计相关性来解释结果。例如，在肺癌数据集中，手术对肺癌严重程度有影响。但这并不是因果关系解释。因果关系解释还能让用户进行反事实思考（counterfactual thinking）和可行决策（actionable decision）。例如，戒烟可以降低肺癌的严重程度，而取消手术并不能改变肺癌的严重程度。这意味着，因果关系解释可以帮助用户制定更具针对性的决策，从而更好地解决实际问题。

基于这些发现，研究员们提出了 XInsight，旨在提供更可解释的数据洞察，减少因果关系的幻觉所带来的误解。XInsight 首先利用因果发现（causal discovery）定性地寻找数据内部变量之间的因果关系。基于定性的因果关系，XInsight 提出了 XTranslator，将因果图（causal graph）上的因果语义转化为可解释数据分析（eXplainable Data Analysis，XDA）的语义。根据 XDA 层面的语义信息，XInsight 再进一步利用实际因果性（actual causality）框架定量地分析每个变量内部不同取值对数据差异的影响，最终获得解释。

举一个使用 XInsight 进行数据洞察的例子，如图 7 所示。在前面展示的肺癌数据集中，分析师观察到一个有趣的现象："区域 A 的患者肺癌的平均严重程度比区域 B 高得多"，然后提出可能出现此现象的归因。通过 XInsight 可以得出一个因果关系的解释，即 "吸烟"是肺癌严重程度的定性因素，并且强调 "吸烟 = 是"（Smoking=Yes）及其作为定量解释的责任（responsibility）。用户能够据此更深入地理解不同区域患者肺癌严重程度之间的差异，并针对吸烟这一因素提出相应的干预措施，如制定更严格的烟草控制政策、推广禁烟宣传等，以达到更好的防控效果。



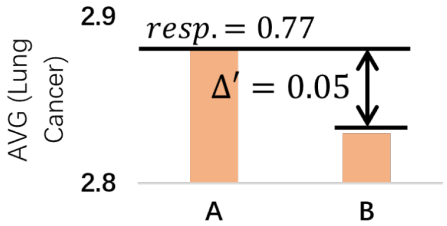
Type	Predicate	Responsibility
Causal	Smoking = Yes	0.77
Causal	Mid (2) ≤ Stress ≤ High (3)	0.61
…	…	…
Non-causal	Surgery = Yes	0.73

(e) explanations identified by XPLAINER

Three modules of XINSIGHT

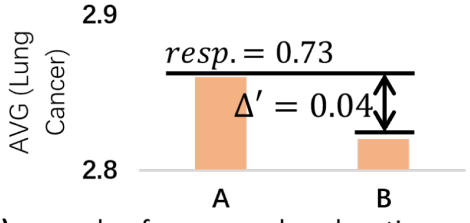
图 6：XInsight 的模型框架

Causal Explanation: "Factor=Smoking. Smoking=Yes" explains the difference on Lung Cancer between Location=A and Location=B.



(f) example of causal explanation

Non-causal Explanation: "Factor=Surgery. Surgery=Yes" is relevant to the difference on Lung Cancer between Location=A and Location=B.



(g) example of non-causal explanation

图 7：因果解释和非因果解释

尽管 XInsight 功能强大，但想要发挥出每个模块的理想功能，仍具有相当大的挑战。首先，大多数现实世界的数据集是在不考虑因果充分性（causal suffciency）的前提下收集的，即我们观察到的变量集合包含了所有的重要因果关系，没有遗漏任何关键的潜在变量。此外，在现实世界中，数据集中往往有一些数据与其他数据相关，比如学生的成绩由学习情况决定。这种数据间的关系称为函数依赖（Functional Dependency，FD），尤其是当它们从关系数据库中实例化时，函数依赖关系更为常见。但是这些 FD 有可能会让分析师误以为数据之间存在因果关系，但实际并没有。这就违反了忠实性假设（faithfulness assumption），这对于许多想要从数据间发现因果关系的算法是一个关隘。简单而言，忠实性假设是指在因果结构学习中，所有的统计依赖性都与某个因果关系相对应。为了应对这些挑战，研究员们建立了 FD-induced Graph 理论。XLearner 会使用 FD-induced Graph 从数据中选择一组不会违反忠实性假设的变量，并采用 FCI （fast causal inference）算法来处理因果不充分（causal insufficiency）的数据，然后将 FCI 的结果与 FD-induced Graph 中的因果关系相结合，进而获得最终的因果图。

其次，从因果原语（causal primitive，即因果图中的结构关系）到 XDA 语义（例如，变量是否提供因果解释）的转换尚未得到充分研究。因此目前尚不清楚如何确定变量 X 是否可以在给定上下文的情况下解释目标，以及 X 是否可以提供因果或非因果解释。XTranslator 从因果图中的各种因果原语（例如，m-separation，祖先 / 后代关系）出发，提供了一种系统性地将它们翻译成 XDA 语义的方法。

最后，在获取量化解释时，同样存在挑战。现有的数据库因果关系（DB Causality）主要用于数据溯源，通常以元（tuple）作为解释。相比于元组，谓词级解释（predicate-level explanation）更加简洁易懂，所以在 XDA 的场景中可能更为理想，由此也需要对原有的数据库因果关系理论进行适配。此外，数据库因果关系理论中，对一个潜在解释的定量分析具有相当高的计算复杂度（NP-Complete），由此也需要一个快速且有理论保证的近似算法来加速计算。XPlainer 针对数据分析中不同的聚合函数（aggregate function）设计了针对性的近似优化算法，既满足了数据分析的实时性要求，也提供了理论保证。

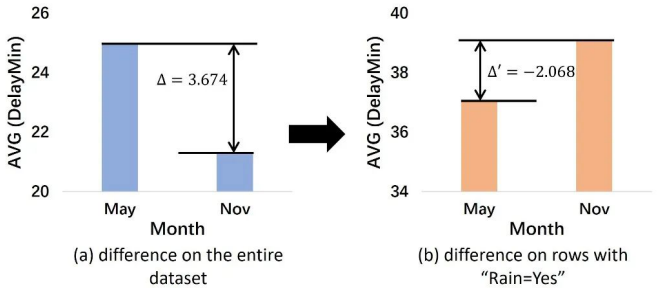


图 8：XInsight 检测出雨季可以对航班延误进行解释

在实验中，研究员们在多个公共数据集、私有数据集和合成数据集（synthetic dataset）上都验证了 XInsight 的有效性。例如在航班延误数据集中，如图 8 所示，XInsight 发现了雨季（Rain=Yes）是不同月份航班平均延误时间差异的原因。

从意图出发：LLM 导航的数据探索

MetaInsight 和 XInsight 所展现出的自动化数据探索工具的巨大潜力，都是基于某个特定的数据洞察，从而触发了数据分析意图（data analysis intent），它们的输出是数据分析意图在数据中的一种体现。但是，从另一个角度来看，这表示现有的解决方案的智能程度仍存在不足，MetaInsight 和 XInsight 还需要依赖人类数据分析师来找到特定的触发条件，并且在多数情况下，单一的数据洞察往往不能满足用户的真实需求。

以一位教育分析师 Alice 为例。她正在通过探索式数据分析，试图了解学生数学成绩的总体趋势。Alice 花费了数十分钟手动筛选和排序数据，并绘制了数学成绩随时间的变化图，发现数学成绩整体上升。接下来，她希望比较不同学校学生的表现，所以她又花费了更多的时间手动筛选数据，对比了 A、B、C 三所学校的学生数学成绩。她发现 A、B 两校的数学成绩呈现上升趋势，而 C 校在 2020 年出现了一个异常值。Alice 好奇地决定深入研究这个异常值。她花费了大量时间来回探索数据，通过各种变量进行筛选和分组，最终发现当排除 "Take-home" 考试时，C 校 2020 年的异常值便不再存在。于是，她记录下这一发现，并得出结论：这个异常值是由于 C 校在 2020 年改变考试形式所导致的。

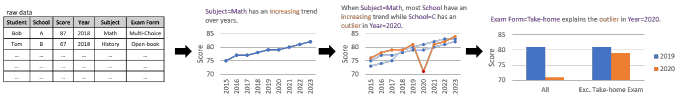


图 9：数据探索示例

正如上面的例子所展现的，数据探索远比提供一个特定的数据洞察要复杂，它既依赖于数据分析师的专业能力，同时也需要对特定数据集中的用户意图和领域知识有所掌握。为了解决这些问题，研究员们提出了一种更为智能的数据洞察工具 InsightPilot。InsightPilot 能够通过自动化的方式，帮助用户更高效地从数据中挖掘有价值的信息，减轻数据分析师的工作负担。相较于 MetaInsight 和 XInsight 等，InsightPilot 不仅可以自动执行数据分析任务，更能主动地从数据中寻找有价值的洞察，并为用户提供更全面、精准的数据分析结果。前面的示例中，Alice 需要花费大量时间对数据进行手动筛选、排序和分组。然而，借助 InsightPilot，这些繁琐的任务将被大大简化，节省了分析师的时间，让他们可以专注于深层次的数据洞察。

为了实现更为自动化的数据探索，InsightPilot 将数据探索的过程抽象为由 context-intent-analysis 组成的序列，即利用大语言

模型（LLM）理解已有的数据洞察（context）、提出一个合理的数据分析意图（intent），并由 Insight Engine 将意图转化为具体的数据分析（analysis）。通过利用数据分析的结果对数据洞察进行更新，让 LLM 推荐新的数据分析意图，从而迭代形成若干个由 context-intent-analysis 三元组完成的序列，从而使 InsightPilot 完成整个数据探索的过程。

图 10 演示了将 InsightPilot 应用于 Alice 的场景。首先，用户通过用户界面用自然语言提出一个问题："请展示学生数学成绩中值得挖掘的趋势。"然后，Insight Engine 会根据用户的问题从数据中生成初始洞察，并以自然语言的形式呈现给 LLM。例如，一个根据用户问题生成的洞察可能是："学校 A 的平均成绩排名第一。"针对用户的问题，LLM 从初始洞察中选择与用户问题最相关的洞察，因此选择了"学生的数学成绩随着时间的推移呈上升趋势。"

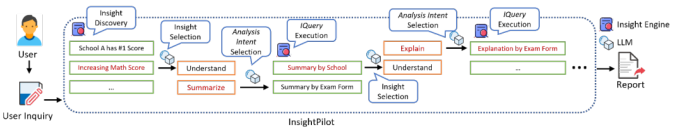


图 10：如何利用 InsightPilot 实现图 9 中的数据探索过程

接下来，当 LLM 选择了一个相关洞察后，Insight Engine 会对这个洞察进行分析，并向 LLM 提供可行的分析意图选项。在这个示例中，LLM 在"理解"和"归纳"两种意图中最终选择了"归纳"。然后，在收到选定的分析意图后，Insight Engine 将执行相应的查询语句并用一组新的洞察来回应分析意图。为了总结所选的洞察，Insight Engine 会尝试按照不同学校和考试形式来总结数学成绩的趋势。它发现："除了学校 C，大多数学校的学生数学成绩随着时间的推移呈上升趋势。学校 C 在 2020 年出现了一个异常值。"为了进一步探索数据，LLM 将再次选择洞察和分析意图与 Insight Engine 交互。此时，LLM 要求 Insight Engine "解释学校 C 在 2020 年的异常值。"

这种基于洞察和意图的交互将持续进行，直到 LLM 对探索结果满意或达到最大 token 数限制。在数据探索结束后，Insight Engine 将输出已探索洞察，由 LLM 将它们总结成一个有意义且连贯的报告呈现给用户。相比于使用 LLM 直接理解数据，InsightPilot 借助了 MetaInsight、XInsight 等解决方案来完成数据分析任务，为 LLM 提供结果，保障了数据探索的可靠性。

数据分析是一个以人为本的应用领域。希望通过上述的案例分析可以从研究的视角启发大家。从认识论角度来讲，人类探索世界的边界在不断扩大，理解世界的内涵也在丰富，这个道理同样适用于数据分析。而从效用角度来讲，数据分析的目标是完成实际生产生活中各种由数据驱动的任务，这些任务是多环节多角度构成的，但目前的数据分析工具，往往是散落在某个环节，从某个角度来解决一个局部问题，是辅助性的。因此，数据分析的舞台是足够大的，相信未来会有更多相关的科研成果诞生，进一步探索数据分析的智能化！

相关论文：

Extracting top-k insights from multi-dimensional data
https://dl.acm.org/doi/10.1145/3035918.3035922

MetaInsight: Automatic Discovery of Structured Knowledge for Exploratory Data Analysis
https://www.microsoft.com/en-us/research/uploads/prod/2021/03/metainsight-extended.pdf

Quickinsights: Quick and automatic discovery of insights from multi-dimensional data
https://dl.acm.org/doi/10.1145/3299869.3314037

XInsight: eXplainable Data Analysis Through The Lens of Causality
https://arxiv.org/abs/2207.12718

Demonstration of InsightPilot: An LLM-Empowered Automated Data Exploration System
https://arxiv.org/abs/2304.00477

语音合成模型 NaturalSpeech 2：只需几秒提示语音即可定制语音和歌声

如果问华语乐坛近期产量最高的歌手是谁，"AI 孙燕姿"一定有姓名。歌迷们先用歌手的音色训练 AI，再通过模型将其他歌曲转换成以歌手音色"翻唱"的歌曲。语音合成技术是"AI 孙燕姿"的背后支持。广义的语音合成包含文本到语音合成（Text to Speech，TTS）、声音转换等。在 TTS 领域，微软亚洲研究院机器学习组和微软 Azure 语音团队早已深耕多年，并在近期推出了语音合成模型 NaturalSpeech 2，只需几秒提示语音即可定制语音和歌声，省去了传统 TTS 前期训练过程，实现了零样本语音合成的跨越式发展。

文本到语音合成（Text to Speech，TTS）作为生成式人工智能（Generative AI 或 AIGC）的重要课题，在近年来取得了飞速发展。多年来，微软亚洲研究院机器学习组和微软 Azure 语音团队持续关注语音合成领域的研究与相关产品的研发。为了合成既自然又高质量的人类语音，NaturalSpeech 研究项目（https://aka.ms/speechresearch）应运而生。

NaturalSpeech 的研究分为以下几个阶段：

1）第一阶段，在单个说话人上取得媲美人类的语音质量。为此，研究团队在 2022 年推出了 NaturalSpeech 1，在 LJSpeech 语音合成数据集上达到了人类录音水平的音质。

2）第二阶段，高效地实现多样化的语音合成，包含不同的说话人、韵律、风格等。为此，该联合研究团队在 2023 年推出了 NaturalSpeech 2，利用扩散模型（diffusion model）实现了 zero-shot 的语音合成，只需要几秒钟的示例语音（speech prompt）模型就能合成任何说话人、韵律、风格的语音，实现了零样本语音合成的重要突破，为语音合成技术的未来发展带来了无限可能。

3）当前，研究团队正在开展第三阶段的研究，为达到高自然度（高质量且多样化）的语音合成这一目标，开创新局面。

三大创新设计，让 NaturalSpeech 2 脱颖而出

于近期发布的新一代语音合成大模型 NaturalSpeech 2，经历了上万小时、多说话人的语音数据集训练，并采用了 zero-shot（预测时只提供几秒钟的目标示例语音）的方式合成新的说话人、韵律、风格的语音，以实现多样化的语音合成。

要想达到良好的 zero-shot 训练效果，面临极大挑战。先前的方法是将语音量化成离散 token，并用自回归语言模型进行建模（例如 AudioLM）。但这种方法存在很大的局限性：自回归模型面临严重的错误传播（error-propagation）问题，导致生成语音质量低下、鲁棒性差，韵律失调以及重复、漏词等问题。同时还容易陷入离散 token 量化和自回归建模的两难困境（如表 1 所示），即要么离散 token 难以以较高质量还原语音，要么离散 token 难以预测。

The Dilemma in Previous Systems	Single Token (VQ)	Multiple Tokens (RVQ)
Waveform Reconstruction (Discrete Audio Codec)	Hard	Easy
Token Generation (Autoregressive Language Model)	Easy	Hard

表 1：先前语音合成系统的两难处境

NaturalSpeech 2 提出了一系列创新设计，如图 1 所示，完美地有效规避了先前的局限，实现了零样本语音合成的重要突破。考虑到语音波形的复杂性和高维度，微软亚洲研究院机器学习组与 Yoshua Bengio 共同提出的 Regeneration Learning 范式，为这个问题提供了创新的参考答案。

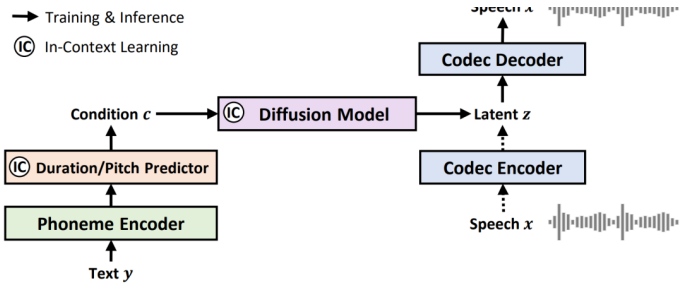


图 1：NaturalSpeech 2 系统概览

NaturalSpeech 2 首先利用神经语音编解码器（Neural Audio Codec，如图 2 所示）的编码器（encoder），将语音波形转换为连续向量并用解码器（decoder）重建语音波形，再运用潜在扩散模型（Latent Diffusion Model）以非自回归的方式从文本预测连续向量。在推理时，利用潜在扩散模型和神经语音解码器从文本生成语音的波形。

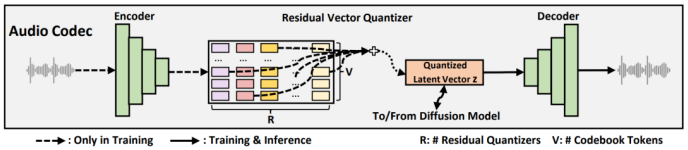


图 2：NaturalSpeech 2 中的 Neural Audio Codec 概览

相比先前的语音合成系统，NaturalSpeech 2 有以下几大优势，如表 2 所示：

Representations Generative Models In-Context Learning	Discrete Tokens Autoregressive Models Both Text and Speech are Needed	Continuous Vectors Non-Autoregressvie/Diffusion Only Speech is Needed
Stability/Robustness?	✗	✓
One Acoustic Model?	✗	✓
Beyond Speech (e.g., Singing)?	✗	✓

表 2：NaturalSpeech 2 相比先前语音合成系统的优势

1. 使用连续向量替代离散 token。离散 token 会导致序列长度过长（例如，使用 8 个残差向量量化器，序列长度会增加 8 倍），增加了预测的难度。而连续向量可以缩短序列长度，同时增加细粒度重建语音所需要的细节信息。

2. 采用扩散模型替代自回归语言模型。通过非自回归的生成方式，能避免自回归模型中的错误累积所导致的韵律不稳定、重复吐次漏词等问题。

3. 引入语音提示机制，激发上下文学习能力。研究员们通过创新设计的语音提示机制（如图 3 所示），让扩散模型和时长 / 音高预测模块能够更高效地学习语音上下文，从而提升了零样本的预测能力。

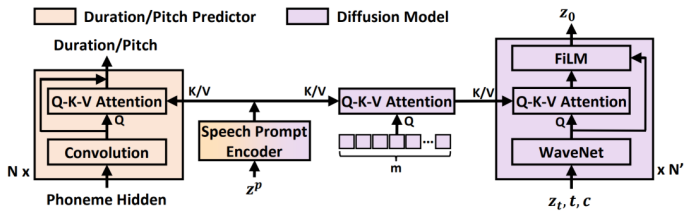


图 3：NaturalSpeech 2 中的语音提示机制

得益于以上设计，NaturalSpeech 2 生成的语音非常稳定、鲁棒，无需要复杂的两阶段模型来预测中间表征序列。同时，非自回归的方式和音高时长预测机制也赋予了 NaturalSpeech 2 扩展到语音之外的风格（例如歌声）的能力。

微软亚洲研究院高级研究员谭旭表示，语音合成是人工智能内容生成的一个非常重要的领域，该研究团队一直致力于构建高自然度的语音合成系统。NaturalSpeech 2 是继去年推出的 NaturalSpeech 后跨越的又一里程碑，利用大数据、大模型和零样本合成技术，极大地丰富了语音合成的音色、韵律、风格的多样性，使语音合成更自然更像人类。

NaturalSpeech 2 的语音合成性能大检测

研究团队将 NaturalSpeech 2 的模型大小扩展到了 400M，并基于 4.4 万小时的语音数据进行了训练。值得一提的是，即使 NaturalSpeech 2 与被模仿人“素昧平生”，只需几秒的语音提示， NaturalSpeech 2 输出的结果也可以在韵律 / 音色相似度、鲁棒性和音质方面都更优于先前的 TTS 系统。这一成果使得 NaturalSpeech 2 的性能达到了新高度，并有望为未来的 TTS 研究提供基础性参考。

首先，在音质方面，NaturalSpeech 2 在 zero-shot 条件合成的语音显著优于先前的 TTS 系统，如表 3 和表 4 所示。

Setting	LibriSpeech	VCTK
Ground Truth	−0.04	−0.30
YourTTS	−0.65	−0.58
NaturalSpeech 2	0.00	0.00

表 3：NaturalSpeech 2 和先前 TTS 系统的主观质量得分（CMOS）对比

Setting	CMOS (v.s. NaturalSpeech 2)
VALL-E	−0.31
NaturalSpeech 2	**0.00**

表 4：NaturalSpeech 2 和 VALL-E 的主观质量得分（CMOS）对比

同时，在相似度方面，NaturalSpeech 2 也能更好地生成和语音提示相似的语音，如表 5 和表 6 所示（评估指标详细介绍参见论文）。

LibriSpeech		Pitch				Duration			
		Mean↓	Std↓	Skew↓	Kurt↓	Mean↓	Std↓	Skew↓	Kurt↓
YourTTS	3s	10.52	7.62	0.59	1.18	0.84	**0.66**	0.75	3.70
NaturalSpeech 2	3s	**10.11**	**6.18**	**0.50**	**1.01**	**0.65**	0.70	**0.60**	**2.99**
YourTTS	5s	9.57	6.61	0.55	0.83	0.81	0.62	0.56	2.82
NaturalSpeech 2	5s	**6.96**	**4.29**	**0.42**	**0.77**	**0.69**	**0.60**	**0.53**	**2.52**
YourTTS	10s	7.13	6.35	0.89	1.46	0.75	0.55	0.61	2.77
NaturalSpeech 2	10s	**6.90**	**4.03**	**0.48**	**1.36**	**0.62**	**0.45**	**0.56**	**2.48**
VCTK		**Pitch**				**Duration**			
		Mean↓	Std↓	Skew↓	Kurt↓	Mean↓	Std↓	Skew↓	Kurt↓
YourTTS	3s	13.67	6.63	0.72	1.54	**0.72**	0.85	0.84	3.31
NaturalSpeech 2	3s	**13.29**	**6.41**	**0.68**	**1.27**	0.79	**0.76**	**0.76**	**2.65**
YourTTS	5s	14.61	6.02	0.70	1.33	0.76	0.70	0.82	3.49
NaturalSpeech 2	5s	**14.46**	**5.47**	**0.63**	**1.23**	**0.62**	**0.67**	**0.74**	**3.40**
YourTTS	10s	10.88	4.79	0.50	0.97	0.75	**0.62**	0.82	3.57
NaturalSpeech 2	10s	**10.28**	**4.31**	**0.41**	**0.87**	**0.71**	**0.62**	**0.76**	**3.48**

表 5：NaturalSpeech 2 与语音提示的韵律相似度比较

Setting	LibriSpeech	VCTK
GroundTruth	3.55	3.63
YourTTS	2.27	2.57
NaturalSpeech 2	3.29	3.16

表 6：NaturalSpeech 2 的主观相似度评分 SMOS 结果

在稳定度方面，相较于既有的 TTS 模型，NaturalSpeech 2 的表现也更为优异，如表 7 和表 8 所示。

Setting	LibriSpeech	VCTK
Ground Truth	1.94	9.49
YourTTS	7.10	14.80
NaturalSpeech 2	**2.26**	**6.99**

表 7：NaturalSpeech 2 合成语音的词错误率

AR/NAR	Model	Repeats	Skips	Error Sentences	Error Rate
AR	Tacotron [3]	4	11	12	24%
	Transformer TTS [5]	7	15	17	34%
NAR	FastSpeech [6]	0	0	0	0%
	NaturalSpeech [11]	0	0	0	0%
NAR	NaturalSpeech 2	0	0	0	0%

表 8：NaturalSpeech 2 合成语音的可懂度测试

研究员们还从互联网上收集了歌声数据，并将其与语音数据混合起来，共同训练模型。令人惊喜的是，无论是语音还是歌声提示，NaturalSpeech 2 都可以进行零样本歌声合成。欢迎访问链接：https://speechresearch.github.io/naturalspeech2/，一起听一听更多 AI 合成的语音和歌声吧！

随着合成语音质量的不断提升，确保 TTS 可被信赖是一个需要攻坚的问题。微软主动采取了一系列措施来预判和降低包括 TTS 在内的人工智能技术带来的风险。微软致力于依照以人为本的伦理原则推进人工智能的发展，早在 2018 年就发布了“公平、包容、可靠与安全、透明、隐私与保障、负责”6 个负责任的人工智能原则（Responsible AI Principles），随后又发布负责任的人工智能标准（Responsible AI Standards）将各项原则实施落地，并设置了治理架构确保团队把各项原则和标准落实到日常工作中。

未来，该研究团队将持续推动符合负责任的人工智能原则的语音合成大模型的研发，在更加多样化的场景中生成质量更高且更自然的语音，让语音合成技术可以赋能更多个人和组织。

更多研究成果请关注该团队研究主页 https://speechresearch.github.io/

相关链接：

论文链接：
https://arxiv.org/abs/2304.09116

项目演示：
https://speechresearch.github.io/naturalspeech2/

相关阅读

扫描二维码查看文章

NaturalSpeech 模型合成语音在 CMOS 测试中首次达到真人语音水平

AI 合成语音如今已经屡见不鲜，然而在用户听来却不能让人产生与真人对话和阅读般的沉浸感。微软亚洲研究院和微软 Azure 语音团队联合推出的全新端到端语音合成模型 NaturalSpeech，在 CMOS 测试中首次达到了真人说话水准。这将近一步提升微软 Azure 中合成语音的水平，让所有合成声音都惟妙惟肖。

CVPR 2023 | 掩码图像建模 MIM 的理解、局限与扩展

掩码图像建模（Masked Image Modeling, MIM）的提出，为计算机视觉模型训练引入无监督学习做出了重要贡献。得益于 MIM 的预训练算法，计算机视觉领域在近年来持续输出着优质的研究成果。然而整个业界对 MIM 机制的研究仍存在不足。

秉持着不断扩展前沿技术边界的探索精神，微软亚洲研究院的研究员们在理解 MIM 作用机制，以及基于这些机制提升现有 MIM 算法的领域，取得了一系列的创新成果，并获得了 CVPR 2023 的认可。这些成果包含：基于 MIM 预训练方法的扩展法则研究、分析 MIM 的具体性质以及有效性背后的原因、通过蒸馏技术将 MIM 模型的优势拓展到小模型中。

预训练 - 微调（Pre-training and Fine-tuning）是过去十年计算机视觉中最重要的学习范式之一，其基本想法是在海量数据的任务中，对神经网络进行训练，然后再将预训练过的模型在下游数据量较少的任务中进行微调。这种方式能够将上游大数据任务中学到的信息迁移至下游数据量较少的任务上，缓解数据量不足的问题，并显著提升模型的性能。

预训练 - 微调范式的成功，源于计算机视觉领域十年来预训练算法的停滞。自 2012 年 AlexNet 提出以来，计算机视觉中的预训练算法在很大程度上被等价于以 ImageNet 数据集为代表的图像分类任务。尽管图像分类的数据标注成本已然较低，但后续的数据清洗、质量控制等步骤仍对扩展图像分类数据产生了挑战，而数据不足的困难也限制了计算机视觉模型的进一步扩大。因此，如何使用无监督学习方法进行视觉模型的预训练逐渐成为了计算机视觉任务中的核心问题。

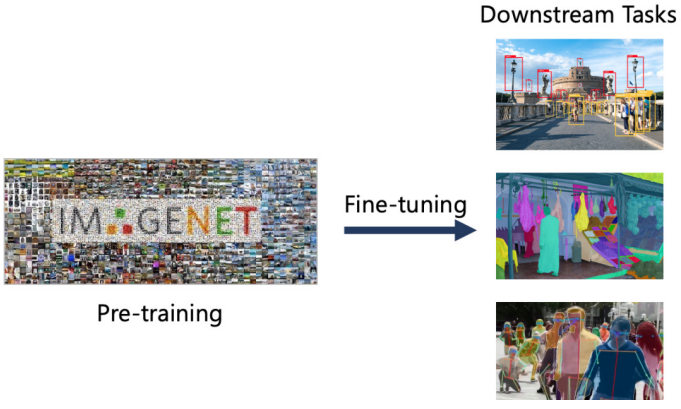


图 1：预训练 - 微调范式

2021 年 6 月，微软亚洲研究院提出的 BEiT 方法，通过引入自然语言处理（NLP）中的掩码语言建模（Masked Language Modeling, MLM）算法，成功地证明了计算机视觉中无监督预训练可以达到与有监督预训练相同甚至更好的效果。2021 年 11 月，微软亚洲研究院提出的 SimMIM 与 Meta 提出的 MAE 进一步简化了 BEiT，并提升了算法性能。自此，掩码图像建模（Masked Image Modeling, MIM）的研究范式正式开启。

虽然基于 MIM 的预训练算法的成果在计算机视觉领域内百花齐放，但对 MIM 机制的探索仍然十分匮乏。今天我们将介绍微软亚洲研究院视觉计算组在理解 MIM 作用机制，以及基于这些机制扩展并提升现有 MIM 算法的系列工作。

探索 MIM 的扩展法则与数据可扩展性

扩展法则（scaling law）的概念最初由 OpenAI 发表于 2020 年的"Scaling Laws for Neural Language Models"，文中提出：测试集上的 Loss 会随着计算（compute）、数据规模（dataset size）与模型参数量（parameters）的增加而呈现可以预测的下降模式。该发现对于如何优化自然语言模型的设计与训练，具有里程碑式的指导意义。

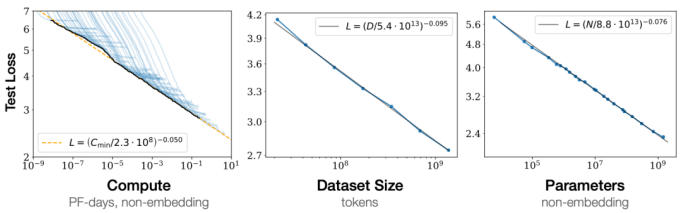


图 2：自然语言处理中的扩展法则：测试集 Loss 随着计算，数据规模以及模型参数的增加呈现可以预测的下降模式

在入选 CVPR 2023 的"On Data Scaling in Masked Image Modeling"（论文链接：https://arxiv.org/abs/2206.04664）一文中，微软亚洲研究院的研究员们也探索了基于 MIM 预训练方法的扩展法则。尽管在计算与模型大小这两个维度中，MIM 预训练算法也呈现了较好的扩展性质，但是在数据维度上，MIM 算法则呈现了与在 NLP 中截然不同的特性：测试集的 Loss 随着数据集大小达

到一定规模后不再降低，呈饱和状。这引发了一个关键问题——作为一个无监督预训练算法，MIM 是否能从更多的数据中受益？换言之，MIM 是否具有数据可扩展性？

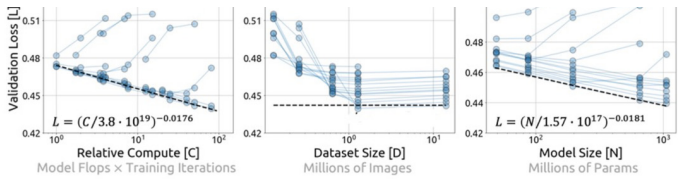


图 3：MIM 中的扩展法则：测试集 Loss 仅随着计算与模型参数的增加呈现可预测的下降模式，而在数据集大小维度上，呈现了饱和的现象

为了回答该问题，研究员们分析了模型大小、数据规模以及训练长度的影响，发现 MIM 具有数据可扩展性，但需要满足两个关键的条件：1）需要更大的模型；2）需要配以更长的训练轮数。进一步的观察表明，该现象是由过拟合（over-fitting）导致的。

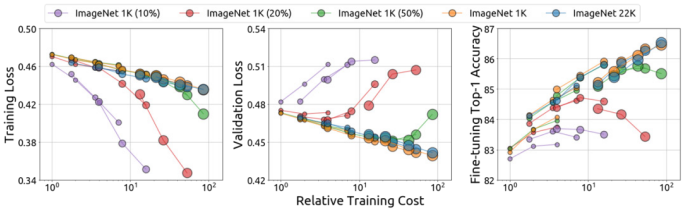


图 4：MIM 中的过拟合现象

如图 4 所示，对一些较大的模型，使用小数据与长训练轮数会使得训练 Loss 异常下降，测试 Loss 与下游任务中的微调性能受损。同时，过拟合状态时，模型更倾向于呈现记忆图像的性质；非过拟合状态时，更倾向呈现推理的性质。基于这些发现，研究员们认为对于 MIM 而言，测试集 Loss 比训练 Loss 更适合作为下游任务迁移能力的代理指标。

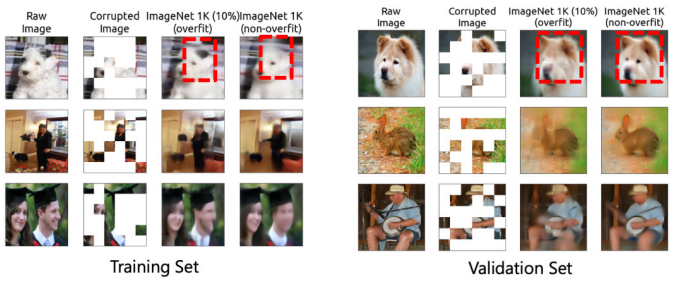


图 5：展示了过拟合模型与非过拟合模型对图片补全的可视化结果。过拟合模型在训练集上会记忆原图，而在测试集上则无法正确的推理内容。非过拟合模型则在训练集与测试集图像上都表现出较好的推理能力。

更深入地理解 MIM 及其有效性

MIM 展示了其在预训练 - 微调范式下的广泛有效性。传统视角通常认为模型的有效性取决于其提取的特征质量。然而，进一步实验发现，在固定网络权重的设定时，MIM 的性能远逊色于其他预训练算法。这说明 MIM 的有效性源自其他因素。

Approach	COCO				ADE20K		Kinetics-400	
	Frozen		Full ft.		Frozen	Full ft.	Frozen	Full ft.
	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}	mIoU	mIoU	acc@1	acc@1
SUP-1K	42.4	38.7	50.5	44.5	49.8	52.3	60.4	77.0
SUP-22K	45.0	41.1	51.9	45.7	51.9	55.3	70.3	79.7
EsViT-1K	42.0	38.5	51.5	45.6	49.7	52.1	62.0	76.5
SimMIM-1K	34.1	32.4	52.9	46.7	42.4	51.7	14.2	75.9
iCAR-Laion	43.3	39.5	51.7	45.5	51.0	55.3	65.1	79.5
iCAR-Laion-22K	44.9	41.2	52.3	46.1	51.1	55.4	69.4	80.2

表 1：在固定网络权重（frozen setting）与微调全网络权重（full fine-tuning setting）下，比较不同预训练算法的性能

于是，在微软亚洲研究院入选 CVPR 2023 的另一篇论文“Revealing the Dark Secrets of Masked Image Modeling”中（论文链接：https://arxiv.org/abs/2205.13543），研究员们对 MIM 的性质以及有效性背后的原因进行了更细致的研究与分析，取得了如下发现：

1） 有监督预训练以及基于对比学习的预训练方法的深层网络仅建模长程信息，相比之下，MIM 能够对局部信息与长程信息同时建模，如图 6 所示。

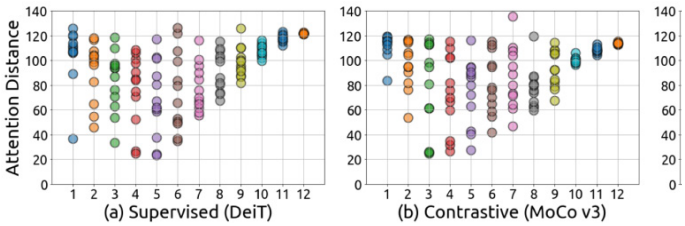


图 6：不同模型中的注意力距离（attention distance）。有监督预训练与对比学习预训练算法在网络的深层只关注长程信息，而 MIM 方法同时关注长程信息与局部信息。

2） MIM 中不同注意力头（attention head）关注的信息具有多样，如图 7 所示，在有监督预训练与对比学习预训练算法中，网络的较深层注意力模块里，不同注意力头关注的信息是趋同的，而在 MIM 中，不同注意力头关注的信息更多样，这在一定程度上避免了模型塌陷（model collapse）的问题。

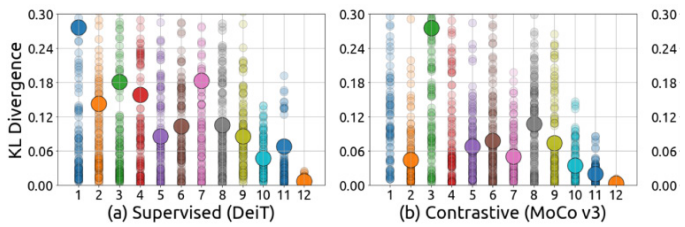


图 7：不同模型中注意力头对应的注意力地图（attention map）的多样性分析

3） MIM 对语义信息的刻画较少，但是对几何信息的刻画较多。因此，研究员们对比了监督学习与 MIM 在语义分类任务中和几何任务中的性能表现。结果显示 MIM 在语义分类任务中的性能表现较差，但是在几何任务中的性能表现较好，如表 2 所示。同时，研究员们还考察了在混合任务（如物体检测）中，MIM 与有监督预训练在分类与定位两个子任务上的性能变化情况（如图 8 所示）。结果也显示 MIM 在分类任务上的收敛速度比有监督预训练差，但是在定位任务上收敛性更好。

	Sup	MIM
K12-Set	**89.7**	86.1
Oxford Pets	**95.9**	90.9
Caltech101	**91.9**	85.5
SUN397	**72.3**	70.8

	Sup	MIM
Pose	75.9	**77.6**
Depth	0.335	**0.304**
Obj. Tracking	67.8	**70.0**

表 2：（左）语义分类任务的性能比较；（右）几何任务的性能比较

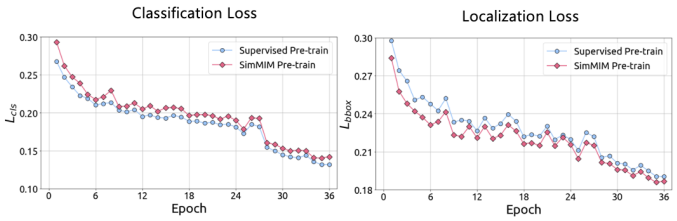


图 8：物体检测任务中的分类损失与定位损失

扩展 MIM 在小模型上的有效性

在关于 MIM 的早期论文中，科研人员普遍发现 MIM 方法对大模型更加友好，直接在小模型中使用的有效性欠佳。如表 3 所示，在 ViT-T 等较小的模型中使用 MIM 预训练算法，其性能甚至落后于随机初始化的模型。如何将 MIM 应用于小模型，是领域中一个重要的开放性问题。在另外一篇入选 CVPR 2023 的工作"TinyMIM: An Empirical Study of Distilling MIM Pre-trained Models"中（论文链接：https://arxiv.org/abs/2301.01296），微软亚洲研究院的研究员们通过蒸馏（distillation）技术，成功将 MIM 模型的优势拓展到了小模型中。

Method	ViT-T	ViT-S	ViT-B	ViT-L
Scratch	72.2	79.9	81.2	82.6
MAE	71.6	80.6	83.6	85.9
Gap	-0.6	+0.7	+2.4	+3.3

表 3：不同大小模型下，使用 MIM 预训练与随机初始化模型的性能比较

TinyMIM 中，研究员们系统性地研究了如何使用经过 MIM 预训练的模型蒸馏至小模型中。其研究对象包括输入形式，蒸馏对象，以及蒸馏方法三个方面。通过广泛的实验，研究人员发现：

1） 直接蒸馏元素间的关系（relation）是 MIM 中最有效的蒸馏方式，其性能可以比蒸馏 CLS Token 在 ViT-T 上好 4.2 Top-1 Acc，在 ViT-B 上好 1.6 Top-1 Acc.

Method	Model Size	Top-1 Acc.
Supervised (DeiT)	ViT-T	72.2
MAE		71.6
Class Token Distillation		70.6
Feature Distillation		73.4
Relation Distillation		75.8 (**+4.2**)
Supervised (DeiT)	ViT-S	79.9
MAE		80.6
Class Token Distillation		79.6
Feature Distillation		80.8
Relation Distillation		83.0 (**+3.1**)
Supervised (DeiT)	ViT-B	81.2
MAE		83.6
Class Token Distillation		82.6
Feature Distillation		83.8
Relation Distillation		85.0 (**+1.6**)

表 4：不同蒸馏对象对结果的影响

2） 在蒸馏时引入 MIM 任务会损害性能。如表 5 所示，使用掩码图像作为输入，以及在蒸馏时引入图像重构任务，都会损害模型的蒸馏效果。

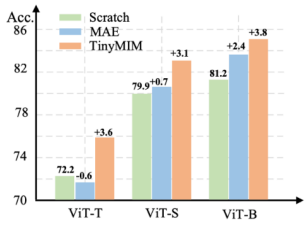
Masked Image	Reconstruction Loss	Top-1 Acc.
		84.6
✓		83.9
✓	✓	84.0

表 5：蒸馏时引入 MIM 任务对性能的影响。

Student	Teacher	Acc.
ViT-S	MAE-ViT-B	82.3
	MAE-ViT-L	82.1
	MAE-ViT-L → TinyMIM-ViT-B	82.6
ViT-T	MAE-ViT-S	74.1
	MAE-ViT-B	74.4
	MAE-ViT-B → TinyMIM-ViT-S	75.0

表 6：序列化蒸馏对结果的影响

3） 序列化蒸馏可以进一步降低难度，提升性能。序列化蒸馏指的是使用小模型蒸馏大模型的过程中，引入中等规模模型进行蒸馏，即先蒸馏出一个中等大小的模型，再利用该模型去蒸馏小模型。这样的蒸馏方式可以获得更好的性能，如表 6 所示。结合上述发现，TinyMIM 在一系列中小型模型中均取得了显著的性能提升，相较于其他直接训练的小模型，如 MobileViT 等，也取得了更好的下游任务迁移能力，如图 9 所示。



Model	Param. (M)	Flops (G)	Top-1 (%)	mIoU
DeiT-T [44]	5.5	1.3	72.2	38.0
PVT-T [46]	13.0	1.9	75.1	39.8
CiT-T [39]	5.5	1.3	75.3	38.5
Swin [32]	8.8	1.2	76.9	40.4
EdgeViT-XS [35]	6.4	1.1	77.5	42.1
MobileViTv1-S [34]	4.9	2.0	78.4	42.7
MobileViTv3-S [45]	4.8	1.8	79.3	43.1
TinyMIM*-T (Ours)	5.8	1.3	**79.6**	**45.0**

图 9：TinyMIM 相较于 MAE 与其他小模型设计方法均取得了显著的性能优势

展望未来

随着计算机视觉中预训练范式从有监督学习逐渐演变至自监督学习，科研人员对视觉智能的认识与理解也在不断改变与深化。现阶段，基于掩码图像建模（MIM）的图像预训练算法已经展现出无监督预训练强大的潜力，但是否存在更适合视觉信号的预训练方法仍然是领域内最重要的开放问题之一。此外，在视觉与语言大一统的发展趋势之下，如何有效利用掩码信号建模等预训练算法高效连接语言与视觉信号的问题也仍需探索。微软亚洲研究院的研究员们希望随着对掩码图像建模预训练算法理解与认识的深化，研究并提出更高效的预训练算法，促进视觉智能迈入下一个发展新阶段。

相关阅读

扫描二维码查看文章

CVPR 2023 | 计算机视觉顶会亮点解读

本文为大家带来 5 篇微软亚洲研究院被 CVPR 2023 收录的论文，主题涵盖手语识别与检索、多模态生成、图像编辑、视频理解任务等。

科研第一线

WWW 2023 | 一键追更互联网技术国际顶会的最新科研进展

国际万维网会议（Proceedings of the ACM Web Conference，简称 WWW）是互联网技术领域的顶级学术会议之一，汇集了国际一流学者与产业界精英，持续关注着互联网技术的学术研究前沿与热门发展方向。在今年的 WWW 2023 大会上，有多篇来自微软亚洲研究院的论文被录用。文章精选了其中的六篇进行简要介绍，研究主题涵盖算法公平、知识蒸馏、推荐系统与图自监督学习等。欢迎扫码二维码，了解互联网技术研究的最新进展。

扫描二维码了解更多信息

ICSE 2023 | 为计算平台的高质量运行保驾护航

ICSE 是软件工程领域公认的权威国际学术顶会。自 1975 年创办以来，ICSE 大会持续为研究人员和相关从业者输送着软件工程领域中最新的理论创新、技术趋势与研究成果。本届 ICSE 大会接受了多篇来自微软亚洲研究院的成果，在云计算高速发展的当下，更需要研究员们不断深耕，发掘研究洞见，为云平台的高质量运行保驾护航。

扫描二维码了解更多信息

科学匠人 | 对话陈卫：为什么 AI 大模型时代更需要计算机理论研究？

2021 年 12 月，微软亚洲研究院成立了理论中心，由微软亚洲研究院高级研究员陈卫博士担任中心主任。从高中第一次接触编程起，陈卫就对计算机产生了浓厚的兴趣并结下了不解之缘。在获得保送北京大学数学系资格的情况下，陈卫还是通过高考进入了清华大学计算机系，开启了计算机理论的学习与研究生涯。

在康奈尔大学读博时，陈卫的研究偏向分布式计算理论，这是他的第一个主要研究方向，也是他加入微软亚洲研究院系统组后持续探索的领域。在微软亚洲研究院工作的十九年中，陈卫参与过众多研究项目，从在系统组对分布式计算进行的论证和分析，到在理论组基于影响力最大化与在线学习和优化展开的系列研究，再到今天的理论中心，理论研究一直是他不变的初心。

那么，在人工智能大模型当道的今天，陈卫如何看待计算机理论研究的意义？微软亚洲研究院理论中心又将有哪些新的研究方向？让我们通过对话听一听计算机理论科学家陈卫的想法。

微软亚洲研究院高级研究员陈卫

Q：如何理解计算机理论研究，它都包含了哪些内容？有怎样的作用？你为何在这个“冷门”领域深耕多年？

陈卫：计算机理论是一个横向概念，它以数学为基础和工具来研究计算机科学的各个方面。在传统的计算机科学中，理论是操作系统、编译原理、网页搜索、图形学等所有这些领域的指导基础，例如，算法理论、复杂性理论等。随着计算机技术的不断发展，理论方向又延伸出了深度学习理论、数据驱动优化理论等。如果特指计算机科学理论的话，那么就是指结合数学工具和计算机应用背景来设计优秀的算法，并在计算机上生成代码去运行和验证。

事实上，计算机科学理论是一个交叉学科，它并不是一个独立的学科。首先它以数学作为基础，需要用严格的数学概念和方法对计算进行建模和分析，数学中的代数、分析、概率论、统计学等都是重要的工具。其次，它也涉及物理学的很多方面，包括涌现、相变理论等。我研究多年的网络科学也是计算机理论的一个分支，而网络科学本身也是一门交叉学科，其研究对象——网络形态，不仅指人的社交网络，也包括神经网络、蛋白质互动网络等。此外，随着人工智能技术的发展，理论研究还要纳入心理学、社会学等社会科学，研究将更趋于综合性。

我从事理论研究 20 余年，对这一领域始终充满热爱的原因是我可以通过自己的分析能力来推导、发现新的算法和新的理论，从而证明技术的可行性。这种依赖理性的思维和分析帮助我们更好地认识世界和改造世界，是让我对理论研究着迷的原因和动力。

Q：GPT-4 的出现，进一步证明了人工智能模型越大性能越强的论断。那么，作为理论研究专家，你如何看待大模型？大模型对理论研究有什么推进作用？理论研究又能帮助大模型解决什么问题？

陈卫：大模型的横空出世带来了很多变革，也第一次在应用上让普通大众切身感受到人工智能的实用性。从理论研究看，对于人工智能这个黑盒子，我们更习惯于问一些问题，比如它为什么如此强大？它的能力边界在哪里？有什么它不能做的？

目前来看，数学计算和推理是大模型的弱项。以加法为例，大模型是从左到右地概率预测每一位数，而人类是从右向左计算，思路是相反的。即使模型明确知道加法规则，但其内部的生成也不会按照规则运行，这就是概率生成模型的一个能力边界。

在微软亚洲研究院理论中心，我们的研究人员认为目前的大模型及其前向预测与信息压缩相关。相当于大模型把反映人类语言的所有“网页”压缩在了一个模型中，其生成过程就类似于解压的过程。所以，我们正在利用基于压缩的计算复杂性理论来理解和分析大模型的训练和生成过程，希望通过这个研究能更准确地认识大语言模型的生成能力。

大模型还有一个重要特点就是它的“涌现”行为。比如在算数四则运算上，两位数以内的计算，初始模型可以给出的结果准确率会高一些，但要计算三、四位或更多位数的，就需要将模型参数变多、训练变长才可以，这就是模型的涌现特性。

我们正在研究探寻这种涌现是否有理论能够解释。一个可能的理论是网络模型，其实涌现行为在网络科学、物理学中经常出现，我在网络科学方面的研究让我对网络中的涌现行为有比较深入的理解。我们正在考虑把大语言生成模型和网络科学建立联系来研究其内在的涌现特性。简单来说，大模型的生成可以看作是一张网，

输入一个词，生成下一个词，两个词之间就连接成一条边，然后再生成下一个词再连接一条边，每条边都是概率生成，并不绝对，所以不太稳定。而网络一旦有概率，在网络科学中就有可能出现涌现特性。

Q：在当前以大模型为主的研究背景下，理论研究将会面临哪些新问题？就个人而言，你会优先关注哪些方面？

陈卫：刚才说到的涌现行为就很重要，还有大模型的能力边界、性能效率、参数规模。例如，GPT-3 拥有 1750 亿个参数，模型是否真的是越大，性能效率就越高？这两者之间应该取得一个平衡。不可否认的是，模型越大能力越强，我们也确实可以通过增加更多参数、数据让模型更强，也可以涌现出更多新能力，但是人脑的工作原理却不是这样的。

《思考，快与慢》（Thinking，Fast and Slow）一书中将人类思考模式分为快思考和慢思考两个系统，即系统 1 和系统 2。系统 1 是常用的、依赖直觉的、无意识思考系统，系统 2 则是需要主动控制的、有意识进行的思考系统。现在的大模型更像是系统 1，凭直觉生成下一个字符，这也是它强大的地方，能够出口成章，但却也是它的弱势所在，它只能生成一次，没有回溯能力，缺乏更系统的有控制的推理和分析能力，这就是模型的限制。

因此，从理论上讲，很重要的一个问题就是：是否需要单独引入新的系统 2 的结构来与现有类似系统 1 的大模型结构合作以提高人工智能的能力，还是说只需要进一步提高模型的规模和训练数据就能提升大模型的性能？我认为，人工智能模型只基于语言模型和单向预测是不够的，新一代模型需要系统 2 的分析推理

Q：成立至今，微软亚洲研究院理论中心主要开展了哪些方向的研究？

陈卫：微软亚洲研究院理论中心会根据最新的人工智能发展趋势来动态调整研究策略。我们并不会限制研究员的研究方向，只要对理论研究感兴趣，研究员们可以从各种方向进行探索。

目前，理论中心主要的研究包括：数据驱动优化理论——如今的大模型都是由数据驱动的，然而数据是时刻变化的，所以需要将传统优化理论与数据结合，从数据角度做优化；深度学习理论——提升人工智能的可解释性、鲁棒性；还有可信计算，以及隐私保护等。其实很早之前在计算机领域并没有针对隐私保护的理论指导，直到 2007 年，微软研究院提出了差分隐私理论概念，随后该概念才被推广到了数据库、云计算等隐私保护场景。这是理论研究对计算机科学研究具有指导意义的一个很好的例证。

理论中心的研究主要集中在新的技术方向上，当然这些研究方向也会根植于传统的理论基础。大模型出现后，我们需要更新的理论，这些都还在摸索的阶段。科学研究初期通常都是应用研究发展较快，理论支持相对滞后，而当技术发展到一定时期就会出现很多问题，比如深度学习的可解释性、运行机制就需要理论指导，就像经典的算法理论一直在指导计算机科学的发展一样。如果我们完全不清楚 AI 大模型的运行机制和它超强能力的边界，就将其应用到生产生活的各个领域，必然会种下隐患。因此，我始终认为理论是计算机科学及相关科学非常重要的基础， 在当今 AI 大模型似乎要一统天下时，更需要理论的研究和支持。

Q：微软研究院有没有针对大模型的新的理论研究方向和成果？

陈卫：大模型确实将人工智能推向了新阶段，改变了原来的研究方法，也让大家站在了统一的起跑线上。我们微软研究院总部的同事 Sebastien Bubeck 和他的团队近期提出了 Physics of AGI 的概念，即通用人工智能物理学。因为现在的大语言模型更像一个黑盒子，对它的研究更像是对一个物理系统、物理现象的研究，就像物理学里的实验物理和理论物理研究一样，通过实验来总结规律。这就像历史上研究天体运行的规律一样，先是开普勒用观测数据找出天体运行的若干经验定律，后来才是牛顿在理论上的突破，提出万有引力定律，再加上他发展的微积分工具，完美地解释了开普勒的经验定律。

Bubeck 团队对大模型进行了实证研究，通过抽象出代数系统来验证大模型核心架构 Transformer 的能力，并给出了一定的理论指导（相关论文：Unveiling Transformers with LEGO: a synthetic reasoning task, https://arxiv.org/abs/2206.04301）。我们计划与他们合作，通过抽象出网络图模型来评估 Transformer 的边界，并结合实证研究，希望能够构建出基于网络的大模型理论模型。

Q：想要从事理论研究，需要具备哪些特质？要如何培养理论研究人才？

陈卫：概括地讲，理论研究人才除了要具备较强的基于数学的分析和推理能力，也要有较高的综合能力，以及交叉学科的背景。从事理论研究需要有开阔的思路、博采众长，不能只局限于数学、分析，或计算机科学中的某一个方面。同时，还要有主动性，主动思考发现新问题，尤其是当下人工智能、大模型的研究是没有固定模式的，不能遵从已有的范式，更需要创新精神。

目前，许多学生更多具有的是竞赛式思维，只要有明确的问题，他们总会找到解决方法。但在研究领域，没有人会告诉你问题是什么，比如涌现行为并不具体，而是需要科研人员自己去明确它是否可以转化成数学问题。然而，如何培养这类人才也是值得思考的问题，尤其是人工智能发展到如今这个程度，如何一方面利用人工智能作为辅助，另一方面充分培养发挥人的创造性， 使人工智能和人相互促进，培养出新一代的学生和科研人员，也是一个重要的研究课题。

扫描二维码查看视频

科学匠人 | 麻省大学副教授熊杰加盟微软亚洲研究院 —— “你相信无线感知吗？”

自 2022 年初，邱锂力博士从美国得克萨斯大学奥汀分校回国加入微软亚洲研究院，已有一年多的时间，在她的带领下微软亚洲研究院（上海）稳步发展，尤其围绕系统、人工智能和无线感知等领域展开了深入研究。与此同时，微软亚洲研究院持续在全球范围内引入行业杰出人才，充实科研力量，不断推动前沿技术的探索，微软亚洲研究院首席研究员熊杰博士就是其中的代表之一。作为新兴领域，无线感知在进行着哪些有趣、新奇且超乎想象的研究？无线感知可以给 AI 带来什么？让我们一起通过文章了解熊杰“信仰”的无线感知领域。

2023 年 2 月，熊杰从美国落地上海，开启了他在微软亚洲研究院的科研之旅。在成为微软亚洲研究院首席研究员之前，熊杰曾任美国麻省大学计算机系副教授一职，并分别在新加坡南洋理工大学、美国杜克大学、英国伦敦大学学院取得学士（一等荣誉学位）、硕士以及博士学位。熊杰多年来致力于无线感知、智慧医疗和移动计算方面的研究，相关学术论文曾获得 ACM MobiCom、ACM SenSys、ACM UbiComp 等多个全球顶会的最佳论文。

微软亚洲研究院首席研究员熊杰博士

是什么缘由促使熊杰走出校园，加入微软亚洲研究院？又是什么吸引他持续深入探索无线感知领域？他所带领的团队又将在无线感知研究中如何施展拳脚？新兴的无线感知领域，还需要哪些新鲜的血液？

选择微软亚洲研究院：遵循内心渴望，跟随前辈步伐

10 年前，熊杰正在伦敦大学学院攻读博士学位，研究领域是无线通信。那时 WiFi、3G 通信已落地应用，4G 产业正慢慢兴起，在可预见的未来几年，无线通信的研究将趋于饱和与成熟。彼时，熊杰和自己的导师也意识到了这一发展趋势，于是他们开始寻找新的研究方向。

无线感知这一交叉研究既是一个全新的领域，又和无线通信的基础知识结构紧密相关，自此，熊杰便踏上了无线感知的研究之路，并很快取得了突破性进展。

2015 年博士毕业后，熊杰选择了留在学术界担任助理教授，但加入微软亚洲研究院，与世界级科研人员共同探索前沿科学技术的想法一直埋在心底。多年间，与他共事的许多老师和同学都有在微软研究院工作的经历，他们都对微软亚洲研院自由、包容和多元的科研环境留下了深刻的印象。2022 年，熊杰看到无线领域的资深前辈邱锂力博士加入了微软亚洲研究院，遵循内心的渴望，跟随前辈的步伐，又恰逢微软亚洲研究院大力引进无线感知领域的研究人才，一年后，熊杰选择回国加入微软亚洲研究院，

那些你意想不到的无线感知应用

过去十几年中，无线技术在通信领域取得了巨大的发展，例如手机与 WiFi、4G/5G 的连接，蓝牙与手持设备的连接；服装店使用 RFID （射频识别技术）来识别顾客所拿的货品及价格；还有可用于大范围 IoT 设备连接的 LoRa（Long Range Radio，远距离无线电）技术，支持长达几公里的通信距离；以及已经应用于手机中，能够实现更精确定位的 UWB (Ultra-Wide Band，超宽带) 技术等等。面对这些无线技术，研究员们在想，除了用于通信外它们是否还可以实现其它有意义和有趣的事情？正是这些前沿的想法促成了无线感知领域的兴起。

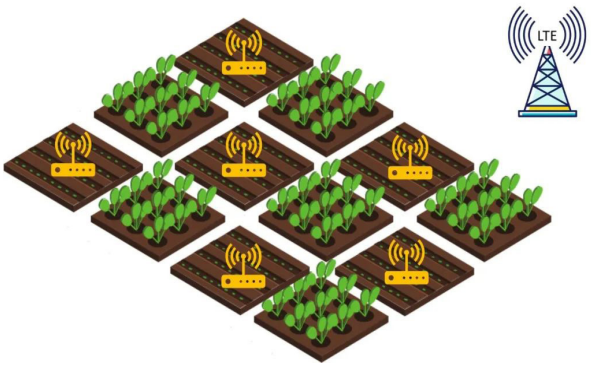
所谓无线感知，就是使用生活中常见的无线信号来进行感知。它具有两个特性：无需接触（Contact-free）、无需传感器设备（Sensor-free）。这就意味着实现无线信号感知后，未来的一些感知应用不再需要传感器。以人的睡眠姿态和质量监测为例，现在人们需要穿戴一些特定的设备才能实现全程监测，然而长期穿

戴这些设备会增加不适感，也可能在改变睡姿时影响监测质量。如果能够实现无线信号感知的监测，就可以摆脱穿戴设备的束缚，这些问题也就迎刃而解。

为了实现无线感知的目标，熊杰一直在探索 4G、LTE、声音、LoRa、UWB 等信号的感知技术。加入微软亚洲研究院后，熊杰正带领团队从三个方向进一步展开相关研究。

开发无线感知新理论。无线感知目前还是一个新兴领域，基础理论尚不完善。熊杰团队期望可以找到或发展出全新的理论来指导无线感知的研究，帮助更多人理解无线感知，比如衡量无线感知的性能、了解影响感知效果的主要因素等。虽然无线感知与无线通信紧密相关，但通信能力可以由信号强度和噪声来衡量，然而在感知中不光信号强度，信号的动态变化量也同样重要，这些都需要基础理论的指导。

探索无线感知新应用。通过创新利用多种信号，无线感知技术可以开发许多与日常生活息息相关、高可用性的落地应用。比如使用 WiFi 信号进行液体的准确识别，现在相关技术已经可以做到区分不同品牌的可乐，鉴别牛奶是否变质，识别水中糖的含量；还可以通过声音信号，实现如呼吸、心跳等生命体征的监测；利用无线信号监测土壤湿度，从而科学指导农业灌溉生产等等。



利用无线信号进行土壤湿度监测，科学指导农业灌溉生产

探索实际问题，创新无线感知平台和方法。当前无线感知研究主要在实验室中进行，目标大都聚焦于提升无线感知的性能指标，比如感知的距离、精细度。然而将无线感知技术应用到实际场景中却面临着新的挑战，如无线感知的稳定性、耗电量，以及对通信的干扰等。“我希望通过对实际问题的研究，来弥合消除实验室原型设计到落地应用之间的距离。”熊杰说。

熊杰认为，在无线感知探索中，不同的信号具有各自独特的优势，手机、汽车等人们高频使用的产品中都有麦克风、扬声器等器件，为利用声音信号感知提供了极大的便利性。通过车辆中的扬声器可以进行车内物体的感知，比如帮助驾驶员在离开车的瞬间识别车内是否还有生命体，如婴幼儿和宠物等，从而避免疏漏。同样也可以通过导航的手机发出超声波探测人的眨眼频次来实现疲劳驾驶监测！

相比声音信号，LoRa 信号的覆盖范围更广且可穿透建筑墙体。如果能够通过 LoRa 信号实现无线感知，那么对于楼房倒塌、地震救灾等场景将有很大帮助。而已经在手机中实现精准定位的 UWB 信号，可以在浴室、卫生间等隐私保护场所进行监测、感知，当老人摔倒时及时向有关人员发送通知。这些技术方向都是熊杰团队下一步攻克的重点。

想从事无线感知研究？需要兴趣导向，脑洞大开

无线感知的发展前景非常广阔，包括移动医疗、无人驾驶、机器人、智慧农业在内的应用场景都将以无线感知技术作为领域未来发展的基础。而在人工智能技术驱动的潮流下，无线感知也将成为人工智能发展必不可少的一部分。熊杰认为“在以大模型为主流人工智能应用的当下，无线感知数据将为 AI 训练提供有别于文本、图像、音视频之外的第四种数据类型，促进新型 AI 模型的设计，也将推动 AI 技术在更多场景得到应用。”

随着邱锂力和熊杰的加入，微软亚洲研究院收到了更多来自世界各地、希望从事无线感知研究的人才的申请。新技术领域的研究必将面临新的挑战，产生新的人才需求。那么无线感知领域的科研人员需要具备哪些特质？“相比于其他计算机领域，无线方向的研究对于专业知识的要求没有那么高，在掌握基本的信号处理、编程知识之外，更需要的是研究者的脑洞大开，敢于提出创新大胆的想法和创意，才能做出创新的研究。另一方面，从事无线感知领域的研究需要拥有极大的兴趣，并有坚持长期研究的耐心和决心。”熊杰说到。

就像熊杰读博期间面试微软研究院实习时，面试官问过他的一个问题——“你相信无线感知吗？”

相关阅读

扫描二维码查看文章

科学匠人 | 对话邱锂力：一起探索未知的科技之美

2018 年世界人工智能大会上，微软宣布成立微软亚洲研究院（上海）。成立至今，微软亚洲研究院（上海）都做了哪些研究，取得了怎样的进展？未来会重点投入哪些研究方向？有哪些人才引进的新计划？与我们一起对话微软亚洲研究院（上海）负责人邱锂力博士，了解微软亚洲研究院（上海）的成长步伐和未来规划。

见证 Ada Workshop 2023：一同创造“我们的时刻”

当计算机领域的女性榜样和女学生们欢聚一堂，她们会讨论什么，点亮什么，创造什么？

4 月 8 日，微软 Ada Workshop 2023 年度活动在春日与大家如期重逢。这是阔别三年后的首场线下活动，200 余位同学从五湖四海共聚现场，会场内座无虚席，笑语不断，提问踊跃，互动热烈。这也是 Ada Workshop 首次开放给高中生，我们期待在更早的时间点，帮助女学生找到自己的心之所向。同时，四大平台 9 小时无间断直播，近 4 万名观众云端观看，共话女性在 STEM 领域的发展之路。

本次活动以“The Moment · 我们的时刻”为主题，以“Empower Women through Connection”为核心理念，邀请到了来自微软和学术界、产业界的优秀女性榜样与男性同盟，为参与活动的女学生分享自己的独家故事、心路历程和技术远见，带领同学们一同体悟个人成长的上下求索、领略技术发展的前瞻视野，展现了女性在驱动科技创新与科技向善中的独特价值和别样视角。该活动由微软亚洲研究院学术合作经理王婧雯组织与主持。

正如梅琳达·盖茨在《女性的时刻》一书中写道：“女人如果在另一个女人的故事中听到自己的声音，她就会变得勇敢，单薄的声音就会变成合唱。”在这里，故事与故事碰撞，温暖与勇气交织。我们最终彼此联结，彼此唱和。

妳是否从女性榜样的故事中听到了自己的声音？抑或是通过她们的分享获得了前行的力量与勇气？关于技术的未来发展，妳眼前的迷雾是否被拨开了一些？只要有一个答案是肯定的，那我们的努力便是值得的。

Ada Workshop 2023 嘉宾与参会学生大合影

创造“我们的时刻”
感受女性科技共同体的同频共振

面对性别偏见的无形壁垒，身为“少数派”的科技女性如何突出重围，成为熠熠生辉的存在？她们一路走来，翻越了怎样的高山，又领略了怎样的风景？

实际上，女性在科技发展的长河之中一直发挥着关键作用，却总在叙事与记忆中被动“隐身”。活动伊始，微软亚洲研究院资深学术合作经理、Women@Microsoft GCR Co-chair、CCF 女工委执行委员孙丽君女士在开场致辞环节中带领大家再度解码 Ada Lovelace、Grace Hopper、夏培肃等女性科技榜样的辉煌故事，“如星辰一般灿烂的女性，推动了计算机近两百年历史的发展。”她重新回溯了 Ada Workshop “予力计算机及相关领域女学生成长”的初心，希望此次活动能够成为一颗火种，点亮在座的每一个人，而每个人也能够成为火种，将多元与包容的价值观一直传递下去。孙丽君还鼓励参会者多听、多问、多连接，共同创造“我们的时刻”。

微软亚洲研究院资深学术合作经理孙丽君

而后，正在科技领域发光发热的女性榜样们依次登场，分享她们一路走来的科研探索、职业发展故事与个人经验。高校教授、企业界的科技女性、研究员和创业者，她们的身份不一，但均以其独特视角为学生们带来了真切的共鸣与鼓舞。

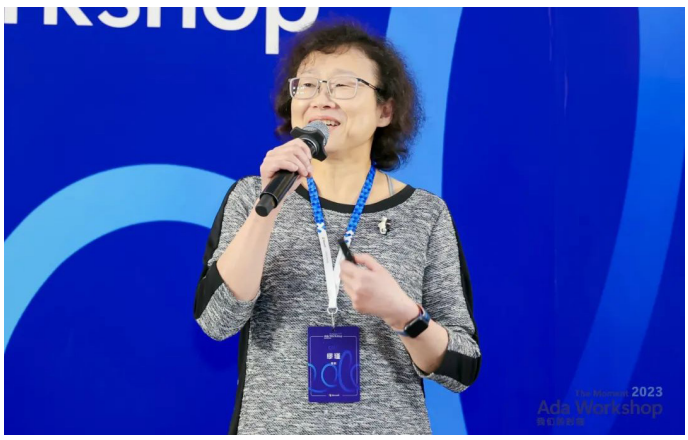
在上午场的主题报告环节中，中国科学技术大学教授、计算机科学技术学院副院长张燕咏老师，分享了她在科研路上作为“背包客”的探索历程。海外任教十余载回归祖国，而今作为国家科技创新重大项目首席科学家的她，结合自己的发展经验，为在座的女学生提供了独到智慧的“参考答案”。比如，在读博期间，选导师是至关重要的一步，成功的关键便是“相信导师”；至于科研题目的选择，则需要结合个人兴趣和当下研究热点；要找到自己最

高效的工作方式，并据此确定和导师的日常沟通方式，身为 “夜猫子” 的她总是在凌晨的小镇上独自漫步，脚踩落叶，沙沙作响；同时，珍惜每一个站在台上展示自己研究的机会，而展示的核心是交流而不是背诵，放下语言的顾虑，充满热情地分享你的研究，才更有机会被看见。

除此之外，她认为，工作与生活的平衡不是关键，自己内心的平衡才是一切的源泉。无论是高铁上、飞机上还是上课路上，要争取一切时间和机会，与自己相处，聆听自己内心的声音。

中国科学技术大学教授张燕咏

接下来，微软资深工程总监缪瑾女士分享了她为何认为“好奇心不应该杀死猫”。横跨硅谷、苏州和上海，从开发到管理，涉及多个领域和多类企业，同时作为微软 D&I 活动的积极参与者，她有着多元的故事与经历。在缪瑾看来，是一以贯之的好奇心驱使着她一步步地走到现在的位置。她主张要带着兴趣和目标去学习，兼顾广度和深度，关注行业新兴发展趋势，主动学习，拥抱变化，构建成长性思维，追寻更大的发展空间。同时，她逐步拆解破除了对女程序员的刻板印象，强调女性在计算机领域的别样视角和优势，指出应当引入性别视角，创造更公平的数字技术环境。在面对数字经济发展带来的机遇和挑战时，女性要相信自己，发挥所长，向前一步——“现在是展示我们力量的时刻，并且，刻不容缓”。



微软资深工程总监缪瑾

之后，从数学转向计算机、自传道授业解惑中助力学子成长的清华大学教授兰艳艳老师，分享了她“行我所行，无问西东”的成长故事。

“很久很久没有见到这么多女生在一个会议室里面，大家就像闺蜜一样聊天。”她的语调轻快，鼓励每一个女生都去做自己，定义自己想要的生活。一方面，要倾听内心的声音，另一方面，要去看世界，思考世界。她认为，变化固然重要，我们要通过终生学习不断“扩大舒适区”；但更重要的是坚守自己的内心，想尽办法做成自己想做的事，这能带给人莫大的幸福感——“我希望我老去后的每一根白发都代表着我年轻时做过的某一篇论文、某一项成果”。此外，她还建议女生不要一个人闷头做科研，而要积极融入学术共同体，发出属于自己的声音，与科研小伙伴一起成长。

清华大学教授兰艳艳

在下午场的主题报告环节中，Microsoft Senior Operations Group Manager, Women@Microsoft Asia Lead Sindhura Sunkara 女士首先分享了她的“世界冒险之旅”，她一向以世界公民为己任，投身于女性赋权的社会公益事业中。

数据显示，女性占据了 STEM 学生席位中的的 50%，但仅仅占据了顶级科技工作者中的 20%。女性在新兴技术角色中的代表性不足，特别是在云计算和人工智能领域。而她结合了各个领域的优秀女性榜样案例，指出了女性潜在的巨大可能性——她可能是领导，是冠军，是成功者，是创造者，是问题的解决者，是变革的推动者。她鼓励女性要认识自己，打开自己，做真实的自己；保持学习，自我刷新，适应变化，敢于冒险；同时要主动寻求帮助，与周围的人建立连接，构建属于自己的同盟，去勇敢地追逐梦想、实现梦想。

之后，香港浸会大学副教授陈黎老师分享了她的“跨学科研究之路”。作为世界前 2% 的科学家，陈老师擅长从交叉视角重新理解人与机器的关系，而她的科研之路则与二十年前微软大厦的一场讲座密不可分。在这场讲座上，她鼓起勇气与讲者建立联系。这位讲者最终成为了她的博士生导师，也带领她开启了人机交互这一研究方向的探索。

Microsoft Senior Operations Group Manager
Women@Microsoft Asia Lead Sindhura Sunkara

通过展示自己的科研成果和探索历程，她鼓励同学们要树立创新思维、批判思维与正向思维，并且指出女生在人机交互方向往往拥有非常敏锐的洞察力，因此她也会有意识地提升学生中的女生比例，并将她们视为科研上的合作者、生活中的好朋友，毫无保留地进行指导与交流。

香港浸会大学副教授陈黎

接下来，微软亚洲研究院高级研究员王希廷博士，分享了她如何"以复原力应对变化"。从"好学生""乖乖女"到"时代的冲浪者"，从恐惧改变到拥抱改变，她坦陈了自己在心态与思维的一系列成长，鼓励大家放弃对不完美的恐惧，遵循内心的感召，实现自我赋能，并通过人与人之间的深度连接，突破认知的局限，去勇敢地拥抱生命的真相。幡然醒悟之后，面对大模型技术飓风所带来的焦虑和迷茫，她反而跃跃欲试，试图打开"黑箱"，对其进行心理测量，探寻它的解释之道。同时，她也分享了自己同各领域专家合作，一同探究可解释、负责任的人工智能的科研之路。

最后一位主旨报告讲者，KodeRover 创始人兼 CEO 李倩女士，则在报告中分享了她如何完成了"从农民到云计算专家创业"的跨越。在电脑还是奢侈品的年代，家中添置了一台小霸王，她从此和计算机结缘。毕业后，她三年内换了五份工作，历经巨头、外企与创业公司，试图找寻自己的心之所向。"学生时代只是一个起点，你要关注的是自己的加速度"，她说，在面临着巨大的生存和发展压力时，她决定向死而生，毅然走上自己所热爱的创业之路，为工程师提供相关基础设施，并进行百分百开源，以便生产出更大的社会价值。在她的团队中，有一半都是女工程师，"不是男女差异，而是每个人都需要成为自己"。作为女性开源创业者，她更有着助力中国产业升级走向全球的愿景。

微软亚洲研究院高级研究员王希廷

KodeRover 创始人兼 CEO 李倩

在主旨报告之外，Ada Workshop 2023 还准备了干货满满的技术远见讲座系列，邀请各领域的女性研究员分享其关于技术未来的洞察，带领在场的女学生立于时代巨浪之前，并鼓励大家跳上甲板，勇当舵手：

微软亚洲研究院人工智能与机器学习组主管研究员骆煦芳博士通过基于脑电图的 ICU 新生儿癫痫检测和帕金森疾病进展预测两个案例，生动展现了机器学习如何帮助解决医疗健康场景中存在的问题；

微软亚洲研究院系统与网络组主管研究员程文雪博士分享了自己和女性研究员的合作成果，介绍了大模型计算机系统研究与 AI 技术发展相互促进的关系，后者为前者既创造了新需求，又提供了新思路；

在微软研究院科学智能中心主管研究员邓攀博士看来，复杂性是生物学研究的最大挑战，而利用人工智挖掘复杂生物规律，能够驱动分子科学研究范式的跨越式发展；

微软（亚洲）互联网工程院主管研究员黄丹青带来有关 AI + 平面设计的分享，这一交叉领域的研究既需要普适大模型的支持，也需要独特设计美感的支撑，其间最核心的还是以"人"为中心；

微软亚洲研究院多模态计算组主管研究员刘蓓联系到具身 AI 这一学科，分享了从互联网数据到具身环境的多模态学习思路；

微软亚洲研究院媒体计算组主管研究员王婧璐从视频会议这一案例引入，呈现了如何通过开发新一代多媒体基础模型，探索人工智能在实时、智能和沉浸式的人 - 机 - 云交互体验中的潜力；微软亚洲研究院主管研发工程师吴文珊分析了微调和上下文两种方法及其局限性，向大家揭示通用大模型在成为基础设施后响应真实世界场景的机遇与挑战。

技术远见讲座系列

我们始终相信，分享本身就会带来力量，所有的故事都将引起回响。女性的声音汇聚在一起，将拼凑出新的科技图景，带来更多的启发与可能性。

见证"她的时刻"
和更多同行者探索科技旷野

有哪一个时刻，让你回想起来心脏依旧砰砰直跳？在故事与故事的交融中，我们可以从女性社群中汲取到怎样的能量？

每个人的故事都值得述说，每一个留下的脚印都是宝贵的行动参考。在"妳的时刻"特别分享环节中，Ada Workshop 2023 特邀年轻女科技工作者、女学生代表，走到聚光灯下，分享属于自己的高光时刻。从大一入学到职场新人，或许，她们就是曾经的妳、现在的妳和未来的妳。

来自清华大学大一的蒋子悦同学，在高中阶段便迈出了科研尝试的第一步，对莫奈干草堆组画进行数学分析，通过其绘画色彩判断出当时的光照情况。而来自中国传媒大学大四的王泓霏同学则分享了自己的酸甜苦辣保研路，用自己的经历鼓舞更多女生能找到属于自己的光，并且成为光，照亮更多人。接下来，来自南京大学研三的童欣同学讲述了自己的曲折"转码"历程，以及其间女性榜样与女性社区带给她的信心与力量。之后，来自中国科学院大学博三的林妙倩同学分享了她去年参加 Ada Workshop 的体验，从中感受到了奇妙的共振，找到了前进的指引，挣脱性别偏见带来的自我束缚，更加从容地面临未来的挑战。最后，微软 Azure 云计算部门开发工程师贺心，用实力回应关于女程序员的质疑，并鼓励大家要敢于开始，持续学习，也分享了自己在微软 Work Hard Play Hard 的生活方式。

清华大学大一学生蒋子悦

在见证个体的闪耀之外，我们更期待着看到女性群体彼此联结所迸发出的无限能量。因此，Ada Workshop 2023 特设 The Moment 午餐会环节，为参会者提供与女性科技榜样和同行伙伴一起共进午餐的机会，一同交流自己的好奇与困惑，建立更深层次的连接。

在本届嘉宾之外，更有往届讲者微软（亚洲）互联网工程院资深软件工程师边冠琳、微软亚洲研究院高级研究员陈琪、微软亚洲研究院主管研究员徐勇重磅回归。美食的香气和滋味帮助大家更好地打开了话匣子，放下顾虑，"饭饭而谈"。无论是生活琐事，还是人生抉择，她们都在珍贵的对话与沟通间寻找到了奇妙的共鸣，也逐渐摸索到了解答人生疑问的金钥匙。

我们始终期待，以故事启发故事，用交流连接孤岛。当女性彼此对话，共同行动，将会产生某种奇妙的纽带，托举着她们一同向上。

探索“我的时刻”
拥抱属于女性的无限可能

身处技术变革的时代，如何应对随之掀起的巨浪？在充满不确定的未来里，如何成为不可取代的人？或许，我们都曾面临相同的困惑。

而 Ada Workshop 2023 圆桌讨论环节便提供了这样一个直面问题真挚交流的契机。从上午场“我们的时刻”到下午场“我的时刻”，我们从大到小串联起大家所提出的困惑，试图带大家找到走出迷茫的方向。

在前述嘉宾之外，微软亚洲研究院系统研究组高级研究员薛继龙、微软亚洲研究院自然语言计算组高级研究员刘树杰、微软亚洲研究院异构计算组主管研究员张丽也加入了这一环节，提供了更加丰富的视角和见解。

在与观众的实时词云互动图中，我们可以看到，迷茫和焦虑成为了最大的关键词，同时也伴随着期待和激动。

而嘉宾们也坦陈了自己作为“局中人”，拥有和同学们相似的感受。张燕咏老师指出，每个人都需要在新的时代重新思考“我”与 ChatGPT 的关系是什么，抓紧机遇，跳上这艘巨轮，结合自己的研究方向做出自己在大模型时代的独特贡献。和产品打交道的缪瑾老师则认为，在面临重新定义生产力的技术时，比起焦虑是否被 AI 取代，更重要的是利用现有技术提升自己的工作效率，不被其他人所取代。兰艳艳老师作为大模型的研究专家，认为超人的智能是时代发展的趋势，鼓励大家在面临“智能涌现”时去积极拥抱变化，多看看星辰大海的世界，解决以前解决不了的问题。而专攻系统方向的薛继龙老师则指出大模型也是新时代的系统，换个视角，放下焦虑。

圆桌讨论上午场

而具体到更加微观的方向探索，王希廷老师认为，好的目标在于能激起人内心的能量，“你想到它，心在砰砰跳，有一点够不着，让自己很兴奋，想一想又有可能会实现，有一些具体的计划”；李倩老师则鼓励大家打破“小镇做题家”的信息壁垒，通过网络积极获取信息，要打开自己的感知去嗅探这个世界，人生其实是一场旷野；在陈黎老师看来，选择的前提是热爱，最佳路径是先宽后窄，在了解全局的情况下再去深入钻研；而王树杰老师则分享了自己参与微软联培博士的经历，从个体经历出发为同学们提供借鉴。

再到更实际的学生招募问题，作为大学导师的陈黎老师认为五大人格测试是可供参考的标准之一，她会更看重候选人好奇与自律的特质；作为公司首席执行官的李倩则偏爱“从娃娃抓起”，看重候选人的基础和素养，以及是否真正热爱自己所做的事业；而在作为研究院导师的王希廷老师看来，自驱力是非常宝贵的品质；最后，作为研究院经理的王树杰老师分享了自己找“三好实习生”的标准（数学好、编程好、态度好），并且再次强调了沟通能力在团队协作中的重要作用。

圆桌讨论下午场

实际上，Ada Workshop 2023 并不提供标准答案，而是鼓励每个人去寻找答案、创造答案。在会议的固有环节之外，参会的同学们发挥自己的能动性，建立起更广泛的连接。

茶歇时，同学们将心爱的讲者团团围住，一个问题连着另一个问题，在向上交流中探寻科研的脉络；午餐时，同学们三两成团，主动交友，鼓起勇气和陌生人 say hi，从而开启一段从未设想过的奇妙缘分；微信群中，同学们呼朋引伴，张罗着相同背景与志向的伙伴一起交流，逐渐建立属于自己的信息网络……

我们始终希望，Ada Workshop 2023 只是一个起点，而接下来的科技女性故事，还将更加异彩纷呈。

扫码二维码查看视频回放

相关阅读

扫描二维码查看文章

选择的时刻，如何做出自己在技术变革时代的贡献？
听听前辈们怎么说

在技术革新的时代，如何理解科技发展的前沿趋势并抓住机遇？
面对纷繁复杂的可选项，如何擦亮双眼，做好自己的关键选择？
遭遇挫折，如何跳出既有的思维定势，重新定义问题？
如何兼顾深度和广度？学术和创业的关系是什么？

完成一幅计算机学术生态的拼图，少不了这些“斜杠女性”

完成一幅计算机学术生态的拼图，少不了这些"斜杠女性"
在微软亚洲研究院学术合作部，来自不同国家、有着不同学术背景和人生阅历的女性，正在以各自所擅长的方式，为跨领域研究搭建沟通协作的桥梁。

树洞回答 | 每一种情绪都值得被看见

那些在聊天框里输入又删掉的话，那些在微博小号里才能记载的心情，那些在朋友圈仅自己可见的动态，那些在计算机领域科研、学习、生活中遇到的难题，树洞接收到了你们的信号。

在将树洞收到的内容筛选、归类后，我们为提问的你匹配到了微软亚洲研究院中合适的解答者。

其中，有来自职场的困惑：如何在工作和生活的双重内卷下自我排解？也有来自未踏入社会的同学们的疑虑：如何快速增强科研实力？研究过程中，如何更好应对难以深入问题核心、找不到方向时的迷茫？还有来自文学、遥感等跨领域的学术思考。

听一听树洞连接到的第一波回复，希望这些真诚的回应能够成为大家科研生活中情绪负担的舒缓解药。

Q：在做研究时，对问题的思考总是浮在表面上，每次需要更深入的思考时，总无法触及问题的核心。如何才能让自己的思维不仅停留在问题表面？

科研小白找 idea 真的好难，对领域并不那么熟悉，也不太知道想到了之后能不能 work，要反复想反复验证，一个 idea 一个 idea 地换，好累呀....

在科研生活中，面对未知，找不到前进的方向时，如何制止自己的慌乱、思想抛锚与死磕（钻牛角尖），以让自己冷静理性？

微软亚洲研究院高级研究员杨凡：

这些都是非常好的问题，非常值得探讨。我深切理解同学们在做研究时所面临的迷茫和焦虑。作为一位"过来人"，我对这些问题感同身受。即使到现在，我也还在不断思考和探索如何将研究做得更深，如何把握未来趋势。这些问题没有绝对正确且统一的答案。不过，身为一名长期在计算机系统领域工作的研究员，我确实积累了一些个人心得。借此机会，我想将这些经验分享出来，希望对大家有所启发。

正视困难，接受挑战

严肃的研究是非常具有挑战性的。我还记得在与沈向洋博士的一次闲谈中，他曾谈到过做研究的困难，大致的观点是：取得扎实的研究成果非常困难，而且这并不会因为你以前做出过好的结果，下一次就会变得容易。他的这些话减轻了我的焦虑，原来一流的研究员也会觉得研究很困难啊！之后我在自己的研究项目中也体会到，在面对一流的研究问题时，人人平等，问题本身并不会因为研究者的资历而变得更容易。有鉴于此，面对困难，我的建议是首先要坦然面对。在科研过程中"面对未知，找不到前进方向"是常态，而找到突破点实际上才是罕见的。接受这一现实，在研究过程中我们的心态就会更加平和，避免浮躁。

积极思考，厚积薄发

在进行研究时，我们经常会遇到对新领域不熟悉，难以找到合适题目，或者老师给了题目也无法深入开展等问题。到现在我和我的同事还有实习生们也时常会陷入这样的困境。但在实际工作中，我发现一些思维技巧可以帮助我们脱困而出。

首先建议采用追根溯源的思考方法——多问为什么。当面临问题时，要问自己为什么会遇上这个问题？当想到某个方法时，可以问自己为什么会考虑这个方法？对某个方案，要问自己为什么它与众不同？为什么它可能行得通，又为什么可能行不通？持续追问，直至问题的答案可以归约为领域中某个公认的设计原则（比如系统设计中的 separation of concern，modularization，layering，minimization 等原则），或者问题 / 答案可以被形式化，并用准确的数学语言加以描述。通常到这个阶段，我们对这个问题的理解就比较深刻了。

另一个思维技巧是将问题抽象化（abstraction）。思考这个想法是否可以用更抽象的方式描述？它是不是代表着一类更普遍、更通用的问题？例如，在研究深度学习集群调度和深度学习编译器的时候，我和同事们常问自己："是否可以采用更通用的技术，而不仅仅局限于深度学习这一特定场景来解决问题？"最终，这种

思考方式指引我们通向了更底层的专门的技术方案。还有一个方法是从多角度思考同一问题。比如想证明一个问题时，多问自己是否可以证伪这个问题。在考虑编译技术时，是否可以从作业调度的角度来解决。

在思考过程中遭遇瓶颈时，与其他同事、同学交流往往也会带来帮助。我发现在一个开放、放松的环境中和同事交流，有时候可能只谈了一下手头的工作，就能激发新的灵感。此外，暂时放下问题，转换思维，也有助于打破思考上的僵局。在工作中，我和我的同事通常都是多线程工作，一个问题解决不了那就去思考其他问题。我个人在某些问题上花了几年都没有进展，但可能在一个契机闪现时突然就有了眉目。

综上所述，保持积极的思考态度并灵活运用各种思维技巧对学术研究大有裨益。通过前面提到的这几个方法，我们可以让自己更深入地思考，真正沉淀下来，最终取得突破，厚积薄发。我相信，幸运之神会眷顾那些准备好的人。

保持激情，勇担失败

研究中遇到挫折和失败，容易让我们患得患失。然而，持续的负面情绪不仅对身心健康有害，还会削弱我们对研究的热情和创造力。因此，如何应对负面情绪至关重要。在面对挑战时，我个人的体会是要保持好奇心，专注于深入理解事物的本质，而非只关注简单的成败。在研究过程中，弄清楚某个现象背后的原因比仅展示结果更为重要。因此，当遭遇失败时，我们应努力理解失败的原因。随着对问题的理解逐渐加深，别人无法感知的见解才能被探寻。这种深刻的洞察力往往是通向新大陆，实现重大突破的起点。

此外，保持研究的激情是长期坚持、实现厚积薄发的重要原生动力。但如何找到并保持研究热情是一个非常个人化的问题，每个研究员可能都有自己独特的答案。我们可以尝试从自己热爱的领域、感兴趣的问题出发，寻找激发研究热情的源泉，从而在研究道路上走得更远。

预测未来，引领潮流

仅凭激情是难以长久的，与科研成果伴生的成就感才是我们不断前行的根本原因。对于新入行的同学，在顶级学术会议上发表论文是一种成就感。但随着时间的推移，相比于论文发表，科研人员们更希望自己的成果能够产生广泛的影响，甚至改变某个领域的思维范式。做出扎实、有影响力的研究工作是每个科研人员孜孜不倦追求的目标。

在这方面，我非常认同微软亚洲研究院"老一辈"研究员，微软亚洲研究院常务副院长、微软杰出首席科学家郭百宁博士的观点。他一直鼓励我们要勇敢地预测未来，努力成为引领潮流的研究员。他会引导我们思考：未来 5 年甚至 10 年内，你所从事的领域将会发展到什么程度？为了实现这样的发展，最重要的技术障碍是什么？你当前的研究是否在扫除这些障碍？妥善地回答这些问题有助于我们开展有影响力的工作。他还特别强调不要害怕预测错误，因为正如前文所述，"勇担失败"，对失败原因的深刻刨析通常会带来突破的契机。相反，若只是跟随别人的步伐，则很难做出开创性的工作。

在勇敢地预测未来这方面，研究院的高级研究员胡瀚是一个出色的榜样。在闲谈中，胡瀚说他相信"注意力"机制（Attention）在计算机视觉（CV）领域将发挥基础性作用。基于这一预测，他和团队成员进行了长达数年的高强度研究。在当今研究节奏极快、项目通常需在数月甚至数周内见效的 CV 领域，这无疑是一个勇敢的举动。正是这种对未来的预测，让胡瀚和他团队的同事们做出了 Swin Transformer，该成果不仅在学术界获得认可，斩获 CV 领域的最高荣誉马尔奖，而且在工业界，CV 领域的各个子任务也常常将 Swin Transformer 作为标配之一。

以上仅仅是我个人的体会，难免存在疏漏和谬误。但我希望它们能够对同学们产生一些帮助。要强调的是，每个研究员都有自己独特的经历和心得。探索科研之路，关键在于不断学习、实践和反思，逐渐形成自己的研究方法和思维模式。在科研生涯中，培养自己的创新能力、批判性思维和解决问题的技巧是一个长期、渐进的过程。与此同时，大家也需要逐渐学会与团队合作、与同行交流。在科研过程中，需要不断调整自己的心态，乐观地承受失败，保持研究的热情和好奇心。

科研是一条充满挑战的道路，但正是这种挑战才使得科学研究如此吸引人。在不断努力、尝试和取得进展的过程中，我感受到科研真正的乐趣在于探索未知、发现新大陆。长期深耕一个有挑战性的问题，最终获得答案，这比简单的世俗认可更能带来满足感。我相信这也是广大科研工作者能够耐住寂寞，长期奋战在各自科研领域的根本原因。

Q：有关于新 paper 的很多 idea，但是经常因为动手能力差就没有实践，总是停留在理论阶段导致进度很慢，有什么好方法能克服这个问题呢？

微软亚洲研究院主管研究员李潇：

首先，建议多刷 arXiv/ Google Scholar 推送 / 近期的顶会论文集，我自己的感受是多读读论文就会发现"很多 idea"往往是"已经被很多人做过的 idea"，不过这也不是一件坏事。每一篇优秀的论文都有它对于一个问题或一个 task 的 insight，也都有它的局限性；对于同一个 idea，不同研究者的实现方法和探索到的内面往往也是有区别的，你反而可以从中更全面、深刻地理解一个 idea。以我所研究的计算机视觉 / 图形学方向来说，每年的 CVPR、SIGGRAPH 等顶会上，也不乏同样一个 idea 分别被不同的 paper 以不同的角度挖掘，把这些相似的 paper 放在一起对比

阅读，同样非常有助于对 idea 的深度发掘，找到本质的研究问题。

其次，如果调研完文献之后真的有很多看上去都可行的 idea，那么（尤其是在早期研究生阶段，经验还不十分丰富的时候）建议仔细思考后集中攻克其中一到两个，而不是全面出击，防止“这个也想做，那个也想做，但是最后哪个都没做成”。我自己的经验是在初期的研究生涯中尽早的得到一些反馈，对于后面做出优秀的工作是非常有帮助的。这个过程中，和导师 / 前辈 / 厉害的师兄师姐交流也十分重要 ，有助于确定哪些 idea 是真正值得做的，哪些则不是特别重要的或者甚至是歧途。

最后，（纯理论方向除外）大多数 idea 并不存在所谓的“停留在理论阶段”，目前计算机还是一门以实验为主的学科，理论都是通过实验不断完善的，需要不断地“learning by doing”。建议可以准备一个便利贴（App 里的也可以，比如 Windows 自带的 sticky notes），把实现 idea 这个大任务拆解成很多小任务，一项一项地完成。在实现过程中，如果真的受限于动手能力碰到了一些困难，建议通过阅读类似方向 / 类似 idea 的经典文章的开源代码学习实现。

Q：计算机行业在国内乃至全球都越来越卷，普通程序员因工作压力所带来的负面情绪也日益增多。作为计算机领域的研究生，或者类似于我的更多同类入门者，有什么是我们需要尽早知道和应对的吗？有什么经验可以从前辈那里借鉴？

微软亚洲研究院高级研究员曹婷：

首先应该了解“卷”和整个经济大环境密不可分，经济周期处于紧缩下行，经济总量增长缓慢，再赶上毕业生数量处于高峰期，在有限的资源竞争中，内卷就会产生。但如同月亮的阴晴圆缺，经济周期有其固有规律，有下行就有繁荣。所以请保有信心，期待经济的下一轮繁荣。但无论处在什么周期，我们都应该让自己拥有核心竞争力，这需要不断思考、学习什么方向才是未来社会真正需要的核心技术，并为之努力，成为专业人才，避免无效忙碌，被其他技术降维打击。拥有核心竞争力的人才无论何时，也不论是去高校、企业还是创业，是在国内还去国外发展，都会找到让自己发光发热的舞台。

Q：找不到实习，想科研实力快快增强却找不到方向。好想有个有经验的科研前辈带带我。

微软亚洲研究院实习生项目负责人张津：

科研是一条不断探索和学习的道路，开始的时候你可能会遇到一些困难和瓶颈，但只要坚持不懈地努力，相信你一定可以找到合适自己的方向，并成为一名成功的研究者。以下是一些可以帮助你继续增强科研实力的建议，以供参考：

首先，合理利用学校的资源是非常重要的。例如参加校内科研项目，或者与教授、导师多多交流，了解他们的科研方向和研究领域。他们的宝贵经验和见解肯定很有参考价值！其次，关注科研相关的公众号，多逛逛国内外知识分享类的社交平台也是很好的选择。这些平台可以让你实时了解前沿的学术工作，从中探索自己可能感兴趣的领域，学习其他研究者的经验和见解，拓宽自己的思路和视野，从而辅助你增强自己的科研实力和研究水平。

Q: 读研所在小组的 GPU 资源有限，idea 不好想，在大模型越来越流行的时期，很焦虑，觉得自己做的课题太小，价值不高。这种焦虑情绪要如何排解？

微软亚洲研究院高级研究员曹婷：

首先，有大量 GPU 资源固然好，但也不是所有工作都需要 GPU 资源，比如不在 GPU 训练的盘古模型，很多基础算法研究也不需要大量 GPU 资源。如果放宽眼界，遇到困难努力思考解决方法，我们不难发现方法总比困难多。另外，课题的大小也和研究能力相匹配，等科研水平逐渐提高，才可以更好、更有能力地承担更重要的项目。

Q：我是一名计算机视觉方向的研二学生。刚入学时只期望学习更多，然而研一除去上课外，大多时间在做项目，有时感觉也学到了不少，有时又感觉只是没技术含量的替老师打工赚钱。我们组要求研二能写好一篇文章，但最近在看论文和做实验的过程中，感觉自己的知识储备真的很不够，我也只有一张卡能跑。现在距离找工作也不远了，可是感觉自己缺乏找到算法岗的能力，师兄也大多是找的开发。有时候感觉很矛盾，既希望学术上真的能做出一点什么东西，又怀疑自己在这有限的时间内也只能再产出点没意义的学术垃圾，再一想这样不如多花点时间多学些开发知识，为工作准备。想法有点乱，不知道研究院的前辈们是否能理解我这种小菜鸟的矛盾心情和焦虑，请给一些建议。

微软亚洲研究院主管研究员李潇：

既然你的想法"有点乱"，那就让我们先来总结一下你面临的主要问题有：

（1）学术氛围不够
（2）计算资源不足
（3）不确定将来的方向（算法 / 开发）

几点建议：

（1）如果有条件，建议利用暑期或者项目相对不忙的时间找一份工业界（这里指的是相对成熟的独角兽 / 大公司）的实习。这样一方面可以体会到工业界实际的工作，另一方面，目前工业界具备更充分的计算资源，周围也有很多有工程 / 研究经验的同事，有助于你学习到更多的代码经验，实现自己的研究想法。如果导师不允许或者项目很忙，也可以试一下目前很多工业界公司提供的兼职 / 远程实习机会。

（2）关于算法 / 开发：目前计算机领域的趋势是算法和工程越来越"不分家"。一方面，算法工程师仍然属于"工程师"，往往也需要具备一定的（甚至很好的）工程开发能力。另一方面，大模型的火爆使得工业界研究的范式会产生很大的改变：以往需要很多独立算法模型设计去完成的任务，会在一定程度上被大模型取代。大模型的模型结构本身相对稳定，而如何把大模型训起来（足够的计算资源），训得快（优秀的系统架构及工程实现），训得好（高质量的数据闭环）将会是一段时间内的重点能力。

（3）鉴于问题中描述的学习及科研氛围，如果真心对科研有兴趣和热情，建议认真且慎重的考虑硕士毕业后攻读博士学位。

Q：非科班同学怎样才能最快进入微软亚洲研究院实习？

微软亚洲研究院实习生项目负责人张津：

微软亚洲研究院一向秉承着多元与包容的理念，期待着来自各种不同学科和文化背景的科研人员和同学们的加入！非科班出身的同学想要进入微软亚洲研究院实习，不妨参考以下建议：

首先，你可以通过关注微软亚洲研究院官方网站（https://www.msra.cn/zh-cn/research-groups）了解各个研究组的研究领域，并通过微软学术合作和微软亚洲研究院公众号获取最新的实习生招聘信息，明确自己感兴趣的研究方向或课题。

接下来，你可以自主学习或参加课程，理解掌握相关的计算机知识，从而更好地提升自己的学习能力和综合素质。

除此以外，最重要的是通过官方渠道大胆投出你的简历。你也可以在线上或线下"学术追星"，在微软亚洲研究院举办的学术活动中，如近期我们举办的 Ada Workshop，与希望合作的导师或实习生建立联系、请教经验。

滴！第一批焦虑与困惑已被树洞吸收。

你的释放以及前辈们的经验分享，都是为了让大家以更好的状态迎接未来的生活！

相关阅读

扫描二维码查看文章

关于内卷与反内卷、建立学术社交网络，听听过来人的建议

作为尚处求学阶段的学生，该如何有意识地建立自己的学术社交网络？面对日益"内卷"的社会环境，该如何平衡科研与生活并达到自洽？

读书日书单 | 等待科技类读物播下的科研种子萌发

在近来仿佛科幻小说照进现实般的人工智能技术的"涌现"浪潮中，我们特别邀请了微软亚洲研究院高级研究员杨凡在世界读书日来分享他的科幻与科技类书单。优秀的科幻作品能够 "预言" 未来，但更是因为这些作品开拓了人们想象力的边界，在思维层面激发了创新思考与不断追寻科技的进步，从而推动现实生活迈向想象中更美好的未来。在这个特别的日子里，让我们跟随杨凡书单打开的入口，在无垠的星空、浩瀚的宇宙面前，超越时空，在科技世界里自由"悦"读。

AI 科技大本营 | 微软研究员联合 Yoshua Bengio 推出 AIGC 数据生成学习范式 Regeneration Learning

在 AIGC 取得举世瞩目成就的背后，基于大模型、多模态的研究范式也在不断地推陈出新。微软研究院作为这一研究领域的佼佼者，与图灵奖得主、深度学习三巨头之一的 Yoshua Bengio 一起提出了 AIGC 新范式——Regeneration Learning。这一新范式究竟会带来哪些创新变革?

AIGC（AI-Generated Content）在近年来受到了广泛关注，基于深度学习的内容生成在图像、视频、语音、音乐、文本等生成领域取得了非常瞩目的成就。不同于传统的数据理解任务通常采用表征学习（Representation Learning）范式来学习数据的抽象表征，数据生成任务需要刻画数据的整体分布而不是抽象表征，需要一个新的学习范式来指导处理数据生成的建模问题。

为此，微软研究院的研究员和深度学习 / 表征学习先驱 Yoshua Bengio 一起，通过梳理典型的数据生成任务以及建模流程，抽象出面向数据生成任务的学习范式 Regeneration Learning。该学习范式适合多种数据生成任务（图像 / 视频 / 语音 / 音乐 / 文本生成等），能够为开发设计数据生成的模型方法提供新的洞见和指导。

Regeneration Learning: A Learning Paradigm for Data Generation

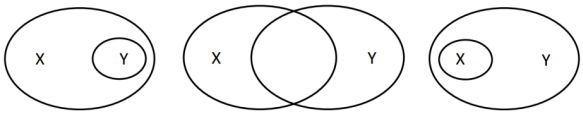
Xu Tan[1]*, Tao Qin[1], Jiang Bian[1], Tie-Yan Liu[1], Yoshua Bengio[2]
[1]Microsoft Research [2]Mila & University of Montreal
[1]{xuta,taoqin,jiabia,tyliu}@microsoft.com
[2]yoshua.bengio@mila.quebec

为什么是 Regeneration Learning ?

什么是数据理解与数据生成?

机器学习中一类典型的任务是学习一个从源数据 X 到目标数据 Y 的映射，比如在图像分类中 X 是图像而 Y 是类别标签，在文本到语音合成中 X 是文本而 Y 是语音。根据 X 和 Y 含有信息量的不同，可以将这种映射分成数据理解（Data Understanding）、数据生成（Data Generation）以及两者兼有的任务。图 1 显示了这三种任务以及 X 和 Y 含有的相对信息。

X 和 Y 的信息差异导致采用不同方法来解决不同的任务：



(a) Data understanding (b) Data understanding/generation (c) Data generation

Types	Information	Tasks
Understanding	$X \gg Y$	image classification, objective detection sentence classification, reading comprehension
Generation	$X \ll Y$	text generation or image synthesis from ID/class
Understanding/Generation	$X \not\gg Y$ and $X \not\ll Y$	text to speech, automatic speech recognition, text to image generation, talking-head synthesis

图 1：机器学习中常见的三种任务类型以及 X 和 Y 含有的相对信息量

对于数据理解任务，X 通常比较高维、复杂并且比 Y 含有更多的信息，所以任务的核心是从 X 学习抽象表征来预测 Y。因此，深度学习中非常火热的表征学习（Representation Learning，比如基于自监督学习的大规模预训练）适合处理这类任务。

对于数据生成任务，Y 通常比较高维、复杂并且比 X 含有更多的信息，所以任务的核心是刻画 Y 的分布以及从 X 生成 Y。对于数据理解和生成兼有的任务，它们需要分别处理两者的问题。

数据生成任务面临独特的挑战包括：

因为 Y 含有很多 X 不含有的信息，生成模型面临严重的一对多映射（One-to-Many Mapping）问题，增加了学习难度。比如在图像生成中，类别标签"狗"对应不同的狗的图片，如果没有合理地学习这种一对多的映射，会导致训练集上出现过拟合，在测试集上泛化性很差。

对于一些生成任务（比如文本到语音合成，语音到说话人脸生成等），X 和 Y 的信息量相当，会有两种问题，一种是 X 到 Y 的映射不是一一对应，会面临上面提到的一对多映射问题，另一种是 X 和 Y 含有虚假关联（Spurious Correlation，比如在语音到说话人脸生成中，输入语音的音色和目标说话人脸视频中的头部姿态没有太大关联关系），会导致模型学习到虚假映射出现过拟合。

为什么需要 Regeneration Learning

深度生成模型（比如对抗生成网络 GAN、变分自编码器 VAE、自回归模型 AR、标准化流模型 Flow、扩散模型 Diffusion 等）在数据生成任务上取得了非常大的进展，在理想情况下可以拟合任何数据分布以实现复杂的数据生成。但是，在实际情况中，由于数据映射太复杂，计算代价太大以及数据稀疏性问题等，它们不能很好地拟合复杂的数据分布以及一对多映射和虚假映射问题。类比于数据理解任务，尽管强大的模型，比如 Transformer 已经取得了不错的效果，但是表征学习（近年来的大规模自监督学习比如预训练）还是能大大提升性能。数据生成任务也迫切需要一个类似于表征学习的范式来指导建模。

因此，我们针对数据生成任务提出了 Regeneration Learning 学习范式。相比于直接从 X 生成 Y，Regeneration Learning 先从 X 生成一个目标数据的抽象表征 Y′，然后再从 Y′ 生成 Y。这样做有两点好处：

X → Y′ 相比于 X → Y 的一对多映射和虚假映射问题会减轻；

Y′→ Y 的映射可以通过自监督学习利用大规模的无标注数据进行预训练。

Regeneration Learning 的形式

Regeneration Learning 的基本形式 /Regeneration Learning 的步骤

Regeneration Learning 一般需要三步，包括：

将 Y 转化成抽象表征 Y′。转换方法大体上可分为显式和隐式两种，如表 1 中 Basic Formulation 所示：显式转换包括数学变换（比如傅里叶变换，小波变换），模态转换（比如语音文本处理中使用的字形到音形的变换），数据分析挖掘（比如从音乐数据抽取音乐特征或者从人脸图片中抽取 3D 表征），下采样（比如将 256*256 图片下采样到 64*64 图片）等；隐式转换，比如通过端到端学习抽取中间表征（一些常用的方法包括变分自编码器 VAE，量化自编码器 VQ-VAE 和 VQ-GAN，基于扩散模型的自编码器 Diffusion-AE）。

步骤 2：从 X 生成 Y′。可以使用任何生成模型或者转换方法，以方便做 X → Y′ 映射。

步骤 3：从 Y′ 生成 Y。通常采用自监督学习，如果从 Y 转化为 Y′ 采用的是隐式转换学习比如变分自编码器，那可以使用学习到的解码器来从 Y′ 生成 Y。

Formulation	Category	Method	Data Conversion ($Y \rightarrow Y'$)
Basic	Explicit	Fourier Transformation	Speech/Image (e.g, Wave→ Spectrogram)
		Grapheme-to-Phoneme	Text (e.g., learning→ˈlɜːrnɪŋ)
		Music Analysis	Music (MIDI→Chord/Rhythm)
		3D Image Analysis	Image (Face to 3D Co-efficient)
		Down Sampling	Speech/Image (e.g., 256*256→64*64)
	Implicit	Analysis-by-Synthesis VAE VQ-VAE/VQ-GAN DiffusionAE	Image/Speech/Text ($Y \rightarrow Z$)
Extended	Factorization	AR	Image/Speech/Text ($Y \rightarrow Y_{1:t}$)
	Diffusion	DDPM	Image/Speech/Text ($Y_0 \rightarrow Y_t$)
	Latent Diffusion	VAE + DDPM	Image/Speech/Text ($Y \rightarrow Z_0, Z_0 \rightarrow Z_t$)

表 1：Y → Y′ 转换的不同方法

如表 1 中 Extended Formulation 所示，一些方法可以看成是 Regeneration Learning 的扩展版本，比如自回归模型 AR，扩散模型 Diffusion，以及迭代式的非自回归模型等。在自回归模型中，Y_{<t} 可以看成是 Y_{<t+1} 的简化表征，在 Diffusion 模型中，Y_{t+1} 可以看成是 Y_{t} 的简化表征，和基础版的 Regeneration Learning 不同的是，它们都需要多步生成而不是两步生成。

Regeneration Learning 和 Representation Learning 的关系

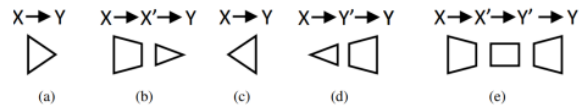


Figure (a): Presentation ($X \rightarrow$). Figure (b): Representation ($X \rightarrow X' \rightarrow$). Figure (c): Generation ($\rightarrow Y$). Figure (d): Regeneration ($\rightarrow Y' \rightarrow Y$). Figure (e): Representation + Regeneration ($X \rightarrow X' \rightarrow Y' \rightarrow Y$).

Paradigm	Original	Compact	Self-Supervised Learning	Easy Mapping
(b) Representation Learning	X	X'	$X \rightarrow X'$	$X' \rightarrow Y$
(d) Regeneration Learning	Y	Y'	$Y' \rightarrow Y$	$X \rightarrow Y'$
(e) Combination	X, Y	X', Y'	$X \rightarrow X', Y' \rightarrow Y$	$X' \rightarrow Y'$

图 2：Regeneration Learning 和 Representation Learning 的对比

如图 2 所示，Regeneration Learning 可以看成是传统的 Representation Learning 在数据生成任务中的对应：

Regeneration Learning 处理目标数据 Y 的抽象表征 Y′ 来帮助生成，而传统的 Representation Learning 处理源数据 X 的抽象表征 X′ 来帮助理解；

Regeneration Learning 中的 Y′ → Y 和 Representation Learning 中的 X → X′ 都可以通过自监督的方式学习（比如大规模预训练）；

Regeneration Learning 中的 X → Y′ 和 Representation Learning 中的 X′→ Y 都比原来的 X → Y 更加简单。

Y′→ Y 的映射可以通过自监督学习利用大规模的无标注数据进行预训练。

Regeneration Learning 的方法研究以及实际应用

Regeneration Learning 的研究机会

Regeneration Learning 作为一种面向数据生成的学习范式，有比较多的研究问题。如表 2 所示，包括如何从 Y 获取 Y′ 以及如何更好地学习 X → Y′ 以及 Y′→ Y 等，详细信息可参见论文。

Perspective	ID	Research Questions
$Y \to Y'$	1	How to design better analysis-by-synthesis methods (beyond VAE, VQ-VAE, DiffusionAE, etc) to learn Y'?
	2	How to design better learning paradigms other than analysis-by-synthesis to learn Y'?
	3	How to leverage unpaired data Y and/or paired data (X, Y) to learn Y'?
	4	How to better trade off the difficulty between $X \to Y'$ and $Y' \to Y$ mappings when learning Y'?
	5	How to disentangle semantic meaning and perceptual details to learn a more semantic instead of detailed Y'?
	6	How to determine the discrete or continuous format of Y' for each data generation task?
$X \to Y'$ $Y' \to Y$	7	How to design better generative models to learn $X \to Y'$ and $Y' \to Y$ mapping?
	8	How to leverage the assumption of semantic conversion and detail rendering to design better methods?
	9	How to leverage large-scale self-supervised learning for $Y' \to Y$ mapping?
$X \to Y' \to Y$	10	How to reduce the training-inference mismatch in regeneration learning?

表 2：Regeneration Learning 的研究问题

Regeneration Learning 在数据生成任务中的应用条件

Regeneration Learning 在语音、音频、音乐、图像、视频、文本等生成中有着广泛的应用，包括文本到语音合成，语音到文本识别，歌词 / 视频到旋律生成，语音到说话人脸生成，图像 / 视频 / 音频生成等，如表 3 所示。

Task	X	Y	Y'	$Y \to Y'$ & $Y' \to Y$
Speech Synthesis	Text	Waveform	Spectrogram / Code	STFT & Vocoder / Codec
Speech Recognition	Speech	Character	Phoneme	G2P & P2G
Text Generation	Text/Knowledge	Text	Template	Text2Template & Template2Text
Lyric/Video to Melody	Lyric/Video	Melody	Music Template	Music Analysis & Generation
Talking-Head Synthesis	Speech	Video	3D Face Parameters	3D Face Analysis & Rendering
Image/Video/Sound Generation	Class/Text	Image/Video/Sound	Latent Code	Codec Extraction & Generation

表 3：一些利用 Regeneration Learning 的数据生成任务

总的来讲，只要满足以下几点要求，都可以使用 Regeneration Learning：目标数据太高维复杂；X 和 Y 有比较复杂的映射关系，比如一对多映射和虚假映射；X 和 Y 缺少足够的配对数据。

最近流行的数据生成模型及其在 Regeneration Learning 范式下的表示

下面简单梳理了近年来在 AIGC 内容生成领域的一些典型的模型方法，比如文本到图像生成模型 DALL-E 1、DALL-E 2 和 Stable Diffusion，文本到音频生成模型 AudioLM 和 AudioGen，文本到音乐生成模型 MusicLM，文本生成模型 GPT-3/ChatGPT，它们都可以被看作是采用了 Regeneration Learning 类似的思想，如表 4 所示。

Model	X	Y	Y'	$X \to Y'$	$Y' \to Y$
DALL-E	Text	Image	Visual Token	AR	VQ-VAE Decoder
DALL-E 2	Text	Image	CLIP Latent	AR/Diffusion	Diffusion
Stable Diffusion	Text	Image	VAE Latent	Diffusion	VAE Decoder
AudioLM	Audio Prompt	Audio	Semantic/Acoustic Token	AR	VQ-VAE Decoder
AudioGen	Text	Audio	Audio Token	AR	VQ-VAE Decoder
MusicLM	Text	Music	MuLan/Semantic/Acoustic Token	AR	VQ-VAE Decoder
GPT-3/ChatGPT	Text Prompt	Text	Text Prompt	Chain-of-Thought Prompting	

表 4：最近比较受关注的数据生成模型及其在 Regeneration Learning 范式下的表示

机器学习 / 深度学习依赖于学习范式指导处理各种学习问题，如传统的机器学习，包括有监督学习、无监督学习、强化学习等。在深度学习中，有针对数据理解任务的表征学习。微软研究员们和深度学习 / 表征学习先驱 Yoshua Bengio 一起面向数据生成任务提出了针对性的学习范式 Regeneration Learning，希望能指导解决数据生成任务中的各种问题。微软亚洲研究院机器学习组的研究员们将 Regeneration Learning 的思想应用到各类生成任务中，比如文本到语音合成，歌词到旋律生成，语音到说话人脸生成等，详情请见：https://ai-creation.github.io/。

结语

本篇文章介绍了微软亚洲研究院机器学习组在 AIGC 数据生成方面的研究范式工作，首先指出了数据生成面临的挑战以及新的学习范式的必要性，然后介绍了 Regeneration Learning 的具体形式、与 Representation Learning 的关系、当前流行的数据生成模型在该范式下的表示，以及 Regeneration Learning 潜在的研究机会。希望 Regeneration Learning 能够很好地指导解决数据生成任务中的各种问题。在这一研究方向上，机器学习组还开展了模型结构和建模方法以及具体的生成任务方面的研究，欢迎继续关注我们的其他文章！

相关论文：

Regeneration Learning: A Learning Paradigm for Data Generation
https://arxiv.org/abs/2301.08846

作者简介

谭旭，微软亚洲研究院高级研究员

研究领域为深度学习及 AI 内容生成。发表论文 100 余篇，研究工作如预训练语言模型 MASS、语音合成模型 FastSpeech、AI 音乐项目 Muzic 受到业界关注，多项成果应用于微软产品中。研究主页：https://ai-creation.github.io/

深科技 | 胡瀚：成功用 Swin Transformer 连接 CV 和 NLP 主流架构的“破壁人”

计算机视觉和自然语言处理分别是计算机科学的重要研究方向，然而长久以来，两个领域遵循着截然不同的研究范式。在 Transformer 架构出现后，能否让其强大的通用性从自然语言处理拓展至计算机视觉的研究与应用中，打破两个领域间的“次元壁”成为了研究者们不断探索的问题。

微软亚洲研究院高级研究员胡瀚，成为了首个“破壁人”。他所提出的 Swin Transformer 促进了视觉 Transformer 取代长期统治视觉骨干网络的卷积神经网络，让计算机能够像理解语言一样看懂世界。凭借这一里程碑式的开创性成果，胡瀚作为“远见者”入选了 2022 年度《麻省理工科技评论》“35 岁以下科技创新 35 人”中国榜单。

语言翻译、回答问题、生成文本……最近大热的 ChatGPT 可以被看作是大型语言模型的里程碑式进步，其展现出来的通用性和可靠性令人赞叹，可以说其基本上解决了自然语言处理领域的所有问题，也让人看到了实现通用人工智能的曙光。

这让计算机视觉领域的相关研究人员不禁设想，是否可以利用类似的方式来解决通用的视觉问题？如果能同样解决视觉问题，那将为现在强大的语言模型装上眼睛，让它能去更广阔的物理世界进行探索。

要想实现这一目标，一个重要的基础是视觉和语言在建模和学习上的统一。然而，长期以来，研究人员一般采用 Transformer 架构解决自然语言领域的问题，而采用卷积神经网络处理各种视觉任务。由于 Transformer 具有很强的通用性，所以能否让 Transformer 在计算机视觉中得到应用，推动这两个领域甚至更广阔的人工智能应用朝着统一的方向发展，助力解决更为广泛的智能问题呢？

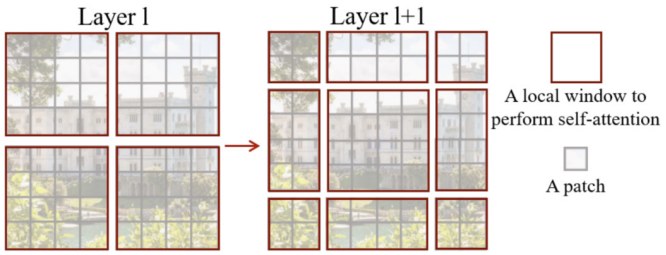
微软亚洲研究院研究员胡瀚，长期从事计算机视觉的研究工作，致力于推进计算机视觉与自然语言处理建模和学习的融合和统一。他所提出的 Swin Transformer，成为了推动视觉 Transformer 取代长期统治视觉骨干网络的卷积神经网络的一个里程碑工作。凭借这一开创性的成果，他成为 2022 年度《麻省理工科技评论》“35 岁以下科技创新 35 人”中国入选者之一。

2022 年度《麻省理工科技评论》“35 岁以下科技创新 35 人”中国入选者

提出 Swin Transformer，助推视觉 Transformer 的大规模研究

在清华大学自动化系读博期间，胡瀚就开始了对计算机视觉的研究。当时，他受到人类视觉机制的启发，尝试使用更全局系统的方式来解决视觉分割问题，并在视觉的基本原则方面有了一些掌握。博士毕业后，他继续从事计算机视觉研究。在很早的时候，他就坚信要想实现更通用的人工智能，不同领域在建模方面的统一将是一个重要的基础。在 2017 年 Transformer 出现后不久，他就看好这一架构的强大通用性，并开始积极尝试将 Transformer 引入到视觉领域中。他早期的尝试包括基于 Transformer 实现学界首个端到端的物体检测器（2017 年），以及在 2019 年首次将 Transformer 用于视觉骨干建模，尽管效果不错，但这一神经网络遇到了实现效率问题而不太实用，也没有成为主流。

两年后，他于 2021 年提出的 Swin Transformer 解决了其中的效率难题，从而推进了这一网络在视觉领域走向实用。在这个工作中，他创造性地提出了“移位窗口”方法，该方法无需同时处理数以千计的局部窗口，可以将需要处理的窗口数量降低 50 倍，这大大提升了计算的并行性，在 GPU 上取得了 3 倍的速度提升。



Swin Transformer 中的一个关键技术：移位窗口方法

胡瀚和团队首次证明了 Transformer 网络能够在非常广泛的视觉问题中大幅超越卷积神经网络，推动该领域大规模兴起了对视觉 Transformer 的研究。“当时我们很快做了开源，把一些实现细节分享给了整个领域。有了这个基础，其他研究者才能更快地去追随并开展进一步研究，进而共同推进该领域的发展。”他说。

作为项目负责人，他以《Swin Transformer: 使用移位窗口的分层视觉 Transformer》（Swin Transformer: Hierarchical Vision Transformer using Shifted Windows）为题发布在预印本平台 arXiv[1]。据悉，该成果获得了计算机视觉国际大会的最佳论文（马尔奖），这一奖项被视为国际计算机视觉领域的最高荣誉之一。同时，相关论文在一年多时间获得超过 5000 次引用和超过 10000 次 GitHub 标星。

目前，Swin Transformer 正作为骨干网络广泛地应用于计算机视觉领域，并为全球亿万人的工作和生活带来了很大的改变。比如，其已在微软产品 PowerPoint 的视觉素材推荐中获得应用，正在帮助用户制作更具美观性的设计和演示文稿。另外，其还被应用于图像搜索、元宇宙、自动驾驶和机器人等诸多领域。

与此同时，Swin Transformer 所代表的大一统趋势，有利于大大简化芯片设计，也有利于更通用的人工智能模型的开发。

“Swin Transformer 所解决的计算机视觉长期与自然语言的主流架构不匹配的问题是一个更宏大目标中的第一步，即实现和人脑一样用一个通用模型和类似的学习机制去解决各种智能问题。”胡瀚表示，目前他正在继续攻克这个更宏大目标上的各种挑战，比如如何有效地扩展计算机视觉和多模态模型并将其稀疏化。

人脑具有的强大智能，以及能从少量样本中学习新智能的能力，很大程度上来自于其海量的百万亿级的神经连接。同时，需要说明的是，连接的稀疏性，又能让大脑变得非常节能。因此，开发有效的视觉大模型和稀疏模型，对实现强大而通用的智能来说非常关键。

他通过解决训练稳定性、视觉任务分辨率鸿沟，以及基于自监督预训练解决海量标注数据需求的问题，成功地训练了拥有 30 亿参数的稠密视觉模型 Swin Transformer v2.0 版本。作为截止 2022 年 8 月世界最大的稠密视觉模型，Swin Transformer v2.0 版本当时在多个重要的代表性视觉评测集中取得了新的记录。此外，胡瀚还参与开发了目前 GPU 上最高效的混合专家框架 Tutel 和用于计算机视觉的 Swin-MoE 模型。

促进 AI 不同领域实现大一统和大融合，赋能人类美好生活

回顾自小的成长背景和求学经历，胡瀚感恩于父母对教育的重视。“小时候尽管家境清贫，而且当地辍学率比较高，但他们仍旧支持我一路读到博士。”他喜欢没日没夜地看书，有时也会睡在图书馆，很小就意识到知识的无边界以及世界之大。

在他看来，到北京也是冥冥之中的缘分驱使。“记得高考完找班主任老师给自己写寄语，他写了‘北上’这两个字，后来真的有幸北上到北京求学。后来，博士期间的导师周杰教授，给我们创造了引导为主，鼓励自由探索的氛围。这很适合我，让我的思维自主创新和自我求索方面的能力有了很大的提升。”

他认为，走上研究之路，是命运的安排。他觉得自己遇上了三个转折点：第一个是能够有幸从小乡村考入清华，和最优秀的同学一起学习生活，接受国内最好的教育。第二个是在读研时选择了人工智能方向，也遇到了适合自己的导师，亲眼见证当时发展不算最火热的人工智能，逐渐变得越来越重要，并不断地改变着世界的面貌。第三个是进入微软亚洲研究院工作，与国内人工智能方向最优秀的研究员和前辈合作，在一个适合做研究的土壤下生根发芽，做出了具有代表性的工作。

作为一个科研工作者，他将自己的主要目标定为研究新的生产力工具和突破性技术，从而赋能每个人、每个机构和整个社会。因此，他希望能够推动人工智能不同领域之间的大一统和大融合，让其在未来能像人脑一样，用一个强大通用的模型，就能解决各种复杂的智能任务。“这样的人工智能系统将有望和 100 年前的电力革命一样，改变整个人类社会的生活和生产方式，促进社会进步，让每一个人的生活都能更加美好。”胡瀚说。

对于未来，他也给自己制定了新的研究目标，他希望攻克通用视觉问题，为实现对图像的可靠（几乎不出现错误）而全面理解和生成而努力。“对于这一个目标来说，方法论并没有本质区别。既然 ChatGPT 可以几乎可靠地解决几乎所有自然语言处理方面的问题，那么我相信，通用的视觉问题也是能够得到可靠解决的。”胡瀚如是说。

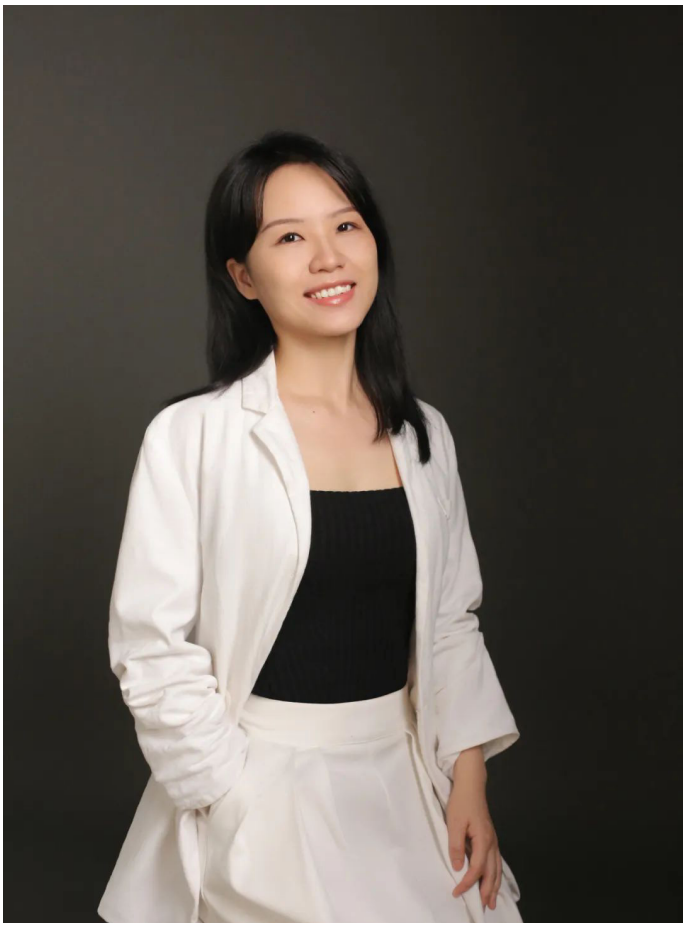
外滩教育｜王希廷：做自己的最优指标创造者

从小学霸、清华本硕博连读，再到人工智能科学家……微软亚洲研究院高级研究员王希廷的经历像极了"别人家的孩子"。虽然如此优秀，她却经历过一段充满迷惘的挫败期。读博、做科研期间，她发现追求极致的卓越和以主流标准优化自己的评价体系失效了，甚至成为了进一步发展的阻碍。想在未知的科研中创新，更需自己从外界指标驱动转向追求内心的驱动力。她是如何减少内耗，听从自己的声音的？又是如何结合人工智能和心理测量，成为微软亚洲研究院负责任的人工智能研究的一部分的？跟随这篇文章，和王希廷一起与自己和解，获得科研与生活中的松弛感。本文经授权转载自公众号"外滩教育" (ID:TBEducation)（文丨吴玉莹，编丨 Iris）。

凭借自己的才能和努力，最终获得社会认可的成功，是一件好事吗？在大多数人想来，这是毋庸置疑的好事。

然而戴着成功的光环一路走来，从一个追求极致的学霸，到清华本硕博连读，再到微软亚洲研究院的人工智能科学家，王希廷用亲身经验得出的结论和大多数人的臆断不同：

即使按照主流标准优化自己到极致，总会出现一个阶段，显示这种优绩主义失效了。



王希廷

这种失效正如哈佛大学政治哲学家迈克尔·桑德尔所指出的，优绩主义将人粗暴地分为胜利者和失败者，但即使作为胜利者，也将面临毁灭性的损失。

"损失"不是外在的失去，而是内在的缺损。希廷这样回忆，"有一天，我发现自己人生重要的时间节点都由考试、论文发表、参加学术会议而标记，内心感到很空虚、无望。"

"获得让自己满意和认可的人生"，成为希廷新的目标，借由对自我内在的挖掘，无论是作为一名女性、一个母亲，还是一名科研工作者，她都重新发现了新的自己。

以下为第一人称自述：

"做到极致"和"以主流标准优化自己"，是加持还是阻碍？

在整个受教育的成长过程中，我从小就把"追求极致"和"以主流标准优化自己"当作自然而然融入灵魂的本能。说是自然而然，应该跟我的家庭背景有关系，我的父母本身就是特别追求极致的人。

爸爸毕业于大连理工大学，他可以自己学砌墙、装修房子，工作上因为形势所迫换了专业，很快能学会电脑，最近我家装修房子都是他自学画的 CAD 图，所以他学习能力很强，同时自我要求很严格。我妈妈也是典型的理工女，她是物理老师，偶像是牛顿，在 90 年代大家还不太用电脑的时候就自己装家里的 Wi-Fi，家里的电线、双联开关都是她自己动手。

正是因为有追求极致的父母，所以追求极致自然而然成为他们对我，以及我对自己的要求。相反，如果没有什么事情做到最优，他们很容易就会发现，然后提出改正建议，尽量达到最优。

那时候也会想既然确定了要优化，优化到什么方向呢？最自

然的就是优化到父母、老师、同学们甚至不熟悉的叔叔阿姨们都认可的方向——学习成绩。可以说我是改进和优化原则的受益者”，因为在这个过程中我获得了父母的肯定和认可、老师的喜爱，以及别人对父母的羡慕、父母对我的自豪等等这些来自外部的赞美。以至于这套做事原则深入到我的灵魂，碰到很多事脑子里都有个声音在提醒我哪里可以改得更好。

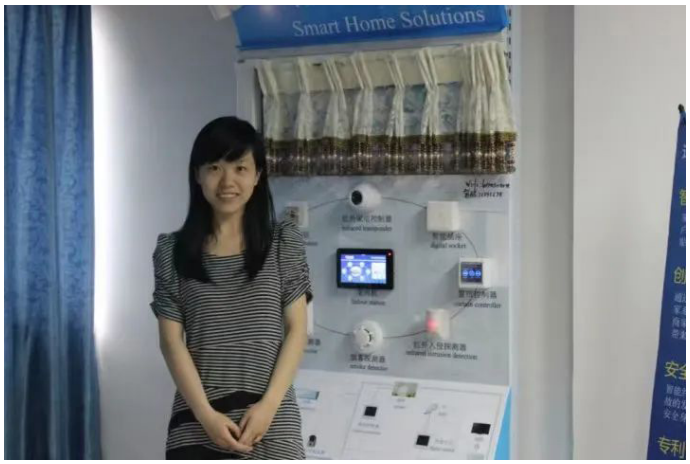
“极致”和“以主流标准优化自己”成为两个正相关的、互相强化的原则，因为不够极致，所以要不断优化，学习成绩第一、上清华、本硕博连读、发顶刊论文、超高的论文引用率在不同阶段成为我不同的优化标准。直到读博、做科研期间，我发现这两个原则不但越来越没有作用，还让我第一次产生很深的挫败感，甚至成为我进一步发展的阻碍。

以前我觉得所有事情只要努力做，就能把它做好，但到了读博写论文的阶段，我觉得非常努力了，天天熬夜去做科研，但是没有成果，整个人就很焦虑和恐惧，觉得自己是一个失败者，做不好任何事情。到现在我知道的‘读博故事’跟我当时的状况都很像。

一方面，博士阶段的科研创新不是“努力”就能完成的，它可能更需要的是有内心感召力的科研方式。但很多人在读博的开始还是依据“努力”准则来做事，比如天天熬夜，比如周末都不休息，甚至我听说过有的女博士周末一个人在实验室，做不出什么，就在那里哭。这个就真的是错误的努力方向。

更重要的是在另一方面，当你做不出科研成果，就觉得自己是失败的、做错的、不行的，长期在心理上的自我否定会带来更多的恐惧和焦虑。

这个就是曾经的“优绩主义”失效，甚至成为阻碍的过程。可能每一个“优绩主义”的受益者多多少少都会尝过“失败者”的滋味。回头来看，到了现在这个成长阶段，外界的主流标准对我做事情起到的作用越来越少，如果只靠外界的主流标准来支撑自己，它就变成了一个非常单薄的数字，一种非常表面的、遥远的东西，跟我的内在关联越来越弱。



王希廷博士暑期实践在佛山留念

同时，在反思过程中也会发现，与其靠外在，不如靠内在，自己内在的感受、思考、驱动力才有可能支撑自己前进，越是到了后面想要在科研上做出创新的、有影响力的事情，就越需要我们找到自己内心的动力。

从自我 PUA 到自我赋能，心理学教会我如何听从内心的感召

如何找到内心的动力？我首先想到的是我们本科心理老师的一句话：你们学理工科的，总觉得理工科的事情特别难，但最难、最复杂的是人性。

当时的我还想代表理工科跟代表“人性”的心理学争一下胜负，等到读博时候恐惧焦虑大爆发，很快意识到心理学的重要性，无论是寻求心理咨询师的帮助，还是自己主动找一些心理学相关的书籍，都为我更长远的发展提供了精神支撑的资源。

心理学带给我更多视角看待“极致”和“优化”。比如读博时候，极致的努力和过分的自我苛刻其实是一种自我 PUA，当我为了主流标准去努力、没有做到就认为自己是失败者的时候，我从来没问过自己的内心，到底什么才是我最想要的。现在回想，其实读博的出路就在于关注自己的内心，当你看清楚了自己内心恐惧和焦虑的是主流标准，那么内耗的阶段、自己内心斗争的阶段会少很多。

同时我也明白自己对以主流标准优化自己带来的挣扎和痛苦，其实也是一种脑子和心的冲突，当我听从内心的力量以后，那种困顿反而消失了。比如成为妈妈后，我不会以主流指标和优绩主义来规范自己的女儿。

小的时候，只要我成绩好，我可以不做任何家务，那时候我认为这是优绩主义给我带来的好处；但现在作为妈妈，我的女儿才两岁，我就很鼓励她自己去劳动，她有自己的小扫把和小簸箕，会自己扫地、拖地，我希望她可以通过体会劳动过程中的收获和成就感，从而感受到自己的力量。

除此之外，在育儿的过程中，对爱的感受让我觉得，外在标准带来的成就感绝对没有爱带来的力量更能让个体受益。有一次，公司的心理咨询师让我想象自己特别爱的一个人，然后再想象对这个特别爱的人说什么，对 TA 有一个什么样的美好愿望。闭着眼睛，我就想到特别爱女儿，很希望她能够快乐幸福地生活。

那个时候我感受到了爱，同时也觉得自己很有能量。事后我就反思，从爱孩子的角度出发，以主流指标优化自己真没有那么重要。我希望她幸福快乐，就是她最自然的成长过程。

其实相对于女儿的独立个体，作为妈妈的我都可以看作是一种外在的价值标准，那么在育儿过程中我察觉到自己“主观的、外

在的"因素，对我和女儿的成长都是至关重要的。

有时候女儿在玩一个游戏，搞了半天她都没有找到诀窍，虽然看着她我很着急，但我会告诉自己尽量"忍一下"。可能过了不到 10 秒钟，她就找到诀窍了，然后玩得很开心，我也会很高兴。在这个过程中，我更了解她，更信任她，同时她也会对自己更有信心，如果我剥夺了她探索的乐趣，那么她以后可能就会以外在为标准，弱化了内在的力量。神奇的是，在感受到对女儿强烈的爱之后，我还学会了好好爱自己，甚至发现了原来不曾感受到的与周围人之间深刻的联结。

比如女研究员之间的联结。在微软亚洲研究院差不多和我同期怀孕、生子的女研究员有四五个，我们自发地吃饭、聊天，从孕期体重增长、孕后身材恢复、再到产后适应工作，因为身边有人互相认同和鼓舞，所以一点也不觉得孤单。

那种彼此之间对母亲和女研究员身份所遇到问题的深刻共鸣，对爱和育儿的共识，让我们紧紧连在一起，这是我从前从未有过的特殊体验。

在微软亚洲研究院举办的 Ada Workshop 2023 活动现场，王希廷与同学们进行分享

将人工智能和人类做对比，重新发现人

可以说心理学不仅帮我解决了读博时期和新手妈妈时期的焦虑、恐惧，从更广泛的视角来看，心理学还给我的科学研究带来新的生机，借由这种生机，可能也给在普遍教育困境中的学生和家长们，一些关于教育的启发。

人工智能时代来临，首先给科研和教育带来冲击。微软亚洲研究院院长周礼栋在一次演讲中提到，"今天我们应该像一名冲浪者一样做科研"。以前的（或者传统的）做科研，更多的是一种保守的、增量式的科研方式，就是你认认真真去做了，按部就班去工作就可以了，它更加像一种平庸的优秀，在舒适圈中做科研。

而冲浪者式的科研是在新时代里，需要一种真正有感召力的、创新的、充满勇气去挑战未知领域的科研方式。它需要调动自己强大的内心力量，做一种引领性质的工作，因此它提出的要求比传统科研要高得多。

在"极致"和"优化"教育原则下成长起来的我，对于一种没有指标的"优化"，或者说创造一种全新指标的"优化"非常陌生。举一个例子，刚到研究院的时候，我有一个不适应的阶段。以前在读博时科研是非常明确的，导师规定研究方向，然后指导你一步步做出来，包括指正你的错误、如何发表文章等等细节。

但在研究院，每个人都是独立的科研贡献者（individual contributor），你要自己找研究方向，自己说明它的研究价值，然后自己发表文章。研究院也不要求发表文章的数量，它要求的是有影响力。我要如何去优化"影响力"这样的抽象指标呢？

我想起了周明老师的例子。周明老师是当时研究院非常有影响力的研究员之一，他在科研中做了自己想要做的事，最自豪的就是为中国自然语言处理领域做出了自己的贡献，他培养了很多学生，主办了中国自然语言处理会议，建立了中国自然语言处理的研究社群。

这样的指标其实已经从数字变成一种精神，首先是发自内心热爱自己的研究，在此基础上，达成一种共同的理念：做负责任的人工智能，真的能为人人服务，真的能打动人的内心。

这是个慢慢体会和理解的过程，在学习自我赋能的过程中，我也逐渐领悟，无论是教育还是科研创新，问其他人是没有答案的，你要从自己的内心去寻找答案，甚至创造答案。

正是在不断学习心理学的契机下，我发现了把人工智能和心理测量相结合的科研方式，这也是微软亚洲研究院负责任的人工智能研究的一部分。

通过测量人工智能某些属于人的稳定的高阶能力，比如创造力、沟通能力、解决问题能力等，我们发现：人工智能更擅长发散性思维，因为它的"脑容量"比正常人类的脑容量大很多，因此它的搜索能力、记忆能力、发散能力都很强。

另外，人工智能在创造力方面还有不足。我觉得这是因为它的优化目标没有人类这么丰富多彩，也就是只有人类真的知道什么东西对我们的社会发展很重要，或者说因为人工智能没有内心的感召这个部分，所以它更多的是在现有数据中做一个能力很强但没有什么情感的专家。

也是在这样的对比中，让我们更为理解"人类社会该往哪个方向发展，我们如何做到对人的关怀，教育中的创造力应该以什么为核心思想"等等问题，是牵扯到未来人类发展的、需要我们非常仔细去思考的部分。

从一个女孩长成女性科研人员，再到成为母亲养育女儿，我自己的成长和教育经过了从外部评价到内部的自我召唤、自我肯定的过程。

在今天这个极速变化的时代，如果我能提出一些建议的话，那就是找到心之所向，活出最优秀的自己。

做到以主流标准来优化自己，在当下的教育体系内自然会得到很多好处，同时，跟随内心的感召、发现内在的力量，才能找到一个人前进中很重要的方向。

如果你不知道自己真正想要什么，就没有办法调动自己内心的全部力量，无法获得让自己认可、满意的人生。

扫描二维码查看视频

相关阅读

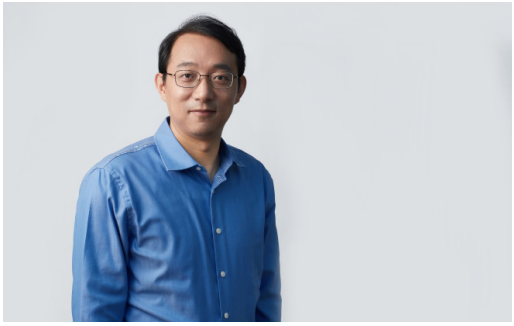
扫描二维码查看文章

科学匠人 | 李琨：执著于高性能计算研究的“别人家的孩子”

优秀的科研和项目履历让李琨获得了多家知名企业的 offer，同时也被网友称为“别人家的孩子”，成了校园内的“网红”。现在，这个“别人家的孩子”已经加入微软亚洲研究院，成为了“异构计算组”的一名研究员。李琨因为他的博士毕业论文《大规模并行多层次不连续非线性可扩展理论研究及应用》获得了 2022 年度“CCF 优秀博士学位论文激励计划”（简称“CCF 优博奖”）。“CCF 优博奖”代表着自己多年来在高性能计算领域的研究工作得到了学术界的肯定，坚定了他继续从事科研探索的动力与决心。

科学匠人 | 梁傑然：长期主义研究者的心法秘诀

微软亚洲研究院高级研究员梁傑然（Mike Liang）关于 AI 模块化研究的论文“On Modular Learning of Distributed Systems for Predicting End-to-End Latency”被国际顶级网络领域学术会议 NSDI 2023 接收。与此同时，梁傑然此前的研究工作“Design and Evaluation of a Versatile and Efficient Receiver-Initiated Link Layer for Low-Power Wireless”还荣获了国际移动计算和感知领域顶级会议 ACM SenSys 2022 时间检验奖（Test of Time Award）。一项研究成果，经受住时间的检验，十二年之后再获认可，这对研究员来说是一种怎样的体验？梁傑然是如何做到持续创新与坚持长期主义研究的？现在的他又有着怎样的研究愿景？



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