微软亚洲研究院院刊

Matrix

NO.64
2023年 1 -3 月

微软CTO对话比尔·盖茨：GPT-4与人工智能的未来

微软3D生成扩散模型RODIN，秒级定制3D数字化身

完成一幅计算机学术生态的拼图，少不了这些“斜杠女性”

01 焦点

微软 CTO 对话比尔·盖茨：GPT-4 与人工智能的未来 2

微软亚洲研究院副院长邱锂力获选 2022 年度美国国家发明家科学院院士（NAI Fellow） 4

02 前沿求索

微软 3D 生成扩散模型 RODIN，秒级定制 3D 数字化身 5

让天下没有难训练的大模型，微软亚洲研究院开源 TorchScale 7

高精度压缩 Transformer，NNI 剪枝一站式指南 10

打破拓扑结构的天花板，深入探究属性图上表示学习的研究蓝海 13

科研第一线

AAAI 2023 | 微软亚洲研究院精选论文 19

微软亚洲研究院项目入选《2022 CCF 技术公益年度案例集》 19

03 文化故事

科学匠人 | 梁傑然：长期主义研究者的心法秘诀 20

科学匠人 | 李琨：执著于高性能计算研究的“别人家的孩子” 23

完成一幅计算机学术生态的拼图，少不了这些“斜杠女性” 25

实习派 | 王一栋：主动就会有故事，高效科研秘诀大公开 28

关于内卷与反内卷、建立学术社交网络，听听过来人的建议 30

这次开学，我们请来了 ChatGPT 和各位前辈指点迷津 30

04 观点

对话 | 为 “冷门绝学” 甲骨文研究插上科技之翼 31

带你读论文 | 了解 AIGC 音频 / 图像数据生成，这几篇论文给你划好了重点 34

05 媒体报道

机器之心 | 微软多模态 ChatGPT 来了？ 16 亿参数搞定看图答题、智商测验等任务 37

机器之心 | DSN-DDI：双视图表征学习实现药物间相互作用预测性能突破 40

央视新闻 | 最潮中国范儿，当甲骨文遇上新科技 41

「AI 中国」机器之心 2022 年度评选结果公布 41

微软 CTO 对话比尔•盖茨：GPT-4 与人工智能的未来

一系列技术变革引领我们走到今天，并深刻影响着人类社会。如今，随着人工智能技术的快速发展，ChatGPT、New Bing、GPT-4 等新产品和新技术的陆续发布，又将如何帮助我们创造未来？在微软与 OpenAI 的密切合作中，微软执行副总裁兼首席技术官 Kevin Scott 一直在思考一个问题：人工智能领域出现的惊人革命对 OpenAI、对微软、对所有利益相关者以及整个世界的意义是什么？

针对这一系列问题，Kevin Scott 与比尔·盖茨进行了一次深入的探讨。本文节选了对话中的部分内容，完整对话请扫描文末二维码观看视频。让我们跟随他们的对话，一起了解比尔·盖茨对 GPT-4 的初体验，以及他对人工智能技术未来发展趋势和影响的看法。

Kevin Scott: 过去几年中，技术领域出现了一些有趣的进展，尤其是微软与 OpenAI 合作研发的 GPT-4 和 ChatGPT。GPT-4 在 OpenAI 外部的第一个实例展示在 2022 年 8 月就进行了，你当时看到 GPT-4 之后的感受是怎样的？

比尔•盖茨： 人工智能一直是计算机科学的圣杯。在机器学习出现之前，人工智能的整体进展相当缓慢，即使是语音识别也只是勉强能做到尚可的程度。之后，在机器学习领域，尤其是感知、语音识别、图片识别等迎来了快速发展。但是这些模型在复杂逻辑上仍然存在不足，无法像人类一样说话、阅读和做事。

早期的文本生成模型缺乏上下文理解能力。它可以生成一个句子，例如它前面会说 “Joe 身处芝加哥”，几句话后又说 “Joe 身处西雅图”。从局部来看，它生成的句子很好，但从人的角度来看，这是前后矛盾的。

当时，OpenAI 和微软的团队对 GPT-3 甚至 GPT-4 的早期版本抱有极大的热情。我对他们说，“如果它能够通过 AP Biology（大学进阶生物学），对训练集之外的问题都能给出充分合理的回答，那么它将是一个重要的里程碑，所以请你们继续努力。”

我以为他们至少需要两三年才能成功。令人惊讶的是，2022 年 9 月初，OpenAI 和微软团队来我家做客并向我演示最新的模型时，他们让我问一些 AP Biology 问题，让人震惊的是，除了一个与数学相关的问题之外，它都能准确作答。我还问它，“对一位生病孩子的父亲会说些什么？”它给出了非常细腻体贴的答案，比当时房间里所有人的回答都好。后来，我得到了一个账户，我让它写大学入学申请、写诗歌、让它根据某些剧情写一集《Ted Lasso（足球教练）》剧本……真的难以想象它的极限在哪里。

尽管还有一些问题有待完善，但这将是一个根本性的变革，现在自然语言可以作为人机交互的主要接口，这是巨大的进步。

Kevin Scott: 对人工智能、GPT-4，可探讨的问题有很多，先说说它不擅长的地方，我们最不希望给人们一种它是一个 AGI（通用人工智能），它是完美的，它不需要做很多额外的工作来改进它的印象。你刚才提到的数学问题就是其中之一，那么随着时间的推移，你认为人工智能系统还需要在哪些方面做提升？

比尔•盖茨： 当向机器提问时，关于上下文的背景知识是个普遍问题。例如，先让机器讲个笑话，接着又问它一个严肃的问题，人类可以从你的面部表情感受变化，知道现在已经不再是玩笑的语境，但是人工智能却还在继续那个笑话。这种语境感在交互中有着极大的作用。

另外，在解决问题的难度方面，当我们做同一道数学方程式时，可能需要五、六次简化才能将其转化为正确形式，并且不断学习这些简化。而机器推理是通过层级线性下降推理链实现的，如果简化需要运行 10 次，机器可能就不会了。数学是非常抽象的推理，这是人工智能最大的弱点。

矛盾的是，它又可以解决很多数学问题。如果你让它以抽象形式解释某些问题，本质上是给出一个与问题相匹配的方程或程序，它会做得很完美，完全可以将它当作一个求解器。然而，如果做数值运算，那么它就会经常出错。无论是哪一个薄弱环节，都需要花时间才能解决，要严肃对待。我们需要更创新的模式，

通过提示词（prompts）或训练对模型进行数学方面的训练。

如何评价 GPT-4 呢？那些说它很糟糕的人错了，那些说它是 AGI 的人也不对。我们的观点介于两者之间，要做的是确保它能用正确的方式被使用。

Kevin Scott：我们都知道，你亲身经历了好几次重大的技术变革并有自己独特的视角。在现今又一个重大变革时刻来临之时，你对于那些正在考虑使用新技术的人有什么建议？他们应该如何使用新技术？这与你在 PC 和互联网时代的想法有什么关联？

比尔•盖茨：最初的计算机并不能为个人所用，之后微处理器的出现和大批公司的努力才有了个人电脑，IBM、苹果和微软又都参与了软件开发。然后，互联网将这些连接起来，再后来又演进出了移动计算、手机。数字世界极大地改变了我们的生活。

能够读与写的计算机的诞生，与上述节点中的任何一步一样意义深远。有一小部分人认为我们可能高估了技术，这也没错。但在这次的变革中，我们低估了自然语言和计算机处理自然语言的能力，以及它对白领工作的影响，包括销售、服务、医生，我也曾认为这会是很多年之后才会发生的事情。

人工智能的新阶段才刚刚开始，我们正处于对它狂热的阶段，就像曾经对互联网的狂热一样，当然现在回过头来看，互联网已经成为了重要的工具。这是一次巨大的突破，是整个数字计算机领域的里程碑。

Kevin Scott：我一直在思考一件事，从 Ada Lovelace 编写出第一个计算机程序至今，让数字机器（digital machine）为人们工作是有技术门槛的，你必须是一位熟练的程序员，要了解客户的需求，然后构建软件才能让机器为你做事。

现在，有了自然语言接口，人工智能可以编写代码启动一整套服务和系统，这让普通人也能使用机器完成复杂的任务，而不必花多年时间学习专业知识，对此你怎么看？

比尔•盖茨：技术的每一次进步都降低了人们使用它的门槛。电子表格就是一个例子，尽管仍然需要理解公式，但却不必深入理解逻辑或符号。有很多程序可以帮你将公司数据进行可视化，或进行复杂查询，从而了解人员流失和销售业绩的情况。你不必去 IT 部门排队等候，再让他们告诉你。

无论是查询、汇报，或者触发工作流和某项活动，你只需要用语言描述就会生成一个程序，有一整套的查询和编程工具，供所有人使用。人工智能正在赋予人们最直接的互动能力，这也是当下我们正在努力的课题。

Kevin Scott：从个人的角度来看，最令你兴奋的事是什么？你非常关心教育、公共卫生、气候和可持续能源等领域，人工智能对这些领域会产生哪些影响？

比尔•盖茨：我们一直在思考健康和教育问题。在医生少、获得医嘱建议困难的卫生系统中，AI 赋能医疗的研究将很有意义。另外，所有人都希望有一个私人教师来提供帮助。比如，在一些特殊的学校中，学生在写作方面会收到教师的逐行反馈，但对大多数孩子来说，并不能得到一对一的指导。

我认为教育会是最有趣的应用领域，其次是健康领域。当然，这些技术在销售和服务类场景中也有很大的商业机会。比如，在一个有着二、三十人的班级中，教师无法单独关注某一个学生，无法同时了解每一个人的行为动向。而在多个学科领域利用人工智能技术对话、反馈，可以有效提升教育水平。

我们必须承认，计算机在彻底改变教育方面还有很多事要做。接下来的 5-10 年里，我们需要从新的角度考虑学习问题，以及如何在教育中提供帮助，而不仅仅是通过计算机查找材料。

Kevin Scott：这是个全球性的问题。我们也看到父母的参与对孩子的教育有很大影响。有的父母工作繁忙，很难与孩子接触，想象一下，有这样一种技术，它不在乎你说什么语言，可以在家长和老师之间架起桥梁，帮家长了解阻碍孩子成长的问题，甚至对孩子进行个性化教育，真正解决眼前的问题，这非常令人兴奋。

那么，你认为在接下来的 5-10 年里，我们还将面临哪些挑战？我们继续努力的方向是什么？

比尔•盖茨：我认为，在算法的执行方面会有一系列创新，很多芯片从硅到光学的转变将可以减少能源和成本。英伟达目前在这方面处于领先地位，将来也会出现更多的挑战者，因为大家希望在运行、训练上的成本越低越好。理想情况下，我们希望模型可以运行在端侧，这样就可以在独立的客户端设备上进行操作，而不必去云端获取。

软件方面也将面临巨大挑战。例如，用户是需要特定版本，还是持续改进的版本？即使是微软也会同时追求这两种目标。理

想情况下，我们希望针对不同的领域，通过训练数据，甚至可能是一些适用于它的前置检查、后置检查的逻辑，来更准确地处理不同的需求。

除此之外还有许多社会问题，包括促进教育、医疗的发展等等。微软一直致力于提高生产力，未来，有些事情将会自动化，最终有的任务可能只需要一个人来完成，但这个人将比以往能够完成更多的事情。由此带来的挑战和机遇也会很多。我看到 OpenAI 的团队正在探索其中，但我相信很多其他机构和组织也在推动相关工作。技术创新的速度将更胜以往，以此为目标的人力、资源和公司的数量远远超过了以前。

Kevin Scott：我职业生涯的早期，大部分时间都是作为一名计算机科学家接受培训的，编写编译器，编写大量汇编语言和设计编程语言，或者在研究生院进行并行优化和高性能计算机体系结构的研究。离开研究生院之后，我想我再也不会使用这些东西了。然而今天我们在建造超级计算机来训练模型时，这些技术又有了用武之地。如果现在你是一个 20 多岁的年轻程序员，你会对哪些技术感兴趣？

比尔•盖茨：这里面有相当多的数学的元素。很幸运，我曾经做了很多与数学相关的事，这是通往编程的大门。有些编程人员没有数学背景，我建议他们去掌握一些数学知识，因为很多计算都不只是编程问题。

最初的 Macintosh 是一台 128K 的机器，其中 22K 是位图屏幕，几乎没有人能够编写出适合的程序，只有微软和苹果成功了。但现在你用数十亿个参数来操作这些模型，那么我们是否可以跳过一些参数，或简化一些参数，或进行预计算？在资源受限的机器上，优化变得尤为重要。

尽管过去半年在计算加速方面的进展比预期要好，但未来几年，又将面临多大的资源瓶颈？我们如何确保企业以更明智的方式分配这些资源？无论如何，在计算机科学的几乎每个领域，包括数据库类型技术、编程技术等方面，都需要我们以一种全新的方式来思考。

Kevin Scott：最后想问一下，你在工作之外会做些什么事？我们都知道你很喜欢阅读，经常提着一个巨大的手提袋随身携带着书籍，无论走到哪里，都会大量阅读，从科学到小说，无所不包。你的阅读节奏是怎样的？

比尔•盖茨：我打匹克球有 50 多年了，我也喜欢打网球和读书。我最近一年读了 80 多本书，包括 Thomas Sowell、Vaclav Smil、Steven Pinker 的书，这些作家的思想重塑着我的思维。同时，阅读也能让我放松心情。我想我该多读些小说，人们向我推荐了很多好的小说，这也是我会在《盖茨笔记》上分享我的书单的原因。

扫描二维码观看完整视频

微软亚洲研究院副院长邱锂力获选 2022 年度美国国家发明家科学院院士（NAI Fellow）

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。

扫描二维码了解更多信息

微软 3D 生成扩散模型 RODIN，秒级定制 3D 数字化身

近日，由微软亚洲研究院提出的 Roll-out Diffusion Network (RODIN) 模型，首次实现了利用生成扩散模型在 3D 训练数据上自动生成 3D 数字化身 (Avatar) 的功能。仅需一张图片甚至一句文字描述，RODIN 扩散模型就能秒级生成 3D 化身，让低成本定制 3D 头像成为可能，为 3D 内容创作领域打开了更多想象空间。相关论文 "RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion" 已被 CVPR 2023 接收。

创建个性化的用户形象在如今的数字世界中非常普遍，很多 3D 游戏都设有这一功能。然而在创建个人形象的过程中，繁琐的细节调整常常让人又爱又恨，有时候大费周章地选了与自己相似的眼睛、鼻子、发型、眼镜等细节之后，却发现拼接起来与自己仍大相径庭。既然现在的 AI 技术已经可以生成惟妙惟肖的 2D 图像，那么在 3D 世界中，我们是否可以拥有一个 "AI 雕塑家"，仅通过一张照片就可以帮我们量身定制自己的 3D 数字化身呢?

微软亚洲研究院新提出的 3D 生成扩散模型 Roll-out Diffusion Network (RODIN) 可以轻松做到。让我们先来看看 RODIN 的实力吧!

(a) 给定的照片 *(b) 生成的虚拟形象*

图 1：给定一张照片，RODIN 模型即可生成虚拟形象

(a) 输入文字"留卷发和大胡子穿着黑色皮夹克的男性" *(b) 输入文字"红色衣着非洲发型的女性"*

图 2：给定文本描述，RODIN 模型可直接生成虚拟形象

与传统 3D 建模需要投入大量人力成本、制作过程繁琐不同的是，RODIN 以底层思路的创新突破与精巧的模型设计，突破了二次元到三次元的结界，实现了只输入一张图片或一句文字就能在几秒之内生成定制的 3D 数字化身的能力。在此之前，AI 生成技术还仅仅围绕 2D 图像进行创作，RODIN 模型的出现也将极大地推动 AI 在 3D 生成领域的进步。相关论文 "RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion" 已被 CVPR 2023 接收。

RODIN 模型首次将扩散模型用于 3D 训练数据

在 3D 生成领域，尽管此前有不少研究利用 GAN（生成对抗网络）或 VAE（变分自动编码器）技术，从大量 2D 图像训练数据中生成 3D 图像，但结果却不尽如人意，"两面派"、"三头哪吒" 等抽象派 3D 图像时有出现。科研人员们认为，造成这种现象的原因在于这些方法存在一个基础的欠定（ill posed）问题，也就是说由于单视角图片存在几何二义性，仅仅通过大量的 2D 数据很难学到高质量 3D 化身的合理分布，所以才造成了各种不完美的生成结果。

对此，微软亚洲研究院的研究员们转变思路，首次提出 3D Diffusion Model，利用扩散模型的表达能力来建模 3D 内容。这种方法通过多张视角图来训练 3D 模型，消除了歧义性、二义性所带来的 "四不象" 结果，创建出更逼真的 3D 形象。

然而，要实现这种方法，还需要克服三个难题：

首先，尽管扩散模型此前在 2D 内容生成上取得巨大成功，将其应用在 3D 数据上并没有可参考的实践方法和可遵循的前例。如何将扩散模型用于生成 3D 模型的多视角图，是研究员们找到的关键切入点。

其次，机器学习模型的训练需要海量的数据，但一个多视图、一致且多样、高质量和大规模的 3D 图像数据很难获取，还存在隐私和版权等方面的风险。网络公开的 3D 图像又无法保证多视图的一致性，且数据量也不足以支撑 3D 模型的训练。

第三，在机器上直接拓展 2D 扩散模型至 3D 生成，所需的内存存储与计算开销几乎无法承受。

多项技术创新让 RODIN 模型以低成本生成高质量的 3D 图像

为了解决上述难题，微软亚洲研究院的研究员们创新地提出了 RODIN 扩散模型，并在实验中取得了优异的效果，超越了现有模型的 SOTA 水平。

RODIN 模型采用神经辐射场（NeRF）方法，并借鉴英伟达的 EG3D 工作，将 3D 空间紧凑地表达为空间三个互相垂直的特征平面（Triplane），并将这些图展开至单个 2D 特征平面中，再执行 3D 感知扩散。具体而言，就是将 3D 空间在横、纵、垂三个正交平面视图上以二维特征展开，这样不仅可以让 RODIN 模型使用高效的 2D 架构进行 3D 感知扩散，将三维图像降维成二维图像也大幅降低了计算复杂度和计算成本。

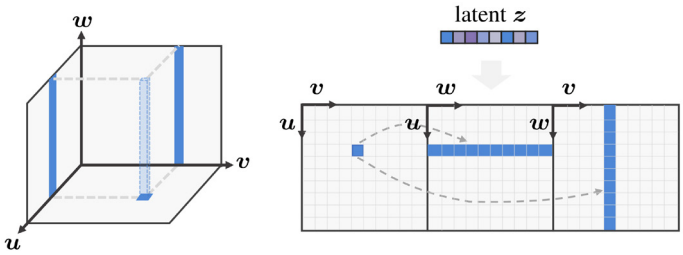


图 3：3D 感知卷积高效处理 3D 特征。（左图）用三平面（triplane）表达 3D 空间，此时底部特征平面的特征点对应于另外两个特征平面的两条线。（右图）引入 3D 感知卷积处理展开的 2D 特征平面，同时考虑到三个平面的三维固有对应关系。

要实现 3D 图像的生成需要三个关键要素：

3D 感知卷积，确保降维后的三个平面的内在关联。传统 2D 扩散中使用的 2D 卷积神经网络（CNN）并不能很好地处理 Triplane 特征图。而 3D 感知卷积并不是简单生成三个 2D 特征平面，而是在处理这样的 3D 表达时，考虑了其固有的三维特性，即三个视图平面中其中一个视图的 2D 特征本质上是 3D 空间中一条直线的投影，因此与其他两个平面中对应的直线投影特征存在关联性。为了实现跨平面通信，研究员们在卷积中考虑了这样的 3D 相关性，因此高效地用 2D 的方式合成 3D 细节。

隐空间协奏三平面 3D 表达生成。研究员们通过隐向量来协调特征生成，使其在整个三维空间中具有全局一致性，从而获得更高质量的化身并实现语义编辑，同时，还通过使用训练数据集中的图像训练额外的图像编码器，该编码器可提取语义隐向量作为扩散模型的条件输入。这样，整体的生成网络可视为自动编码器，用扩散模型作为解码隐空间向量。对于语义可编辑性，研究员们采用了一个冻结的 CLIP 图像编码器，与文本提示共享隐空间。

层级式合成，生成高保真立体细节。研究员们利用扩散模型先生成了一个低分辨率的三视图平面（64×64），再通过扩散上采样生成高分辨率的三平面（256×256）。这样，基础扩散模型集中于整体 3D 结构生成，而后续上采样模型专注于细节生成。

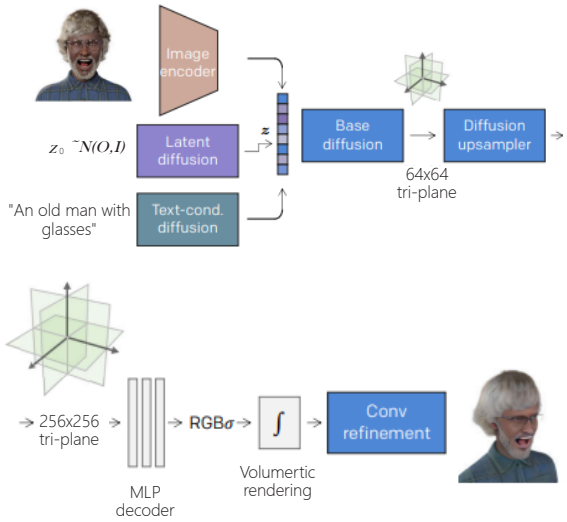


图 4：RODIN 模型概述

此外，在训练数据集方面，研究员们借助开源的三维渲染软件 Blender，通过随机组合画师手动创建的虚拟 3D 人物图像，再加上从大量头发、衣服、表情和配饰中随机采样，进而创建了 10 万个合成个体，同时为每个个体渲染出了 300 个分辨率为 256*256 的多视图图像。在文本到 3D 头像的生成上，研究员们采用了 LAION-400M 数据集的人像子集训练从输入模态到 3D 扩散模型隐空间的映射，最终让 RODIN 模型可以只使用一张 2D 图像或一句文字描述就能创建出逼真的 3D 头像。

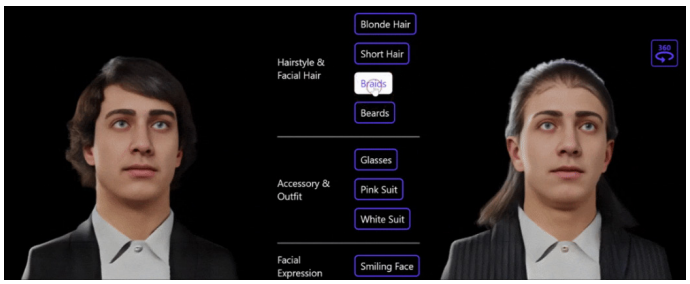


图 5：利用文字做 3D 肖像编辑

图 6：更多随机生成的虚拟形象

微软亚洲研究院主管研究员张博表示，“此前，3D 领域的研究受限于技术或高成本，生成的 3D 结果主要是点云、体素、网格等形式的粗糙几何体，而 RODIN 模型可创建出前所未有的 3D 细节，为 3D 内容生成研究打开了新的思路。我们希望 RODIN 模型在未来可以成为 3D 内容生成领域的基础模型，为后续的学术研究和产业应用创造更多可能。”

让 3D 内容生成更个性、更普适

现如今，虚拟人、数字化身在电影、游戏、元宇宙、线上会议、电商等行业和场景中的需求日益增多，但其制作流程却相当复杂专业，每个高质量的化身都必须由专业的 3D 画师精心创作，尤其是在建模头发和面部毛发时，甚至需要逐根绘制，其中的艰辛历程外人难以想象。微软亚洲研究院 RODIN 模型的快速生成能力，可以协助 3D 画师减轻数字化身创作的工作量，提升效率，促进 3D 内容产业的发展。

“目前，3D 真人化身的创建耗时耗力，很多项目背后可能都有一个上百人的团队在做支持，实现方法更多的是借助虚幻引擎、游戏引擎，再加上画师的专业绘画能力，才能设计出高度逼真的真人定制 3D 化身，普通大众很难使用这些服务，通常只能得到一些现成的、与本人毫无关连的化身。而 RODIN 模型低成本和可定制化的 3D 建模技术，兼具普适性和个性化，让 3D 内容生成走向大众成为可能。”微软亚洲研究院资深产品经理刘潇说。

尽管当前 RODIN 模型生成结果主要为半身的 3D 头像，但是其技术能力并不仅限于 3D 头像的生成。随着包括花草树木、建筑、汽车家居等更多类别和更大规模训练数据的学习，RODIN 模型将能生成更多样的 3D 图像。下一步，微软亚洲研究院的研究员们将用 RODIN 模型探索更多 3D 场景创建的可能，向一个模型生成 3D 万物的终极目标不断努力。

相关链接：

论文链接：
RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion
https://arxiv.org/abs/2212.06135

项目页面：
https://3d-avatar-diffusion.microsoft.com

扫描二维码了解更多信息

让天下没有难训练的大模型，微软亚洲研究院开源 TorchScale

近期，微软亚洲研究院从深度学习基础理论出发，研发并推出了 TorchScale 开源工具包。TorchScale 工具包通过采用 DeepNet、Magneto 和 X-MoE 等最先进的建模技术，可以帮助研究和开发人员提高建模的通用性和整体性能，确保训练模型的稳定性及效率，并允许以不同的模型大小扩展 Transformer 网络。

如今，在包括自然语言处理 (NLP)、计算机视觉 (CV)、语音、多模态模型和 AI for Science 等领域研究中，Transformer 已经成为一种通用网络结构，加速了 AI 模型的大一统。与此同时，越来越多的实践证明大模型不仅在广泛的任务中能产生更好的结果、拥有更强的泛化性，还可以提升模型的训练效率，甚至衍生出新的能力。因此，学术界和产业界都开始追求更大规模的模型。

然而随着模型的不断扩大，其训练过程也变得更加困难，比如会出现训练不收敛等问题。这就需要大量的手动调参工作来解决，而这不仅会造成资源浪费，还会产生不可预估的计算成本。

与其扬汤止沸，不如釜底抽薪。微软亚洲研究院从深度学习基础理论出发，创新推出了 TorchScale 工具包，并已将其开源。TorchScale 是一个 PyTorch 库，允许科研和开发人员更高效地训练 Transformer 大模型。同时，它有效地提升了建模的性能和通用性，提高了 Transformer 的稳定性和训练效率。

“我们希望通过 TorchScale 的系列工作从更底层出发做一些基础性的研究创新，通过数学或者理论上的指导和启发，在 Transformer 模型扩展的工作中取得更好的效果，而不是单纯的调参或仅从工程层面去部分缓解某些问题。TorchScale 能够支持任

意的网络深度和宽度，实验验证它可以轻松扩大模型规模，而且只需要几行代码就能够实现多模态模型的训练。”微软亚洲研究院自然语言计算组首席研究员韦福如表示。

TorchScale 主要从以下三个方面帮助科研人员克服了扩展 Transformer 大模型时的困难：

DeepNet——提升模型的稳定性；Magneto——提升模型的通用性；X-MoE——提升模型训练的高效性。

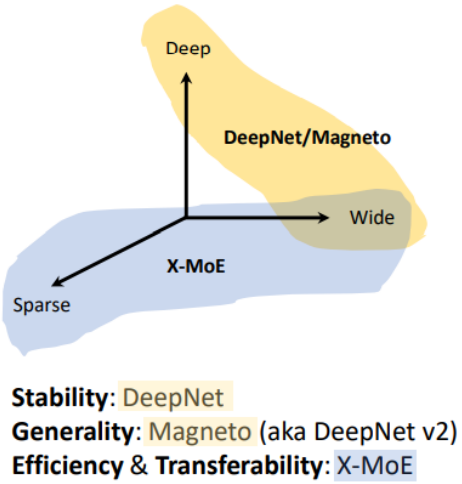


图 1：TorchScale 解决了大模型在稳定性、通用性、高效性方面的问题

DeepNet：让 Transformer 训练深度超过 1000 层

尽管近年来模型参数的数量越来越大，已经从百万级扩展到万亿级，但参数的深度却一直受限于 Transformer 训练的不稳定性。为了解决这一问题，一些科研人员尝试通过更好的初始化或架构来提升 Transformer 的稳定性，但这也只能让 Transformer 在百层级别的深度下保持稳定。

研究员们发现，模型输出的剧烈变化是导致模型不稳定的重要原因。为此，研究员们在残差连接处使用了一种新的归一化函数——DeepNorm。新的函数由理论推导而来，能把模型输出的变化限制在常数范围内。此方法只需要改变几行代码，就能大幅提升 Transformer 的稳定性。通过引入新的 DeepNorm 函数，研究员们训练了超深的 Transformer 网络 DeepNet，在保证模型稳定的同时，可以将模型深度扩展到 1000 层以上。

DeepNorm 同时具备 Post-LN 的性能优势和 Pre-LN 的稳定训练优势。这个新方法或将成为 Transformer 的首选替代方案，它不仅适用于深模型，更适用于大模型。值得一提的是，与具有 120 亿参数的 48 层模型相比，微软亚洲研究院 32 亿参数的 200 层模型在 100 多个语言、超 10000 个语言对和 130 亿个文本对的多语言机器翻译实验中实现了 5 BLEU 的提升。在大规模多语言翻译任务上，随着 DeepNet 模型深度从 10 层扩展至 100 层和 1000 层，模型也获得了更高的 BLEU 值。

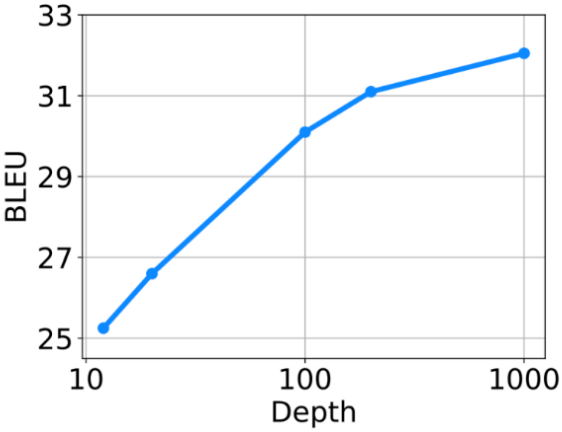


图 2：随着模型深度从 10 层扩展至 100 层和 1000 层，DeepNet 有效提升了多语言翻译结果

韦福如说，“此前，科研人员在训练更大规模的模型时，往往需要投入大量的精力在模型调参上，无形中增加了实验成本，有的模型在训练中途就无法继续下去了，即使给模型打上补丁也还是会影响模型性能。DeepNet 可以帮助科研人员大幅降低调参的负担，在提升模型性能的同时降低实验成本。”

Magneto：真正实现多模态模型架构统一

跨语言、视觉、语音和多模态的模型在模型结构上走向大一统的趋势如今已经愈发明显。具体而言，从 NLP 领域开始，Transformer 已成为 AI 各领域的主流结构。然而，尽管都使用了 Transformer，但不同模态任务的模型结构在具体实现时仍存在显著差异。例如，GPT 和 ViT 模型采用了 Pre-LN Transformer，而 BERT 和机器翻译模型使用的是 Post-LN 来获得更好的性能。更重要的是，对于多模态模型，不同输入模态的最优 Transformer 变体通常是不同的。以微软亚洲研究院推出的多模态预训练模型 BEiT-3 为例，其使用 Post-LN 对于视觉部分是次优的，而 Pre-LN 对于语言部分是次优的。

要想让多模态预训练真正实现大一统就需要一个统一的架构，该架构需要在不同任务和模态上都能有良好的性能表现。另外，如之前所述，Transformer 架构训练的稳定性也是一个痛点。微软亚洲研究院的研究员们意识到，通用模型的开发需要更基础的 Transformer，即 Foundation Transformer。首先它的建模能够作为各种任务和模式的统一架构，这样就可以使用相同的主干而无需反复魔改。其通用的设计原则也应该支持多模态基础模型的开发，在不牺牲性能的前提下将统一的 Transformer 用于各种模态。其次，它的网络结构应能够保障训练的稳定性，从而降低基础模型大规模预训练的难度。

为了实现这些目标，微软亚洲研究院的研究员们提出了一个 Foundation Transformer——Magneto。在 Magneto 中，研究员们引入了 Sub-LN，为每个子层（即多头自注意力和前馈网络）添加了额外的 LayerNorm，并且提出了一种新的初始化方法，为从根本上提高训练的稳定性提供了理论保证。

通过对 Magneto 在不同任务和模态上的评测，包括掩码语言建模（即 BERT）、因果语言建模（例如 GPT）、机器翻译、掩码图像建模（即 BEiT）、语音识别和视觉语言预训练（即 BEiT-3），结果显示在下游任务上，Magneto 显著优于各种 Transformer 变体。此外，得益于训练稳定性的提高，Magneto 还允许使用更高的学习率来进一步提高结果。

Language Pretraining

	GLUE
BERT/PostLN	85.7
Magneto	**86.3**

Average GLUE score

Speech Recognition

	Dev-Clean	Dev-Other	Test-Clean	Test-Other
PreLN	2.97	6.52	3.19	6.62
Magneto	**2.68**	**6.04**	**2.99**	**6.16**

Results on LibriSpeech benchmark. Lower is better.

Vision/BEiT Pretraining

	Base	Large
BEiT (ViT)	84.5	86.2
BEiT (Magneto)	**84.9**	**86.8**

ImageNet Top 1 Accuracy

Multimodal/BEiT-3

	VQA	NLVR2
BEiT-3 (ViT)	78.37/78.50	82.57/83.69
BEiT-3 (Magneto)	**79.00/79.01**	**83.35/84.23**

Results on Vision-language benchmarks

图 3：Magneto 在语言、图像、语音和多模态任务上的实验结果

X-MoE: 优于基线 SMoE 模型，助力模型高效训练

在有关大模型训练的研究中，除了将网络深度做得更深和将宽度即隐藏维度扩大以外，还可以利用混合专家系统（Mixture of Experts, MoE）。尽管 MoE 可以在诸如语言模型和视觉表示学习等广泛问题上获得更好的性能，但也会导致更高的计算成本，这促使越来越多的科研人员开始探索稀疏混合专家模型（Sparse Mixture-of-Experts, SMoE）。SMoE 主要通过构建稀疏激活的神经网络来增加模型容量。在不显著增加计算开销的情况下 SMoE 模型在各种任务（包括机器翻译、图像分类和语音识别）上的性能都优于稠密模型。

在 SMoE 模型中，路由机制发挥着重要的作用。给定输入 token，路由机制会测量每个 token 与专家之间的相似度分数，然后再根据路由得分将 token 分配给最匹配的专家。因此，近年来许多研究都集中在如何设计 token 专家分配算法上。然而，微软亚洲研究院的研究员们发现，当前的路由机制倾向于以专家为中心来推动隐藏表示聚类，这容易引起表征坍塌（Representation Collapse），损害模型性能。

为了缓解现有的路由机制引起的表征坍塌问题，微软亚洲研究院的研究员们提出了新的方法 X-MoE，为 SMoE 模型引入了一种简单而有效的路由算法。具体来说，区别于现有 SMoE 模型直接使用隐藏向量进行路由，X-MoE 先将隐藏向量投射到低维空间中，再对 token 表示和专家表示进行 L2 归一化，来测量低维超球面上的路由分数。此外，研究员们还提出了软专家门（soft expert gate），以学习控制专家的激活。

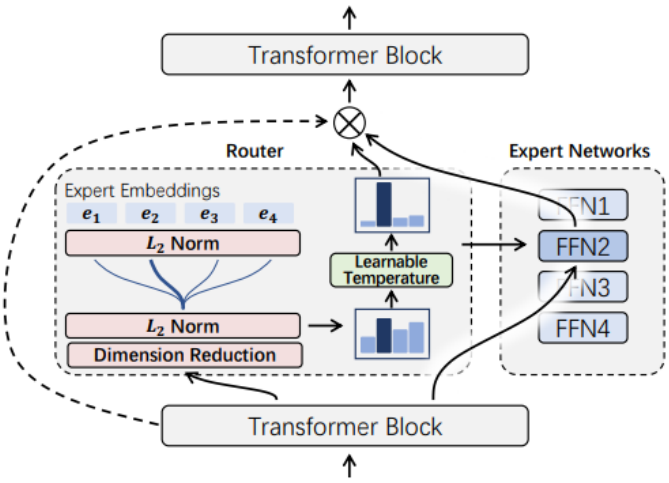


图 4：X-MoE 流程图

微软亚洲研究院的研究员们对这一新方法在跨语言模型预训练任务上进行了评测。实验结果表明，在语言建模和微调性能方面，基于 X-MoE 的模型始终优于基线 SMoE 模型。实验分析还表明，与 SMoE 基线相比，X-MoE 方法有效缓解了表征坍塌问题。该方法在预训练和微调期间也实现了更一致的路由行为，证实了 X-MoE 路由算法的有效性。

“随着技术的持续演进，大模型的训练不仅仅是工程层面的工作。我们应该从基础研究的角度出发，探索下一代 Transformer 网络架构。与此同时，在 AI 模型大一统趋势的推动下，我们更应该追求同一结构来支持不同模态的输入，并在不同语言和模态的任务上获得良好的性能。通过理论指导让模型变得更大、更稳定、更通用。”韦福如说。

相关链接：

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

高精度压缩 Transformer，NNI 剪枝一站式指南

无论在学术界还是产业界，今年人工智能大模型都是爆款话题。但面对这些动不动就数十亿级别参数的模型，使用传统方法微调，宛如水中捞月、海底捞针。作为微软亚洲研究院为科研人员和算法工程师量身定制的一站式 AutoML（自动机器学习）工具，NNI（Neural Network Intelligence）在过去的三年间不断迭代更新，加强了对各种分布式训练环境的支持，成为了最热门的 AutoML 开源项目之一。

微软亚洲研究院对 NNI 进行了更新。在最新的版本中，NNI 集成了大量前沿的剪枝算法，如 TaylorFO Weight、Movement 等。基于现有的经典预训练模型，研究员们通过大量实验，发现了既能降低模型参数量和计算量，又能保持模型较高精度的剪枝步骤与算法组合，获得超越 SOTA 的模型剪枝效果。

本文以 Transformer 系列的预训练模型和数据集 GLUE-MNLI 为例，介绍 NNI 的 pruner 剪枝流程和使用的剪枝算法组合。

剪枝流程

在正式介绍剪枝流程前，我们需要先了解什么是 pruner，mask 和 SpeedUp。

pruner：使用具体的剪枝算法实例化的剪枝器。

mask：在剪枝过程中，pruner 会生成一个和目标子模块大小相同的 mask（全 1）矩阵，并在 mask 矩阵中将目标子模块中需要剪掉的部分的对应位置置为 0。最后通过将目标子模块和对应的 mask 矩阵相乘，即可得到模拟剪枝后的模型效果。

SpeedUp：从上述描述可以看出，在剪枝过程中，实际上只是将需要剪枝的部分用 0 进行了替换，因此使用 SpeedUp 模块是修剪上述目标子模块中需要剪掉的参数，而不是用 0 替代，从而实现真正意义上的减少参数量。

在使用 NNI Compression 模块中的 pruner 进行剪枝操作时，用户只需完成数据 / 模型等的准备、pruner 的构建，以及模型剪枝和再训练，即可为模型构建一个剪枝的 pipeline。

以 Transformer 系列的预训练模型为例，其剪枝流程共包含 4 步：首先准备数据 / 模型等，接着针对多头自注意力机制（Multi-head Attention）、嵌入层（embedding）和前馈神经网络（FFN）分别剪枝和再训练模型。

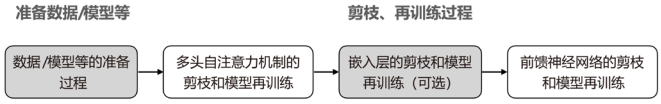


图 1：Transformer 系列模型的剪枝流程示意图

1. 准备数据 / 模型等

在正式构建剪枝过程之前，用户需要加载预训练模型，对数据预处理并创建相应的 dataloader，同时设计相应的训练 / 评估函数，以用于后期对模型的训练和评估。其流程如图 2 所示：

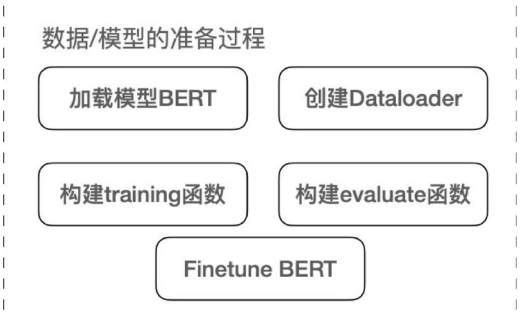


图 2：数据 / 模型准备过程的流程示意图

具体来说，首先需要从 Transformers 库中加载预训练模型，然后对数据 GLUE-MNLI 进行处理，并得到相应的 dataloader。随后，针对模型和数据集 GLUE-MNLI，构建相应的训练 / 评估函数。最后将模型在 GLUE-MNLI 数据集上进行微调。

完成以上步骤就相当于完成了数据 / 模型等的准备工作，可以得到预训练模型在 MNLI 数据集上微调后的模型。考虑到 Transformer 系列预训练模型的模型参数中的大头为嵌入层，且编码层 / 解码层中包含了多头自注意力机制和前馈神经网络。因此，在之后的步骤中需要分别对多头自注意力机制、嵌入层和前馈神经网络剪枝，并引入动态蒸馏机制对剪枝后的模型再训练。

2. 多头自注意力机制的剪枝和基于动态蒸馏机制的模型再训练

多头自注意力模块的剪枝和模型再训练分为 3 步，如图 3 所示：首先要构建 pruner，接着对多头自注意力模块进行剪枝，最后使用动态蒸馏机制再训练模型。

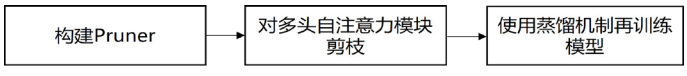


图 3：多头自注意力机制的剪枝和再训练流程示意图

在进行剪枝前，用户需要选定一个剪枝算法并实例化相应的 pruner。所有的剪枝算法均需向模型中传入 config_list 参数，因为其定义了需要剪枝的运算名、运算类别及稀疏度等。具体到 Movement 剪枝算法，还需要设置其他的一些参数，如：evaluator 参数，用于训练感知的模型压缩过程；movement_mode 参数，共有 "soft" 和 "hard" 两种模式，若为 "soft"，则难以精确地控制模型剪枝后的稀疏度，但是可以得到性能更好的模型。参数 regular_scale 用于控制剪枝的稀疏度，regular_scale 越大，模型剪枝后的稀疏度越高。更多其他参数可参阅 https://nni.readthedocs.io/zh/stable/reference/compression/pruner.html#movement-pruner。

接下来，要使用构造的剪枝算法实例 pruner 对多头自注意力模块进行剪枝。用户只需调用 pruner.compress() 即可执行对模型的剪枝过程，并得到剪枝后的模型和 attention_mask。其中 attention_mask 给出了需要剪枝的子模块的参数剪枝范围，0 代表该位置被剪掉，1 代表该位置被保留。

NNI 的 SpeedUp 模块可以将被 mask 住的参数和计算从模型中删除，具体的删除逻辑如图 4 所示，以 Query Linear 层的 weight（记作 Q）为例，其维度为 [768,768]，那么 Q 的 weight 的 mask 矩阵维度也为 [768, 768]，将其记作 mask。首先将该 mask 矩阵的维度进行变换，第一维是多头数目 8，其余的则是第二维，将变换后的 mask 矩阵记作 reshaped mask 矩阵。接着，对 reshaped mask 矩阵在第二维度上求和，并判断求和后的值是否为 0，此时的 mask 矩阵维度变为 [8]，每个位置对应着一个多头。对于变换后的 mask 矩阵，若位置 i 的值为 0，则代表在 Q 中的第 i 个多头需要被剪掉。在图中，位置 0、3、7 的值均为 0，因此，在 Q 中的第 0、3、7 个多头需要被剪掉。最后，将 [0,3,7] 作为参数传入 prune_heads 函数中，对 Q 进行修剪。修剪后，Q 的维度为 [576,768]。对 SpeedUp 更加全面的介绍可以参考发表于 OSDI 2022 的论文 SparTA。在即将发布的 NNI 3.0 中 SpeedUp 会对更多模型提供更加完善的支持。

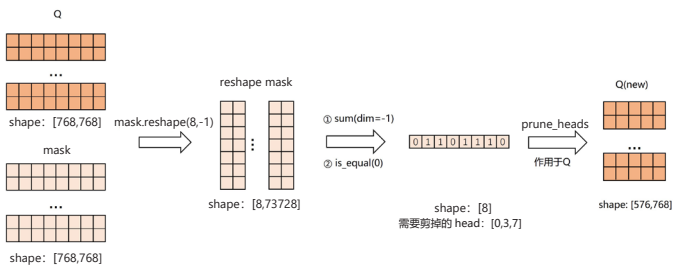


图 4：利用 prune_heads 函数修剪自注意力模块的过程示意图

在对多头自注意力模块剪枝后，以微调后的模型作为教师模型，以剪枝后的模型作为学生模型，然后借鉴 CoFi 中的动态蒸馏机制 [1] 对模型进行再训练，就可以得到新的模型。这里的动态蒸馏机制，是指教师模型的层和学生模型的层之间不是一个静态对应关系，每次蒸馏教师都可以选择从自身的高层动态蒸馏信息到学生模型低层中的一层里。

3. 嵌入层和前馈神经网络的剪枝，以及基于动态蒸馏机制的模型再训练

嵌入层和前馈神经网络的剪枝过程与多头自注意力模块的剪枝过程类似。此处使用 Taylor 剪枝算法（https://nni.readthedocs.io/zh/stable/reference/compression/pruner.html#taylor-fo-weight-pruner）对嵌入层和前馈神经网络进行剪枝。同样地，研究员们定义了 config_list、evaluator 参数及 taylor_pruner_steps 参数。由于嵌入层的维度与后续模型中的维度具有相关性。因此，基于上述参数，在嵌入层的剪枝过程中研究员们将剪枝模式 mode 设置为了 "dependency-aware" 模式，并传入模型的输入 dummy_input，以帮助 pruner 捕捉和嵌入层维度具有依赖关系的子模型。

接下来，使用分别构造的 pruner 对前馈神经网络和嵌入层进行剪枝。和多头自注意力模块的剪枝不同的是，此处使用了迭代式剪枝法，即在模型基于动态蒸馏的再训练过程中，每 2000 步分别使用 pruner 对前馈神经网络和嵌入层剪枝一次，其中，前馈神经网络共剪枝 19/24 次，嵌入层共剪枝 3 次。每次剪枝后，使用 ModelSpeedUp 对前馈神经网络层进行剪枝，以实现真正意义上的修剪参数，而不是将需要修剪的参数用 0 替换。

实验结果

通过调整 regular_scale 参数的值和前馈神经网络的剪枝次数，研究员们得到了具有不同稀疏度和性能的模型。该过程使用了 1 张 A100 进行实验，并设置 batch_size 为 32。

Attention Pruning Method	Embedding Pruning Method	FFN Pruning Method	Total Sparsity(%)	Accuracy(%)
BERT-base				85.07
Movement Pruner (sparsity=0.1, regular_scale=3)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝19次	57.83	84.6
Movement Pruner (sparsity=0.1, regular_scale=5)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝19次	63.2	84.18
Movement Pruner (sparsity=0.1, regular_scale=5)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝24次	66.77	84
Movement Pruner (sparsity=0.1, regular_scale=10)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝19次	69.66	83.44
Movement Pruner (sparsity=0.1, regular_scale=15)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝19次	72.53	83.12
Movement Pruner (sparsity=0.1, regular_scale=10)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝24次	72.83	83.04
Movement Pruner (sparsity=0.1, regular_scale=20)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝19次	74.33	82.82
Movement Pruner (sparsity=0.1, regular_scale=15)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝24次	75.2	82.44
Movement Pruner (sparsity=0.1, regular_scale=20)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝24次	76.78	82.06
Movement Pruner (sparsity=0.1, regular_scale=30)	Taylor Fo Weight Pruner 剪枝3次	Taylor Fo Weight Pruner 剪枝24次	78.89	80.39

图 5：实验结果

从上图实验结果可以看出：

a. 随着 regular_scale 的增加，模型总的稀疏度有所增加。当 regular_scale 大于等于 10 时，模型总的稀疏度超过了 69%，性能损失超过 1%。

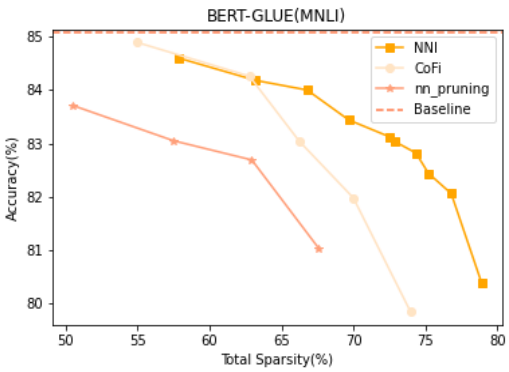
b. 随着前馈神经网络剪枝次数的增加，模型总的稀疏度有所增加，同时模型的性能有所下降，且随着模型总稀疏度的增加，模型的性能下降程度逐渐增大。

c. 对嵌入层剪枝 3 次，能够将模型的维度从 768 减小至 561，在一定程度上提升了模型总的稀疏度。

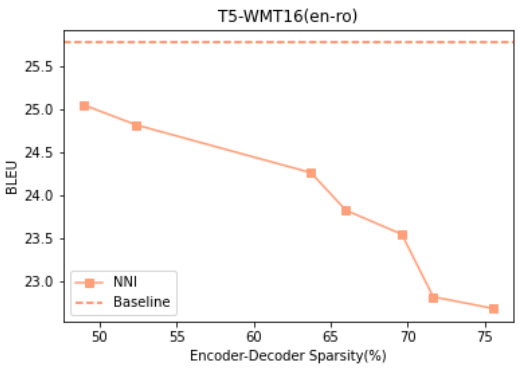
实验结果与平台对比

进一步分析实验结果可以发现，使用 NNI 对 BERT 在 MNLI 数据集上剪枝后的性能好于 nn pruning 框架（图 6(a)），且当模型总的稀疏度低于 65% 时，NNI 和 CoFi 对 BERT 在 MNLI 数据集上剪枝的性能差距较小，当模型总的稀疏度大于 65% 时，使用 NNI 对 BERT 在 MNLI 数据集上剪枝后的性能好于 CoFi。图 6(b) 和图 6(c) 分别展示了 NNI 在 T5 和 ViT 模型上的剪枝性能。如图所示，当模型相应部分的稀疏度超过 75% 后，模型性能下降约为 3%，当模型相应部分的稀疏度低于 50% 时，模型性能下降较少。

(a)



(b)



(c)

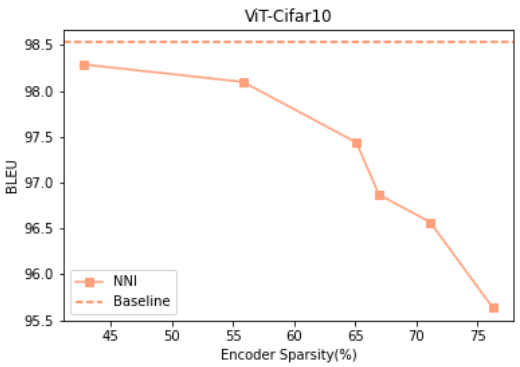


图 6：NNI 在经典预训练模型下的剪枝性能示意图

三个平台（Paper）的详细比较结果，如表 1 所示。可以看出，NNI 的 Compression 模块不仅具有完整的教程实例，同时还提供了 SpeedUp 模块，能够实现真正意义上的减少模型参数量，而非将需要修剪的参数置为 0。

同时，NNI 支持 BERT、RoBerta、GPT、BART、T5、ViT 等主流模型，并提供了 Taylor、Movement、ADMM、Slim、AGP、Activation APoZ、Activation Mean 等 16 种前沿剪枝算法，能够更好地满足用户的需求，具有较强的通用性。

	实验表现	SpeedUp	教程实例	支持模型	算法种类	备注
nn_pruning	√			BERT/BART/T5	Movement	
CoFi	√		√	BERT/RoBerta	CoFi	Paper 形式，不是一个针对剪枝的工具平台
NNI	√	√	√	BERT，RoBerta，GPT，BART，T5，ViT 等主流模型	Taylor，Movement，ADMM，Slim，AGP，Activation APoZ，Activation Mean 等 16 种前沿算法	通用的 AutoML 平台

表 1：各平台（Paper）功能对比总结

展望未来

在 NNI 3.0 版本中，微软亚洲研究院的研究员们还将引入蒸馏模块，更好地为用户提供集剪枝、蒸馏为一体的压缩工具，同时 SpeedUp 模块也将更全面地支持对 Transformer 的修剪。

相关链接：

论文链接：
SparTA: Deep-Learning Model Sparsity via Tensor-with Sparsity-Attribute
https://www.usenix.org/conference/osdi22/presentation/ zheng-ningxin

最新版 NNI 的完整代码和 tutorial：
https://nni.readthedocs.io/zh/stable/tutorials/ pruning_bert_glue.html

参考文献：

[1] Structured Pruning Learns Compact and Accurate Models
https://arxiv.org/pdf/2204.00408.pdf

打破拓扑结构的天花板，深入探究属性图上表示学习的研究蓝海

作者：社会计算组

图是一种通用的数据表示形式，来自不同领域的数据均可以表示为图，例如文本数据可以看成是一维图，图像可以看成是二维图，分子和蛋白质等实体也可以天然地的用图表示。通俗而言：万物皆可图。

而图表示学习（Graph Representation Learning，GRL）能够将图中的节点或者整个图转化为低维可计算的向量，为机器学习模型处理图这种高维复杂的数据形式提供了合适的计算接口。根据粗略统计，近年来，每一年各大顶会上收录的图神经网络（Graph Neural Networks，GNNs）论文数目增长迅速，相关研究与应用也蓬勃发展。

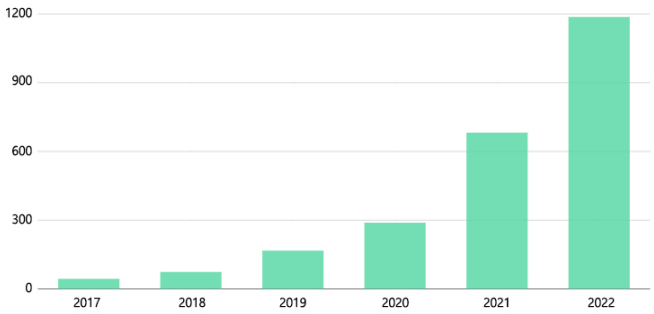


图 1：近年来各大会议上 GNN 相关的论文数目

但是，与自然语言处理（NLP）领域的语言模型 LM 和计算机视觉（CV）领域的 ViT (Vision Transformer) 相比，GNN 在实际应用中的影响仍相对较弱。对此，微软亚洲研究员社会计算组的研究员们深入探究了目前 GNN 的机制和原理，从根源上分析拓扑结构形成的原因，提出针对节点属性的理解建模有可能比拓扑结构建模更加重要，进而形成了一系列（文本）属性图上的表示学习的工作。

在本篇文章中，研究员们将从新的视角定义和利用拓扑结构，揭示语言模型与拓扑结构之间的关联。

基于拓扑结构建模的图神经网络

目前常见的 GNN 模型一般着眼于挖掘拓扑结构中独特的知识。GNN 模型通常会先将节点表示为一个特征向量，然后设计复杂而精细的聚合函数来捕捉中心节点周围的拓扑结构信息。以知识图谱这个常见的图数据为例，微软亚洲研究院发表在 ICML 2022 的《HousE: Knowledge Graph Embedding with Householder Parameterization》工作中提出了基于拓扑结构、具有强大综合建模能力的知识图谱表示学习（Knowledge Graph Embedding, KGE）模型 [1]。

KGE 模型旨在学习知识图谱中实体和关系的表示。这些模型的性能好坏很大程度取决于对 KG 中关系模式（relation pattern）和关系映射属性（relation mapping property）建模的能力。知识图谱中重要的关系模式有：(1) 对称，如 is_friend_of 就是一种对称关系；(2) 非对称，如 is_father_of 就是一种非对称关系；(3) 逆，如 is_teacher_of 和 is_student_of 就是一对互逆的关系；(4) 组合，如 is_grandmother_of 就是 is_father_of 和 is_mother_of 的组合关系。关系的映射属性则有一对一关系、一对多关系、多对一关系和多对多关系。

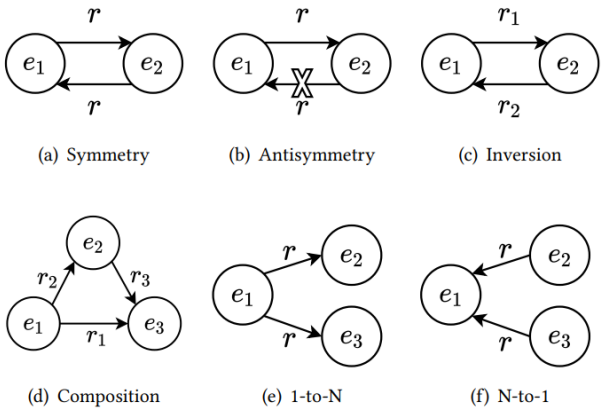


图 2：四种重要关系模式和两种复杂的关系映射属性

表 1 总结了现在有代表性的 KGE 方法的建模能力，其中还包含对这些方法局限性的分析：(1) 虽然将关系视为实体间的旋转变换是对多种关系模式建模的有效方法，然而关系的旋转无法突破低维空间（2、3、4 维）很大程度地限制了模型的建模能力；(2) 还未出现能完美演绎知识图谱中的重要关系模式与复杂映射属性的"六边形战士"模型。

Model	Symmetry	Antisymmetry	Inversion	Composition	Mapping Properties	Dimension of Rotation
TransE	---	✓	✓	✓	---	---
TransX	✓	✓	---	---	✓	---
DistMult	✓	---	---	---	✓	---
ComplEx	✓	✓	✓	---	✓	---
RotatE	✓	✓	✓	✓	---	2
Rotate3D	✓	✓	✓	✓	---	3
QuatE	✓	✓	✓	---	✓	4
HousE	✓	✓	✓	✓	✓	k

表 1：现有典型的 KGE 模型对重要关系模式和复杂映射属性的建模能力

现有方法的局限性促使研究员们思考：如何设计一个具有建模能力更全面的 KGE 模型？为此，本研究引入了 Householder 反

射变换作为基本数学工具，并基于此设计了两种线性变换作为知识图谱中的关系表示。在 Householder 框架下，研究员们提出了 HousE。HousE 是现有基于旋转的 KGE 模型的升级版。如表 2 所示，HousE 能够自然地将旋转扩展到任意 k 维，并且其建模能力能够覆盖表 1 中的所有关系模式与映射属性。在多个公开知识图谱基准数据集进行验证的实验表明，HousE 明显优于现有的基准模型。

	WN18RR					FB15k-237					YAGO3-10				
Model	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10	MR	MRR	H@1	H@3	H@10
TransE†	3384	.226	-	-	.501	357	.294	-	-	.465	-	-	-	-	-
DistMult⋄	5110	.43	.39	.44	.49	254	.241	.155	.263	.419	5926	.34	.24	.38	.54
ComplEx⋄	5261	.44	.41	.46	.51	339	.247	.158	.275	.428	6351	.36	.26	.4	.55
ConvE⋄	4187	.43	.40	.44	.52	224	.325	.237	.356	.501	1671	.44	.35	.49	.62
RotatE	3340	.476	.428	.492	.571	177	.338	.241	.375	.533	1767	.495	.402	.55	.67
Rotate3D	3328	.489	.442	.505	.579	165	.347	.250	.385	.543	-	-	-	-	-
QuatE	3472	.481	.436	.500	.564	176	.311	.221	.342	.495	-	-	-	-	-
DualE	-	.482	.440	.500	.561	-	.330	.237	.363	.518	-	-	-	-	-
Rot-Pro	2815	.457	.397	.482	.577	201	.344	.246	.383	.540	1797	.542	.443	.596	.669
HousE-r	1885	.496	.452	.511	.585	165	.348	.254	.384	.534	1449	.565	.487	.616	.703
HousE	**1303**	**.511**	**.465**	**.528**	**.602**	**153**	**.361**	**.266**	**.399**	**.551**	**1415**	**.571**	**.491**	**.620**	**.714**

表 2：不同 KGE 模型在标准数据集上的效果

基于拓扑结构的建模真的没问题么？

随着图神经网络的快速发展，GNN 被普遍应用于 NLP、推荐系统、自然科学等领域。例如，在推荐系统中，每个用户和商品都可以看成是图中的两类节点，两者之间的历史交互行为（例如点击、购买等）构成了图上的边。基于图的推荐系统有助于捕捉高阶的协同过滤信号（Collaborative Filtering, CF），因此可以更好的刻画用户的兴趣偏好。

考虑到“万物皆可图”的普适性，难道基于图拓扑结构的建模真的无懈可击么？

其实不然。基于拓扑结构的建模面临着以下几个问题：

（1）不可知性。图拓扑结构的复杂性使得人很难对其进行深入且准确的认知。判断两个序列是否相似，例如两段文本是否相关，并不难。但是判断两个图是否相似或者同构（例如 Graph Matching），则是 NP-hard 问题。例如，图 3 中的两张图看起来区别很大，但实际上是等价的。如果准确理解判断图拓扑结构对人来说都算难题，那么很难期待机器学习模型能够在这种不可知的数据上达到更好的效果。

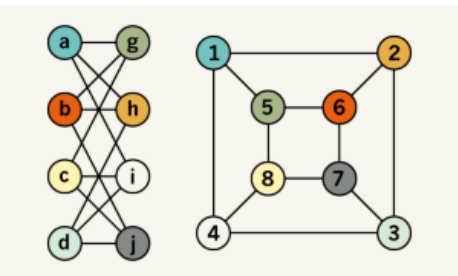


图 3：图匹配问题的一个例子 [2]

（2）低计算效率和高资源消耗。图神经网络的一大特征是可以利用高阶的远距离邻居来提供丰富的上下文信息。但同时，这种远距离的依赖带来了“邻居爆炸”（neighborhood explosion）的挑战。假设节点的平均度数为 k，那么 l 跳之内的节点数目为 N=k^1+k^2+...+k^l。GNN 的聚合过程需要 N 个节点的信息来学习单个节点的表示，一方面这需要更大的内存来存储这种指数规模的邻居节点，另一方面也会极大影响模型的训练速度。

（3）拓扑结构的可靠性。例如在社交网络中，由于用户行为的随机性和不确定性，某个社交用户可能错误地关注或点击一些不相关的其他用户，带来了拓扑结构中的噪声。这类噪声会随着逐条的信息聚合而逐渐增强，使学习到的用户表示不够准确。

根据以上分析，基于拓扑结构的建模可能并不是最优解，而且 GNN 在如短序列推荐（Session-based recommendation, SBR）等场景中也并不一定有效。例如，短序列推荐（SBR）是针对用户在短期、动态的会话（即用户一段时间内的活动序列）中的行为进行推荐。与传统的基于用户或物品的推荐系统不同，SBR 更注重当前会话中用户的实时需求，从而更好地应对用户兴趣的快速变化和长尾商品的推荐问题。最近的 SBR 研究（SR-GNN[4]，SGNN-HN[5] 和 DHCN[6] 等）也出现了大量使用基于 GNN 的模型。但与模型复杂度的指数级增长相比，这些模型在基准测试中带来的性能提升却微乎其微。鉴于这种现象，研究员们的疑问是：基于 GNN 的模型对于 SBR 来说，是过于简单，还是过于复杂了？

为了回答该问题，研究员们尝试剖析现有的基于 GNN 的 SBR 模型，并分析其在 SBR 任务上的作用。如图 4，典型的基于 GNN 的 SBR 模型可以分解为两个部分：GNN 模块和 Readout 模块。分别在这两个部分上应用 Sparse Variational Dropout（SparseVD）在训练模型时计算参数的密度比（density ratio）。

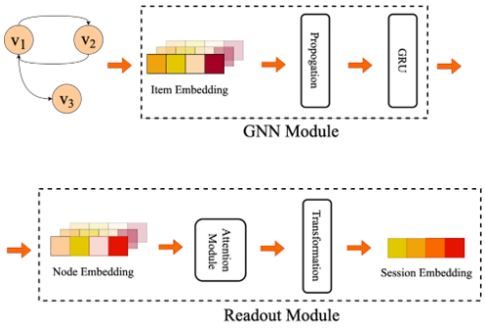


图 4：基于 GNN 的 SBR 模型的建模范式

可以从图 5 左图中 GNN 模块的密度比看出，随着训练趋于稳定，该密度比趋于 0。而右图所表示的 Readout 模块显示随着训练的进行，注意力池化权重的密度比可以保持在一个较高水平。在其他数据集和其他 GNN-based SBR 模型上，也出现了相同的趋势。因此得出结论：GNN 模块的参数很有可能是冗余的。

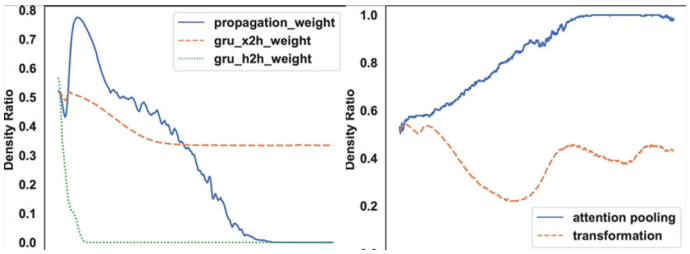


图 5：SBR 模型中不同模块的参数重要度分析

基于上述发现，研究员们提出了以下用于 SBR 的更简单但更有效的模型设计准则：

（1）不强调复杂的 GNN 设计，更倾向于删除 GNN 传播部分，仅保留初始嵌入层；

（2）SBR 模型应该更加关注基于注意力的 Readout 模块。由于注意力池化权重参数保留了较高密度比，研究员们推测，在基于注意力的 Readout 方法上进行先进的架构设计将会更有效。由于放弃了对 GNN 传播部分的要求，Readout 模块应该承担更多模型推理上的责任。考虑到现有基于实例视图（instance-view）的 Readout 模块的推理能力不足，所以需要设计具有更强大推理能力的 Readout 模块。

因此，研究员们提出了一个名为 Atten-Mixer 的模型，相关论文发表于 WSDM 2023，并获得 Best Paper Runner-up 奖项 [3]。如表 4 所示，即使移除了 GNN 模块，Atten-Mixer 在多个常用的数据集上的实验结果仍优于复杂的基于 GNN 的模型。

Model	*Diginetica*			*Gowalla*			*Last.fm*		
	HR@20	MRR@20	Time (s)	HR@20	MRR@20	Time (s)	HR@20	MRR@20	Time (s)
NextItNet	35.60	9.66	91.06	38.69	16.48	67.94	21.02	6.46	413.27
NARM	48.27	16.43	107.61	49.67	22.14	80.52	21.73	6.87	427.14
SR-GNN	51.16	17.67	341.68	50.16	24.58	338.62	22.49	8.30	1626.94
GC-SAN	50.63	17.37	437.27	50.35	24.65	398.04	22.63	8.40	1814.78
SGNN-HN	51.57	17.54	365.38	50.72	24.97	326.91	23.66	8.34	1595.59
LESSR	51.71	18.15	440.84	51.34	25.49	511.68	23.37	8.84	1927.20
NISER+	54.18	18.36	292.15	53.89	25.73	278.65	23.82	8.36	279.80
DHCN	53.85	18.50	2169.87	53.77	24.13	2452.76	22.86	7.78	21059.94
DSAN	54.02	18.62	273.48	54.09	26.64	279.17	24.17	8.42	1203.81
Atten-Mixer	**55.66**	**18.96**	288.12	**55.12**	**27.01**	267.37	**24.50**	**9.05**	1140.09

表 4：不同的 SBR 模型在常见数据集上的效果

节点属性建模可能更加重要

节点和边是构成图的关键元素。其中节点代表了其内在的天然属性，边刻画了图的拓扑结构。目前常用的 GNN 更多关注于对拓扑结构的建模，忽视了对节点属性的理解。例如，GCN/ GAT/ GraphSAGE 等通用的 GNN 模型通常会将节点利用先验知识或者预先训练好的编码器转变为节点表示向量。这些初始的节点表示向量在后续的 GNN 聚合过程被视为静态不可学习的参数，这种学习范式中的节点属性建模和拓扑结构建模是相互独立的，并不能进行端到端的图表示学习。另外，节点属性建模一般是基于先验知识或者预先得到的编码器，无法保证其得到的节点特征能够和下游的图挖掘任务保持一致，因此微软亚洲研究院的研究员们从节点属性建模出发，开辟了新途径。

不同于以往的二步式学习范式，研究员们尝试对节点属性和拓扑结构进行协同训练（co-training）。从图衍变的角度来看，节点属性有可能比拓扑结构更加重要。拓扑结构中的每一条边都有其存在的原因，而节点之间的关联关系是边形成的根本原因和内在动力。例如在图 6 中，左图表示了一个已经存在的社交网络和一个新用户 Alex，每个用户都有社交属性和偏好，Alex 在加入社交网络之后，大概率会关注和自己属性类似或者兴趣爱好相同的其他用户（例如同事或者同学）等，进而形成了社交网络上的拓扑结构。

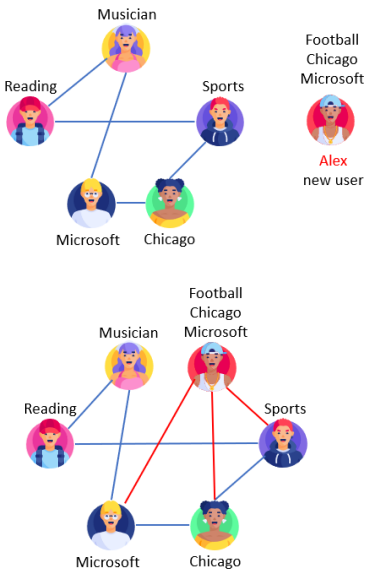


图 6：社交网络中关注关系形成的一个示例

大家可以简单的理解为：拓扑结构可能会受到其他外界因素（例如用户行为和偏好）等的影响有所偏差，但是始终受节点属性的影响和制约。拓扑结构和节点属性之间的关系有些类似于"先有鸡还是先有蛋"的困境（The Chicken-and-egg Conundrum）。从另外一个角度来看，如果没有拓扑结构的话，节点以及其自身的属性信息其实还是会存在，例如社交网络中某个用户没有关注任何人，他依然是社交网络中的一个实体。但是，如果没有了节点，拓扑结构就会消失，正所谓"皮之不存毛将焉附"。因此，研究员们认为节点属性建模更加重要。

那么问题又来了：如果节点属性建模是拓扑结构建模的基础的话，那么拓扑结构是不是就没用了呢？

这个答案是否定的。

首先，由于节点属性经常会不完整或者遗失，所以拓扑结构可以从节点之间的关系上对其进行补充。其次，拓扑结构是从用户行为等其他层次刻画节点之间的关联关系，这种关联关系会有自己独特的知识，并不一定和节点属性相似性完全一致，因此不如“全都要”，同时对节点属性和拓扑结构联合建模。

如何做到节点和边的联合表示学习？研究员们着眼于文本属性图（Text Attributed Graph, TAG）上的表示学习，以融合中心与邻域文本特征的图节点向量表示为破局之法。为了同时训练节点内的文本和节点间的拓扑结构，一个直接的方式是利用预训练语言模型（LM）对文本进行建模，然后利用 GNN 对 LM 学习得到的节点表示进行基于拓扑结构的融合。

如图 7 左图所示，针对中心节点 c 以及其邻居节点 N1 和 N2，分别利用多层 Transformer 架构对其包含的文本信息进行建模，然后利用 GNN 聚合生成的表示向量作为最终中心节点的表示。下游的图分析任务（例如节点分类）的优化目标得到的梯度会反传回来同时更新 LM 和 GNN，形成一个端到端的表示框架。相关工作已入选 SIGIR 2021[7]、EMNLP 2021[8] 等会议。同时，该系列工作也在赋能微软必应（Bing）搜索、必应（Bing）广告、Shopping feeds 等微软产品为其提质增效。

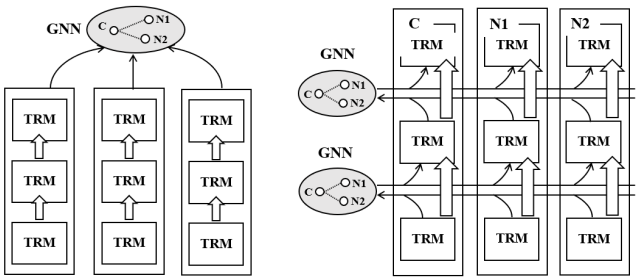


图 7：松耦合和紧耦合的 LM-GNN Co-training 范式

虽然模型能够实现联合建模，但拓扑结构和文本属性仅仅是通过级联方式进行了松耦合。对此，研究员们进而提出了紧耦合的 LM-GNN 共同训练范式（LM-GNN Co-training 模型）。如图 7 右图所示，紧耦合范式采取了层级化的 LM-GNN 整合方式：在每一层中，每个节点先由各自的 Transformer Block 进行独立的语义编码，编码结果汇总为该层的特征向量（默认由 CLS 所关联的 hidden state 来表征）；各节点的特征向量汇集到该层的 GNN 模块进行信息整合；信息整合的结果被编码至对应各个节点的图增广（graph augmented）特征向量中，并分发至各个节点；各节点依照图增广特征向量进行下一层级的编码。

相较于此前的松耦合架构，紧耦合在 LM 编码阶段便充分参照了邻域信息，从而大大提升了各节点文本表示的质量。同时，考虑到节点间的信息交互是借由特征向量在极其轻量 GNN 模块中进行，每层整体的运算开销与单纯利用 Transformer Block 进行各节点独立的编码相差无几。相关的研究工作已在 NeurIPS 2021[9]、KDD 2022[10] 等会议发表。

虽然与之前的 GNN 模型相比，LM-GNN Co-training 模型能够在文本属性图上取得优越的效果，但该训练方法大大提高了训练的难度。如图 7 所示，联合训练需要同时对中心节点和 k 个邻居节点的文本信息进行建模，相当于 (1+k) 次 LM 的前向计算和梯度的后向传递。考虑到预训练语言模型的参数量，这种计算复杂度和资源消耗难以承受。一种简单的方式是减少邻居节点的数目 k，但是会带来拓扑结构信息的丢失。因此，如何在保证效率的前提下进行 LM-GNN 的有效共同训练，是个亟需解决的问题。

因此，研究员们提出了一个交替训练的框架 [11]。如图 8 所示，在 LM 训练阶段，GNN 学习到的拓扑结构知识能够有效地转移到 LM 中，使其能够了解拓扑结构和下游任务的相关信息。而在 GNN 训练阶段，上一次训练得到的融合拓扑结构信息的 LM 能够为 GNN 提供更高质量的初始节点特征，有助提升其性能。上述两个阶段相互增强，进而可以得到高质量的节点表示向量。上述学习范式的时间复杂度约等于单个 GNN 和单个 LM 的训练复杂度之和，所以能够有效处理大规模节点属性图上的表示学习。该研究入选了 ICLR 2023（Notable-top-5%），并且在 OGBN leaderboard[12] 上登顶多个数据集的首位。

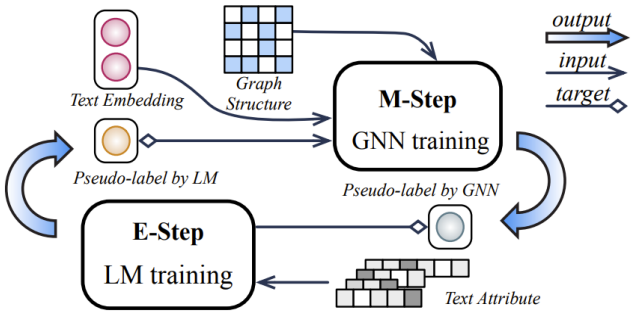


图 8：交替学习的 LM-GNN 训练范式

Leaderboard for ogbn-products

The classification accuracy on the test and validat

Package: >=1.1.1

Rank	Method	Ext. data	Test Accuracy	Validation Accuracy
1	GLEM+EnGCN	Yes	0.9014 ± 0.0012	0.9370 ± 0.0004
2	EnGCN	No	0.8798 ± 0.0004	0.9241 ± 0.0003
3	GLEM+GIANT+SAGN+SCR	Yes	0.8737 ± 0.0006	0.9400 ± 0.0003
4	GIANT-XRT+R-SAGN+SCR+C&S	Yes	0.8684 ± 0.0005	0.9365 ± 0.0003
5	GIANT-XRT+SAGN+SCR+C&S	Yes	0.8680 ± 0.0007	0.9357 ± 0.0004

Leaderboard for ogbn-arxiv

The classification accuracy on the test and valida

Package: >=1.1.1

Rank	Method	Ext. data	Test Accuracy	Validation Accuracy
1	GLEM+EnGCN	Yes	0.7966 ± 0.0006	0.8017 ± 0.0008
2	EnGCN	No	0.7798 ± 0.0007	0.7876 ± 0.0005
3	GLEM+RevGAT	Yes	0.7694 ± 0.0025	0.7746 ± 0.0018
4	GIANT-XRT+AGDN+BoT+self-KD	Yes	0.7637 ± 0.0011	0.7719 ± 0.0008
5	GIANT-XRT+R-RevGAT+KD	Yes	0.7635 ± 0.0006	0.7692 ± 0.0010

Leaderboard for ogbn-papers100M

The classification accuracy on the test and validat

Package: >=1.2.0

Rank	Method	Ext. data	Test Accuracy	Validation Accuracy
1	GLEM+GIANT+GAMLP	Yes	0.7037 ± 0.0002	0.7354 ± 0.0001
2	GIANT-XRT+GAMLP+RLU (use raw text)	Yes	0.6967 ± 0.0005	0.7305 ± 0.0004
3	GAMLP+RLU+SCR	No	0.6842 ± 0.0015	0.7188 ± 0.0007
4	SAGN+SLE (4 stages)	No	0.6830 ± 0.0008	0.7163 ± 0.0007
5	GAMLP+RLU	No	0.6825 ± 0.0011	0.7159 ± 0.0005

图 9：GLEM 在 OGBN leaderboard 上取得了多个数据集的第一名

大规模语言模型也许是图表示学习的未来

基于上述的一系列研究工作，微软亚洲研究院的研究员们证明了节点属性建模有可能比拓扑结构建模更重要。随着最近 ChatGPT 所带来的强大认知能力的涌现，一个问题也随之而来：大规模语言模型（LLM）能否成为处理图数据的基础模型？研究员们也对这个问题进行了初步的探究和展望。

这里举一个简单的例子。TUDataset 是图分类任务中经常使用的数据集，其中一个任务是给定一个化合物，根据结构组成判断其是芳香族还是杂芳族。不同于主流 GNN 的做法，如图 10 所示，研究员们利用 ChatGPT 作为基准模型，将化合物转化成 ChatGPT 能够理解的语言，然后询问该输入的图是否属于芳香族。令人惊讶的是，ChatGPT 给出了正确的判断结果和原因解释。

从表面来看，图和语言是不同的两种表示形态，它们两者之间也许会有难以逾越的鸿沟。但是，如果从人类的认知来看，当面对这种化学分类问题的时候，人类也是先将其转化成隐式的语义信号，然后与自己学习到的知识进行对比，进而最终得到判断结果。因此，语言模型对图的处理可能更加符合人类的认知习惯。

现有的 GNN 模型在处理图分析任务的时候，一般会根据它的性质，比如碳原子的电子极性属性等，把节点转化成固定的表示向量，然后再根据 GNN 的聚合函数学习到图的表示。GNN 的训练信号来源于少量的标注数据集。但是，目前这种通用的实验设置天然具有两大缺陷：(1) 节点属性的大量缺失。例如，碳原子可能除了电子极性之外还有别的特性，如质子数目和电子数目等。如果利用简单的经验知识仅仅将其部分属性转化为节点表示向量，那么就会忽略掉大部分的节点属性信息。(2) 就像 Kate Bowler 说过的，Everything happens for a reason。图的构建和表现出来的化学性质都是有原因的，这种原因可能已经在化学课本或者资料上被准确定义了。而目前这种实验设置抛弃了准确定义的先验知识，完全依赖于少量的标注数据中的微弱信号进行推理，等价于自我抛弃了豪华的机甲套装而强行拿着长矛盾牌出战。

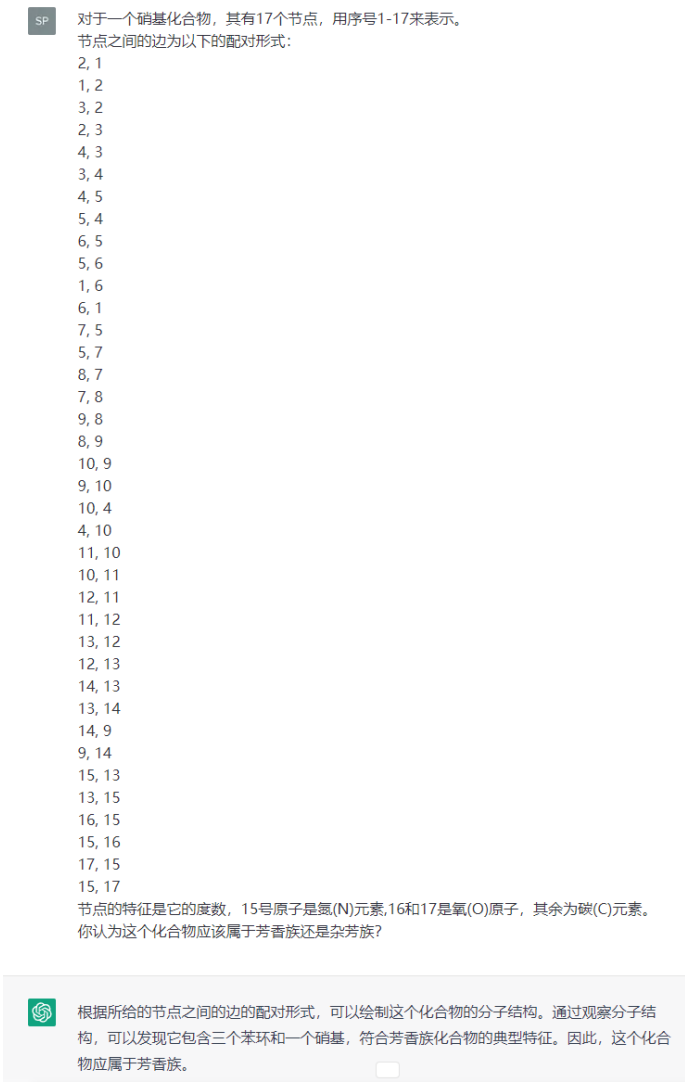


图 10：ChatGPT 能够正确处理化合物图的一个例子

那么，为什么大规模语言模型有助于解决上述问题呢？其原因在于处理方式更加符合人类认知，毕竟返璞归真可能会更加贴近事物的本质。(1) 在 LLM 里输入图时，节点是被直接表示为其原始文本，比如 “碳原子” 和 “氮原子”。如果 LLM 足够强大，那么它的知识里已经有能力记住碳原子在维基百科、化学教材里提到的特征，这样能够提供更全面更强大的节点属性；(2) 在 LLM 预训练学习的过程中，它梳理了大量的化学书籍并熟悉了很多背景知识，因此它进行判断的依据不仅仅是依靠标注数据，而是人类已经总结好的宝贵知识。

LLM 和 GNN 的对比可以类似于：两个人同时参加化学考试，在做一个分类题的时候，GNN 是把其他几十个分子以及对应的标签给列出来，从 0 开始推断里面可能包含的知识；而 LLM 则是记忆了很多已有的知识，然后基于知识对当前的分子进行判断。某种程度上来说，LLM 是站在了巨人的肩膀上，并且也更加符合我们人类的认知和行为习惯。

因此，研究员们相信，LLM 有希望在图表示学习领域成为未来。但是，目前也存在这一些困难和挑战。最主要的挑战可能在于图结构的多样性，例如在社交网络中学习得到的知识可能很难应用于化学分子的判断中，所以能否有一个通用的 LLM 模型来处理不同类型的图上五花八门的分析任务，还值得科研人员深思。

参考文献

[1] Li R, Zhao J, Li C, et al. House: Knowledge graph embedding with householder parameterization[C]//International Conference on Machine Learning. PMLR, 2022: 13209-13224.

[2] Savage N. Graph matching in theory and practice[J]. Communications of the ACM, 2016, 59(7): 12-14.

[3] Zhang P, Guo J, Li C, et al. Efficiently Leveraging Multi-level User Intent for Session-based Recommendation via Atten-Mixer Network[C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023: 168-176.

[4] Wu S, Tang Y, Zhu Y, et al. Session-based recommendation with graph neural networks[C]//Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 346-353.

[5] Pan Z, Cai F, Chen W, et al. Star graph neural networks for session-based recommendation[C]//Proceedings of the 29th ACM international conference on information & knowledge management. 2020: 1195-1204.

[6] Xia X, Yin H, Yu J, et al. Self-supervised hypergraph convolutional networks for session-based recommendation[C]// Proceedings of the AAAI conference on artificial intelligence. 2021, 35(5): 4503-4511.

[7] Li C, Pang B, Liu Y, et al. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 223-232.

[8] Bi S, Li C, Han X, et al. Leveraging Bidding Graphs for Advertiser-Aware Relevance Modeling in Sponsored Search[C]// Findings of the Association for Computational Linguistics: EMNLP 2021. 2021: 2215-2224.

[9] Yang J, Liu Z, Xiao S, et al. GraphFormers: GNN-nested transformers for representation learning on textual graph[J]. Advances in Neural Information Processing Systems, 2021, 34: 28798-28810.

[10] Pang B, Li C, Liu Y, et al. Improving Relevance Modeling via Heterogeneous Behavior Graph Learning in Bing Ads[C]// Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 3713-3721.

[11] Zhao J, Qu M, Li C, et al. Learning on Large-scale Text-attributed Graphs via Variational Inference[J]. ICLR, 2023.

[12] https://ogb.stanford.edu/docs/leader_nodeprop/

[13] https://chrsmrrs.github.io/datasets/docs/datasets/

相关阅读

扫描二维码查看文章

微软亚洲研究院深入探索图深度学习领域两大挑战，以图深度学习赋能知识计算

在图深度学习领域的持续深耕，让微软亚洲研究院 DKI 组提出了一系列新方法和新思路，为多项研究成果的突破奠定了基础。那么对于图深度学习技术在知识计算领域的应用，微软亚洲研究院的研究员们有哪些独到的理解？又预见了哪些前沿的研究方向？

科研第一线

AAAI 2023 | 微软亚洲研究院精选论文

AAAI（Association for the Advance of Artificial Intelligence）是由美国人工智能协会主办的人工智能领域的顶级学术会议之一。2023年度的 AAAI 大会于 2 月 7 日至 2 月 14 日举办，微软亚洲研究院有多篇论文入选，欢迎扫码二维码了解这些最新学术成果。论文主题：工业应用中的人工智能 （左侧二维码）；人工智能理论、负责任的人工智能和人工智能创作等（右侧二维码）。

扫描二维码了解更多信息

微软亚洲研究院项目入选《2022 CCF 技术公益年度案例集》

2023 年 1 月，微软亚洲研究院助力公益组织制作视频这一案例被中国计算机学会（CCF）收入了《2022 CCF 技术公益年度案例集》。2022 年 10 月，山水自然保护中心急需在 5 天内赶制用于重大奖项的申报视频。来自微软亚洲研究院的研究员们在得知此事之后，利用多模态模型 NUWA、基于深度学习的 AI 音乐创作项目 Muzic 等先进的 AI 技术生成了视频素材，并对原始的信息图形设计进行了优化。同时，研究员们基于这些 AI 创作成果，还融入了自己的理解和创造力，最终成功在截止日期前完成了内容清晰、画面丰富的视频制作，帮助山水自然保护中心获得了该重大奖项。

扫描二维码了解更多信息

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。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？

2010 年，还在博士求学阶段的梁傑然（Mike Liang）成为了微软雷德蒙研究院刘劼博士和赵峰博士的一名实习生。实习期间，梁傑然了解到研究院有一个研究课题是借助传感器实现数据中心环境的数字化，而这恰好也是他的专业方向和兴趣所在。"物理世界中有太多的现象，比如声、光、热、力、电，虽然我们能够看到和感受到，但却无法更进一步地理解。我博士期间的专业就是传感器网络。当时这个领域的同学们都有一个梦想，希望可以在物理世界部署大规模传感器网络来达到实时数字化，并透过无线方式将感知的数据传输存储起来，从而更深度地理解物理世界。"谈及开展相关研究工作的初心时，梁傑然说。

微软亚洲研究院高级研究员梁傑然（Mike Liang）

那时对于数据中心来说，温度的精确感知和散热是一个丞待解决的大问题。为了防止数据中心过热，业界通常的做法是将冷却系统的温度调至最低，但这会产生高昂的电力成本，几乎一半的电费都花在了冷却系统上，造成了巨大的资源浪费。因此，微软雷德蒙研究院的研究员希望通过设计和部署上千个无线传感器，来理解数据中心热分布和预测变化，精准控制冷却系统的温度。

然而，利用传感器收集数据，再通过无线网络传输数据，这一过程本身也是一个巨大的挑战，如何实现超大规模的低功耗无线通讯又成了新问题。对此，不同的研究机构提出了五花八门的解决方法，底层架构研究的混乱也让上层的应用变得困难。"我们通过系统化的研究和梳理，将我们的发现和洞察提炼总结，最终给无线研究人员提供了一个统一且优化过的无线通信网络架构。其他研究人员可以直接在这个底层架构的基础上进一步进行创新研究。" 梁傑然介绍道。

最终，这项开创性和基础性兼具的研究工作 "Design and Evaluation of a Versatile and Efficient Receiver-Initiated Link Layer for Low-Power Wireless" 经受住了时间的检验，在国际移动计算和感知领域顶级学术会议 ACM SenSys 2022 上获得了时间检验奖（Test of Time Award），得到了研究界的肯定。正如 ACM SenSys 大会对这项工作所做的评价："2010 年，该研究工作率先实现了在低功率无线通讯中利用同步传输在 MAC 层的优势，来突破低功率无线电的极限。在过去 12 年的时间里，这项成果为许多物联网和嵌入式系统奠定了无线通讯协议的基础。"

梁傑然 2010 年的研究工作荣获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）

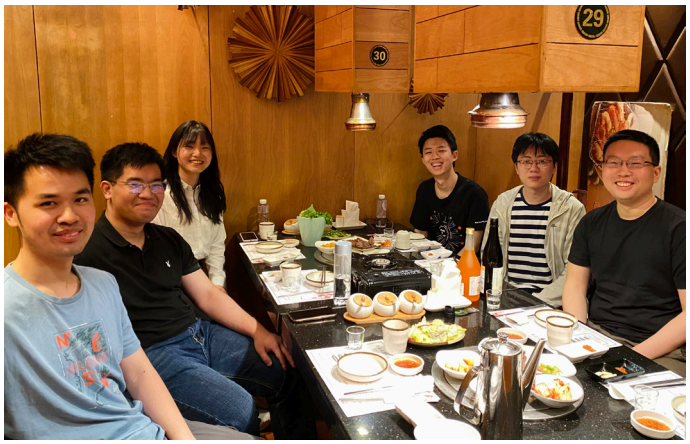
加入微软亚洲研究院，与有趣的人做有趣的事

"毕业后，虽然有多个选择，但我只想加入微软亚洲研究院。我想与更多有趣的人做有趣的事，这里也满足了我对人生的期许。"

微软亚洲研究院对梁傑然的吸引力，一方面来自于身边那些背景各异、专业不同的同事们。他们有的喜欢硬件，有的专攻操作系统，还有的深入算法研究，等等。在梁傑然看来，这会是一个“有趣”的组合。另一方面，微软亚洲研究院自由的科研氛围给每一位研究员都提供了足够的空间和资源去做自己“感兴趣”的研究，让梁傑然可以一展身手，尽情施展自己的理想和抱负。

“一方水土养育一方人”。梁傑然在宝岛出生，加拿大成长，美国完成博士学业。经历丰富的他一直非常同意这句话的含义——每一个人都有着自己独特的性格、喜好，和擅长的技能。而现在，在微软亚洲研究院多元包容的研究氛围中，他对这句话有了更多的理解。他意识到不同背景的人，思维观念不同，对于同一个问题的解读方式也会有所不同。

底层思维的转变让梁傑然对研究工作有了新的认识，他愿意花更多的时间去倾听别人的想法，而不是一味输出自己的观点，并激发大家把自己的优点发挥到极致。比如，团队中有实习生对代码有很极致的追求，这无形之中促使了团队在系统工程上有了更深的认知，发现很多新问题。再比如，来自统计学专业的实习生给团队正在做的 AI for Systems 研究带来了数学思维。梁傑然非常享受这种能和这些“有趣”的人一起做更多“有趣”的事的状态。



梁傑然（右一）与实习生合照

梁傑然（左一）与系统和网络研究组同事合照

多角度研究一以贯之，让人人皆可编程的云成为现实

2011 年至今，梁傑然在微软亚洲研究院度过了十余个春秋，取得了多项研究成果。作为低功耗无线通信研究的延续，梁傑然在正式加入研究院后又实现了从硬件到软件的多项创新，并和微软雷德蒙研究院刘劼博士团队一起将其应用在微软多个数据中心，推动了微软 Azure 对数据中心在环境监控与低碳排放的研究。

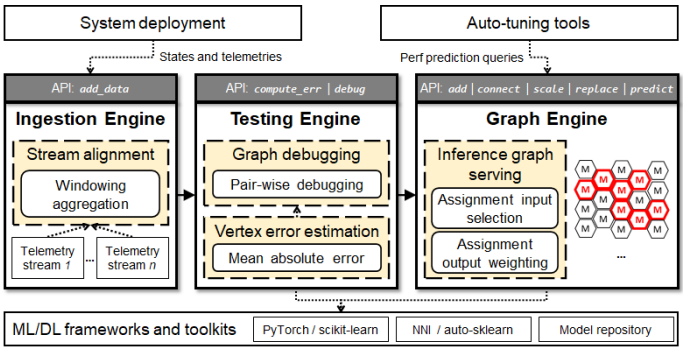
伴随着智能手机的流行，梁傑然找到了新的研究兴趣。“智能手机最酷的一点是其上可以承载各种各样的应用程序，而这让‘人人皆可编程成为了一种趋势。”但由于每个人的编程质量参差不齐，大量的应用程序难免会产生很多 bug。为了保证最终用户的应用体验，就需要人工对应用商店中成千上万的应用进行审核，但这不仅效率低且质量无法保证。对此，团队和微软雷德蒙研究院的 Ranveer Chandra 博士萌生了自动化审核的想法，并开始借助机器学习技术开发审核工具。利用自动化审核工具，机器会进行一轮初筛，淘汰那些问题明显的应用程序，然后再将剩余应用交由人工审核，大大减轻了审核人员的压力。梁傑然和团队还曾发表题为“How to Smash the Next Billion Mobile App Bugs?（如何解决下 10 亿个 App 的漏洞？）”的论文，以期籍由机器学习驱动的技术，让人人都能开发出高质量的应用程序。相关的一系列研究也推动了业界对应用程序审核机制的创新。

近年来，梁傑然又看到了人工智能技术在“人人皆可编程”下的更大潜力，转入系统和网络研究组，主攻 AI for Systems 方向的研究。在梁傑然看来，云计算的发展让每个人都能方便、快速地获取计算资源。未来，云计算势必是世界上最强大的计算系统之一。虽然云看起来使用简单，只要根据需求购买相应数量的虚拟服务器即可，但事实并非如此，它还会涉及一系列的分布式编码、配置、运维的专业问题。例如，使用哪种类型的虚拟服务器更能满足业务需求？每台虚拟服务器适合运行哪些业务应用？虚拟服务器之间如何分布式运行？业务高峰低谷时如何平衡资源？不仅如此，还需要优化配置应用参数以更好地利用云资源，更要防止一台虚拟服务器发生意外对其它虚拟服务器造成影响。解决这一系列问题都需要专业知识。随着机器学习算法愈渐成熟，梁傑然在 AI for Systems 的研究可以利用机器学习和 AI 等技术，使云上的资源能自主地适配用户应用的负载需求。最终，云的这个自主性将帮助所有人都能更方便地编程世界上最强大的计算系统。

与此同时，从和产品组合作中，梁傑然和团队深刻地体会到机器学习理论和实际系统问题的差距。云原生系统普遍有着高度的复杂度、规模和行为动态变化。当学习对象发生变化时，需要花费数小时甚至数天重新收集系统数据再训练 AI 模型，而且在这个过程中还会产生巨大的成本。在 AI for Systems 的模型学习中，他们发现云系统的复杂变化其实有规律可循。比如系统上有十项云服务，某一周更新其中的一项，虽然这会影响整个系统，但理论上更新只是更改了其中仅一个服务的编码配置。同样的道理也适用于云服务的扩容。因此，AI 模型也只要相应地修改变化部分

即可，这就是模块化。“模块化的思维方法让我们重新思考以往 AI for Systems 的落地，从而促使我们在范式上进行革新。”梁傑然说。

基于这些发现，梁傑然和团队提出了 Fluxion，一个通过模块化学习建模端到端系统延迟的框架。Fluxion 引入了新的抽象学习分配，允许对单个子组件进行建模，而不用对整个系统进行端到端延迟建模。并且通过统一的界面，该方法可以将多个异构学习任务组合成一个推理图，动态地对复杂的分布式系统进行建模，显著降低了成本和延迟。相关论文 “On Modular Learning of Distributed Systems for Predicting End-to-End Latency” 已被国际顶级网络领域学术会议 NSDI 2023 接收。



Fluxion 的架构图

此外，梁傑然和团队还在着手进行其他的研究，来实现云上资源的自主适配性。除了大规模的自动扩容技术，还包括 AI for Processors 技术。例如，云上运行的数据库和网页服务器对芯片有着不同的要求，而它们却被同一块通用芯片以同样的方式运行。如果芯片不能及时做出优化，那么理论上这些软件的性能就会被极大影响。梁傑然和团队希望使用 AI 技术让通用芯片更深刻地理解正在被执行的指令（或低阶机器语言），来自主地针对不同的场景做出不同的调整。

为实现 “人人皆可编程” 的愿景，梁傑然和团队从多个角度不同方向探讨云资源的自主适配性。其中部分研究成果如今已经融入到微软的产品和服务中，如微软必应（Bing）搜索产品中，通过自主优化数据缓存，最终提升了终端用户的体验，让梁傑然的愿景逐渐成为现实。

以三年为界，不断地试错与调整

无论是获得 ACM SenSys 2022 时间检验奖的无线感知研究，还是 “人人皆可编程” 的愿景，每次研究赛道的转换，梁傑然都选择将长期主义的理念灌注于研究工作之中。为什么梁傑然一次又一次地选择这种短期内无法实现突破性成果的研究？

“这还是要回归到我的理念：与有趣的人做有趣的事。计算机行业瞬息万变。但我觉得更有趣的是三年之后有可能发生的行业趋势。这一定程度上是个赌注：赢了，我们就比其他人早走了几步，甚至有可能带来范式上的转变；即使失败了，三年的时间也有一定的容错空间，可以让我们再次调整，重新选择赛道。” 梁傑然特别喜欢微软杰出科学家 Phil Bernstein 对于科研的反思——我们应该关注研究成果究竟能对学术界或产业界的未来 3-5 年带来什么样的推动和改变，而不是每年发了多少篇论文。

“更重要的是，微软亚洲研究院一直鼓励长期投资，做有影响力的研究，并创造了一个多元包容的科研氛围。这让我可以和背景不同的同事们形成‘有趣’的组合，一起做 ‘有趣’ 的研究。”

梁傑然（右一）和微软中国网球俱乐部的同事一起在海淀区重点企业网球俱乐部比赛中获得佳绩

工作之余，梁傑然还是一位网球高手，这项 “有趣” 的运动他也坚持了十多年，既获得过阶段性的荣誉，也在努力向着长远的联盟球队第一的宝座进军。

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？对话微软亚洲研究院（上海）负责人邱锂力博士，一起了解微软亚洲研究院（上海）的成长步伐和未来规划。

科学匠人 | 李琨：执著于高性能计算研究的“别人家的孩子”

2022 年 4 月 30 日，对于微软亚洲研究院研究员李琨来说是个很特殊的日子——“一夜爆火”成为学校风云人物，这让本就“社恐”的他花了好几天时间才习惯。原来在前一天，李琨拍摄了一段自己面试求职经验分享的 vlog，并引发了众多网友的围观。优秀的科研和项目履历让他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”加入了微软亚洲研究院，成为了“异构计算组”的研究员。

而就在 2023 年 2 月 18 日，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”），前去参加了 CCF 颁奖典礼。领奖的那一刻，他无比激动，“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。作为 up 主，李琨也以 vlog 的形式，记录了这一难忘的“高光”时刻。

“别人家的孩子”在博士期间坐过四年“冷板凳”

2012 年 ，刚刚进入大学的李琨和很多人一样，并没有明确的专业方向。机缘巧合之下，他被调剂到了计算机专业，在接触到计算机之后，李琨发现通过编程，以代码的形式可以实现很多有趣、奇妙的事情，也可以开发出不同的软件和硬件的新功能。这让原本就对理工科较为擅长的他，开始对计算机尤其是并行计算、高性能算法和软件的设计产生更浓厚的兴趣。至此，李琨踏上了高性能计算（HPC）研究的“不归”路。

2016 年，李琨以优异的成绩直博进入中国科学院计算技术研究所，继续从事高性能计算方面的研究。但是博士的求学之路并不是一帆风顺的，回顾那几年，李琨总结了十六字箴言：但行好事，莫问前程；道阻且长，行则将至。

微软亚洲研究院研究员李琨

李琨用“一直在坐冷板凳”来形容自己读博后的前四年。尽管他不断参与各类科学研究和工程项目，但相关成果一直未在领域内的知名会议或期刊上得到接收。从学术的角度来看，这意味着自己一直以来所做的工作并没有得到同行和学界的认可。对大部分博士生来说，这会是非常令人沮丧和焦虑的，心态也可能会崩溃。李琨坦言，在博士前四年的屡试屡败中，自己也时常游走在崩溃的边缘。不过即使如此，他依然坚定着对高性能计算方向的研究。在李琨看来，不断试错的过程也是经验积累的过程。为了更好地推进研究，李琨会及时与同学、朋友和导师沟通，探讨研究工作中遇到的难点，讨论可行方案，避免研究偏离正确轨道。在交流过程中，他也能纾解心情，让自己保持一个平和的工作心态。

功夫不负有心人，付出总会有回报。博士阶段后期，李琨的研究工作陆续开花结果，多项工作相继被高性能计算领域的顶级会议与期刊接收，相关研究也推动了高性能算法和软件在国产超算平台上的大规模研发，产生了较强的影响力。因为李琨坚持不懈的探索和出色的研究成果，2022 年 6 月，他获得了中国科学院院长奖，该奖项是中国科学院研究生奖学金中含金量最高的奖项之一，旨在表扬和鼓励那些在科学研究和技术创新方面做出突出成绩的研究生。

近期，他的博士学位论文《大规模并行多层次不连续非线性可扩展理论研究及应用》还获得了 2022 年度 CCF 优博奖和 2022 年度 ACM SIGHPC 中国优博奖。该论文系统性地总结了李琨在博士六年期间的工作和成果：通过对大规模并行多层次不连续非线性可扩展理论展开研究，深入分析可扩展性发展规律，提出了多层次协同设计理论，在多种硬件并行规模、不同软件并行粒度、各级交叉并行应用上开展了多种可扩展性的优化设计。

李琨表示，“很荣幸能够获得这些奖项。其实，我是站在了巨人的肩膀上。博士阶段的很多工作都是实验室里多年的科研和项目沉淀的结果，这离不开导师和各位指导老师的培养和指导。同时，博士论文的撰写工作很多也是在微软亚洲研究院实习期间完成的，这期间系统与网络组的各位前辈也给我提出了很多宝贵的意见。这些奖项是对我这些年工作的认可，给予了我更大的信心继续从事高性能计算研究，并激励着我在这个方向持续探索。同时，这也能够让更多人了解高性能计算，吸引更多的优秀人才加入到这一研究的行列中来。”

Big Science on Cloud 让“科学计算”更普及

2022 年的毕业求职季中，李琨收到了多家知名企业和研究机构的 offer。最终，李琨选择了微软亚洲研究院，正式成为“异构计算组”的研究员。

谈到选择微软亚洲研究院的原因时，李琨说："我对研究院早有耳闻，一直向往来这里工作。博士期间我有幸加入系统与网络组进行实习，并了解到微软亚洲研究院对科学计算（Scientific Computing）非常重视，这刚好和我的研究方向匹配。所以，我在实习过程中就与各领域的科研前辈共同探索将高性能计算与机器学习驱动的科学研究进行融合，这些交流与合作再次加深了我对科学计算研究的兴趣，也让我对研究院有了更多的了解。研究院充满着自由、包容的研究氛围，而且有许多优秀的人才，大家都会基于平等的身份互相学习、交流、探讨问题。更重要的是，研究院鼓励长期思维，不会催促你短期内尽快出成果，也不会以论文数量为导向，这给了年轻学者更多的试错机会，让我们能够勇于尝试，开拓新的研究方向。"

在研究院的实习生社团活动中，李琨体验 VR 游戏设备

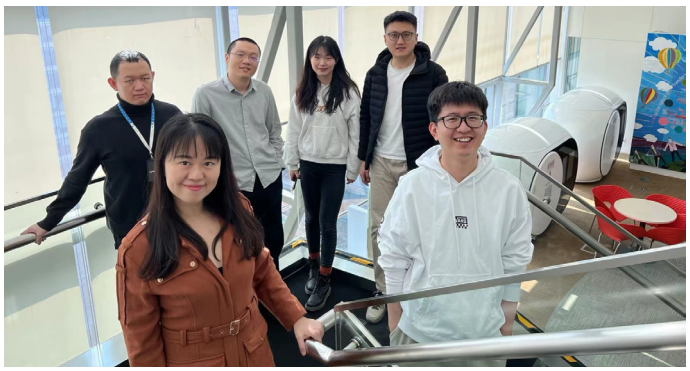
读博期间，李琨就开始关注科学计算的研究。在他看来，很多生物学、化学、物理学上的科学发现和探索，不仅需要科学家的努力，还需要与计算机相结合才能充分发挥算力来加速大规模计算。比如李琨参与过的核材料辐照损伤观测，现实中它需要几十年甚至上百年的时间才能观察到结果，这显然是不现实的，而有了超级计算机和高性能计算技术的帮助，则能通过数值模拟，大大加速科学研究的进度。

过去几十年，科学计算领域也正朝着这个方向蓬勃发展，在发现遗传学奥秘、地震模拟预测、新冠肺炎疫情预测上都取得了进展。但这些模型严重依赖于运维成本高、软件移植效率低、硬件可扩展性不佳的超级计算机，更重要的是，并不是所有机构和研究人员都能轻易获得这些计算资源，这严重阻碍了科学计算研究的普及化。

凭借硬件和网络技术的进步，像微软 Azure 云平台这样的云基础设施广泛采用统一的 CPU 和 GPU 异构架构，从而推动了人工智能的繁荣发展。与传统超级计算机相比，云平台可以为科学计算提供一个高性能、低成本且弹性扩展的计算平台。但是，云平台上也存在一些未解决的问题，例如如何让 AI 驱动的 HPC 应用获得更高精度，如何提升对 AI 基础设施的利用率，如何降低 HPC 优化技术的门槛等等。因此，云平台上的 "HPC+AI" 也成为了李琨和异构计算组当前研究的主要课题之一。

为此，微软亚洲研究院异构计算组提出了 Big Science on Cloud 的理念，其目标是通过创新性地协调云上高精度和高性能计算单元，实现面向所有用户的通用一站式 "HPC+AI" 的科学解决方案，从而更高效地实现广泛的科学计算。

"对于 HPC+AI 的研究，目前我们主要关注两点：一是专注于科学计算中经典高性能算法的异构优化研究。以 Stencil 算法为例，它是高性能计算中的七大关键算法之一，也是许多科学计算应用中的关键算法。我们以这个传统的 HPC 算法切入，进而提升云上科学计算的性能；二是以算法为基础，逐渐将其拓展至应用层，如分子动力学的大规模模拟。事实上，我们不仅仅是算法的设计，更多地是以算法为用例，结合 AI for Science 领域，探索云上高性能、低成本、可扩展的大规模异构并行研究。" 李琨说。



李琨（第一排右一）和异构计算组同事们

身份的转变代表着更大的责任

从实习生到研究员，不仅仅是角色的转变，更是心态的转变。李琨认为，做实习生时还是保持着在校学生的心态，研究时更多地是遵循导师或 mentor 的建议，沿着较为明确的方向或想法前进实现即可。

然而，成为正式的研究员之后，肩上的责任变得更大了。除了要对自己负责，还要对实习生、对共事的同事负责，工作内容也会变多变广。探索研究时也要开拓新思路，形成新想法，全面思考问题，并建立长远思维，让研究工作具有更深远的影响。

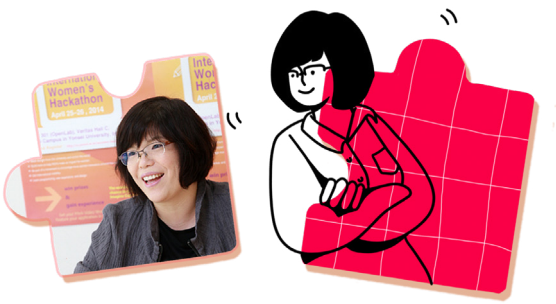
"无论是上学期间还是现在工作，每一个新想法都让我感到兴奋。科学家们在探索未知过程中 '攻坚莫畏难' 的钻研精神也深深影响着我。希望借助微软亚洲研究院这一广袤的科研腹地，我可以与各领域优秀的科研人员一起做出更多有意义、有价值的创新成果。" 李琨说道。

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。这些“斜杠女性”有着不拘一格的职场哲学与处世之道，她们也更加乐于在开放、多元、包容的氛围中，为研究院连接众多合作伙伴，以自己的热情和智慧，为广阔的计算机学术生态拼图填补着不可或缺的板块。

多元与包容是创新和创造的动力

Miran Lee，微软亚洲研究院学术合作总监
国籍：韩国
关系管理能手 / 自主创业和经理人双 buff 加成 / 徒步旅行爱好者



Q：你的主要工作职责是什么，其中哪些部分最具挑战？

Miran Lee：我负责微软亚洲研究院在韩国和亚太地区的学术合作，主要的工作内容包括制定学术合作的战略和方向，发现业务机会，设计各类合作项目，以及管理维护合作伙伴的关系。在学术合作部，我们与合作伙伴通过研究合作、课程开发、学术交流等方式建立稳固的合作关系，推动计算机技术的进步并助力计算机领域的人才培养与发展。

在这个过程中，由于我的工作十分具有多样性，所以我认为最有挑战性的部分是，如何有效协调多个项目的优先级；如何通过良好沟通和管理维系牢固的伙伴关系；以及如何及时地把握学术界趋势从而为学术合作部设计具有足够吸引力的项目或计划。

Q：如果将计算机学术生态看作一幅拼图，那么你是如何与他人共同完成这幅拼图的？在推进学术交流与合作的过程中，有什么成功的秘诀或建议？

Miran Lee：作为计算机学术生态系统的一员，与他人共同完成这幅“拼图”的最佳方式就是合作和沟通。与学生、科研人员和教职员工建立强有力的伙伴关系必须要采取通力合作的方式，所有各方都为实现共同的目标而努力。而有效的沟通则是在合作伙伴之间建立信任和理解的关键，它有助于创造一个积极的工作环境，让每个人都感到被重视和倾听。

我认为，促进学术交流与合作，首先要为每一次合作设定明确的目标，并确保每个参与者都可以清楚地了解自身的定位。其次要营造一个开放、包容的氛围，鼓励大家沟通、反馈，让每个人都可以自在地表达想法和意见。倾听伙伴们的想法，认可他们的专业知识，对于伙伴间建立信任和尊重也十分重要。而对新想法保持开放的态度，并积极适应不断变化的环境，可以确保合作始终保持相关性和有效性。

Q：你曾经在科技领域多个公司工作过，还创过业，为什么后来选择加入了微软亚洲研究，并在这里工作十余年？

Miran Lee：微软亚洲研究院提供能够充分发挥我在技术和业务方面优势，并与业内优秀人才合作的机会。而且研究院致力于推动计算机科学领域前沿技术的发展并造福整个人类社会的这一理念，非常令我向往。这里还拥有独特的多元与包容文化，让我享受与不同文化背景的同事共同工作，共同改变世界的成就感。

Q：以往那些丰富的跨行业经验，对你现在的工作有怎样的帮助？

Miran Lee：我曾经在一些优秀的科技公司担任过管理者职位，这让我在推动建立长期合作伙伴关系以及关键领域创新时，能够制定并采取更合理的策略。在韩国安阳大学担任兼职教授的经历，使我对师生和科研人员的需求有更深的理解，有助于我设计出更好的合作计划。除此之外，我在软件开发和技术方面的经验也让我能与研究院的各个研究组进行有效的协作。总的来说，跨行业的经验给我带来了独特的视角，为我在目前的岗位上所取得的成绩做出了贡献。

Q：微软亚洲研究院的员工有着不同国籍及文化背景。你对跨文化的沟通和合作有何感受?

Miran Lee：跨文化交流是微软亚洲研究院至关重要的组成部分，也是我们进行创新和创造的重要动力之一。身处多样性文化交汇之中，我们可以获得新的视角和见解，并学会理解及欣赏那些不同的思维与工作方式。

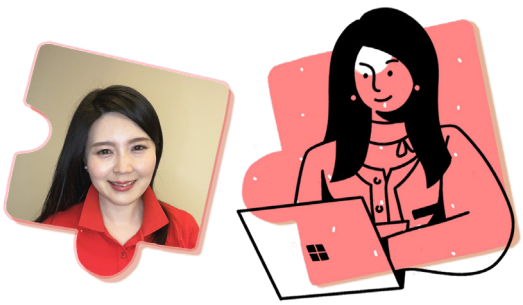
跨文化的沟通合作也让我们锻炼出了更强的人际交往能力，并且善于倾听、拥有同理心。其中最为关键的是拥有开放包容的思想，尊重不同的意见和观点，愿意对沟通或工作方式进行调整，从而适应不同文化的特性。

Q：你有哪些业余爱好，这些爱好给你带来了哪些收获?

Miran Lee：我的爱好是阅读和徒步旅行。阅读对一个人的成长、智力发展和生活乐趣都大有裨益。置身于大自然中的徒步旅行不仅能改善健康、增强力量和耐力，还有助于减轻压力、改善情绪，让思维清晰。这些爱好既能让我放松心情、享受生活，还经常让我获得一些意外惊喜的新技能、新兴趣，或是结识志趣相投的朋友。

多元与包容是创新和创造的动力

Mawo Kamakura，微软亚洲研究院资深学术合作经理
国籍：日本
传统文化遗产保护者 / 美食爱好者 / 茶文化拥趸



Q：是什么样的契机，让你选择加入了微软亚洲研究院?

Mawo Kamakura：这一切开始于我研究生导师的一封电子邮件。我实验室的许多学长和同事都曾在微软亚洲研究院实习过，所以我对研究院“闻名已久”，但因为自己的研究范畴是人文社科专业，因此我从未预料到会和研究院产生交集。

但当我从导师的推荐邮件中了解到研究院学术合作部的工作性质后，就开始认真考虑这个机会。一方面此前来往于埃及、柬埔寨、意大利等多个国家的工作经历让我非常适应与不同文化背景的人进行合作；另一方面我一直与文化遗产组织（包括联合国机构、政府组织）和计算机科学等领域的人共事。我感到这些独特的经历能让我在研究院发挥不同的作用。

事实证明这是个正确的选择。微软亚洲研究院给了我发挥自己专长的空间，也让我获得了“活出自我”的满足感!

Q：你一直致力于通过 3D 数字技术保护传统文化，如何看待 AI 等创新技术在此领域的作用?

Mawo Kamakura：大概从 2005 年开始，我陆续参与了一些关于文化遗产的 3D 数字化保护项目，包括在柬埔寨、意大利、埃及和日本的联合国教科文组织世界遗产，比如，将第二艘胡夫太阳船中发掘的物件进行 3D 数据储存等。

由于 AI 和存储技术的发展，以及显示设备的小型化和多样化，文化遗产的大规模数据存储变得越来越容易。同时，新技术也让我们能够更加便利地获得更多未知的知识，比如破译文物上的古代语言，识别模糊的文字和图画。总之，AI 技术给我们带来了更丰富、更灵活的选择，让人类文化瑰宝得以长存于世。

Q：你目前主要负责什么工作？如果把你涉猎的不同领域比作一幅拼图，你希望这个拼图最终会形成怎样的画面？

Mawo Kamakura：我负责微软亚洲研究院在日本地区学术合作的相关事宜，同时我还是东京大学空间信息科学中心 (Center for Spatial Information Science, University of Tokyo) 的访问学者。我的工作可以概括成三个方面：学术合作、产业合作和连接沟通。

我希望让更多不同领域的研究者感受到与微软亚洲研究院合作的价值和意义，并且尽我所能为双方建立连接，推动合作创新，用自己擅长的“连接 A 与 B”的能力为不同领域牵线搭桥。

Q：你是如何运用不同领域的知识和技能来解决当前工作中所遇到的问题?

Mawo Kamakura：当遇到问题我会考虑两件事：寻找能将过往经验用于解决眼前问题的方法，以及期待从新的经历中获得新知识来获得自我提升。我也会尝试把一个任务分解成多个相对简单的小环节，并发现其中最重要的部分。

Q：在微软亚洲研究院与不同国家和文化背景的同事合作，给你带来了怎样的跨文化交流感受?

Mawo Kamakura: 微软亚洲研究院的很多同事都来自亚洲，在文化上有很多相似之处。例如除英文外，一些同事能和我使用汉字交流。即便我们对汉字有着截然不同的读法，但这依然是非常有趣的经历。这些交流能让我了解不同的观点并且获益良多。

对多元文化的包容是一个优秀组织的特征之一。因此，微软亚洲研究院才能聚集众多拥有不同背景、个性和专长的杰出人才。在这里每个人都能自由追求自我，也能结交良师益友，这正是研究院最具吸引力的一点。

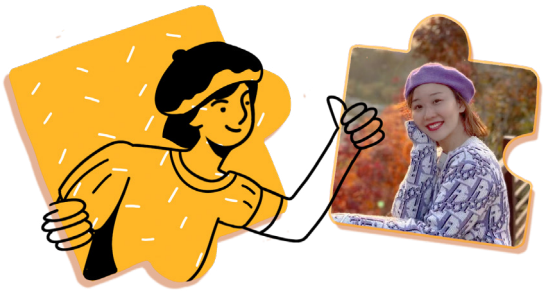
Q：你有哪些业余爱好可以分享？它们带你给哪些收获？

Mawo Kamakura：在行走于各国的旅程中，我发现自己对不同国家的饮食非常有兴趣。比如品茶就是我最喜爱的饮食文化之一。同事送给我的五味子茶，和我在研究院的一次活动中发现的小青柑茶，至今都令我难以忘怀。

品尝不同文化的饮食，能让我真切地感受到自己的“存在感”。另外我也热衷于自己烹饪，这对我来说是一个重整思绪、保持工作与生活平衡的好方法。

好的技术就如同投身环境科学的初心——为了让世界更美好

石贝贝，微软亚洲研究院资深学术合作经理
国籍：中国
环境科学 / 计算机入门到精通 / 带着家人去旅行



Q：你现在的工作内容是什么，其中哪个部分最具挑战？

石贝贝：我主要负责微软亚洲研究院的开放科研合作计划，重点负责“可持续发展”与“信任”研究主题，另外也包括铸星计划，以及多个区域的高校合作等工作。

总结起来，我的工作就像为两个原本孤立的地方架设桥梁，为两种独立物质创造反应催化剂，为合作关系奠定稳固地基。而其中的挑战就是在不同时境下，找到需要架桥、催化、稳固的目标，并用最恰当的方式来实现。

Q：拥有环境科学专业背景的你，如何看待科技发展对自然环境和社会带来的影响？

石贝贝：我从小就很喜欢去到自然景观非常丰富的地方，当初选择环境科学这个专业就是希望可以保护环境，保卫地球家园。上学期间，我发现环境科学的很多课程与计算机科学存在交叉融合。后来一次生态与计算机交叉学科的科研项目，成了我加入计算机行业的契机。此后我承担过许多环境科学和计算机领域的跨界项目，在这个过程中我感受到，好的技术与我最初投入环境科学的初心是相同的，都是为了让世界更美好。

不可否认，在技术的发展过程中造成了对自然环境的破坏和资源的过度消耗，而解决这些问题其实更加需要技术创新的力量，比如用与环境友好的技术来替代破坏环境的发展方式。但这样的科技创新是非常不易的，它需要各学科的学者们通力合作、长期钻研、融汇创新。在微软亚洲研究院已经有一批研究员通过我们的负碳计算科研合作计划与亚洲许多高校的环境科学、地球科学和气候学等领域的专家一起投身其中，为这个艰巨但极具意义的研究方向而努力。我也非常有幸可以与他们一起并肩向前。

Q：跨学科的背景，给你的工作和生活带来了怎样的影响？

石贝贝：在我的职业生涯中，一直在不停地从舒适圈跨越到新领域。从环境科学到计算机科学，从研究员到研究项目经理，我需要不断学习那些以往可能非常陌生的知识。幸运的是，一路走来身边优秀的同事给予了我很大地帮助和信任，帮我快速成长。

好在我独有的专业背景以及跨领域的视角和思考，也常常给团队带来有价值的贡献。在这个过程中，我逐渐能够以更好的心态去面对生活和工作的变化，应对新事务和新领域的挑战，并且以更宽广的视野发现广阔的精彩世界。

Q：你如何看待 AI 对社会的影响？

石贝贝：AI 的发展和普及，尤其是最近随着大模型时代的到来，也许我们很快就会看到先进 AI 技术推动的新一轮“工业革命”。回顾历史，虽然每次技术变革都会出现喜忧交织的情况，但技术本身是中性的，积极掌握技术并用于正途一定会为社会发展创造巨大价值。

在这个过程中，从事 AI 研究的我们比任何时候都更加看重技术能推动社会向着可持续且互信的方向发展。正因为此，微软亚洲研究院从 2020 年就开始推出负碳计算科研合作计划，希望充分发挥 AI 技术在数据和计算领域的优势，助力全球的可持续发展。2021 年，我们又开启了负责任的人工智能跨学科探索计划，与法学、心理学和社会学等人文科学领域的专家们一起，在打造可靠可信的人工智能技术的同时促进相关人文学科的发展，从而更好地支撑未来社会发展的需求。

Q：如果把 AI 带来的美好未来看成一幅拼图，你认为负责任的 AI 和可持续发展应该在拼图中扮演怎样的角色？

石贝贝：我想负责任的人工智能会是 AI 这幅图画中的基本“构图”之所在，确保 AI 呈现的画面是安全、可靠且稳定的。而可持

续发展应是图中的“色彩”，让 AI 系统的发展在不会对人类、环境和未来造成不可逆转伤害的基础上，丰富、真实地呈现人们心中最美的画面。总之，负责任的人工智能和可持续发展在 AI 美好的未来发展中都扮演着非常重要的角色，只有负责任的 AI 和可持续发展共同实现，AI 给全球社会所带来的美好未来才能真正到来。

Q：你有什么业余爱好，从中又收获了什么？

石贝贝：我最大的爱好是旅行。小时候，我就梦想环游世界，想去看看不同的山河湖海，体验各异的风土人情。从小到大，我身边的旅伴从父母，到朋友、伴侣、再到带着家人和女儿，每次旅行无论是前期准备还是旅程结束后的回味，都是段幸福的经历。

旅行让我收获了很多新知识和新本领，这些收获滋养了我多元的思维模式，让我可以更好地面对生活与工作。

相关阅读

扫描二维码查看文章

科技女性的 N 种可能

放眼科技圈，如今已经有越来越多女性在各自的领域里发光发热。CSDN《开谈》栏目联合《新程序员》杂志、微软公司，邀请到三位来自微软不同领域的代表性人物，进行了一场温暖而有力量的对谈。听听她们作为科技行业从业者的感受与思考，看看科技圈中不同领域的女性发展现状，从她们身上看到当代科技女性的缩影，从当代科技女性的视角探索更多职业发展的可能性。

实习派 | 王一栋：主动就会有故事，高效科研秘诀大公开

在科研中，如果主动向前一步，是否会有不同的故事？来自东京工业大学的王一栋的实习经历给出了肯定的答案。他尝试了“先合作，再实习”的新路径——主动联系导师达成远程科研合作后，正式来到微软亚洲研究院社会计算组实习。在将近一年的实习时间里，他作为核心成员参与了四个科研项目，并在 NeurIPS、ICLR、COLING、ACML 等国际顶会中发表了五篇论文，其中四篇为第一作者。在这里，王一栋确认了自己对科研工作的胜任力，并逐步搭建起科研兴趣驱动的全球研究社群，从实习中收获的友谊甚至助推他做出了回国读博的决定。如此高效产出高质量科研工作的秘诀是什么？他又有哪些在论文发表之外的实习收获？相信你会通过这一期的故事有所启发。

先合作再实习，展开一场师徒之间的双向奔赴

早在正式实习之前，因研究兴趣的契合，经由东京工业大学同组师兄侯汶昕介绍，王一栋就开始以远程的方式与微软亚洲研究院主管研究员王晋东展开科研合作。他们聚焦鲁棒机器学习中的半监督学习，期望用以此提高模型的泛化性。

经过科研探索，他们针对半监督提出了课程伪标签的方法，其能被简单地应用到多个半监督方法上并提升其精度，同时不会引入新的超参数和额外的计算开销。基于此，他们提出了 FlexMatch 这一新算法，在多个图像分类数据集上取得了 SOTA 的效果，相关成果发表于 NeurIPS 2021 且在学术界反响良好。此外，他们还开源了一个统一的基于 Pytorch 的半监督方法库 TorchSSL，公平地实现了诸多流行的半监督方法，方便相关领域进行进一步研究。

“王老师全程悉心指导了这个项目的开发与投稿。经过这个项目的合作，我也对微软亚洲研究院产生了很大的向往。”成果丰硕的科研合作后，王一栋申请来到研究院实习，对之前合作的问题做进一步的探索。

“研究院的 Mentor 都是比较开放包容的，在邮件里表明自己的合作意向，很可能得到积极的回应”，王一栋从“先合作，再实习”的方式中受益良多。王晋东也认为这是对彼此更负责的一种方式，有助于 Mentor 和学生之间的双向了解。通过这次合作，王一栋确认了 Mentor 在方向、性格等方方面面与自己的契合度。与常规模式相比，前期的合作经历使得他正式来这里实习后更加“如鱼得水”，迸发出极大的科研力量。王一栋的硕士导师、东京工业大学的 Takahiro Shinozaki 教授也非常支持王一栋积极向外探索，并持续对他的科研工作进行细致的指导。

与 Mentor 的及时沟通是王一栋不走弯路的一大保障，在他看来，研究院的实习生数量相对较少，每个 Mentor 都是等待每一位实习生挖掘的宝藏般的存在。王晋东重视并尊重学生，不对工作时长做硬性规定，从做什么方向到投什么会议，每个环节都给予学生充分的尊重，持续为王一栋提供着科研想法指导以及实验和写作所需的资源支持。"有了想法之后我不会马上去做，而会寻求王老师的意见，即使是有些幼稚的想法，王老师也会耐心听我讲完并指出其中的不足。王老师也经常把自己的想法讲给我听，让我试着做一些课题。达成一致后再共同规划实验，大大节约了试错成本。" 王一栋说。

受到研究员们严谨的科研态度的感染，王一栋愈发精益求精。在社会计算组组会的报告时间，微软亚洲研究院首席研究员谢幸要求参会成员使用英文交流展示，高度专业且具有全球视野的交流帮助王一栋提升了沟通能力和英文表达能力。同时，在王晋东的影响下，他以严谨的态度杜绝 "虎头蛇尾" 的科研。在完成 FreeMatch 的论文时，师徒一起在一个多月里进行了近十次的大版本改动，"经常是写完一个版本后，我们共同讨论出更好的写作思路，不断大篇幅地修改"。该成果最终发布于 ICLR 2023。

线下实习也使王一栋解锁了 Mentor 在科研之外的另一面。共同讨论问题、吃饭聊天、周末一起在王晋东家打 "双人成行"……在密切的交往中，他们共享着许多快乐记忆，也成为了无话不谈的好朋友。亦师亦友，这是王一栋和王晋东关系的最佳描述。

王一栋（中）与谢幸（左一）、王晋东（右一）

厚积薄发，在问题意识中绘就体系化的研究版图

博观而约取，厚积而薄发。经历了本科阶段对科研的懵懂接触以及硕士阶段对课题的自由探索，王一栋在研究院完成了从科研菜鸟到合格科研工作者的蜕变，逐渐形成体系化的研究版图。

将问题意识贯穿在科研工作的每一个环节是王一栋高效科研的秘诀之一。在 FlexMatch 这项工作的进行过程中，研究者们意识到半监督学习存在着两大痛点：训练时间过长，环境不友好；且阈值的选择对半监督学习算法性能影响很大，但是挑选阈值又很费时费力。

针对以上问题，他们提出了通过在训练期间保持大量和高质量的伪标签来有效利用未标记的数据的方法 SoftMatch，能够根据模型的学习状态以自适应方式调整置信度阈值的模型 FreeMatch，首个将视觉、语言和音频分类任务进行统一的半监督分类学习基准 USB，以及用于细粒度情感分类 TOWE 任务的多粒度半监督算法，从各种角度应对在标签不足的场景中让预训练模型进行高精度适配的挑战。

完成了对半监督学习的阶段性探索后，王一栋又将目光投向长尾视觉识别任务，在分类间隔的视角下提出了间隔校准算法 Margin Calibration，用三行代码解决了长尾不平衡类别分类的问题。目前，关于这些问题的探索仍在持续进行中。

在论文写作中，王一栋通过王晋东传授给他的技巧，积累了 "如何让自己科研想法能够更容易地被审稿人所看重" 的经验。大道至简，在写作时应突出所解决问题的重要性，而非花大量笔墨描述方法的复杂性。"如果之前没有人想过去解决这个问题或者解决得不好，即使所用方法很简单也能脱颖而出。" 王一栋说。

对于合格的科研工作者而言，成熟的心态管理也是学术修炼中不可或缺的重要一环。在微软亚洲研究院，王一栋养成了不因一时得失而情绪起伏的稳定心态。他回忆起 FreeMatch 这篇论文在收获审稿人的一致好评后却最终遭遇拒稿时的复杂心情，在 Mentor 王晋东的开导下，王一栋重拾对工作价值的信心，根据审稿人的意见继续改进后，通过把精力投入其他工作的方式来转移注意力。在反复打磨中，被拒稿的工作也相继收到好消息。

生活中的王一栋

“广结善缘”，优秀工作助推全球研究社群搭建

在微软亚洲研究院所在的这条善缘街上，王一栋持续“广结善缘”，通过自己的优秀工作吸引了来自全球顶尖研究机构的研究者，拓展关系边界并促成稳定合作，逐步搭建起基于共同研究兴趣的科研社群。

论文 Flexmatch 公布后，在学术界受到了许多关注，其开源项目 TorchSSL 在 GitHub 上获得了 1000 多颗星，他也收到了来自领域内其他研究者的许多邮件。在与不同人的交流碰撞中，王一栋受到诸多启发，也以此为契机与这些遍布全球的优秀研究者们展开进一步的科研合作。除了王一栋、王晋东和谢幸，这些工作还融淬了卡内基梅隆大学的陈皓、马克斯·普朗克信息学研究所的范越、北卡罗来纳大学的衡强、东京工业大学的吴昊，以及南京大学的吴震、西湖大学的张岳等研究者的成果。“特别感谢他们，我的进步离不开他们的帮助。”王一栋说。

在研究院开放包容的环境中，王晋东乐见这样的合作，并发挥牵线搭桥的作用，帮助王一栋整理思路以更好地与全球研究人员分享，同时更好地理解合作者的反馈以修正研究思路。在齐心合作下，一篇篇优秀的论文在头脑风暴、互信互补中相继诞生。

除了通过研究工作吸引志趣相投的合作者，王一栋也主动联系敬仰的老师交流学习，科研的每一步都充满主动性。萌生出把半监督学习应用到自然语言处理场景中的想法后，他又前往西湖大学学习交流。无论是硕士导师、东京工业大学的 Takahiro Shinozaki 教授，还是西湖大学的张岳教授，都给予了王一栋详尽的指导，并纷纷成为论文合作者。一个由研究兴趣驱动，集结了具有不同背景、来自不同文化的全球研究者的科研社群逐渐成型。

在微软亚洲研究院，王一栋还收获了许多好友。在劳逸结合的工作节奏中，他们的友谊在一起吃饭、看电影、打台球的过程中不断加深。今年 9 月，他即将进入北京大学攻读博士学位，“决定回国读博最主要的契机就是在研究院认识了许多好朋友”，他们的友谊也将继续在北京延续。对于性格开朗、交友广泛的王一栋来说，这些弥足珍贵的情感甚至在无形中影响着他的人生航向。

王一栋时常将自己取得的成绩归结为“运气好”，并对一路走来为这份运气加成的恩师与挚友心怀感恩。在微软亚洲研究院，他在科研压力并不大的状态中按部就班地工作，最终得以“守得云开见月明”。事实上，好运气的背后更是主动出击的行动、持之以恒的努力、独立敏锐的思考、乐观稳定的心态、以及开放共赢的合作精神。研究院之于他像是“梦开始的地方”，收获满满的实习使他对这里充满归属感，“未来有机会的话非常愿意再回来实习或工作。”王一栋说。

王一栋（右三）与伙伴们

Mentor 寄语

王晋东
微软亚洲研究院
主管研究员

王一栋同学在实习期间深耕鲁棒机器学习，特别是半监督学习领域。他在理论、算法、应用和开源框架方面均取得了卓越的进展，并将研究成果发表于国际顶尖学术会议 ICLR、NeurIPS、COLING 和 ACML 中。这些工作成功解决了在标签不足的场景中如何让预训练模型进行高精度适配的挑战，并使得微软亚洲研究院在该领域拥有了国际领先的研究成果。

王一栋同学思维敏捷、行动力强，并且具有很强的团队合作精神。他善于团结不同背景的研究人员，通过广泛合作的方式做出更具影响力的工作。我对他的科研潜力深感信心，相信他未来会有更大的发展，并希望他能在其他领域做出更多成果。

关于内卷与反内卷、建立学术社交网络，听听过来人的建议

扫描二维码查看文章

这次开学，我们请来了 ChatGPT 和各位前辈指点迷津

扫描二维码查看文章

对话 | 为“冷门绝学”甲骨文研究插上科技之翼

近期，微软亚洲研究院主管研究员武智融与首都师范大学甲骨文研究中心莫伯峰教授团队合作，为人工智能“入驻”甲骨文研究领域成功打开了一扇大门——基于自监督学习技术构建的甲骨文校重助手 Diviner 大幅提升了甲骨文校重工作的效率，并获得了学术界的肯定与认可。这项合作成果让人工智能技术在甲骨文这个“冷门绝学”中找到了用武之地，也为历史传承与文化遗产保护插上了科技之翼。

甲骨文研究的意义是什么？计算机领域研究员与甲骨文研究者，对于人工智能与甲骨文的结合有怎样不同的看法？人工智能技术成功落地甲骨文研究，将对甲骨文等历史文化研究界产生怎样的影响？下面让我们通过武智融博士与莫伯峰教授的对话，了解这一跨界科研成果背后的故事，以及未来的发展前景。

武智融：关于甲骨文的研究，我们这些领域的门外汉只是粗浅地知道甲骨文研究是一项非常有意义的工作。所以先请您给我们科普一下甲骨文发现的历史，以及现存的甲骨收藏情况。

莫伯峰：甲骨文的发现者叫王懿荣，他是晚清一位著名的金石学家。1899 年，这位金石学家身患疾病，他在购买的中药里发现一味名为“龙骨”的药材上刻画着奇怪的符号，对文字颇有研究的他意识到，这可能是早期留下的文字，自此展开了对甲骨的购藏，这便是广为流传殷墟甲骨文被发现的故事。

虽然甲骨文被发现是在十九世纪，但直到 1928 年，殷墟甲骨的科学挖掘才正式起步。此前，殷墟甲骨曾经历了一段漫长的私人挖掘和倒卖时期。因此，甲骨流散到了很多地方，不仅国内有很多公私藏家，海外也收藏有大量甲骨，比如日本、加拿大、英国、美国、德国、俄罗斯、新加坡、瑞典、瑞士、法国等国家都收藏有甲骨。当然，最大宗的收藏还是在中国，像中国国家图书馆、台北中研院和故宫博物院，所藏的甲骨都达到了数万件。

以上是甲骨实物的收藏情况，实际上甲骨文研究通常并不直接接触这些实物，而是利用它们的图像资源，主要是甲骨拓本。相较于甲骨实物的收藏情况，甲骨拓本的情况更为复杂。伴随着甲骨在不同藏地间流转，同一片甲骨有时会流下很多拓本，拓本总数远大于甲骨总数。因此，对甲骨文图像资料进行整理就变得十分重要。比如由郭沫若任主编、胡厚宣任总编辑的《甲骨文合集》就是系统整理甲骨图像资料的一部集成。这是一项非常困难的工作，中国社科院历史研究所专门成立了一个“《甲骨文合集》编辑工作组”，从 1959 年开始直到 1982 年才完成。之后编辑出版的《甲骨文合集补编》，是对《甲骨文合集》的补充，也同样耗时很久。

武智融：那我们研究甲骨文的重要意义是什么？关于甲骨文研究都有哪些具体的内容？您的主要研究方向是什么？

莫伯峰：甲骨文是中华文化的瑰宝，正是因为甲骨文的发现，中国信史被向上推进了约一千年。同时，它也是现代汉字的源头，是我国已知最早的成系统文字。汉字是世界上唯一从古代一直沿用至今的文字，甲骨文又是这个“唯一”的源头，这就决定了它无可取代的地位和独一无二的价值。作为人类共同的宝贵财富，甲骨文已被列入《世界记忆遗产名录》，甲骨文发现地河南安阳殷墟也被列入了《世界文化遗产名录》。由于甲骨文的时代离今天已经三千多年了，甲骨文所传递的信息已经笼上了一层迷雾，我们只有通过研究才能把时间附着其上的迷雾去掉，解开这些谜团。

甲骨学已经成为一门综合性学问，历史学家、语言文字学家、考古学家都能从甲骨文中汲取“营养”。对于历史学家而言，甲骨是商代的史料；对于语言文字学家而言，甲骨是宝贵的商代语料；对于考古学家而言，甲骨是古代人类的文化遗存。当然，甲骨文还可以与其他各种学科相结合，比如利用甲骨文中的日食、月食、流星雨的记录，与天文学相结合；利用古生物学知识，又可以为龟甲、牛骨这些原料的来源提供线索。追索各种事物的早期历史，很多都能从甲骨文中找到痕迹。

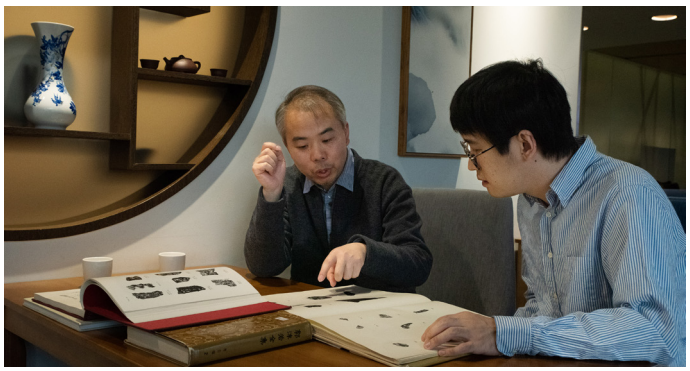
从研究的角度来看，史料和语料研究有着密切的关系，如果不识字就无法将其当作史料进行研究，反之如果对商朝历史变迁不了解，那么在认字和语言理解上也会存在困难。所以我们在研究中这两方面都会涉及。尽管在外人看来这些研究晦涩难懂，但当兴趣使然深入其中时，你会发现它其乐无穷。将甲骨文字与后代文字进行关联就好像是一个探案、解谜的过程。

武智融：在我和您取得联系，提出想将人工智能技术应用于甲骨文研究之前，您已经做过一些相关的尝试了，您为什么会考虑将人工智能引入到甲骨文研究中？

莫伯峰：这些年，人工智能取得的进展有目共睹，它的很多研究领域都与古文字研究非常契合，比如计算机视觉所处理的问题与图形有关，甲骨文字也是一种图形，两者存在共同之处；再比如机器翻译、语音识别是语言处理任务，甲骨文是记录商代语言的书面符号，本质上也就是语言问题；还有知识图谱的应用，以形式化的方式表示知识结构，同样非常有利用于应用在古文字研究这种专业化非常强的领域。综合多方面的考量，我认为二者结合大有可为。

武智融：而且从计算机领域的角度来看，利用人工智能技术探索甲骨文研究也将对通用人工智能的研究产生有意义的影响。目前，人工智能的落地应用以有监督的训练为主。随着近些年自监督学习的巨大进展，大大拓宽了人工智能的落地场景。在甲骨文领域，由于人类的知识是有限的，能够提供的标签和监督也是极其有限的，这成为了不得不应用自监督学习才可以解决的问题。此外，甲骨文与多模态识别也紧密相关，因为甲骨文本身就是多模态的数据，它的文字本身既是语言也是图形，相当于一种实体上存在两种表现形式。

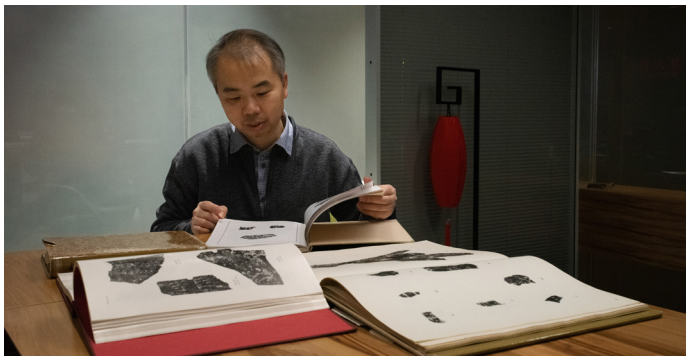
莫伯峰：的确是这样，要走向通用人工智能，人工智能需要面向一些探索性的问题，获得一些发明、发现。如果只是利用人工智能进行甲骨文的文字识别，那么人类认知将会是机器的天花板，因为大部分机器学习都是有监督训练的，可能最后的结果只是开发出一些学习甲骨文字之类的小程序，以提高公众对甲骨文的认识，但这对甲骨文研究的意义非常有限。我们更希望使用人工智能技术对甲骨文研究有直接、具体的推动作用，哪怕是一小点的进步，在甲骨文研究中也会是创新性的成果。



武智融：那么您认为 Diviner 模型在甲骨文校重工作中发挥了怎样的作用？促进了哪些新成果的出现？

莫伯峰：从我自身参与过的甲骨校重工作经历来看，纯靠人工来做那是相当痛苦，数据量太大了。甲骨校重作为甲骨文研究的基础性题目，引入人工智能技术的意义至少有两方面：一是面向过去。那些已经发表的甲骨拓本都做过人工校重，Diviner 在人工已经做过的工作上还能发现一批新成果，这十分不容易，也非常有价值。没有 Diviner 的介入，这些成果至少不会那么迅速地被获得。二是面向未来。现在还有很大数量的甲骨拓本没有发表出来，未来针对这些甲骨拓本的校重工作，利用 Diviner 模型将会节省大量的人力，并促使整个甲骨校重的工作模式发生改变。

在这次甲骨文校重工作的初步尝试中，我们仅花一周时间就利用 Diviner 模型和部分数据，校对出了 200 多组重片。这只是一个初步的结果，数据还没有全部整理完，最后校对出的成果肯定将远远超出这一数量。

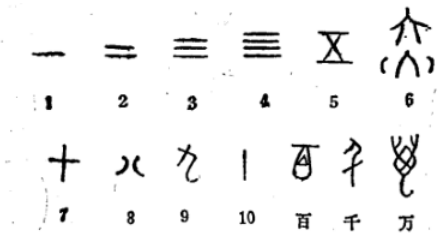


武智融：这次合作的研究成果有些是单纯的校重；有些校重成果帮助了甲骨缀合；还有的帮助到一些不同时期、不同完整度的重片实现拼接，让信息更完整；更有两个拓片内容几乎一样，但却不是重片的情况。那么，这些拓片在历史上是怎么形成的？是原来完整甲骨脱落分裂成两半，之后又被别人拼接起另一半的么？还是有其它的原因？

莫伯峰：这正体现了自监督学习在研究性课题中的独特价值，Diviner 模型发现的结果有些是我们事先都没有预计到的，所以特别具有创新性。如果是有监督学习，我们肯定提供不了那种情况的样本。像你提到的校重推动缀合、不同完整度重片拼接、近似度极高的“伪重片”问题，都是模型结果跑出来以后我们才注意到的。这些情况的出现，基本上都与甲骨流传和甲骨拓本制作过程中的一些特殊情况有关。比如不同完整度重片的现象，就是由于甲骨传拓方法不一致导致的：早期的拓本有些并不是把所有甲骨骨面全部拓全，而是只拓印有文字的部分，所以拓本可能很小，但它原来完整的甲骨可能是很大的一块。而后期拓本制作时，虽然会将整个骨面都完整拓印，但有可能甲骨本身已经破碎不全了，所以拓本也不会完整。这样就出现了重片之间部分重合，部分不重合的现象。

武智融：而且即使两片拓片差异不大，查重也并不是一项简单的工作，因为时间的流逝加上经过多次流转，甲骨表面的物理情况、纹理都会发生变化，所以即使是完全一样的甲骨，在不同时期拓出来的样子也会很不一样。

莫伯峰：是的，在机器判定为重片后，这些结果还需要研究者再进行检验。甲骨学是一个系统性的科学，涉及方方面面，解决一个问题需要从不同维度综合考虑。虽然在机器看来，重片问题就是相似度匹配，但研究者还会从其它维度来综合考虑。比如甲骨文的数字，一是一横，二是两横，三是三横，四是四横，对计算机而言，三与四只有一横之差，相似度非常高，但从人的认知来看是完全不一样的。我们这次还遇到过一个极端特殊的例子——曾有一批甲骨遭受过火灾，导致了形态发生了很大变化，火灾前后的拓本差异很大，我们研究者了解背景，所以能排除这种干扰进行校重，而机器就非常困难了。所以说，研究者可以从多维度去验证，这就展现了人机协作的重要性。



甲骨文的数字符号

武智融：这里我再概括介绍一下 Diviner 模型的技术实现。在传统的计算机视觉研究中，现有的对比模型和技术主要用于几乎完全相同的图片上，在甲骨文问题中，不同的重片在外观上可能有很大的差异。因为一块完整的甲骨可能会碎裂成多片，重片需要从大骨片中找出小骨片。因此，基于全局外观表示的传统对比模型将不再起作用。Diviner 模型是一个基于自监督学习的高级人工智能模型，可从局部寻找匹配关系，再拓展到全局，为甲骨文构建标准化的数据库提供了新的可能。

而且因为自监督学习的匹配算法，Diviner 模型具有强大的泛化能力。模型可以通过图像增强技术模拟同一块甲骨在不同时期制作成拓片或者因年深日久造成的图像变化，例如磨损、模糊等。在大规模无标注数据上获取密集的自我监督，远远比稀疏的、基于整体的人工监督更有效。同时，Diviner 模型还能够精确地预测出重片之间点对点的对应关系，将重片拼合或拼接在一起。

武智融：通过这次合作，以及 Diviner 模型在甲骨文校重工作中的一些新发现，您认为人工智能技术对未来的甲骨文研究工作有怎样的意义和影响？

莫伯峰：我们这次的合作项目可以说是人机协作的典型案例，人和机器各自发挥了自己的特点和优势，所以解决了一些过去解决不好的问题。通过这次合作让我们对人工智能与甲骨文的跨领域研究更有信心。甲骨文或者说古文字方面的研究还有各种各样的课题，我希望通过人和机器的协作，未来让我们能够对这些问题逐一进行探索，取得更多的进展。

另外，人工智能技术的应用也让我们对一些古文字问题有了新的思考。因为在计算机没有进入这个领域之前，我们对很多问题的思考都是只从人类的角度出发的。计算机的加入会给我们一些新的启示，就是每当机器出现和人不一样的结果时，都给我们提供了一个重新思考的契机——为什么机器会得出这样的结论？它与我们人类答案的差别背后反映了什么实质问题？通过审视我们和机器的不同，也为我们重新思考某些学术问题，提供了非常好的视角。

与微软亚洲研究院的这次合作，虽然只是甲骨文和人工智能交叉研究的一个小序幕，但也给人工智能“进驻”甲骨文研究领域推开了一扇门，能起到一定的示范意义。我相信，未来人工智能和古文字的交叉研究将有广阔的发展前景。

带你读论文 | 了解 AIGC 音频 / 图像数据生成，这几篇论文给你划好了重点

作者：微软亚洲研究院高级研究员谭旭

作为近期人工智能领域内的顶流之一，AIGC（AI-Generated Content 或 Generative AI）早已火爆出圈，频登各大互联网平台热搜。基于深度学习的内容生成在图像、视频、语音、音乐、文本等生成领域都取得了令人瞩目的成果。

由于现实世界中的信息在多数情况下呈现文本、图像和语音等多种模态，人类会通过综合运用多种感官来感知和理解现实世界，因此，如何赋予计算机这种综合理解多种模态的能力也成为了学术界的研究热点。

与文本生成更加关注抽象语义不同，声音和视觉模态还需要生成更多的细节信息。所以，声音和视觉内容（语音、音效、音乐、图像、视频等）的生成面临着一系列挑战：如何刻画声音视觉内容中复杂且高频的数据分布；如何建模生成过程中的一对多映射问题；如何利用大规模无标注数据解决数据稀疏性问题；在基于其它模态生成时，如何解决跨模态对齐问题等。

本文送上一个可以击破 AIGC 数据生成中这些难题的论文锦囊，希望大家可以在入坑 AIGC 领域之初能有所启发。

学习范式（Learning Paradigm）—— 高屋建瓴

一个好的学习范式能为研究者在探索复杂的深度学习问题时，指导设计方法和模型。在传统的数据理解任务中，深度学习先驱 Yoshua Bengio 等人倡导的表征学习 Representation Learning 非常值得参考。表征学习可以指导深度学习模型提高学习数据表征的能力，以增强对数据的理解。

相关论文：

[1] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. https://arxiv.org/abs/1206.5538

而在 AIGC 的数据生成任务中，微软研究院的研究员们同 Yoshua Bengio 提出的 Regeneration Learning 的学习范式能为各个数据生成任务提供指导。它将复杂的带条件的数据生成任务 X—>Y 分解成两个阶段，X—>Y′ 和 Y′—>Y，其中 X 是条件信息，Y 是目标数据，而 Y′ 是 Y 的抽象表征，通过自监督的方法比如自编码器学到。

Regeneration Leaming 有几个好处：1) X—>Y′ 相比于 X—>Y 的一对多映射和虚假映射问题会大大减轻；2) 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。

相关论文：

[2] Tan, X., Qin, T., Bian, J., Liu, T. Y., & Bengio, Y. (2023). Regeneration Learning: A Learning Paradigm for Data Generation. arXiv preprint arXiv:2301.08846. https://arxiv.org/abs/2301.08846

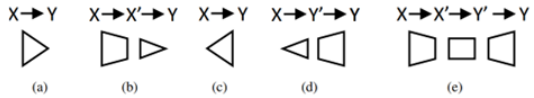


Figure (a): Presentation ($X \rightarrow$). Figure (b): Representation ($X \rightarrow X' \rightarrow$). Figure (c): Generation ($\rightarrow Y$). Figure (d): Regeneration ($\rightarrow Y' \rightarrow Y$). Figure (e): Representation + Regeneration ($X \rightarrow X' \rightarrow Y' \rightarrow Y$).

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 1：Regeneration Learning 和 Representation Learning 的对比

编解码器 (Codec)——化繁为简

声音和视觉内容（语音、音效、音乐、图像、视频等）往往含有复杂的高频细节信息，因此科研人员们利用 Codec（编解码器）等方法，将承载高频细节的声音和视觉内容转化为抽象紧致的表征（离散 Token 或者连续向量），以降低后续数据生成的难度。相关论文，包括图像里的 Codec [3][4][5] 以及声音里的 Codec [6]。

论文 [3] 是较早的一篇将连续图像音频数据通过 VQ-VAE（向量量化自编码器）转成离散 Token 的工作，而后续论文 [4] 将 VQ-VAE 和 GAN 结合进一步提升效果。

相关论文：

[3] Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1711.00937

[4] Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12873-12883). https://arxiv.org/abs/2012.09841

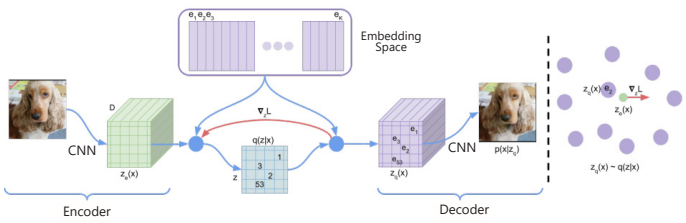


图 2：VQ-VAE

论文 [5] 是文本到图像生成大火的 Stable Diffusion，和 VQ-VAE 和 VQ-GAN 不同的是，它更加偏向利用 VAE 将图像转为连续向量形式的抽象表征。

相关论文：

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695). https://arxiv.org/abs/2112.10752

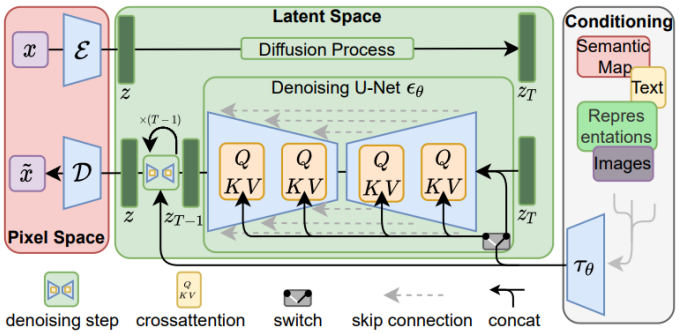


图 3：Stable Diffusion

论文 [6] 则利用 VQ-VAE 将语音波形转成离散 Token，为了增加重建质量，它采用了 Residual Vector Quantizers（残差向量量化器）将一帧语音量化成多个残差 Token。

相关论文：

[6] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., & Tagliasacchi, M. (2021). SoundStream: An end-to-end neural audio codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 495-507. https://arxiv.org/abs/2107.03312

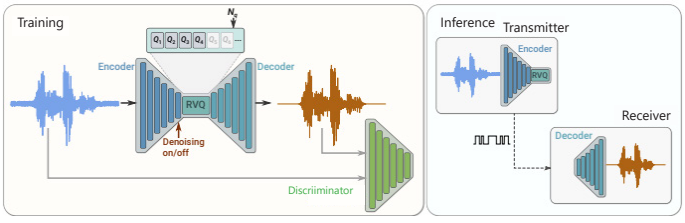


图 4：SoundStream

生成模型（Generative Model）——无中生有

强大的生成模型能细致精准地刻画数据中的复杂分布，让模型能更好地从学习到的分布中采样，以实现数据的从无到有生成。

在当前流行的数据生成模型中，文本生成 GPT 系列比如 GPT 1/2/3 以及 ChatGPT 采用的是 Transformer 自回归模型，而在图像和音频生成中，有些采用的是扩散模型（比如 DALL-E 2，Imagen，Stable Diffusion，及 DiffWave/ WaveGrad/ GradTTS），也有些采用的是自回归模型（比如 DALL-E，Parti，AudioLM）。关于各种生成模型的比较分析，可参考文章 https://zhuanlan.zhihu.com/p/591881660。

以下论文总结了典型的生成模型，包括变分自编码器 VAE [7]，生成对抗网络 GAN [8]，标准化流 Flow [9]，扩散模型 Diffusion [10][11]，以及自回归模型 AR [12]。

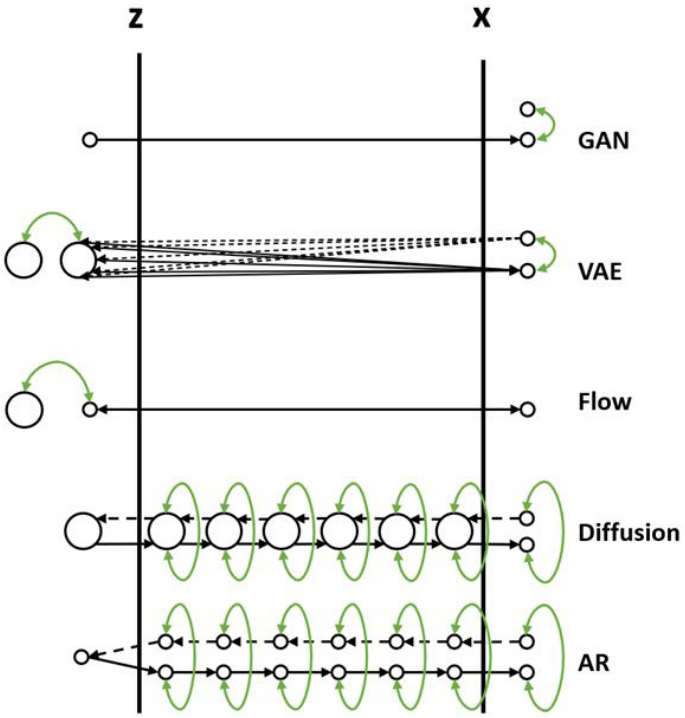


图 5：生成模型

相关论文：

[7] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. https://arxiv.org/abs/1312.6114

[8] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial networks. Advances in Neural Information Processing Systems. https://arxiv.org/abs/1406.2661

[9] Dinh, L., Krueger, D., & Bengio, Y. (2014). Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516.
https://arxiv.org/abs/1410.8516

[10] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning (pp. 2256-2265). PMLR.
https://arxiv.org/abs/1503.03585

[11] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.
https://arxiv.org/abs/2006.11239

[12] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
https://arxiv.org/abs/2005.14165

跨模态对齐（Cross-Modal Alignment）——牵线搭桥

当利用条件信息作为输入来生成数据的时候，条件信息往往和生成数据的模态不一致。因此需要一个跨模态对齐模型来拉近两个模态之间的关系。

文本到图像生成模型 DALL-E 2 [13]，通过文本 - 图像对齐模型 CLIP [14] 来拉近图文之间的距离；文本到音乐音频生成模型 MusicLM [15]，则通过文本 - 音乐音频对齐模型 MuLan [16] 来拉近音乐和文字之间的距离。

通过利用对齐模型将输入模态转为共享的表征作为生成模型的条件输入，可大大降低生成模型处理不同模态输入的成本，使其专注于数据生成，提高生成效果。下列论文采集了 DALL-E 2 都在用的文本 - 图像对齐模型 CLIP [14] 以及 MusicLM 在用的文本 - 音频对齐模型 MuLan [16]，这些方法值得一试。

相关论文：

[13] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.
https://arxiv.org/abs/2204.06125

[14] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.
https://arxiv.org/abs/2103.00020

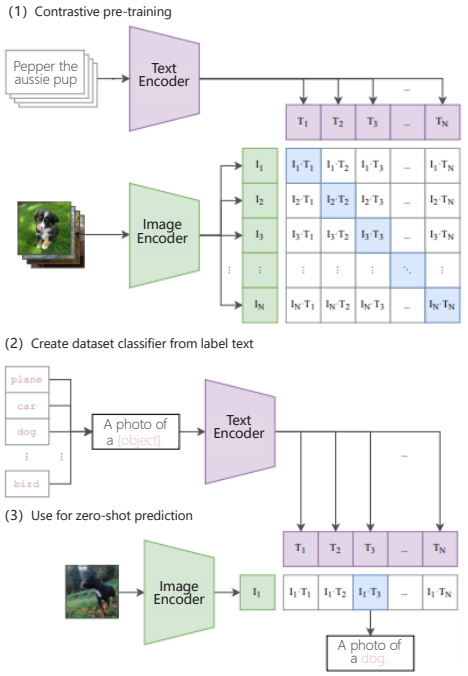


图 6：CLIP

相关论文：

[15] Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., ... & Frank, C. (2023). Musiclm: Generating music from text. arXiv preprint arXiv:2301.11325.
https://arxiv.org/abs/2301.11325

[16] Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., & Ellis, D. P. (2022). MuLan: A joint embedding of music audio and natural language. arXiv preprint arXiv:2208.12415.
https://arxiv.org/abs/2208.12415

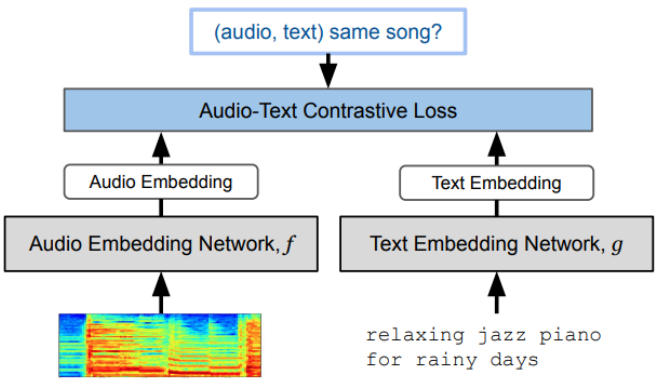


图 7：MuLan

机器之心 | 微软多模态 ChatGPT 来了？ 16 亿参数搞定看图答题、智商测验等任务

在 NLP 领域，大规模语言模型（LLM）已经成功地在各种自然语言任务中充当通用接口。只要我们能够将输入和输出转换为文本，就能使得基于 LLM 的接口完成一个任务。例如，对于摘要任务，将文档输入到语言模型，语言模型就可以生成摘要。

尽管 LLM 在 NLP 任务中取得了成功的应用，但研究人员仍努力将其原生地用于图像和音频等多模态数据。作为智能的基本组成部分，多模态感知是实现通用人工智能的必要条件，无论是对于知识获取还是与现实世界打交道。更重要的是，解锁多模态输入能够极大地拓展语言模型在更多高价值领域的应用，比如多模态机器人、文档智能和机器人技术。

因此，微软亚洲研究院在论文《Language Is Not All You Need: Aligning Perception with Language Models》中介绍了一个多模态大规模语言模型（MLLM）——KOSMOS-1，它可以感知一般模态、遵循指令（即零样本学习）以及在上下文中学习（即少样本学习）。研究目标是使感知与 LLM 保持一致，如此一来模型能够看到（see）和说话（talk）。研究员们按照 METALM（参见论文《Language models are general-purpose interfaces》）的方式从头开始训练 KOSMOS-1。

研究员们将一个基于 Transformer 的语言模型作为通用接口，并将其与感知模块对接。他们在大规模多模态语料库上训练模型，语料库包括了文本数据、任意交错的图像和文本、以及图像描述数据。此外，研究员们还通过传输纯语言数据来校准跨模态的指令遵循能力。

Dataset	Task description	Metric	Zero-shot	Few-shot
Language tasks				
StoryCloze [MRL+17]	Commonsense reasoning	Accuracy	✓	✓
HellaSwag [ZHB+19]	Commonsense NLI	Accuracy	✓	✓
Winograd [LDM12a]	Word ambiguity	Accuracy	✓	✓
Winogrande [SBBC20]	Word ambiguity	Accuracy	✓	✓
PIQA [BZB+20]	Physical commonsense	Accuracy	✓	✓
BoolQ [CLC+19]	Question answering	Accuracy	✓	✓
CB [dMST19]	Textual entailment	Accuracy	✓	✓
COPA [RBG11]	Causal reasoning	Accuracy	✓	✓
Rendered SST-2 [RKH+21]	OCR-free sentiment classification	Accuracy	✓	
HatefulMemes [KFM+20]	OCR-free meme classification	ROC AUC	✓	
Cross-modal transfer				
RelativeSize [BHCF16]	Commonsense reasoning (object size)	Accuracy	✓	
MemoryColor [NHJ21]	Commonsense reasoning (object color)	Accuracy	✓	
ColorTerms [BBBT12]	Commonsense reasoning (object color)	Accuracy	✓	
Nonverbal reasoning tasks				
IQ Test	Raven's Progressive Matrices	Accuracy	✓	
Perception-language tasks				
COCO Caption [LMB+14]	Image captioning	CIDEr, etc.	✓	✓
Flicker30k [YLHH14]	Image captioning	CIDEr, etc.	✓	✓
VQAv2 [GKSS+17]	Visual question answering	VQA acc.	✓	✓
VizWiz [GLS+18]	Visual question answering	VQA acc.	✓	✓
WebSRC [CZC+21]	Web page question answering	F1 score	✓	
Vision tasks				
ImageNet [DDS+09]	Zero-shot image classification	Top-1 acc.	✓	
CUB [WBW+11]	Zero-shot image classification with descriptions	Accuracy	✓	

表 1：在零样本和少样本下测试了 KOSMOS-1 在自然语言、多模态和视觉任务的能力。

最终，KOSMOS-1 模型原生支持零样本和少样本学习设置下的语言、感知语言与视觉任务，具体如上表 1 所示。

研究员们在下图 1 和图 2 中展示了一些生成示例。除了各种自然语言任务，KOSMOS-1 模型能够原生处理广泛的感知密集型任务，如视觉对话、视觉解释、视觉问答、图像描述生成、简单的数学公式、OCR 和带描述的零样本图像分类。他们还根据瑞文推理测验（Raven's Progressive Matrices, RPM）建立了一个 IQ 测试基准，用来评估 MLLM 的非语言推理能力。

图 1：KOSMOS-1 生成的示例（蓝色框是输入提示，粉色框是 KOSMOS-1 输出）。(1)-(2) 图像描述生成，(3)-(6) 视觉问答，(7)-(8)OCR，(9)-(11) 基于视觉的对话。

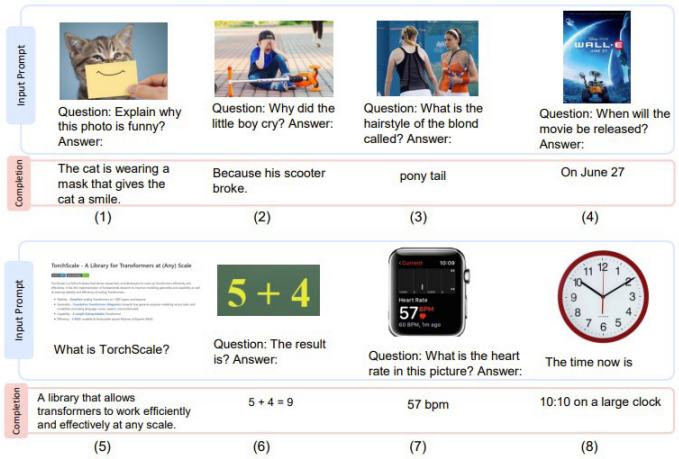


图 2：KOSMOS-1 生成的示例（蓝色框是输入提示，粉色框是 KOSMOS-1 输出）。(1)-(2) 视觉解释，(3)-(4) 视觉问答，(5) 基于网页的问答，(6) 简单的数学公式，(7)-(8) 数字识别。

这些示例表明，多模态感知的原生支持为将 LLM 应用于新任务提供了新的机遇。此外与 LLM 相比，MLLM 实现了更好的常识推理性能，表明了跨模态迁移有助于知识获取。

由于 KOSMOS-1 模型的参数量为 16 亿，因此有网友表示有望在自己的电脑上运行这个多模态大模型。

KOSMOS-1：一个多模态大型语言模型

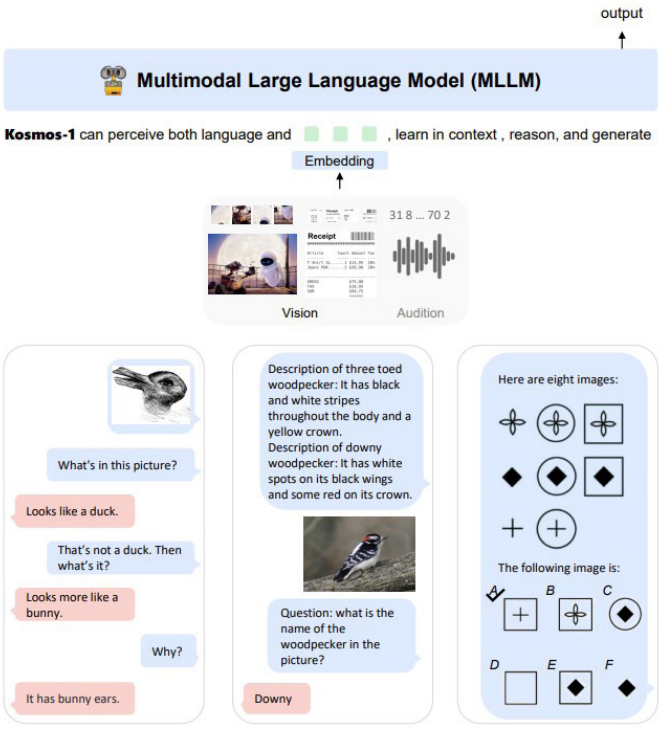


图 3：KOSMOS-1 是一个多模态大规模语言模型，KOSMOS-1 能够感知多模态输入、遵循自然语言指令、在语言任务和多模态任务中完成上下文学习。本工作将视觉信息与大规模语言模型对齐，推进从大规模语言模型到多模态大规模语言模型。

如图 3 所示，KOSMOS-1 是一个多模态语言模型，它既可以感知一般的模态、遵循指令、还能在上下文中学习并生成输出。具体来说，KOSMOS-1 的主干是一个基于 Transformer 的因果语言模型。除了文本之外，其他模态也能被嵌入并输入到该模型中，如下图中，除了语言还有视觉、语音等的嵌入。Transformer 解码器用作多模态输入的通用接口。一旦模型训练完成，KOSMOS-1 在零样本和少样本中也能对语言任务和多模态任务进行评估。

Transformer 解码器以统一的方式感知模态，输入信息会被 flatten 为带有特殊 token 的序列。例如 <s> 表示序列开始、</s> 表示序列结束。特殊 token <image> 和 </image> 表示编码图像嵌入的开始和结束。

Datasets	Format Examples
Text	<s> KOSMOS-1 can perceive multimodal input, learn in context, and generate output. </s>
Image-Caption	<s> <image> Image Embedding </image> WALL-E giving potted plant to EVE. </s>
Multimodal	<s> <image> Image Embedding </image> This is WALL-E. <image> Image Embedding </image> This is EVE. </s>

表 2：KOSMOS-1 训练阶段使用的数据格式

嵌入模块将文本 token 和其他输入模态编码成向量表示，对于输入 token，该研究使用查找表将其映射到嵌入中。对于连续信号模态（例如图像和音频），也可以将输入表示为离散编码。

之后，获得的输入序列嵌入会被馈送到基于 Transformer 的解码器。然后因果模型以一种自回归的方式处理序列，从而产生下一个 token。总而言之，MLLM 框架可以灵活地处理各种数据类型，只要将输入表示为向量即可。

模型训练

首先是训练数据集。数据集包括文本语料库、图像 - 字幕对、图像和文本交叉数据集。具体而言，文本语料库包括 The Pile 、Common Crawl (CC)；图像 - 字幕对包括 English LAION-2B、LAION-400M、COYO-700M 以及 Conceptual Captions；图像和文本交叉多模态数据集来自 Common Crawl snapshot。

Hyperparameters	
Number of layers	24
Hidden size	2,048
FFN inner hidden size	8,192
Attention heads	32
Dropout	0.1
Attention dropout	0.1
Activation function	GeLU [HG16]
Vocabulary size	64,007
Soft tokens V size	64
Max length	2,048
Relative position embedding	xPos [SDP+22]
Initialization	Magneto [WMH+22]

表 3：KOSMOS-1 训练阶段使用的模型参数

数据集有了，然后是训练设置。MLLM 组件包含 24 层、隐藏维度是 2048、8192 个 FFN 和 32 个注意力头、参数量为 1.3B。为了使模型更好的收敛，图像表示是从具有 1024 个特征维度的预训练 CLIP ViT-L/14 模型获得的。图像在训练过程中被预处理为 224×224 分辨率，此外，训练期间除了最后一层，所有的 CLIP 模型参数被冻结。KOSMOS-1 的参数总数约为 1.6B。

实验结果

该研究进行了一系列丰富的实验来评价 KOSMOS-1：语言任务（语言理解、语言生成、 OCR-free 文本分类）；跨模态迁移（常识推理）；非语言推理（ IQ 测试）；感知 - 语言任务（图像描述生成、视觉问答、网页问答）；视觉任务（零样本图像分类、带有描述的零样本图像分类）。

图像描述生成。下表给出了不同模型在 COCO 和 Flickr30k 上的零样本性能。相比其他模型，KOSMOS-1 均取得了显著效果，甚至在参数量远小于 Flamingo 的基础上，性能也不错。

Model	COCO		Flickr30k	
	CIDEr	SPICE	CIDEr	SPICE
ZeroCap	14.6	5.5	-	-
VLKD	58.3	13.4	-	-
FewVLM	-	-	31.0	10.0
METALM	82.2	15.7	43.4	11.7
Flamingo-3B*	73.0	-	60.6	-
Flamingo-9B*	79.4	-	61.5	-
KOSMOS-1 (1.6B)	**84.7**	**16.8**	**67.1**	**14.5**

表 4：KOSMOS-1 在图像描述生成任务（COCO 和 Flickr30k）中的零样本测试结果

下表为少样本性能对比：

Model	COCO			Flickr30k		
	$k=2$	$k=4$	$k=8$	$k=2$	$k=4$	$k=8$
Flamingo-3B	-	85.0	90.6	-	72.0	71.7
Flamingo-9B	-	93.1	**99.0**	-	72.6	**73.4**
KOSMOS-1 (1.6B)	**99.6**	**101.7**	96.7	**70.0**	**75.3**	68.0

表 5：KOSMOS-1 在图像描述生成任务（COCO 和 Flickr30k）中的少样本测试结果

视觉问答。KOSMOS-1 比 Flamingo-3B 和 Flamingo-9B 模型具有更高的准确率和鲁棒性：

Model	VQAv2	VizWiz
Frozen	29.5	-
VLKDViT-B/16	38.6	-
METALM	41.1	-
Flamingo-3B*	49.2	28.9
Flamingo-9B*	**51.8**	28.8
KOSMOS-1 (1.6B)	51.0	**29.2**

表 6：KOSMOS-1 在视觉问答任务（VQAv2 和 VizWiz）中的零样本测试结果

下表为少样本性能对比：

Model	VQAv2			VizWiz		
	$k=2$	$k=4$	$k=8$	$k=2$	$k=4$	$k=8$
Frozen	-	38.2	-	-	-	-
METALM	-	45.3	-	-	-	-
Flamingo-3B	-	53.2	55.4	-	34.4	38.4
Flamingo-9B	-	**56.3**	**58.0**	-	34.9	**39.4**
KOSMOS-1 (1.6B)	**51.4**	51.8	51.4	**31.4**	**35.3**	39.0

表 7：KOSMOS-1 在视觉问答任务（VQAv2 和 VizWiz）中的少样本测试结果

IQ 测试。瑞文推理测验是评估非语言推理最常见的测试之一。图 4 显示了一个示例。

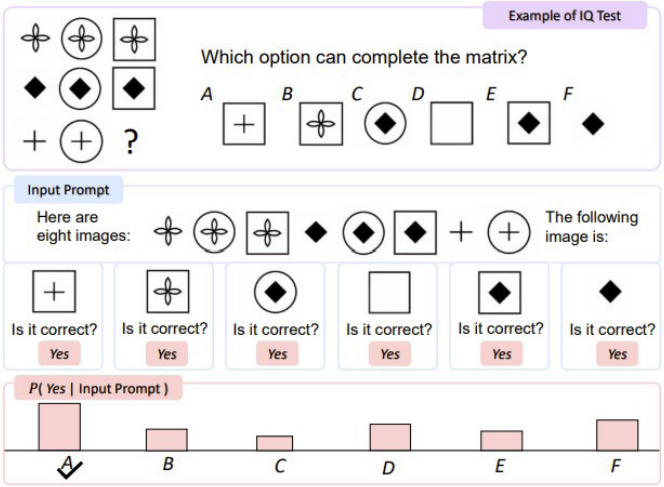


图 4：上图为 Raven IQ 测试的一个例子，下图为 KOSMOS-1 如何完成 IQ 测试。提示包括 IQ 测试输入图像和自然语言指令。我们将每个候选图像分别附加到输入提示上，并询问模型是否正确。选择预测 Yes 概率最高的候选图像作为最终答案。

表 8 显示了在 IQ 测试数据集上的评估结果。KOSMOS-1 能够在非语言环境中感知抽象概念模式，然后在多个选择中推理出之后的元素。据了解，这是首次有模型可以执行此类零样本 Raven IQ 测试。

Method	Accuracy
Random Choice	17%
KOSMOS-1	22%
w/o language-only instruction tuning	**26%**

表 8：KOSMOS-1 在 Raven IQ 测试中的零样本测试结果

网页问答。网页问答旨在从网页中找到问题的答案。它要求模型既能理解文本的语义，又能理解文本的结构。结果如下：

Models	EM	F1
Using extracted text		
LLM	7.6	17.9
KOSMOS-1	**15.8**	**31.3**
Without using extracted text		
KOSMOS-1	3.8	10.6

表 9：KOSMOS-1 在基于网页问答的零样本测试结果

多模态思维链提示。受思维链提示的启发，本文对这方面进行了实验。如图 5 本文将感知语言任务分解为两个步骤。在第一阶段给定图像，使用提示来引导模型生成合理的理由或者所需要的图像描述，以产生最终结果。

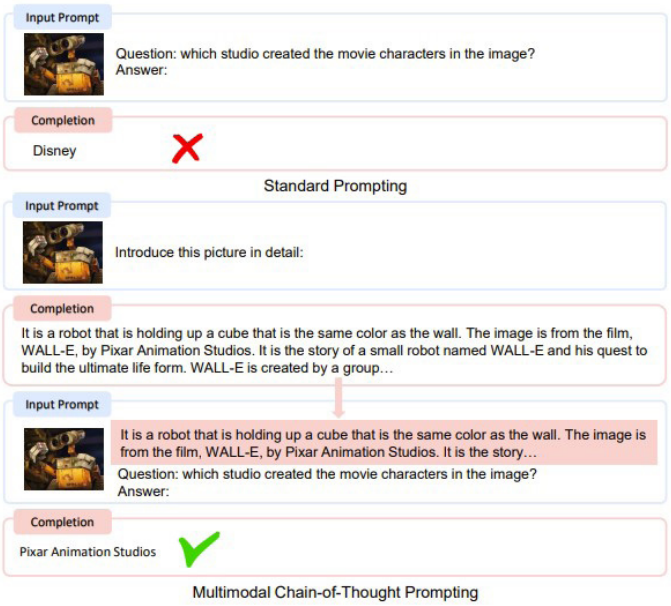


图 5：多模态思维链使 KOSMOS-1 首先生成合理的理由或者所需要的图像描述，然后将其用于解决复杂的问答和推理任务。

从表 10 可以看出，多模态思维链提示的得分为 72.9 分，比标准提示高出 5.8 分：

Models	Accuracy
CLIP ViT-B/32	59.6
CLIP ViT-B/16	59.8
CLIP ViT-L/14	64.0
KOSMOS-1	67.1
w/ multimodal CoT prompting	**72.9**

表 10：KOSMOS-1 利用多模态思维链提升 Rendered SST-2 零样本结果

相关链接：

论文地址：
https://arxiv.org/pdf/2302.14045.pdf

项目地址：
https://github.com/microsoft/unilm

机器之心 | DSN-DDI：双视图表征学习实现药物间相互作用预测性能突破

药物 - 药物相互作用（drug-drug interaction, DDI）预测用于识别药物组合之间的相互作用，其中由于理化不相容性而引起的不良反应已引起广泛关注。微软研究院科学智能中心的研究员和湖南大学 DrugAI 团队首次提出了一种新的用于 DDI 预测的双视图药物表示学习网络（“DSN-DDI”），显著提高了现有药物的 DDI 预测性能，显示出在现实世界中 DDI 应用的有用性，在协同药物组合预测方面表现出良好的可转移性，可作为药物发现领域的通用框架。

扫描二维码查看文章

央视新闻丨最潮中国范儿，当甲骨文遇上新科技

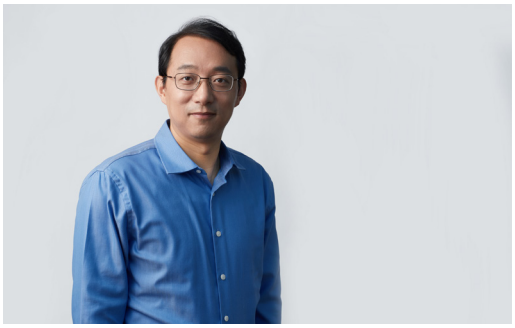
来自商朝晚期的甲骨文，距离它首次被发现已经过去 123 年，目前仍有三分之二的甲骨文字仍未破解。面对如此珍贵的历史文化遗产，甲骨文研究工作也插上了科技的翅膀，寻找新的突破口。 现在，用人工智能将一张拓片与另外 18 万张拓片逐一比对只需要 3 到 5 分钟，结束了拓片整理需要耗费数年的历史。目前，甲骨文智能拼合、甲骨文智能识别等不少新技术已经进入应用阶段。相信在不久的将来，我们对甲骨文的研究还能不断提速，一页页揭开埋藏地下数千年的灿烂文明长卷。

扫描二维码了解更多信息

「AI 中国」机器之心 2022 年度评选结果公布

「AI 中国」机器之心 "2022 年度评选" 结果正式公布。本次评选中，微软亚洲研究院成功获选 "最强技术实力公司 Top 20"。 该榜单主要关注企业的技术实力、对新一代人工智能技术的研发布局与技术储备。入选企业具有强大的技术实力，重视对新一代人工智能技术的研发布局，在研发人员、论文发布、专利储备上拥有强大的储备，并在相应人工智能技术细分领域中处于顶尖水平。

扫描二维码了解更多信息



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，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