BAAI

2021-2022 年度

智源人工智能前沿报告

BAAI AI Frontiers

北京智源人工智能研究院

2021 年 12 月

目录

审定专家[1]

黄铁军　北京智源人工智能研究院，北京大学

余　山　中国科学院自动化研究所，智源学者

刘知远　清华大学，智源青年科学家

黄　高　清华大学，智源青年科学家

张祥雨　旷视科技，智源青年科学家

鲁继文　清华大学，智源青年科学家

赵　鑫　中国人民大学，智源青年科学家

魏哲巍　中国人民大学

王树徽　中国科学院计算技术研究所

卢策吾　上海交通大学

庄福振　北京航空航天大学

王静远　北京航空航天大学

燕博南　北京大学

苗　旺　北京大学

高　阳　清华大学

王昊奋　同济大学

袁进辉　OneFlow 一流科技

刘鹏飞　卡耐基梅隆大学

曹　越　微软亚洲研究院

韩　凯　华为诺亚方舟实验室

[1] 排名不分先后

编者

北京智源人工智能研究院

戴一鸣　李梦佳　廖　璐　卢　凯　张冬敏　周岷峰　苟　瑜　靳虹博　殷靖东

刘克宇　张大魁　袁　莎　马　雷　赵小帆　闫亚琼　李　敏　刘方正　梁杨晓

李静云

智源社区

陈　旭　马瑞军　赵万铖　熊宇轩

感谢以下专家为报告提供建议

智源人工智能数理基础方向专家学者

智源机器学习方向专家学者

智源智能信息检索与挖掘方向专家学者

智源智能体系架构与芯片方向专家学者

智源自然语言处理方向专家学者

智源人工智能的认知神经基础方向专家学者

智源青年科学家

智源人工智能青年科学家俱乐部（青源会）专家学者

报告贡献情况[2]

黄铁军对报告提出方向性建议和指导。

余山、刘知远、黄高、张祥雨、鲁继文、赵鑫、魏哲巍、王树徽、卢策吾、庄福振、王静远、燕博南、苗旺、高阳、王昊奋、袁进辉、刘鹏飞、曹越、韩凯对报告中的科研、平台和工具等领域的发展情况提供内容、案例和建议。

戴一鸣、李梦佳策划本报告，制定大纲，组织并实施工作。

戴一鸣撰写第一章内容，**李梦佳、陈旭、马瑞军、赵万铖**撰写第二章内容。**廖璐**对部分翻译进行了审校，**熊宇轩**对内容进行了审校。

廖璐、李梦佳、戴一鸣、卢凯、赵小帆、闫亚琼、李敏、刘方正协调联络、收集和汇总素材。

张冬敏、周岷峰、苟瑜、靳虹博、殷靖东、刘克宇提供人工智能技术、产业、投融资等方面的内容和素材。

袁莎、张大魁、马雷对报告内容提出了专业建议。

梁杨晓、李静云提供有关悟道技术、产业等素材和内容。

智源各方向首席科学家、智源学者，智源青年科学家，青源会专家学者为报告提供了建议。

[2] 分工有重复

摘要

本报告总结 2021 年人工智能前沿科技主要趋势如下：

1. 信息模型、具身模型和脑模拟机器人的结合将诞生超级人工智能。
2. 系统研究超大规模智能模型发展和影响的新兴领域已经形成，超大规模预训练模型研发竞赛进入白热化阶段，多模态预训练模型成为下一个大模型重点发展领域。
3. Transformer 成为计算机视觉领域的重要网络架构，并开始向强化学习、图神经网络等领域渗透。
4. 加速方法创新提升了超大参数规模模型的训练效率，催生更大规模参数的巨型模型。
5. Prompt Tuning 成为自然语言处理领域预训练语言模型新型训练范式，预训练语言模型发展的新路线是提升训练和推理的效率。
6. 遮盖图像建模、非 Transformer 架构、神经辐射场等技术快速发展，成为计算机视觉的热点研究领域；脉冲视觉领域发展，将开辟机器视觉新路线。
7. 生物神经元与芯片结合成为类脑芯片的研究热点。
8. 高性能、低能耗 AI 芯片不断涌现的同时，由 AI 辅助设计成为芯片发展新趋势；存算一体 AI 芯片设计、应用步伐加快。
9. Web 模型成为新型信息搜索范式的核心支撑，预训练语言模型助力信息检索性能提升。
10. 借鉴脑神经和认知科学研究成为启发类脑智能研究的重要来源。
11. 无线高带宽、微创、结合 AI 算法等成为脑机接口的发展重点。

12. 传统科研领域成为人工智能发展的“新战场”，人工智能在辅助基础和应用科学研究的同时，也提升了智能产品和服务的性能。
13. 强化学习环境成为发展泛化性更强、适应复杂环境智能体的重要支撑，而提升训练效率成为强化学习领域的研究重点。
14. 因果推断在经济学、社会学研究中实现突破。
15. 基于超大规模预训练模型的平台和系统成为研发机构和企业的发展思路。
16. 面向更为复杂任务和需求的基准测试和数据集不断涌现。
17. AI 为人类科学家提供领域数据集，助力基础科学研究。
18. AI 算力成为超算性能比拼的“新擂台”。

本报告总结 2021 年人工智能产业主要趋势如下：

1. 智能医疗赛道持续火热，各大医疗 AI 企业纷纷冲刺 IPO，“烧钱”成为今年这一赛道最鲜明的标签。
2. 国家开始逐步发放各类医疗影像 AI 软件三类证，为医疗影像的发展提供了契机。
3. 资本助力下，新兴 AI 创企、互联网科技巨头和传统药企在 AI 制药领域百花齐放。
4. 2021 年被业界公认为数字疗法产业元年，一批数字疗法企业崭露头角。
5. 医保的推进可为手术机器人打开市场，全民可用的时代或可指日可待。
6. 脑机接口不再只是“意念打字”的融资噱头，逐渐从实验室走向临床实践，从科幻照进了现实。
7. 自动驾驶行业迎来新的投融资热潮，2021 年是十年来自动驾驶赛道资本热度最高的一年。

8. 今年，国内大批 Robotaxi 企业已进入车队测试及服务试运营的阶段，未来行业的竞争核心也将会转向运营规模与测试里程的比拼。
9. 乘用场景以外，物流、港口、矿区、城市环卫等细分场景成为自动驾驶落地新风口。
10. 今年，国内外激光雷达企业也得到了资本市场大力支持。新旧车企纷纷表示，其新车将首次搭载激光雷达，引发激光雷达量产落地的新纪元。
11. 计算机视觉，在技术成熟度、商业化进程、市场增长速度、投融资热度等方面，是人工智能产业当前热门的发展赛道。2021 年，我国计算机视觉产业快速发展，企业加快上市步伐，争夺“视觉 AI 第一股”。
12. 随着 AI 芯片技术的不断发展，芯片制程不断优化，工艺逐步提升，AI 芯片功能的细分程度进一步提升，形成异构形态的计算格局。
13. 高效、节能成为 AI 芯片发展的长期目标。追求在提升算力的前提下降低功耗，是近年来企业关注的重点。
14. GPU 依然是 AI 芯片企业研发关注的重点方向。GPU 性能较高，且兼具计算的灵活性，适用于构建大规模的 AI 计算集群，在研发超大规模 AI 模型方面具有应用前景。

前言

2021 年对于人工智能技术和产业，依旧是不平凡的一年。随着算力、数据、算法等要素逐渐齐备，先进的算法结构不断涌现，各个研究方向研究成果不断突破，成熟的 AI 技术逐渐向代码库、平台和系统发展，实现产业和商业层面的落地应用，推动人工智能发展迈向新的阶段。

科研方面，2021 年，人工智能基础理论逐渐成形，研究者对于超级人工智能的发展路径，以及深度学习模型基础理论有了更深刻的见地。2021 年也是超大规模智能模型大发展的一年，在 GPT-3 的影响下，一大批参数规模更大，训练数据量更为惊人，性能表现更强，通用任务更丰富的模型涌现出来，形成了面向“大模型”研究的新兴领域，大模型研发竞赛进入白热化阶段，多模态预训练、模型加速和应用等领域的研究如火如荼展开。Transformer 作为一种具有优势的神经网络算法架构，在计算机视觉、强化学习、图神经网络等领域逐渐渗透，展现出人工智能多学科领域通用架构的可能性。在机器学习、自然语言处理、计算机视觉等领域，新算法、新模型、新范式持续推动领域研究推陈出新。在芯片领域，将生物大脑与芯片结合，研发类脑芯片的势头更为惊人；同时，以电子元器件为基础的传统芯片不断改进，实现更高性能和更低的功耗，存算一体芯片设计快速发展，产品化步伐加快；AI 辅助设计芯片成为新趋势。预训练模型对于信息检索挖掘领域产生深远影响，有望形成基于 Web 大模型的新型信息检索范式。同时，认知神经科学研究对启发人工智能研究起到了不可忽视的作用，脑机接口等新型技术也逐渐从实验室走向实用。此外，AI for Science 的新兴领域逐渐形成，

物理学、材料学、生物学等学科已成为人工智能的下一个战场，人工智能在推动科学研究和智能产品服务进步等方面起到了更加重要的作用；

平台和工具方面，基于超大规模智能模型的开放平台对于研发先进算法和模型更加重要，极大降低应用的研发门槛，超大规模智能模型支撑的行业应用快速进入落地阶段；同时，面向复杂任务和基础科研的数据集和基准层出不穷，对于塑造 AI 科研和产业的标准，为人们提供客观、前沿的评价标准奠定基础；而人工智能算力基础设施已成为世界各国超算关注的发展重点，更大规模的 AI 超算集群落地，有助于在大尺度条件下探索人工智能的性能边界，并支持 AI 在国家战略和国民社会经济等领域实现新突破。

产业方面，今年值得关注的人工智能产业领域中，基础层重点关注 AI 芯片和数据服务领域；技术层关注自然语言处理、计算机视觉、机器学习等领域；应用层关注智能医疗和自动驾驶等领域。在上述领域中，国际国内头部、独角兽及初创企业快速发展，在产品、融资、商业模式等方面取得新的进步。

本报告分为两部分。第一章为人工智能科研和技术发展情况，其中包括科研领域、平台和工具发展情况两部分，重点梳理 2021 年度人工智能领域的科研和技术发展趋势、热点内容及案例。科研部分包括人工智能数理基础、机器学习、预训练模型、计算机视觉、自然语言处理等十余个领域，选择案例多为具有研究思路和方法论的创新性，或在人工智能领域引起热议

的研究成果及论文。平台和工具发展情况介绍包括 AI 系统和开源库、基准测试和数据集，以及算力平台三个方面的发展情况。

第二章为人工智能产业发展情况，具体包括基础层、技术层和应用层三个部分，基础层部分详述了 2021 年度 AI 芯片领域的融资事件，值得关注的 AI 芯片企业，技术层部分聚焦在自然语言处理、计算机视觉、智能语音、知识图谱等领域头部企业和创业企业的融资发展情况，应用层详述了在医疗健康、自动驾驶以及内容产业三个赛道中头部企业和创企的融资发展情况。由于人工智能产业研究范围广、事件多，本报告只列举本年度亮点领域案例，不追求大而全的描述。

研究方法

本报告采用案例征集、专家咨询等方法。首先向高校和科研机构人工智能专家学者及企业从业者征集 2021 年度人工智能领域发展的动态、案例等内容，并通过向专业人士咨询的形式汇总观点及建议，形成 2021-2022 年度人工智能前沿报告。征集案例的时间从 2021 年 1 月起，截至 2021 年 12 月 15 日。除特殊说明外，文中案例均为 2021 年内发生的事件。

第一章

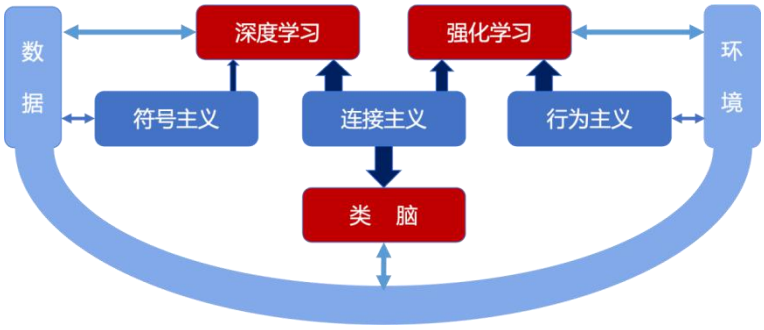
人工智能科研和技术发展情况

科研发展情况

人工智能基础理论

信息模型、具身模型和脑模拟机器人的结合将诞生超级人工智能

2021 年 11 月，在智源研究院成立三周年演讲中，智源研究院黄铁军提出了超级人工智能发展演进的思路。黄铁军认为，当前，人工智能的发展主要基于三个流派的思路，一是符号主义和连接主义结合，在大数据的支撑下发挥作用；二是连接主义和行为主义支撑下的新型强化学习方法，通过与环境的互动发挥作用；三是直接以生物进化而来的神经网络作为基础，即类脑。但不论是什么样的方法，都是实现人工智能的手段，实现智能，本质上来源于数据和环境，什么样的环境就能够创造什么样的智能。



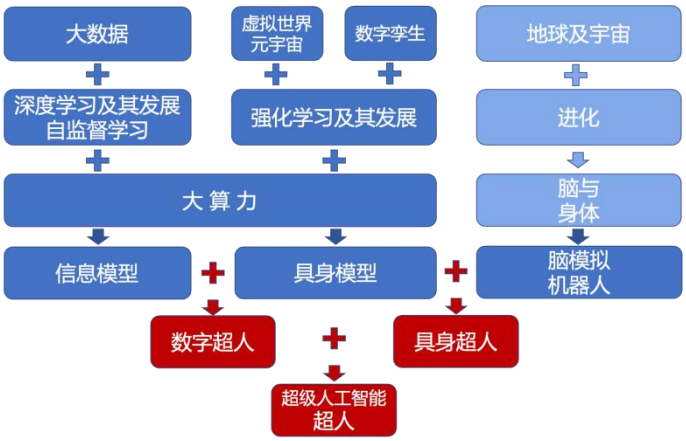
图注：通用智能的三条实现路径

来源：智源研究院

目前，人工智能在实现方法上已经明晰，主要包括近几年兴起的自监督学习为代表的基本算

法，以及强化学习领域的发展为两条主要路径。然而在数据层面，将会有新的变化。近来数字孪生、虚拟世界、元宇宙等技术快速发展，这些是比数据更高一级的数字环境，在数字环境下，让算力发挥作用。

未来几年，一是已经相对成熟的“大数据+大算力+深度学习算法”的信息模型将从研究进入实际应用；二是基于虚拟世界、实时时空环境训练的具身模型也会取得较大的发展，如自动驾驶、机器人、游戏中数字人；从更长远的角度出发，将人类大脑、生物大脑、机器人等研究方向结合，形成物理世界中具有真正实体性的机器人技术也会取得大发展。未来三年，这些技术将持续演变。未来五到十年，信息模型和具身模型将会结合，成为“数字超人”，在知识能力以及跟环境的互动程度上，将比以往的人类都要强。在元宇宙中，我们甚至不知道面对的是人工智能的化身还是真正的人类。具身模型和机器人也将结合，在物理世界出现能力比人类还要强的无人系统，即“具身超人”。乐观估计，在未来三十年，数字超人和具身超人可能会结合，最终诞生超级人工智能，这是人工智能的终极梦想，实现超越人类的智能系统。



图注：实现超级人工智能的具体路径

来源：智源研究院

对深度神经网络工作机制的理论研究热度上升

深度神经网络是当前人工智能领域的重要组成部分，但其依然是“黑盒”模型，其内部的工作机制机理仍有待进一步研究。近来，一些研究提出，除了研究“无限宽神经网络”（如神经正切核、NTK 等）之外，还可以从其它理论的角度理解神经网络的功能和工作机制。

加州大学伯克利分校研究者提出“深度学习第一性原理”

6 月，加州大学伯克利分校马毅等研究者公开了一项研究，尝试从数据压缩和区分性表征（Discriminative Representation）的原理出发，为理解深度（卷积）神经网络提供理论框架。研究者认为，如果将最大编码率衰减（Maximal Coding Rate Reduction：MCR^2）作为优化目标，可以构建一种类似神经网络架构的白盒深度学习模型，其中包括了矩阵参数、非线性层、归一化和残差连接等神经网络中的组成要素，如果引入“群不变性”，可以直接推导出多通道卷积神经网络的结构，研究者称这种模型为 ReduNet。

Facebook 研究者提出从“第一性原理”解释深度神经网络

8 月，Facebook[3]、MIT 研究者提出用于解释深度神经网络的“第一性原理”思路。“第一性原理”指回归事物最基本的条件，将其拆分成各要素进行解构分析，从而找到实现目标最优路径的方法。该研究提出了用于理解更为贴近现实的深度神经网络的有效理论。研究从“第一性原理”出发，解释通过层到层之间的迭代和非线性学习动态如何能够精确地描述训练后网络的输出。同时，从近似核方法的角度，发现模型对于学习函数的依赖可以用一种简单而

普适的方法来表达。研究者还为神经网络中的梯度爆炸和梯度消失提出了解决方案。研究表明，神经网络的深度-宽度比值决定了有效训练网络的模型复杂度。通过使用信息论技术，研究者预估了优化的深宽比值，能够使得模型更为有用。研究者同时也研究了如何使用残差连接能够让模型更深。采用以上的工具，研究者还探究了模型架构、超参数和优化器所带来的归纳偏置问题。论文地址：https://arxiv.org/pdf/2106.10165.pdf。

[3]Facebook 公司已于 2021 年 10 月更名为 Meta, 本报告中的 Facebook 泛指 Meta 公司及 Facebook AI Research(FAIR)实验室等相关机构

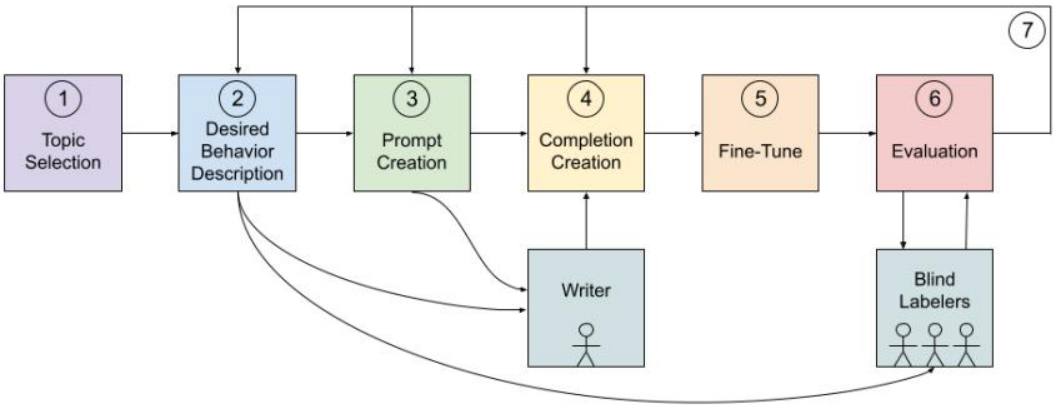
预训练模型技术

系统研究超大规模智能模型发展和影响的新兴领域已经形成

随着 BERT、GPT-3、DALL·E 等超大模型的兴起，“自监督学习+预训练模型微调”适配方案逐渐成为主流。然而，随着超大规模预训练模型在科研、产业、社会、经济等领域的作用日益凸显，其带来的深远影响成为科学家们关注的重点。

OpenAI 提出 PALMS 数据集构建和模型微调方法

6 月，OpenAI 提出名为“PALMS”的数据集构建和模型微调方法，可构建出“具有价值导向的数据集”（Values-Targeted Datasets），使其能够修正 GPT-3 偏见，对解决大模型带来的伦理问题起到了推动作用。

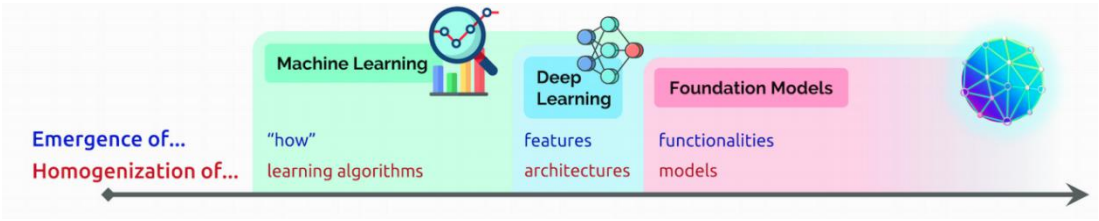


图注：OpenAI 提出的纠正 GPT-3 偏见的整体流程

来源：https://cdn.openai.com/palms.pdf

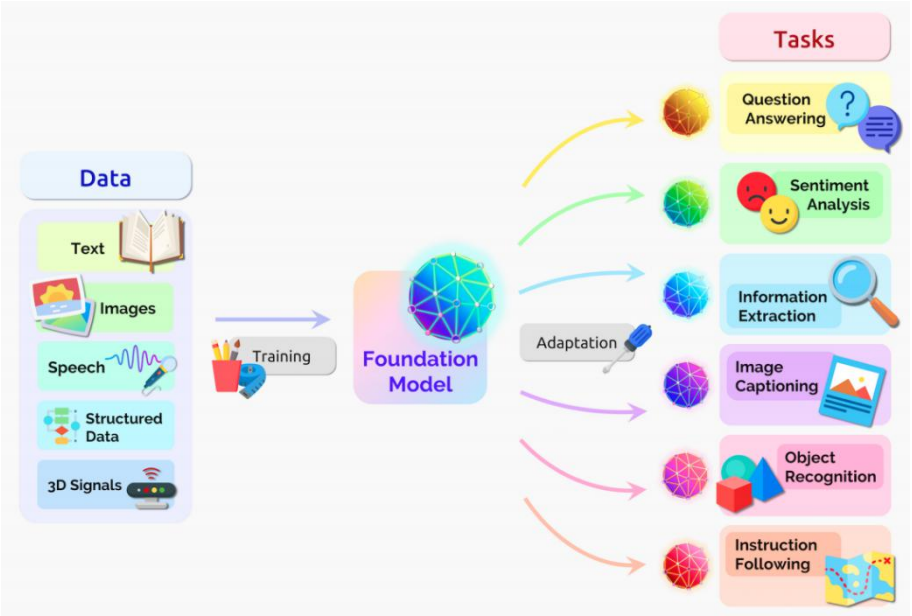
Percy Liang、李飞飞等学者提出基础模型概念

8 月，Percy Liang、李飞飞等学者将大规模预训练模型统一命名为基础模型（Foundation Models），并撰文讨论基础模型面临的机遇和挑战。论文分为四个部分，分别阐述了基础模型的能力、应用领域、技术层面和社会影响。



图注：基础模型的涌现和同质化现象

来源：https://arxiv.org/pdf/2108.07258.pdf



图注：基础模型在多种模态数据的训练和下游任务应用中处于中心地位

来源：https://arxiv.org/pdf/2108.07258.pdf

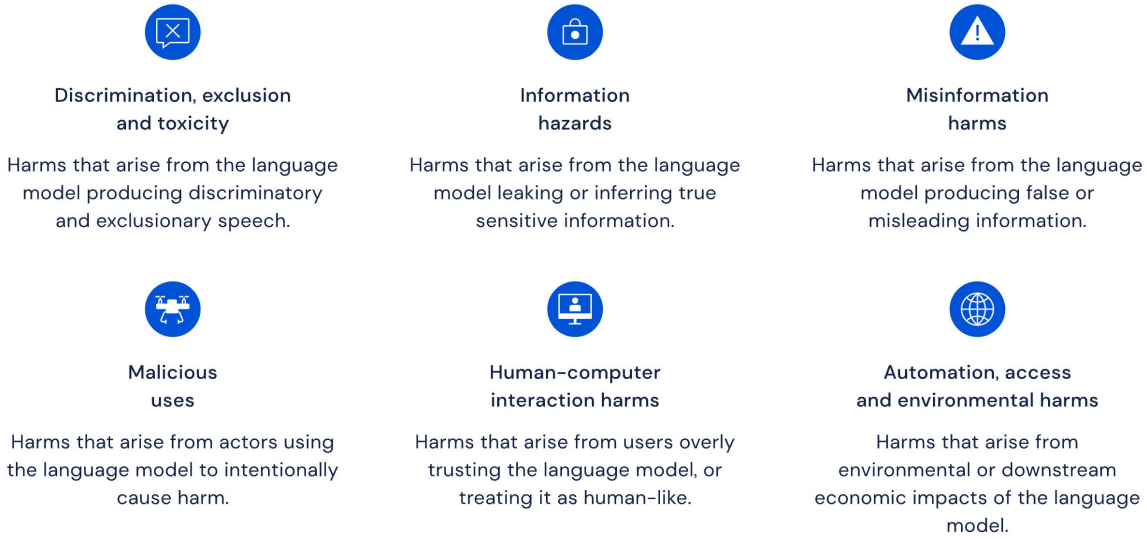
图注：基础模型涉及的议题

来源：https://arxiv.org/pdf/2108.07258.pdf

DeepMind 发表语言模型社会危害评估论文

12 月，DeepMind 发表论文，研究预训练语言模型带来的伦理和社会危害。研究者主要探究了模型在六大方面的不良影响，并谈到两个伦理和社会影响方面需要研究者持续关注。一是当前的基准测试工具不足以评估一些伦理和社会危害。例如，当语言模型生成错误信息，人类会相信这种信息为真。评估这种危害需要更多与语言模型进行人机交互。二是对于风险控

制的研究依然不足。例如，语言模型会学习、复现和放大社会偏见，但是关于这一问题的研究仍处于早期阶段。



图注：DeepMind 论文研究的六大语言模型伦理和社会危害

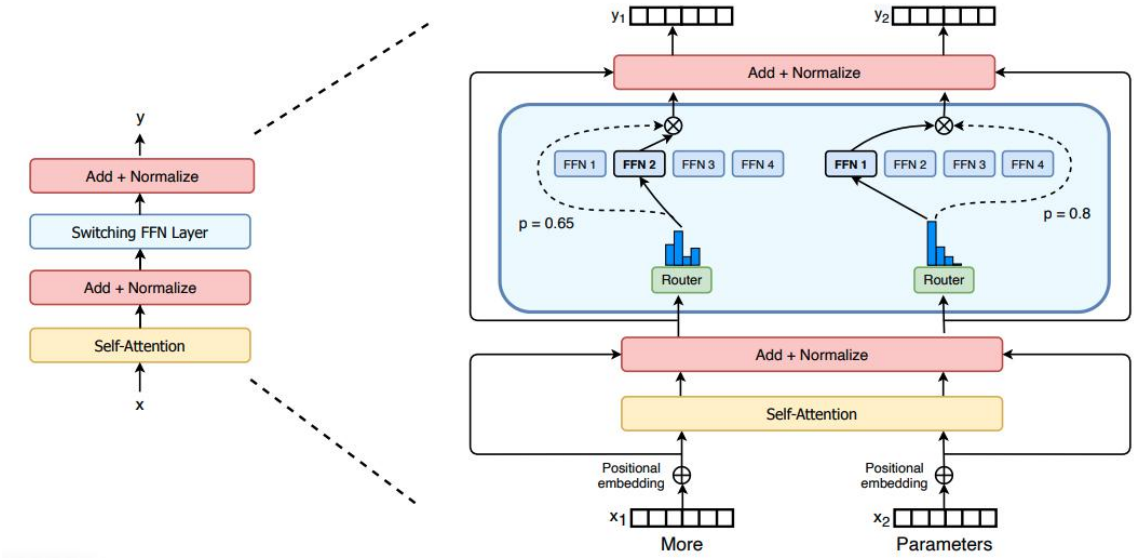
来源：https://deepmind.com/blog/article/language-modelling-at-scale

超大规模预训练模型研发竞赛进入白热化阶段

GPT-3 的问世，激发研究者探索规模更大、性能更惊人的超大规模预训练模型。国内外大型科研机构和企业纷纷投入巨量算力进行研发工作，将算力规模推升至万亿规模，探索模型的参数、性能和通用任务能力边界。目前，已有 OpenAI、谷歌、FaceBook、微软、英伟达、智源研究院、阿里达摩院、华为、百度、浪潮等研发机构和企业加入“军备竞赛”。

谷歌研发万亿规模预训练模型 Switch Transformer

1 月，谷歌研究人员研发出新的语言模型 Switch Transformer，包含 1.6 万亿个参数，是包含 1750 亿参数的 GPT-3 的九倍。研究者将 Switch Transformer 与谷歌研究的 T5-Base 和 T5-Large 模型进行了对比，结果表明，在相同的算力资源下，新模型实现了最高 7 倍的预训练速度提升。



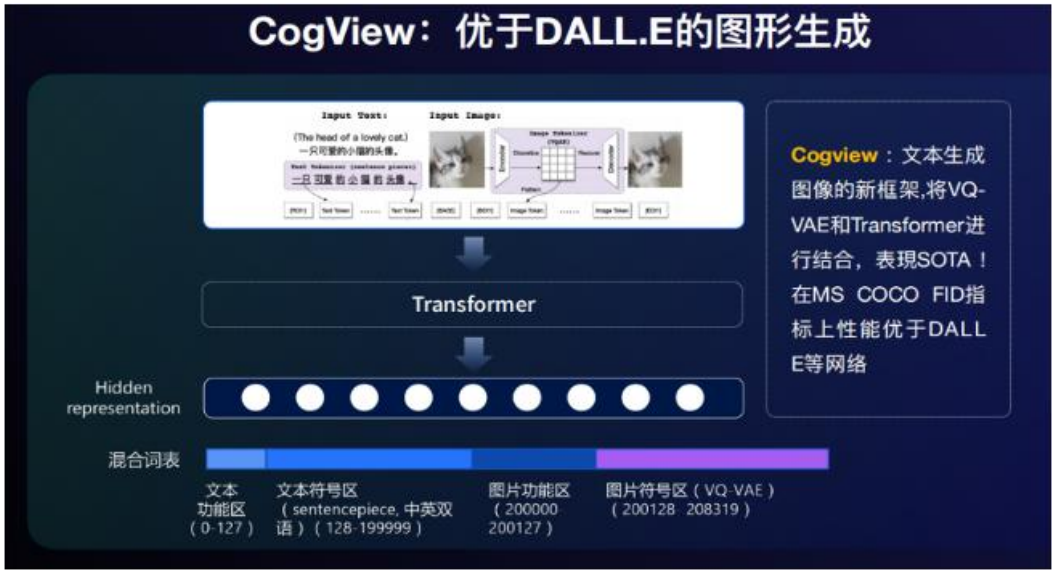
图注：Switch Transformer 编码块结构

来源：https://arxiv.org/pdf/2101.03961.pdf

智源发布超大规模智能模型悟道 1.0/2.0

3 月 20 日，智源研究院发布我国首个超大规模智能信息模型“悟道 1.0”，训练出包括中文、多模态、认知、蛋白质预测在内的系列模型，并在模型预训练范式、规模和性能扩增技术、训练语料数据库建设等方面取得了多项国际领先的技术突破。6 月 1 日，智源研究院发布“悟

道 2.0”模型，参数规模达到 1.75 万亿，是 GPT-3 的 10 倍，打破由 Switch Transformer 预训练模型创造的 1.6 万亿参数记录，是中国首个万亿级模型。



图注：悟道 2.0 中的技术创新成果

来源：智源研究院

微软、英伟达发布预训练模型 Megatron-Turing

10 月，微软联合英伟达推出了 Megatron-Turing（MT-NLP）预训练模型。该模型是微软的 T-NLG（Turing-NLG）和英伟达 Megatron-LM 模型结合的下一代版本，包含 5300 亿参数。研究者选择了五个领域中的 8 项任务来评估 MT-NLG 的效果。实验中，该模型在其中一些任务上实现了最佳的性能表现。

Dataset	Dataset source	Tokens (billions)	Weight (%)	Epochs
Books3	Pile dataset	25.7	14.3	1.5
OpenWebText2	Pile dataset	14.8	19.3	3.6
Stack Exchange	Pile dataset	11.6	5.7	1.4
PubMed Abstracts	Pile dataset	4.4	2.9	1.8
Wikipedia	Pile dataset	4.2	4.8	3.2
Gutenberg (PG-19)	Pile dataset	2.7	0.9	0.9
BookCorpus2	Pile dataset	1.5	1.0	1.8
NIH ExPorter	Pile dataset	0.3	0.2	1.8
Pile-CC	Pile dataset	49.8	9.4	0.5
ArXiv	Pile dataset	20.8	1.4	0.2
GitHub	Pile dataset	24.3	1.6	0.2
CC-2020-50	Common Crawl (CC) snapshot	68.7	13.0	0.5
CC-2021-04	Common Crawl (CC) snapshot	82.6	15.7	0.5
RealNews	RealNews	21.9	9.0	1.1
CC-Stories	Common Crawl (CC) stories	5.3	0.9	0.5

图注：MT-NLG 模型采用的数据集

来源：微软官网

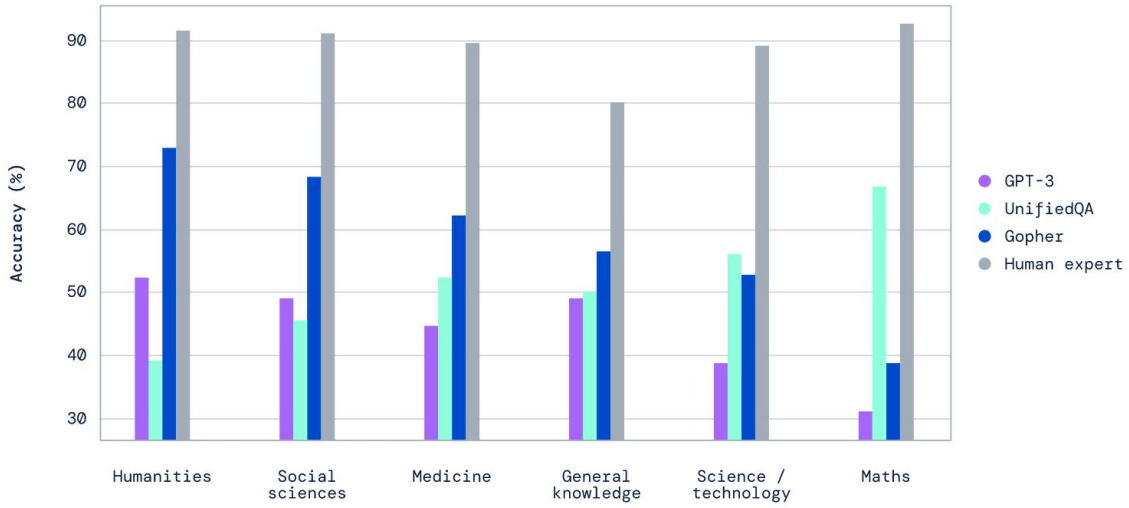
Category	Tasks	Zero-shot	One-shot	Few-shot
Completion prediction	**Lambada**	0.766*	0.731*	0.872*
Reading comprehension	**BoolQ**	0.782	0.825	0.848
Reading comprehension	**RACE-h**	0.479	0.484	0.479
Commonsense reasoning	**PiQA**	0.820*	0.810*	0.832*
Commonsense reasoning	**HellaSwag**	0.802	0.802	0.824
Commonsense reasoning	**WinoGrande**	0.730	0.737	0.789
Natural language inference	**ANLI-R2**	0.366	0.397	0.396
Natural language inference	**HANS**	0.607	0.649	0.702
Word sense disambiguation	**WiC**	0.486	0.513	0.585

图注：MT-NLG 在零样本、单样本和小样本条件下在不同任务中的表现

来源：微软官网

DeepMind 发布预训练模型 Gopher

12 月，DeepMind 发布预训练语言模型 Gopher，参数规模达 2800 亿。该模型采用 4096 块 TPUv3 加速芯片进行训练，并结合了多种并行加速策略。该研究主要用于探索不同规模的模型的优势和不足，了解在模型参数规模增长后，在哪些领域上能够得到更好的性能表现。研究者发现，模型规模的增长对于阅读理解、事实核查、毒害言论辨认等任务有较大提升，但是逻辑推理和常识任务上的提升并不显著。此外，研究者也研究了 Gopher 模型在对话等领域的能力以及缺陷。



图注：Gopher 和其他模型在大规模多任务语言理解（Massive Multitask Language Understanding，MMLU）基准上在不同类别下的表现

来源：https://deepmind.com/blog/article/language-modelling-at-scale

其他企业持续研发超大规模预训练模型

除以上案例外，4 月，华为云联合循环智能发布盘古 NLP 超大规模预训练语言模型，参数规模达 1000 亿，联合北京大学发布盘古α超大规模预训练模型，参数规模达 2000 亿；阿里达

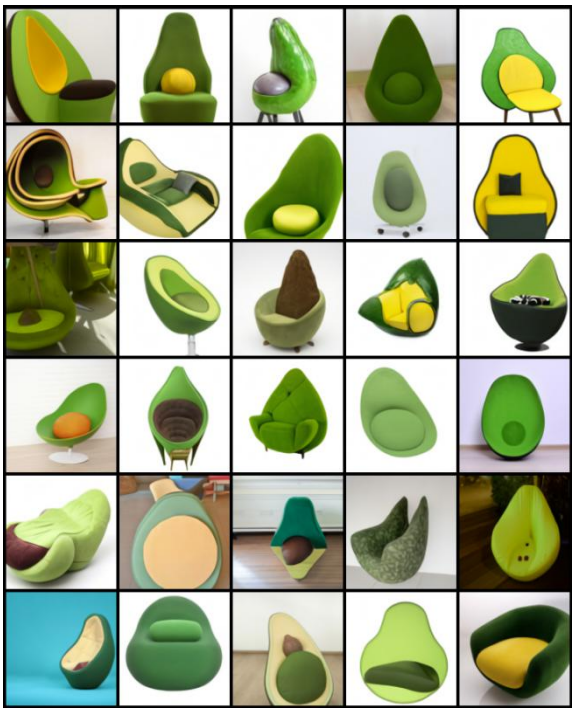
摩院发布 270 亿参数的中文预训练语言模型 PLUG，联合清华大学发布参数规模达到 1000 亿的中文多模态预训练模型 M6，目前已突破十万亿参数规模；7 月，百度推出 ERNIE 3.0 知识增强大模型，参数规模达到百亿；10 月，浪潮发布约 2500 亿的超大规模预训练模型；12 月，百度推出 ERNIE 3.0 Titan 模型，参数规模达 2600 亿；谷歌训练参数规模达 4810 亿的巨型 BERT 模型，结果公布在 MLPerfv1.1 训练榜单上；此外，谷歌还提出了 1.2 万亿参数的通用稀疏语言模型 GLaM，在 7 项小样本学习领域的性能超过 GPT-3。

多模态预训练模型成为下一个大模型重点发展领域

在大数据、大参数和大算力的支持下，预训练模型能够充分学习文本中的表征，掌握一定的知识。如果模型能够学习多种模态的数据，在图文生成、看图问答等视觉语言（Vision Language）任务上具有更强表现。多模态预训练模型是 2021 年的重点研究方向，OpenAI、微软、智源、清华大学、中科院自动化所等机构均发布了多模态预训练模型。

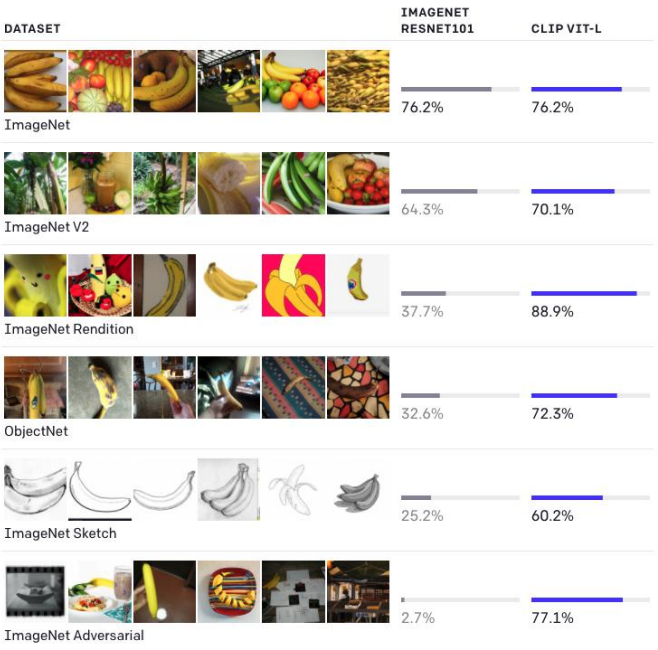
OpenAI 提出大规模多模态预训练模型 DALL·E 和 CLIP

1 月，OpenAI 同时发布了两个大规模多模态预训练模型——DALL·E 和 CLIP。DALL·E 可以基于短文本提示（如一句话或一段文字）生成对应的图像，CLIP 则可以基于文本提示对图片进行分类。OpenAI 表示，研发多模态大模型的目标是突破自然语言处理和计算机视觉的界限，实现多模态的人工智能系统。



图注：DALL·E 生成的“牛油果形状的椅子”

来源：OpenAI 官网

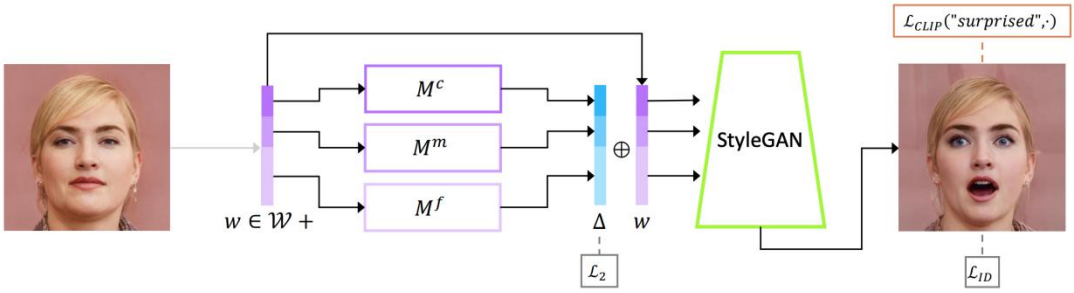


图注：CLIP 模型在多项 ImageNet 测试中取得优秀水平

来源：OpenAI 官网

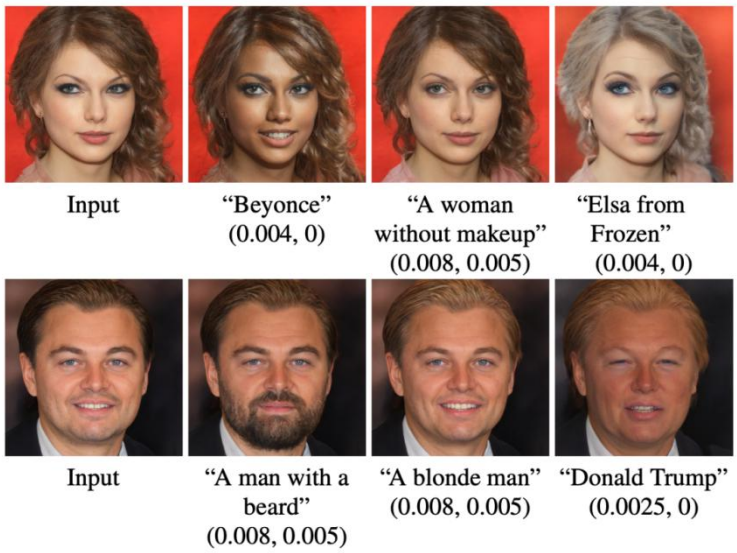
以色列希伯来大学等提出文生高清图模型 StyleCLIP

3 月，以色列希伯来大学、Adobe 研究院等将 StyleGAN 和 CLIP 模型结合，提出了一种能够根据文本提示生成高清晰度图像的模型，名为 StyleCLIP。研究者认为，StyleCLIP 能够结合预训练模型学习到的语义知识，加上生成对抗网络的图像生成能力，能够创造出更逼真的图像，在实际应用中有一定的优势。



图注：StyleCLIP 的处理图像的流程

来源：https://arxiv.org/pdf/2103.17249.pdf

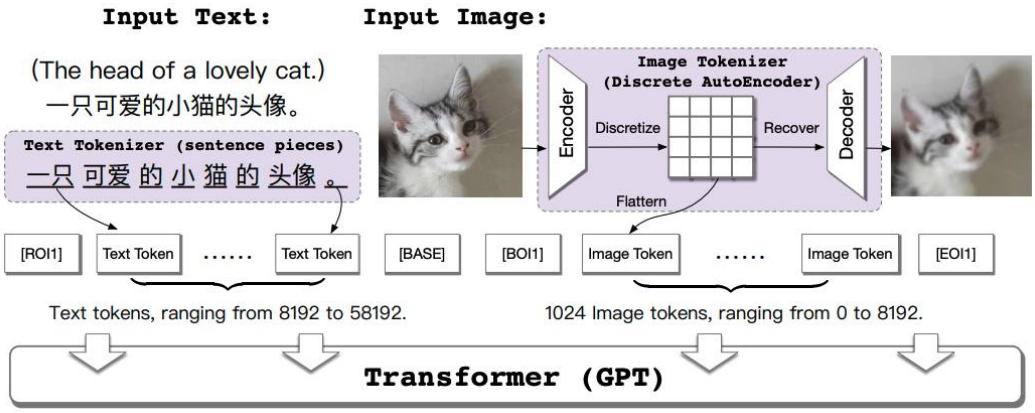


图注：根据文本提示进行的图像 PS 操作结果

来源：https://arxiv.org/pdf/2103.17249.pdf

智源、清华等研究者提出文生图模型 CogView

5 月，智源研究院、清华大学、阿里达摩院的研究者发布了 CogView 文生图模型论文，其将 VQ-VAE 和 40 亿参数的 Transformer 模型结合，通过在风格学习、超高清图像生成、文-图排序和时尚设计等多个下游任务上进行微调，并采用了消除 NaN 损失等稳定预训练的方法。实验结果显示，CogView 在模糊化后的 MS COCO dataset 数据集上取得了最高的 FID 结果，高于以往的 GAN 和 DALL·E。



图注：CogView 的架构

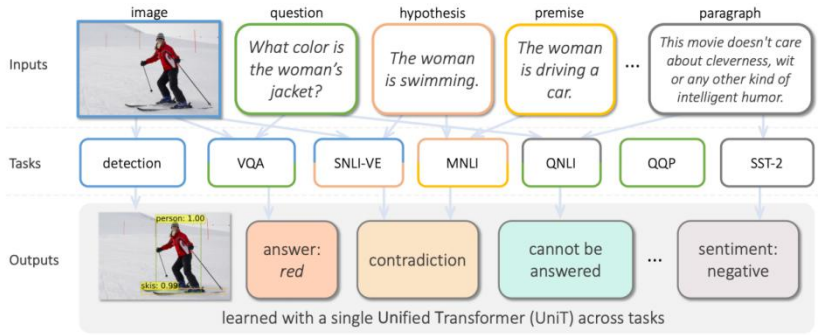
来源：https://arxiv.org/pdf/2105.13290.pdf

图注：CogView 按照提示语生成的图像

来源：https://arxiv.org/pdf/2105.13290.pdf

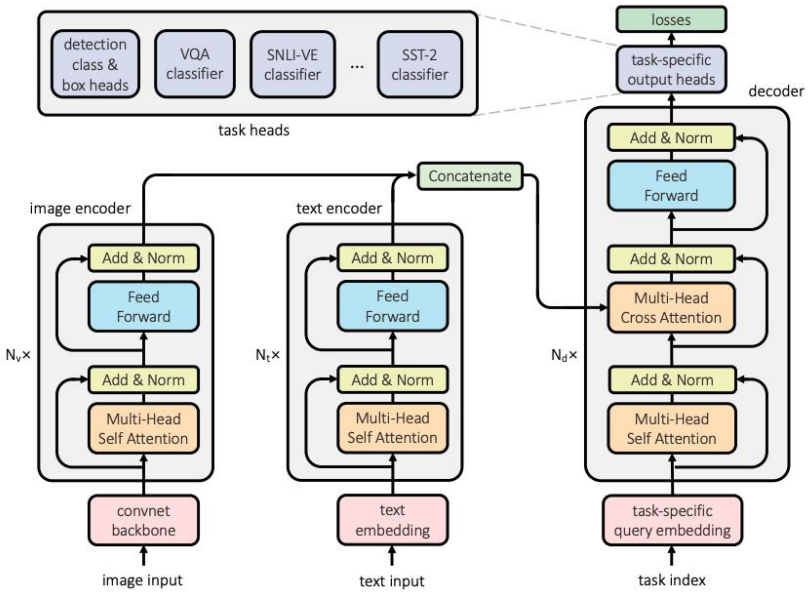
Facebook 研究者提出多任务多模态统一模型 UniT

8 月，Facebook 研究团队提出了名为 UniT 的多任务多模态统一 Transformer 模型，其基于统一的 Transformer Encoder-Decoder 架构，能够同时解决视觉、多模态、语言等领域中的一系列任务，包括目标检测、视觉-文本推理、自然语言理解等。论文表示，该模型在 7 个任务上都有较强的性能。



图注：UniT 模型能够学习的数据和完成的任务一览

来源：https://arxiv.org/pdf/2102.10772.pdf

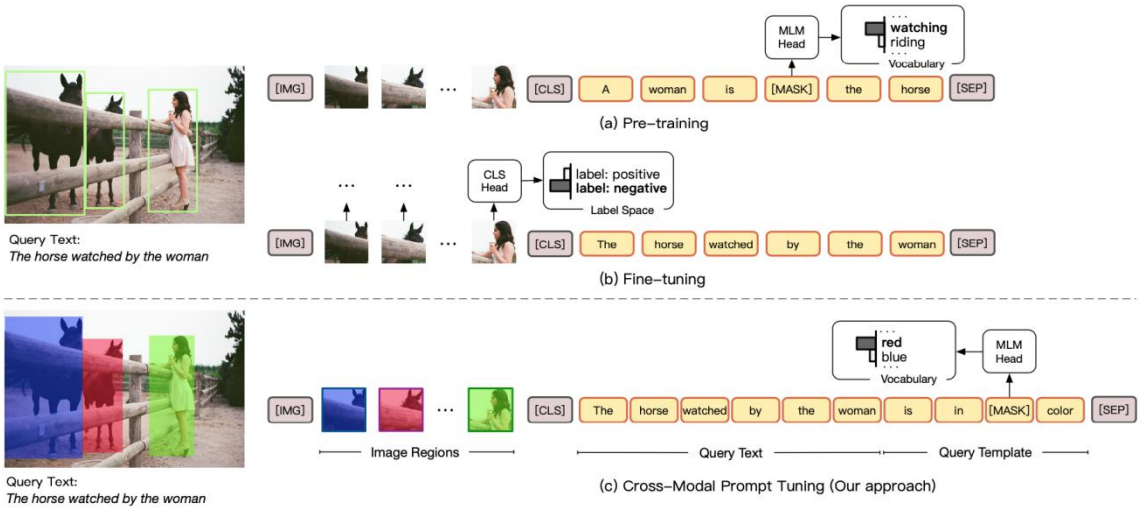


图注：UniT 模型架构

来源：https://arxiv.org/pdf/2102.10772.pdf

清华等研究者提出跨模态提示学习模型 CPT

9 月，清华和新加坡国立大学的研究者提出了跨模态提示学习模型 CPT，其利用颜色对跨模态预训练模型进行基于提示学习的微调，在视觉定位、场景图生成任务的少次学习场景下较基线模型取得显著提升。

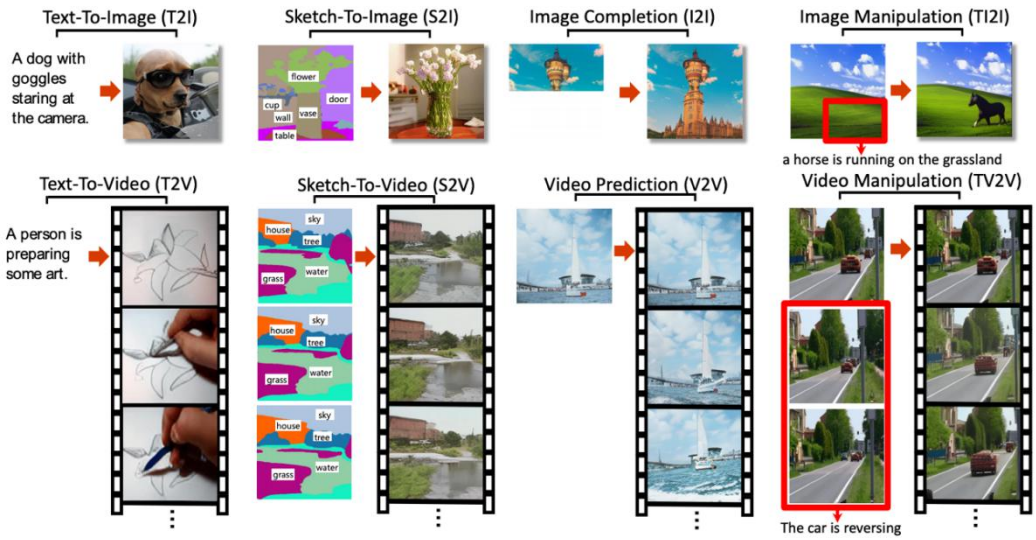


图注：CPT 跨模态提示学习框架

来源：https://arxiv.org/pdf/2109.11797.pdf

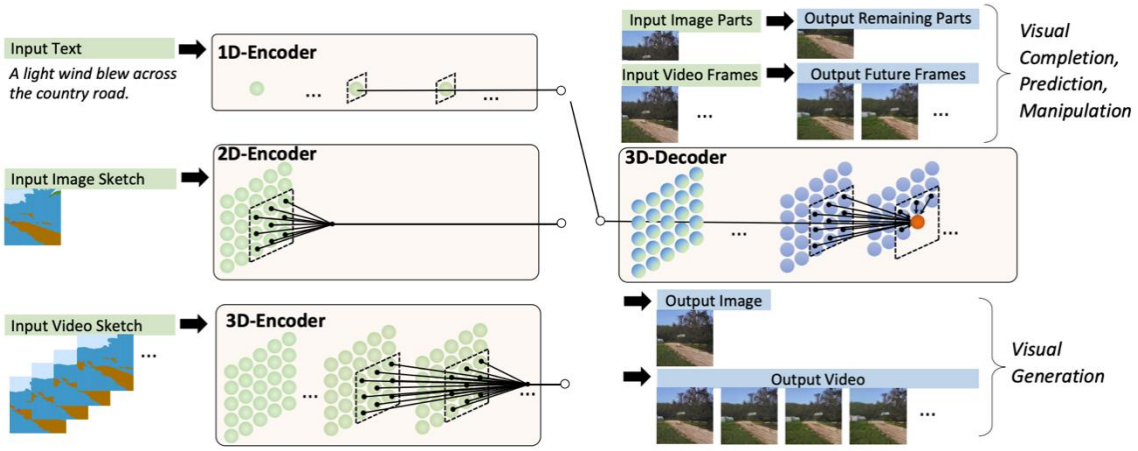
微软亚洲研究院、北大研究者提出涵盖三种模态数据的预训练模型 NÜWA（女娲）

11 月，微软亚洲研究院、北大研究者提出统一多模态预训练模型 NÜWA。该模型采用 3D Transformer 架构，能够生成视觉（图像或视频）信息。通过将该模型在 8 个下游任务上进行试验，女娲模型在文生图、文生视频、视频预测等任务上实现最佳性能。



图注：女娲模型支持的下游任务

来源：https://arxiv.org/pdf/2111.12417.pdf

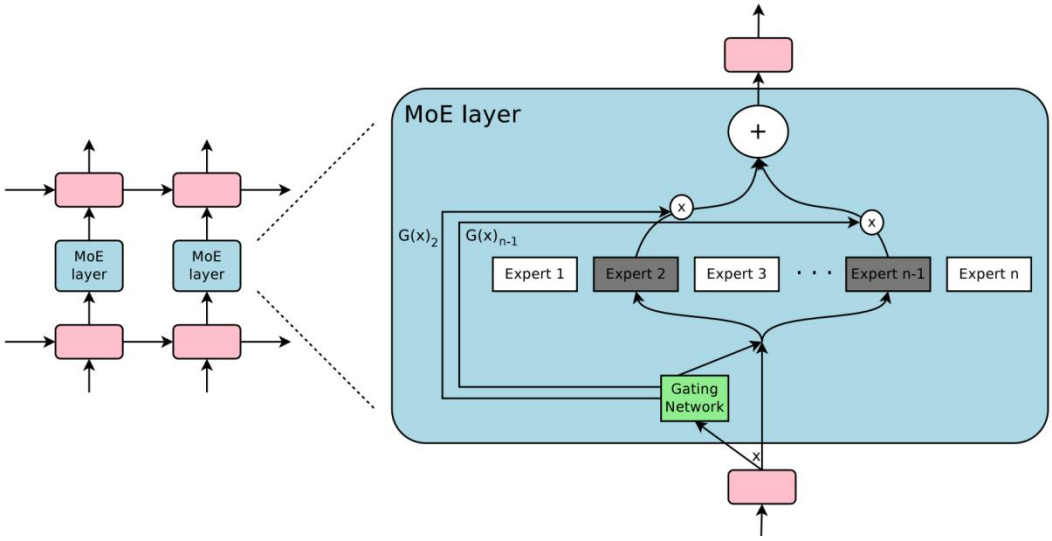


图注：女娲模型的架构

来源：https://arxiv.org/pdf/2111.12417.pdf

加速方法创新提升超大参数规模模型的训练效率

受制于算力资源，超大规模预训练模型的训练和推理面临严重的瓶颈。在 GShard 和 Switch Transformer 的研究中，谷歌通过采用混合专家技术（Mixture of Experts，MoE），通过在神经网络中引入多个专家网络（Expert Network），降低需要激活的神经元数量，提升模型的计算效率，将预训练语言模型的参数提升至万亿规模。



图注：MoE 的架构，采用稀疏门控函数（Sparse Gating Function）来决定执行计算的专家网络

来源：https://arxiv.org/pdf/1701.06538.pdf

微软等研究者提出 ZeRO-Offload 异构训练技术

随着超大规模预训练模型参数规模的增加，今年出现了更多大模型计算加速和优化方法，着力提升模型的计算效率。1 月，微软、加州大学默塞德分校（University of California, Merced）的研究者提出了一种名为“ZeRO-Offload”的异构深度学习训练技术，使用相同的硬件能够训练比以往规模大 10 倍的模型。在 32GB RAM 的 V100 GPU 上，用户可以通过 ZeRO-offload

训练 130 亿参数的 GPT-2；在单个 DGX-2 服务器上，ZeRO-offload 能够训练参数量超 700 亿的模型，在原有的硬件基础上实现了 4.5 倍的模型规模提升。

智源、清华研究者联合研发 FastMoE 加速系统

由于 MoE 技术和谷歌软硬件绑定，其无法直接应用于 PyTorch 等开源算法框架。为了解决这一问题，3 月，智源研究院和清华大学联合研发了名为 FastMoE 的加速系统，使普通用户可以通过改写代码的方式，直接使用 MoE 模块。相比原版，FastMoE 实现了 47 倍的提速优化。FastMoE 系统既可以作为 PyTorch 网络中的一个模块使用，也可用于改造现有网络中某个层。用户只需要几行代码便可调用 MoE 模块。FastMoE 也支持将任意神经网络模块作为专家网络，并包含了一些专门优化的 CUDA 代码，更加充分地利用了 GPU 大规模并行计算的能力。

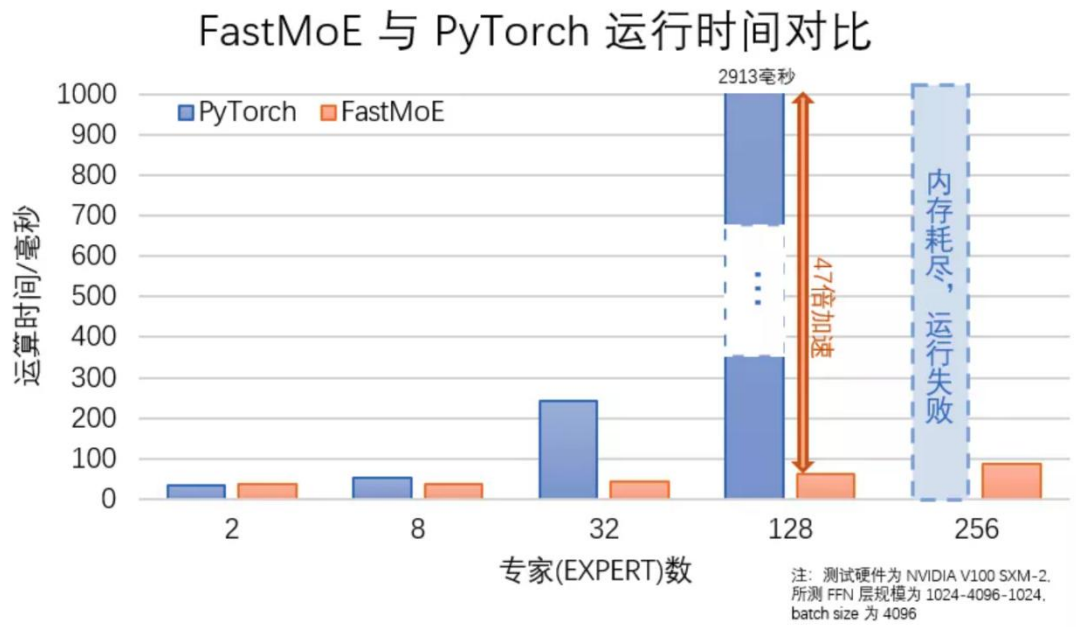
```
model = ...

from fmoe.megatron import fmoefy
model = fmoefy(model, num_experts=<number of experts per worker>)

train(model, ...)
```

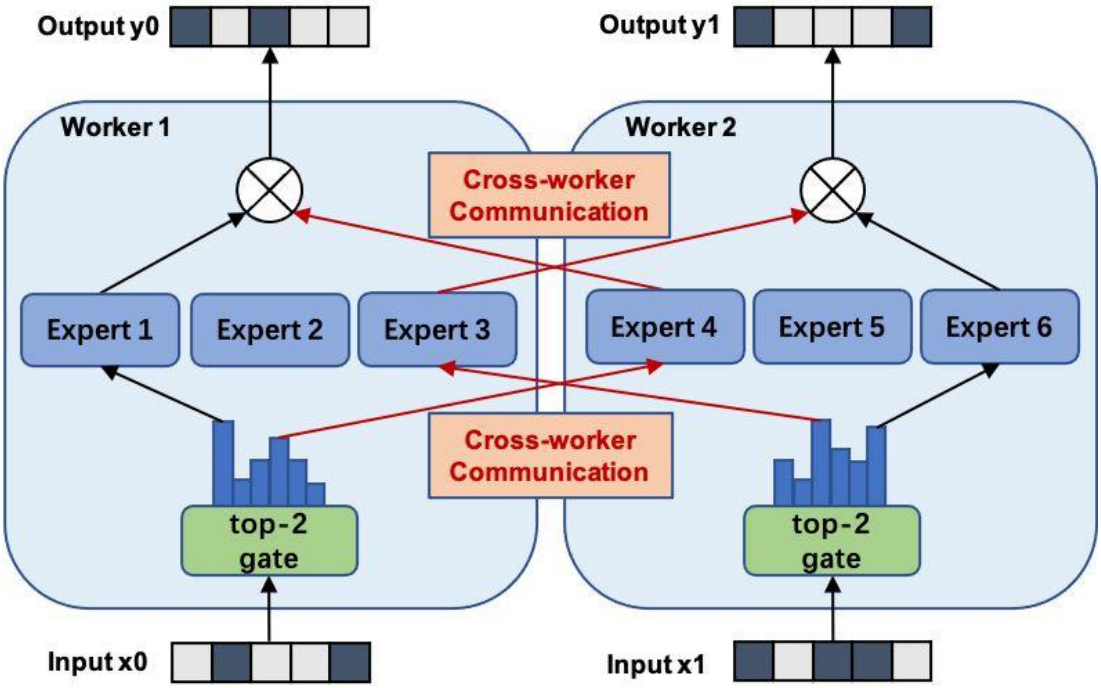
图注：调用 FastMoE 代码的方式

来源：https://github.com/laekov/fastmoe



图注：FastMoE 和原版 PyTorch 性能的对比

来源：https://mp.weixin.qq.com/s/9Quf_sfxHlugj-91XKQxJg

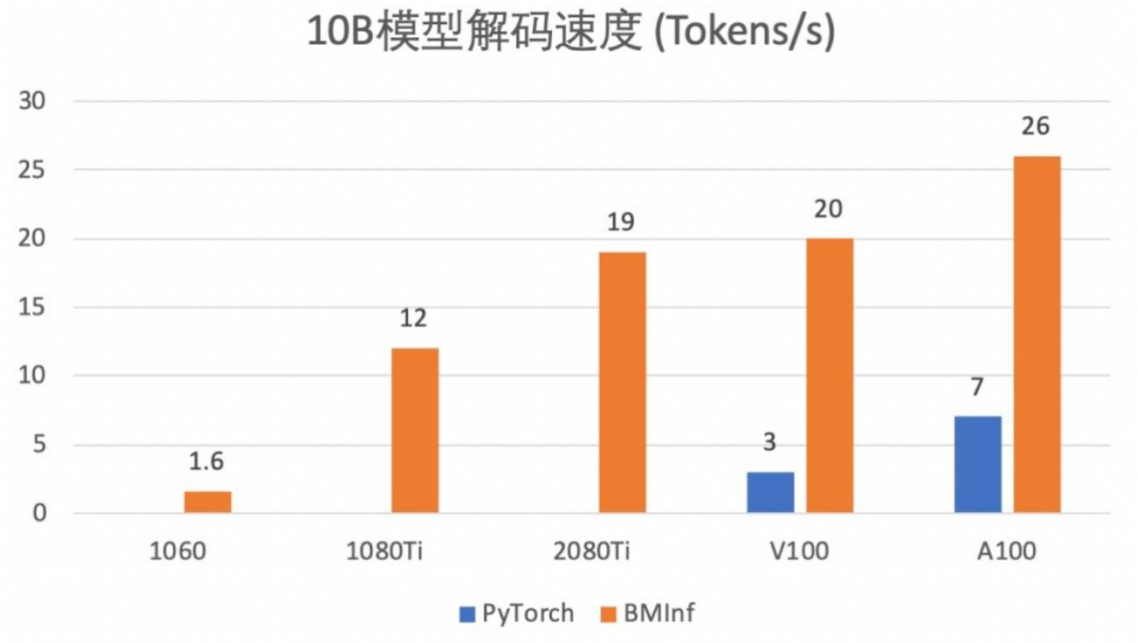


图注：FastMoE 的数据并行模式

来源：https://mp.weixin.qq.com/s/9Quf_sfxHlugj-91XKQxJg

智源、清华研究者联合研发 BMInf 加速系统

预训练大模型在各个领域均取得了惊人的效果，但大模型的应用却具有较高的算力门槛，较长的模型响应速度。9 月，清华与智源研究者联合发布了低资源大模型推理工具包 BMInf，在消费级显卡上也可以进行百亿大模型的高效推理。



图注：BMInf 和原版 PyTorch 性能的对比

来源：https://github.com/OpenBMB/BMInf

微软、英伟达联合提出 PTD-P 加速方法

10 月，微软和英伟达联合提出了 PTD-P（Inter-node Pipeline Parallelism, Intra-node Tensor Parallelism, and Data Parallelism）训练加速方法，通过数据并行、张量并行和 Pipeline 并行“三管齐下”的方式，将模型的吞吐量提高 10%以上。该并行方法可以在 3072 个 GPU 上，以 502P 的算力对一万亿参数的 GPT 架构模型进行训练，实现单 GPU 吞吐量

52%的性能提升。利用该技术，微软和英伟达在 3000 多块 GPU 上训练出 5300 亿参数的超大规模预训练语言模型 Megatron-Turing。

Number of parameters (billion)	Attention heads	Hidden size	Number of layers	Tensor model-parallel size	Pipeline model-parallel size	Number of GPUs	Batch size	Achieved teraFlOP/s per GPU	Percentage of theoretical peak FLOP/s	Achieved aggregate petaFLOP/s
1.7	24	2304	24	1	1	32	512	137	44%	4.4
3.6	32	3072	30	2	1	64	512	138	44%	8.8
7.5	32	4096	36	4	1	128	512	142	46%	18.2
18.4	48	6144	40	8	1	256	1024	135	43%	34.6
39.1	64	8192	48	8	2	512	1536	138	44%	70.8
76.1	80	10240	60	8	4	1024	1792	140	45%	143.8
145.6	96	12288	80	8	8	1536	2304	148	47%	227.1
310.1	128	16384	96	8	16	1920	2160	155	50%	297.4
529.6	128	20480	105	8	35	2520	2520	163	52%	410.2
1008.0	160	25600	128	8	64	3072	3072	163	52%	502.0

图注：采用 PTD-P 技术训练模型时达到的参数规模和性能水平

（单位：teraFlOP/s per GPU；petaFLOP/s）

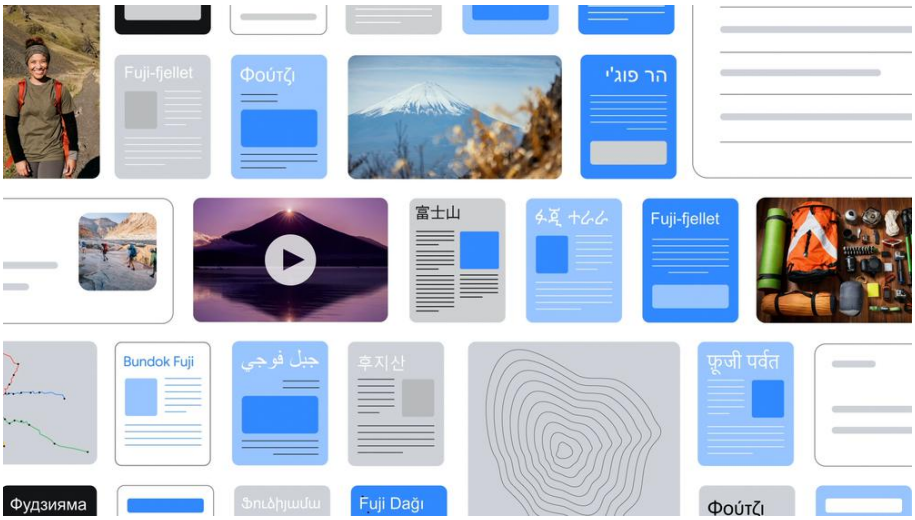
来源：https://arxiv.org/pdf/2104.04473.pdf

预训练模型在生物学研究和互联网等场景实现应用

随着数据规模逐渐扩大，数据模态进一步丰富，预训练模型将向更多领域渗透，通过“预训练-微调”的范式，完成多种类型的任务。在科研领域，预训练模型将与领域内的数据结合，成为一种完成下游任务的“基础模型”，助力诞生更多科学研究发现。在产业领域，面向更为复杂的智能决策场景，基于多种互联网数据进行预训练，具有决策能力的大模型可能是下一步发展的重点。

谷歌提出多任务统一模型 MUM

5 月，谷歌在 2021 IO 大会上公开了多任务统一模型（Multitask Unified Model，MUM）的发展情况。MUM 模型能够理解 75 种语言，并预训练了大量的网页数据，擅长理解和解答复杂的决策问题，并能够从跨语言多模态网页数据中寻找信息，在客服、问答、营销等互联网场景中具有应用价值。



图注：MUM 模型能够根据用户提问从多种源头的网页信息中搜索出对应的旅行攻略

来源：https://blog.google/products/search/introducing-mum/

清华、智源等研究者提出中文核心语言模型 CPM

6 月，清华、智源等研究者在北京智源大会上公开了以中文为核心的多语言预训练模型 CPM，兼具中英文语言的理解和生成能力，在识记、阅读、分类、推理、跨语、生成、概括等七大机器语言能力测试中，与现有开源预训练模型相比整体性能显著最优。公开可下载的 CPM-2 模型分为 3 个不同版本：110 亿参数中文模型、110 亿参数中英模型以及 1980 亿中英 MoE 模型。

	CCPM	C^3	Sogou-Log	WMT20	Math23K	LCSTS	LCQMC	AdGen	CUGE
	Acc	Acc	MRR/NDCG	BLEU	Acc	Rouge-L	Acc	BLEU/Distinct	Score
mT5-small	$87.7_{(100)}$	$41.5_{(100)}$	$29.2/29.2_{(100)}$	$9.1_{(100)}$	$18.4_{(100)}$	$33.1_{(100)}$	$82.1_{(100)}$	$10.2/32.3_{(100)}$	100
mT5-large	$89.9_{(102)}$	$56.3_{(136)}$	$32.2/31.1_{(108)}$	$11.1_{(122)}$	$34.3_{(186)}$	$34.4_{(104)}$	$85.0_{(104)}$	$10.0/35.5_{(104)}$	126
mT5-XXL	$90.6_{(103)}$	**$86.4_{(208)}$**	**36.9**/$34.9_{(123)}$	$24.0_{(264)}$	$61.6_{(335)}$	$34.8_{(105)}$	$88.3_{(108)}$	$9.8/68.7_{(154)}$	190
CPM-2	**$91.6_{(104)}$**	$86.1_{(207)}$	$36.3/$**$35.5_{(123)}$**	**$26.2_{(288)}$**	**$69.4_{(377)}$**	**$35.9_{(108)}$**	**$89.2_{(109)}$**	**$10.6/70.2_{(161)}$**	**198**

图注：CPM 模型在下游任务中的性能表现

来源：https://arxiv.org/pdf/2106.10715.pdf

智源、清华等研究者提出蛋白质预训练模型 ProteinLM

8 月，智源研究院悟道团队联合清华大学、腾讯量子实验室提出蛋白质预训练模型 ProteinLM，目前已开源 2 亿和 30 亿参数规模的模型。该模型支持蛋白质二级结构预测、荧光性预测、接触预测、折叠稳定性预测和远缘同源性检测任务。相较于基线模型 TAPE（3800 万参数），ProteinLM 在下游任务上表现有所提升，尤其是在蛋白质折叠预测问题上，模型较基线模型提高了 39%。

Task	Metric	TAPE	ProteinLM (200M)	ProteinLM (3B)
contact prediction	P@L/5	0.36	0.52	**0.75**
remote homology	Top 1 Accuracy	0.21	0.26	**0.30**
secondary structure	Accuracy (3-class)	0.73	0.75	**0.79**
fluorescence	Spearman's rho	0.68	0.68	0.68
stability	Spearman's rho	0.73	0.77	**0.79**

图注：ProteinLM 模型在下游任务中的性能表现

来源：https://github.com/BAAI-WuDao/ProteinLM

清华大学研究者提出基于大模型的 EVA 对话系统

8 月，清华大学唐杰等学者提出了基于超大规模预训练模型的开放域中文对话系统 EVA。该系统有 28 亿参数，基于 14 亿对话对的 WDCDialogue 数据集训练。实验显示，EVA 相比其他中文的预训练对话系统在多轮人机对话场景下有更好的表现。

图注：EVA 对话系统的界面

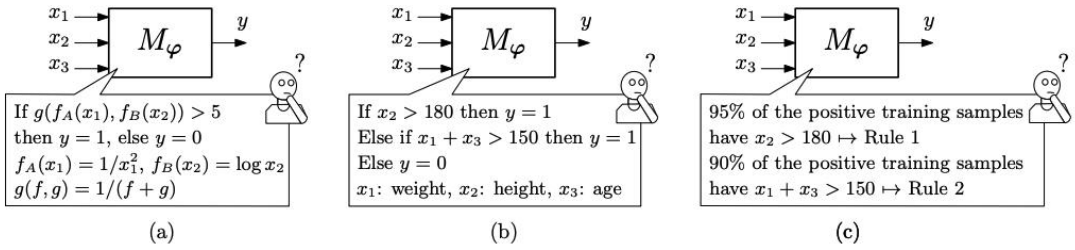
来源：https://arxiv.org/pdf/2108.01547.pdf

机器学习

深度学习模型事后可解释性研究出现新范式

可解释性是近些年来研究者探索的重要领域。人工智能存在的风险之一在于其作出的决策结果不透明，无法基于其输出给出直接的解释。缺乏可解释性导致很多 AI 算法和模型无法应用在敏感领域，如医学诊断、航空航天、国防、金融等。根据论文《Explainable Artificial

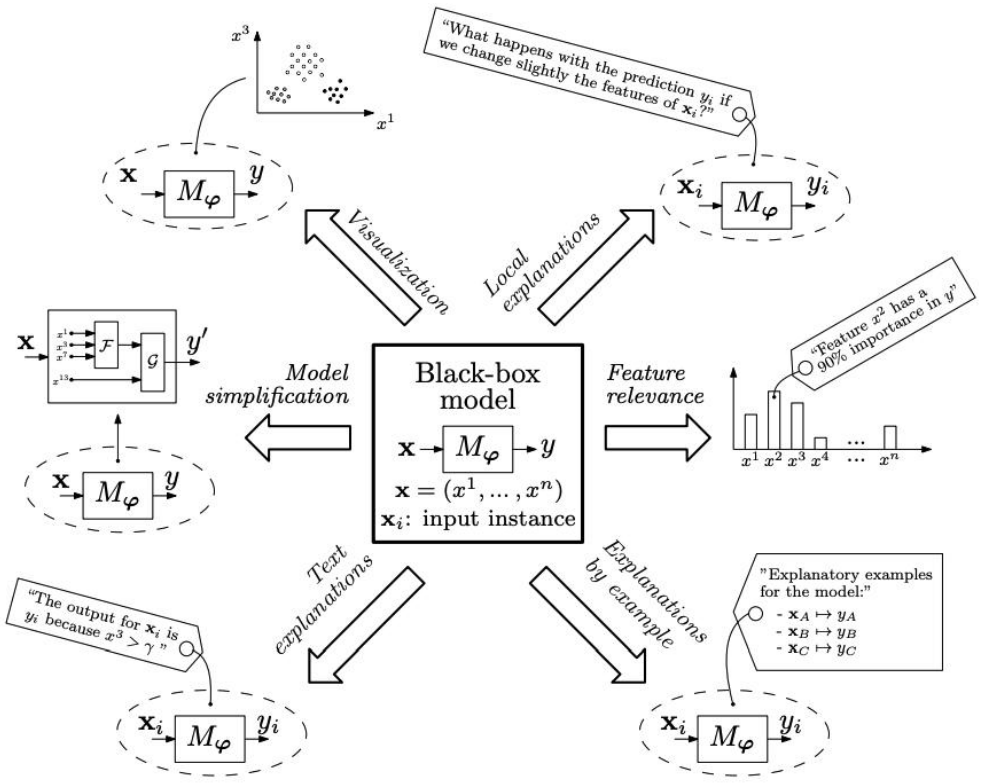
Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI》，机器学习可解释性研究主要关注两个方面，一是机器学习模型的透明性（Transparency in Machine Learning Models），涵盖不同层级的机器学习透明度，包括可模拟的、可分解的和算法层面的透明性。



图注：不同层级的机器学习透明度

来源：https://arxiv.org/pdf/1910.10045.pdf

二是机器学习模型的事后可解释性（Post-hoc Explanability Techniques for Machine Learning Models），包括文本解释、模型简化、可视化、局部解释、特征相关性、由样本解释等不同的方法，提升模型的可解释性。

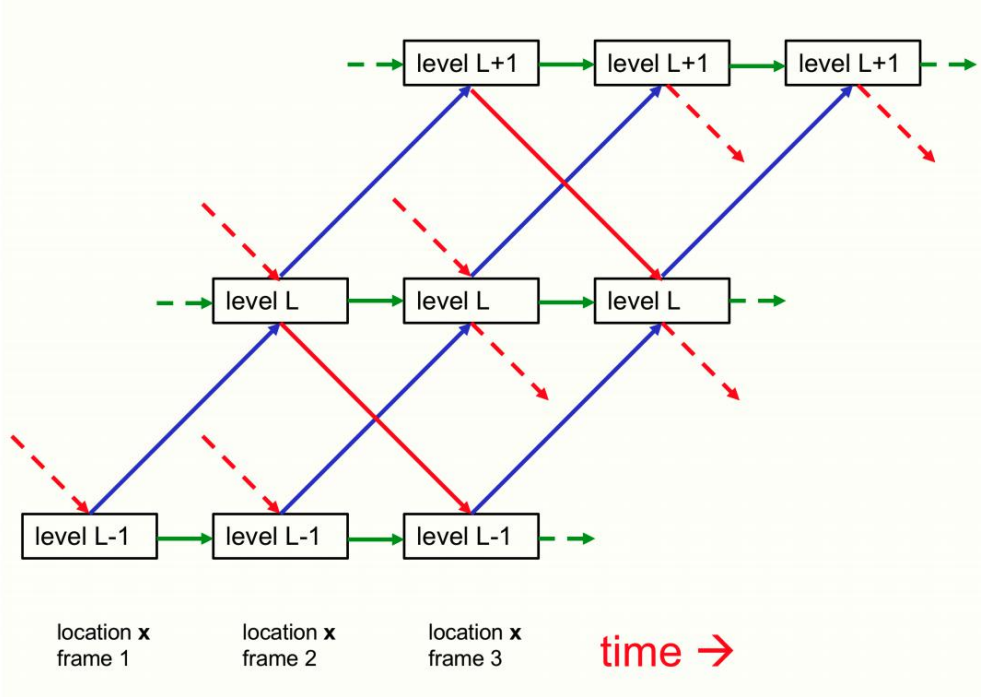


图注：机器学习模型的事后可解释性方法一览

来源：https://arxiv.org/pdf/1910.10045.pdf

Hinton 提出神经网络将图像解析为“部分-整体”层次结构的思路

近年来，在深度学习的可解释领域，研究者重点通过对神经网络结构进行改变的方式，提升模型的可解释性。2 月，图灵奖得主 Geoffery Hinton 提出了一个假想的系统——GLOM，这一系统融合了 Transformer、神经场（Neural Field）、对比表示学习、模型蒸馏和胶囊网络等结构，试图回答这样一个问题：具有固定架构的神经网络如何将图像解析为“部分-整体”的层次结构，而每个图像的层次结构不同？如果 GLOM 被证实可行，其可以提升 Transformer 类型的系统在视觉或语言任务上的可解释性能力。



图注：在相毗邻的三层 GLOM 架构中从上到下、从下到上，同一层之间的交互关系

来源：https://arxiv.org/pdf/2102.12627.pdf

特拉维夫大学、Facebook 研究者提出提升注意力模型可解释性方法

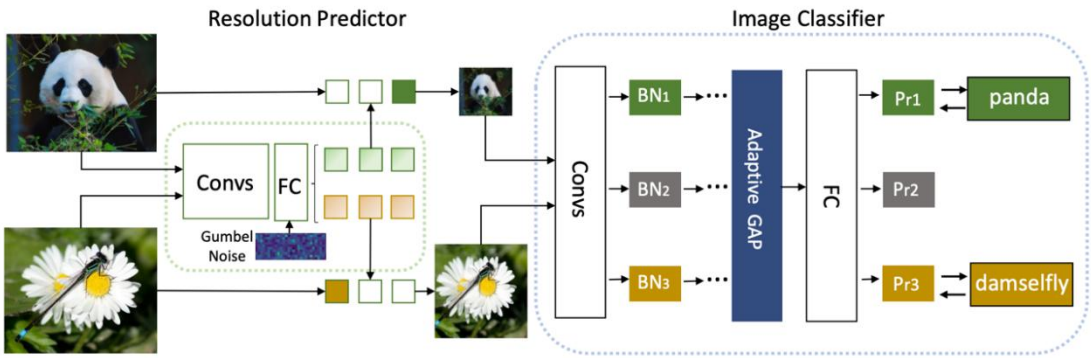
当前，Transformer 正在成为多种领域普遍使用的神经网络架构，并在多种任务上取得了最佳性能，但是目前对于 Transformer 的 Post-hoc 解释还停留在单层自注意力解释性上。3 月，以色列特拉维夫大学和 Facebook 的研究者提出了一种解释任意 Transformer 架构模型预测结果的计算范式，通过累计多层注意力图的的信息传播，捕捉输入中对输出影响最大的部分。该研究针对多种不同的 Transformer 架构都提出了对应的注意力传播计算方法，包括自注意力、自注意力和互注意力组合、自编码的注意力等。论文地址：https://arxiv.org/pdf/2103.15679.pdf。

神经网络算法持续改进优化，降低算力依赖并提升任务性能

基于神经网络结构的深度学习算法在过去十年取得了惊人的进展，但算力投入大、依赖大量标注数据的问题依然存在。为了解决这些问题，在现有神经网络算法架构上的改进仍有发展空间。

华为诺亚实验室等研究者提出动态分辨率网络 DRNet

深度卷积神经网络通畅采用精细的设计，有着大量的可学习参数，在视觉任务上实现很高精确度要求。为了降低将网络部署在移动端成本较高的问题，近来发掘在预定义架构上的冗余已经取得了巨大的成果，但对于 CNN 输入图像清晰度的冗余问题还没有被完全研究过，即当前输入图像的清晰度都是固定的。10 月，华为诺亚实验室、中国科学院大学等机构研究者提出一种新型的视觉神经网络 DRNet（Dynamic Resolution Network）。基于每个输入样本，该网络可以动态地决定输入图像的清晰度。该网络中设置了一个清晰度预测器，其计算成本几乎可以忽略，能够和整个网络共同进行优化。该预测器可以对图像学到其需要的最小清晰度，甚至能够实现超过过去识别准确率的性能。实验结果显示，DRNet 可以嵌入到任何成熟的网络架构中，实现显著的计算复杂度降低。例如，DR-ResNet-50 在实现同样性能表现的前提下可以降低 34%的计算，相比 ResNet-50 在 ImageNet 上提升 1.4 个点的性能同时能够降低 10%的计算。



图注：DRNet 的架构

来源：https://arxiv.org/pdf/2106.02898.pdf

Model	α	Params	FLOPs	↓FLOPs	Acc@1	Acc@5
ResNet-50-baseline	-	25.6 M	4.1 G	-	76.1%	92.9%
DR-ResNet-50	-	30.5 M	3.7 G	10%	77.5%	93.5%
DR-ResNet-50	2.0	30.5 M	2.3 G	44%	75.3%	92.2%
DR-ResNet-50	2.5	30.5 M	2.7 G	34%	76.2%	92.8%
DR-ResNet-50	3.0	30.5 M	3.2 G	22%	77.0%	93.2%
DR-ResNet-50	3.5	30.5 M	3.7 G	10%	77.4%	93.5%
ResNet-101-baseline	-	44.5 M	7.8 G	-	77.4%	93.5%
DR-ResNet-101	-	49.4 M	7.0 G	10%	79.0%	94.3%

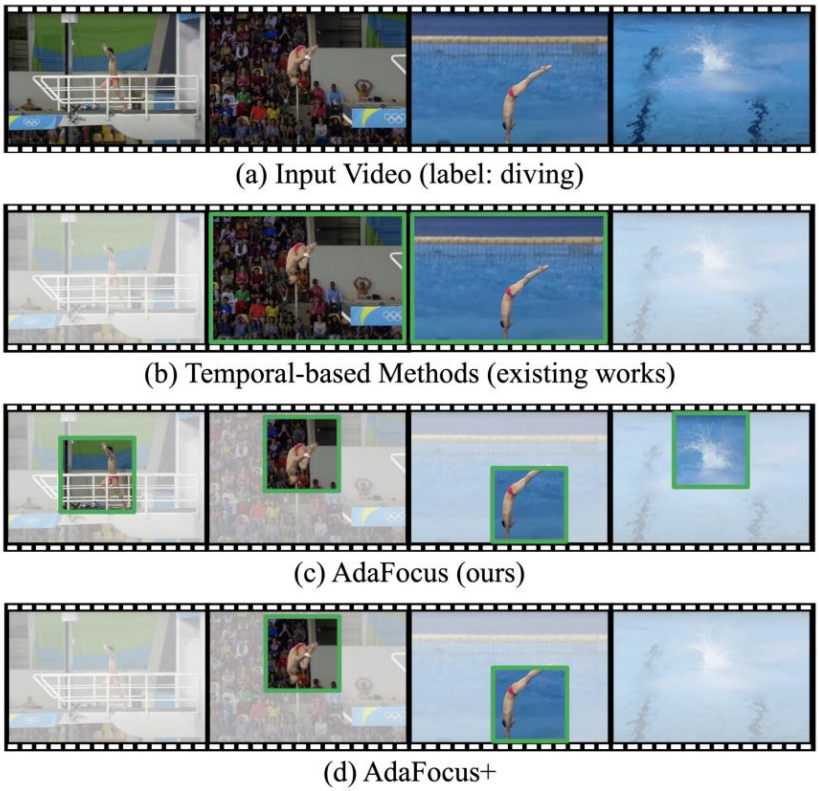
图注：ResNet-50 和 ResNet-101 在 ImageNet-1K 数据集上的性能表现以及与 DRNet 的对比

来源：https://arxiv.org/pdf/2106.02898.pdf

智源、清华研究者提出时空自适应动态神经网络 AdaFocus

10 月，来自北京智源人工智能研究院和清华大学的研究者提出高效视频分析框架 AdaFocus。该方法突破了传统深度神经网络的静态推理范式，实现了在时间、空间两个维度自适应定位与目标任务相关性最强的视频帧和关键区域，有效降低了基于深度视频分析方法的计算冗余性。在 Sth-Sth V2、ActivityNet 等主流视频处理数据集上，AdaFocus 可将模型总体推理

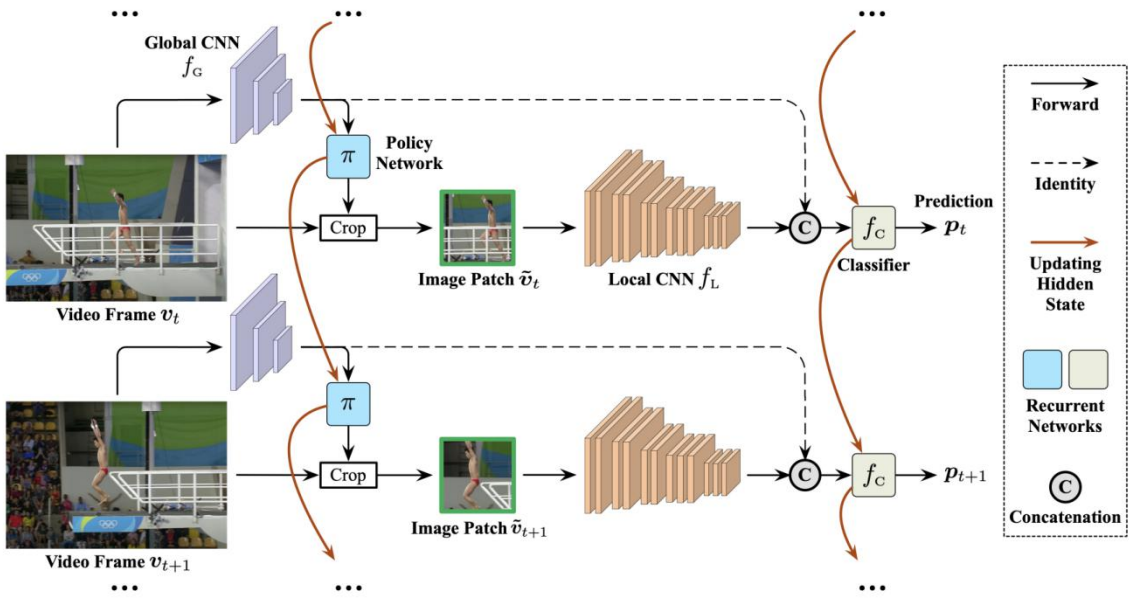
效率相较现有方法提高 2-3 倍。该方法在边缘计算、视频监控、视频推荐等场景有较大的应用前景，也为设计低延迟、低能耗的深度学习基础模型提供了启发性的思路。



图注：AdaFocus 与现有方法的对比

来源：

https://openaccess.thecvf.com/content/ICCV2021/papers/Wang_Adaptive_Focus_for_Efficient_Video_Recognition_ICCV_2021_paper.pdf



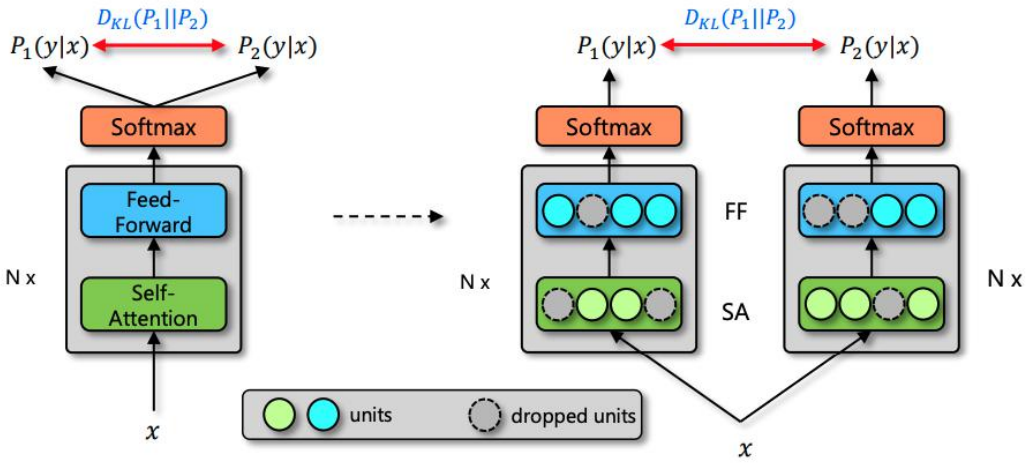
图注：AdaFocus 网络结构示意图

来源：

https://openaccess.thecvf.com/content/ICCV2021/papers/Wang_Adaptive_Focus_for_Efficient_Video_Recognition_ICCV_2021_paper.pdf

微软亚洲研究院研究者提出 R-Drop

10 月，微软亚洲研究院的研究者提出了 R-Drop 方法，用于解决 Dropout 的随机性导致训练和推理产生不一致结果的问题。通过最小化两个经由 Dropout 采样的子模型输出分布的双向 KL-散度，R-Drop 能够强制不同的子模型产生的结果有着相同的分布。从实验结果来看。R-Drop 可以提升 ViT、RoBERTa-large、BART 等模型的微调表现，并在 Vanilla Transformer 上实现了 WMT14 英译德任务的最佳表现，性能甚至超过许多超大规模预训练模型。

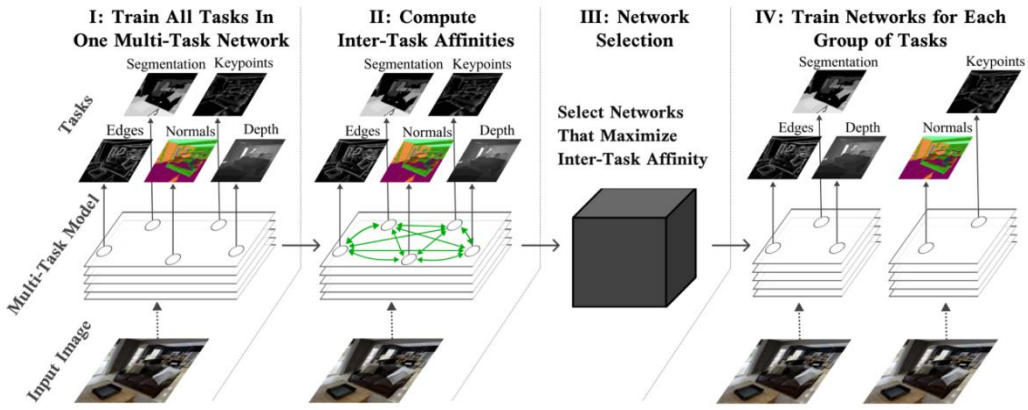


图注：R-Drop 的原理

来源：https://arxiv.org/pdf/2106.14448.pdf

谷歌研究者提出多任务训练策略 TAG

多任务学习能够让模型通过在一个任务上学习信息，提升在其他任务上训练的性能。然而，简单地让模型在所有任务上一块训练可能导致模型性能的下降，且完全搜索所有的任务组合的成本很高。因此，高效地找到对于训练有提升的任务是一个重要的研究问题。10 月，谷歌的研究者提出了名为 TAG（Task Affinity Groupings）的多任务训练策略，能够通过一次运行训练所有任务，并量化单个任务的梯度对于其他任务损失的影响。通过在视觉任务上的实验，研究者发现这一方法相比单纯同时训练所有任务降低了 10%的测试损失，并且比当前最佳的任务分组策略快 11.6 倍。



图注：TAG 方法的流程

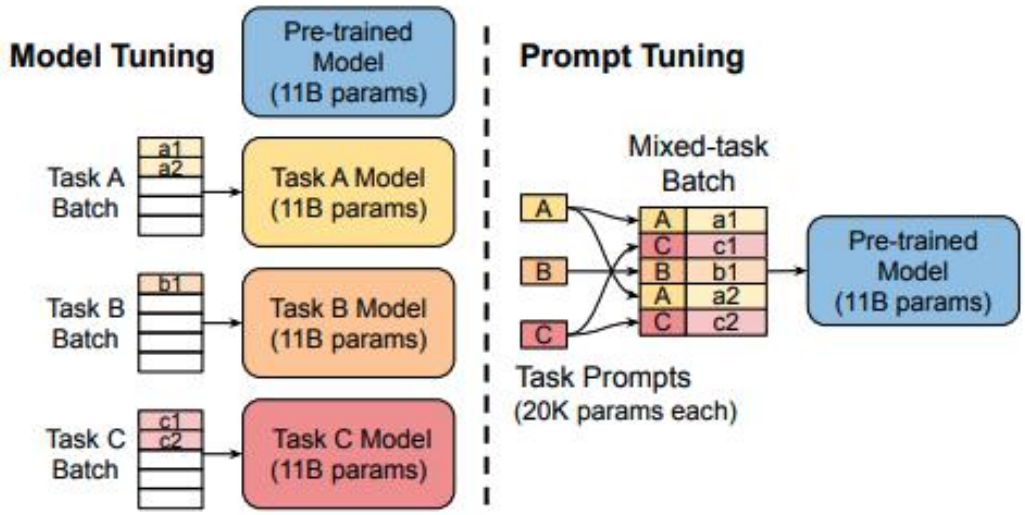
来源：https://arxiv.org/pdf/2109.04617.pdf

自然语言处理

Prompt Tuning 成为预训练语言模型新型训练范式

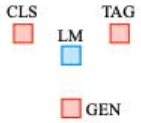
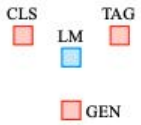
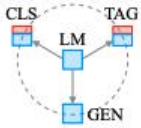
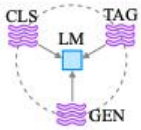
2018 年以来，预训练模型的体量不断增大，如 T5、GPT-3、悟道等，大模型成为 NLP 领域一项非常重要的技术突破。但是，预训练模型精调过程中所需的硬件和数据需求也在不断增长，丰富的下游任务也使微调阶段的设计更为复杂。为了解决这些问题，新型的“预训练-精调”正在快速发展，其中 Prompt Tuning 等方法已经崭露头角，成为当前的研究热点。

Prompt Tuning 是一类对预训练模型进行精调的方式，将人为的规则给到预训练模型，使模型更好地理解人的指令的技术，可简单理解为给任务的输入加入补充文本。



图注：模型微调和 Prompt Tuning 的区别

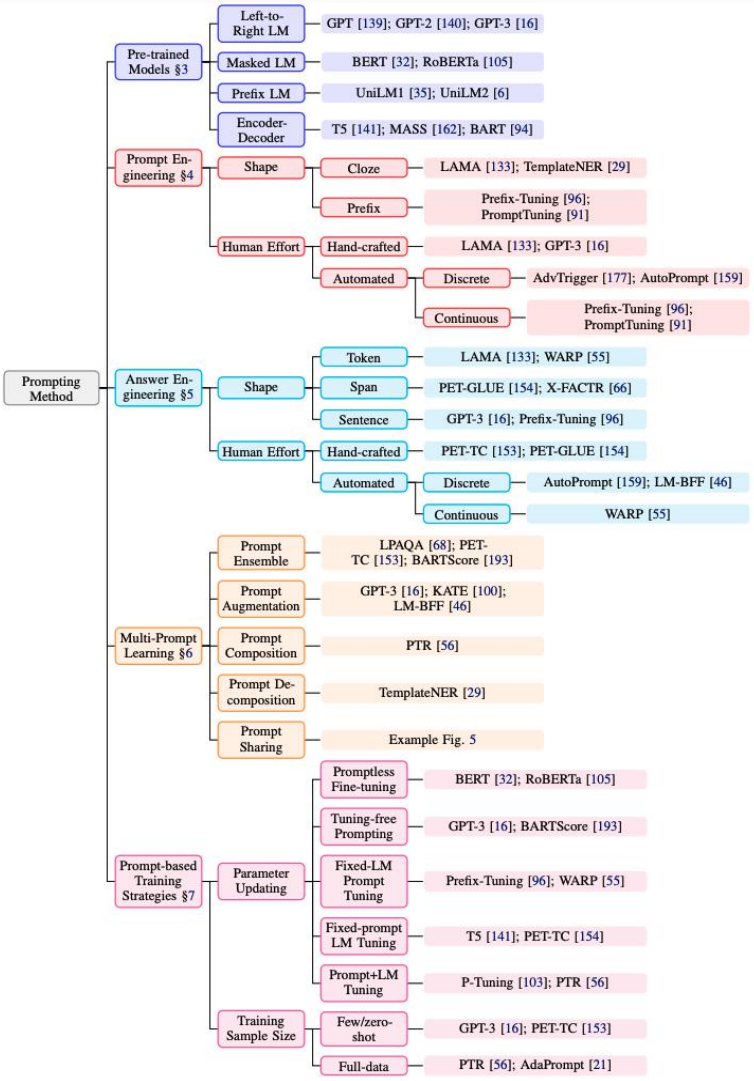
来源：https://arxiv.org/pdf/2104.08691

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

图注：NLP 研究领域经历的四种范式变革，包括完全监督学习（非神经网络）、完全监督学习（神经网络）、预训练-微调，以及 Pre-train, Prompt, Predict

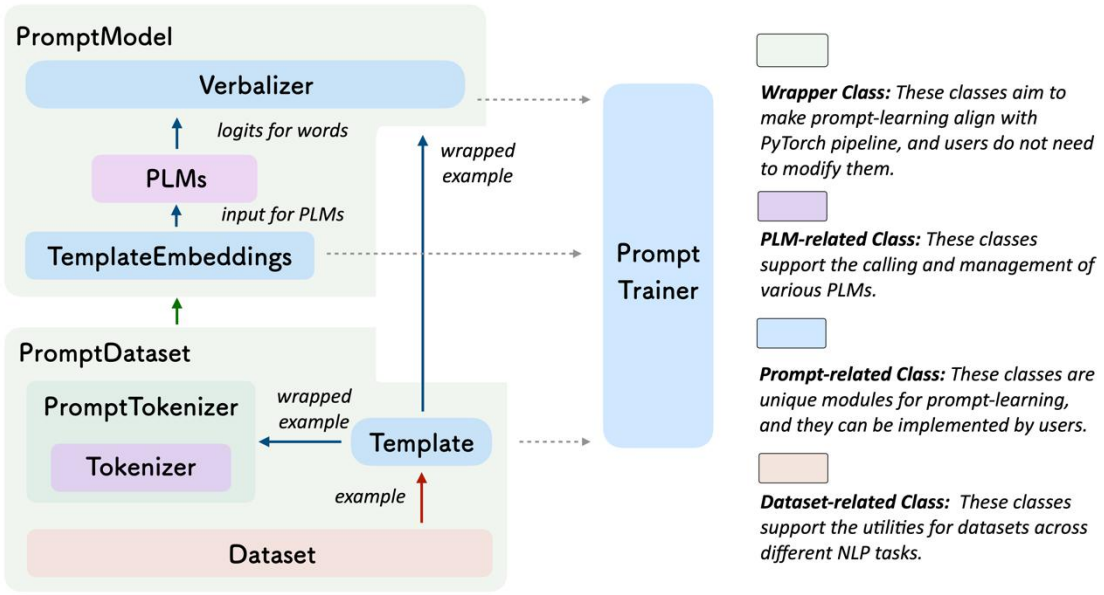
来源：https://arxiv.org/pdf/2107.13586.pdf

相比 Fine-tune 范式，Prompt Tuning 使预训练模型在下游任务中发挥更好的性能，实现更深层次的参数共享；同时，几乎所有 NLP 任务都能够放在 Zero-shot 情境下处理，不再需要训练样本；此外，目前也有不少研究表明 Prompt 在 Few-shot 情景下的性能也更好。清华大学副教授刘知远认为：“Prompt Tuning 也许会是深度学习时代的 Feature Engineering 问题，如何给各大任务设计合理的 Prompts 将会是很有意思的科学问题。”此外，清华大学还发布了统一范式的 Prompt-learning 编程工具包 OpenPrompt，以推动此领域的发展。



图注：Prompt Tuning 领域的方法一览

来源：https://arxiv.org/pdf/2107.13586.pdf

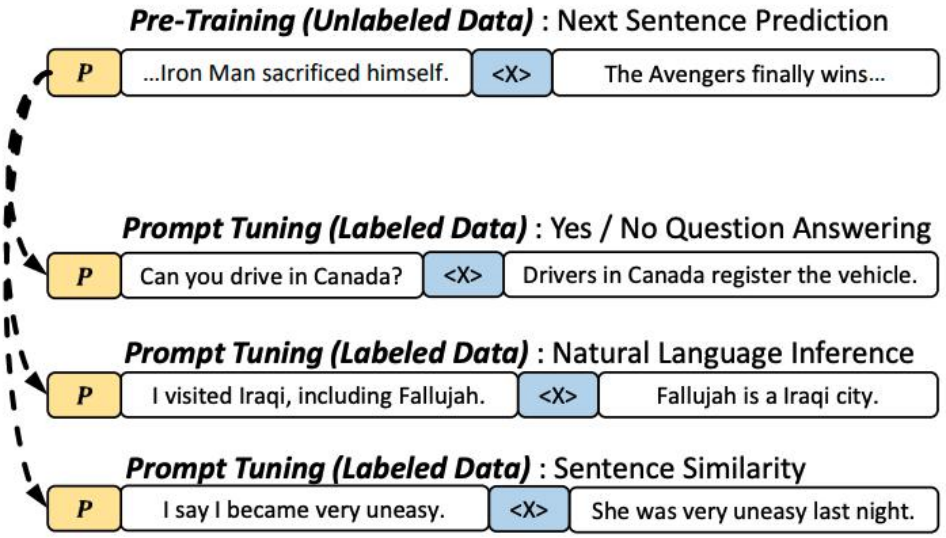


图注：OpenPrompt 工具包架构图

来源：https://github.com/thunlp/OpenPrompt

智源、清华大学研究者提出 Pre-trained Prompt Tuning 框架

9 月，清华大学黄民烈、刘知远等研究者提出了名为 Pre-trained Prompt Tuning（PPT）的方法，目的是为了改进 Prompt Tuning 在小样本任务上的性能弱于 Fine-tune 方法的问题。研究者将分类任务（sentence-pair、multiple-choice、single-text）都统一到一种任务中，并预训练 Soft Prompt。测试中，研究者采用了 T5、mT5 和 CPM-2 三种模型，对比了微调和多种 Prompt Tuning 训练策略的结构。实验表明，PPT 在大多数任务上具有明显的性能优势。



图注：PPT 处理文本的方法

来源：https://arxiv.org/pdf/2109.04332.pdf

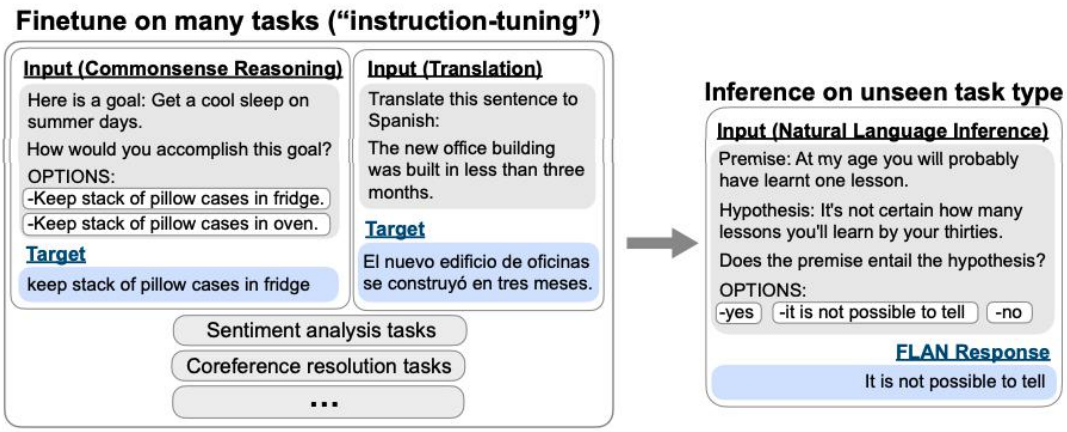
	Model	Method	SST-2 Acc.	SST-5 Acc.	RACE-m Acc.	RACE-h Acc.	BoolQ Acc.	RTE Acc.	CB F1
English Tasks									
FT (11B)	T5-Small	-	$72.8_{3.1}$	$31.1_{0.4}$	$26.4_{0.6}$	$26.3_{0.5}$	$59.2_{0.6}$	$54.0_{1.7}$	$70.1_{4.6}$
	T5-Base	-	$74.6_{2.7}$	$28.8_{1.8}$	$27.2_{0.5}$	$26.7_{0.2}$	$61.9_{2.1}$	$56.1_{2.3}$	$70.4_{2.6}$
	T5-Large	-	$89.1_{2.2}$	$42.4_{1.2}$	$48.2_{1.6}$	$43.2_{1.7}$	$74.6_{0.9}$	$64.4_{3.4}$	$82.3_{2.2}$
	T5-XL	-	$89.6_{3.2}$	$38.4_{5.1}$	$55.0_{2.8}$	$50.9_{2.6}$	$77.2_{2.1}$	$62.3_{6.8}$	$81.9_{9.0}$
	T5-XXL	-	$91.4_{0.8}$	$40.6_{2.0}$	$\mathbf{62.9_{3.9}}$	$\mathbf{54.8_{3.0}}$	$80.8_{2.4}$	$64.1_{2.0}$	$\mathbf{86.5_{5.3}}$
PT (410K)	T5-XXL	Vanilla PT	$70.5_{15.5}$	$32.3_{8.3}$	$34.7_{8.2}$	$31.6_{3.5}$	$61.0_{5.3}$	$53.5_{3.5}$	$50.7_{4.1}$
		Hybrid PT	$87.6_{6.6}$	$40.9_{2.7}$	$53.5_{8.2}$	$44.2_{6.4}$	$79.8_{1.5}$	$56.8_{2.6}$	$66.5_{7.2}$
		LM Adaption	$77.6_{7.5}$	$36.2_{3.6}$	$27.3_{0.2}$	$26.5_{0.4}$	$62.0_{0.3}$	$55.3_{1.0}$	$61.2_{1.7}$
		PPT	$93.5_{0.3}$	$\mathbf{50.2_{0.7}}$	$60.0_{1.2}$	$53.0_{0.4}$	$66.43_{5.7}$	$58.9_{1.6}$	$71.2_{6.2}$
		Hybrid PPT	$93.8_{0.1}$	$50.1_{0.5}$	$62.5_{0.9}$	$52.2_{0.7}$	$\mathbf{82.0_{1.0}}$	$59.8_{3.2}$	$73.2_{7.0}$
		Unified PPT	$\mathbf{94.4_{0.3}}$	$46.0_{1.3}$	$58.0_{0.9}$	$49.9_{1.3}$	$76.0_{2.7}$	$\mathbf{65.8_{2.1}}$	$82.2_{5.4}$
Chinese Tasks									
	Model	Method	ChnSent Acc.	Amazon Acc.	CCPM Acc.	C^3 Acc.	LCQMC Acc.	CMNLI Acc.	OCNLI Acc.
FT (11B)	mT5-Small	-	$76.1_{2.6}$	$29.9_{1.9}$	$31.9_{1.2}$	$29.6_{0.5}$	$52.4_{2.5}$	$36.5_{0.2}$	$34.9_{1.3}$
	mT5-Base	-	$78.2_{0.6}$	$36.4_{0.9}$	$40.4_{6.8}$	$29.4_{0.6}$	$50.9_{1.0}$	$36.3_{0.5}$	$35.4_{0.6}$
	mT5-Large	-	$79.1_{0.6}$	$31.0_{1.4}$	$46.0_{4.0}$	$29.9_{0.8}$	$52.1_{0.6}$	$35.8_{1.2}$	$35.2_{1.1}$
	mT5-XL	-	$82.7_{2.6}$	$35.5_{1.7}$	$68.3_{5.1}$	$29.7_{1.2}$	$52.9_{2.4}$	$36.8_{1.6}$	$35.6_{0.5}$
	mT5-XXL	-	$83.6_{1.5}$	$42.1_{0.8}$	$79.7_{1.1}$	$37.2_{3.3}$	$53.1_{1.0}$	$39.0_{0.4}$	$37.4_{1.2}$
	CPM-2	-	$86.1_{1.8}$	$42.5_{2.0}$	$81.8_{1.6}$	$38.4_{3.7}$	$58.8_{1.8}$	$40.7_{1.0}$	$38.5_{1.5}$
PT (410K)	CPM-2	Vanilla PT	$62.1_{3.1}$	$30.3_{4.8}$	$31.0_{9.7}$	$28.2_{0.4}$	$51.5_{3.4}$	$35.4_{0.5}$	$37.0_{0.5}$
		Hybrid PT	$79.2_{4.0}$	$39.1_{3.8}$	$46.6_{15.0}$	$29.2_{0.5}$	$54.6_{2.3}$	$37.1_{0.6}$	$37.8_{1.4}$
		LM Adaption	$74.3_{5.2}$	$35.2_{2.4}$	$33.7_{12.8}$	$30.2_{1.5}$	$51.4_{2.9}$	$35.1_{0.3}$	$38.0_{1.1}$
		PPT	$90.1_{0.8}$	$48.6_{0.6}$	$\mathbf{85.4_{0.6}}$	$43.8_{2.2}$	$59.1_{0.6}$	$\mathbf{43.0_{0.5}}$	$40.1_{0.4}$
		Hybrid PPT	$89.5_{0.3}$	$\mathbf{48.8_{2.0}}$	$83.9_{0.5}$	$46.0_{0.5}$	$\mathbf{67.3_{0.9}}$	$41.3_{0.8}$	$38.7_{0.6}$
		Unified PPT	$\mathbf{90.7_{0.2}}$	$44.6_{1.1}$	$83.4_{0.9}$	$\mathbf{50.2_{0.6}}$	$55.0_{0.4}$	$40.6_{0.4}$	$\mathbf{41.5_{1.5}}$

图注：PPT 相比微调和其他 Prompt Tuning 方法的性能对比

来源：https://arxiv.org/pdf/2109.04332.pdf

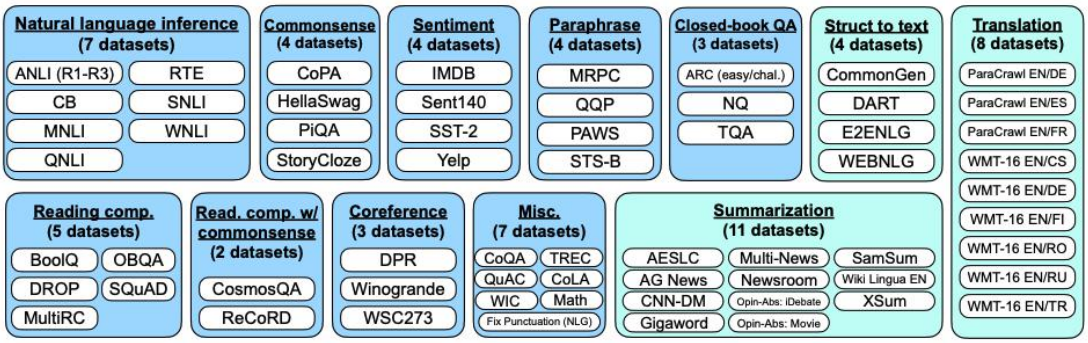
谷歌研究者提出指令微调技术 Instruction Tuning

10 月，谷歌 Quoc V. Le 团队提出了一种名为 Instruction Tuning（指令微调）的新型技术，本质是将 NLP 任务转换为自然语言指令，再将其投入模型进行训练，通过给模型提供指令和选项的方式，使其能够提升零样本任务的性能表现。



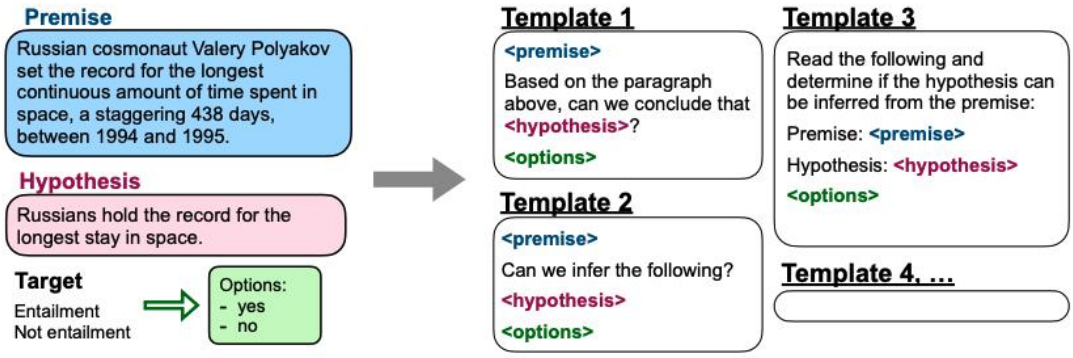
图注：Instruction Tuning 的具体流程

来源：https://arxiv.org/pdf/2109.01652.pdf



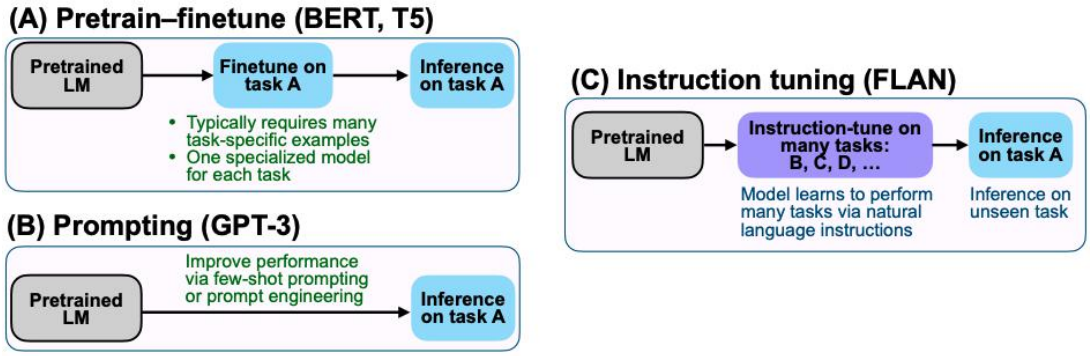
图注：指令微调论文中采用的数据集以及数据集所归属的任务类型

来源：https://arxiv.org/pdf/2109.01652.pdf



图注：将数据集构建为模板的流程

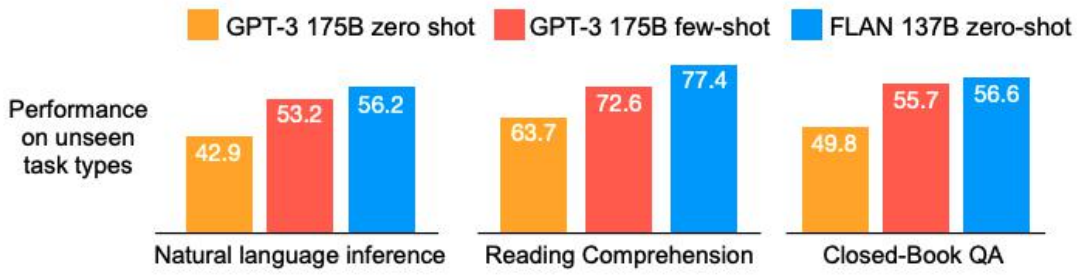
来源：https://arxiv.org/pdf/2109.01652.pdf



图注：指令微调和预训练-微调、Prompt 等方法的区别

来源：https://arxiv.org/pdf/2109.01652.pdf

为了进行测试，研究者构建了一个名为“Finetuned LAnguage Net（FLAN）”的预训练模型，基于 Transformer 架构，有 1370 亿参数。从测试结果来看，FLAN 在 25 个零样本任务上超越了 GPT-3 的性能表现。此外，在一些任务中，该模型的零样本任务性能甚至超过 GPT-3 的小样本性能。

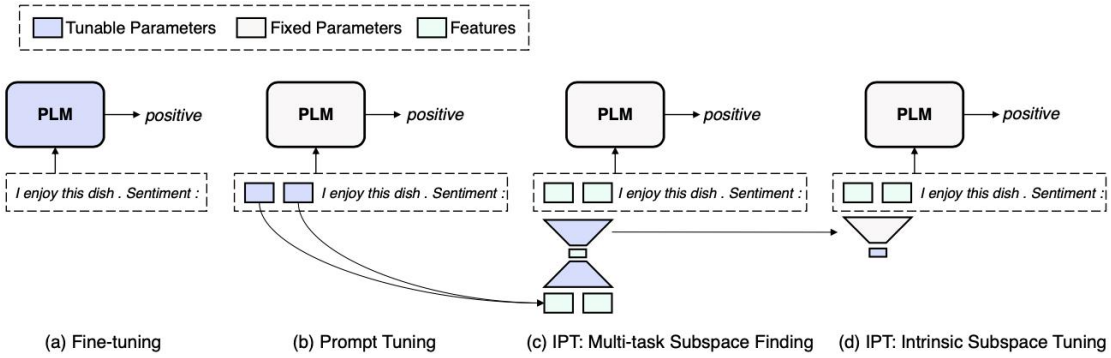


图注：FLAN 模型和 GPT-3 在一些零样本、小样本任务中的性能对比

来源：https://arxiv.org/pdf/2109.01652.pdf

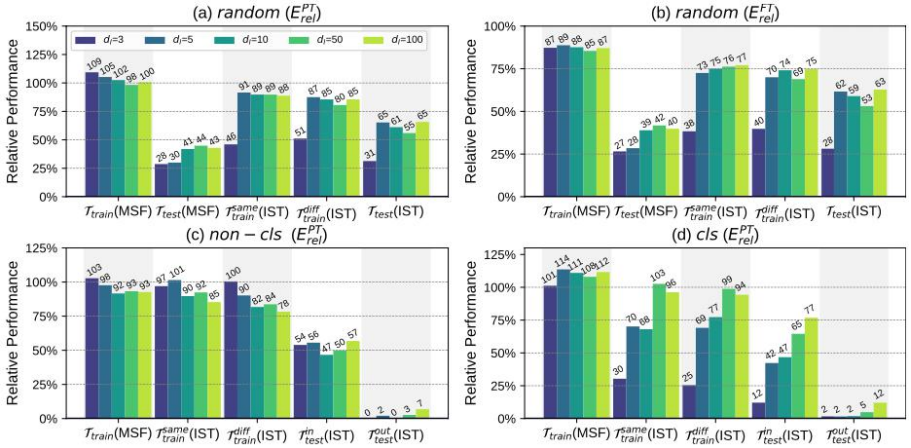
清华大学研究者提出 Prompt 迁移框架 IPT

10 月，清华大学刘知远、李涓子等研究者研究了 Prompt 的迁移性，探索如何更好的利用训练完的 Prompt 在不同任务、模型之间做迁移，并减少 Prompt Tuning (PT)的训练开销以及提升性能。具体提出了名为 Intrinsic Prompt Tuning (IPT)的框架，将各类 NLP 任务的优化过程在同一个低维本征子空间进行刻画。研究人员对 120 个 NLP 任务进行实验，成功找到了这样一个子空间，并发现在其中仅微调一个 5 维的向量，可以分别在从未见过的训练数据和新任务上进行迁移，达到 87%和 65% Prompt Tuning 的性能，并具有极高的训练稳定性。



图注：IPT 分析框架的训练流程

来源：https://arxiv.org/pdf/2110.07867.pdf



图注：IPT 框架在微调不同维度的向量时，还原 Prompt Tuning 性能的比例

来源：https://arxiv.org/pdf/2110.07867.pdf

提升性能和效率成为预训练语言模型发展的新路线

超大参数规模的预训练模型有着海量算力的需求，对于中小型机构而言难以承担。当参数规模的提升达到顶峰后，研究者们将开始思考在现有的参数规模下提升模型的性能和效率，降低训练和推理的难度。9 月，OpenAI 创始人兼 CEO 萨姆·阿尔特曼（Sam Altman）表示，下一代预训练模型 GPT-4 的参数可能小于 GPT-3 的 1750 亿，研究者将主要探索在相对较小的参数规模下获得更大的收益。不追求极致的参数规模，或将是下一阶段预训练模型领域的发展重点。

澜舟科技等研发中文语言模型孟子

7 月，澜舟科技-创新工场团队与上海交通大学、北京理工大学等单位联合研发了中文语言模型“孟子”，参数规模仅 10 亿，在 CLUE 中文理解评测的总排行榜，以及分类排行榜和阅读理解排行榜均位列榜首。其中，总排行榜分数突破 84 分，逼近人类基准分数（85.61）。

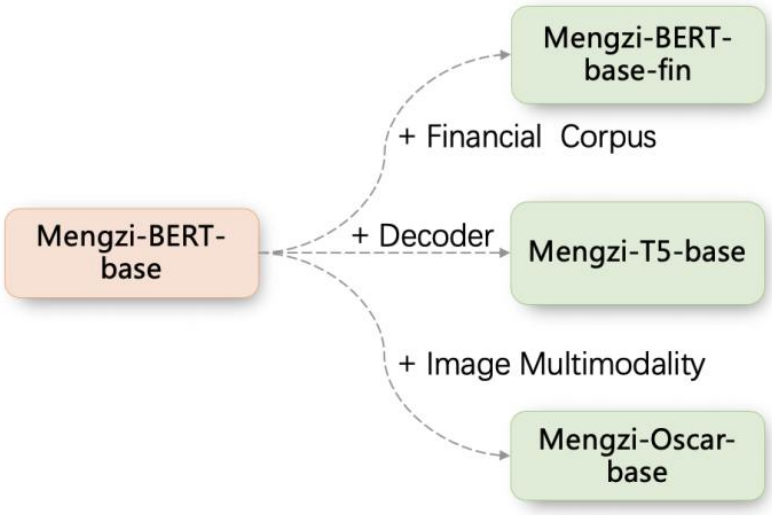
CLUE 排行榜

排名	模型	规模	总分	AFQMC	TNEWS	IFLYTEK	OCNLI	WSC2020	CSL	CMRC2018	CHID	C3
1	孟子	十亿	82.90	79.82	64.68	65.08	81.87	96.55	89.87	82.25	96.00	89.98
2	Motian	十亿	82.15	78.30	57.42	65.46	84.97	94.83	90.17	85.30	94.43	88.49
3	BERTSG	百亿	81.80	79.85	57.42	64.54	85.93	95.17	89.00	83.80	93.06	87.44
4	Pangu	千亿	81.18	78.11	57.42	65.19	83.30	95.52	87.73	84.45	93.25	85.64
	人类水平		86.68	81.00	71.00	80.30	90.30	98.00	84.00	92.40	87.10	96.00

排名截至2021年7月30日

图注：“孟子”模型在 CLUE 总榜的排名情况，截至 2021 年 7 月 30 日

来源：澜舟科技官网



图注：孟子系列模型

来源：https://arxiv.org/pdf/2110.06696.pdf

微软提出预训练模型 T-NLRv5

12 月，微软提出大规模预训练模型 T-NLRv5。该模型在 GLUE 和 SuperGLUE 基准总榜上位列榜首，并在 MNLI-m（Multi-Genre Natural Language Inference-matched）和 RTE（Recognizing Textual Entailment）两项测试上超过人类水平。

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Microsoft Alexander v-team	Turing NLR v5		91.2	72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9	94.1	95.9	57.0
2	ERNIE Team - Baidu	ERNIE		91.1	75.5	97.8	93.9/91.8	93.0/92.6	75.2/90.9	92.3	91.7	97.3	92.6	95.9	51.7
3	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.5	95.2	49.1
4	liangzhu ge	DEBERTa + CLEVER		90.9	73.9	97.5	92.8/90.4	93.2/92.9	76.4/90.9	92.1	91.7	96.7	93.1	96.6	35.2
5	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.2	94.5	53.2
6	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.0	94.5	52.6
+ 7	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.7	94.5	51.2
8	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.8	94.5	53.1
9	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
+ 10	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.3	94.5	47.9
+ 11	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.0	94.5	51.6
+ 12	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.8	91.8	50.7
+ 13	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.7	89.0	50.1
14	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.9	89.0	49.3
15	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.2	89.0	48.7
+ 16	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.3	89.0	42.8
17	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

图注：GLUE 榜单排名（12 月 3 日）

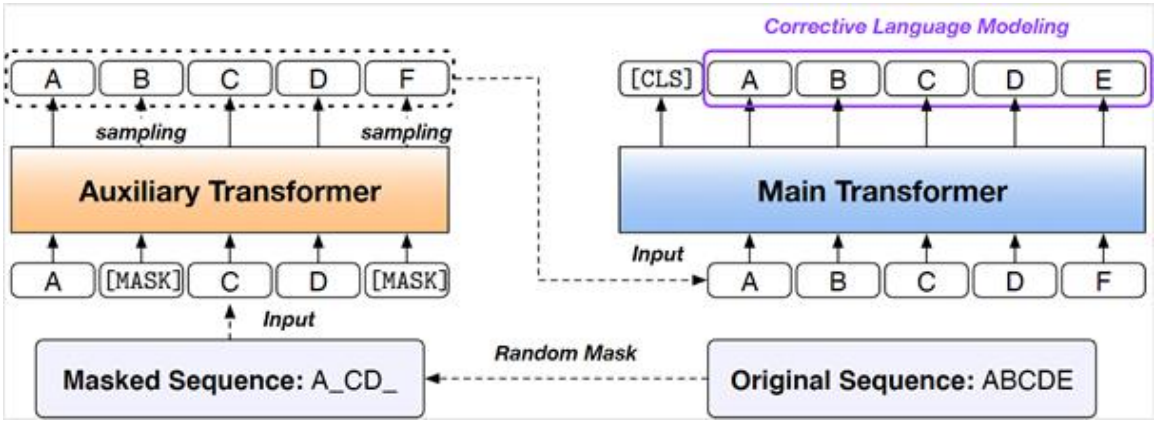
来源：https://gluebenchmark.com/leaderboard

Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
1	Microsoft Alexander v-team	Turing NLR v5		90.9	92.0	95.9/97.6	98.2	88.4/63.0	96.4/95.9	94.1	77.1	97.3	67.8	93.3/95.5
2	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
+ 3	Zirui Wang	T5 + UDG, Single Model (Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+ 4	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
5	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+ 6	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9

图注：SuperGLUE 榜单排名（12 月 3 日）

来源：https://super.gluebenchmark.com/leaderboard

T-NLRv5 基于 COCO-LM 模型结构，使用一个辅助 Transformer 语言模型来对其输入文本进行破坏，而在主 Transformer 模型中，采用纠正语言模型任务来预训练。为了提升模型的效率，研究者采用了为快速预训练 （FastPT）定制化的 CUDA 核，使用混合精度预训练。研究者还使用了 ZeRO 优化器，能够降低在多级并行预训练过程中的 GPU 内存。最终，T-NLRv5 的参数规模最大仅有 54 亿。

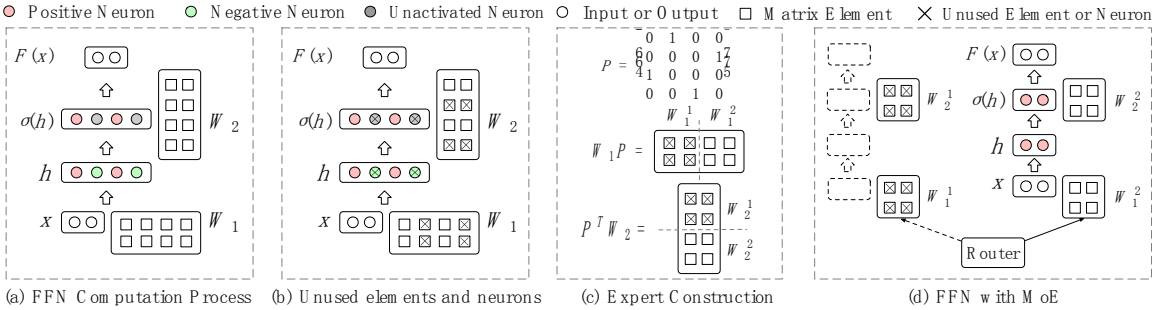


图注：T-NLRv5 的模型架构

来源：微软

清华、微信等提出预训练模型条件激活技术

10 月，清华大学、微信联合提出大模型加速技术 MoEfication，基于模型的稀疏激活现象，将 Transformer 计算中占主要部分的前馈网络转换为稀疏激活的专家网络（Mixture-of-Expert）。过参数化给大规模预训练模型带来了强大的学习能力，同时也造成了稀疏激活的现象。研究者发现大模型的计算过程中每次仅有 3%至 4%的神经元被激活。虽然每个神经元学习到了不同的信息，但对于一个特定的输入，仅有极少量信息被使用。MoEfication 技术利用了稀疏激活特性，显著减少了大模型推断过程中的计算开销。



图注：MoEfication 算法架构

来源：https://arxiv.org/pdf/2110.01786.pdf

计算机视觉

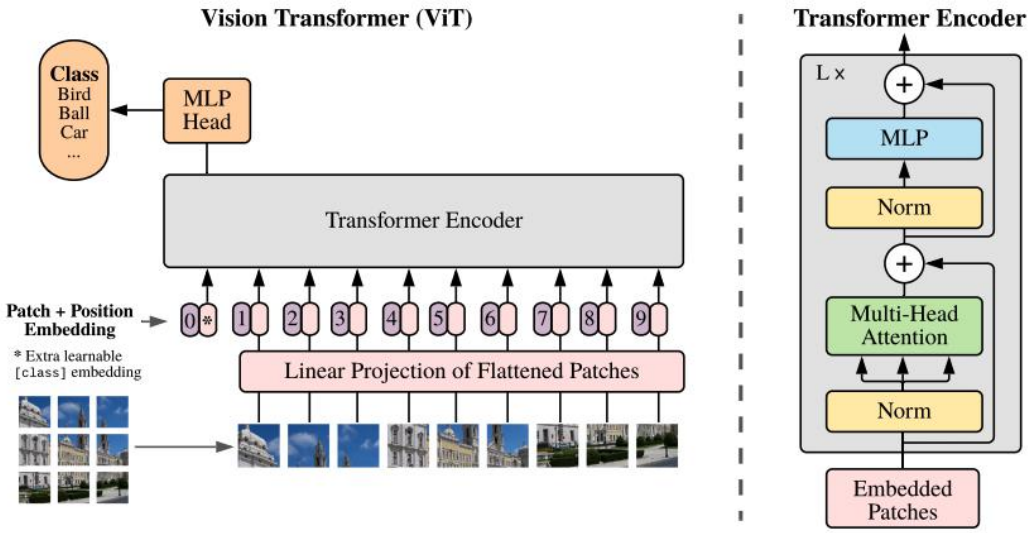
Transformer 成为计算机视觉领域的重要网络架构

由于 Transformer 架构在注意力机制建模方面的优势，许多研究者期待其在计算机视觉领域也能够取得进展。自预训练语言模型 BERT 问世后，许多研究者已经开始推动 BERT 在视觉领域实现应用，如 VL-BERT、VideoBERT 等。GPT-3 问世后，掀起了采用 Transformer 架构进行自然语言处理研究的热潮。今年，大量基于 Transformer 架构的网络涌现，为计算机视觉研究注入了新的活力。

谷歌大脑团队提出 ViT/Facebook 提出 MViT

2020 年 10 月，谷歌大脑团队首次尝试将标准 Transformer 应用于图像，提出了视觉 Transformer（ViT）模型，并在多个图像基准上接近甚至优于最佳性能。6 月，ViT 团队尝

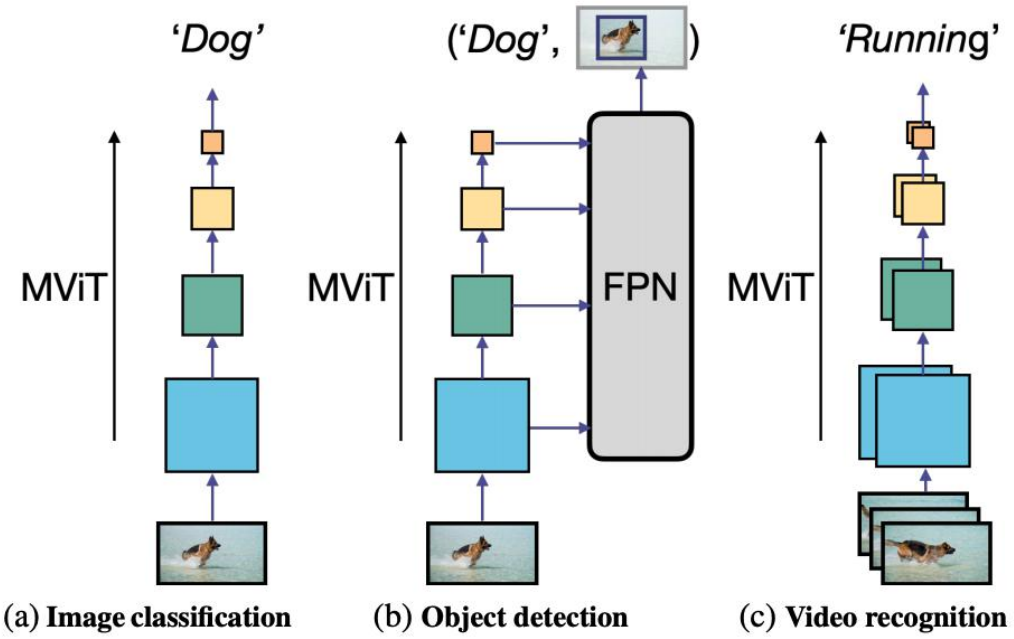
试将 ViT 模型进行扩展，训练出具有 20 亿参数的变体模型 ViT G/14，在 ImageNet 数据集上达到了新的最佳性能。



图注：ViT 模型的架构

来源：https://arxiv.org/pdf/2010.11929.pdf

此外，12 月，Facebook 和加州大学伯克利分校的研究者提出了 ViT 的改进版模型 MViT，在 ImageNet-1k 分类、COCO 目标检测和 Kinetics-400 视频分类三个基准上实现了最佳的性能。

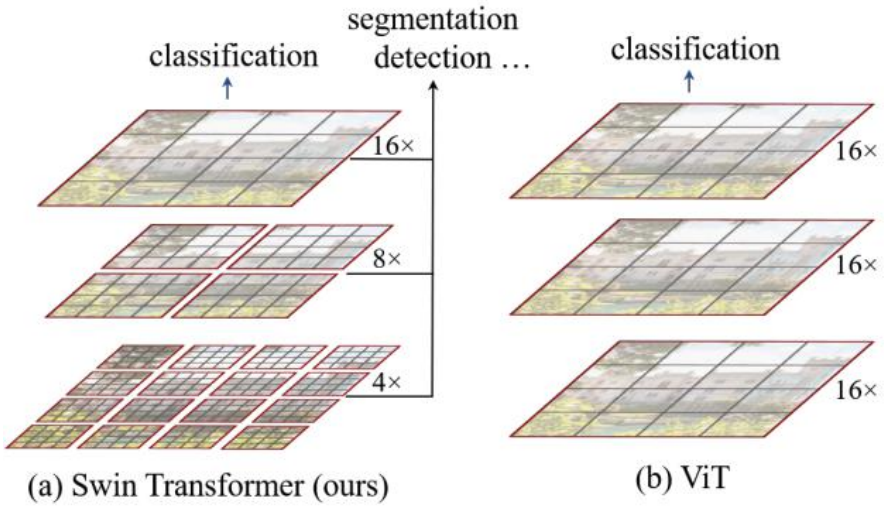


图注：MViT 在图像分类、目标检测、视频感知等任务上的架构

来源：https://arxiv.org/pdf/2112.01526.pdf

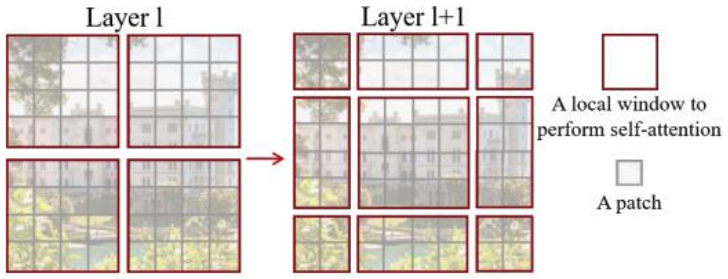
微软亚洲研究院研究者提出 Swin Transformer

8 月，微软亚洲研究院研究者提出了 Swin Transformer 的视觉架构，一是采用 CNN 中常用的层次化构建方式，构建层次化 Transformer；二是引入局部性（Locality）的思想，采用多种尺度的窗口，对无重合的窗口区域内进行 Self-Attention 计算。实验结果表明，Swin Transformer 在 COCO 的分割和检测任务以及 ADE20K 的语义分割任务上都超越了 CNN，达到了最佳性能。Swin Transformer 因其在计算机视觉领域的贡献获得 ICCV2021 最佳论文奖（马尔奖）。11 月，Swin Transformer 升级，可以训练分辨率达 1536x1536 的图像，在 4 个视觉基准上刷新纪录。



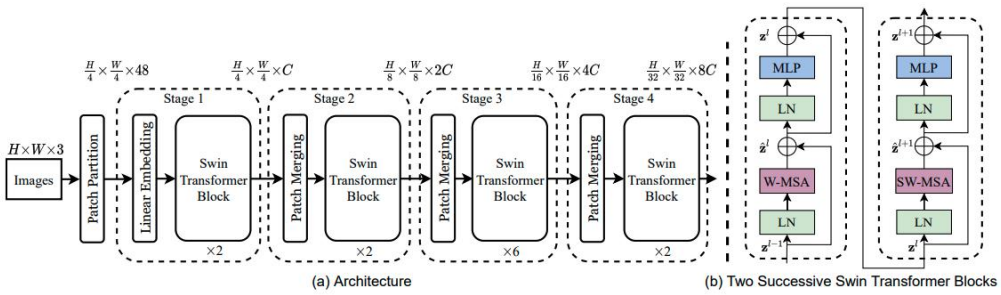
图注：Swin Transformer 采用的层次化构建方式，并采用了多种尺度的窗口进行图像采样

来源：https://arxiv.org/pdf/2103.14030.pdf



图注：Swin Transformer 中采用了多种尺度的窗口来捕捉图像信息

来源：https://arxiv.org/pdf/2103.14030.pdf

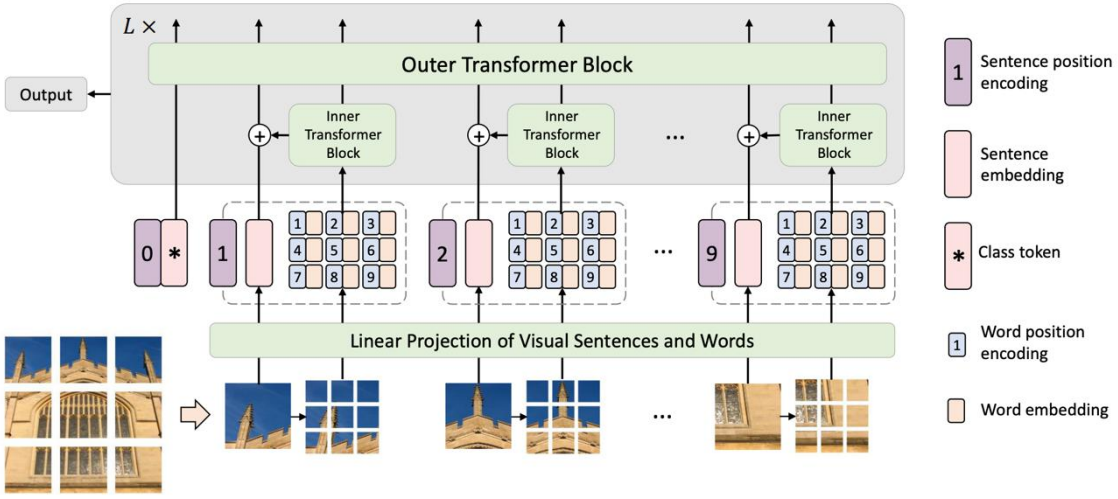


图注：Swin Transformer 的整体架构

来源：https://arxiv.org/pdf/2103.14030.pdf

华为诺亚实验室研究者提出 TNT 视觉架构

3 月，华为和诺亚实验室研究者提出名为 TNT（Transformer iN Transformer）视觉网络架构。该架构采用结构嵌套的 Transformer 结构，将图像切分为 Patch 后输入 Transformer 进行处理。与 ViT 不同的是，图像并不拉直为序列。TNT 将切分为 Patch 的图像视为“视觉句子”（Visual Sentence），将每个 Patch 进一步细分为更小的 Patch，视为“视觉单词”（Visual Word）。在预训练过程中，TNT 采用外 Transformer 架构对图像 Patch 之间进行关系建模，并采用内 Transformer 对每个 Patch 中的更细分的 Patch 之间的关系进行建模。这样能够显著提升模型对于图像局部结构的建模能力，提升了模型的性能。实验显示，TNT 在 ImageNet 和下游任务上的性能超越 DeiT、ViT 等架构的表现。此外，TNT 相比 DeiT 和 PVT 等视觉模型在相同推理速度和更高的精确度下具有更高计算效率。



图注：TNT 的总体架构

来源：https://arxiv.org/pdf/2103.00112.pdf

Model	Resolution	Params (M)	FLOPs (B)	Top-1	Top-5
CNN-based					
ResNet-50 [13]	224×224	25.6	4.1	76.2	92.9
ResNet-152 [13]	224×224	60.2	11.5	78.3	94.1
RegNetY-8GF [28]	224×224	39.2	8.0	79.9	-
RegNetY-16GF [28]	224×224	83.6	15.9	80.4	-
EfficientNet-B3 [32]	300×300	12.0	1.8	81.6	94.9
EfficientNet-B4 [32]	380×380	19.0	4.2	82.9	96.4
Transformer-based					
DeiT-Ti [35]	224×224	5.7	1.3	72.2	-
TNT-Ti	224×224	6.1	1.4	**73.9**	**91.9**
DeiT-S [35]	224×224	22.1	4.6	79.8	-
PVT-Small [40]	224×224	24.5	3.8	79.8	-
T2T-ViT_t-14 [44]	224×224	21.5	5.2	80.7	-
TNT-S	224×224	23.8	5.2	**81.5**	**95.7**
ViT-B/16 [10]	384×384	86.4	55.5	77.9	-
DeiT-B [35]	224×224	86.4	17.6	81.8	-
T2T-ViT_t-24 [44]	224×224	63.9	13.2	82.2	-
TNT-B	224×224	65.6	14.1	**82.9**	**96.3**

图注：TNT 架构相比其他视觉架构的性能表现

来源：https://arxiv.org/pdf/2103.00112.pdf

Model	Indices of TNT blocks	FLOPs (B)	Throughput (images/s)	Top-1
DeiT-S [35]	-	4.6	907	79.8
DeiT-B [35]	-	17.6	292	81.8
PVT-Small [40]	-	3.8	820	79.8
PVT-Medium [40]	-	6.7	526	81.2
TNT-S	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	5.2	428	81.5
TNT-S-1	[1, 4, 8, 12]	4.8	668	81.4
TNT-S-2	[1, 6, 12]	4.7	704	81.3
TNT-S-3	[1, 6]	4.7	757	81.1
TNT-S-4	[1]	4.6	822	80.8

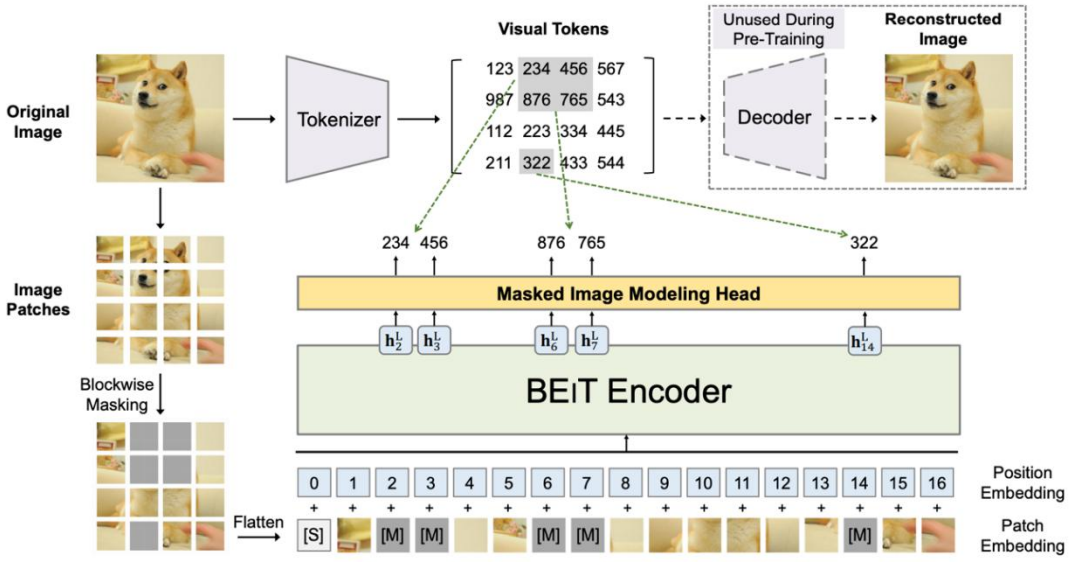
图注：TNT 模型和 DeiT、PVT 的模型在 GPU 计算效率上的对比

来源：https://arxiv.org/pdf/2103.00112.pdf

遮盖图像建模替代对比学习成为视觉自监督学习的新热点

微软研究者提出采用遮盖图像建模预训练的视觉模型 BEiT

6 月，微软研究者提出一种自监督视觉表示模型 BEiT，其借鉴了 BERT 预训练模型的自掩码预训练机制，提出了遮盖图像建模（Masked Image Modeling）任务。在预训练阶段，图像被分割为 Patch 并转换为视觉 Token。随后，研究者对默写图像 Patch 进行随机遮盖，并将其输入到 Transformer 网络中。预训练目标是让模型基于被遮盖过的图像 Patch 对原始的视觉 Token 进行重建。实验说明，Base-size 的 BEiT 模型在 ImageNet-1K 数据集上实现了 83.2% 的 Top-1 准确率。而更大规模的 BEiT 模型在仅使用 ImageNet-1K 数据集的情况下实现了 86.3%的准确率，超过了 ViT-L 模型在 ImageNet-22K 上的预训练表现。

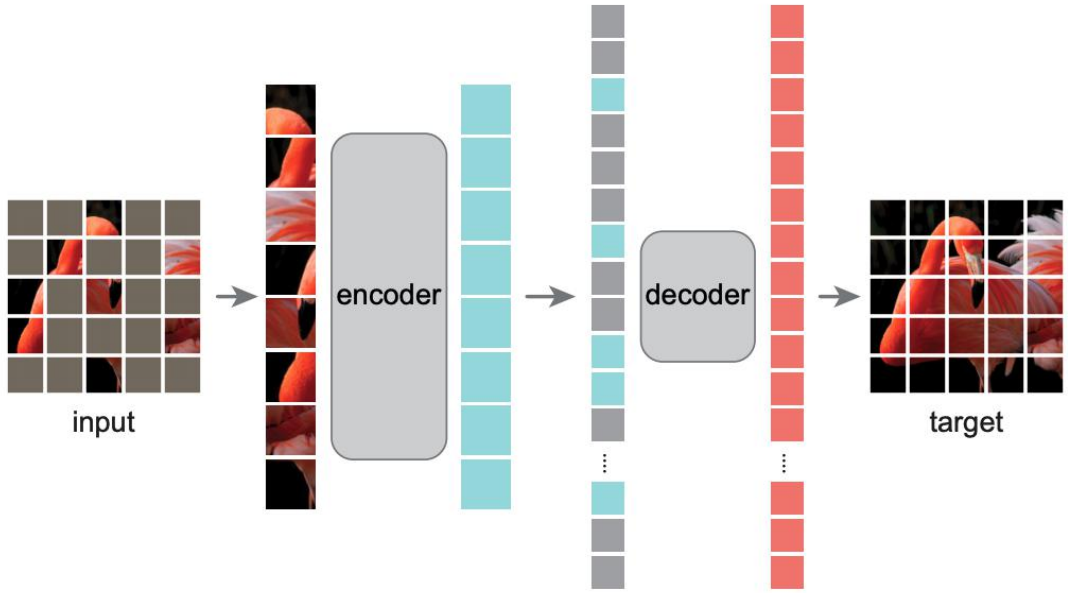


图注：BEiT 模型的预训练流程

来源：https://arxiv.org/pdf/2106.08254.pdf

Facebook 团队提出 Masked Autoencoder 掩码自动编码器方法

11 月，Facebook 何恺明团队提出了一种名为掩码自动编码器（Masked Autoencoder，MAE）的视觉训练方法。该方法在对于输入图像的局部进行了遮盖，并通过不对称的编码器-解码器结构对缺失像素进行重建。预训练后，撤除解码器，可将完整的图片输入编码器，使其完成视觉任务。实验结果显示，该方法在多种任务上都可以用更少的数据实现较高性能。



图注：掩码自动编码器的结构，编码器可在训练后用于下游视觉任务

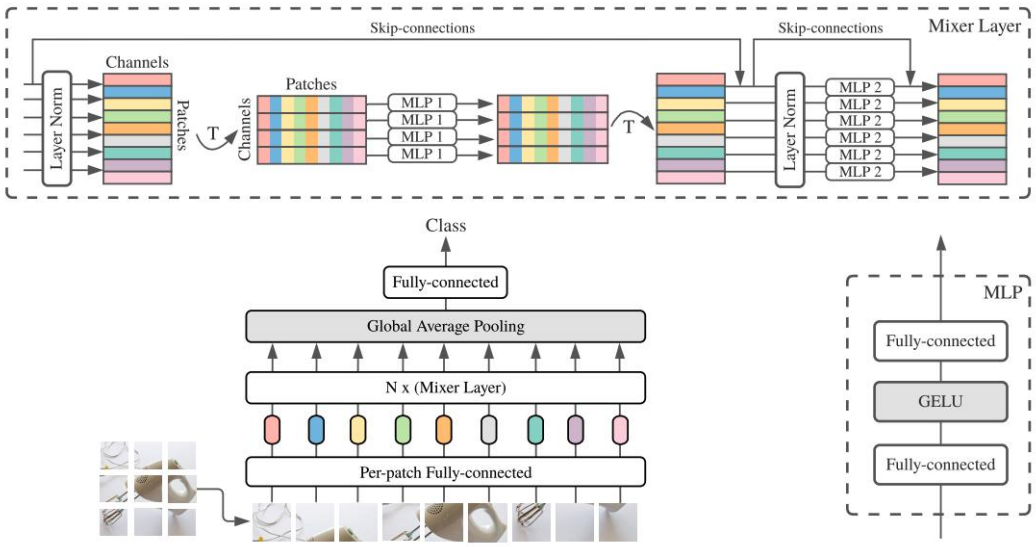
来源：https://arxiv.org/pdf/2111.06377.pdf

非 Transformer 架构在视觉任务上仍有发展潜力

尽管 Transformer 在计算机视觉领域体现出了优势，但许多研究者也在探索替代性的解决方案，以实现计算速度更快、效率更高的模型。近来的一些研究发现，多层感知机结构在视觉任务方面能够平衡性能和计算成本，对于实际应用而言仍是一种具有吸引力的方案。

谷歌研究者提出完全基于多层感知机的视觉架构 MLP-Mixer

6 月，谷歌研究团队提出了一种完全基于多层感知机的视觉网络架构 MLP-Mixer。MLP-Mixer 包括 Per-patch 线性嵌入、Mixer 层和分类器头。Mixer 层中包括一个 Token-mixing MLP 和一个 Channel-mixing MLP，采用两个 MLP 层和一个高斯误差线性单元（Gaussian Error Linear Units，GELU）。根据实验，MLP-Mixer 在图像分类任务上实现了接近最优的性能，在 ImageNet 数据集上取得了 87.94%的 Top-1 准确率。



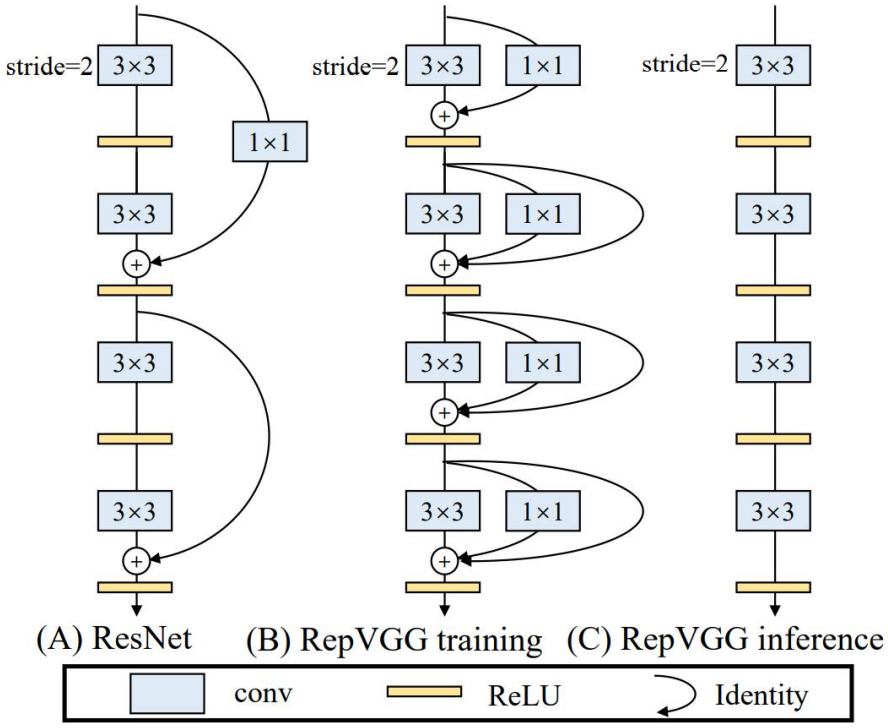
图注：MLP-Mixer 架构

来源：https://arxiv.org/pdf/2105.01601.pdf

清华、旷视研究者提出极简 CNN 设计新范式 RepVGG

清华大学、旷视在 CVPR 2021 上提出了一种高效极简卷积网络架构，称为 RepVGG，仅由 3x3 卷积和 ReLU 激活函数构成，甚至没有任何分支。应用一种创新的模型设计与优化方法，即“结构重参数化”，RepVGG 在训练时具有分支结构，可以在训练后等价转为无分支极简

结构，同时兼顾了多分支架构的高精度和单路架构的高推理效率。RepVGG 在图像分类和语义分割等任务上可以持平或超越 EfficientNet、RegNet 等主流 CNN 架构，甚至以更高的精度和速度超越 Swin Transformer 等部分最新的 Transformer 架构，展现了简单架构的生命力和结构重参数化方法的有效性。

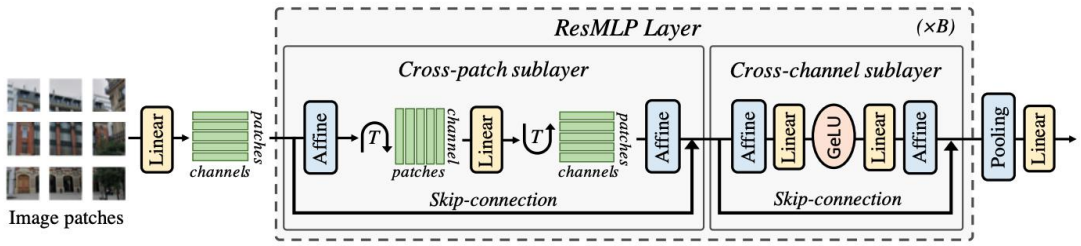


图注：RepVGG 架构

来源：https://arxiv.org/pdf/2101.03697.pdf

Facebook 研究者提出多层感知机架构 ResMLP

同月，Facebook 的研究者提出了 ResMLP，一种用于图像分类的多层感知机架构，不采用任何形式的注意力机制，仅包含 MLP 和高斯误差线性单元，不需要特定 Batch 或者跨通道的标准化（如 Batch-Norm 和 LayerNorm）。

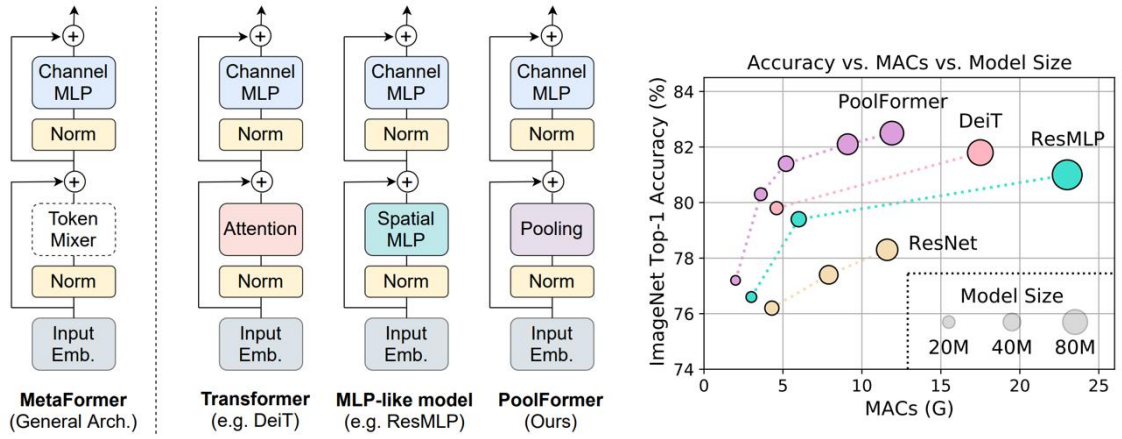


图注：ResMLP 的整体架构

来源：https://arxiv.org/pdf/2105.03404.pdf

新加坡 Sea AI Lab 等的研究者提出视觉任务通用架构 MetaFormer

11 月，新加坡 Sea AI Lab 和新加坡国立大学研究者提出了一种面向视觉任务的 MetaFormer 架构。研究者猜想，Transformer 通用架构，而非具体的 token mixer 模块，对于提升模型性能更为重要。通过将 Attention 模块替换为空间池化（Spatial Pooling）模块，构建基于名为“PoolFormer”架构的模型，能够在视觉任务上取得较高的性能表现。基于实验结果，研究者提出了 MetaFormer 的概念，一种通用的视觉架构。



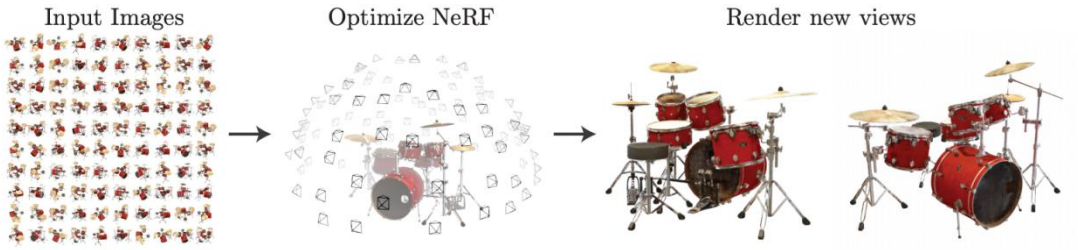
图注：a）MetaFormer 和 PoolFormer 与其他模型的架构对比；

b）不同架构模型在视觉任务上的性能对比

来源：https://arxiv.org/pdf/2111.11418.pdf

神经辐射场（NeRF）启发图像生成、三维重建等研究

用神经辐射场（NeRF）去替代点云、体素等传统三维数据形式来表示三维场景，是近年来三维视觉领域的热点研究话题。NeRF 创造性地将物体的三维信息通过隐式表示（Implicit Representation）编码在神经网络的参数中，通过输入视角信息，便可解码出任意视图下渲染的结果。NeRF 获 ECCV 2020 Best Paper 荣誉提名奖，同时也启发了一系列工作。

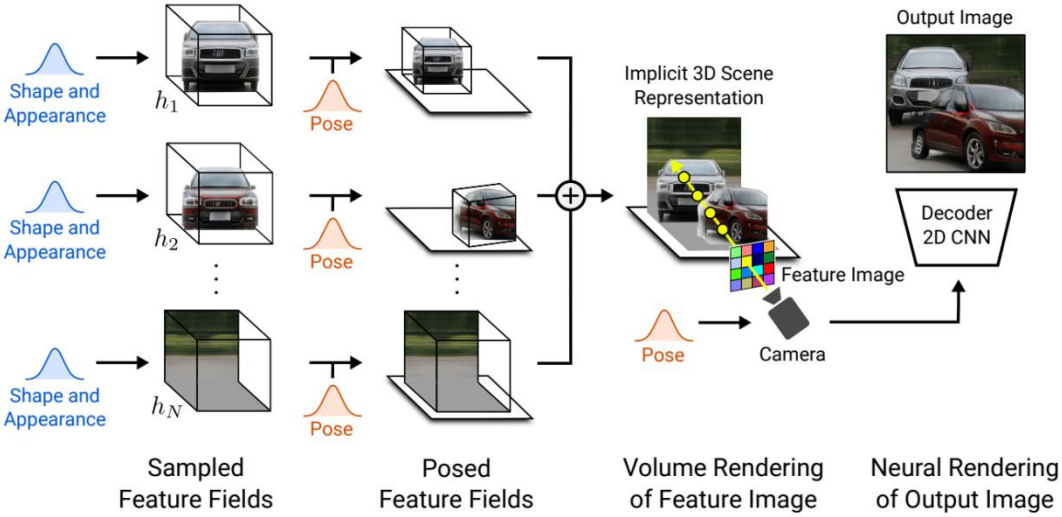


图注：NeRF 在重建三维图像新视角的过程

来源：https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123460392.pdf

德国马普所等研究者提出图像生成方法 GIRAFFE

4 月，德国马克思普朗克研究所和蒂宾根大学的研究者基于 NeRF 的思想，提出一种新型的图像生成方法，可控制生成图像中的内容，如对象的位置、方向，以及图像的背景等。该方法通过将输入图像进行编码，将图像转换为三维场景，通过解耦场景和对象，实现对对象进行编辑，最后再将图像生成出来即可。GIRAFFE 论文获得了 CVPR 2021 的最佳论文奖。



图注：GIRAFFE 的总体结构

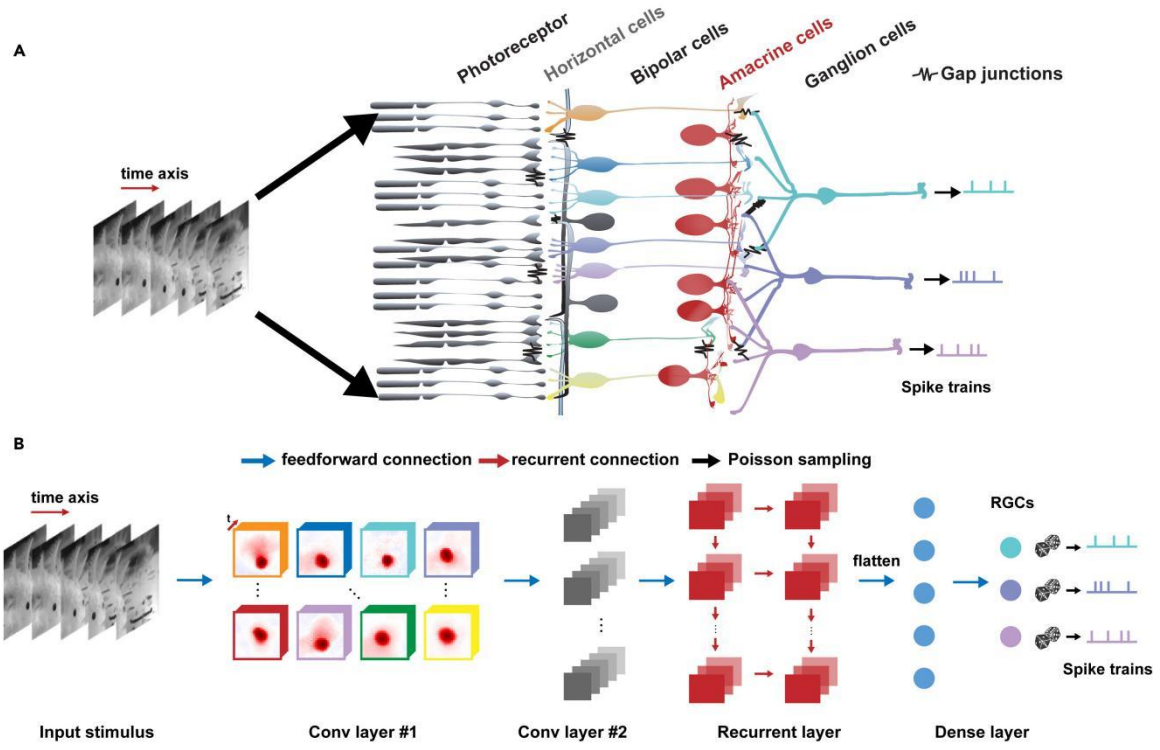
来源：https://arxiv.org/pdf/2011.12100.pdf

脉冲视觉开辟机器视觉新路线

北京大学团队提出模拟灵长类视网膜中央凹编码机理的脉冲视觉模型

深度学习支撑机器视觉在过去十年取得了巨大进步，但与生物视觉相比还存在巨大差距，例如对抗攻击脆弱、计算复杂度随分辨率线性增长等。近来，北京大学团队提出了模拟灵长类视网膜中央凹编码机理的脉冲视觉模型，推翻了沿用近两个世纪的相机和视频概念，专利获得中美日韩欧授权，研制了比人类视觉和影视视频快千倍的脉冲视觉芯片和相机，用普通器件实现了高铁会车、瞬态电弧、风洞激波等高速物理过程的连续成像，并结合脉冲神经网络，在笔记本算力条件下实现了超高速目标的实时检测跟踪和识别，在硬件和算力相当的情况下将机器视觉性能提升了三个数量级。团队还深入研究了生物视网膜编码复杂动态场景的神经

网络结构和信号编码机理，提出并实现了一种基于卷积循环神经网络（CRNN）的视网膜编码模型，能够高精度地预测大规模视网膜神经节细胞对动态自然场景的响应，可学习出视网膜神经节细胞感受野的形状及位置，模型结构更接近生物视网膜，可以使用更少的参数学习出精度更高的编码模型。还提出了评估刺激时空复杂度和感受野时空规律性的定量指标，实验结果揭示了网络的循环连接结构是影响视网膜编码的关键因素，这一模型不仅具有生物学价值，而且对设计新一代脉冲视觉模型、芯片乃至研制视网膜假体都具有重要意义，论文已在《细胞·模式》（Cell·Patterns）发表。



图注：视网膜结构与对应的卷积循环编码网络

来源：https://www.cell.com/patterns/fulltext/S2666-3899(21)00205-1

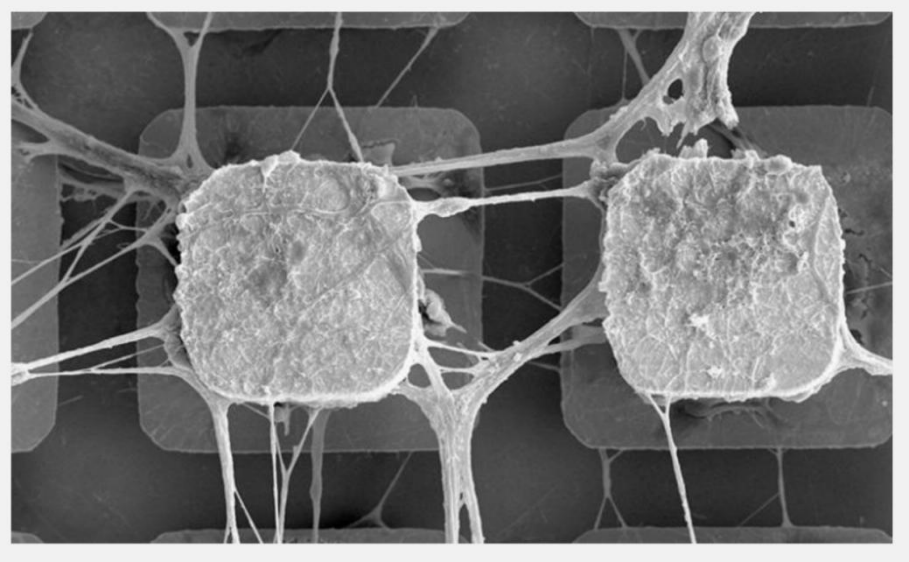
智能体系架构与芯片

生物神经元与芯片结合成为类脑芯片的研究热点

类脑智能是指受大脑神经和认知机制启发，采用软硬协同方式实现的机器智能。类脑芯片对于实现类脑智能至关重要。近年来，在硬件层面模拟大脑神经元的研究成为热点，有望突破电子元器件受到物理条件的固有局限。

英国阿斯顿大学研究将人脑干细胞“编织”在芯片上

1 月，英国阿斯顿大学研究者发起名为“Neu-ChiP”计划，期望能够将人脑干细胞应用于计算等方面，驱动人工智能的发展。该计划由欧盟委员会未来和新兴技术（FET）项目资助，合作者来自英国、法国、西班牙、瑞士、以色列等国家。目前研究者正在探索在微芯片上培养人脑干细胞，将类似于人脑皮层的干细胞网络分层放在微芯片上，向细胞发射不断变化模式的光束，来刺激细胞生长，并使用 3D 计算建模的方法来观测细胞的变化，这些技术用于模拟人脑的可塑性——这是其能够快速适应新信息的关键。这项研究将让这些干细胞能够从数据中解决问题，以此推动机器学习范式的转变（Paradigm Shift）。

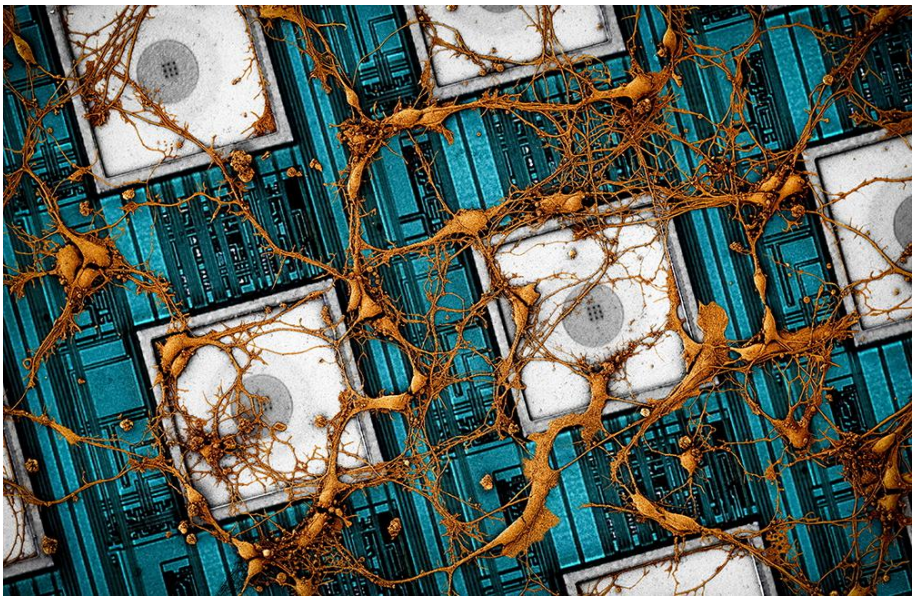


图注：大鼠神经元培养物的 SEM 图像，神经元位于结构电极的顶部和底部

来源：https://www.eurekalert.org/news-releases/631234

哈佛大学等研究者提出基于纳米电极阵列复制大脑神经元连接图方法

9 月，哈佛大学和三星高等技术研究院等机构的研究者发表于《自然 · 电子》（Nature · Electronics）杂志上，论文提出了一种基于纳米电极阵列复制大脑的神经元连接图的做法，并将复制的连接图粘贴在一个高密度固态存储器的高密度三维网络上。通过这一做法，作者期望能够构建一种能够近似大脑计算模式的记忆芯片，具有低功耗、快速学习、能适应环境，甚至具有自主性和认知能力。



图注：鼠的神经元在 CNEA（CMOS 纳米电子阵列）上生长的图像

来源：

https://news.samsung.com/global/samsung-electronics-puts-forward-a-vision-to-copy-and-paste-the-brain-on-neuromorphic-chips

高性能、低能耗 AI 芯片不断涌现

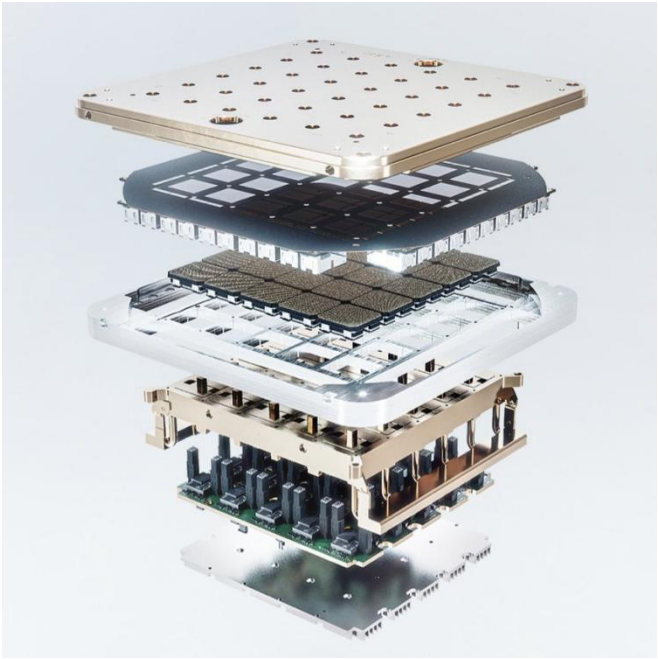
随着人工智能模型参数规模快速增长，其所需算力迅速增加，对于硬件提出了更高的性能要求。同时，大算力带来的能耗问题也受到人们的关注，研究者们开始探索研发算力更高，能耗较低的“节能芯片”。

IBM 公布节能型 AI 芯片

2 月，在 2021 年国际固态电路虚拟会议（ISSCC）上，IBM 团队公布了全球首个节能型 AI 芯片，采用 7nm 制程，有 4 个核心，支持包括 FP16、FP8 等训练精度，以及 INT4、INT2 等多种推理精度，可达到 80%以上的训练利用率和 60%以上的推理利用率。IBM 表示，这一芯片技术可用于多种商业应用和场景，如大规模模型训练、语音转文本服务、NLP 服务、金融交易欺诈检测等。

特斯拉发布自研 Dojo D1 高性能芯片

8 月，特斯拉在 AI 日活动上发布自研 Dojo D1 高性能芯片。该芯片采用 7nm 制程，在 FP32 精度下具有 22.6TFLOPS 的算力，4TB/s 的带宽。



图注：由 D1 芯片组成的计算区块，最终用于组成 AI 超算
来源：https://www.tomshardware.com/news/tesla-d1-ai-chip

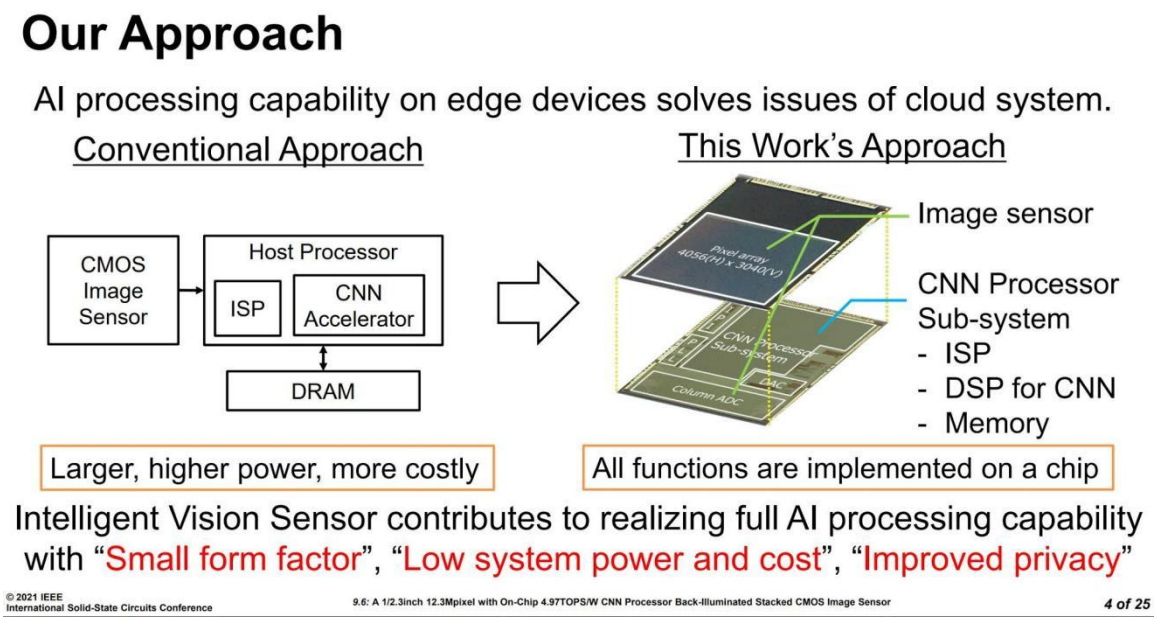
OPPO 发布 6nm 工艺 AI 芯片，能效 11TOPS/W

12 月，Oppo 发布自研 NPU 芯片，采用 6nm 制程，采用专用芯片架构，算力最高达 18TOPS，能耗比为 11.6TOPS/W。预计该款芯片将于 2022 年在 Oppo 手机产品中搭载。

存算一体 AI 芯片设计、应用步伐加快

索尼发布感存算一体化设计近光学传感器 AI-ISP 芯片

随着物联网、零售、智慧城市等产业的发展，在相机产品中搭载 AI 处理能力的需求快速增长。边缘端芯片的 AI 处理能力能够解决只在云计算系统中出现的问题，如延迟、云端通讯、处理开销，以及隐私问题等。当前市场对边缘端智能相机要求包括小型、低功耗、低成本、易部署等，但目前传统的 CMOS 图像传感器只能输出原始图像数据。因此，在设计具有 AI 能力的智能相机时，将图像信号处理器（ISP）、神经网络处理能力、DRAM 等结合在一起十分重要。在 2021 IEEE 国际固态电路会议（ISSCC）上，索尼发布了其背照式堆叠型 CMOS 图像传感器芯片，芯片能耗比达到 4.97TOPS/W。通过将图像传感器、CNN 处理器，以及 ISP、DSP、内存等子系统进行堆叠设计，在单芯片上实现完整的 AI 图像处理能力。



图注：索尼芯片的整体架构

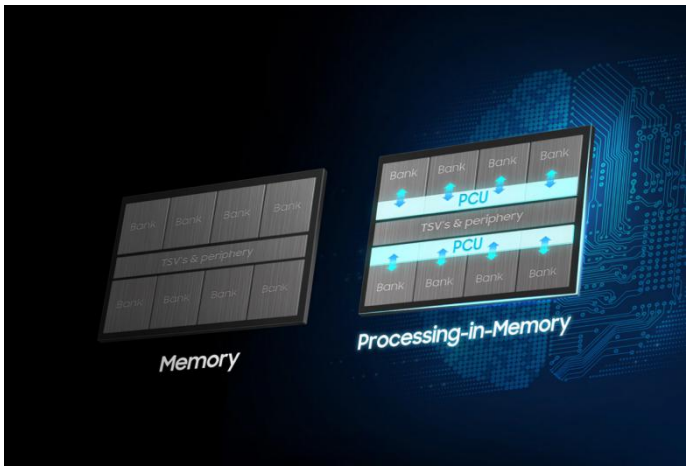
来源：

http://www.f4news.com/2021/03/27/isscc-2021-on-line-sony-sensor-with-integrated-ai-processor/

三星发布 HBM-PIM 芯片开启端侧存算一体化产品落地

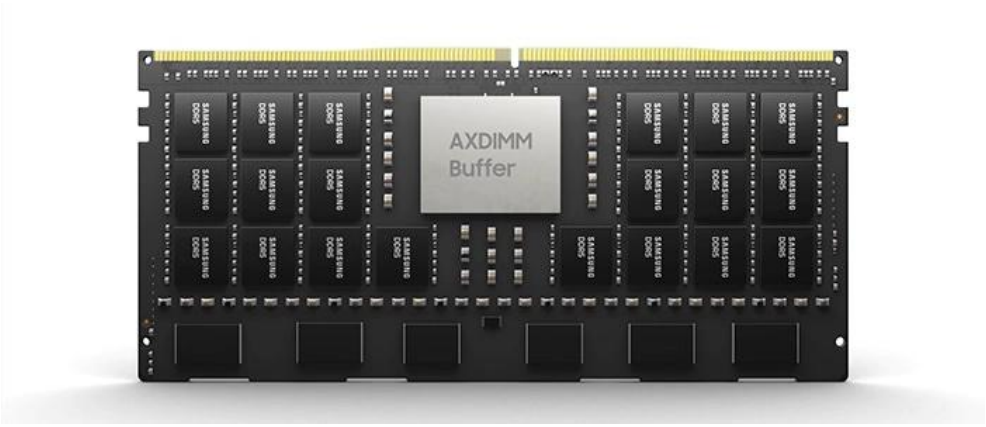
存算一体芯片和传统的冯诺依曼架构芯片不同，其在片上集成内存和计算，不需要先将数据加载到内存中再传递给计算单元，因此可以最大程度减少延迟，提升处理速度和能源效率。当前，包括美国 Mythic、国内知存科技、苹芯科技等都在积极研发相关芯片产品。2 月，三星发布了名为“HBM-PIM”的存算一体芯片，其基于三星的 HBM 高带宽内存芯片，集成赛灵思的 AI 加速系统，系统总体的性能提升 2.5 倍，并降低了约 70%的能耗。三星称其为工业界首款在高带宽内存芯片上集成存算一体能力的芯片。部署该芯片不需要对现有的内存生态环境进行改动，并能够和 LPDDR 和 GDDR 等高带宽内存芯片集成。三星认为，该款芯片

可以在多种场景中部署，包括移动端、数据中心和超级计算机等。此外，三星还发布了 AXDIMM 芯片模组，结合 DRAM 模组，并通过降低 CPU 和 DRAM 之间的数据移动，从而提升 AI 加速系统的能耗效率。



图注：三星 HBM-PIM 和内存芯片的对比

来源：三星官网



图注：三星 AXDIMM 模组

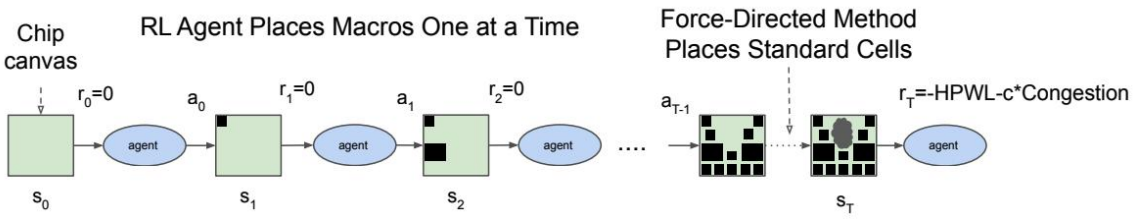
来源：三星官网

由 AI 辅助设计成为芯片发展新趋势

芯片设计是一项系统工程，其中包括布局、布线，以及多种参数的选择。以往芯片主要由人类专家根据经验知识进行设计，时间耗时较长。

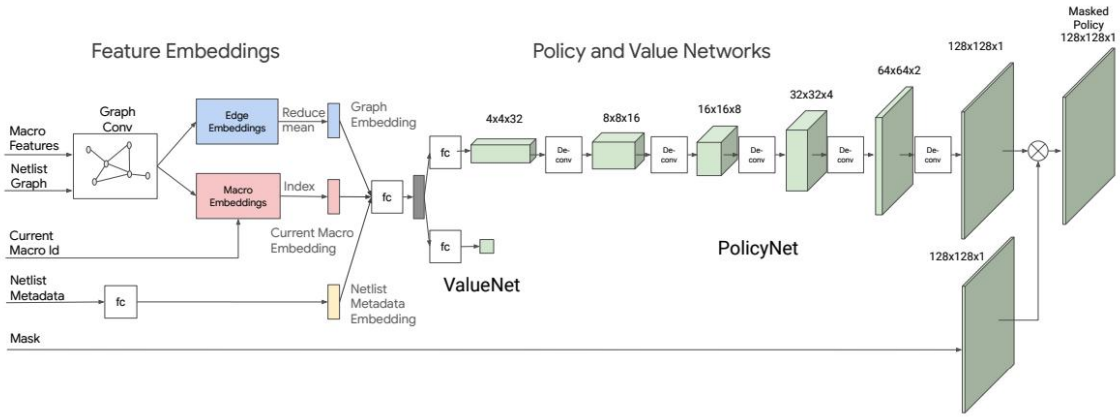
谷歌大脑团队提出基于 AI 的芯片布局设计方法

2020 年 4 月，谷歌大脑研究者 Jeff Dean 等提出了一种基于 AI 的芯片设计方法。该方法可以在 6 小时内完成设计工作，而人类需要数周时间。今年 6 月，谷歌联合斯坦福大学的研究者对这一方法进行了改进，并应用于下一代 AI 加速芯片的设计中。该方法可自动生成的芯片布局设计，并在功耗、性能和芯片面积等关键指标上媲美人类工程师。目前该研究已发表于《自然》杂志上。



图注：该论文提出的方法用于设计芯片的流程

来源：https://arxiv.org/pdf/2004.10746.pdf



图注：该算法的架构

来源：https://arxiv.org/pdf/2004.10746.pdf

Name	Method	Timing		Area	Power	Wirelength	Congestion	
		WNS (ps)	TNS (ns)	Total (μm^2)	Total (W)	(m)	H (%)	V (%)
Block 1	RePlAce	374	233.7	1693139	3.70	52.14	1.82	0.06
	Manual	136	47.6	1680790	3.74	51.12	0.13	0.03
	Ours	84	23.3	1681767	3.59	51.29	0.34	0.03
Block 2	RePlAce	97	6.6	785655	3.52	61.07	1.58	0.06
	Manual	75	98.1	830470	3.56	62.92	0.23	0.04
	Ours	59	170	694757	3.13	59.11	0.45	0.03
Block 3	RePlAce	193	3.9	867390	1.36	18.84	0.19	0.05
	Manual	18	0.2	869779	1.42	20.74	0.22	0.07
	Ours	11	2.2	868101	1.38	20.80	0.04	0.04
Block 4	RePlAce	58	11.2	944211	2.21	27.37	0.03	0.03
	Manual	58	17.9	947766	2.17	29.16	0.00	0.01
	Ours	52	0.7	942867	2.21	28.50	0.03	0.02
Block 5	RePlAce	156	254.6	1477283	3.24	31.83	0.04	0.03
	Manual	107	97.2	1480881	3.23	37.99	0.00	0.01
	Ours	68	141.0	1472302	3.28	36.59	0.01	0.03

图注：与基线方法的结果对比（数值越小越好；H: Horizontal, V: Vertical）

来源：https://arxiv.org/pdf/2004.10746.pdf

智能信息检索与挖掘

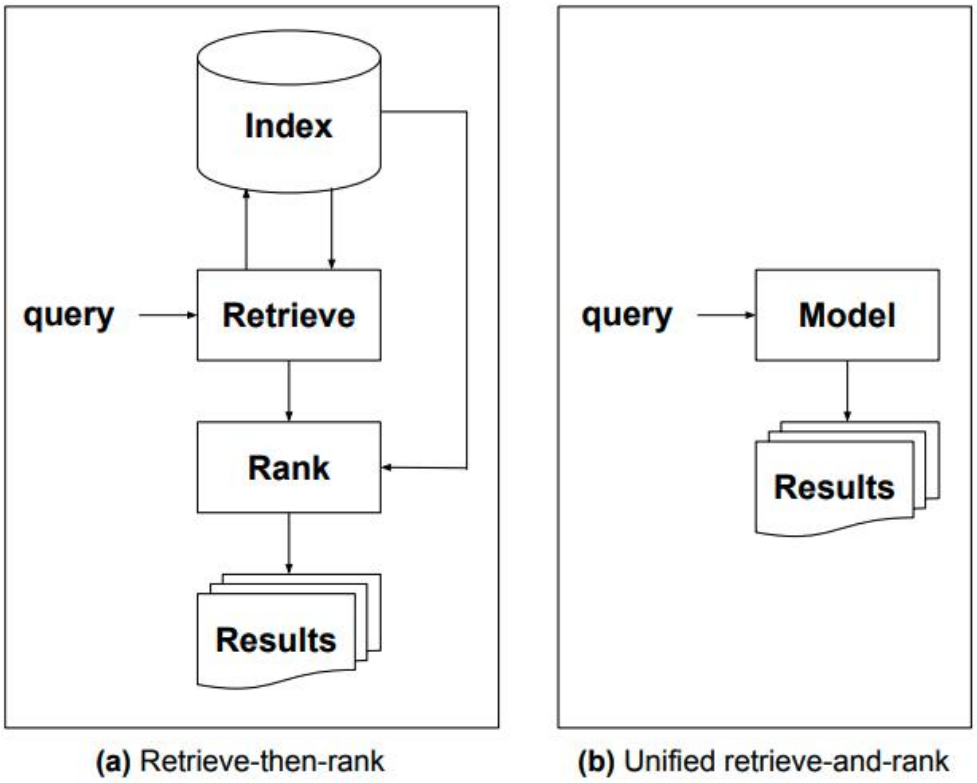
Web 模型成为新型信息搜索范式的核心支撑

Web（互联网）是人类知识和经验的汇总，有着各种模态的数据，不同的数据模态之间有着丰富的协作形式。如果有了大模型的方法，可以利用 Web 数据构建汇聚人类知识和经验的模型，名为“Web 大模型”，并在下游完成多种问答、生成任务，成为基础模型的一种。

在搜索领域，将出现以大模型为中心的搜索/信息检索新范式。以模型为核心的搜索范式，其思想核心在于通过大模型全面地学习知识，将各种零散的、多模态的数据知识化、系统化，形成一个大脑。当用户提问的时候，直接通过模型进行交互，省去了传统搜索中需要的编号（Index）、抽取（Retrieve）和排序（Rank）等步骤。

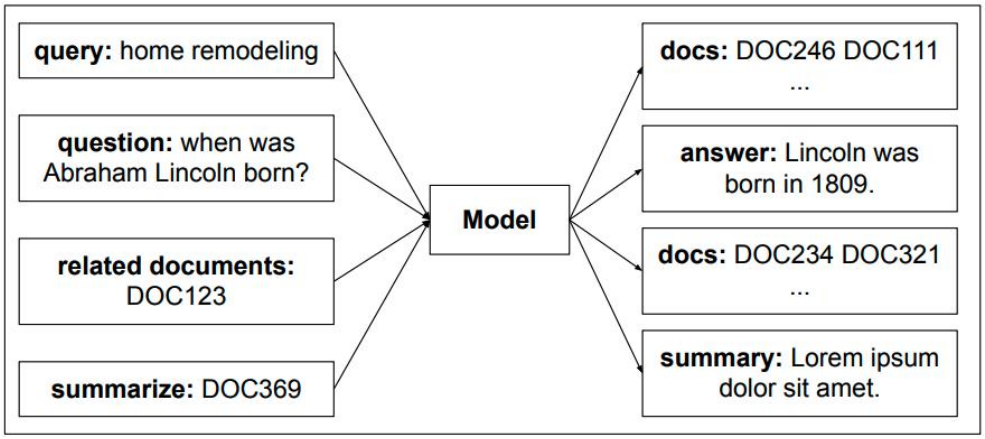
谷歌研究者发表以大模型为核心的搜索模式

7 月，谷歌研究者发表以大模型为核心的搜索模式。给定模型搜索的请求（Query），模型可以自动返回结果。除了检索以外，模型可以完成各种知识获取任务，甚至包括生成和翻译任务。



图注：a）传统的“抽取后排序”（Retrieve-then-rank）搜索方法；b）以模型为核心的搜索方法

来源：https://arxiv.org/pdf/2105.02274.pdf

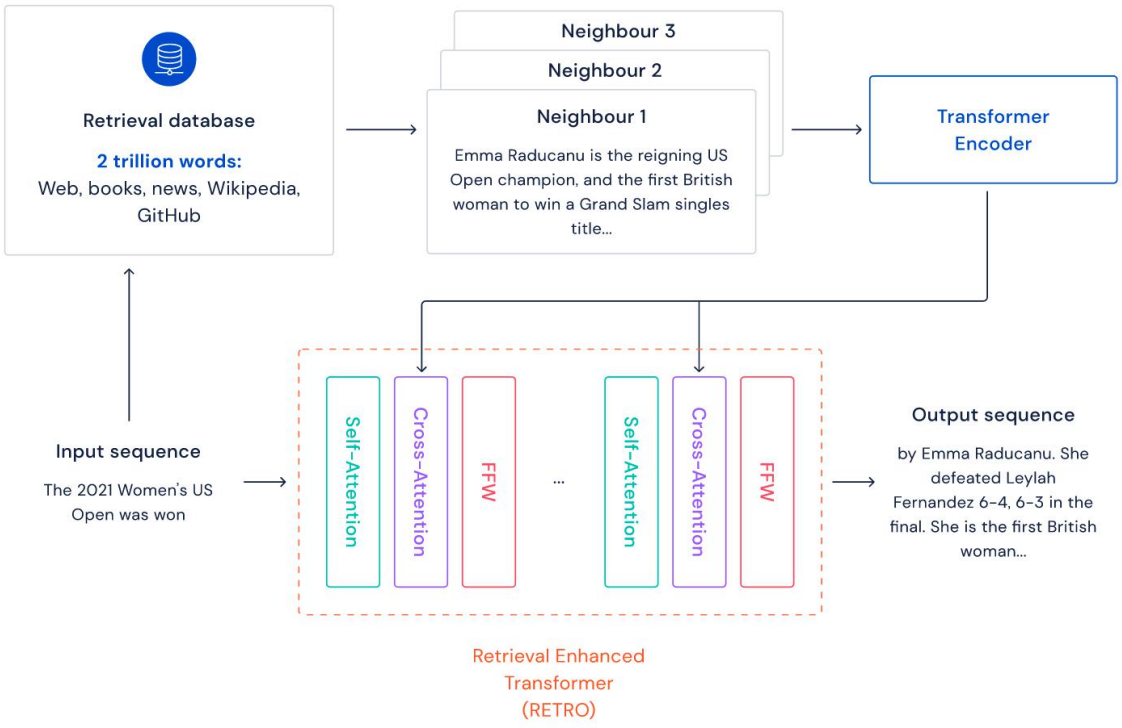


图注：以模型为核心的搜索方法通过学习各种形式和模态的数据，能够完成包括文档检索、问答、摘要等多种下游任务

来源：https://arxiv.org/pdf/2105.02274.pdf

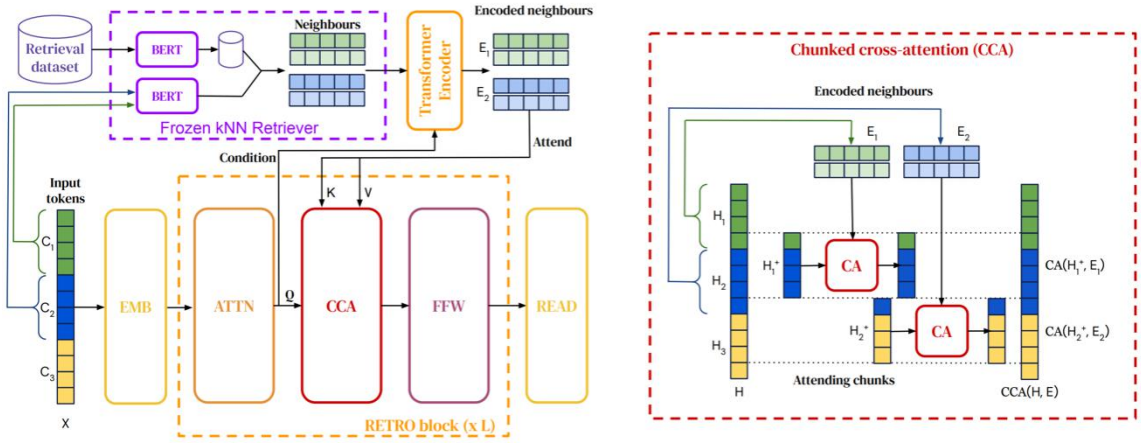
DeepMind 提出 RETRO 信息检索预训练模型

12 月，DeepMind 基于预训练模型 Gopher 等研究成果，提出名为 RETRO（Retrieval-Enhanced Transformer）的信息检索预训练模型，该模型采用互联网规模的信息抽取机制进行了预训练。受大脑在学习时依赖专用记忆机制的启发，RETRO 能够有效地查询文本段落，并改进其预测结果。通过将生成的文本与 RETRO 的输入的原始文本进行比较，可以解释模型做出某些预测的原因及其来源。与常规的 Transformer 相比，该模型取得了与其相当的性能，但参数少了一个数量级，并且在多种语言建模基准上获得最先进的性能。



图注：RETRO 进行信息检索、抽取和生成的流程机制

来源：https://deepmind.com/blog/article/language-modelling-at-scale



图注：RETRO 模型的架构。左：整体架构；右：CCA（Chunked Cross-attention）模块的内部架构

来源：

https://storage.googleapis.com/deepmind-media/research/language-research/Improving%20language%20models%20by%20retrieving.pdf

预训练语言模型助力信息检索性能提升

近年来，预训练模型在信息检索中取得了一定的成果，但性能提升并不高，构建面向信息检索的预训练任务是一个重要问题。当前，一些研究者正在探索是否可以构建自监督预训练任务来模拟信息检索的相关性匹配。

智源学者提出基于 BERT 的信息检索方法 B-PROP

6 月，中科院计算所研究员，智源学者郭嘉丰团队提出了基于 BERT 的信息检索方法 B-PROP。该研究提出了针对信息检索的预训练任务，名为“代表词预测”（Representation Words Prediction），并应用 BERT 预训练模型代替 Unigram 语言模型进行代表性词预测任务的

构建。实验结果表明，B-PROP 相比其他方法在小数据集上的性能更优。在 MS MARCO 文档排序任务上，B-PROP 成为首个 MRR@100 指标上超过 0.4 的成果。

Model Type	Model Name	Robust04		ClueWeb09-B		Gov2		MQ2007		MQ2008	
		nDCG@20	P@20	nDCG@20	P@20	nDCG@20	P@20	nDCG@10	P@10	nDCG@10	P@10
Traditional Retrieval Models	QL	0.413	0.367	0.225	0.326	0.409	0.510	0.423	0.371	0.223	0.241
	BM25	0.412	0.363	0.230	0.334	0.421	0.523	0.414	0.366	0.220	0.245
Neural IR Models	DRMM	0.425	0.371	0.246	0.349	0.457	0.545	0.441	0.382	0.221	0.248
	Conv-KNRM	0.414	0.360	0.238	0.336	0.462	0.552	0.431	0.377	0.215	0.239
Pre-trained Models	BERT	0.459	0.389	0.295	0.367	0.495	0.586	0.506	0.419	0.247	0.256
	Transformer$_{ICT}$	0.460	0.388	0.298	0.369	0.499	0.587	0.508	0.420	0.245	0.256
	PROP$_{Wiki}$	0.502	0.421	0.316	0.384	0.519	0.593	0.523	0.432	0.262	0.267
	PROP$_{MARCO}$	0.484	0.408	0.329	0.391	0.525	0.594	0.522	0.430	0.266	0.269
Our Approach	B-PROP$_{Wiki}$	**0.519***	**0.430***	0.331	0.393	0.534*	0.599*	**0.529***	0.436*	0.271*	0.273
	B-PROP$_{MARCO}$	0.510*	0.429*	**0.353***	**0.407***	**0.552***	**0.606***	**0.529***	**0.439***	**0.273***	**0.275***

图注：在五个小数据集上将 B-PROP 和其他方法进行比较的结果

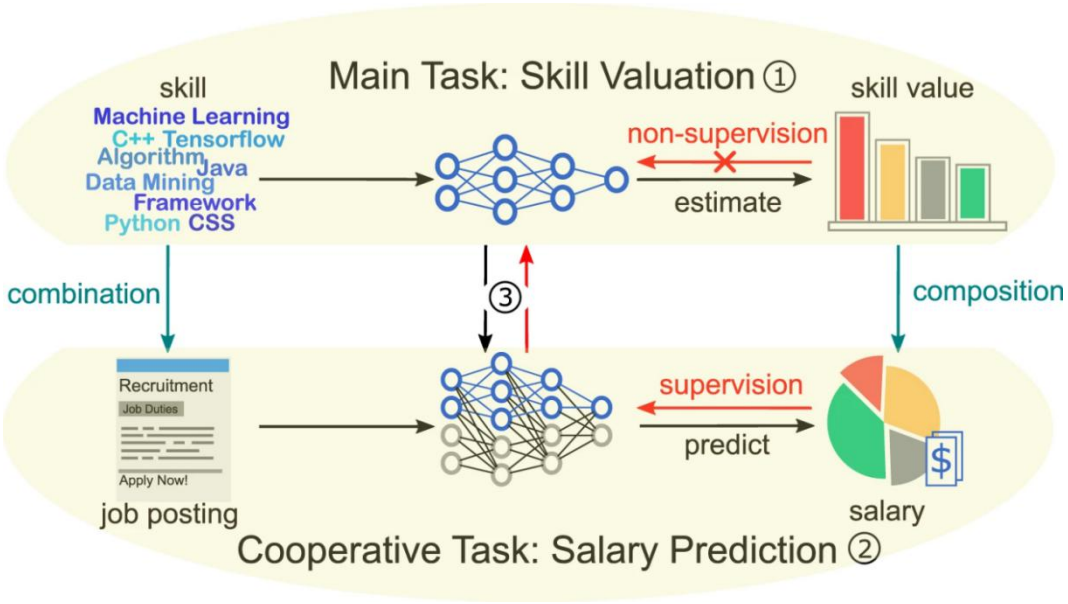
来源：https://arxiv.org/pdf/2104.09791.pdf

数据驱动的新方法推动定量分析在社会科学领域实现应用

在数据规模快速增长的背景下，许多社会科学领域，如管理学、人口学、社会学、政治学等学科，都在寻求通过数据驱动的方法进行研究，以求通过定量方法精确地解答定性问题，实现研究成果的突破。

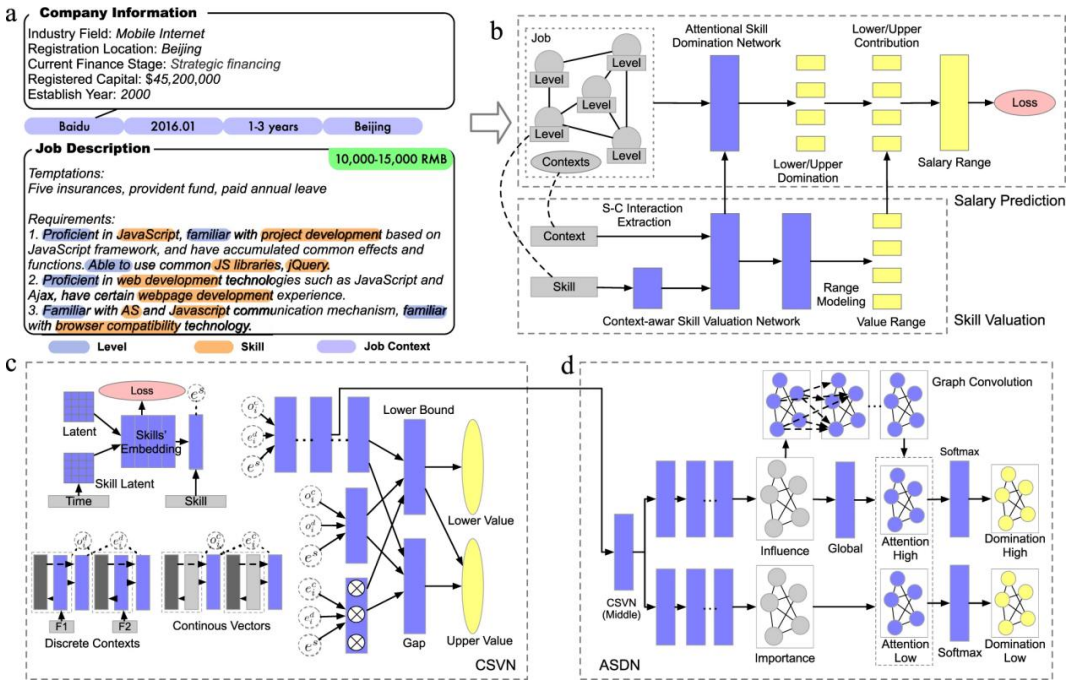
百度等团队提出劳动力市场技能价值评估方法及框架

3 月，由百度人才智库团队牵头，联合中科院计算所、北京航空航天大学、中科大等机构的研究者提出了一种利用合作协同神经网络（Cooperation Composition Neural Network）对劳动力市场的技能价值进行评估的方法，并进一步定量分析了技能和薪资之间的关系，形成了一套体系化的评估框架。



图注：SSCN 的基本原理

来源：https://www.nature.com/articles/s41467-021-22215-y



图注：SSCN 的基本架构和处理数据的方法

来源：https://www.nature.com/articles/s41467-021-22215-y

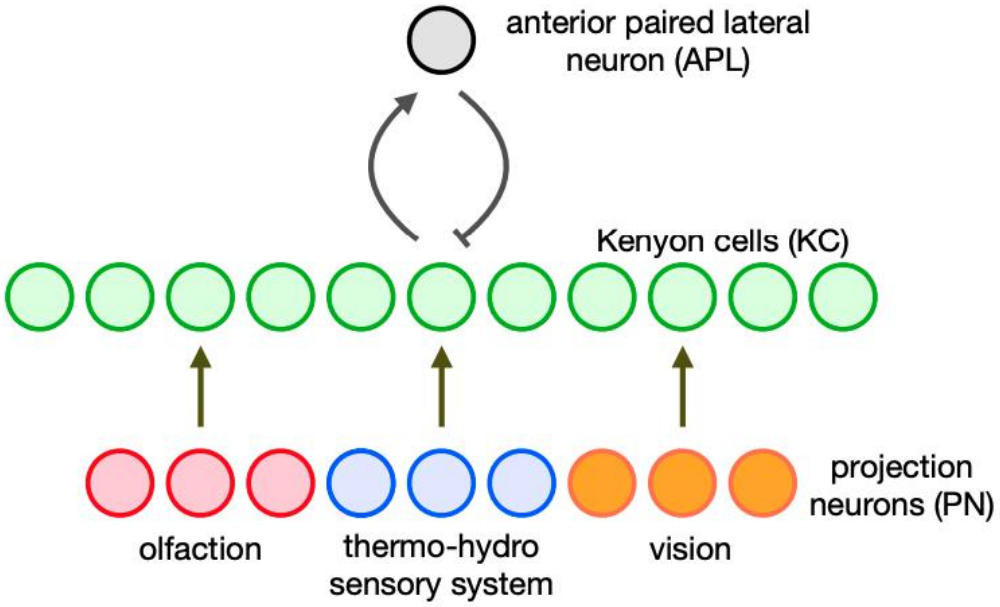
人工智能的认知神经基础

借鉴脑神经和认知科学研究成为启发类脑智能研究的重要来源

近年来，借鉴脑神经结构和认知科学成果，对设计类脑的神经网络架构，推动通用智能的发展具有重要意义。近年来，越来越多的研究者期待通过生物仿脑的方法，发掘更为高效、更低功耗的结构和算法。

MIT-IBM 联合实验室基于果蝇大脑构建神经网络学习 NLP 任务

3 月，MIT-IBM 联合实验室的研究者基于果蝇大脑中的成熟神经生物学网络模体（Motif），将结构进行数学形式化后构建神经网络。该网络可以学习语义表征，生成静态的、依赖于上下文的词嵌入。根据实验，该网络的性能不仅可以与现有 NLP 方法相媲美，内存占用率也更小，需要的训练时间更短。在上下文单词任务中，果蝇网络的表现比 GloVe 高出近 3%，比 Word2Vec 高出 6%以上。

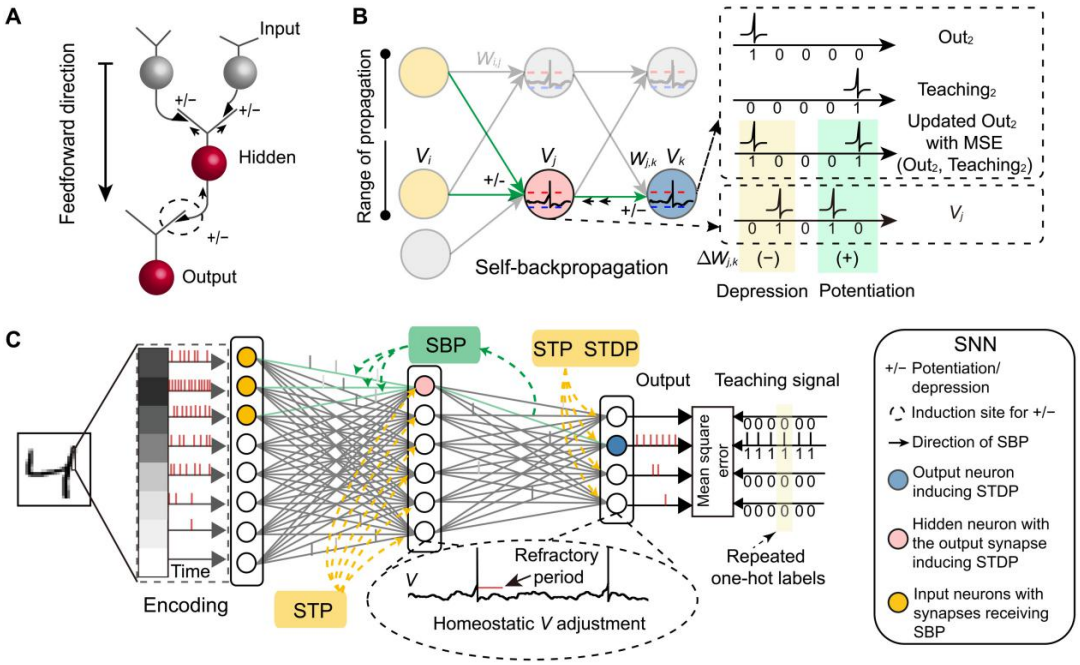


图注：受果蝇大脑结构启发设计的网络结构

来源：https://openreview.net/pdf?id=xfmSoxdxFCG

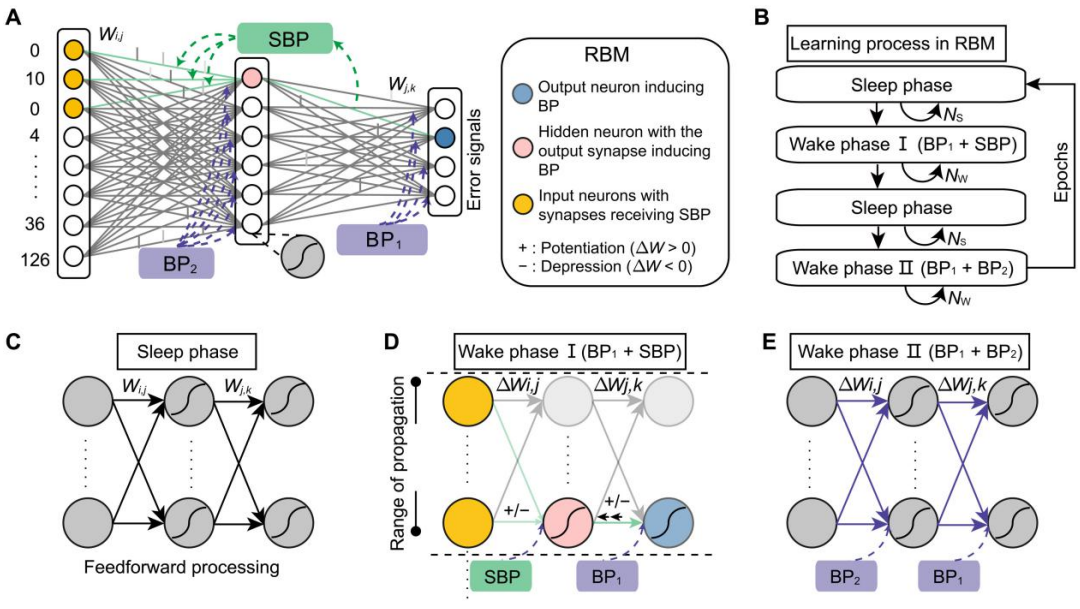
中科院研究者提出基于生物大脑介观尺度自组织反向传播机制启发神经网络设计思路

10 月，中科院脑智卓越中心徐波、蒲慕明团队在《科学》杂志子刊《Science Advances》上发表研究成果，团队发现，根据生物大脑网络中发现介观尺度自组织反向传播机制（Self-Backpropagation，SBP），能够启发设计更为高效的脉冲神经网络（SNN）和人工神经网络（ANN），实现高效的全局优化学习效果。实验显示，SBP 可以提升 SNN 和 ANN 在 MINIST、NETalk 和 DvsGesture 三个基准测试中的性能表现。



图注：SBP 在脉冲神经网络中的使用

来源：https://www.science.org/doi/10.1126/sciadv.abh0146



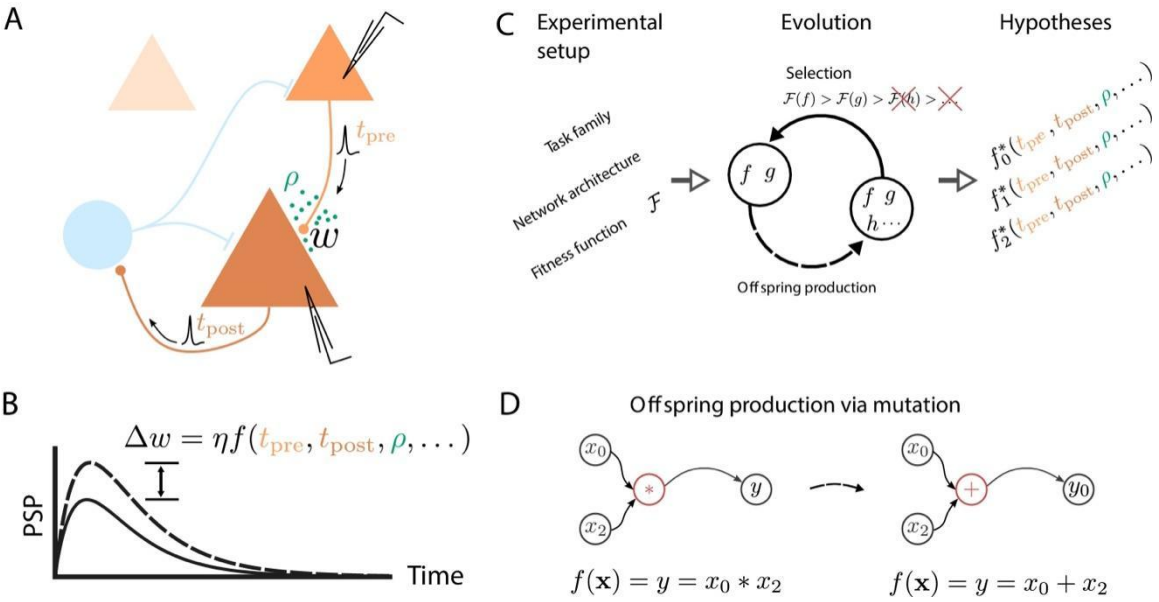
图注：SBP 在人工神经网络中的使用

来源：https://www.science.org/doi/10.1126/sciadv.abh0146

瑞士伯尔尼大学研究者提出基于进化过程发掘生物物理可塑性规则的方法

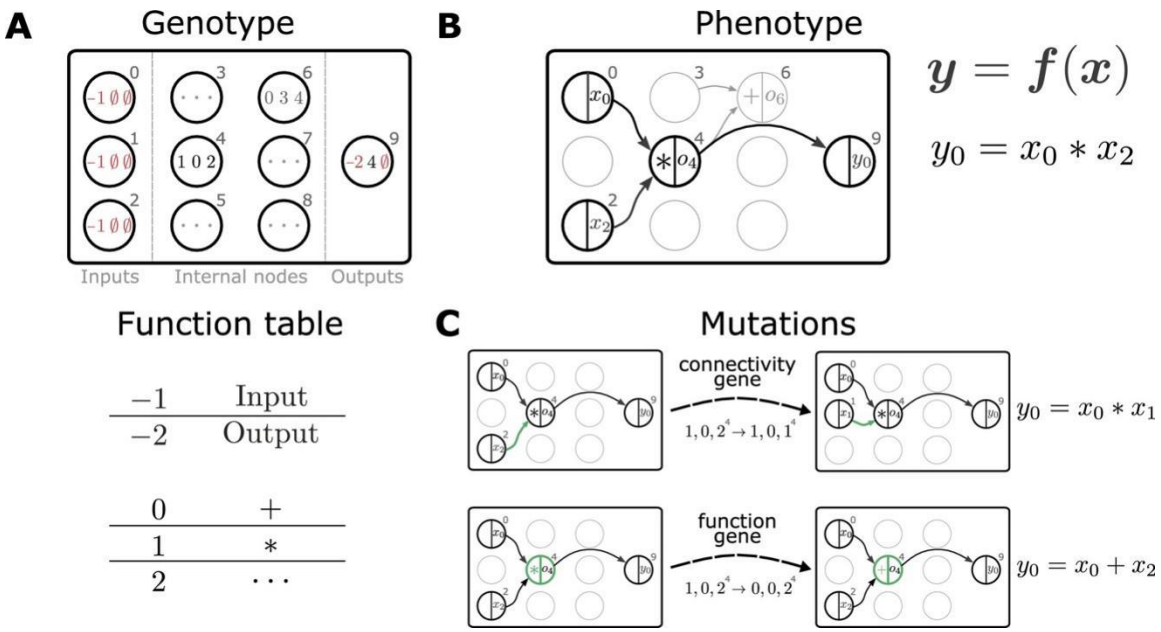
生物通过适应环境生存，而神经元突触之间耦合强度的改变对于生物的环境适应能力非常重要，这些变化被称为是可塑性规则（Plasticity Rules）。如何用数学方法描述这些变化，对于理解生物处理信息的机制，开发具有认知能力的人工智能具有重要意义。

10 月，瑞士伯尔尼大学研究团队提出一种自动化方法，通过模仿生物进化的过程（如自然选择）来寻找任务的解决方案，这种方法能够发现生物物理上的可塑性规则，在搜索脉冲神经网络的可塑性规则方面有很大潜力。



图注：脉冲神经网络中突触可塑性规则的人工进化过程

来源：https://elifesciences.org/articles/66273



图注：笛卡尔遗传编程（Cartesian Genetic Programming，GCP）中的表示（Representation）和突变（Mutation）的数学表示

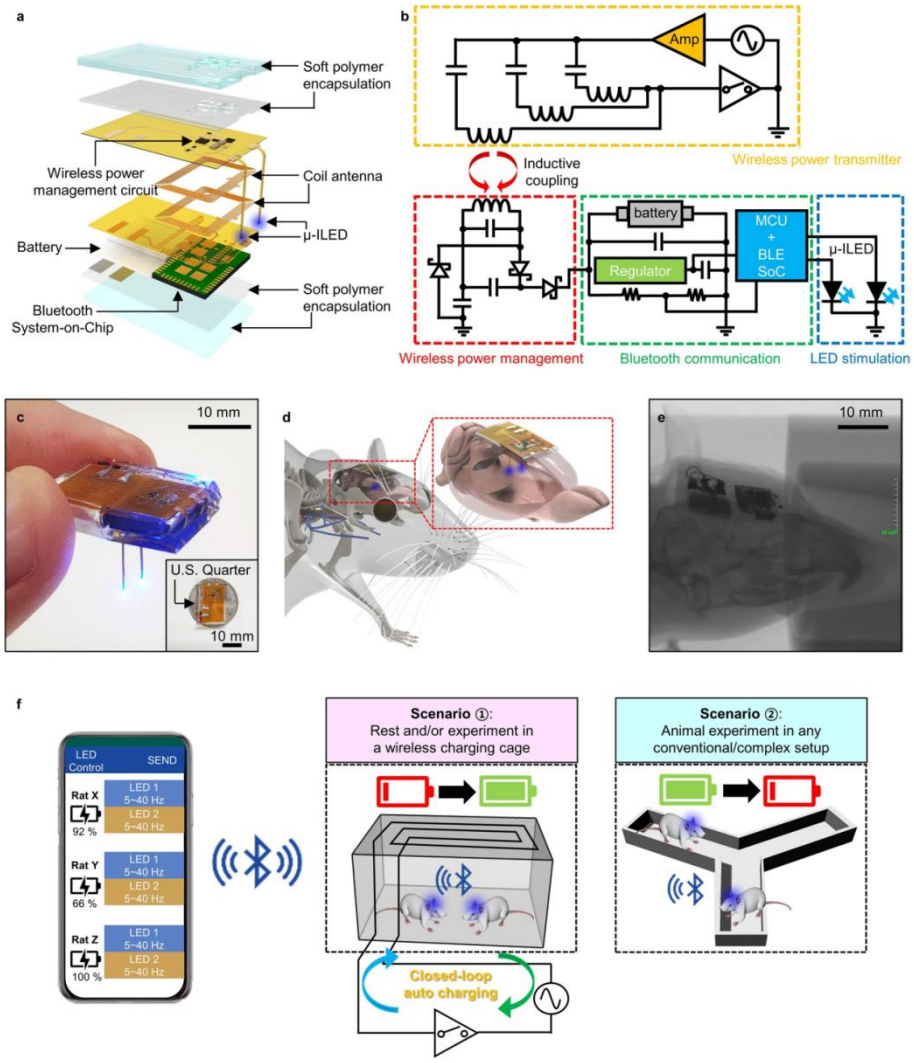
来源：https://elifesciences.org/articles/66273

无线高带宽、微创、结合 AI 算法等成为脑机接口的发展重点

脑机接口被认为是实现生物与机械智能融合的重要途径，在医疗康复等领域具有应用价值，但脑机接口现在仍存在一些问题没有得到解决。一是很多侵入式和半侵入式脑机接口都采用有线连接，设备笨重，而无线脑机接口系统的信号保真度较差；二是要获得大脑皮层中更多的信号，往往需要采用侵入式，即将电极阵列植入大脑皮层中，这对手术的要求很高，而且对大脑有一定的损伤；三是对于高度灵巧的运动脑电信号进行解码还比较困难，相关数据也较为匮乏，对于训练算法不够充足。

韩国科学技术院研究者发明可体外无线充电软脑植入物

1 月，韩国科学技术院（KAIST）的研究人员发明了一种可以直接在体外进行无线充电的软脑植入物。可以实现长期的神经回路操控，不需要定期进行破坏性手术来更换植入物的电池。这款设备由超软且符合生物标准的聚合物支撑，具有与生物组织长期兼容的特性。同时，它还配置了安装在超薄探针上的微米级 LED，可以使用光来无线操控深层大脑中的目标神经元。

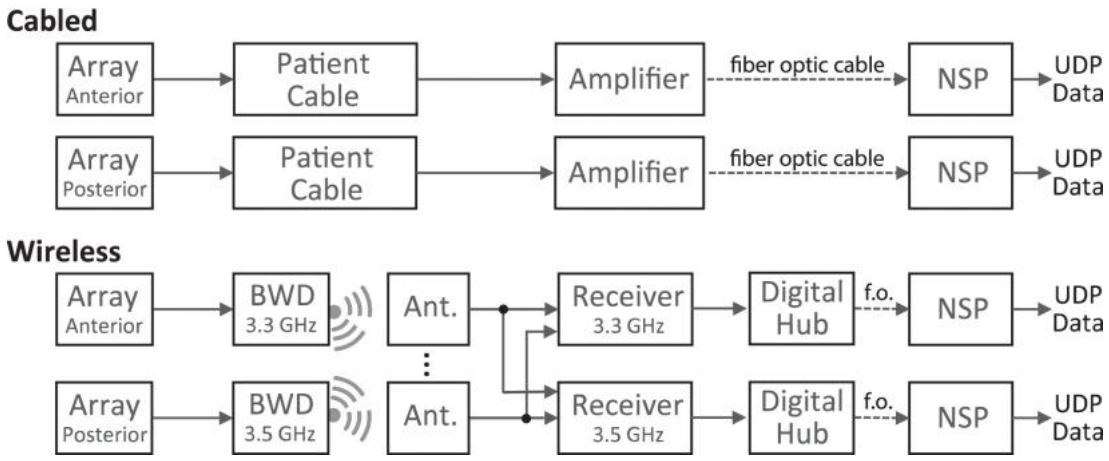


图注：采用软脑植入物的脑机接口结构和工作原理

来源：https://www.nature.com/articles/s41467-020-20803-y

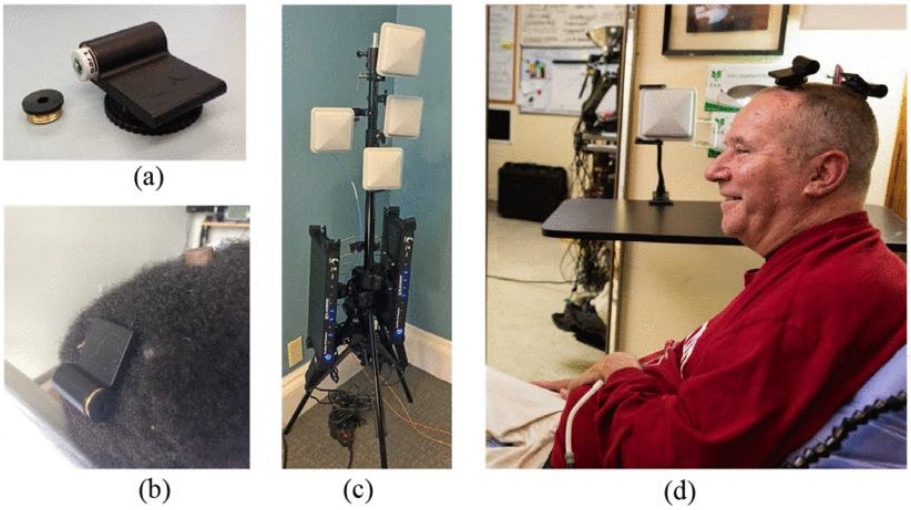
美国 BrainGate 研发无线脑机接口系统

2 月，美国 BrainGate 的研究人员首次在临床试验中，为两名瘫痪患者配置了无线脑机接口系统，首次实现了大脑信号与计算机之间的无线高带宽传输。该系统采用侵入式脑机接口，将两个由 96 个电极的组成电极阵列植入患者的大脑皮层，用于捕捉全频谱信号。信号发射器采用重量约 43g 的无线发射器，可以固定在使用者头部。该脑机接口无需用户被有线连接束缚在解码系统上，并且能让患者达到很高的点击精度和打字速度（约 13.4 个字/分钟）。



图注：无线和有线脑机接口在组成上的差别

来源：https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9390339



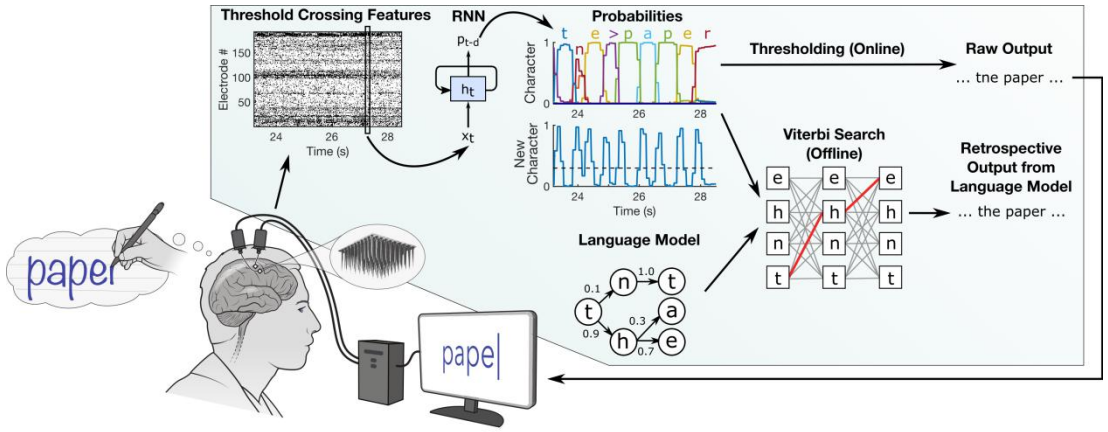
图注：该款无线脑机接口的组成部件和实际使用情形

来源：https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9390339

斯坦福大学研究者利用 AI 和脑机接口实现“意念写字”

当前，脑机接口已经可以实现一定程度的大脑和计算机之间的沟通，如让瘫痪患者能够操控光标打字等，但是对于一系列高度灵巧的行为来说，解码这些行为背后的脑电信号需要更高精度的脑电解码方法。

5 月，斯坦福大学的研究者提出了一种新型的脑机接口系统，采用循环神经网络，能够将来自运动皮层的手写字脑电信号解析为文本。在线情况下，该脑机接口实现了 90 字/分钟的速度，准确率为 94.1%，在有自动纠错软件的支持下，离线的准确率高达 99%。

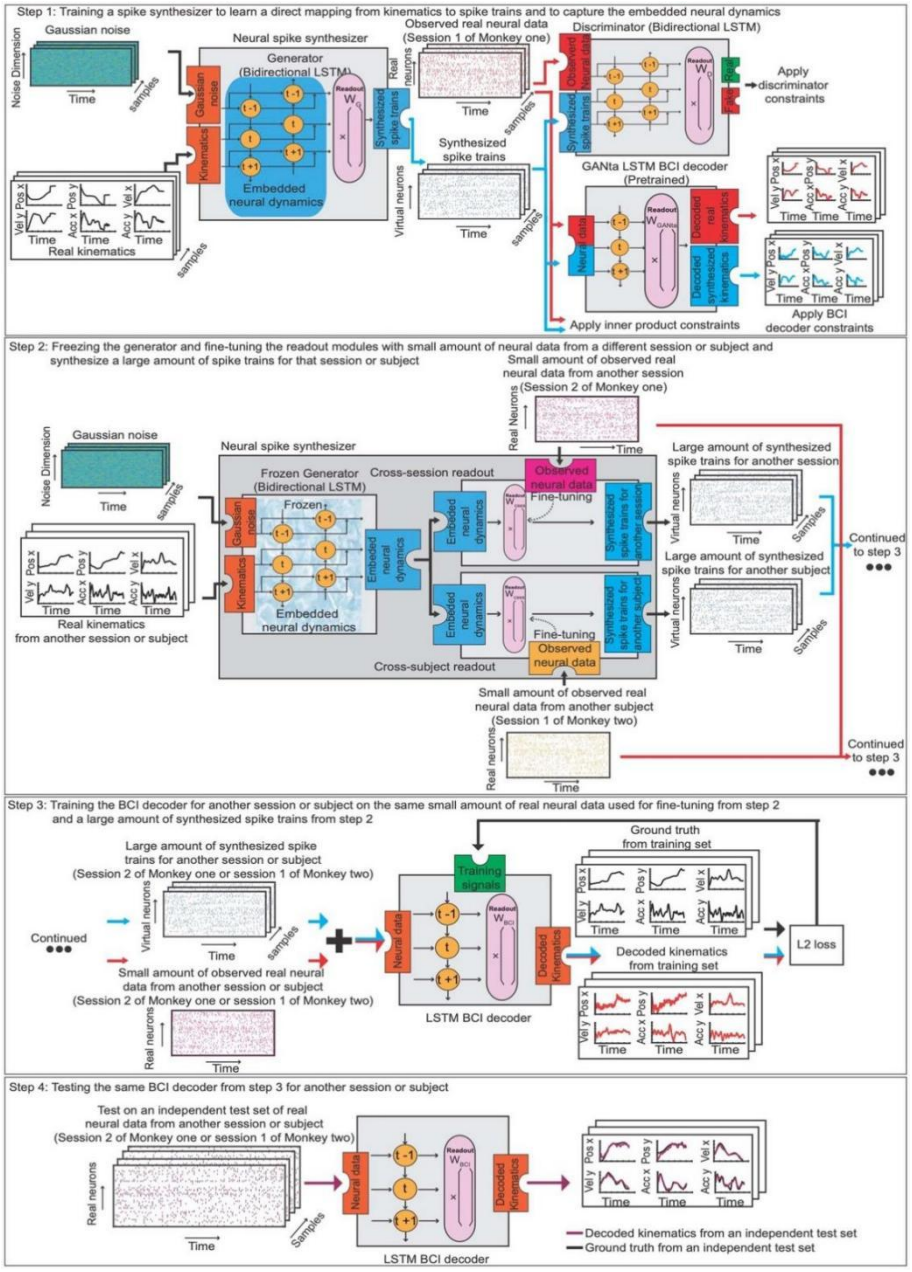


图注：该款脑机接口的整体结构

来源：https://github.com/fwillett/handwritingBCI

南加州大学研究者提出合成运动控制信号数据的方法

脑机接口研究中，构建神经信号和肢体行动关系的算法需要大量的神经信号数据，但这些数据往往难以获得。11 月，南加州大学团队提出了一种生成式脉冲训练合成器。研究者采用实验猴子手臂中植入的电极阵列，记录了其在类似贪吃蛇的游戏任务中的运动控制信号（Spike Train），然后将这些数据输入合成器中，并结合该猴子在不同时间或从另一只实验猴子上采集的数据进行微调，最终生成大量用于训练算法的数据。此外研究者表示，由于该方法是完全基于数据驱动的，因此可以用于除了运动控制以外的其他脑机接口应用。

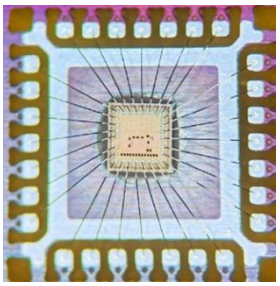


图注：该算法的架构，以及生成大量数据的步骤

来源：https://www.nature.com/articles/s41551-021-00811-z

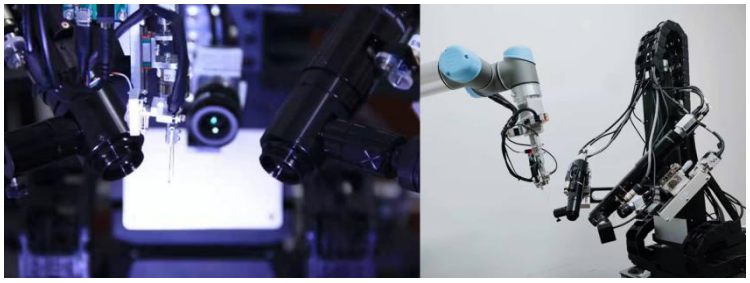
中科院自动化所在植入式脑机接口领域取得重要进展

高通量植入式脑机接口近年来引起了广泛的关注，其核心技术包括高生物相容性柔性电极及其植入，高性能低功耗信息处理/传输芯片与系统，以及高精度神经编解码算法等。中科院自动化所 2021 年研制了柔性电极自动植入机器人系统，机器人能在显微图像引导下将柔性电极精准植入动物大脑皮层；研制了 64 通道植入式脑机接口芯片，具备信号采集和刺激的功能；并研制了用于植入式系统信号传输的人体信道通信芯片，传输速率最高可达 60Mbps。这些技术突破为植入式脑机接口的未来发展打下了重要基础。



图注：人体信道通信芯片

来源：学者供图



图注：柔性电极植入机器人

来源：学者供图

AI for Science

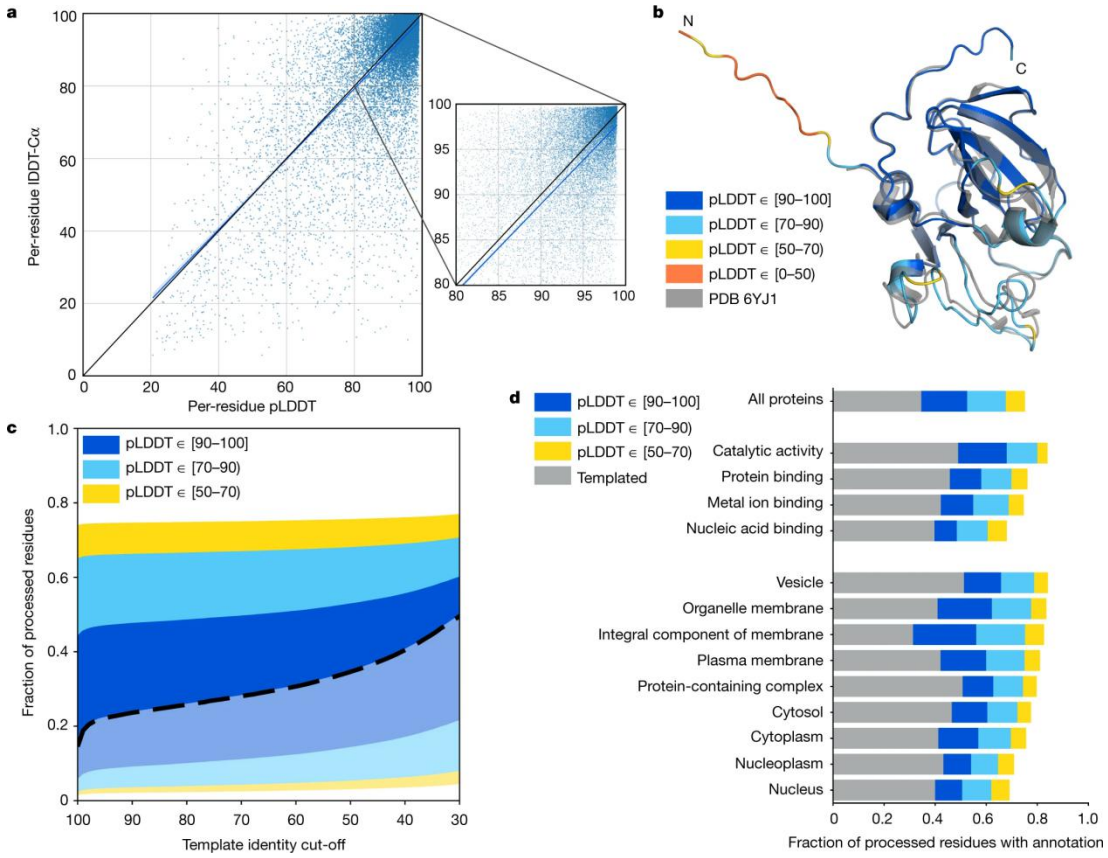
传统科研领域成为人工智能发展的“新战场”

近年来，人工智能在科研领域被初步应用，越来越多的科学家自研或采用成熟的人工智能算法，辅助进行数据挖掘分析、建模、仿真、预测等科研工作，加快发现自然科学新规律、新模式，减少重复性人力工作，提升科学发现的准确性，显著提高科研人员的工作效率。目前，人工智能已在物理学、化学、材料学、生物学等领域得到应用，如 2020 年 11 月“深度势能”团队获得戈登·贝尔奖（Gordon Bell Prize），AlphaFold2 破解蛋白质折叠预测难题等。随着人工智能技术和科学研究的结合愈发紧密，已出现了“AI for Science”（人工智能科学研究）的新兴研究领域。中科院院士，北京大学教授鄂维南在 2021 年智源大会发言中提出：“传统科研领域应当成为人工智能的主战场，要利用人工智能全面提升科研能力，加快进入‘智能化科研’时代，推动对当前的工业和技术升级。”

DeepMind 开源 AlphaFold2 蛋白质预测算法和数据库/华盛顿大学等提出并开源蛋白质预测算法 RoseTTAFold

7 月，DeepMind 使用新开发的 AlphaFold2 算法预测出了 35 万种蛋白质的结构，其中包括人类基因组表达的约 2 万种蛋白质，以及其他 20 种生物学研究中的常用模式生物（如大肠杆菌、酵母和果蝇）表达的蛋白质，是过去用实验方法解决的蛋白质数量的两倍多。研究发现，

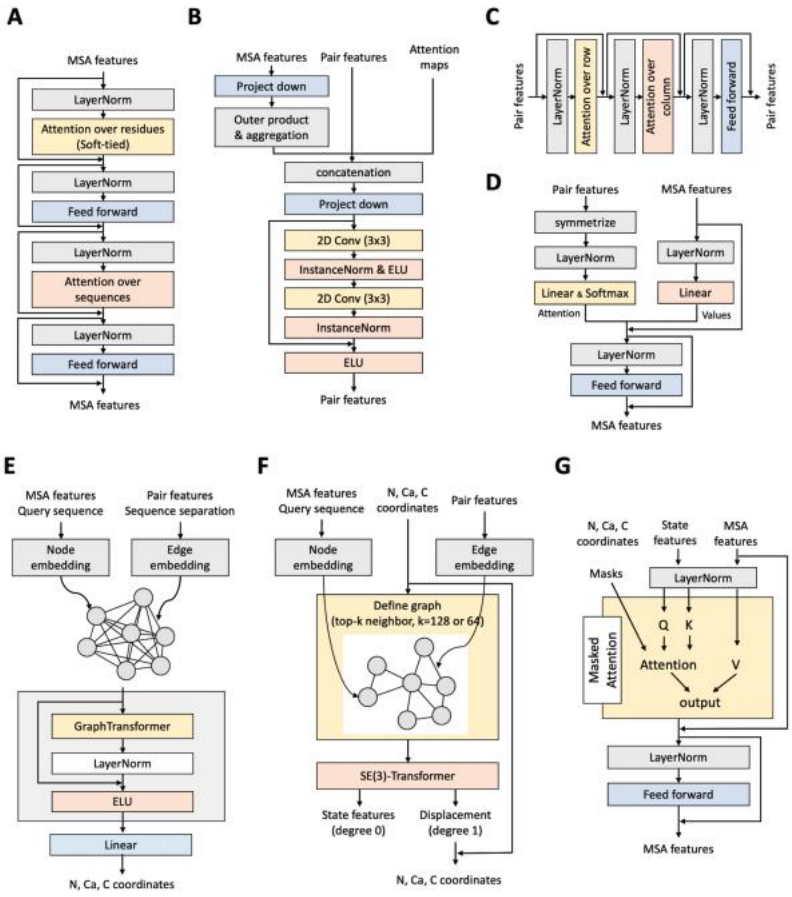
AlphaFold2 能对人类蛋白质组中 58%的氨基酸结构位置给出可信预测。35.7%的结构位置的预测达到了高置信度，是实验方法覆盖结构数量的两倍。在蛋白层面，AlphaFold2 对 43.8%的蛋白中至少四分之三的氨基酸序列都给出了可信预测，该研究于 7 月 22 日登上《自然》杂志。DeepMind 宣布，已与欧洲生物信息研究所（EMBL-EBI）合作建立 AlphaFold DB 蛋白质结构数据库，将覆盖 98.5%的人类蛋白质信息，预测结果免费开放。12 月，这项研究被《自然》杂志评为 2021 年度技术突破。



图注：AlphaFold2 模型在多种物质上预测结果的置信区间

来源：https://www.nature.com/articles/s41586-021-03828-1

7 月，华盛顿大学、哈佛大学等的研究者提出蛋白质结构预测算法 RoseTTAFold，该方法基于深度学习，通过在蛋白质序列信息的学习，能够快速生成蛋白质的精确结构，减少传统方法在实验测定等方面投入的时间和精力。目前该算法已开源。



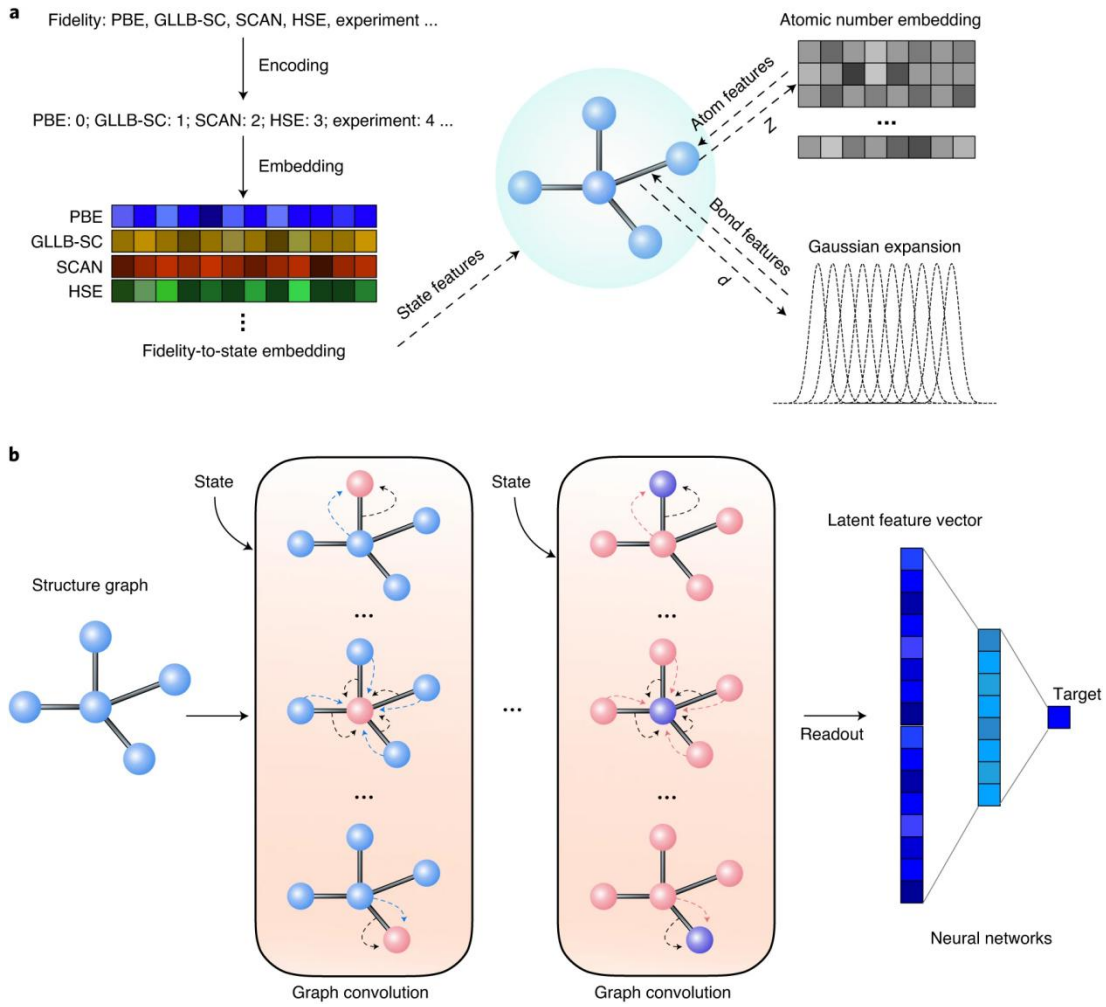
图注：RoseTTAFold 中用于预测蛋白质结构的一系列算法架构

来源：https://www.science.org/doi/abs/10.1126/science.abj8754

加州大学圣地亚哥分校研究者提出基于机器学习的材料筛选方法

1 月，来自加州大学圣地亚哥分校等机构的研究者提出了一种名为“Multi-fidelity Materials Graph Networks”（多精度材料图网络）的机器学习方法，通过学习来自多种测量和仿真

来源的数据，通过 AI 模型预测材料的特性。该方法能够构建出具有普遍意义、更准确的“材料属性模型”，从而帮助科学家筛选有研究前景的候选材料。



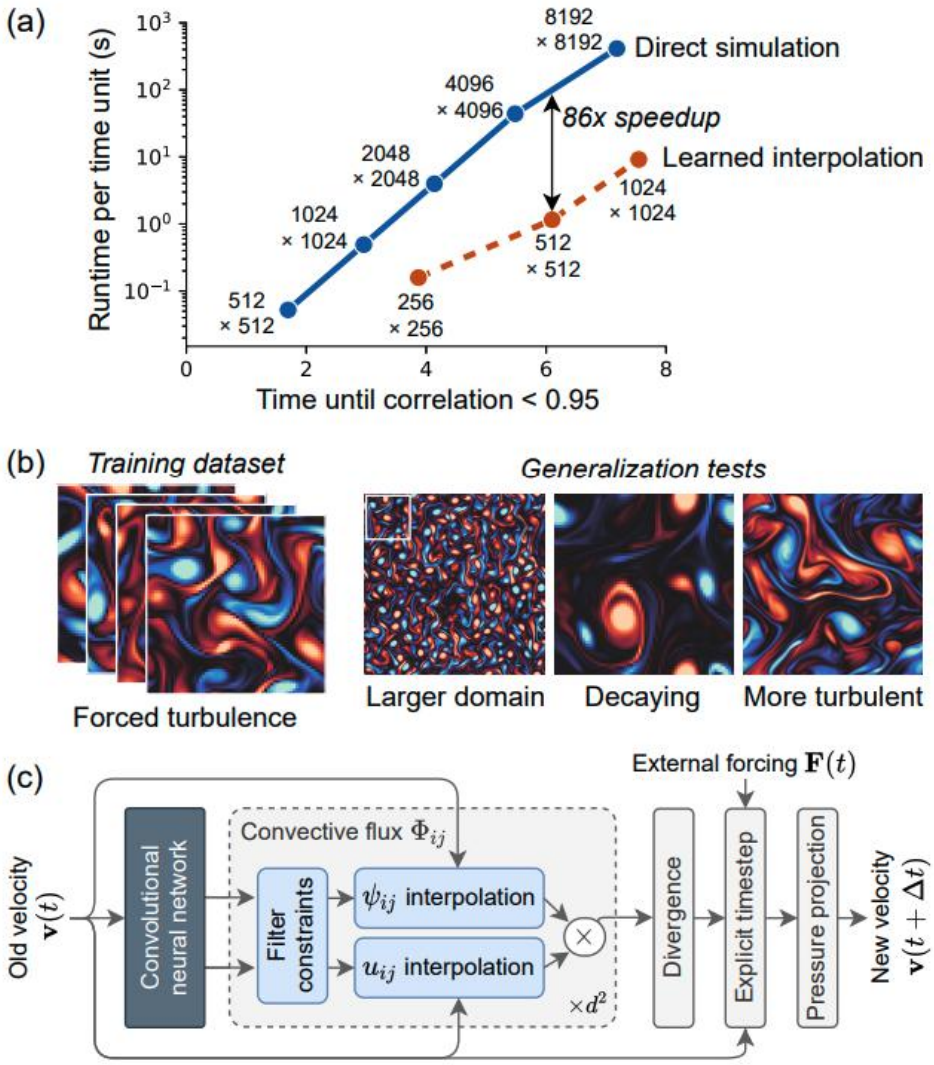
图注：多保真度材料图网络处理材料学数据和进行属性建模的方法

来源：https://www.nature.com/articles/s43588-020-00002-x

谷歌研究者利用深度学习改进用于建模二维湍流的计算流体动力学中的近似结果

1 月，谷歌的研究者使用端到端深度学习方法改进用于建模二维湍流的计算流体动力学中的近似结果（Approximations）。在湍流的直接数值仿真（Direct Numerical Simulation,

DNS）和大涡仿真（Large Eddy Simulation，LES）方面，该方法在每个空间维度的分辨率是前者的 8-10 倍，实现了 40-80 倍的计算加速。

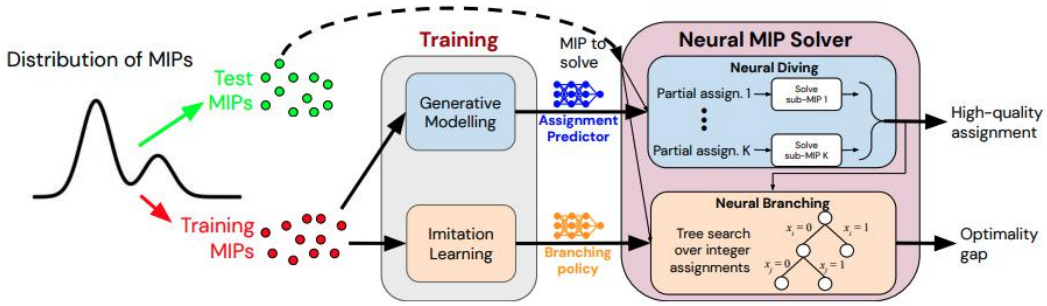


图注：该研究提出方法与结果概览：a）精确度和计算成本之间的关系（蓝线：基线方法；红线：机器学习方法）；b）训练数据集和生成数据集的结果；c）该方法采用的端到端深度学习网络和处理数据的流程。

来源：https://arxiv.org/pdf/2102.01010.pdf

DeepMind、谷歌研究者基于神经网络算法改进混合整数编程求解器

7 月，DeepMind 和谷歌的研究者提出神经网络算法，对常见的混合整数编程（Mixed Integer Programming：MIP）求解器（如 SCIP）进行改进，并应用于 MIP 求解器的两个关键子任务。该工作根据求解器生成的高质量变量分配（Assignment）和最优（Optimal）分配之间的目标值差异（Gap in Objective Value）来衡量性能。相比原有的求解器的平均原始对偶差（Primal-dual gap），通过机器学习增强的 SCIP 在 3 个具有最大 MIP 的数据集（一共有 5 个数据集）上实现了 1.5x、2x 和 104x 的较优差（Better Gap），在第 4 个数据集上以 5x 的速度更快实现 10%的差距，在第 5 个数据集上取得了与 SCIP 不相上下的表现。

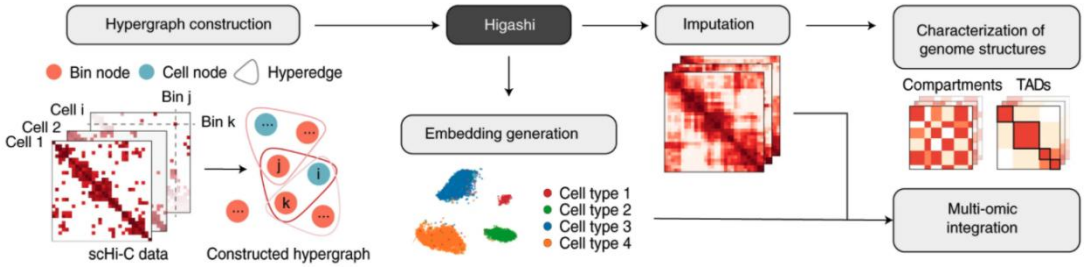


图注：该方法采用的算法架构

来源：https://arxiv.org/pdf/2012.13349.pdf

CMU 研究者提出超图表示学习算法诠释人体基因组折叠方式

10 月，卡内基梅隆大学（CMU）计算生物学系（CBD）的研究者提出了一种基于超图表示学习（Hypergraph Representation Learning）的算法“Higashi”，可以诠释人体细胞核中基因组的折叠方式，以及这些折叠方式如何影响基因表达，论文发表于 10 月的《自然》杂志上。

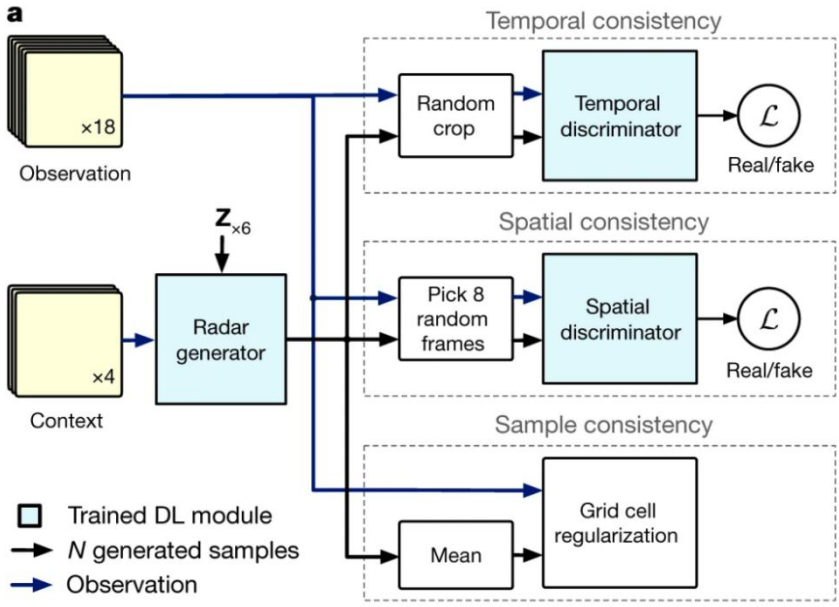


图注：Higashi 框架的整体结构

来源：https://www.nature.com/articles/s41587-021-01034-y

DeepMind 提出 AI 应用于降雨预测的深度生成模型

10 月，DeepMind 在《自然》杂志发表论文，通过与英国气象局合作，将 AI 技术应用于降雨预测。研究者采用深度生成模型，可提前 5-90 分钟预测 1536km×1280km 区域内的降水情况。与其他方法对比，该模型在 89%的情况下中具有最高的准确度和实用性（Usefulness）。

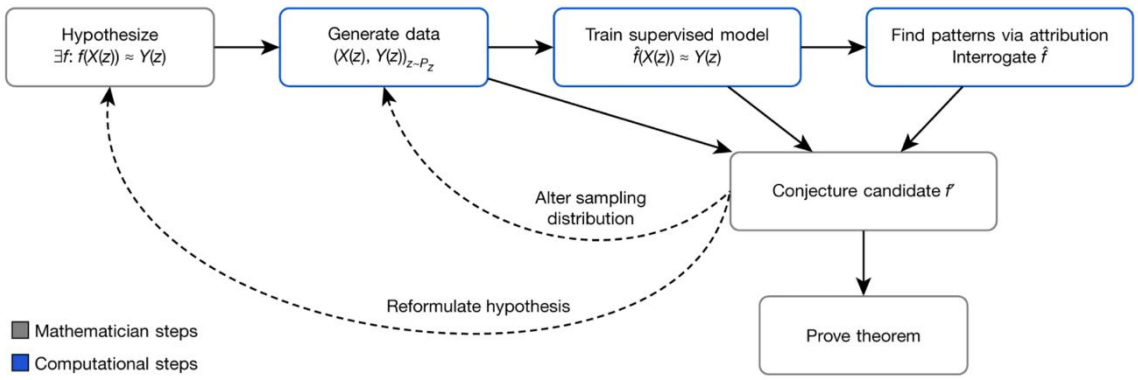


图注：DeepMind 提出的 AI 模型架构

来源：https://www.nature.com/articles/s41586-021-03854-z

DeepMind 研究者提出通过机器学习辅助发现数学猜想的方法

12 月，DeepMind 研究者提出采用机器学习辅助发现数学猜想和定理的方法。通过人工智能技术，我们能够发现数学目标中潜在的模式和关系，理解这些目标之间的属性机理，并帮助数学界寻找直觉和模式。研究者通过人工智能技术成功帮助数学家进一步解决数学问题。例如：（1）扭结的代数和几何结构的新关联；（2）由对称群的组合不变性猜想预测出的候选函数。研究者认为，这一研究有助于形成一种数学和人工智能领域之间的协作模式，通过利用两个领域之间的优势，得到更加具有突破性的发现。



图注：DeepMind 提出的用 AI 发现数学定理的流程，其中蓝色框为计算步骤，灰色框为数学家参与的步骤

来源：https://www.nature.com/articles/s41586-021-04086-x

DeepMind 用深度学习算法解决密度泛函理论中的分数电子问题

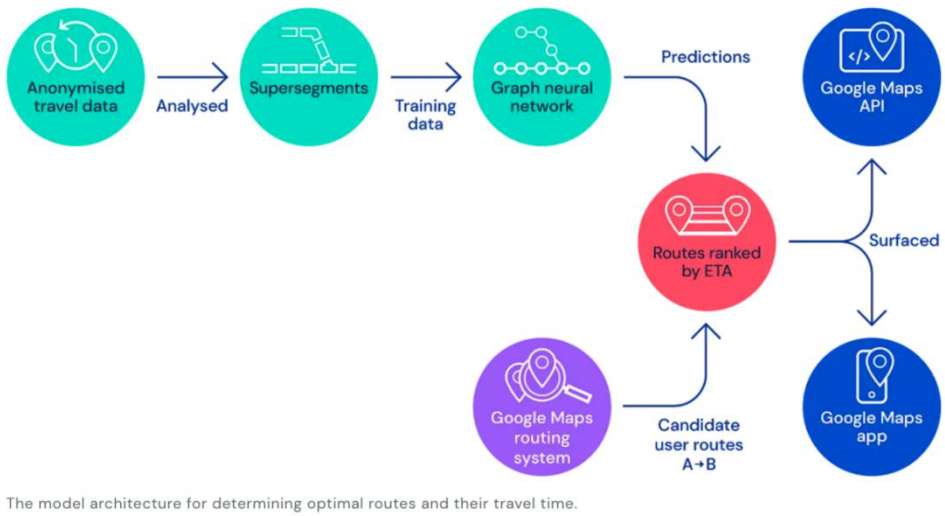
过去 30 年，密度泛函理论（Density Functional Theory，DFT）已经成为最广泛使用的电子结构方法，用于预测多种化学、生物学、材料科学领域系统的属性。但是当前最先进的密度泛函理论方法存在局限。12 月，DeepMind 研究者开发了深度学习算法，使用精确的化学

数据和分数电子约束数据进行训练。训练后的模型函数能够超越传统函数在主族原子和分子基准上的表现。论文认为，这项工作为 DFT 中长期存在的关键问题提供了一个解决方案，并展示了将 DFT 和现在机器学习方法结合的成功性。

人工智能技术提升智能产品和服务的性能

DeepMind 联合谷歌发布基于图神经网络的交通预测算法

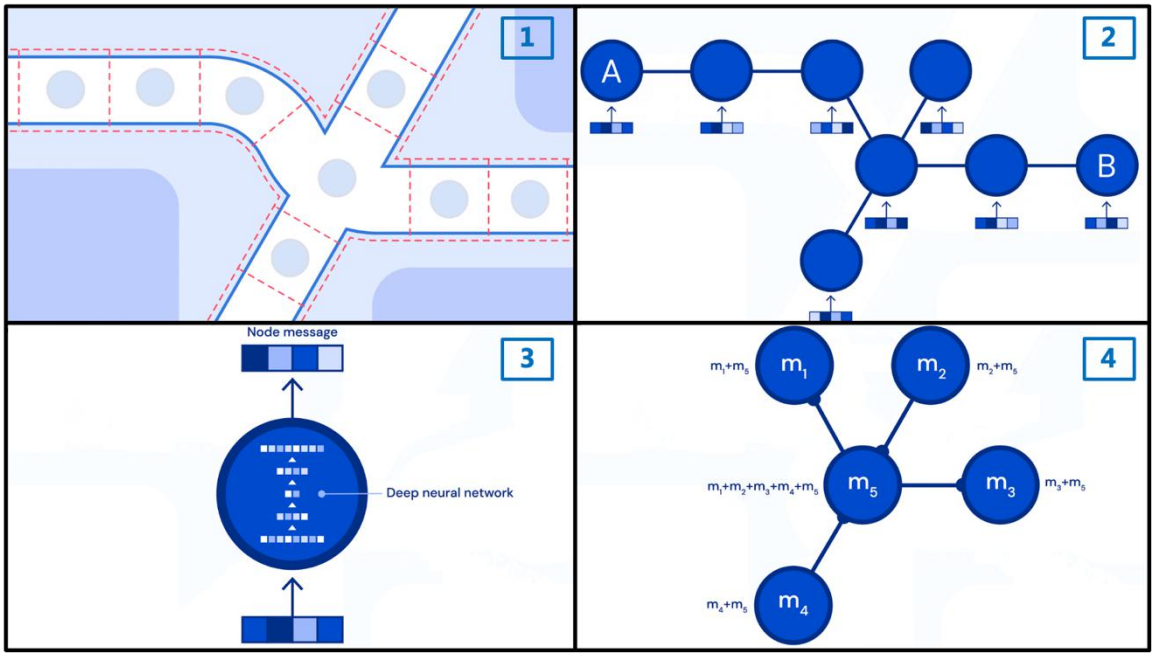
人工智能能够催生出性能更强的产品和服务。9 月，DeepMind 联合谷歌发布了新的交通预测算法，其基于图神经网络，通过将地图信息转换为图节点和边，并使用图神经网络进行处理，能够大幅提高预计到达时间（Estimated Times of Arrival，ETAs）的准确度。



图注：采用图神经网络改善 ETA 的系统流程

来源：

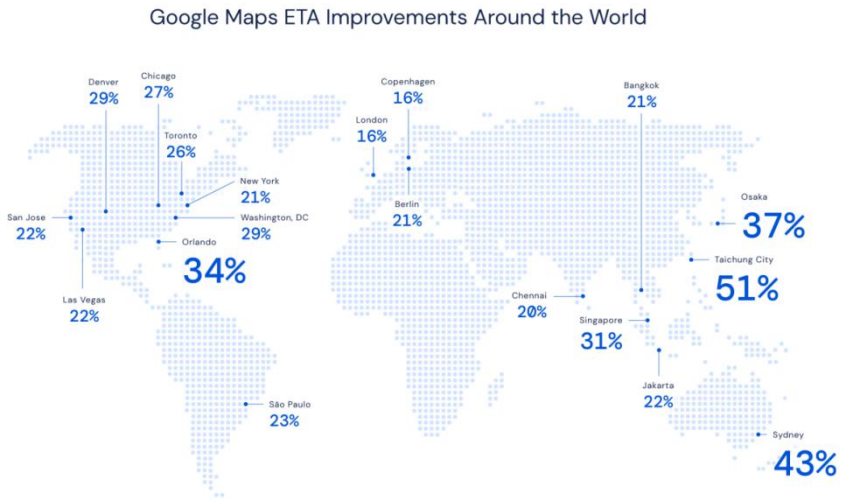
https://deepmind.com/blog/article/traffic-prediction-with-advanced-graph-neural-networks



图注：将道路转换为图的节点和边，并采用图神经网络进行处理的流程

来源：

https://deepmind.com/blog/article/traffic-prediction-with-advanced-graph-neural-networks



图注：图神经网络对 Google Map 中全球多个城市的 ETA 预测的提升情况

来源：

https://deepmind.com/blog/article/traffic-prediction-with-advanced-graph-neural-networks

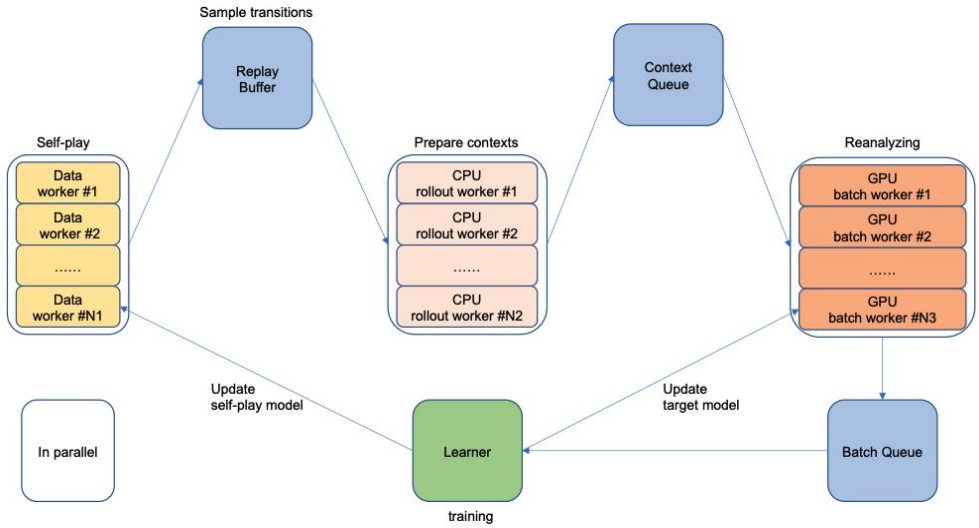
强化学习

提升训练效率成为强化学习领域的研究重点

近来，许多研究者期望能够探索出更为高效的强化学习算法，一是具有较好泛化能力，适用于多种场景；二是在输入数据较少或较为简单，类似真实环境的情形下，智能体依然能够取得较好的表现。目前已有 MuZero 等实现了这一目标。然而，强化学习也面临样本效率的挑战。从零开始训练智能体，往往需要通过成百上千万的步骤才能达到预期的性能表现，这会增加智能体对于算力的需求，不适合在真实场景下部署应用。

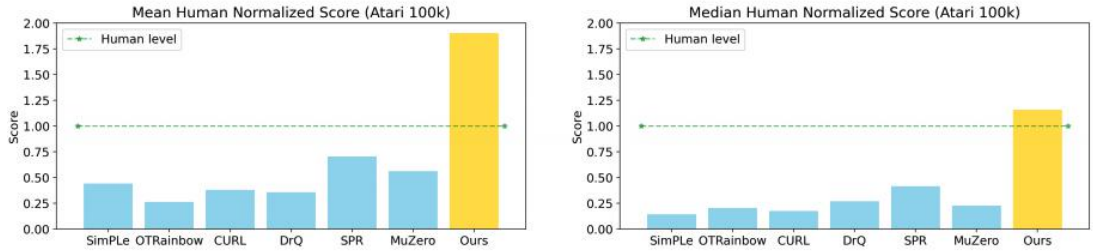
清华大学研究者提出小数据强化学习算法 EfficientZero

11 月，清华大学交叉研究院高阳课题组发表论文，提出小数据强化学习算法 EfficientZero，仅需要两个小时的真实时间训练，该算法比人类在雅达利 100k 数据集上的评价表现高了 190.4%，比中值表现高了 116%。同时，EfficientZero 已接近 DQN 在 2 亿帧上的性能，但数据需求量降低了 500 倍。



图注：EfficientZero 的部署 Pipeline

来源：https://arxiv.org/pdf/2111.00210.pdf



图注：Atari 100k 数据集上 EfficientZero 和其他算法在同等规模训练数据下的性能对比

来源：https://arxiv.org/pdf/2111.00210.pdf

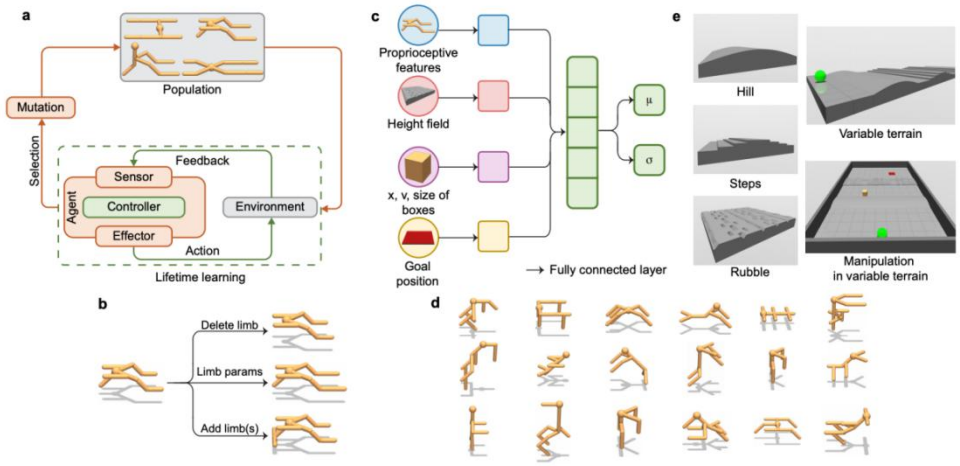
强化学习环境成为发展泛化性更强、适应复杂环境智能体的重要支撑

在强化学习的发展过程中，为智能体打造适合的训练环境，提供丰富多样的环境反馈，全面评价智能体的表现，是许多科研机构关注的问题。OpenAI 曾推出 Gym 和 Universe 两个强

化学习平台，为训练新一代智能体提供了丰富的游戏、环境和评测支持。近年来，能够模拟更为真实和复杂的训练环境，具有智能体配置、环境设置、训练、评价一条龙服务的强化学习平台不断涌现。

斯坦福大学李飞飞等学者提出深度进化强化学习框架

在自然界中，动物利用其形态来学习复杂的任务，获得显着程度的具身智能（Embodied Intelligence）。具身智能假设智能行为可以被具有对应形态的智能体通过适应环境的方式学习到。在强化学习中，创建具有特定形态的智能体，使其通过具身性获得智能能力是一大挑战。2 月，斯坦福大学李飞飞等学者提出了名为深度进化强化学习（Deep Evolutionary Reinforcement Learning，DERL）框架。该框架可以让智能体通过在复杂的任务和环境中，仅依赖低层次自我中心（Low Level Ego-Centric）传感信息的方式，逐步进化出多样的智能体形态。通过 DERL，研究者发现了一些环境复杂性和形态智能，控制学习能力等之间的关系。

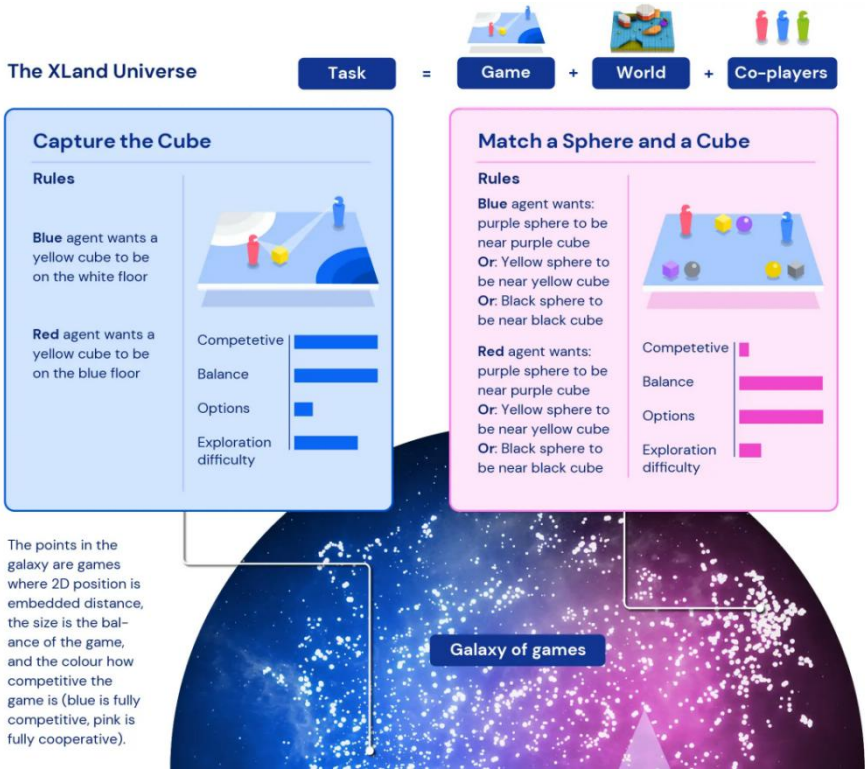


图注：DERL 框架的总体架构和构建智能体的方式

来源：https://arxiv.org/pdf/2102.02202.pdf

DeepMind 研究者提出 XLand 通用智能体强化学习训练环境

7 月，DeepMind 研究者提出一种名为 XLand 的通用智能体强化学习训练环境。DeepMind 认为，泛化能力不足是限制当前强化学习算法应用的一大障碍。由于泛化能力并不是一蹴而就形成的，人类是从简单的任务开始，逐渐掌握复杂的任务。受此启发，DeepMind 提出 XLand，其中包含了数十亿个任务，涵盖雅达利、夺旗、Dota2、捉迷藏等不同的游戏、世界和玩家对象。AI 智能体首先学习简单任务，不断完善，然后逐渐在更为复杂的任务上训练。智能体在 XLand 的 4000 个独立世界中能够玩大约 70 万个独立游戏，涉及 340 万个独立任务。



图注：XLand 涵盖的任务、游戏类型和使用方法

来源：

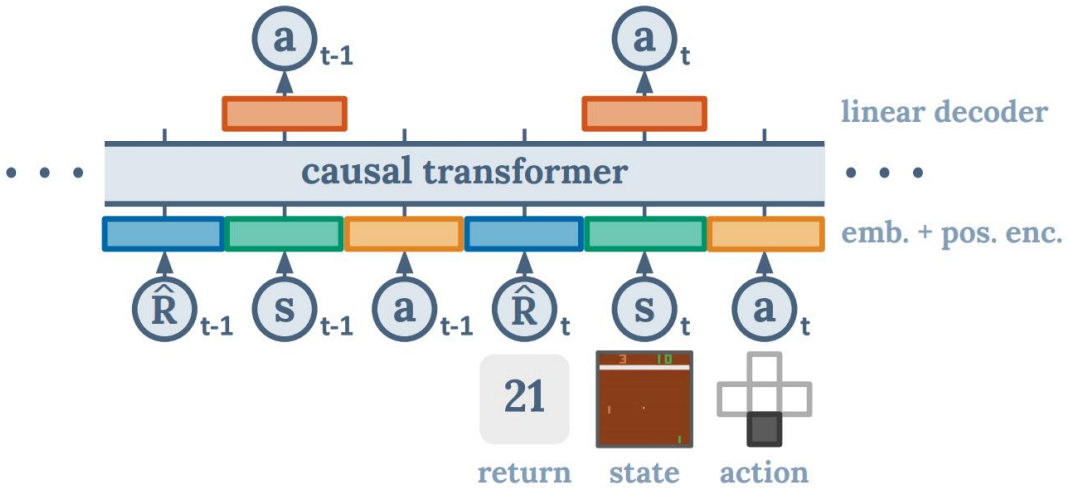
https://deepmind.com/blog/article/generally-capable-agents-emerge-from-open-ended-play

Transformer 渗透强化学习领域

Transformer 的快速发展，有望成为人工智能领域通用算法架构。许多研究者认为，将智能体的行为转换为序列，并进行建模，就可以在 Transformer 架构中进行学习和训练，因此目前有许多研究者也在探究其对构建更高效智能体所带来的影响。

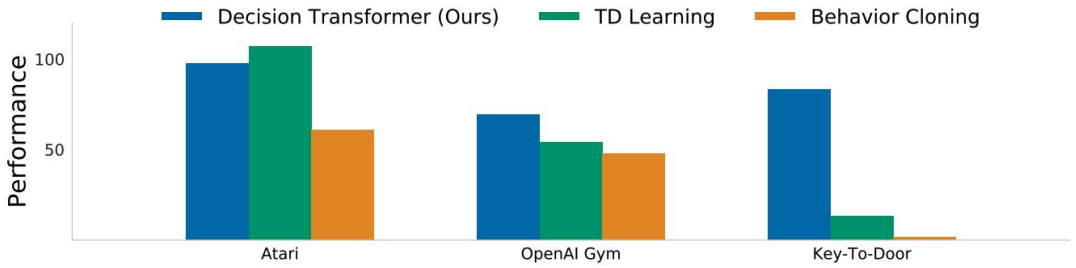
加州大学伯克利分校等研究者提出基于 Transformer 的强化学习架构

6 月，加州大学伯克利分校、Facebook、谷歌的研究者提出了一种序列建模强化学习的方法，构建了基于 Transformer 的强化学习架构。实验显示，在 Atari、OpenAI Gym、Minigrid 进行测试，Decision Transformer 均可达到与其他算法媲美甚至超越的性能表现。



图注：Decision Transformer 的架构

来源：https://arxiv.org/pdf/2106.01345.pdf



图注：将 Decision Transformer 和其他强化学习算法比较的结果（采用 Normalized Episode Return）

来源：https://arxiv.org/pdf/2106.01345.pdf

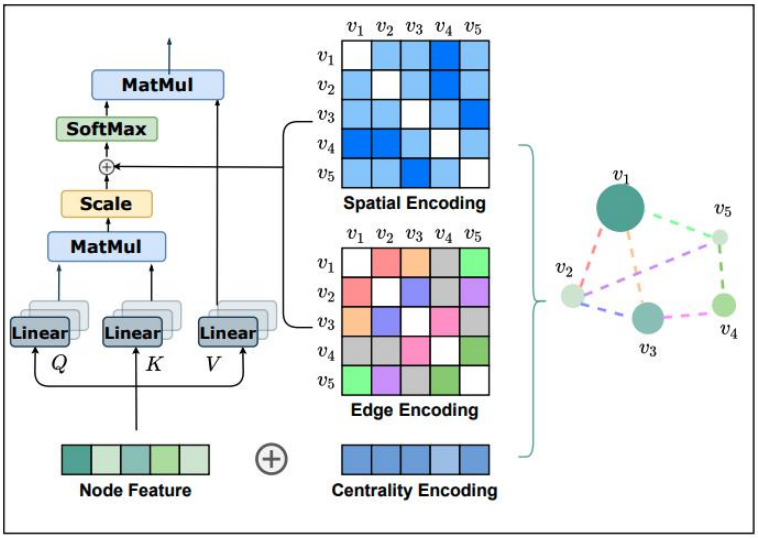
其他值得关注的 AI 研究和热点

Transformer 和图神经网络结合产生更强的性能表现

当前，因其具有更强的表征学习能力，Transformer 在深度学习领域快速发展，正在逐渐取代多种传统的神经网络结构。在图神经网络领域，研究者也尝试将 Transformer 和图网络结合，更好地捕捉图节点和边的关系数据。

7 月，大连科技大学、普林斯顿大学、北大、微软亚洲研究院的研究者提出了名为 Graphormer 的图神经网络模型，在一系列图表示学习任务上取得了优异的结果。研究者采用 Transformer，输入节点中心性编码（Centrality Encoding）、节点间空间结构信息编码（Spatial Encoding）和边编码（Edge Encoding）等信息，使模型对于图具有更好的表征能力。研究者还提出了一些简单但有效的结构编码方法，帮助 Graphormer 更好地建模图结构数据，此外，研究还

通过数学方法展示了Graphormer的表示能力,认为采用该研究提出的图结构信息编码方式,许多流行的图神经网络变体可以被认为是 Graphormer 的特殊案例。研究者在 OGB-LSC 量子化学回归挑战（Quantum Chemistry Regression Challenge）、OGBG-MolHIV、OGBG-MolPCBA 和 ZINC 图表示学习任务上进行了实验，取得了最佳的性能表现。Graphormer 在 2021 KDD 数据挖掘挑战赛上获得了 PCQM4M-LSC 赛道的冠军。



图注：Graphormer 的架构和将节点、边转换进行编码的方法

来源：https://arxiv.org/pdf/2106.05234.pdf

Awardees of PCQM4M-LSC Track (Leaderboard)

Winners

1st place: MachineLearning (contact)

- **Team members**: Chengxuan Ying (Dalian University of Technology), Mingqi Yang (Dalian University of Technology), Shengjie Luo (Peking University), Tianle Cai (Princeton University), Guolin Ke (MSRA), Di He (MSRA), Shuxin Zheng (MSRA), Chenglin Wu (Xiamen University), Yuxin Wang (Dalian University of Technology), Yanming Shen (Dalian University of Technology)
- **Method**: Graphormer (10 ensemble) + ExpC (8 ensemble)
- **Short summary**: We adopt Graphormer and ExpC as our basic models. We train each model by 8-fold cross-validation, and additionally train two Graphormer models on the union of training and validation sets with different random seeds. For final submission, we use a naive ensemble for these 18 models by taking average of their outputs.
- **Learn more:** Technical report, code
- **Test MAE**: 0.1200

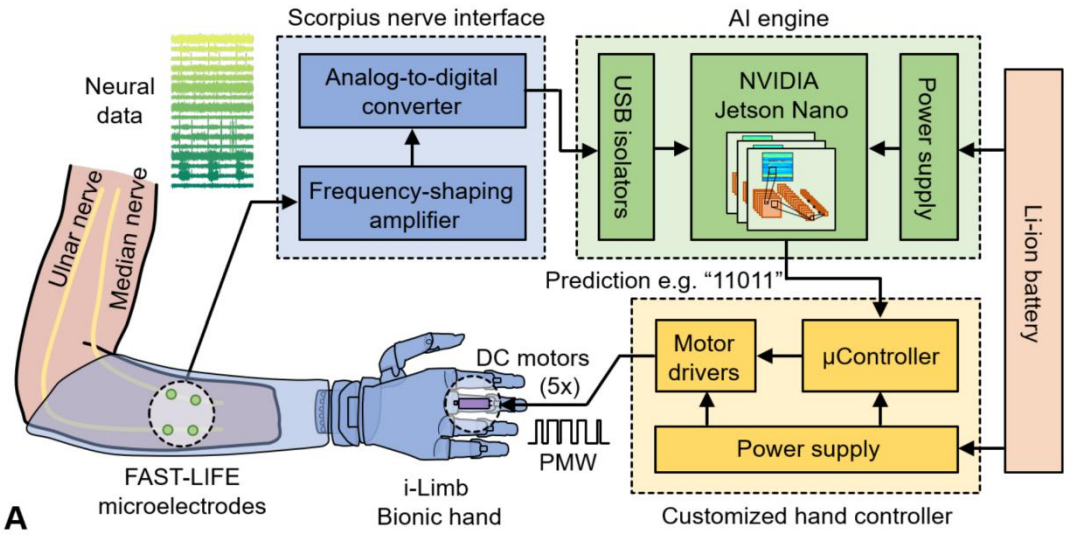
图注：采用 Graphormer+ExpC 方法的团队获得了 KDD Cup 2021 挑战赛 PCQM4M-LSC 赛道的冠军

来源：https://ogb.stanford.edu/kddcup2021/results/#awardees_pcqm4m

神经网络解码脑电信号，有望提升机器控制能力

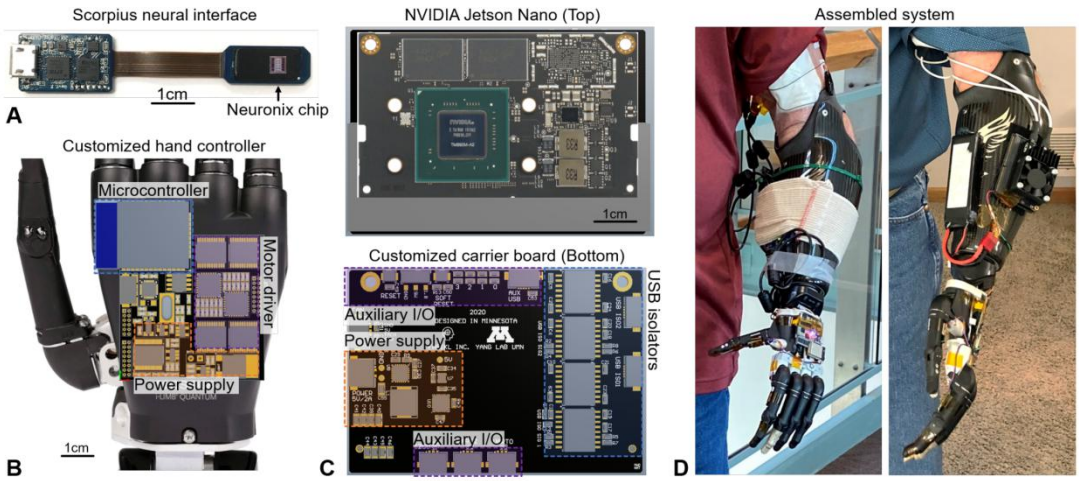
基于深度学习的解码器已能够实现对神经假肢的直观控制，但受限于算力限制，目前仍没有在临床条件下的实验。

3 月，美国明尼苏达大学等的研究者提出了一种有嵌入式深度神经网络控制的假肢手。实验显示，在多种实验室和真实环境下，该假肢手可以提供鲁棒、高准确度（95%-99%）和低延迟（50-120 毫秒）的单手指控制。



图注：该假肢手的系统构成

来源：https://arxiv.org/pdf/2103.13452.pdf



图注：假肢手的各部分零件和组成实物

来源：https://arxiv.org/pdf/2103.13452.pdf

因果推断在经济学、社会学研究中广泛应用

挖掘因果关系是众多科学研究的目标。近年来在各个科学领域，特别是大数据和人工智能领域对因果推断研究的热情高涨，图灵奖获得者 Judea Pearl 和 Yoshua Bengio 都认为因果推断是大数据和人工智能研究的一个突破口，人们需要一场“因果革命”来推动人工智能的发展。但因果推断也面临观察性研究、混杂因素、缺失数据带来的挑战。

2021 年诺贝尔经济学奖授予加州大学伯克利分校的 David Card、MIT 的 Joshua Angrist、斯坦福大学的 Guido Imbens，以表彰他们在经济学研究的实证研究和因果推断方法方面的贡献，其科学背景是观察性数据的因果推断。观察性研究是因果推断的主要数据来源，但不能进行人为控制的试验，其核心难题是混杂因素——处理和结果的公共原因未被观测，这种情

况下，因果作用无法识别，再精深的模型也无能为力。工具变量（Instrumental Variable）是观察性研究中推断因果作用和消除混杂因素的有效方法，自 Philip Wright 在经济学中率先使用已有近百年历史。一个有效的工具变量需要和关心的处理有强的相关性，但和混杂因素独立，对结果变量没有直接因果作用。在观察性研究中找到一个有效工具变量很困难。因此，人们也质疑使用工具变量能否作为推断因果作用的一个普遍方式。Card 使用自然试验作为工具变量分析劳动经济学中一系列重要的因果问题，重塑或加深了人们对这些因果关系的认识，如发现提高最低工资并不会减少就业，推翻了人们对最低工资和就业之间关系的广泛认识。不仅如此，自然试验在劳动经济学中的成功运用也使工具变量、重差法等成为实证研究中推断因果作用的普遍方式。在很长一段时间里，经济学家使用工具变量推断因果作用的主要依赖结构方程模型（Structural Equation Model），结构方程模型在形式上与回归模型相似，但结构方程模型非常隐晦地包含了刻画因果关系需要的假定，以至于经常被和表示相关关系的回归模型混为一谈，而其中的因果假定难以表示和验证。统计学家提出使用潜在结果定义因果作用（Potential Outcome Framework），有更强的表示能力，Angrist 和 Imbens 将工具变量与潜在结果模型结合，使用潜在结果模型刻画工具变量假定和相应的统计模型，定义新的因果概念，发展新的统计推断方法—相当于重建了工具变量方法。

这并不是诺贝尔经济学奖第一次颁发给因果推断的研究成果，1989 年 Trygve Haavelmo 和 2000 年 James Heckman 获诺贝尔经济学奖的主要贡献都与因果研究密切相关。Haavelmo 将数理统计引入经济学，明确经济学模型如联立方程组的因果意义，为计量经济学做出奠基性的工作，被称为计量经济学之父。Heckman 选择模型对观察性研究处理缺失数据和选择

偏差，以及因果推断消除混杂因素影响非常深远。Card，Angrist 和 Imbens 在工具变量的理论和实证方面的工作将因果推断研究推向新的高潮。

The Prize in Economic Sciences 2021

The Royal Swedish Academy of Sciences has decided to award the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021

with one half to

David Card

University of California, Berkeley, USA

"for his empirical contributions to labour economics"

and the other half jointly to

Joshua D. Angrist

Massachusetts Institute of Technology, Cambridge, USA

Guido W. Imbens

Stanford University, USA

"for their methodological contributions to the analysis of causal relationships"

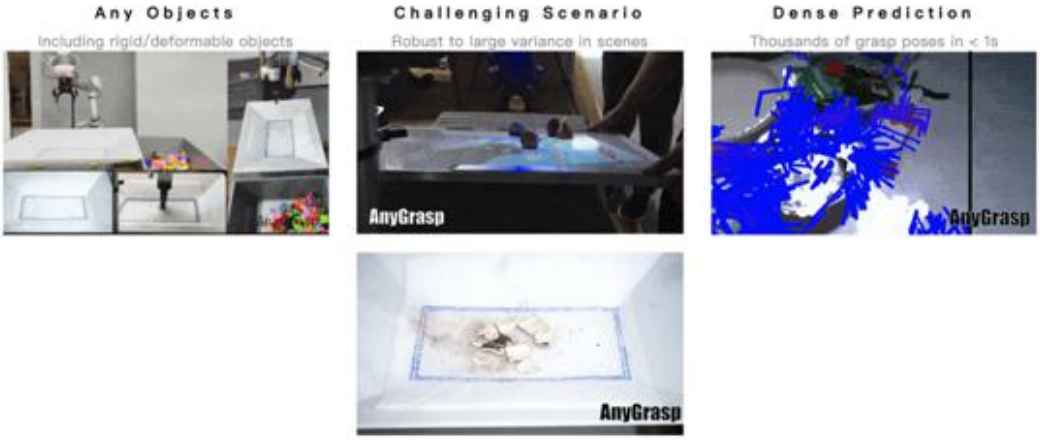
图注：2021 年诺贝尔经济学奖获奖名单

来源：https://www.nobelprize.org/uploads/2021/10/press-economicsciencesprize2021-2.pdf

基于视觉的机器人通用抓取研究实现突破

基于视觉的机器人通用抓取是研究和产业界关注的一个重点问题。目前已有研究实现了在实验室和简单日常场景中物体的抓取，但仅限于形状规则的物体，对于日常生活中大量堆叠、形态复杂、材质多样的物体来说，目前没有相关研究可以完全解决。此外，基于视觉的机器人抓取还需要面对光线干扰、抓取物体所在平面变化等挑战。6 月，上海交通大学团队提出了 AnyGrasp 抓取方法，无需物体的 CAD 模型，在不限制抓取硬件构型和检测相机的情况下，实现普适、高速、稳定、低成本的人类级别的抓取能力。

上海交通大学团队开发的 AnyGrasp 算法可以在 50ms 的时间内为场景生成数千个抓取候选点，同时能保证时序上相邻帧的抓取姿态连续性与稳定性。团队进行了大规模的真机实验，搭载 AnyGrasp 的机械臂在超过 500 个未训练过的物体上进行了超过 1000 次的抓取。物品集包括大量对抗样本，可变形样本，细小物品等。AnyGrasp 算法的抓取完成率高达 99.7%，抓取成功率达到 92%，该指标与人类抓取表现相当。机器人拥有人类级别的抓取能力有助于后续机器人进行灵巧物体操作。



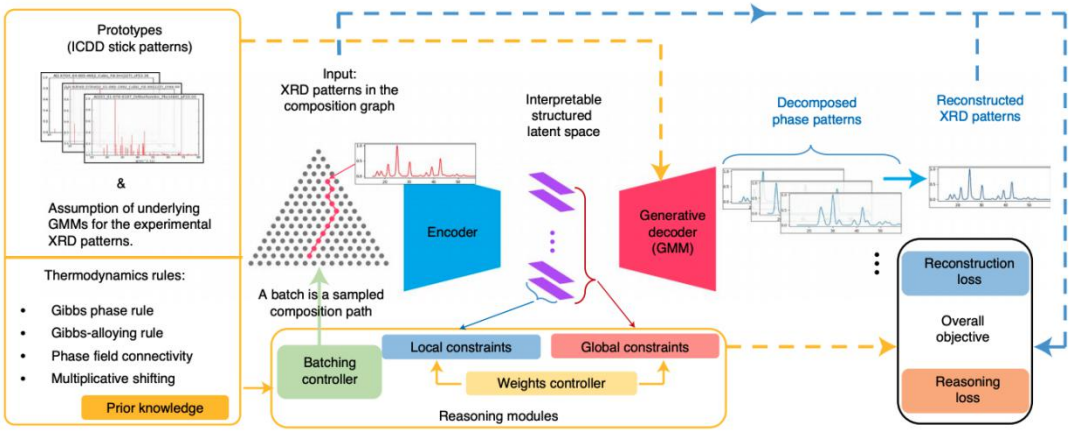
图注：AnyGrasp 适用的条件和场景

来源：https://graspnet.net/anygrasp.html

AI 在环境和可持续发展研究实现应用

为了推动经济、社会和环境的健康发展，联合国在 2015 年提出了可持续发展的 17 项目标（SDGs）。为了实现 2025 年联合国可持续发展目标，研究者近年来持续探索人工智能技术在这一领域的应用，如利用约束优化、动态建模、仿真、机器学习、多智能体系统等计算方法，用于生物多样性保护、环境和社会经济发展、材料和可再生能源研发等领域。

例如，在可再生能源材料研发方面，美国康奈尔大学 Carla Gomes 团队提出了基于 AI 的能源材料(燃料电池和太阳能燃料)研发方法。研究者将材料研发任务建模为无监督的模式解离问题，根据同步 X 射线数据推断晶体结构，通过构建深度推理网络来推理有关热力学规则的先验知识，并使用该模型提升高通量材料发现的速度。该研究已发表在 9 月的《自然 · 机器智能》（Nature · Machine Intelligence）杂志上。



图注：采用深度推理网络对晶体结构进行建模的过程

来源：https://www.cs.cornell.edu/gomes/pdf/2021_chen_nmi_drnets.pdf

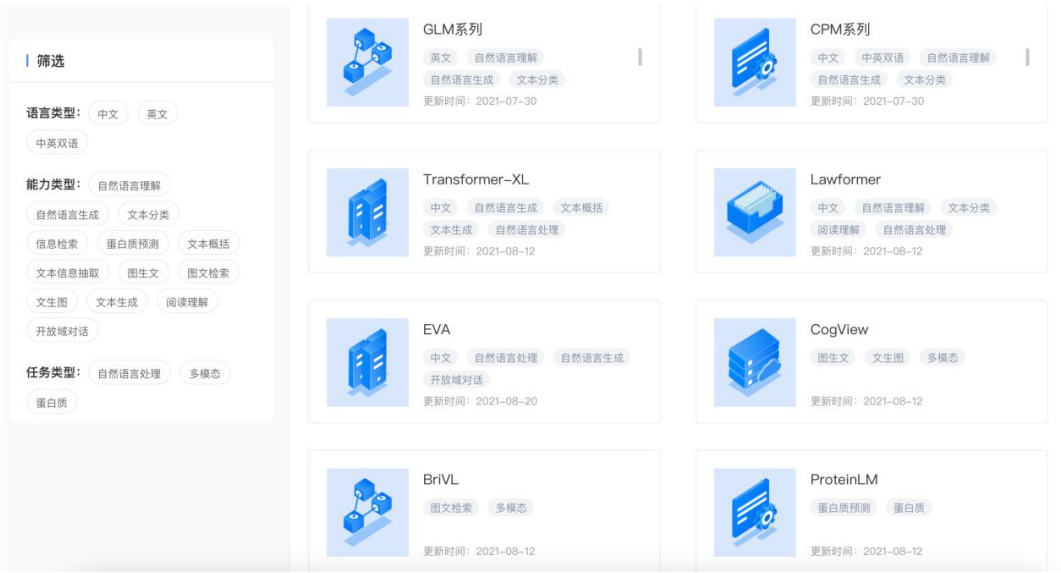
平台和工具发展情况

AI 系统

构建基于超大规模智能模型的 AI 开放平台成为研发机构和企业的重点发展思路

智源悟道大模型开放平台

9 月，智源研究院发布悟道大模型开放平台，包括一系列相关能力和资源，供开源开发者使用。模型能力方面，开放 GML、CPM 等通用语言模型，Transformer-XL 等语言生成模型，以及 EVA 中文对话模型、Cogview 图文生成模型、BriVL 图文检索模型，以及 ProteinLM 蛋白质序列模型等。此外，悟道平台也开放了 P-tuning 和 Inverse Tuning 算法，以及 FastMoE、InfMoE 等一系列框架和数据集。



图注：悟道开放的模型和能力

来源：悟道官网

阿里达摩院 AliceMind 平台

6 月，阿里达摩院开源 AliceMind 平台，平台包括了 PLUG、StructBERT、PALM 等多种预训练模型，用户可直接使用。此外，AliceMind 提供了模型微调、蒸馏、压缩等方面的工具。目前 AliceMind 平台上的应用案例包括对话机器人、企业知识库、医疗领域的信息抽取等。

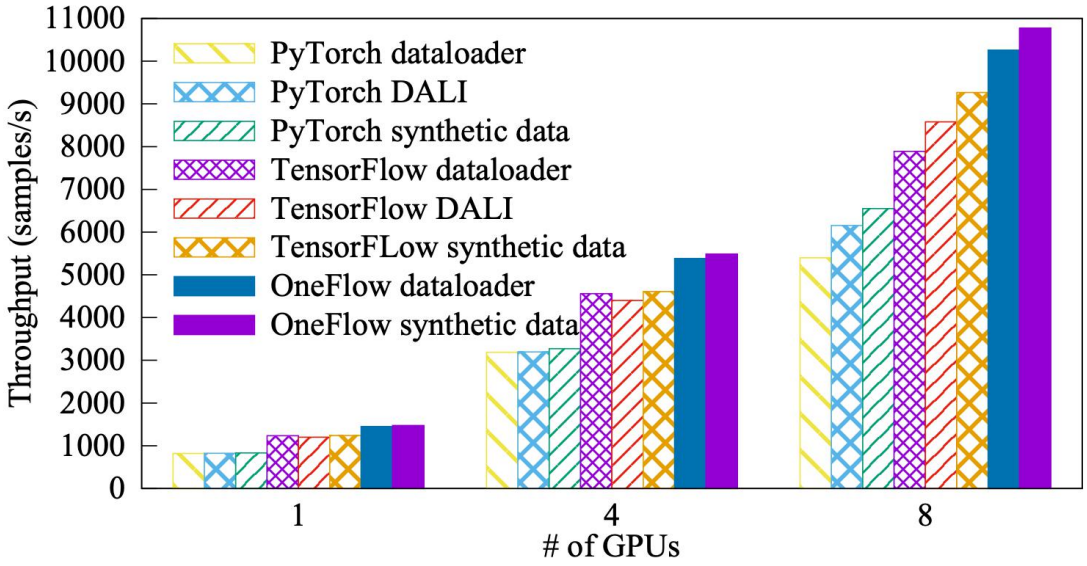
图注：AliceMind 平台支持的落地领域

来源：阿里达摩院

大规模深度学习的分布式训练势在必行

OneFlow 深度学习框架论文正式公开

10 月，一流科技公开了 OneFlow 深度学习框架的论文。基于 SBP（Split, Broadcast 和 Partial-value）抽象和 Actor 模型，一流科技研发出拥有多种并行范式的 OneFlow 分布式深度学习框架 。SBP 使数据并行和模型并行的编程比现有框架更容易，Actor 提供了一套简洁的运行时机制来管理分布式深度学习中的资源约束、数据搬运和计算所施加的复杂依赖关系。实验证明，OneFlow 在训练大型深度学习模型方面的性能优于很多框架。



图注：OneFlow 与深度学习框架在吞吐量上的对比

来源：https://arxiv.org/pdf/2110.15032.pdf

潞晨科技、新加坡国立大学研究者发布大规模并行 AI 训练系统夸父

10 月，潞晨科技、新加坡国立大学研究者发布大规模并行训练系统夸父（e Colossal-AI），是一个统一的并行训练系统，可与不同的并行范式（如数据并行、Pipeline 并行、多张量并行等）结合。通过多维并行、大规模优化器、自适应任务调度、消除冗余内存、降低能量损耗等方式，夸父分布式人工智能训练系统可以降低研究人员的开发门槛，让研究者像编写模型代码一样编写分布式模型，使得研究者不需要在开发过程中过分关注分布式训练。目前相关代码已开源。

```python
import colossalai
from colossalai.utils import get_dataloader


# my_config can be path to config file or a dictionary obj
# 'localhost' is only for single node, you need to specify
# the node name if using multiple nodes
colossalai.launch(
    config=my_config,
    rank=rank,
    world_size=world_size,
    backend='nccl',
    port=29500,
    host='localhost'
)

# build your model
model = ...

# build you dataset, the dataloader will have distributed data
# sampler by default
train_dataset = ...
train_dataloader = get_dataloader(dataset=dataset,
                                  shuffle=True,
                                  )
```

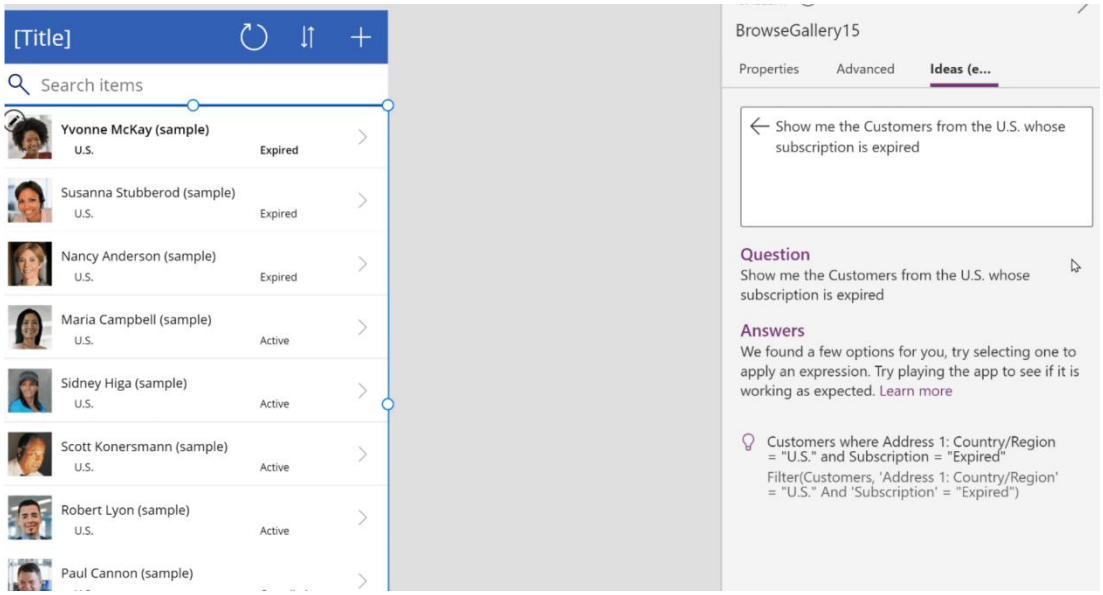
图注：调用夸父系统的示例代码

来源：https://github.com/hpcaitech/ColossalAI

超大规模智能模型支撑的行业应用进入探索落地阶段

微软将 GPT-3 应用于旗下产品

当前，超大规模智能模型已进入行业探索阶段，许多企业正在尝试引入这一技术，为其产品和服务提供更为强大的智能能力支持。5 月，微软宣布将 GPT-3 技术应用于 Microsoft Power App 数据分析平台中，能够自动生成代码，执行数据检索、计算等操作。



图注：GPT-3 能够在微软数据分析平台中根据自然语言指令生成对应的代码，加快数据检索效率

来源：微软官网

OpenAI 等推出 Copilot 代码生成插件和 Codex 模型

6 月，OpenAI 联合微软、GitHub 推出了基于大规模预训练语言模型研发的 Copilot 代码生成插件，能够集成在微软的编辑器 VS Code 中。Copilot 主要可以辅助三种开发任务，包括代码生成、代码补全、测试用例生成等。由于基于“大参数+大算力”进行训练，Copilot 具有非常惊人的性能，一些评论甚至认为 Copilot 带来了一种新的代码开发模式。

8 月，OpenAI 开放 Codex 模型 API 申请，用户可以申请试用。Codex 具有和 GPT-3 相似的架构，参数规模达 120 亿。研究者将 Codex 在数十亿的 GitHub 公开代码和自然语言数据上进行训练，使 Codex 可以理解自然语言，并根据用户的输入完成代码补全、函数注释补写等方面的任务。

图注：Copilot 的技术架构

来源：OpenAI 官网

OpenAI 开放 GPT-3 模型微调工具

12 月，OpenAI 宣布，开发人员可以使用提供的 API 和工具，创建基于自有数据集的 GPT-3 微调模型。用户可直接采用调用 Python 代码或命令行的形式构建定制化的 GPT-3 模型。

OpenAI CLI:

```
openai api completions.create -m <FINE_TUNED_MODEL> -p <YOUR_PROMPT>
```

cURL:

```
curl https://api.openai.com/v1/completions \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{"prompt": YOUR_PROMPT, "model": FINE_TUNED_MODEL}'
```

Python:

```
import openai
openai.Completion.create(
    model=FINE_TUNED_MODEL,
    prompt=YOUR_PROMPT)
```

图注：创建定制化 GPT-3 模型的过程

来源：https://beta.openai.com/docs/guides/fine-tuning

AI 算法和代码库

开源社区复现超大规模预训练模型

GPT-3 社区复现版开源

受制于研发成本、算力等因素中小企业、科研机构和开源开发者无法直接使用开源超大规模智能模型，而有些模型也没有开源，仅提供开放 API。因此，研发超大规模智能模型的开源替代版逐渐兴起。去年 GPT-3 诞生后，OpenAI 并未对外开源 GPT-3 模型本身，仅提供了免费试用 API 的渠道。

今年 3 月，GPT-3 的社区复现版本——GPT-Neo 对外开源。该模型最大达到 27 亿参数规模，在一些任务上有着和原版模型接近的性能表现。

Model and Size	Pile BPB	Pile PPL	Wikitext PPL	Lambada PPL	Lambada Acc	Winogrande	Hellaswag
GPT-Neo 125M	-----	-----	**32.285**	**30.266**	**37.36%**	**50.43%**	**28.67%**
GPT-3 125M	-----	-----	-----	18.6	42.7%	52.0%	33.7%
GPT-Neo 350M	-----	-----	**22.5657**	**13.876**	**47.27%**	**51.14%**	**32.16%**
GPT-3 350M	-----	-----	-----	9.09	54.3%	52.1%	43.6%
GPT-3 Ada	0.9631	-----	-----	9.954	51.60%	52.90%	35.93%
GPT-Neo 1.3B	**0.7527**	**6.159**	**13.10**	**7.498**	**57.23%**	**55.01%**	**38.66%**
GPT-3 1.3B	-----	-----	-----	5.44	63.6%	58.7%	54.7%
GPT-2 1.5B	1.0468	-----	17.48	10.634	51.21%	59.40%	40.03%
GPT-Neo 2.7B	**0.7165**	**5.646**	**11.39**	**5.626**	**62.22%**	**56.50%**	**42.73%**
GPT-3 2.7B	-----	-----	-----	4.60	67.1%	62.3%	62.8%

图注：GPT-Neo 和同样规模的 GPT-3 模型在同样任务上的性能对比

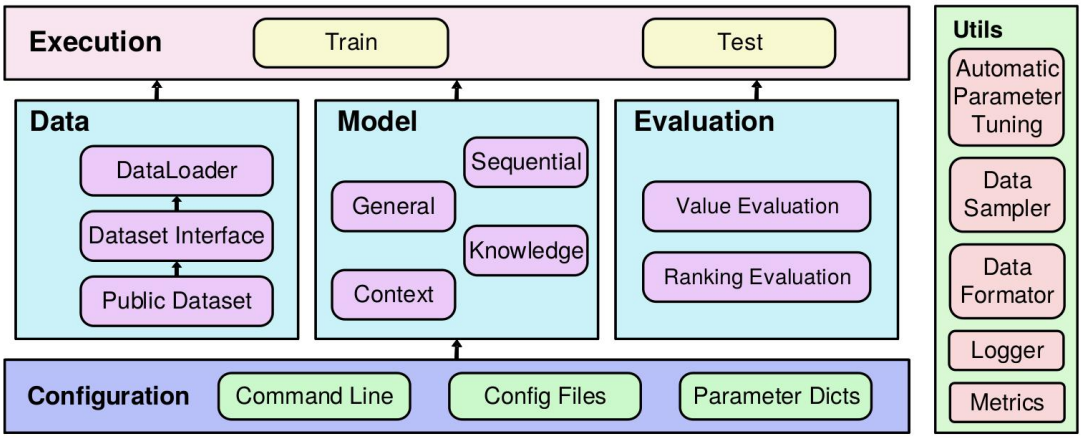
来源：https://github.com/EleutherAI/gpt-neo

多个领域开放 AI 代码库助力研究应用发展

在 AI 模型规模更加庞大，算法架构更为复杂，所需训练数据更为巨大和繁杂的情况下，从零开始从头研发效率较低，消耗科研人员的精力。因此，通过将已有的研发成果集成在开源代码库中，帮助其他开发者进行技术研发工作，已是 AI 领域的惯例。今年，在推荐、强化学习、机器人等领域，更多 AI 代码库开源，形成了垂直、活跃的交流社区。

人大、北邮等研究者开源 RecBole 推荐系统代码库

2020 年底，中国人民大学、北京邮电大学、华东师范大学的研究者开源了 RecBole 推荐系统代码库，目前已更新至 1.0 版本，GitHub 获星 1400 余次。RecBole 基于 PyTorch 实现，面向研究者，具有易于开发、复现、统一、全面、高效等优点。目前 RecBole 实现了 72 个推荐系统模型，覆盖了常见的推荐系统类别，如通用推荐、序列推荐、基于背景信息（Context-aware）的推荐、基于知识的推荐等。RecBole 约定了一个统一、易用的数据文件格式，并已支持 28 个基准测试数据集。 用户可以选择使用 RecBole 提供的数据集来预处理脚本，或直接下载已被处理好的数据集文件。此外，针对 GPU 环境，RecBole 也使用了一系列的优化技术来提升代码库的效率。



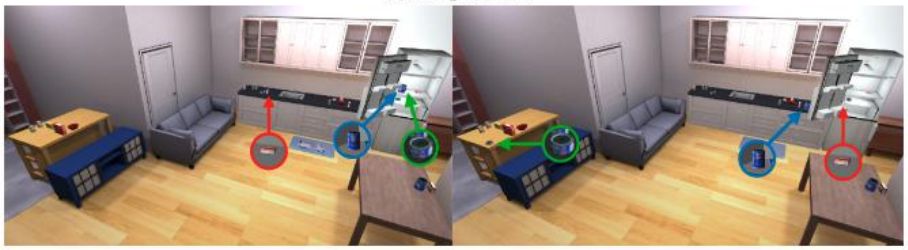
图注：RecBole 的架构

来源：https://github.com/RUCAIBox/RecBole/blob/master/README_CN.md

Facebook 推出机器人训练模拟平台 Habitat2.0

在机器人领域，6 月，Facebook 推出机器人训练模拟平台 Habitat2.0，支持研究者在虚拟环境中训练机器人的导航能力，并和虚拟环境中的其他物体进行交互。在虚拟环境中，智能体可以执行拾取物品、打开和关闭抽屉和门等动作。Habitat 2.0 还包括一个新的完全互动室内空间、3D 数据集和新基准，用于在这些复杂的物理功能的场景中训练和评价虚拟机器人的表现。

(a) TidyHouse



(b) PrepareGroceries

(c) SetTable

图注：Habitat 2.0 为机器人设置的一些任务，包括整理房间、准备杂货、布置餐桌等

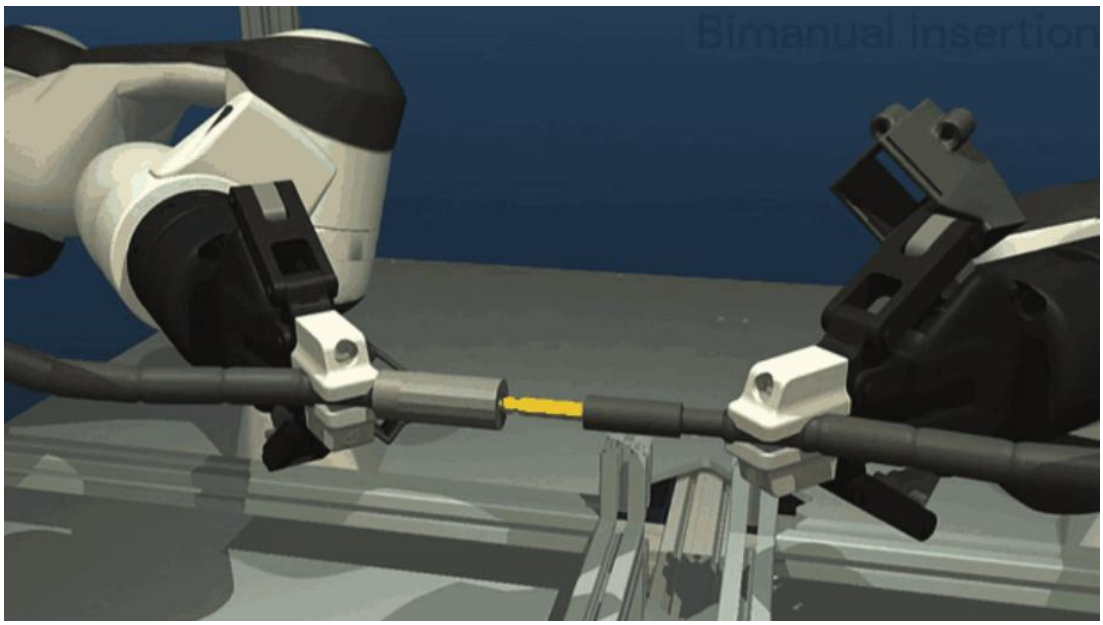
来源：https://arxiv.org/pdf/2106.14405.pdf

图注：虚拟机器人在 Habitat 2.0 环境中学习

来源：https://arxiv.org/pdf/2106.14405.pdf

DeepMind 收购物理模拟引擎产品 MuJoCo

10 月，DeepMind 收购物理模拟引擎产品 MuJoCo，并宣布将开源多关节接触动力学（Multi-Joint Dynamics with Contact）技术。MuJoCo 技术由华盛顿大学研究者开发，曾是付费产品，其结合了广义坐标模拟和优化接触动力学，能够模拟完整的物理运动和物体间的交互关系，是机器人研究者使用的模拟器之一。DeepMind 表示，相关代码库将在 2022 年发布。

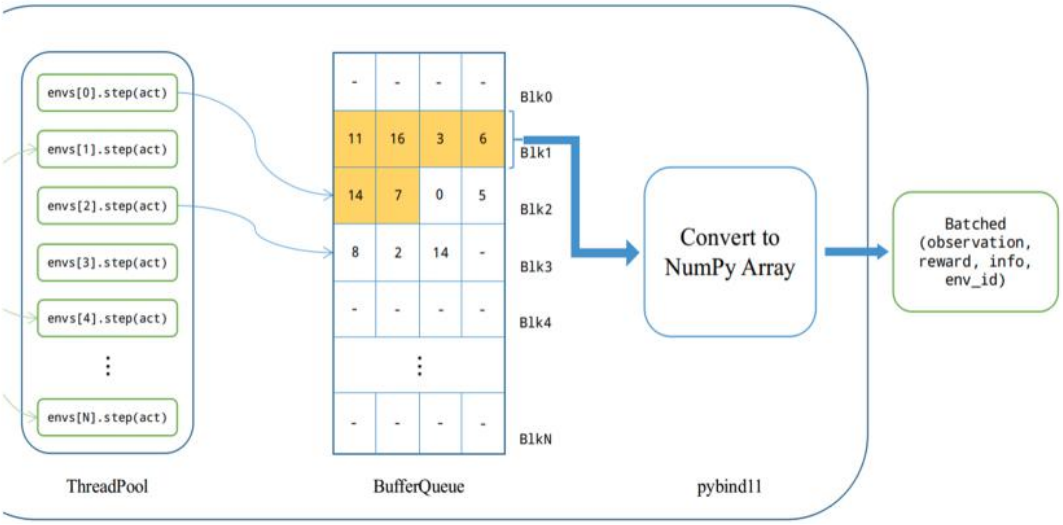


图注：MuJoCo 可以对复杂的物体交互进行模拟

来源：https://deepmind.com/blog/announcements/mujoco

新加坡 Sea AI Lab 团队开源强化学习环境模拟执行引擎 EnvPool

强化学习领域，11 月，新加坡 Sea AI Lab 团队提出强化学习环境模拟执行引擎 EnvPool，能够在 256 个 CPU 的条件下实现每秒一万帧的模拟，支持的强化学习环境包括雅达利、VizDoom 等，并与 DeepMind 和 OpenAI 的强化学习 API 兼容。



图注：EvnPool 的整体架构

来源：

https://events.rainfocus.com/widget/nvidia/nvidiagtc/sessioncatalog/session/1630239583490001Z5dE

英伟达发布 Faster Transformer4.0

在 AI 模型训练工具方面，4 月，英伟达发布了 Faster Transformer4.0 版，增加了对千亿级参数规模 GPT-3 的支持，让 Faster Transformer 采用张量并行和 Pipeline 并行进行多 GPU 推理。

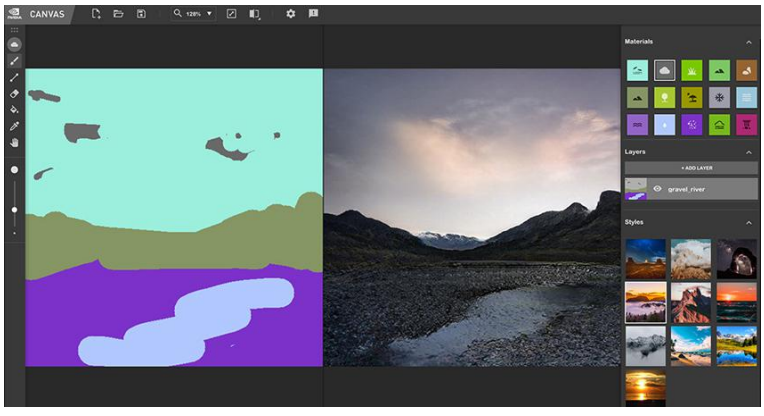
Batch size	Input SeqLen	Output SeqLen	Megatron (ms)	FT* (ms)	FT* speedup
1	128	8	660.38	488.86	1.35
2	128	8	687.34	509.47	1.35
4	128	8	1004.88	629.64	1.6
8	128	8	1705.07	749.86	2.27
16	128	8	3111.57	1037.47	3
1	512	32	2384.78	1719.96	1.39
2	512	32	2503.24	1830.56	1.37
4	512	32	3658.65	2092.56	1.75
8	512	32	6238.79	2629.97	2.37
16	512	32	11409.53	3706.23	3.08

图注：Faster Transformer 与 Megatron 在 GPT-175B 上的时延与加速比

来源：https://zhuanlan.zhihu.com/p/363517823

英伟达发布 Canvas 图像生成程序

图像生成方面，6 月，英伟达基于 2019 年提出的 GauGAN 框架，发布了名为“Canvas”的图像生成程序，用户可以采用材料“画笔”在虚拟画布上描绘形状，然后由 AI 将这些形状转换为真实的图像组合。

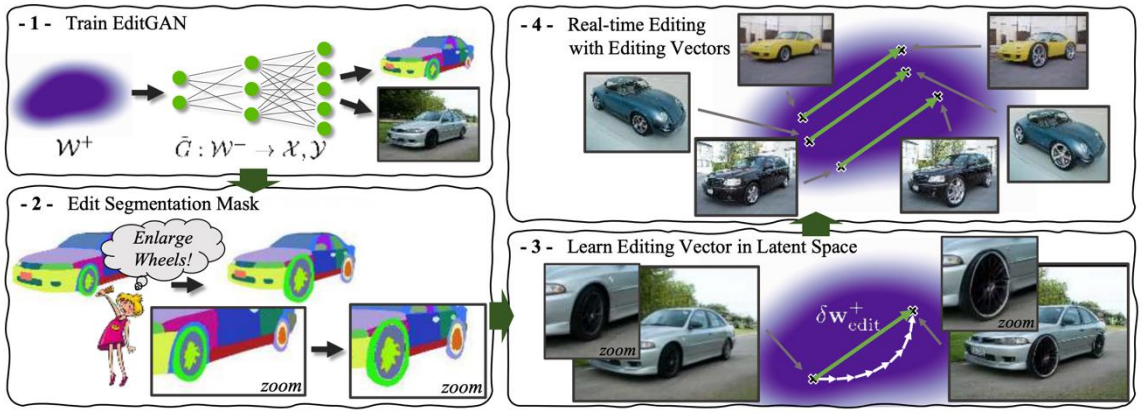


图注：Canvas 的虚拟画布形状和由 AI 生成的对应的图像

来源：https://www.nvidia.com/en-us/studio/canvas/

多伦多大学等提出图像编辑框架 EditGAN

11 月，多伦多大学、向量学院、英伟达等的研究者提出了基于生成对抗网络的图像编辑框架 EditGAN，只需要很少的标注数据，就可以对图像进行高精度的编辑，并能够实时交互运行，直接合成多种图像编辑效果。EditGAN 基于 GAN 框架，对图像和语义分割表示进行联合建模。用户可以对分割遮罩进行编辑，而 GAN 可以在隐空间中学习到改变的图像向量，从而改变图像中的某个部分，实现实时的图像编辑。研究者表示，将很快开放相关代码和工具。

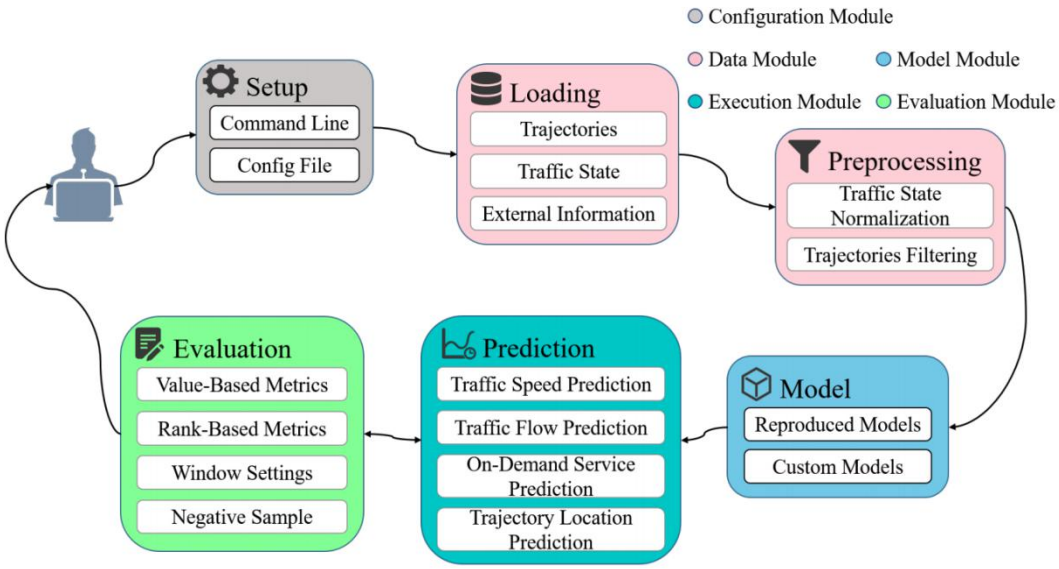


图注：EditGAN 进行图像编辑的流程和原理

来源：https://arxiv.org/pdf/2111.03186.pdf

北航等研究者开源城市时空预测算法库 LibCity

城市计算方面，11 月，北京航空航天大学、人民大学的研究者提出开源的城市时空预测算法库 LibCity，能够为研究者提供可信的城市计算模型对比工具，并推动相关研究的发展。LibCity 整合了 30 多个时空数据集和 50 多个时空算法模型，涵盖交通状态预测、路网匹配、轨迹位置预测等 9 种城市场景任务，用户可以进行算法的评估、开发等。目前 LibCity 已开源，该研究被 ACM SIGSPATIAL 2021 会议收录。



图注：LibCity 的总体组成

来源：https://dl.acm.org/doi/10.1145/3474717.3483923

智源发布生物智能开源开放平台

认知神经研究方面，11 月，智源研究院发布生物智能开源开放平台。其是一个多尺度、多精度、多模态、多认知任务和多模型的智能科学研究基础设施，其中包括：包含超过 30 种认知范式及超过 1000 人次的人类行为范式数据库 CogNet；多尺度，多精度，多模态、多认知任务，涵盖从斑马鱼到人脑结构与功能连接组的开放生物基础数据库 BioDB；面向计算神经科学和类脑计算的自研开源编程工具 BrainPy；深度神经网络和脑影像交叉双脑融合工具 DNNBrain，以及一系列涵盖感知与认知功能的类脑视觉信息处理模型与算法库。

图注：智源生物智能开源开放平台

来源：https://brain.baai.ac.cn/index.html

目前，该平台已推出首个版本，计划 2022 年中进行二次迭代，后续将不断完善，持续推动交叉领域学术资源的开源开放与领域科学家和研究者们共同建设支撑新一代脑启发的人工智能发展的数据、模型、算法、工具软件等开源基础设施，引领我国在该交叉学科领域的前沿探索和创新发展。

算力平台

AI 算力成为超算性能比拼的新“擂台”

近年来，人工智能对算力的需求迅猛增长，已成为当前主要的的算力资源需求来源。AI 计算是智能时代发展的核心动力，以人工智能算力为主的人工智能计算中心应运而生。过去，超算性能的比拼主要是在 TOP500 榜单上进行的，主要是用超算求解数学问题，对超算的计算精度要求高（32 位精度）。但人工智能计算并不要求较高的精度，因此出现了专门面向 AI 超算的榜单。很多超算开始在这些榜单上发力，力求争夺榜首。

富岳超算正式投入使用

3 月，日本理化学研究所（RIKEN）正式启用 AI 超级计算机“富岳”，其是全球首台搭载 ARM 架构处理器高性能计算集群，在 2021 年 6 月 TOP500 榜单上测试性能达 442 PetaFLOPS，HPL-AI 榜单上算力达 2000 PetaFLOPS（2 ExaFLOPS），位列全球第一。目前，富岳已敲定在 100 多项基础和应用研究项目中使用，如医学、药理学、材料学等，研究者将探索利用 AI 技术进行新材料研发、药物研发、医疗诊断算法研究等。

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	706,304	64,590.0	89,794.5	2,528

图注：2021 年 6 月全球 TOP500 超算榜单 Top5 情况

来源：https://top500.org/lists/top500/2021/06/

美国国家能源研究科学计算中心发布 AI 超算帕尔穆特

5 月，美国国家能源研究科学计算中心（NERSC）发布新一代 AI 超算帕尔穆特（Perlmutter），该超算采用 6159 个英伟达 A100 GPU，目标是在混合精度下算力达到 4000 PetaFLOPS（4 ExaFLOPS）。目前帕尔穆特的阶段一（Phase 1）建设已完成，在 2021 年 6 月的全球 TOP500 榜单中测试性能达 64.6 PetaFLOPS，位列世界第五，HPL-AI 榜单中的测试性能达 590 PetaFLOPS，位列第四。目前，该超算已确定了 20 余个科研应用，包括天体物理学、气候科学、材料学等。

Rank	Site	Computer	Cores	HPL-AI (Eflop/s)	TOP500 Rank	HPL Rmax (Eflop/s)	Speedup
1	RIKEN, Japan	Fugaku	7,630,848	2.0	1	0.442	4.5
2	ORNL, USA	Summit	2,414,592	1.15	2	0.149	7.7
3	NVIDIA, USA	Selene	555,520	0.63	6	0.063	9.9
4	LBNL, USA	Perlmutter	761,856	0.59	5	0.065	9.1
5	FZJ, Germany	JUWELS BM	449,280	0.47	8	0.044	10

图注：HPL-AI 2021 年 6 月榜单 Top5

来源：https://hpl-ai.org/doc/results

特斯拉宣布建设人工智能超算平台 Dojo

6 月，特斯拉宣布建设人工智能超算平台 Dojo，采用 5760 个英伟达 A100 显卡，组成 720 个节点，在混合精度下算力突破 1800 PetaFLOPS（1.8 ExaFLOPS），拥有 10PB 的存储空间。特斯拉研发 Dojo 的目的是支持自动驾驶神经网络的训练，使其能够达到足够高的可靠性。

德国 JUWELS Booster 超算位列 Top500 榜单第八

6 月，德国于利希超算中心、冰岛大学、杜伊斯堡-埃森大学研究者公布了超算 JUWELS Booster 的相关情况。该超算搭载了 3744 块英伟达 A100 GPU 和 HDR200 InfiniBand 通信系统，在今年 11 月的 Top500 榜单上位列第八，算力达到 44.1 HPL PetaFLOPS。该超算是今年刚上榜的 AI 超算之一，也是目前 Top500 榜单中欧洲性能最强的 AI 超算系统。

基准测试和数据集

面向复杂语言理解任务的基准测试涌现

基准（Benchmark）在机器学习中，指的是用来比较 AI 算法或模型性能的基准点，其通常包含一个或多个数据集、一个或多个评价指标，以及计算性能的方法。设定基准的意义在于，针对领域内各种各样的 AI 系统，基准能够让研究者使用共同的标准来评价，指导研究者针对模型存在的问题，进行针对性的改进；对于开发者或非专业人士来说，基准能给他们提供一个相对客观的比较方法，让他们快速了解这个领域的进展，识别出有用的模型。

智源研究院发布“智源指数”基准

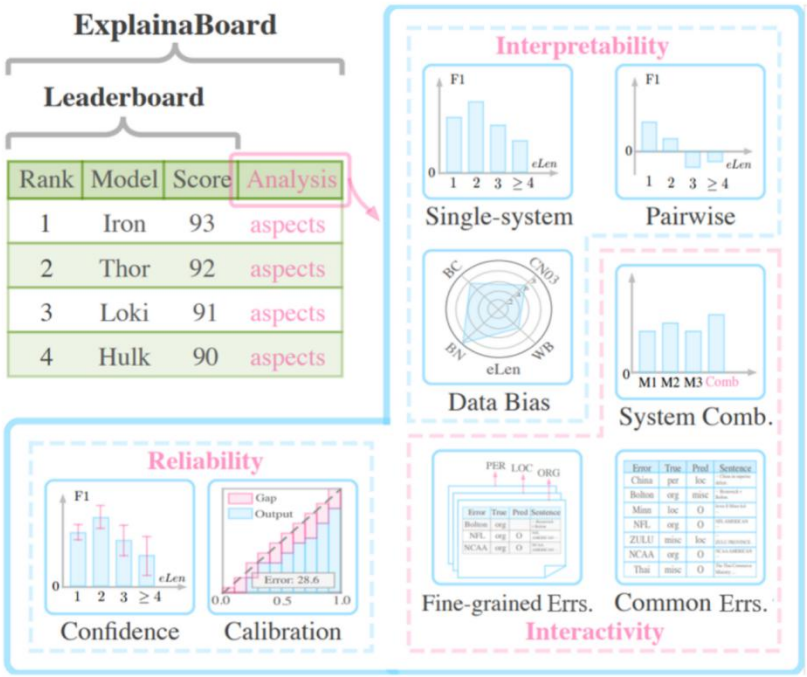
12 月，智源研究院发布“智源指数 GUGE”基准，包含高质量中文自然语言处理数据集、排行榜与在线评测平台，旨在构建全面系统的中文机器语言能力评测体系，形成多层次维度的评测方案。智源指数按照“能力-任务-数据集”的层次结构筛选和组织高质量的代表性数据集，涵盖词句级和篇章级语言理解、信息获取及问答、语言生成、对话交互、多语言、数学等共 17 个领域。

图注：智源指数涵盖的测评能力和数据集

来源：http://cuge.baai.ac.cn/#/task

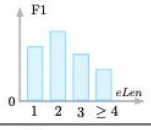
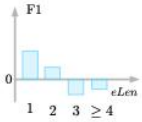
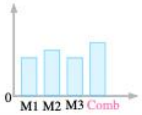
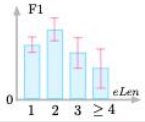
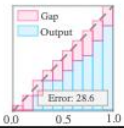
CMU、复旦等研究者提出模型可解释分析和评价排行分析辅助工具 ExplainaBoard

自然语言处理快速发展的同时，基准测试作为跟踪模型在多种任务上性能的工具，发挥了重要作用。但是当前的基准对于提交测试的模型往往采用单一维度的分析角度，采用准确性等较为单一的指标来衡量模型的性能。7 月，CMU、复旦、俄亥俄州立大学的研究者提出了名为“ExplainaBoard”的基准测试排行榜，在标准的基准测试基础上增加了以下功能：一是对于单一系统优劣势进行分析，如“位列榜首的系统在哪些方面存在不足？”；二是解释多个 AI 系统之间的关系，如“A 系统比 B 系统在哪些层面具有性能优势，将 A、B、C 系统结合后会怎么样？”；三是对预测结果进行进一步分析，如“多种系统出现的普遍错误是什么？”。目前，ExplainaBoard 已经涵盖 400 个 AI 系统、50 个数据集、40 种语言，以及 12 个任务。



图注：ExplainaBoard 的基本概念

来源：https://arxiv.org/pdf/2104.06387.pdf

Aspect	Functionality	Input	Output	
Interpretability	Single-system Analysis	One model		**Performance Histogram**: the input model is good at dealing with short entities, while achieving lower performance on long entities.
	Pairwise Analysis	Two models (M1,M2)		**Performance Gap Histogram** (M1−M2): M1 is better at dealing with short entities, while M2 is better at dealing with long entities.
	Data Bias Analysis	Multi-dataset		**Data Bias Chart**: For the `entity length` attribute, the average entity length (We average the length of all test entities on a given data set.) of these datasets order by descending is BN> BC> CN03> WB.
Interactivity	Fine-grained Error Analysis	Single- or Pairwise-system diagnostic results		**Error Table**: Error analysis allows the user to print out the entities that are incorrectly predicted by the given model, as well as the true label of the entity, the mispredicted label, and the sentence where the entity is located.
	System Combination	Multi-models (M1,M2,M3)		**Ensemble Chart**: The combined result of model M1, M2, and M3 is shown by the histogram with x-label value `comb`. The combined result is better than the single models.
Reliability	Confidence	One model		**Error Bars**: the error bars represent 95% confidence intervals of the performance on the specific bucket.
	Calibration	One model		**Reliability Diagram**: Confidence histograms (red) and reliability diagrams (blue). that indicate the accuracy of model probability estimates

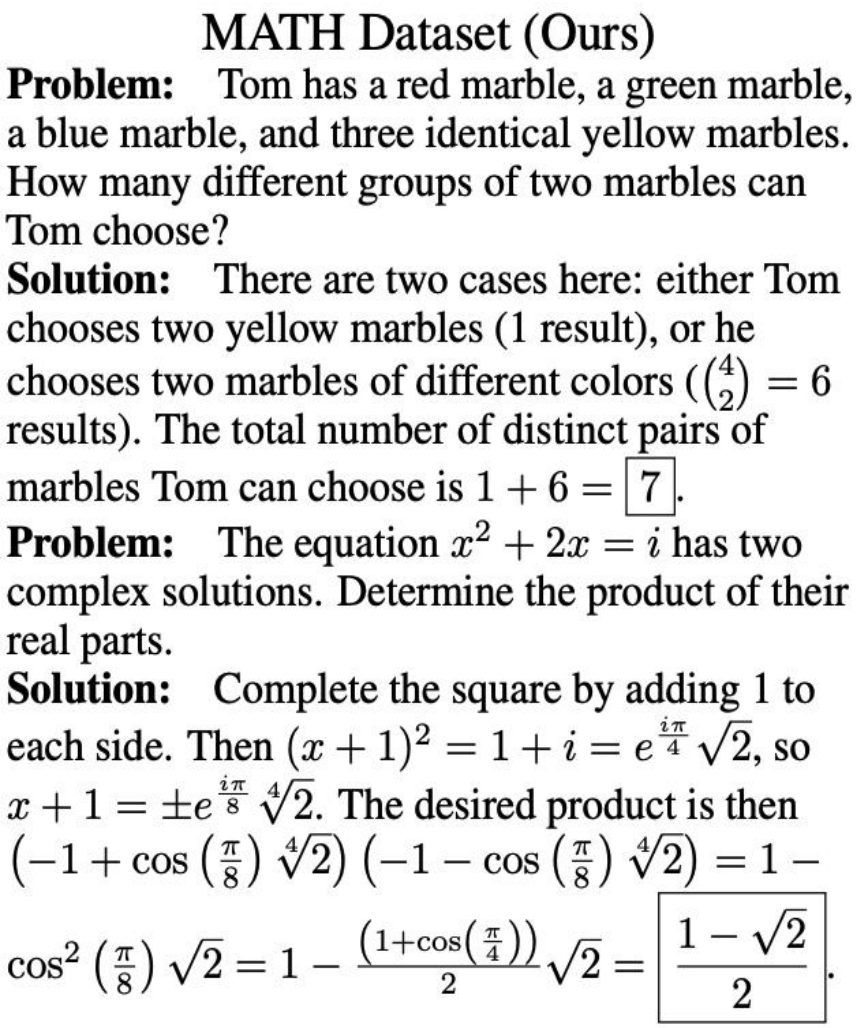
图注：ExplainaBoard 的主要功能，以命名实体识别（NER）任务为例

来源：https://arxiv.org/pdf/2104.06387.pdf

加州大学伯克利分校等研究者提出数学竞赛求解数据集 MATH

11 月，加州大学伯克利分校和芝加哥大学的研究者提出数学竞赛求解数据集 MATH，包含 12500 道数学竞赛难题，每个题目都有完整的逐步求解过程，可用来教机器学习模型生成答案和解释。此外，研究者还创建了大型辅助预训练数据集 AMPS（Auxiliary Mathematics Problems and Solutions），其中包括 10 万个来自可汗学院数学问题的解题方法，以及约

500 万由 Mathematica 脚本生成的问题。论文作者发现，包括 GPT-3 在内的大型语言模型在 MATH 数据集中只能完成 2.9%-6.9%的问题。



MATH Dataset (Ours)

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $\left(-1 + \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\frac{\pi}{8}\right)\sqrt[4]{2}\right) = 1 - \cos^2\left(\frac{\pi}{8}\right)\sqrt{2} = 1 - \frac{\left(1+\cos\left(\frac{\pi}{4}\right)\right)}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

图注：MATH 数据集的内容

来源：https://arxiv.org/pdf/2103.03874.pdf

MIT 等研究者提出 Python Programming Puzzles 编程题目数据集

11 月，MIT、艾伦人工智能研究院和微软的研究人员提出的一种全新的编程题目描述方法。将变成题目用一个 Python 函数定义，目标是找到一个输入 x，使函数输出 True（真值）。基于这种出题形式，研究者提出了名为 Python Programming Puzzles (P3)的编程题目数据集，包含全面、多种难度、多领域和多种算法的编程题目，共 208 种题型，题目数量超过 14 万。

```
# Find a string that when reversed and concatenated with "world" gives "Hello world"
def f1(y: str):
    return y[::-1] + "world" == "Hello world"

# Tower of Hanoi, often teaches recursion. Move [i, j] means move top disk on tower i to j, with 1 ≤ i,j ≤ 3
def f2(moves: List[List[int]], num_disks=8):
    state = [1] * num_disks  # All disks start at tower 1.
    for [i, j] in moves:
        assert state.index(i) <= (state + [1, 2, 3]).index(j), "bigger disk on top"
        state[state.index(i)] = j  # Move smallest disk from tower i to tower j.
    return state == [3] * num_disks  # All disks must end on tower 3.

# Find a non-trivial integer factor d of a large number n
def f3(d: int, n=100433627766186892221372630609062766858404681029709092356097):
    return 1 < d < n and n % d == 0
```

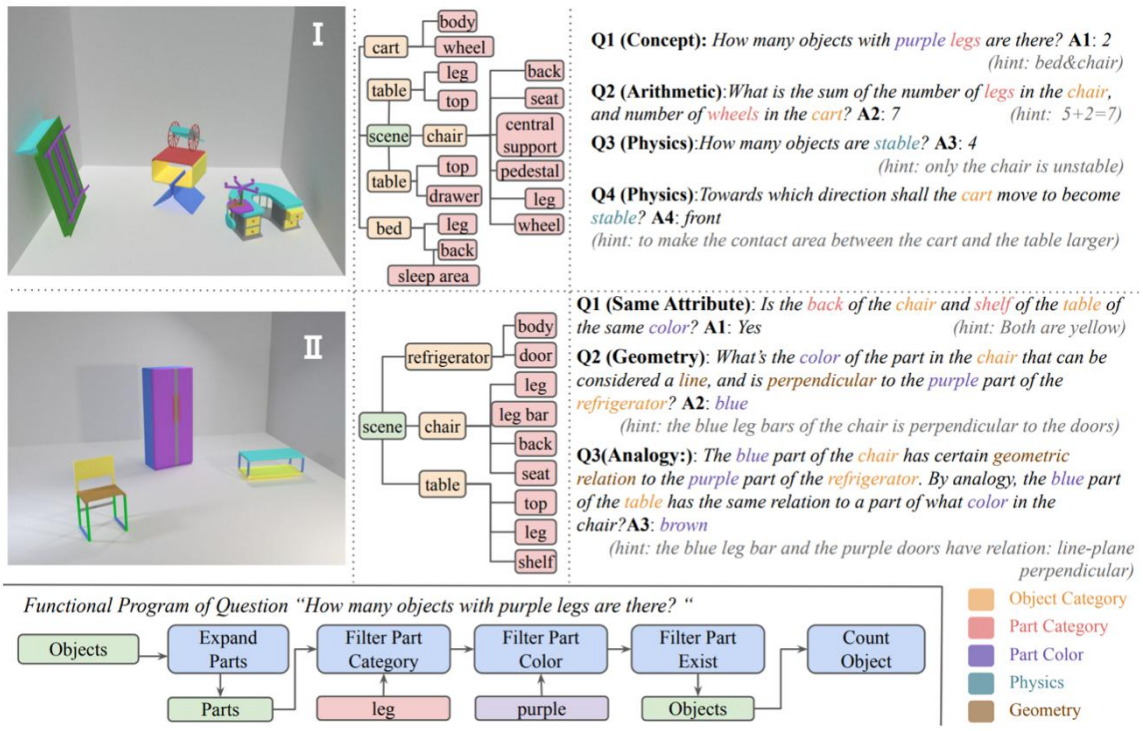
图注：将语言描述的编程题目转换为函数的过程

来源：https://arxiv.org/pdf/2106.05784.pdf

UCLA、斯坦福、MIT 等研究者提出新一代视觉推理数据集

12 月，加州大学洛杉矶分校、斯坦福大学、MIT 等机构的研究者提出一种基于十句局部的概念、关系和物理推理基准，名为“PTR”。PTR 提出了五种人类认知推理问题类型，分别为概念、关系、类比、数学和物理推理，包含 7 万张 RGBD 合成图像，其中包括 Ground Truth 目标、局部级别的标注（如语义示例分割、颜色数学、空间和集合关系、特定的物理属性等）。

通过对一些当前视觉推理领域最佳性能的模型进行测试，研究者发现在很多人类可以轻松完成的任务上，这些模型依然会出现错误。



图注：PTR 基准中的数据类型和将图像处理为推理任务的过程

来源：https://arxiv.org/pdf/2112.05136.pdf

HuggingFace 提出自然语言处理通用数据集

9 月，HuggingFace 开放大规模开源 NLP 数据集和资源库，其中包括 650 个数据集，涵盖自然语言处理的多种任务和基准。该项目采用分布式、社区驱动的方法来进行构建，贡献者超过 250 人。

```python
from datasets import list_datasets, load_dataset, list_metrics, load_metric

# Print all the available datasets
print(list_datasets())

# Load a dataset and print the first example in the training set
squad_dataset = load_dataset('squad')
print(squad_dataset['train'][0])

# List all the available metrics
print(list_metrics())

# Load a metric
squad_metric = load_metric('squad')

# Process the dataset - add a column with the length of the context texts
dataset_with_length = squad_dataset.map(lambda x: {"length": len(x["context"])})

# Process the dataset - tokenize the context texts (using a tokenizer from the 🤗 Transformers libra
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained('bert-base-cased')

tokenized_dataset = squad_dataset.map(lambda x: tokenizer(x['context']), batched=True)
```

图注：调用数据集的示例代码

来源：https://github.com/huggingface/datasets

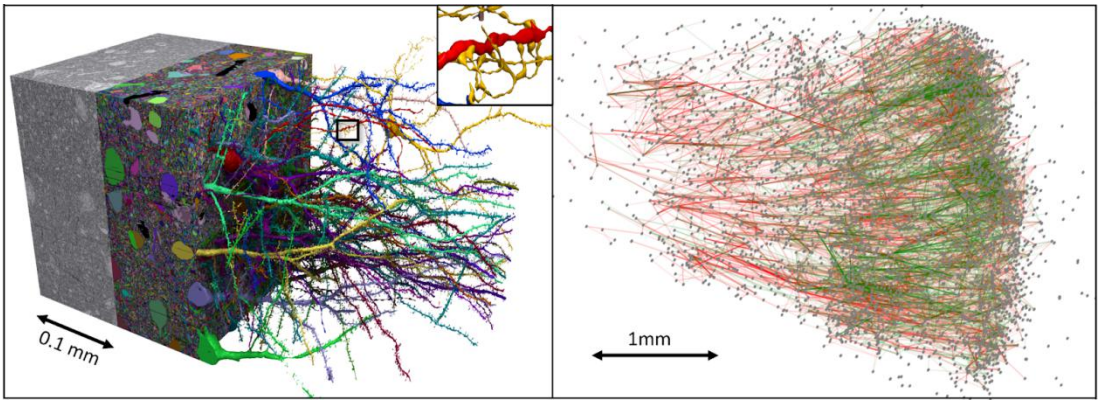
AI 为人类科学家提供领域数据集，助力基础科学研究

人工智能与传统科研领域结合，能够全面提升科研效率，助推技术和产业升级。在人工智能的辅助下，可以在暂不知晓原理的情况下，通过 AI 强大的拟合能力，解决一些建模、模拟方面的问题，由科学家逐步明晰背后的原理，并在 AI 完成的工作基础上进行深入的研究和应用开发。

谷歌、哈佛等发布人类脑组织“H01”数据集

6 月，谷歌与哈佛大学研究者合作，发布了人类脑组织“H01”数据集，通过连续切片电子显微镜以 4nm 分辨率成像，并由计算机进行重建和自动注释，形成了覆盖大约一立方毫米的脑

皮质组织“地图”，其中包括带有数万个神经元、1.3 亿个带注释突触、104 个校对细胞以及许多其他亚细胞注释和结构。H01 是迄今为止对大脑皮层进行该种精细程度成像和重建最大的样本，也是第一个大规模研究人类大脑皮层的“突触连接性”的样本，跨越了大脑皮层中所有层面的多种细胞类型。

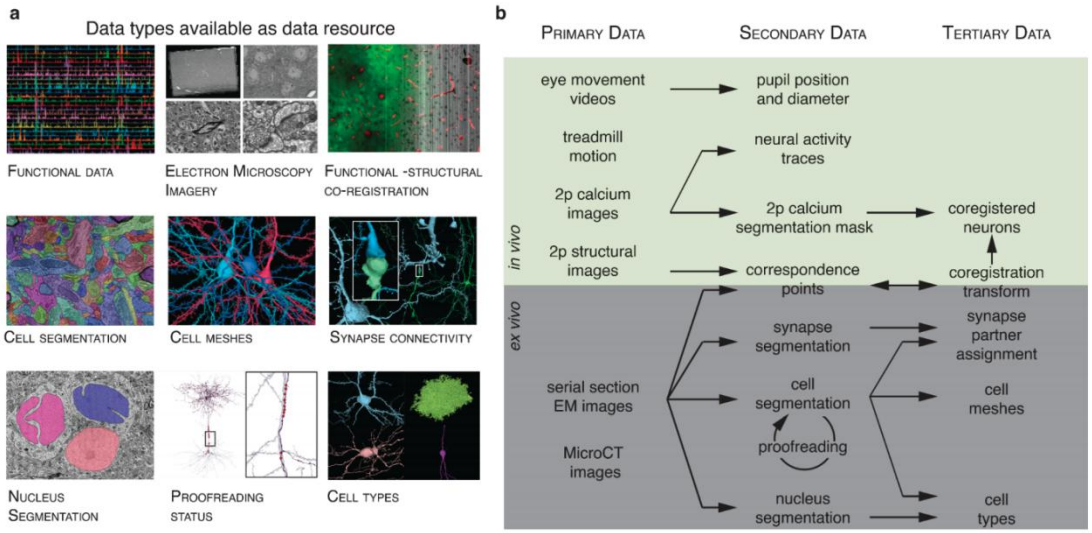


图：人类新皮质连接体重建的示意图。左：神经连接 3D 结构。右：数据集中 5000 个神经元中，兴奋性（绿色）和抑制性（红色）神经元连接的子图。

来源：https://ai.googleblog.com/2021/06/a-browsable-petascale-reconstruction-of.html?m=1

美国大脑皮层网络的机器智能计划发布哺乳动物大脑回路数据连接图

大脑皮层网络的机器智能计划（MICrONS） 属于美国情报系统下设的高级智慧研究计划（Intelligence Advanced Research Projects Activity，IARPA），于 2016 年启动，旨在通过设计一立方毫米的大脑结构，研究其计算方式，并应用于机器学习和人工智能研究中。因其目标远大，MICrONS 也被称作“阿波罗大脑计划”。7 月，MICrONS 发布了一个数据集，建立了哺乳动物大脑回路最详细的数据连接图，包含当前数量最高的神经细胞和连接数，有 20 万个脑细胞和近 5 亿个突触的精细结构和连通特征，数据集现已公开。



图注：该计划开放的 9 种数据类型及每种数据类型之间的关系

来源：https://www.biorxiv.org/content/10.1101/2021.07.28.454025v2.full.pdf

第二章

人工智能产业发展情况

人工智能应用层企业——智能医疗

总体趋势

2021 年，AI 医疗赛道持续火热，各大医疗 AI 企业纷纷冲刺 IPO。“烧钱”已然成为今年这一赛道最鲜明的标签。

2021 年初，医渡科技在港交所上市；3 月，科亚医疗港交所递交 IPO 申请；8 月，推想医疗港交所递交 IPO 申请；9 月，数坤科技港交所递表；11 月，鹰瞳科技成为“医疗 AI 第一股”。从医疗 AI 细分领域来看，2021 年，AI 辅助医疗影像领域逐渐趋于饱和，深睿、推想、数坤等几家头部公司已经站稳了各个细分市场；而随着 5G 通信的铺开，以前受时延限制的 AI+手术机器人开始走上潮头，DeepMind AlphaFold2 的催化加上超大规模算力的发展，让 AI+药物研发也成为炙手可热的方向，资本大力加码生物计算相关的平台公司致力精准药物研发；去年，马斯克旗下 Neuralink 占尽风头，国内脑机接口公司也踏上了资本化发展的快车道，AI+脑科学也开始成为医疗、科研等场景的新热点。

医疗影像

AI 医学影像，即 AI 辅助传统的人工阅片，辅助医生进行相关疾病的临床诊断和早期筛查。医学影像数据几乎占据临床数据 90%，是临床诊断和疾病治疗的基石。根据亿欧智库预测，中国医学 AI 影像市场将从 2020 年的 3.12 亿元（人民币），在 2023 年增至 20.60 亿元（人民币），年复合增长率高达 88.85%，2030 年将达到千亿。2021 年以来，疫情后的医院智能化建设加速。

政策方面，国家开始逐步发放各类医疗影像 AI 软件三类证，都为医疗影像的发展提供了契机。

医疗影像领军企业纷纷开启 IPO 申请，产品陆续获批上市。共有 4 家领军企业向上市发起冲刺。3 月，科亚医疗正式向港交所提交招股书，拟在香港主板挂牌上市。6 月，鹰瞳科技向港交所主板提交上市申请（现已成功上市）。8 月，推想医疗向港交所主板提交上市申请。9 月，数坤科技正式向港交所提交招股书，拟在香港主板挂牌上市。

三类证审批方面，推想科技、数坤科技、科亚方舟、深睿医疗等多家医疗影像企业在三类证审批、融资并购等方面动态频繁。整个行业商业化的繁荣态势凸显。

产品方面，国内目前已获批上市的 AI 医学影像产品超过 15 款。其中包括鹰瞳科技的糖网眼底图像辅助诊断软件，数坤科技的 CT 造影图像血管狭窄辅助分诊软件，科亚医疗的冠脉血流

储备分数计算软件，推想科技的肺结节 CT 影像辅助检测软件等。今年在医疗影像赛道值得关注的企业包括鹰瞳科技、推想科技、数坤科技、深睿医疗、睿心医疗等，以及国内头部互联网企业。

鹰瞳科技成立于 2015 年，是中国首批提供人工智能视网膜影像识别的早期检测、辅助诊断及健康风险评估解决方案的公司之一。目前，鹰瞳科技核心产品包括三个版本的 Airdoc-AIFUNDUS。其中 Airdoc-AIFUNDUS（1.0）是一款人工智能医疗器械软件（SaMD），获批用于辅助诊断糖尿病视网膜病变，以协助医生做医疗诊断，在同类产品中首个获得国家药监局第三类医疗器械证书。今年 11 月，鹰瞳科技成功上市，成为国内医疗 AI 第一股。招股书显示，鹰瞳科技已将 Airdoc-AIFUNDUS（1.0）提供给中国 23 家医院及 3 家社区诊所。

推想科技成立于 2016 年。推想医疗的脑卒中产品“InferRead CT Stroke. AI”已经通过 FDA 认证。获批后，推想科技成为中国第一家同时拥有两张 FDA 认证的 AI 医疗公司。截至目前，推想科技在肺部疾病方面同时拥有美国 FDA、欧盟 CE、日本 PMDA 及中国 NMPA 全球四大市场准入认证。今年 7 月，推想医疗完成 D2 轮融资，D 轮总融资额达人民币 9 亿元。

数坤科技成立于 2017 年。数坤数字医生平台目前包括 12 款产品及 25 款候选产品，主打包括数字心、脑、胸、骨、腹等“数字人体”产品矩阵。其中，心脑血管 AI 的准入门槛比较高。2020 年 11 月，数坤科技研发的“冠脉 CT 造影图像血管狭窄辅助分诊软件”（简称“数坤心血管 AI”）正式获批中国国家药品监督管理局医疗器械三类证，这是心脏冠脉狭窄人工智能辅助诊断领域的全球首张医疗器械注册证。今年 8 月，数坤科技完成人民币 7 亿元融资，9 月向港交所递交招股书。

深睿医疗成立于 2017 年。目前，公司已经获得肺结节和肺炎 AI 产品两张 NMPA 三类证，是为数不多拥有两张三类证的医疗 AI 公司，并于今年 8 月收购了同为头部企业的依图医疗。这是国内 AI 医学影像赛道上迄今为止最大规模的并购事件。今年 5 月，深睿医疗完成 C3 轮融资。

睿心医疗成立于 2017 年，是另一家拥有两项三类证的医疗影像公司。今年 4 月，睿心医疗第一款成熟产品睿心分数（RuiXin-FFR）获批：国家药监局（NMPA）创新医疗器械三类证，5 个月后，用于冠脉 CT 影像智能分析的——睿心冠脉智能后处理平台也收获了 NMPA 认证。该公司于今年 1 月完成人民币 3 亿元 B 轮融资，于今年 8 月完成数亿元人民币 C 轮融资。

同时，国内互联网企业对医疗影像领域也颇多关注。9 月 10 日，腾讯发布消息称，腾讯医疗健康（深圳）有限公司的“肺炎 CT 影像辅助分诊及评估软件”获得国家药品监督管理局批准的三类医疗器械注册证，腾讯也正式成为国内互联网科技行业首个获得医疗人工智能三类证的企业。

此外，今年获得融资以及获批三类证的医疗影像企业还包括：

1.上海点内科技完成人民币数千万元 A 轮融资，专注 AI 肺癌全病程解决方案

2.北京致远慧图完成人民币近亿元 B 轮融资

3.北京汇医慧影获批国内首张 X 射线骨折 AI 产品 NMPA 三类证

4.上海联影智能获批肺结节 CT 影像辅助检测软件三类证

厂商	获批时间	获批产品
科亚医疗	2020.1	冠脉血流储备分数计算软件
鹰瞳科技	2020.8	糖尿病视网膜病变眼底图像辅助诊断软件
数坤科技	2020.11	CT造影图像血管狭窄辅助分诊软件
推想科技	2020.11	肺结节CT影像辅助检测软件
	2021.3	肺炎CT影像辅助分诊与评估软件
汇医慧影	2021.4	骨折X射线图像辅助检测软件
联影智能	2021.6	肺结节CT影像辅助检测软件
深睿医疗	2020.12	肺结节CT影像辅助检测软件
	2021.3	肺炎CT影像辅助分诊与评估软件
睿心医疗	2021.4	无创冠脉血流储备分数计算软件睿心分数
	2021.9	睿心冠脉智能后处理平台
依图科技	2021.3	儿童手部X射线影像骨龄辅助评估软件
联影智能	2021.6	肺结节CT影像辅助检测软件
腾讯觅影	2021.9	肺炎CT影像辅助分诊及评估软件

图注：医疗影像企业获批三类证的时间和产品

来源：智源研究院整理

AI 药物研发

AI 药物研发包括药物发掘、临床前研究、临床试验等阶段，新药研发无异于大浪淘沙，漫长而复杂，且成本高昂。AI 技术可深入参与新药研发从靶点发现到新药上市的各个环节，大幅缩减成本及研发周期，甚至促进生物学研究、发现新的生物靶点机制和开发新的疾病模型。AI 在已有数据积累的药物设计领域，作为效率工具的巨大商业价值在全球生物医药创新实践中已得到验证。

全球人工智能制药市场潜力巨大。从 2019 年至 2024 年，全球人工智能制药市场将以 40.8% 的复合年增长率增长。中信证券今年 7 月研报显示，2021 年上半年，AI 制药是国内数字健康领域增速最快的细分赛道，仅上半年融资额就超过了 10 亿元（人民币），保持高景气度。

资本助力下，新兴 AI 创企、互联网科技巨头和传统药企在 AI 制药领域百花齐放。

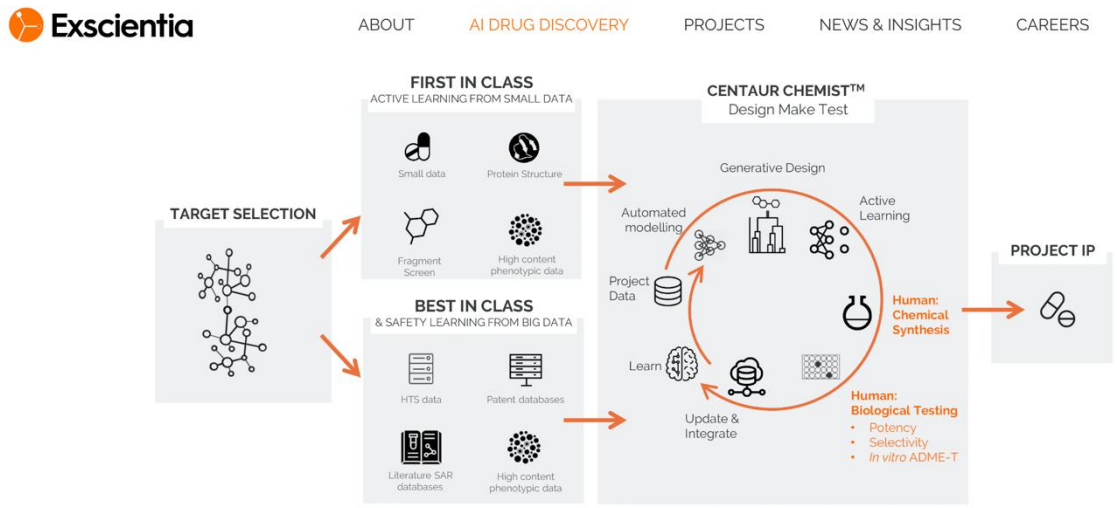
Exscientia、Insilico Medicine（英矽智能）、Insitro 等几家海外领军 AI 制药公司在今年均有大额融资动向，其中 Exscientia、Recursion 已经登陆纳斯达克。

与此同时，晶泰科技等国内创业企业也在 AI 制药赛道上疯狂聚集。4 月，望石智慧（StoneWise）宣布完成了 B+轮和 B 轮融资，融资总额 1 亿美元；7 月，百图生科完成上亿美元 A 轮融资；8 月，晶泰科技完成 4 亿美元 D 轮融资，星药科技完成 B 轮融资。

谷歌母公司 Alphabet、Facebook、华为等科技巨头纷纷入局。今年 4 月，Facebook AI 与合作伙伴共同发布了一种开源 AI 模型，旨在确定将现有药物重新用于新的药物鸡尾酒中的可行性；今年 9 月，华为发布“华为云盘古”药物分子大模型；今年 11 月，谷歌母公司 Alphabet 正式宣布成立新的 AI 制药公司 Isomorphic Laboratories。该公司的创始人正是 DeepMind CEO、AlphaFOLD 发明者 Demis Hassabis。Isomorphic 的使命是“利用人工智能加速药物发现，并最终找到治疗人类一些最具破坏性疾病的方法”。

今年在 AI 药物研发领域值得关注的企业包括 Exscientia、英矽智能、Insitro、Recursion、Schrodinger、晶泰科技、百图生科等。

英国人工智能新药研发企业 Exscientia 成立于 2012 年。该平台的亮点为将新药研发的过程分为“设计-制造-测试-分析”四个步骤，缩短了新药研发周期。该公司在平均一年的时间内，将 7 种精密设计药物从项目启动阶段推进到开发候选阶段。

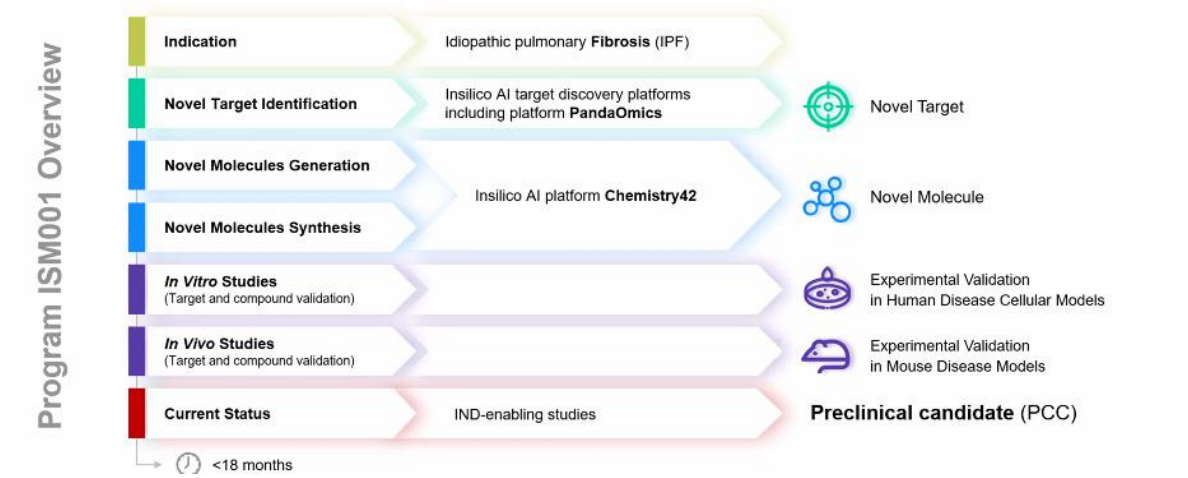


图注：Exscientia 新药研发过程介绍

来源：Exscientia 官网

今年 4 月，Exscientia 宣布首款 AI 设计的肿瘤免疫新药进入人体临床试验，专为成人晚期实体瘤患者开发。迄今为止，该公司已有两款 AI 设计的药物进入临床阶段，包括去年与大日本住友制药研发的治疗强迫症患者的新药。5 月，Escientia 宣布完成 5.25 亿美元 D 轮融资。9 月，Exscientia 成功登陆纳斯达克。

英矽智能成立于 2014 年，早期总部设立于美国马里兰州巴尔的摩，后期设立在中国香港。2021 年 2 月，英矽智能实现了从机制发现、靶点发现及找到新化合物的整个新药研发周期缩短至 18 个月的成果，将药物研发成本降至 260 万美元，远优于传统新药物研发耗时 2-5 年、投入 1098 万美元的平均数据。为了构建 AI+新药研发的生态，围绕靶点发现、生物研发、临床二期试验预测，英矽智能分别开发了 PandaOmics、Chemistry42 和 InClinico 平台，助力新药研发工作。6 月 22 日，英矽智能获 2.55 亿美元 C 轮融资。



图注：英矽智能实现的 AI 药物研发成果

来源：英矽智能

Insitro 由斯坦福著名学者 Daphne Koller 于 2018 年创立。Insitro 在神经科学和肝脏疾病方面进行了深入研究，和吉利德等制药大厂达成了合

作，预计 2022 年搭建首条管线。今年 3 月，Insitro 获得 4 亿美元 C 轮融资，成为全球 AI 制药领域单笔融资金额最高的公司。

美国 AI 制药独角兽 Recursion Pharmaceuticals（Recursion）成立于 2013 年，公司推出的多面集成化系统 Recursion OS，主要包括用于支持硬件和软件的基础层，多样化且可关联的数据库 Recursion Data Universe，以及专用于药物发现、设计和研发的工具 Recursion Map。Recursion 于今年 4 月登陆纳斯达克。

Schrodinger 薛定谔公司于 2020 年 2 月登陆纳斯达克，为行业内第一家 AI 药物研发上市公司。Relay 于 2020 年 7 月登陆纳斯纳克，为第二家 AI 药物研发上市公司。

AI 制药独角兽晶泰科技成立于 2015 年。晶泰科技是典型的从一个环节入手的公司，主要聚焦于药物固态研发环节，包括晶型预测、固态筛选、结构确定等。其核心技术平台为 ID4 (Intelligent Digital Drug Discovery and Development) 智能药物研发平台，它通过结合量子物理、人工智能与云计算技术，能准确预测药物的多种重要特性，除早期的晶型外，现在还包括活性、成药性、毒性等等指标，从而综合加速药物临床前研究的效率与成功率。今年 8 月，晶泰科技完成 4 亿美元 D 轮融资。

百图生科于 2020 年由李彦宏牵头创立，定位为生物计算平台企业。百图生科致力于通过先进的计算和生物技术，从多组学生物数据、高通量验证实验、药物开发经验中高效抽提知识，绘制关于疾病靶点和药物设计的图谱，把药物发现从“大海捞针”变成“按图索骥”，从而提升自身与合作伙伴的药物研发效率，最终想实现 Global First-in-class 原创药物的研发。今年 5 月，百图生科推出“免疫图谱卓越计划”，并于今年 7 月完成上亿美元 A 轮融资。

星药科技成立于 2019 年。目前已基于原研算法，结合药物化学、计算化学和生物学的领域知识、工具及经验建立了药物发现平台 PyxirTM，该平台已经过验证并取得阶段性技术进展。星药科技于今年 8 月完成 B 轮融资。星药科技的创始人兼 CEO 李成涛毕业于清华姚班，获得麻省理工计算机博士学位，是人工智能+药物研发领域的顶尖学者之一，曾入选西贝尔学者 （Siebel Scholar)。

其他值得关注的 AI 制药企业融资事件有：

1. 6 月，宇道生物完成 2000 万美元 A 轮融资，建设国际领先变构药物智能化研发平台和管线
2. 7 月，星亢原 neoX 完成数千万美元 A+轮融资
3. 8 月，智化科技（Chemical.AI）完成近 1500 万美元 A+轮融资

4. 8 月，深势科技完成数千万美元 A 轮融资，持续探索科学计算平台建设

5. 8 月，科因生物获人民币数千万元 A 轮融资

6. 9 月，普瑞基准完成人民币 1.5 亿元 B 轮融资，多组学数据挖掘驱动新药研发

7. 9 月，燧坤智能完成人民币超亿元 A 轮融资

8. 10 月，华深智药完成千万美元级天使轮融资，由清华大学人工智能产业研究院孵化

9. 10 月，硕迪生物完成 1 亿美元 B 轮融资，专注突破 G 蛋白偶联受体靶点家族

10. 其他值得关注的 AI 制药企业融资事件有：

11. 6 月，宇道生物完成 2000 万美元 A 轮融资，建设国际领先变构药物智能化研发平台和管线

12. 7 月，星亢原 neoX 完成数千万美元 A+轮融资

13. 8 月，智化科技（Chemical.AI）完成近 1500 万美元 A+轮融资

14. 8 月，深势科技完成数千万美元 A 轮融资，持续探索科学计算平台建设

15. 8 月，科因生物获人民币数千万元 A 轮融资

16. 9 月，普瑞基准完成人民币 1.5 亿元 B 轮融资，多组学数据挖掘驱动新药研发

17. 9 月，燧坤智能完成人民币超亿元 A 轮融资

18. 10 月，华深智药完成千万美元级天使轮融资，由清华大学人工智能产业研究院孵化

19. 10 月，硕迪生物完成 1 亿美元 B 轮融资，专注突破 G 蛋白偶联受体靶点家族

数字疗法

数字疗法是 2021 最新兴的医疗细分赛道。数字疗法（Digital Therapeutics：DTx）是由软件程序驱动，以循证医学为基础的干预方案，用以预防、治疗或管理疾病。数字疗法可以单独使用，也可以与药物、医疗器械或其他疗法配合使用。通过信息（如 App 上的文字、图片、视频等数据）、物理因子（如声音、光线、电流、磁场及其组合）、以及药物等对患者施加影响，优化患者护理和健康结果。数字疗法产品，在自闭症、睡眠，成瘾类疾病，以及阿尔兹海默这类传统打针吃药和手术疗法等收效甚微的疾病上可能有所突破。

今年，数字疗法在资本市场也受到了广泛关注。

截止今年 7 月，2021 年全球数字疗法的赛道上，共有 86 起融资事件，融资总额超过 30 亿美元。

国际方面，Pear Therapeutics、Akili Interactive Labs 等企业成立较早，技术和产品相对成熟。国内，数字疗法起步较晚，处于刚刚兴起的阶段。2020 年 11 月，我国国家药品监督管理局(NMPA)审批通过了首款数字疗法产品，标志着国内数字疗法新赛道的开启。

2021 年被业界公认为数字疗法产业元年，一批数字疗法企业崭露头角。

今年值得关注的数字疗法企业包括 Pear Therapeutics、Akili Interactive Labs、博斯腾科技、望里科技、尚医科技、妙健康等。

Pear Therapeutics 于 2013 年在美国波士顿成立，是数字疗法最早的先驱之一，其研发的产品 reSET™是全球首个获得 FDA 批准的数字疗法产品。reSET™以治疗课程为核心，帮助患者培养戒断物品成瘾（包括酒精、大麻、可卡因和兴奋剂）的技能。Pear Therapeutics 于今年 3 月完成 2000 万美元 D 轮 II 期融资（去年 12 月 I 期），目前已累计融资 2.84 亿美元。

美国企业 Akili Interactive Labs 成立于 2011 年，该公司主打通过高质量的动作视频游戏体验，治疗认知缺陷，改善神经病和精神病相关症状，

包括 ADHD、抑郁症、自闭症谱系障碍（ASD）和各种炎症性疾病。去年 6 月，由 Akili 开发的 EndeavorRx 获 FDA 批准，成为世界首款“视频游戏处方药”，用于提高 8-12 岁 ADHD 儿童的注意力。Akili 于今年 5 月完成 1.6 亿美元 D 轮融资。

上海博斯腾网络科技有限公司（博斯腾科技）成立于 2016 年。该公司专注阿尔兹海默症的数字化筛查和干预，并形成了一套系统性的数字化干预训练。今年 6 月，博斯腾完成 A 轮近亿元人民币融资。

专注精神疾病的望里科技成立于 2017 年。目前，望里科技已经与国内多家精神疾病泰斗级医疗机构进行深度合作，核心产品已推进入院及临床试验。今年 8 月 30 日，望里科技宣布完成千万级美元 A+轮融资。

尚医科技成立于 2014 年，该公司开发出国内首款“数字疗法”术康 APP，已获得国家药品监督管理局审批，并已成功进入美国市场。今年 9 月，尚医科技完成人民币近亿元 C 轮融资。

妙健康成立于 2015 年，是专注于个人健康行为管理的综合性平台，于今年 5 月正式通过国际数字疗法联盟（Digital Therapeutics Alliance：

DTA）认证，并与 DTA 达成长期战略合作，成为首家国际数字疗法联盟认证会员的中国企业。妙健康于今年 7 月完成人民币数亿元融资。

手术机器人

医疗机器人发展历史悠久，最早在 1985 年，美国就已经开始尝试使用 Puma560 工业机器人来辅助进行脑组织活检手术。按应用场景划分，医疗机器人可分为手术机器人、康复机器人、服务机器人、辅助机器人 4 类。其中，手术机器人占到 37%左右。

手术机器人可分为腹腔镜手术机器人、骨科手术机器人、神经外科手术机器人、血管介入手术机器人等不同类型。其中，腔镜机器人的技术成熟度相对较高，商业化也最成功，市场占比 60%，是行业研发重点，代表公司有美国直觉外科公司，旗下产品“达芬奇”手术机器人一骑绝尘，占据全球一半的手术机器人市场。该产品可以应用于心脏、前列腺、胸腺等部位的软组织手术，其他国外玩家还包括美国 Asensus Surgical、英国 CMR Surgical、韩国 Meerecompany（Revo-I）等。此外，骨科机器人占比 15%，主要应用于创伤骨科、脊柱外科和关节外科，代表企业有美敦力（Medtronic）、捷迈邦美（Zimmer Biomet）、史赛克、施乐辉等。此外，医疗器械巨头强生等也在骨科机器人领域积极布局。神经外科和血管介入机器人方面，技术和产品在不断发展，而其他手术领域的机器人技术则比较空白。今年以来，手术机器人领域的融资金额近 14 亿美元，A 轮及 A 轮之前的早期融资项目占比 60%，

不乏 6 亿美元（CMR Surgical）、1 亿欧元（eCential Robotics）等大额融资。据艾媒咨询数据，预计 2025 年全球手术机器人市场规模将超 2 万亿美元。

骨科手术机器人产业日渐成熟，龙头企业产品纷纷获批上市。

今年 1 月，强生 DePuy Synthes 宣布，旗下的骨科手术机器人产品 VELYS 获得 FDA 批准上市；3 月，史赛克公司的骨科手术机器人 Mako 也在国内获批全膝关节置换适应症；10 月，美敦力宣布其 Hugo RAS 机器人辅助手术系统获欧盟 CE 认证。同时，天智航、微创、键嘉、元化智能等国产骨科手术机器人企业目前均有核心产品投入临床应用，或获批上市。

同时，国内政策利好手术机器人，为打开市场提供助力。今年 10 月 23 日，机器人辅助骨科手术将被纳入北京甲类医保支付目录，患者可获 100%全额报销。与此同时，上海已经将达芬奇机器人的手术和耗材纳入了医保支付范围。虽然手术机器人医疗费用高昂，但

医保的推进可为手术机器人打开市场。全民可用的时代或可指日可待。

美国直觉外科公司（Intuitive Surgical）是全球领先的手术机器人龙头，1995 年成立于美国，2000 年于纳斯达克上市。旗下产品达芬奇手术机器人 2000 年获批上市，2008 年进入中国市场，目前占据全球一半的手术机器人市场。因其垄断地位，每台售价高达 3000 万元人民币。

英国公司 CMR Surgical 2014 年成立于英国剑桥，核心产品 Versius 于 2017 年研发成功，专为微创手术设计，是全球最小的手术机器人，仅有达芬奇机器人的三分之一大小。目前该款机器人已经打入欧洲、澳大利亚、印度和中东等地（尚未在美国获得批准），迄今为止参与手术超 1000 起。今年 6 月，英国公司 CMR Surgical 获 4.25 亿英镑（人民币近 40 亿元）融资，创下全球医疗技术领域规模最大的单笔融资记录。

强生 DePuy Synthes 由强生原骨科业务线与 Synthes 于 2011 年合并而成，提供在关节重建、创伤、脊柱、运动医学等领域的服务方案。其营收领跑全球骨科领域，今年 1 月，旗下骨科手术机器人产品 VELYS™获得 FDA 批准上市。其他美国骨科领域龙头企业还包括史赛克（成立于 1941 年）和美敦力（成立于 1949 年，全球 500 强）。

上海微创医疗机器人 2014 年诞生自微创医疗集团内部的一个孵化项目。目前，微创机器人三款旗舰产品均已纳入国家药监局创新医疗器械绿色通道，包括图迈腔镜机器人、蜻蜓眼三维电子腹腔内窥镜和鸿鹄骨科手术机器人。11 月，上海微创医疗机器人在港挂牌上市，为港交所上市的第一家手术机器人企业。

山东威高始建于 1988 年，该公司旗下的“妙手”手术机器人也属于腹腔镜手术系统，对标“达芬奇”。目前已经投入临床试验阶段，已在中南大学湘雅三医院、青岛医科大学附属医院两医院完成 168 例临床实验。今年 10 月，山东威高宣布完成人民币十数亿的 A 轮融资。

天智航成立于 2005 年，是国内首家取得医疗机器人注册许可证的企业。目前天玑骨科手术机器人已研发至第三代，对标美敦力 Mazor 机器人，并已在国内 100 余家医疗机构进行了常规临床应用，累计完成手术超过 2 万例。天智航已经于 2020 年 7 月上市。

柏惠康维成立于 2010 年，专注神经外科+口腔手术机器人。在已经获批上市的“睿米”神经外科手术机器人以外，柏惠康维也于今年正式发布了 NMPA 认证口腔手术机器人“瑞医博”，探索了手术机器人应用的新场景。

其他今年获得融资的国产手术机器人企业包括：

1. 1 月，罗森博特完成人民币数千万元 A 轮融资

2. 2 月，维卓致远完成人民币亿元 Pre-A 轮融资

3. 3 月，元化智能完成人民币 2 亿元 A 轮融资

4. 4 月，三坛医疗完成人民币数千万元 B 轮融资

5. 5 月，键嘉机器人完成人民币数亿元 C 轮融资

6. 7 月，医达健康向港交所递交招股书

7. 8 月，柳叶刀机器人完成人民币数千万元 Pre-A 轮融资

8. 8 月，梅奥心磁完成 Pre-A 轮融资

9. 11 月，长木谷完成人民币 5.4 亿元 B 轮融资

脑机接口

脑机接口作为 2020 年讨论度最高的医学概念之一，随着去年 8 月马斯克旗下 Neuralink 发布的概念产品（侵入式脑机接口）达到了高峰。马斯克的概念产品属于脑机接口中的侵入式技术，即通过有创方式将电极插入大脑皮层采集信号。另一种非侵入式，则通过脑电帽的形

式，通过导电胶等接触头皮采集脑电信号。2021 年，脑机接口行业得到了巨大的发展，大笔融资随之跟进。3 月，脑机接口平台公司 NeuraMatrix、博瑞康完成新一轮融资；7 月，Neuralink 获脑机接口领域史上最大融资；8 月，优脑银河、柔灵科技等国内脑机接口公司完成新一轮融资。此外，从今年的趋势来看，

脑机接口不再只是“意念打字”这样的融资噱头，而逐渐开始从实验室走向了临床实践，从科幻照进了现实。

脑机接口的应用场景在医学上可用于肢体运动障碍、意识与认知障碍诊疗、精神疾病诊疗以及癫痫和神经发育障碍诊疗等。

今年脑机接口领域亮点企业较多，行业进入快速发展阶段。成立于 2016 年的 Neuralink 是脑机接口领域的先行者。今年 4 月，Neuralink 发布的视频中，展示了猴子脑中植入芯片后玩模拟乒乓球的电子游戏“Mind Pong”的场景。该公司目前致力于将首款产品 N1 Link 推向市场，据创始人马斯克称，第一款 Neuralink 产品将使瘫痪患者用他们的头脑比用拇指的人更快地操作智能手机。今年 7 月，Neuralink 宣布完成 2.05 亿美元（约 13 亿人民币）C 轮融资，是目前在所有涉及脑机接口的公司中出现最大规模的融资。

图注：猴子用脑机接口玩电子游戏的图示

来源：特斯拉

Neuralink 主要竞争对手是位于美国德州奥斯汀的 Paradomics, 该公司成立于 2015 年, 其首个商用产品 Connexus 通信设备将为因严重瘫痪而失去说话能力的患者恢复通信。该公司于今年 7 月获得 2000 万美元种子轮融资。

国内方面，已经有多家公司在不同的分赛道大力布局，其中有企业专注脑科学基础研究，也有企业拓展了睡眠监测等其他新兴应用市场。

北京人工智能+脑科学创业公司优脑银河成立于 2019 年，在脑科学基础研究方面通过个体精准脑功能区剖分技术（personalized Brain Functional Sectors, pBFS），实现精确量化全脑 200 多个功能区。目前，优脑银河的“优点”创新疗法采用无创脑机交互方式，通过对人脑进

行“读信号、解信号、写信号”，在个体层级进行精准检测分析、诊断，并在个性化脑功能区的指导下，实现对脑疾病的精准干预。公司目前针对抑郁症、偏瘫、失语、帕金森、毒品成瘾、强迫症和自闭症等适应症的检测和治疗，与国内多家医院联合展开临床实验。今年 8 月，优脑银河完成人民币 5 亿元 A 轮融资。

成立于 2011 年的脑科学公司博瑞康已经在脑机领域深耕十年，而今又被资本市场看好。早在 1999 年，清华大学医学院神经工程实验室便在脑机接口领域展开了探索,而博睿康的核心团队成员便是来自脑机接口全球 Top5 的清华大学神经工程实验室以及临床神经领域的医疗市场专家。目前，博瑞康团队正在与脑科学和临床神经诊疗专家合作，通过研发微创脑机接口进行癫痫疾病的治疗。今年 3 月，博睿康完成人民币过亿元 B 轮融资。

成立于 2020 年的浙江脑机接口技术公司柔灵科技专注于非侵入式脑机接口在睡眠监测领域的应用。目前，该公司主打的 C 端产品纳米级 AIoT 脑电睡眠贴片也已迭代至 3.0 版本，据报道，该产品预计今年年底小规模量产，明年年初上市销售。今年 8 月，浙江脑机接口技术公司柔灵科技完成人民币数千万元天使轮融资。

图注：柔灵科技纳米级 AIoT 脑电睡眠贴片

来源：柔灵科技

同样入主睡眠监测市场的还包括成立于 2015 年的云睿智能。据悉，云睿智能核心产品优梦思 UMindSleep 额贴式睡眠记录仪已通过国家药品监督管理局（NMPA）医疗器械注册认证，可直接进入各医疗机构临床科室使用，为睡眠障碍症患者的诊疗提供了便利。云睿智能于今年 8 月完成人民币数千万元战略融资。

北京宁矩科技有限公司（NeuraMatrix）是国内唯一具备侵入式技术能力的脑机接口平台公司。该企业成立于 2019 年，由清华大学孵化，据报道，NeuraMatrix 自研的双向脑机接口商用芯片已完成流片，预计明年初实现搭载自研芯片设备的量产，其首个无线侵入式脑机接口设备样机也已开始供客户使用。目前该企业已经与多家科研院所、医院及药企签订了上千万元人民币设备采购意向订单，合作对象包括清华大学附属长庚医院、天坛医院、宣武医院等。NeuraMatrix 已经于今年 11 月完成人民币亿元 A 轮融资。

人工智能应用层企业——自动驾驶

总体趋势

2021 年，自动驾驶行业迎来新的投融资热潮。截止 2021 年 11 月，创投机构在自动驾驶技术领域的投资达 94 笔，远超 2019 年（54 笔）和 2020 年（56 笔）。覆盖领域包括自动驾驶解决方案、芯片、雷达等领域。

截至 2021 年三季度末，国内自动驾驶赛道的融资总额已经超过 1000 亿元人民币（含新造车）。这是十年来自动驾驶赛道资本热度最高的一年。

其中引人注目的投融资事件包括，自动驾驶卡车企业图森未来于今年 4 月正式登陆纳斯达克，成为自动驾驶全球第一股。11 月，硅谷无人车创企 Aurora 借壳上市。福特和大众支持的 Argo AI、智加科技等也陆续公布了上市计划。与此同时，大额融资十分密集，国际领军企业 Waymo、Cruise 均完成几十亿美元级别巨额融资，小马智行、文远知行、滴滴自动驾驶、Momenta 等国内独角兽创企均于今年完成大额融资。自动驾驶芯片代表企业地平线也完成 15 亿美元大 C 轮融资，堪称业界之最。

2021 年自动驾驶领域投融资趋势包括：

Robotaxi 赛道大热，成为当前最具市场空间的自动驾驶落地模式。

基于测试里程积累，AutoX、小马智行、百度等国内第一梯队企业在 Robotaxi 技术及商业化进程上处于领先地位；

矿区物流等特种场景成为企业关注的垂直新兴领域。

除了乘用车场景以外，港口、货运、城市物流、环卫等特种作业场景也是自动驾驶企业寻求商业落地的主要路径。大批细分赛道企业完成融资；

芯片和激光雷达等自动驾驶硬件企业成为投融资新热点。

下文分别从 Robotaxi 赛道、硬件（芯片、激光雷达）赛道、其他细分场景赛道等三方面进行产业情况和亮点公司概述。

Robotaxi

2021 年，随着自动驾驶落地加速，Robotaxi（自动驾驶出租车）带有“共享出行+自动驾驶”的双重光环，引得各路玩家争相入局。当中既包括 Waymo、百度在内的科技巨头，也有小马智行、文远知行等一众自动驾驶解决方案提供商，滴滴出行等出行服务公司，以及特斯拉，通用（子公司 Cruise）在内的整车企业。

图注：Robotaxi 赛道主要参与者

来源：智源研究院整理

要实现 Robotaxi 全无人驾驶出租车（取消安全员）的概念，Robotaxi 自动驾驶等级必须在 L4 级以上。其技术难度高，落地难度大，但因其独特的商业模式和可观的单位利润，市场空间巨大，各方参与者纷纷入局 Robotaxi 军备竞赛，融资用于扩张车队规模，力求在更多的城市开启路测并落地，甚至试水商业化。今年，国内大批 Robotaxi 企业已进入车队测试及服务试运营的阶段，未来行业的竞争核心也将会转向运营规模与测试里程的比拼。

公司	车队规模	部署地点	测试里程
百度	500辆+	北京、上海、广州、长沙、沧州	超1600万公里
小马智行	200辆+	北京、广州、美国加州	超800万公里
文远知行	100辆+（2020年）	广州	超700万公里
滴滴自动驾驶	100辆+	上海、北京、美国加州	超530公里开放道路测试
AutoX	100辆+	上海、深圳、武汉、美国硅谷	未公布
Waymo	615辆	在美超过10个州	超2000万英里（约合3200万公里）
Cruise	201辆	旧金山	123.2万公里（加州2020年测试里程）

图注：部分 RoboTaxi 主要参与者车队运营情况对比图

来源：智源研究院整理

同时，

政策法规推动 Robotaxi 的商业落地。

今年 9 月，美国加州车辆管理局（DMV）宣布颁发自动驾驶部署许可证给 Waymo 和 Cruise，允许其就对大众提供的自动驾驶服务收取费用。11 月，北京市智能网联汽车政策先行区对外发布细则，向百度、小马智行等企业颁发国内首批自动驾驶车辆收费通知书。此举意味着 Robotaxi 从产品到商品的转变。今年值得关注的企业包括 Waymo、Cruise、Momenta、AutoX、Nuro 等。

谷歌自动驾驶部门 Waymo 成立于 2009 年，原为谷歌内部汽车自动驾驶部门，于 2016 年独立，如今已经发展为无人驾驶领域的领军者。和特斯拉以视觉为核心的技术路线不同，Waymo 属于激光雷达主导的技术路线。

目前，Waymo Driver 自动驾驶系统已迭代至第五代，为一款由多种高性能、互补传感器组成的单一集成系统。

商业模式方面， Waym 于 2018 年 12 月在美国凤凰城推出首个商业自动驾驶叫车服务 Waymo One。目前 Waymo One 车队已经累计提供叫车服务数万次。今年 8 月，该自动驾驶叫车服务在旧金山特定服务区开启试运营。除了 Robotaxi，Waymo 还于去年 3 月针对城市配送、干线物流等场景推出重卡自动驾驶服务 Waymo Via。今年 6 月，Waymo 完成新一轮 25 亿美元融资。

图注：Waymo 无人驾驶车

来源：Waymo 官网

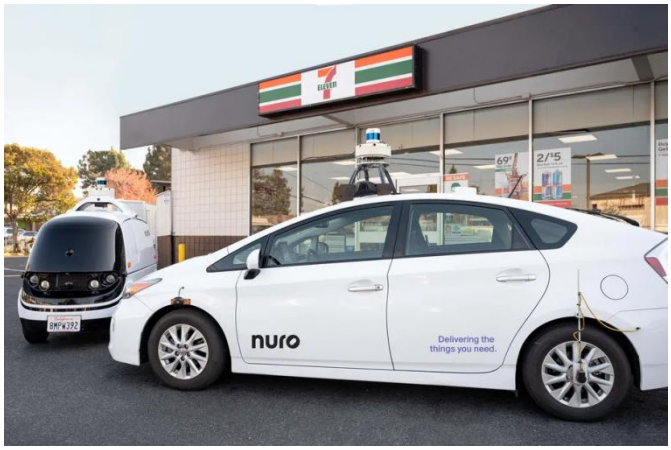
通用汽车旗下自动驾驶公司 Cruise 成立于 2013 年，前称 Cruise Automation，于 2016 年 3 月被通用汽车收购。产品方面，2016 年 Cruise 推出 Bolt EV，由雪佛兰 Bolt EV 轿车搭载 Cruise 自动驾驶方案，并进行了路测。2020 年 1 月，Cruise 联合通用、本田推出无人驾驶共享出租

车 Cruise Origin，没有方向盘、油门和刹车，车身布置激光雷达、毫米波雷达、摄像头等传感器。今年 6 月，Cruise 已经获得通用 50 亿美元的多年期信用贷款额度，开始试量产 100 辆 Origin 车。今年 10 月，Cruise 拿到全自动驾驶汽车（L4）路测执照。今年 1 月，Cruise 获得 20 亿美元融资，4 月，Cruise 再获 27.5 亿美元新融资。新一轮融资后，Cruise 的估值将首次超越 Waymo。

图注：Cruise Origin 产品

来源：Cruise 官网

自动驾驶初创企业 Nuro 成立于 2016 年。Nuro 提供以全自动配送为中心的解决方案，通过无人驾驶送货车，实现外卖、生鲜、百货、药物等配送。目前 Nuro 正与 7-Eleven 公司合作在加州推出自动驾驶送货服务，产品上已经推出自家研发的 R2 车款。今年 11 月，Nuro 完成 6 亿美元 D 轮融资。



图注：Nuro 自动驾驶车

来源：Nuro 官网

自动驾驶初创企业 Momenta 成立于 2016 年。产品方面，Momenta 采用”量产自动驾驶 + 完全无人驾驶路线”的产品战略。其中，量产自动驾驶(Mpilot)，是针对私家车前装可量产的高度自动驾驶全栈式解决方案，覆盖高速、城市快速路、泊车和城区等场景，主要的核心产品有 Mpilotx 等。完全无人驾驶（MSD），广泛应用于出租车和私家车等场景，致力于 L4 级完全无人驾驶技术。去年 10 月，Momenta 正式发布 Robotaxi 产品 Momenta GO。今年 12 月 8 日，由上汽乘用车、上汽人工智能实验室、Momenta、享道出行合作的享道 Robotaxi 在上海嘉定启动运营。融资方面，Momenta 在 9 月获通用汽车投资后，又于 11 月完成 C+轮超 5 亿美元融资，C 轮累计融资超 10 亿美元，是中国自动驾驶领域 2021 年以来最大规模的融资。

图注：享道 Robotaxi 自动驾驶

来源：Momenta 官网

AutoX 成立于 2016 年，是中国目前唯一一家实现城市公开道路完全无人驾驶 RoboTaxi 商业化运营的公司，拥有较大规模车队。2020 年，AutoX 获得了全球第二张、中国首张加州全无人驾驶 RoboTaxi 牌照。目前，Autox 已经在深圳、上海、广州等多地建立了测试车队。今年 7 月，AutoX 正式发布了其第五代全无人驾驶系统 AutoXGen5。

轻舟智航成立于 2019 年，核心创始团队来自 Waymo，同样布局 Robotaxi 赛道，在 Robotaxi 产品形态的基础上拓展出 Robobus 无人小巴的落地场景。目前已逐步推出龙舟系列无人驾驶车，搭载自主研发的

“Driven-by-QCraft”无人驾驶方案，覆盖多种车型，可应用于网约车、公交车及接驳车等多个场景，预计今年车队规模将超过百台，其中龙舟 ONE（轻舟无人巴）在深圳、武汉、苏州等多个城市落地。该公司于 3 月获得数千万 A1 轮融资，并于今年 8 月获得 1 亿美元 A+轮融资。

图注：轻舟智航轻舟无人巴产品

来源：轻舟智航

小马智行 2016 年在硅谷成立，在中美两地运营，也是早期入局 Robotaxi 的自动驾驶创企之一，早在 2018 年 12 月，小马智行便推出了 Robotaxi 项目 PonyPilot。小马智行是除百度外首个在北京获得无人化道路测试牌照的独立自动驾驶公司，目前其 Robotaxi 出行服务已经在北上广三大城市及美国加州多个城市落地，其 L4 级自动驾驶车辆在全球积累了超过 800 万公里的测试里程。此外，小马智行也切入了自动驾驶重卡业务，其卡车业务 PonyTron（小马智卡）已开展商业运营。小马智行于今年 2 月初获得 1 亿美元 C+轮融资。

图注：小马智行无人车

来源：小马智行

文远知行 WeRide 成立于 2017 年。该公司采用 Robotaxi 与无人小巴（Mini Robobus）并行模式。2019 年 11 月在广州推出 Robotaxi 运营服务，截止目前，文远知行自动驾驶里程已累计超过 700 万公里，车队规模超 300 辆。今年 1 月，文远知行和宇通集团合作研发的无人驾驶小巴 Mini Robobus 已在郑州、广州投放测试。此外，文远知行在 9 月联合江铃汽车、中通快递正式发布其首辆自动驾驶货车-WeRide Robovan，进军城市物流行业。文远知行于今年多次获得连续融资，包括 1 月获得 3.1 亿美元 B 轮融资、5 月获得数亿美元 C 轮融资。投后估值 33 亿美元，仅次于小马智行。

图注：文远知行发布首款 L4 级自动驾驶轻客（WeRide Robovan）

来源：文远知行官网

元戎启行成立于 2019 年，是一家 L4 级自动驾驶解决方案提供商，主要为车企、出行公司等提供定制化的自动驾驶解决方案。先后在深圳、武汉、杭州等地展开了自动驾驶测试和试运营，积累超过 200 万公里的测试和运营里程。公司规划 RoboTaxi 自运营车辆在 2021 年底达到 100 辆，包括合作运营车辆在内总体车辆达到 150 辆。今年 9 月，元戎启行完成 3 亿美元 B 轮融资。

百度 Apollo 成立于 2017 年，是国内最早布局自动驾驶的企业之一，截至目前测试总里程超过 1600 万公里，其旗下“萝卜快跑”已在北京、上海、广州、长沙、沧州五地开放常态化运营。去年 10 月，百度自动驾驶出租车服务 Apollo Go 在北京全面开放，用户可在北京经济技术开发区、

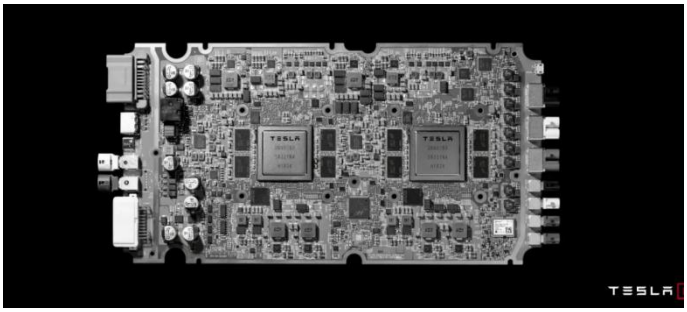
海淀区、顺义区的数十个自动驾驶出租车站点，无需预约，直接下单免费试乘自动驾驶出租车服务。今年 11 月 25 日，百度 Apollo 实现国内首个自动驾驶收费订单，一位北京亦庄居民使用百度 Apollo 自动驾驶出行服务平台“萝卜快跑”完成首单付费。据悉，该名乘客此次乘坐全程 2.1 公里，消费 1.06 元。标志着自动驾驶商业化运营的正式开启。

滴滴自动驾驶于 2019 年拆分成立（原为滴滴自动驾驶部门），在国内外已拥有超过 100 辆自动驾驶车辆，已经获得上海、北京、苏州、合肥、美国加州等地的道路测试牌照。今年 1 月，滴滴自动驾驶完成 3 亿美元 A 轮融资，又在今年 5 月完成 3 亿美元战略融资。

车载芯片

芯片方面，当前全球汽车产业面临“缺芯”困境。在整个汽车 AI 芯片供应市场，高端自动驾驶芯片长期由英伟达和 Mobileye 等巨头把控，但国产芯片企业近年来快速崛起，受到资本市场青睐。2021 年，自动驾驶芯片算力最强的特斯拉、英伟达均有新产品消息传出。特斯拉预计量产 HW4.0，英伟达则推出 Atan 芯片，国内值得关注的芯片企业地平线、黑芝麻等均完成大额融资。

特斯拉在芯片方面专注自研，目前拥有 HW3.0、HW4.0、DOJO 三款自研芯片。今年 8 月，特斯拉在 AI Day 上发布超级计算机 DOJO，算力可达 362TOPS，同时特斯拉宣布将与台积电合作量产新一代自动驾驶芯片控制器 HW4.0，预计其性能为 HW3.0 的 3 倍，算力将达到 432TOPS 以上。2019 年特斯拉推出 HW3.0 时，144TOPS 的算力冠绝行业。特斯拉旗下纯电动皮卡 Cybertruck 将成为首款搭载 HW4.0 的车型，将于 2022 年正式交付。作为自动驾驶行业的风向标，特斯拉的芯片和传感器硬件升级一直备受关注。首先纯视觉的方案，摆脱了对激光雷达的依赖，成本大降，但并没有牺牲太多系统稳定性。从 HW1.0 到 HW2.5，自动驾驶传感器套件没有太大的升级，但是到了 HW3.0 有了重大更新。HW3.0 采用特斯拉自研的自动驾驶芯片，抛弃英飞凌/英伟达的芯片产品，自研高度集成的 SoC+MCU 芯片，全套芯片设计、多传感器融合、自动驾驶算法等方面实现了全自主。但是要满足 L4-L5 级别的自动驾驶，就需要 HW4.0。没有使用激光雷达，而且算法日益成熟，FSD 更新后的算力也将有大幅提升，特斯拉的纯视觉技术方案将成为自动驾驶的重要一隅。



图注：特斯拉 HW3.0 解决方案

来源：特斯拉官网

以色列企业 Mobileye 成立于 1999 年，早期从事自动驾驶芯片与 ADAS 产品的研发，于 2017 年被英特尔收购。二十多年来，Mobileye 以视觉感知技术为基础，推出了算法＋EyeQ 系列芯片组成的一系列解决方案。目前，Mobileye EyeQ 系列产品已在福特、上汽、宝马、沃尔沃、威马、长城、广汽、一汽等传统老牌车企，以及蔚来、理想、小鹏等造车新势力上上车。最新产品为 2020 年推出的 EyeQ5 芯片，单颗算力达 24TOPS，算力方面和英伟达、高通及地平线等产品相比有所逊色。今年 12 月，英特尔表示计划于 2022 年年中推动 Mobileye 在美国上市，估值或超过 500 亿美元。

图注：Mobileye EyeQ5 芯片

来源：Mobileye 官网

厂商	产品	算力	量产时间
特斯拉	HW4.0	432 TOPS 以上	2021-2022
	Dojo	362 TOPS	
英伟达	Orin	254 TOPS	2022
	Atlan	1000 TOPS	
Mobileye	EyeQ5	24 TOPS	2021
地平线	征程5	128 TOPS	2022
黑芝麻智能	A1000 Pro	106-196 TOPS	2022

图注：部分自动驾驶芯片供应商产品对比情况

来源：智源研究院整理

成立于 1993 年的英伟达在自动驾驶上的技术路线不断迭代，产品囊括 Xavier、Orin、Atlan 系列芯片，以及 Hyperion、Drive AGX 系统平台，可支持 L2-L5 级别的自动驾驶。英伟达于今年 4 月的 GTC 大会上发布了全新的自动驾驶 SoC Atlan 芯片，单颗 SoC 的算力能够达到 1000TOPS，相比上一代 Orin 芯片算力提升接近 4 倍（254TOPS），可用于搭载 L4、L5 级别自动驾驶。此前英伟达推出的 Orin 及 Xavier 仍是当前车企搭载的主流，英伟达于今年 GTC 大会上宣布威马、蔚来、小鹏等 7 家车企将搭载 Orin-X 芯片。

地平线成立于 2015 年，是全球领先的嵌入式人工智能核心技术和系统级解决方案提供商，致力于为自动驾驶汽车、智能摄像头等终端设备安装“大脑”（AI 芯片），让它们具有从感知、交互、理解到决策的智能。2021 款理想 ONE 便使用了两颗最新款征程 3 芯片，也是该芯片的首次量产上车。2021 年 5 月，地平线第三代车规级产品，面向 L4 高级别自动驾驶的大算力征程 5 系列芯片宣布流片成功。2021 年 6 月 10 日，地平线已完成高达 15 亿美元大 C 轮融资，投后估值达 50 亿美元。

图注：地平线征程 5 芯片

来源：地平线官网

自动驾驶计算平台企业黑芝麻智能成立于 2016 年，该公司于今年 4 月发布了最新产品车规级自动驾驶芯片华山二号 A1000 Pro，同年 7 月流片成功，算力可达 106－196TOPS，功耗为 25w，单颗芯片可以支持 L3/L4 高级别自动驾驶功能，从泊车、城市内部到高速场景的无缝衔接，预计最

快将于 2022 年底实现车型量产上市。此外，黑芝麻智能还基于华山二号 A1000 自动驾驶芯片，发布了针对车路协同场景应用的新一代车路协同路侧感知计算平台——FAD Edge。9 月 22 日，黑芝麻智能宣布完成数亿美元的战略融资及 C 轮融资，投后估值近 20 亿美元。

图注：黑芝麻 A1000Pro 芯片

来源：黑芝麻智能官网

激光雷达

激光雷达是一种通过发射激光来测量物体与传感器之间精确距离的主动测量装置，被誉为机器人的“眼睛”，从扫描模块结构分为机械式、半固态式和固态式。2021 年，国内包括小鹏、蔚来、威马以及北汽、长安等

新旧车企纷纷表示，其新车将首次搭载激光雷达，引发激光雷达量产落地的新纪元。

今年，国内外激光雷达企业也得到了资本市场的支持。今年以来，Innoviz、Ouster、AEye 等国外激光雷达公司先后通过 SPAC 方式上市。国内，速腾聚创今年 2 月完成战略投资，并且计划 2022 年赴美上市；禾赛科技于今年 6 月完成超 3 亿美元 D 轮融资；图达通于今年 8 月完成 6600 万美元 B+轮融资。此外，完成融资的还有挚感光子、安智汽车、欢创科技、行易道等企业。预计到 2025 年，全球范围内具有 L2 或 L3 级自动驾驶功能的汽车销量将达到近 6000 万辆，届时汽车级激光雷达传感器的市场规模将接近 3200 万台。今年值得关注的激光雷达企业如下。

成立于 2016 年的以色列企业 Innoviz, 是全球高端固态激光雷达领导者，主打 MEMS（半固态式）激光雷达方案。2020 年 10 月，Innoviz 推出新一代激光雷达产品 InnovizTwo，同时其首款面向 L3 至 L5 级别的车规级产品 InnovizOne 也将正式量产落地。其产品助力宝马等车企，以及 Robotaxi、Robobus 等项目打造自动驾驶量产车。今年 4 月，Innoviz 宣布与 Collective Growth Corporation（纳斯达克代码：CGRO）完成业务合并，新公司市值超过 10 亿美元。Innoviz 成为 2021 年首家上市的激光雷达企业。

国内企业也有不少押注激光雷达。图达通成立于 2016 年，图达通的客户之一是蔚来汽车，同时其图像级远距离激光雷达也已部署在百度 Apollo

项目中。今年 8 月，图达通完成 6600 万美元 B+轮融资，图达通方面表示，本轮融资将主要用于力挺面向前装量产的蔚来 ET7 激光雷达的大规模量产交付。

激光雷达制造商禾赛科技成立于 2014 年，至今已推出十款激光雷达产品，其 Pandar 系列激光雷达产品广泛应用于 Robotaxi 和机器人市场。禾赛科技于今年 6 月完成超 3 亿美元 D 轮融资。今年 1 月，禾赛科技向科创板递交招股书，但一个月后 IPO 终止。

图注：禾赛科技 Pandar128 激光雷达

来源：禾赛科技

细分场景

除了通用的乘用车场景以外，港口、机场、货运、矿区、城市物流等特种作业场景也是自动驾驶企业寻求商业落地的主要路径，相比于乘用场景，特种场景因为技术难度低、行驶环境

更简单可控，因此实现商业落地可能性更高。今年以来，各细分赛道玩家也纷纷获得资本青睐。其中，图森未来正式登陆纳斯达克，成为自动驾驶全球第一股；智加科技和西井科技等在港口场景持续发力；易控智驾，踏歌智行，慧拓智能则是矿区自动驾驶的领头羊。

干线物流

图森未来成立于 2015 年，创立之初主打计算机视觉应用，后聚焦在自动驾驶领域，专攻自动驾驶卡车产品与技术研发，目前在北京和北美多地设有研发中心，旗下主要产品为一款 L4 级别无人驾驶卡车，可用于高速公路货运和港内集装箱码头运输等场景。今年 4 月，图森未来正式登陆纳斯达克，成为自动驾驶全球第一股。招股书显示，在自运营车队上，图森未来上市时有 70 辆 L4 卡车运行。

图注：图森未来自动驾驶卡车

来源：图森未来

智加科技成立于 2016 年，是全球领先的重卡自动驾驶公司，致力于自动驾驶在干线物流和港口领域的应用。智加科技总部设在美国硅谷，并先后在北京、上海等地成立分公司及研发中心。智加科技第一款商用产品 PlusDrive 是一款自主研发的自动驾驶系统，并已经交付给卡车制造商。2021 年底，搭载 PlusDrive 的一汽解放智能量卡已经量产，其中搭载 PlusDrive 的解放 J7 完成苏州到敦煌往返自动驾驶长途测试。今年 5 月，智加科技计划将在纽交所上市，但 11 月上市计划终止。

赢彻科技成立于 2018 年，是自动驾驶卡车技术和运营公司，聚焦于将自动驾驶卡车技术应用于干线物流。今年 3 月发布 L3 自动驾驶系统 “轩辕”，计划于 2021 年底实现 L3 自动驾驶重卡量产。今年 8 月，赢彻科技完成 2.7 亿美元 B 轮融资。

港口物流

西井科技成立于 2015 年，主打智慧港口场景，2018 年发布全时无人驾驶电动重卡 Q-Truck 及电动重卡 Well-Truck，电动重卡 Well-Truck 已经在深圳盐田国际码头进行路测。

主线科技成立于 2017 年，聚焦 L4 级自动驾驶卡车技术研发与应用，主要面向港口物流与高速干线物流场景。截止今年 10 月，主线科技已累计

向国内港口客户交付超百台无人驾驶电动集卡。今年 11 月，主线科技获人民币数亿元新一轮融资。

图注：主线科技车队

来源：主线科技

机场物流

驭势科技成立于 2016 年，专注 L4 级自动驾驶算法和系统研发，在厂区物流和机场物流场景已经实现商业落地。目前已经在香港国际机场落地无人驾驶拖车，并在湖南机场落地了中国内地首个空港货运无人驾驶技术。同时也与东风汽车合作研发 Robotaxi。今年 1 月，驭势科技宣布完成累计超 10 亿元人民币新融资。

城市配送

美团聚焦末端配送场景中的城市配送，随着疫情常态化发展，无人配送的需求不断增加，今年 4 月 19 日美团发布了新一代自行研制的无人配送车“魔袋 20”， 并在北京顺义区投放进行测试运行。

图注：美团新一代无人配送车魔袋 20

来源：美团官网

毫末智行成立于 2019 年，独立自长城汽车技术中心智能驾驶前瞻分部。末端物流无人车方面，毫末智行已经与美团、物美多点、阿里达摩院等伙伴达成战略合作，目前已经将末端物流无人车陆续配置到商超配送等特定场景中，已经有 1000 辆车量产下线。毫末智行预计将在 2022 年年中发

布辅助驾驶系统 HPilot 的全新功能“城市 NOH”，并计划在 2022 年下半年交付全场景 NOH，同时在 2023 年推出拥有 HSD（HAOMO Self-Driving）的车队。12 月，毫末智行完成 A 轮近 10 亿元融资。

矿区物流

矿区是自动驾驶落地最快的场景之一，矿区人力成本极其高昂，无人化运输能够极大地提升生产效率，前景十分广阔。从今年矿区无人驾驶企业的融资情况可见一斑。目前，国内至少有踏歌智行、慧拓智能、易控智驾、伯镭科技、希迪智驾（CIDI）、盟识科技等矿区无人化解决方案提供商在内蒙、河南、北京等地做测试或运营。

踏歌智行成立于 2016 年，是国内最早开始关注矿区无人驾驶的企业。目前已推出车-地-云架构的露天矿无人驾驶运输全栈式解决方案，打造由云端智能调度管理、车联网通信、智能路侧设备和车载控制系统组成的全套产品谱系，已开始批量化部署。今年 1 月，踏歌智行宣布完成了数千万美元 B+轮融资。

矿山无人驾驶的运营商易控智驾成立于 2018 年，于今年 6 月完成数千万美元 B1 轮融资。慧拓智能成立于 2014 年，是中科院自动化所孵化的全栈式无人矿山整体解决方案、产品和运营服务提供商。慧拓智能于今年 8 月完成超过 2 亿人民币的 B1 轮融资，创赛道内单笔最大融资纪录。

图注：易控智驾矿区卡车

来源：易控智驾

城市环卫

仙途智能成立于 2017 年，主要聚焦城市环卫场景。在环卫场景中，环卫作业高强度、低技术难度，因此无人驾驶替代可行性高。仙途智能目前在全球范围内已经投入近 100 台自动驾驶车辆，覆盖 1 吨-18 吨所有环卫车型，已经在国内一二线城市和瑞士等发达国家相继落地自动驾驶环卫车。仙途智能于今年 5 月获人民币 1.2 亿元 B1 轮融资。

图注：仙途智能卡车

来源：仙途智能

酷哇机器人成立于 2015 年，从 2017 年开始就发力市政环卫和城市配送，已经在 8 个省 23 个地级市部署了规模化的无人驾驶环卫车辆，其中在西安、长沙、成都和芜湖等地，酷哇承接了区一级的基于智能网联和自动驾驶无人驾驶车队的日常清扫保洁任务。同时在城市物流配送和智慧出行方面也有所布局。今年 9 月，酷哇机器人完成 2.5 亿美元 C 轮融资。

其他领域

新能源汽车领域，蔚来、小鹏、理想汽车、威马、零跑、哪吒（360 旗下）以及大众、福特、丰田等传统车企在自动驾驶领域均有布局。今年以来，包括理想和小鹏在内的造车新势力竞速 IPO，相继开始上市步伐。7 月，小鹏汽车在港交所上市。8 月，理想汽车在港交所上市，此前已赴美上市。

人工智能技术层企业

人工智能产业技术层企业主要关注 AI 产业中和技术研发、解决方案供应和集成等相关的企业，包括但不限于自然语言处理、计算机视觉、机器学习、知识图谱、智能语音等技术领域。下面将分别介绍国内有代表性的企业案例。

自然语言处理

自然语言处理是人工智能领域的一个重要技术领域，主要处理文本等数据，能够完成翻译、文本生成、对话、情感分析等任务，落地场景包括文化娱乐、传媒、客服、营销、企业服务等。今年在自然语言处理领域值得关注的企业有 HuggingFace、微软小冰、循环智能、澜舟科技、幂律智能等。

Hugging Face 创立于 2016 年，最早为一家开发聊天机器人的企业。Hugging Face 专注于 NLP 技术，拥有大型的开源社区。其在 GitHub 上开源的 NLP 预训练模型库 Transformers，已被下载超过一百万次，GitHub 上超过 24000 个 Star。今年 3 月，Hugging Face 完成 4000 万美元 B 轮融资。

小冰公司前身为微软（亚洲）互联网工程院人工智能小冰团队，于 2013 年 12 月在中国组建，是微软全球最大的人工智能独立产品研发团队，也是全球承载交互量最大的完备人工智能框架之一。目前，小冰公司主要提供三类产品线和解决方案：人工智能交互主体，品牌为覆盖中国的小冰及日本、印度尼西亚等国家地区的 Rinna；完整的人工智能交互框架操作系统；为部分垂直领域提供的行业解决方案。小冰单一品牌已覆盖 6.6 亿在线用户、4.5 亿台第三方智能设备和 9 亿内容观众，商业客户覆盖金融、零售、汽车、地产、纺织等多个垂直领域。今年 7 月，微软小冰完成 A 轮融资，估值已超过 10 亿美元规模。

澜舟科技成立于 2021 年 6 月，是创新工场孵化的一家认知智能公司，提供针对商业场景数字化转型、以自然语言处理为基础提供商业洞见类产品。主要产品包括基于预训练模型的功能引擎（包括搜索、生成、翻译、对话等）和针对垂直行业场景的 SaaS 产品。 7 月，澜舟科技-创新工场团队与上海交通大学、北京理工大学等单位联合研发了中文语言模型“孟子”。

幂律智能成立于 2017 年 9 月，主要业务是基于自然语言处理（NLP）、数据挖掘等技术处理法律文本，提供智能合同审查、法律智能问答、知识管理、法律检索等多种产品和服务。幂律智能现阶段的主要产品是智能合同审查工具 MeCheck，服务于企业法务部门，已获得互联网、制造业、

零售等多个领域的客户。幂律智能在今年 4 月宣布完成人民币近 6000 万元 A 轮融资。

循环智能成立于 2016 年，是一家 AI 企业服务公司，主要为企业提供从全渠道沟通数据采集到录音转写、内容挖掘、NLP 语义建模、会话分析洞察和沟通实时辅助的完整解决方案。其产品是一款新生代的 SaaS 智能化销售系统，包含线索推荐系统、客户心声分析以及服务质检三大核心模块。今年 4 月，循环智能与华为云联合开发的首个千亿参数中文模型——盘古 NLP 大模型。目前，循环智能已经将语言大模型的技术能力逐步应用到多个领域，其中在教育领域，已服务了包括新东方在线、51Talk、火花思维等教育机构。今年 12 月，循环智能完成 3800 万美元新一轮融资。

计算机视觉

计算机视觉是 AI 产业中的一大分支领域，在技术成熟度、商业化进程、市场增长速度、投融资热度等方面，是人工智能产业当前热门的发展赛道。2021 年，我国计算机视觉产业快速发展，企业加快上市步伐，争夺“视觉 AI 第一股”。今年值得关注的计算机视觉企业包括“AI 四小龙”，以及云天励飞、格灵深瞳、扩博智能等。

商汤科技创立于 2014 年，专注计算机视觉技术与深度学习底层算法，当前商汤科技正在集中于计算机视觉的后端市场，专注于云端、平台的搭建，通过开发智慧商业、智慧城市、智慧生活和智能汽车平台，覆盖商业、生活、出行等落地应用场景。此外，商汤科技还推出了其新型人工智能基础设施——SenseCore 商汤 AI 大装置。目前，商汤正在上海临港建设大型人工智能计算中心（AIDC），以支持基于云端的 AI 模型生产及部署服务。今年 11 月，商汤科技已获港交所批准在港上市，此次 IPO 计划筹资逾 10 亿美元。

旷视科技成立于 2011 年，聚焦金融安全、城市安防、手机 AR、商业物联，工业机器人五大行业。旷视科技的发展战略正在向产业链上游的硬件端延伸，推出了人工智能摄像头、边缘计算设备等。今年 9 月，旷视科技科创板 IPO 审核通过。此次旷视科技计划募资人民币 60.18 亿元，将用于基础研发中心建设、AI 视觉物联网解决方案及产品开发与升级、智能机器人研发与升级、传感器研究与设计等。

依图科技成立于 2012 年，将计算机视觉人工智能创新性研究与行业应用相结合，将安防和医学影像分析作为主要发展战略，推出了 care.ai 医疗智能全栈式产品解决方案，为医院提供跨科室的多场景应用系统和数据分析平台。今年 6 月上交所披露，依图科技申请撤回科创板上市申请文件。

云从科技成立于 2015 年，自主研发金融行业解决方案，主要应用在银行、互联网金融、证券、基金、保险、汽车金融等场景。目前，云从科技在金融领域布局，先期主要布局银行线下网点智慧化改造业务，逐步切入金融机构核心风控业务。今年 7 月，云从科技科创板首发过会。

除“AI 四小龙”之外，我国计算机视觉初创企业在计算机视觉下游场景谋求发展空间。今年其他值得关注的企业包括：云天励飞、格灵深瞳和扩博智能。

云天励飞成立于 2014 年，以人工智能算法、芯片技术为核心，提供算法软件、芯片等自研核心产品，并可根据客户需求，为客户提供定制化或标准化硬件产品、安装施工服务等打包解决方案，目前主要收入来源为数字城市运营管理及人居生活智慧化升级业务。云天励飞现已在深圳、上海、成都、青岛、杭州等数十个大中型城市实现了人工智能技术、产品和解决方案落地。今年 8 月，云天励飞科创版首发过会。

格灵深瞳成立于 2013 年，具备计算机视觉和深度学习技术以及嵌入式硬件研发能力，主要关注公共安全、智能交通、金融安防等领域，主要产品包括深瞳人眼摄像机、威目车辆大数据系统、威目视图大数据分析平台、威目人脸识别系统和皓目行为分析仪等，在无人驾驶、机器人和智能医疗方面也进行布局。今年 11 月，格灵深瞳科创板首发过会。

扩博智能成立于 2016 年，主要关注风电和零售行业。在风电领域，扩博智能在风机行业巡检产品融入了“三维无人驾驶”技术，研发异形机器人检修技术，实现从巡检到维修的覆盖；建立了风电行业巡检维修数字化管理平台和行业数据平台，未来将向设备预测性维护方向发展。在零售领域，扩博智能以机器视觉和智能硬件为核心，提供包括货架冰柜洞悉、商品情报分析、销售执行监测等智能化服务。今年 3 月，扩博智能完成人民币 2 亿元 Pre-B 轮融资。

机器学习

机器学习是人工智能中的一项主要技术，通过学习数据集中的相关关系，构建对于行业、领域情况的分析，辅助人类进行决策和判断。随着硬件设备的更迭，处理速度的加快，算力日渐提高，机器学习模型规模逐渐增大，可以学习的数据更为丰富，具备更强更精准的智能能力，在企业应用中能够处理更加深层和复杂多变的问题，具有更广阔的发展前景。今年值得关注的机器学习研发企业包括第四范式、杉树科技和瑞莱智慧（RealAI）等。

第四范式成立于 2014 年，主要关注人工智能平台与技术服务，目前已开发端到端的企业级人工智能产品 4Paradigm AIOS-企业级 AI 操作系统、AI 应用开发平台 4Paradigm Sage Studio、HyperCycle 人工智能学习平台、4Paradigm SageOne-软件定义算力平台等。已应用于金融、

零售、制造、能源、电信、医疗等领域。今年 1 月，第四范式完成 7 亿美元 D 轮融资。

北京瑞莱智慧科技有限公司（RealAI）成立于 2018 年。，业务内容主要包括两方面。一是提升 AI 安全性，检测和防范 AI 滥用风险；二是用 AI 提升产业安全，具体产品包括人工智能系统安全性检测、AI 防火墙、AI 虚假内容检测、金融风控和工业设备资产运营管理等解决方案。2021 年 10 月 28 日，瑞莱智慧完成超 3 亿元人民币 A 轮融资。

智能语音

智能语音是国内人工智能领域的热门赛道。在教育、医疗、金融、电商、客服等领域，以及生活、办公、家居、驾驶等场景中均有广泛应用。12 月，中国语音产业联盟发布《2020-2021 中国语音产业发展白皮书》，预计 2021 年我国智能语音市场规模可达 285 亿元。

国内企业中，科大讯飞等大型智能语音科技企业，凭借常年的技术积累，优势明显。而出门问问、云知声、赛舵智能、竹间智能、追一科技等初创企业在智能语音领域积极创新，通过打造差异化产品形态，在垂直领域和细分场景中占据一席之地。国外企业当中，以 PolyAI 为例的 AI 客服企业暂露头角。

科大讯飞成立于 1999 年，长期从事语音及自然语言理解、机器学习推理及自主学习等 AI 技术研究。2010 年，科大讯飞发布讯飞开放平台，为开发者提供人工智能解决方案。 截至 2021 年 7 月 31 日，讯飞开放平台已开放 437 项平台能力。今年 10 月，科大讯飞发布了开放平台的 2.0 战略，在平台基础上，持续拓展行业赛道，推动在消费者、智慧教育、智慧城市、智慧司法、智能服务、智能汽车、智慧医疗、运营商等领域的深度应用。2021 年 8 月，科大讯飞发布公告，拟分拆子公司讯飞医疗赴港上市，计划集资约 5 亿美元(约 39 亿港元)。讯飞医疗成立于 2016 年，主要业务包括智联网医疗平台、智医助理、智慧医院三部分。分拆上市完成后，科大讯飞持有讯飞医疗 51%的股份，仍将维持对讯飞医疗的控股权。

出门问问成立于 2012 年，主要研发智能语音相关的产品和解决方案，建立完整的“端到端”人机交互技术栈，向市场推出一系列软硬结合的 AI 产品，如智能手表、无线智能耳机、智能后视镜、智能音箱等。出门问问也推出了基于其虚拟个人助理的免费 AI 平台(ai.chumenwenwen.com)，开放出门问问语音助手给合作企业，构建完整的 AI 生态。

云知声成立于 2012 年，是一家以语音技术为核心的 AI 独角兽企业，核心业务为 AIoT 服务。云知声利用机器学习平台——云智云 AIoT 平台，在语音技术、语言技术、知识计算、大数据分析等领域构建了人工智能技术图谱，在家居、医疗、教育、车载市场行业均有应用。今年 2 月，云知声撤回科创板 IPO 申请。今年 6 月，云知声完成 D1 轮近 1 亿美元融资。

智齿科技成立于 2014 年，是一家提供智能在线客户机器人、智能语音机器人、呼叫中心等服务为主的客户全生命周期营销与服务解决方案提供商。目前已经拓展了智齿客服、智齿营销、智齿智客及 BPO 四大业务线，其产品矩阵覆盖客户联络中心的“营销+服务+管理”三大场景。今年 4 月，智齿科技宣布完成人民币 2 亿元 C+轮融资。

赛舵智能成立于 2019 年，是一家智能客服产品研发商，主要聚焦语音识别、语音合成、自然语言处理和语义理解相关技术和应用。目前赛舵智能提供的产品形态主要为催收、质检、销售及客服四方面智能机器人，支持私有云+SaaS 的部署方式。赛舵智能于 2021 年 3 月完成数千万人民币 Pre-A 轮融资；2021 年 8 月获超 1000 万美元 A 轮注资。

竹间智能成立于 2015 年，以独特的情感计算、自然语言处理、深度学习、知识工程、文本处理等人工智能技术为基础，致力于打造中国首款情感机器人。竹间智能通过云化的人机交互产品，为企业提供全场景的客户交互解决方案。2021 年，竹间智能推出 AI 云平台“竹间云”。今年 4 月，竹间智能完成人民币 1 亿元 C+轮融资。

追一科技成立于 2016 年，基于自然语言处理、多模态技术，追一科技打造了对话与分析 AI 应用平台及 AIForce 数字员工产品族，并已应用于企业营销、销售、服务、运营场景。其中，AIForce 数字员工产品族涵盖在线机器人 Bot、语音机器人 Call、多模态数字人 Face、智能培训机器人 Learn 等多款产品。今年 6 月，追一科技完成数亿元人民币战略融资。

PolyAI 成立于 2017 年，该公司聚焦对话式人工智能技术开发，旨在打造和部署自动化客服语音助手，以此帮助客户公司获得高性价比、快速响应的客服系统。今年 9 月，PolyAI 完成 1400 万美元融资。

知识图谱

行业中应用知识图谱，可以发掘实体之间的关联，整合数据，解释现象，知识推理，从而发掘深层关系，实现智慧搜索与智能交互。目前知识图谱广泛应用在银行、保险、证券、法院、物流、财税、搜索、电商、社交等领域。除了互联网大厂在知识图谱方向布局的平台，今年值得关注的企业包括明略科技、智谱华章、擎盾信息和创邻科技等。

明略科技成立于 2006 年，是全球企业级数据分析与组织智能服务平台，致力于通过大数据分析挖掘和认知智能技术，推动知识和管理复杂度高的大中型企业进行数字化转型。今年明略科技与腾讯云一起打造了行业知识图谱解决方案，大型案例已经超过 50 个。

北京智谱华章科技有限公司成立于 2019 年 6 月，主要业务是打造数据和知识双轮驱动的下一代人工智能框架，并在此之上开发各类智能应用。目前智谱华章构造了高质量大规模知识图谱、研发了深度隐含关联挖掘算法和认知图谱等核心关键技术，服务政府、企业、科研机构。2021 年 9 月 14 日，智谱华章获得了 A 轮融资，融资额达到 1 亿元人民币。

南京擎盾信息科技有限公司成立于 2009 年，研究发展通用法律人工智能技术，为各级司法系统提供智慧法律解决方案。旗下的智能法律问答机器人以有法律咨询需求的普通用户为服务对象，运用人工智能算法+法律知识图谱技术，以智能硬件为载体，为用户提供法律咨询服务。2021 年 2 月 9 日，擎盾数据完成总额近 1 亿元 C 轮融资。

创邻科技成立于 2016 年，是一家商业图数据库供应商，其自主研发分布式原生并行图平台产品“Galaxybase”，可以提供从数据迁移、数据建模、数据存储、数据查询、数据运算到数据分析的一站式解决方案，实现海量数据的实时深链查询、分析。目前，Galaxybase 已在金融、互联网、电力、政企等行业多个客户落地使用。2021 年 2 月，创邻科技宣布完成数千万元 A+轮融资。

人工智能基础层企业

人工智能基础层是人工智能产业的基础，主要包括 AI 芯片等硬件设施及云计算等服务平台的基础设施、数据资源，为人工智能提供数据服务和算力支撑。下面将分别介绍今年国内 AI 芯片和数据服务领域有代表性的企业案例。

AI 芯片

总体趋势

根据功能，AI 芯片可以分为 AI 训练芯片和 AI 推理芯片。随着 AI 芯片技术的不断发展，芯片制程不断优化，工艺逐步提升，形成以下发展趋势。一是 AI 芯片功能的细分程度进一步提升，CPU 成为计算体系中的控制和调度中心，而 AI 加速任务主要交由 GPU、FPGA、ASIC 等执行，形成异构形态的计算格局。二是高效、节能成为 AI 芯片发展的长期目标。追求在提升算力的前提下降低功耗，是近年来企业关注的重点。三是 GPU 依然是 AI 芯片企业研发关注的重点方向。GPU 性能较高，且兼具计算的灵活性，适用于构建大规模的 AI 计算集群，在研发超大规模 AI 模型方面具有应用前景。

AI 训练芯片

AI 训练芯片是指为机器学习和深度学习模型训练提供加速支持的高性能芯片。

今年在 AI 训练芯片领域值得关注的国内企业有昆仑芯（原百度芯片业务部门）、燧原科技、上海天数智芯、摩尔线程、沐曦集成电路等。国外 AI 训练芯片领域值得关注的有 Cerebras Systems、英国 AI 芯片独角兽 Graphcore、Testorrent 等 AI 芯片独角兽。

英国 AI 芯片独角兽企业 Graphcore 成立于 2016 年，为 IPU 智能处理器制造商。2020 年 7 月，Graphcore 推出第二代云端训练芯片 GC200。今年 10 月，Graphcore 发布最新产品 IPU-POD128 和 IPU-POD256，分别能够提供 32 petaFLOPS 和 64 petaFLOPS 的 AI 计算。

Cerebras Systems 成立于 2015 年，专注于为数据中心训练提供芯片产品，曾被 CB Insights 评为“全球最值得期待的 100 家芯片公司”。Cerebras 曾于 2019 年推出全球最大 AI 芯片 Wafer Scale Engine，2020 年又推出了新一代的 7nm 的 WSE-2，晶体管数量达 2.6 万亿个。今年 8 月，Cerebras 宣布推出世界上第一个人类大脑规模的 AI 解决方案-CS-2 AI 计算机，可支持超过 120 万亿参数规模的训练。今年 11 月，Cerebras 宣布完成 2.5 亿美元 F 轮融资。

美国 AI 芯片独角兽 SambaNova Systems 成立于 2017 年，今年 4 月，SambaNova Systems 推出了其第二代云端 AI 芯片可重构数据流单元 Cardinal SN10 RDU，包含 400 亿晶体管，采用台积电 7nm 制程，由一系列可重构节点组成，用于数据、存储和交换。今年 4 月，SambaNova Systems 宣布完成 6.76 亿美元 D 轮融资。

加拿大芯片初创企业 Testorrent 成立于 2016 年。2020 年 4 月，Tenstorrent 推出其第一款 AI 芯片 Grayskull，兼顾 AI 训练和推理任务。今年 5 月，Testorrent 完成 2 亿美元 C 轮融资。

今年 8 月份，百度发布“昆仑 2”AI 芯片并开始量产。昆仑 2 采用 7nm 制程，搭载第二代 XPU 构架，INT8 算力在 512-768TOPS 之间，INT/FP32 算力在 32-48TOPS 之间，性能是昆仑一代芯片的 2-3 倍。该芯片可以应用在云边端计算等多种场景中，包括智慧城市、智能制造、自动驾驶等领域。

燧原科技成立于 2018 年。该公司主要产品为云端算力平台芯片，今年推出面向人工智能、具有高能耗比的 AI 芯片“邃思 2.5”，支持单精度、半精度和整型多种运算，TF32 的算力为 128TFLOPS，INT8 算力达到 256TOPS。目前公司多个产品已经在云计算中心、超算中心、金融以及

智慧城市等行业内得到应用，合作伙伴主要有工信部标准化研究院、上海国际汽车城等。今年 1 月，燧原科技完成人民币 18 亿元 C 轮融资。

上海天数智芯半导体有限公司成立于 2015 年。该公司目前已经成功研发出支持并行云端计算的 7nm 芯片，可同时用于训练和推理，实现在计算机视觉、智能语音、智能推荐等深度神经网络模型，面向云端和超算场景。此外，天数智芯开发了面向 GPU 的驱动软件、编译器、函数库和工具链。11 月，天数智芯云端 7nm GPGPU 产品卡——天垓 100 已进入量产，该芯片容纳 240 亿晶体管，采用 2.5D CoWoS 晶圆封装技术，支持 FP32、FP16、BF16、INT8 等多精度数据混合训练，单芯算力在 FP16 精度下达到 147TFLOPs。目前公司的合作伙伴包含 H3C、上海超算中心、世纪互联以及闻泰科技等在内的多家企业及机构。今年 3 月，天数智芯获得人民币 12 亿元 C 轮融资。

摩尔线程成立于 2020 年，主要关注 GPU 研发、GPU SoC IP 研发，以及国产 GPU 生态等方面的业务，已经于今年公布研制成功首颗 GPU 芯片。该公司主要创始成员来自英伟达，目前已与国网电商科技、统信软件在内的多个客户展开合作，助力高性能桌面云服务、工业元宇宙等行业发展。今年 8 月和 11 月分别获得数十亿人民币融资，其中 11 月获得人民币 20 亿元 A 轮融资。

沐曦集成电路成立于 2020 年。该公司主要研发高性能 GPU 芯片，并提供 GPU 架构定义、GPU IP 设计、GPU SOC 设计及 GPU 系统解决方案服务等业务。今年 8 月，沐曦获得人民币 10 亿元 A 轮融资。

登临科技成立于 2017 年，其首款基于 GPU + 架构的 Goldwasser 系列产品已开始量产，并得到客户的积极反馈。登临科技分别于今年 2 月和 11 月完成 A+轮融资和战略融资。

壁仞科技成立于 2019 年，聚焦云端通用智能计算，逐步在人工智能训练和推理、图形渲染、高性能通用计算等多个领域赶超现有解决方案，其首款通用 GPU——BR100，采用 7nm 制程，于 10 月正式交付开始流片，预计将于明年面向市场发布。壁仞科技于今年 3 月完成人民币数十亿元 B 轮融资。

AI 推理芯片

AI 推理芯片主要用于人工智能模型推理过程中的计算加速，在云边端多种场景和领域具有广泛引用。2021 年 AI 推理芯片应用场景分布在边缘计算、AIoT、智慧城市等领域。今年在 AI 推理芯片领域值得关注的国内企业有深圳睿思芯科、爱芯元智等。国外企业包括以色列 AI 芯片独角兽 Hailo Technologies、美国云端 AI 芯片初创公司 Groq 等。

美国云端 AI 芯片初创公司 Groq 成立于 2017 年。目前，Groq 已推出其首个云端推理芯片 GroqChip，并正在研发第二代新型芯片，用于推理任务。今年 4 月，Groq 完成 3 亿美元战略融资。

以色列 AI 芯片独角兽 Hailo Technologies 成立于 2017 年，目前已推出一款 AI 推理芯片 Hailo-8 及 M.2 和 Mini PCIe 加速模块，AI 推理芯片 Hailo-8 已经实现量产。今年 10 月，Hailo 完成 1.36 亿美元 C 轮融资。

深圳睿思芯科有限公司成立于 2018 年。该公司主要业务是推动 RISC-V 架构处理器技术的商业化，产品线分为 Pygmy 和 Pygmy-E 两种，分别面向 AIoT 和 IoT 的高性能 AI SOC，面向终端设备上的 AI 推理任务，落地场景包括机器人、语音、视觉等。今年 9 月，睿思芯科获得数千万美元 A+轮融资。

爱芯元智（原名“爱芯科技”）成立于 2019 年。该公司主要研发高能耗比的 AI 视觉芯片，面向多种视觉任务。目前主要有 AX630A 和 AX620A 两款芯片。AX630A 已成功流片，该芯片可以达到 28.8TOPS 的 INT4 计算能力，能够在保持高性能的同时维持低能耗。例如，在 4K@30fps 条

件下开启画质增强和智能分析，功耗低于 3W。产品主要应用场景有智慧城市、智能社区、驾驶、零售、家居以及穿戴设备等。今年 8 月获得人民币数亿元 A 轮融资。

瀚博半导体成立于 2018 年，该公司主打 AI 推理+视频加速卡，其首款服务器级别 AI 推理芯片 SV102 及通用加速卡 VA1 已于今年 7 月发布，即将量产上市。此外，产品矩阵还包括图形 GPU 等。今年 12 月，瀚博半导体完成人民币 16 亿元的 B-1 和 B-2 轮连续融资。

其他 AI 芯片

DPU（数据处理器）是一种专用处理器，既可以单独作为嵌入式处理器，又可以集成在板卡中。DPU 在数据处理方面具有优势，获得 AI 芯片企业关注。今年在 DPU 芯片方面，值得关注的有英伟达、星云智联、大禹智芯、益思科技、芯启源等。

今年 4 月，英伟达公司发布 BlueField-3 DPU，该芯片基于 NVIDIA DOCA 架构，16 核 Arm Cortex-78 CPU 设计，包含 220 亿组电晶体，传输量 400Gbps，运算效率是上一代的 4 倍。BlueField-3 预计于 2022 年第一季度实现样品发布。

星云智联成立于 2021 年 3 月，在今年 3 月、7 月和 8 月获得数十亿人民币融资。该公司专注于数据中心通信架构和 DPU 芯片研发，具有 DPU 相关的 FPGA/SOC 芯片设计与验证、硬件产品架构设计、软件配套在内的全栈核心能力。

芯启源成立于 2015 年，是一家针对超大规模电信和企业级的智能网络提供核心芯片和系统的高科技公司。其核心产品芯启源智能网卡是目前国内先进的基于 SoC 架构的成熟 DPU 完整解决方案，已获得了中国移动苏研院的首批智能网卡订单。今年 6 月，芯启源宣布完成数亿元 Pre-A3 轮融资。

其他今年获得融资的 DPU 芯片企业包括：

1. 4 月，云豹智能完成天使轮融资
2. 7 月，中科驭数完成人民币数亿元 A 轮融资
3. 7 月，大禹智芯完成人民币数千万元 Pre-A 轮融资

类脑芯片基于神经形态架构设计，通过存算一体的方式，能够突破存储瓶颈，具有高性能、低功耗等优势，是近年来 AI 芯片领域研发的重点。类脑芯片产业仍然处于发展阶段，2021 年获得大额融资的企业相对较少，值得关注的企业有灵汐科技、知存科技、九天睿芯，以及英韧科技。

灵汐科技成立于 2018 年，主营业务包括类脑芯片、基于类脑芯片的类脑计算板卡和服务器、软件工具链和系统软件。产品可广泛应用于云端和边缘端的 AI 应用场景，以及脑科学研究等。主要产品为天机芯类脑芯片，采用众核并行、存算一体的异构融合架构，同时支持脉冲神经网络和人工神经网络计算。在算力平台、安防、互联网等场景均有落地前景。目前灵汐科技类脑芯片 KA200 已流片，采用 12nm 工艺，单芯片集成 25 万神经元和 2500 万突触，共 30 个类脑计算核，支持混合精度计算。此外，灵汐科技还开发类脑板卡、类脑模组、类脑服务器及开发工具、软件栈等。

知存科技成立于 2017 年 10 月，该公司使用 Flash 存储器技术开发存算一体加速系列芯片和存算一体 SOC 系列芯片，可以将存储和运算集成在同一芯片内，支持主流的神经网络，具有高性能和低功耗，可用在穿戴设备和智能终端设备中。公司的合作伙伴主要有科大讯飞、北京大学、芯来科技以及北京航空航天大学等公司和机构。今年 6 月份和 9 月份共计获得超过一亿元人民币融资。

英韧科技成立于 2017 年 6 月。该公司主要关注下一代全球存储技术和数据处理系统，提高数据的存储和传输效率，目前的主要产品包括消费级 SSD 控制器 Shasta IG5208 及 RainierQX 等。今年 6 月获得数亿元人民币 C 轮融资。RainierQX 采用四通道 PCIe Gen4 接口，全面支持 NVMe1.4 协议，4K 随机读取速度超过 1M IOPS，顺序读取速度超过 7GB/s，顺序写入速度超过 6GB/s。RainierQX IG5220 控制器目前已经应用在新发布的威刚（ADATA）传奇 840 和惠普 FX90 系列。

九天睿芯成立于 2018 年，今年 6 月份获得数亿元融资。该公司致力于高效模数混合计算研究，提高产品在低功耗延时方面的性能，在 AIoT 等领域内得到应用。目前该公司 SRAM 感存算一体架构芯片 ADA20X 已流片，具有 20TOPS/W 的能耗比。

数据服务

总体趋势

数据服务和平台研发是 AI 基础层软件方面资本的主要着力方面，2021 年行业投资事件共计 37 起，融资额度达到 227.85 亿元，融资的细分行业包含泛娱乐和媒体、图数据、第三方数据服务等领域。

泛娱乐和媒体

在泛娱乐和媒体方面值得关注的企业是数数科技。数数科技成立于 2015 年，该公司的主要业务领域为泛娱乐领域，为客户提供数据采集、处理和分析服务。目前已经开展合作的案例包含椰岛游戏在内的多种游戏。今年 3 月份和 11 月份共计获得 4.76 亿元人民币融资。

安全风控

风险控制方面值得关注的有数美科技和面向内容安全的中科闻歌。

中科闻歌成立于 2017 年，该公司通过多语言、跨模态和深度认知技术打造出天湖大数据智算平台、红旗融媒体平台、闻海全球大数据平台、金安云金融科技平台等多个数据平台，应用于娱乐、智慧城市、安防等多领域。公司目前的合作伙伴有中华人民共和国工业和信息化部、中华人民共和国海关总署、国家互联网应急中心、中华人民共和国公安部等。中科闻歌今年 5 月完成人民币 2 亿元 D 轮融资。

数美科技成立于 2015 年 6 月，该公司打造全栈智能风控系统——天网、天净，以帮助客户解决金融支付、文本、图片、音频及视频相关的多种风险需求，今年 4 月份获得 1.35 亿美元融资。

第三方数据标注

第三方数据标注及服务公司值得关注的有 TalkingData（腾云天下）和海天瑞声。

TalkingData 成立于 2011 年，借助 SmartDP 技术搭建数据采集、标注和服务平台，帮助企业完成智能化和数字化转型。今年 1 月获得 1 亿美元新一轮融资，

海天瑞声成立于 2005 年，该公司是一家专注于数据采集、处理及交付的全栈服务型数据资源和数据服务公司，公司产品包含语音识别数据、语音合成数据、文本数据、图像数据等多个模块。今年 8 月份获得 3.95 亿元融资并上市。

Scale AI 成立于 2016 年，是一家以数据标注为核心业务的创业公司，致力于处理和标注图像、激光雷达和地图数据。初期客户主要为自动驾驶相关客户，Airbnb、Pinterest 和 OpenAI 等公司也在使用该平台。今年 4 月，Scale AI 完成 3.25 亿美元 E 轮融资，目前是全球估值最高的人工智能数据标注公司。

图数据

图数据领域值得关注的企业是世通亨奇。

世通亨奇成立于 2016 年 4 月该公司主要通过知识图谱、文本推理等技术对原始数据集进行提纯，主要的产品有 Plat-X 堇青、琥珀、紫晶、黑曜、情感标注辅助工具以及千河智能搜索，应用场景涉及国防、政企、医疗、零售、金融等领域。今年 11 月该公司获得近亿元融资。

数据工具

在数据工具研发方面值得关注的公司是英国 Rossum，成立于 2017 年 1 月，在今年 10 月份获得 1 亿美元的融资。该公司开发出基于云端的数据提取平台，用以理解半结构化的文档。

Hyperscience 成立于 2014 年，是一家 AI 数据处理解决方案提供商，致力于开发基于 AI 的企业软件，使办公流程自动化。该公司的智能文档处理解决方案有助于企业降低成本。

关于智源研究院

北京智源人工智能研究院（简称“智源研究院”）是落实“北京智源行动计划”的重要举措，是在科技部和北京市委市政府的指导和支持下，由北京市科委和海淀区政府于 2018 年 11 月推动成立的新型研发机构。智源研究院的愿景和目标是按照国家新一代人工智能发展规划总体部署，聚焦原始创新和核心技术，建立自由探索与目标导向相结合的科研体制。支持科学家勇闯人工智能科技前沿“无人区”，挑战最基础的问题和最关键的难题，推动人工智能理论、方法、工具、系统和应用取得变革性、颠覆性突破。营造全球最佳的学术和技术创新生态，推动北京成为全球人工智能学术思想、基础理论、顶尖人才、企业创新和发展政策的源头，率先成为国际领先的人工智能创新中心。推动人工智能产业发展和深度应用，改变人类社会生活，促进人类、环境和智能的可持续发展。

智源研究院积极打造智源社区平台，通过汇聚北京和全国科研、产业领域的人工智能创新人才，构建高度合作，紧密交流的科学研究社区，充分发挥成员的协同效应，推动人工智能的跨学科融合，在科学发现与技术研发循环的过程中，鼓励成员开展紧密协作。智源社区坚信，人工智能是深刻改变人类社会生活的颠覆式创新技术，需要依靠极具创造力的人才支持。我们希望围绕发展人工智能的共同愿景，吸引更多科研工作者、行业从业者和开源开发者加入，为我国人工智能的科研、产业发展贡献自身的力量！

智源研究院微信公众号

智源社区微信公众号

免责声明

本报告所含内容为一般性信息，不构成任何专业业务和财务业务的判断依据，同时本报告的信息来源于公开的资料，我们对该等信息的准确性、完整性或可靠性作尽可能的追求，但无法做任何保证和承诺，编者团队成员及所在单位不对任何因此报告导致的直接或间接损失或损害承担责任。

技术板块联系人：戴一鸣|智源社区分析师

邮　箱：ymdai@baai.ac.cn

产业板块联系人：李梦佳|智源社区主编

邮　箱：mjli@baai.ac.cn

联系我们

电话：（010）6893 3383
邮箱：press@baai.ac.cn
官网：https://www.baai.ac.cn/
地址：北京市海淀区成府路 150 号 智源大厦

智源研究院公众号