大模型发展趋势及国内外研究现状

◎ 撰文 | 熊子晗 李雨轩 陈军 陈大北

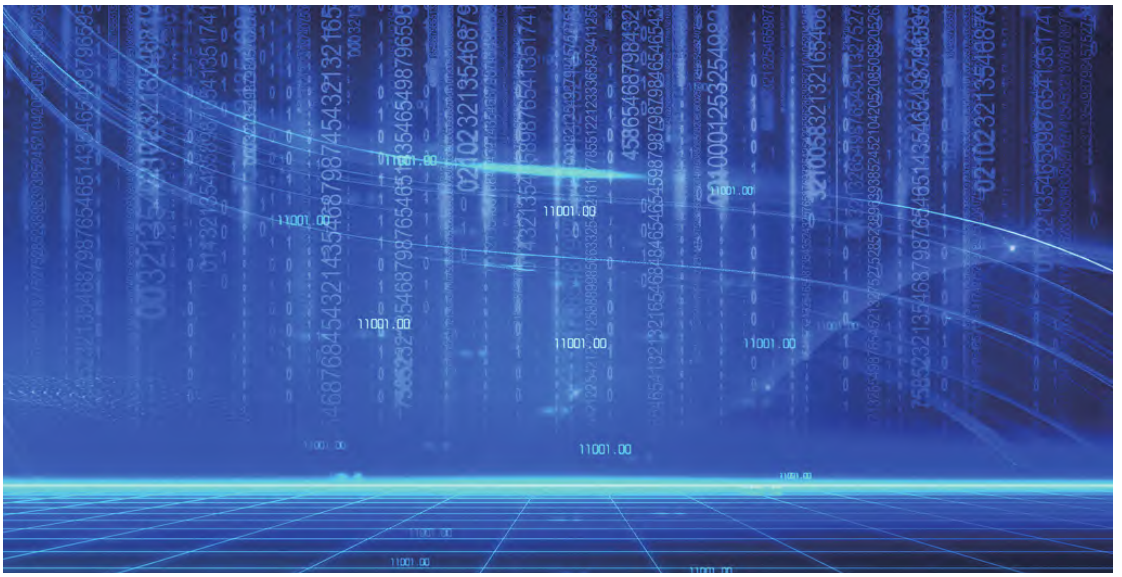
大模型是“大算力 + 强算法”结合的产物，通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于大模型进行应用开发时，将大模型进行微调，如在下游特定任务上的小规模有标注数据上进行二次训练，或者不进行微调，就可以完成多个应用场景的任务。

大模型发展趋势

华为认为，首先，大模型的出现和繁荣既是当前深度学习的顶峰，也代表着深度学习算法的瓶颈。对大模型的需求本质上是对大数据的需求。当前的人工智能算法尚无法高效地建模不同数据之间的关系，并以此解决模型泛化的问题，取而代之，通过收集并处理大量训练数据，人工智能算法能够通过“死记硬背”的方式一定程度上提升泛化能力。从这一角度看，大模型对数据的应用依然处于比较初级、低效的水平。可以预见，这种方式的边际效应明显，数据集越大模型越大，提升同等精度所需要的代价就越大。要想通过预训练大模型真正解决人工智能问题，看来不太现实。其次，除了在数据集构建、模型设计乃至评测标准方面持续演进，业界首先需要做的是抛弃预训练大模型“参数量至上”的评判标准。因此，参数量并不是评判模型能力的最好标准——如何用好参数并将模型的鲁棒性做得更好才是大模型发展真正应该关注的。

2020 年华为云预判 AI 发展有两大趋势：① AI 会从传统小模型发展到大模型，对应算力需求过去 10 年增加了 40 万倍。大模型成为应对 AI 应用碎片化的一种方式，可能收编高度定制化的小模型，导致市场向大公司集中，产业规则集格局也可能改变。② AI for Science（AI 赋能科研），AI 与科学计算交汇。包括传统的气象、海洋、农业、地球科学、航空航天等领域开始从偏微分方程的方法拓展到 AI 方法。

国际数据公司（IDC）认为，大模型的发展是大势所趋。首先，未来大小模型会协同进化，推动端侧化发展。大模型负责向小模型输出模型能力，小模型更精确地处理自己“擅长”的任务，再将应用中的数据与结果反哺给大模型，让大模型持续迭代更新，形成大小模型协同应用模式，达到降低能耗、提高整体模型精度的效果。其次，大模型通用性持续增强，实现 AI 开发“大一统”模式。大模型由于其泛化性、通用性为人工智能带来了新机遇。目前，在通用模型的基础上，各行业正利用精调或提示语 prompt 的方式加入任务间的差异化内容，从而极大地提高了模型的利用率，推动 AI 开发走向“统一”。最后，大模型从科研创新走向产业落地，通过开放的生态持续释放红利。大模型最重要的优势是推动 AI 进入大规模可复制的产业落地阶段，仅需零样本、小样本的学习就可以达到很好的效果，以此大大降低 AI 开发成本。国际数据公司建议各行业尽早拥抱大模型。在合作方面，主要关注大模型与自身业务的适配性以及与头部厂商联手打造行业标杆；在技术方面，建议大模型供应商持续探究大模型的生成可控性；在安全方面，大模型的技术安全以及伴随着大模型落地所带来的伦理问题是关注的重点；在商业化方面，大模型的路径仍不明确，海外市场发展较早，国内厂商可以重点借鉴。



国外大模型研究现状

OpenAI

OpenAI 于 2018 年 6 月公布了一个在诸多语言处理任务上都取得了很好结果的算法，即著名的 GPT。GPT 是第一个将 transformer 与无监督的预训练技术相结合的算法，其效果好于当前已知的算法。该算法是 OpenAI 大语言模型的探索性先驱，也使得后面出现了更强大的 GPT 系列。2019 年 2 月，OpenAI 官宣 GPT-2 模型，GPT-2 模型有 15 亿参数，基于 800 万网页数据训练，并于 2019 年 11 月发布 15 亿参数的完整版本的 GPT-2 预训练结果。2020 年 5 月，OpenAI 正式公布了与 GPT-3 相关的研究结果，其参数高达 1750 亿，这也是当时全球最大的预训练模型。2022 年 1 月，OpenAI 发布 InstructGPT，这是比 GPT 3 更好的遵循用户意图的语言模型。2022 年 3 月，OpenAI 新版本的 GPT-3 发布。

2022 年 11 月，OpenAI 发布 ChatGPT，这是一个 AI 对话系统，一款人工智能技术驱动的自然语言处理工具。它能够通过学习和理解人类的语言进行对话，还能根据聊天的上下文互动，真正像人类一样聊天交流，甚至能完成撰写邮件、视频脚本、文案、代码等任务。ChatGPT 经历了 OpenAI 开发的四代 GPT 模型的进化。此前的三代模型数据质量和数据规模不断提升，使得其生成能力不断精进，已能够执行阅读理解、机器翻译、自动问答等任务，但本质上只是语言模型，不具备回答问题的能力。针对 GPT-3，OpenAI 引入了 1750 亿训练参数，开启了超大模型时代。专家普遍认为，在封闭、静态和确定性环境中，该模型已可以达到人类的决策水平。而 ChatGPT 模型基于 GPT-4 优化引入了新的算法——从人类反馈中强化学习（RLHF），在训练中，训练师会对答案进行排序、打分或给出高质量答案，令 ChatGPT 具备一定的逻辑和常识，成为现阶段全球发布的功能最全面的 AI 模型，远超同类产品的智能化水平。

Google

Google 于 2019 年 10 月提出 T5，全称是 Text-to-Text Transfer Transformer，是谷歌研究人员在 2019 年提出的一个研究框架和预训练模型。当时谷歌研究人员已意识到基于未标注的大量文本数据训练大模型作为下游任务的基础是一种

十分高效的自然语言处理方法。该方法的主要目的是使模型开发通用能力和知识，然后将其转移到下游任务。但是快速发展的预训练模型让大家难以比较不同的方法。为此，谷歌提出将 NLP 领域的预训练任务当作一个 text-to-text 任务，然后基于这个框架研究 NLP 预训练模型。区别于之前的模型，由于谷歌将预训练任务当作一个 text-to-text 任务，因此不需要标注数据，也就不需要 BERT 那种模型，于是谷歌提出了 T5 模型，将 NLP 领域的问答系统、语言模型等任务都当作 Text-to-Text 任务。

2022年1月，Google推出Switch Transformer，声称能够训练包含超过 1 万亿个参数的语言模型的技术，直接将参数量从 GPT-3 的 1750 亿拉高到 1.6 万亿，速度是 Google 以前开发的最大语言模型（T5-XXL）的 4 倍。Switch Transformer 是一种“稀疏激活（sparsely activated）技术，仅使用了模型权重的子集，或是转换模型内输入数据的参数，即可达成相同的效果。

2022 年 10 月，Google 发布了基于 Pathways 架构、拥有 5400 亿参数的转换器语言模型 PaLM（Pathways Language Model）。研究人员称，PaLM 模型在语言理解等方面的评估测试表现十分出色，甚至还在语言和推理类的测评中超过了人类。2023 年 5 月，Google 正式发布新的通用大语言模型 PaLM2。PaLM2 是驱动 AI 机器人 Bard 的模型的升级版，可生成多种文本回应用户。谷歌称其可以使用 100 种语言，擅长数学、软件开发、语言翻译推理和自然语言生成。

BigScience

2022 年 3 月，BigScience 提出了 Bloom 模型。BigScience 是 HuggingFace 与法国国家科学研究中心（CNRS）下的两个高性能计算部门 GENCI 和 IDRIS 联合发起的项目，以开放科研研讨会（workshop）的形式组织。凭借其 1760 亿个参数，Bloom 能够生成 46 种自然语言和 13 种编程语言的文本。对于其中大部分语言模型，如西班牙语、法语和阿拉伯语，Bloom 将是第一个创建超过 1000 亿参数的语言模型。该项目由来自 70 多个国家和 250 多个机构的 1000 多名研究人员参与实施了一年时间，最终在法国巴黎南部的 Jean Zay 超级计算机上完成了 117 天 Bloom 模

型的训练。随着继续试验和调整模型，Bloom 的能力将继续提高。

Facebook/Meta

2022 年 6 月，Fackbook 的研究人员发布了开源的大语言模型 OPT（Open Pre-trained Transformer Language Models），参数规模达 1750 亿。从与 GPT-3 在 14 个任务的对比情况看，OPT 几乎与 GPT-3 的水平一致。但在运行时产生的碳足迹为 GPT-3 的 1/7。为了方便研究，Meta AI 公开了各种大小的 OPT 模型，从 125M 参数到 1750 亿参数的大小模型都有。

2023 年 2 月，Meta AI 发布大型语言模型 LLaMA，宣称可帮助研究人员降低生成式 AI 工具可能带来的“偏见、有毒评论、产生错误信息的可能性”等。Meta 声称其仅用约 1/10 的参数规模实现了匹敌 OpenAI GPT-3、DeepMind Chinchilla、谷歌 PaLM 等主流大模型的性能表现。Meta 目前提供有 70 亿、130 亿、330 亿和 650 亿四种参数规模的 LLaMA 模型。 根据论文，在一些基准测试中，仅有 130 亿参数的 LLaMA 模型性能表现超过了拥有 1750 亿参数的 GPT-3，而且能跑在单个 GPU 上；拥有 650 亿参数的 LLaMA 模型能够与拥有 700 亿参数的 Chinchilla、拥有 5400 亿参数的 PaLM“竞争”。

作为一个基础模型，LLaMA 不是为特定任务而设计的，Meta 研究人员通过标记一些 Tokens 等训练基础模型，从而更容易针对特定的潜在产品应用进行再训练和微调。不同于 Chinchilla、PaLM、GPT-3 等大模型，LLaMA 只使用公开可用的数据集进行训练，其中包括开放数据平台 Common Crawl、英文文档数据集 C4、代码平台 GitHub、维基百科、论文预印本平台 ArXiv 等。

微软 / 英伟达

微软于 2020 年 2 月发布的 TNLG（Turing Natural LanguageGeneration，图灵自然语言生成）是一款 170 亿参数语言模型，它在许多下游 NLP 任务上的表现超过了当时的顶尖水平，在各种语言建模基准测试上的成绩优于之前最顶尖的水平，并且在许多实际任务（包括摘要和问题解答等）上的表现也很出色。TNLG 是基于 Transformer 的生成式语言模型，它可以生成单词以完成开放式文本任务，可以生成直接回答问题和输入文档摘要。当模型越大、预训练数据越多样化和全面，它在泛化到多个下游任务时的表现也越好，即使有更少的训练示例。微软团队认为训练一个大型集中的多任务模型并共享其能力跨多个任务比为每个任务单独训练一个新模型更有效率。

DeepMind

2021 年 12 月，DeepMind 发布 Gopher 模型，该模型具有 2800 亿参数量。经过 152 个任务的评估，Gopher 比当时最先进的语言模型提高了大约 81% 的性能，特别是在知识密集领域，如事实检测和常识上。

2022 年 4 月， 由 DeepMind 在 Gopher 基础上研究 Chinchilla 模型，其训练数据量是 Gopher 的 4 倍，但参数数量仅是 Gopher 模型的四分之一（700 亿个）。在语言建模方面，在对 Chinchilla 和 Gopher 模型进行权威的 Pile 测评之后，结果表明，参数数量更少的 Chinchilla 在所有评估子集上的表现显著优于 Gopher。在大规模多任务语言理解（MMLU）方面，Chinchilla 模型在相关测试的结果也明显优于 Gopher。此外，Chinchilla 在 4 个不同的单独任务上达到了超过 90% 的准确率。此外，Chinchilla 在阅读理解、常识、闭卷问答、性别平等与有毒性语言、性别偏见等方面的测评结果也优于 Gopher。因此，尽管随着计算能力的增强，语言模型的规模可以做得越来越大，但 DeepMind 的分析表明，增加语言规模的大小需要更加关注训练数据集相应的缩放。DeepMind 指出，在对训练数据集进行扩展时，需要重点关注数据集的质量管理，尤其是其中的伦理和隐私等问题。

国内大模型研究现状

华为盘古

由华为云团队于 2021 年 4 月首次以盘古预训练大模型（简称盘古大模型）的名称对外发布，可用于 NLP、CV、多模态、科学计算以及图网络的大模型。2021 年 4 月华为发布了盘古 NLP 大模型、盘古视觉大模型、盘古科学计算大模型；2021 年 9 月，推出用于药物研发细分场景的大模型；2022 年，与能源集团合作发布了盘古矿山大模型、盘古气象大模型、盘古海浪大模型、盘古金融 OCR 大模型。对应到华为大模型赋能千行百业的层次，其基于底层一站式 AI 开发平台 ModelArts 建立了 L0 基础大模型、L1 行业大模型、L2 场景模型多层服务，通过系统化工程赋能行业。

在训练华为盘古大模型时使用到的参数有 1000 亿，数据超过 40TB，资源使用到鹏城云脑Ⅱ。视觉大模型典型的任务包括图像分类、物体检测、物体分割、物体追踪、姿态估计等。语音语义大模型典型的任务包括两个部分，即自然语言处理和语音处理，再细分任务为理解和生成。多模态大模型需要在海量的多模态数据上完成预训练，然后迁移到下游任务中，典型的多模态大模型包括跨模态检索、视觉问答、视觉定位等。科学计算大模型负责人类无法解决的问题，如湍流模拟、天气预报、大形变应力建模等。图网络大模型负责处理极度异质化的数据，通过图的形式对通用数据进行建模，以利用图结构表达数据元素间的相关性。截至目前，盘古大模型已经在很多领域落地实施，如 TFDS 图像自动识别、赋能智慧销售、一网统管事件工单分配、水泥生产系统的自动控制、炼焦系统的自动控制等。

腾讯混元

腾讯集团于 2022 年 4 月首次对外披露混元 AI 大模型（又称 HunYuan）的研发进展。混元集 CV（计算机视觉）、NLP（自然语言理解）、多模态理解能力于一体，先后在 MSR-VTT、MSVD 等五大权威数据集榜单中登顶，实现跨模态领域的“大满贯”。2022 年 5 月，在 CLUE（中文语言理解评测集合）三个榜单同时登顶，一举打破三项纪录。2022 年 8 月，混元又迎来全新进展，推出国内首个低成本、可落地的 NLP 万亿大模型，并再次登顶自然语言理解任务榜单 CLUE。

基于腾讯强大的底层算力和低成本高速网络基础设施，混元依托腾讯领先的太极机器学习平台推出的 HunYuan-NLP 1T 大模型，作为业界首个可在工业界海量业务场景直接落地应用的万亿 NLP 大模型，先后在热启动和课程学习、MoE 路由算法、模型结构、训练加速等方面研究优化，大幅降低了万亿大模型的训练成本。用千亿模型热启动，最快仅用 256 卡在一天内即可完成万亿参数大模型 HunYuan-NLP 1T 的训练，整体训练成本仅为直接冷启动训练万亿模型的 1/8。

HunYuan 协同了腾讯预训练研发力量，旨在打造业界领先的 AI 预训练大模型和解决方案，以统一的平台实现技术复用和业务降本，支持更多的场景和应用。当前 HunYuan 完整覆盖了 NLP 大模型、CV 大模型、多模态大模型、文生图大模型及众多行业 / 领域任务模型，HunYuan-NLP 1T 大模型已在腾讯多个核心业务场景落地，并带来显著的效果提升。

智源悟道

由北京智源人工智能研究院于 2020 年 10 月正式启动，为超大规模预训练模型研究项目，旨在以原始创新为基础实现预训练技术的突破，填补以中文为核心预训练大模型的空白，探索通向通用人工智能的实现路径。项目组由来自清华大学、北京大学、中国科学院计算技术研究所、中国人民大学等超过 100 位顶尖 AI 科学家组成，共同进行悟道

预训练模型的研发工作。2021 年 3 月，智源研究院发布了中国首个超大规模智能模型悟道 1.0，训练出中文、多模态、认知、蛋白质预测等系列模型。2021 年 6 月，悟道项目在北京智源大会上发布 2.0 版本科研成果，其中包括 1.75 万亿参数的全球最大通用预训练模型和其他一系列模型、算法、应用等，将中国预训练模型推向新高度。

悟道 2.0 模型的参数规模是 GPT-3 的 10 倍，打破了之前由 Google Switch Transformer 预训练模型创造的 1.6 万亿参数纪录，是截至当时中国首个、全球最大的万亿级模型。同时，悟道 2.0 模型是在中英双语共 4.9T 的高质量大规模清洗数据上进行的训练，训练数据包含 WuDaoCorpora 中的 1.2TB 中文文本数据、2.5TB 中文图文数据，以及 Pile 数据集的 1.2TB 英文文本数据。另外，悟道 2.0 模型一统文本与视觉两大阵地，支撑更多任务，更加通用化。为了促进预训练成果共享应用，悟道项目还将包括模型、算法、工具、API 和数据的系列科研成果在悟道官方平台进行开源开放。

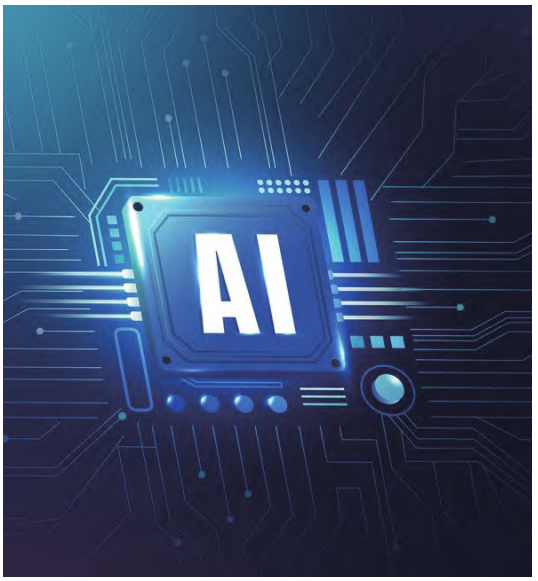
阿里巴巴

达摩院 M6。由阿里达摩院于 2020 年 1 月正式启动，于 2020 年 6 月公布 M6 3 亿参数基础模型，2021 年 1 月公布 M6 百亿参数多模态预训练模型，2021 年 5 月公布 M6 万亿参数模型，2021 年 10 月公布 M6 十万亿参数模型，为当时全球最大的预训练模型。M6 是中文社区最大的跨模态预训练模型，模型参数达到十万亿以上，具有强大的多模态表征能力。M6 通过将不同模态的信息经过统一加工处理，沉淀成知识表征，为各个行业场景提供语言理解、图像处理、知识表征等智能服务。训练数据为 1.9TB 文本和 292GB 图像。另外，为了应对模型扩展到千亿及以上参数超大规模时的多模态预训练模型快速迭代训练难题，达摩院在阿里云 PAI 自研 Whale 框架上搭建 MoE 模型，并通过更细粒度的 CPU offload 技术，最终实现将 10 万亿参数放进 512 张 GPU。（1）自研 Whale 分布式深度学习训练框架，针对数据并行、模型并行、流水并行、混合并行等多种并行模型进行统一架构设计，让用户在仅添加几行 API 调用的情况下就可以实现丰富的分布式并行策略。（2）在 Whale 架构中实施 Mixture-of-Experts（MoE）专家并行策略，在扩展模型容量、提升模型效果的基础上，不显著增加运算 Flops（每秒所执行的浮点运算次数），从而达到高效训练大规模模型的目的。（3）在自研的分布式框架 Whale 中通过更细粒度的 CPU offload，解决了有限资源放下极限规模的难题，并通过灵活选择 offload 的模型层进一步提高 GPU 利用率。

通义千问。通义千问（Tongyi Qianwen）是由阿里巴巴集团旗下的云端运算服务的科技公司阿里云开发的聊天机器人，能够与人交互、回答问题及协作创作。目前该模型主要定向邀请企业用户进行体验测试，仅允许获得邀请码的企业用户在官网加入体验，需要登录阿里云账号。2023 年 4 月 7 日，阿里巴巴集团旗下的云端运算服务公司阿里云正式宣布通义千问对已受邀的企业用户开启内测。2023 年 4 月 11 日，阿里巴巴董事局主席张勇在阿里云峰会上正式发布了大语言模型工具通义千问，并宣布此语言模型会接入阿里旗下的所有业务中。

百度 ERNIE

百度于 2019 年 3 月发布预训练模型 ERNIE1.0，2019 年 7 月发布 ERNIE2.0，2021 年 5 月开源四大预训练模型，包括多粒度语言知识模型 ERNIE-Gram、超长文本双向建模预训练模型 ERNIE-Doc、融合场景图知识的跨模态预训练模型 ERNIE-ViL 和语言与视觉一体的预训练模型 ERNIE-UNIMO，2021 年 12 月发布多语言预训练模型 ERNIE-M。百度持续投入大模型的技术创新与产业应用，布局了 NLP、CV、跨模态等大模型，率先提出行业大模型，构建大模型工具与平台，探索产品与社区，在企业端和用户端均有



不同程度的突破。训练参数为 100 亿。

文心 ERNIE 核心技术采用百度 NLP 自研的基于知识增强的语义理解技术，其创新性地将大数据预训练与多源丰富知识相结合，通过持续学习技术，不断吸收海量文本数据中词汇、结构、语义等方面的新知识，实现模型效果不断进化，显著提升了产品智能化水平。基于文心 ERNIE 核心技术，百度于 2023 年 2 月公开发布文心一言（ERNIE Bot）聊天机器人，这是百度全新一代知识增强大语言模型，能够与人对话互动、回答问题、协助创作，高效便捷地帮助人们获取信息、知识和灵感。文心一言基于飞桨深度学习平台和文心知识增强大模型，持续从海量数据和大规模知识中融合学习，具备知识增强、检索增强和对话增强的技术特色。目前已开放用户申请加入体验，但当前仅支持百度账号绑定中国大陆电话号码的企业级用户。

中科院自动化所紫东太初

其是中国科学院自动化研究所研发的跨模态通用人工智能平台，于 2021 年 7 月发布，是全球首个视觉 – 文本 – 语音三模态预训练模型，同时具备跨模态理解与跨模态生成能力，取得了预训练模型突破性进展。紫东太初跨模态通用人工智能平台以多模态大模型为核心，基于全栈国产化基础软硬件平台，可支撑全场景 AI 应用。多模态预训练模型被广泛认为是从限定领域的弱人工智能迈向通用人工智能路径的探索。

紫东太初兼具跨模态理解和生成能力。与单模态和图文两模态相比，其采用一个大模型就可以灵活支撑图 – 文 – 音全场景 AI 应用，具有在无监督情况下多任务联合学习并快速迁移到不同领域数据的强大能力。引入语音模态后的多模态预训练模型可实现共性图文音语义空间表征和利用，并突破性地直接实现三模态的统一表示。特别是首次使 “以图生音”和“以音生图”成为现实，对更广泛、更多样的下游任务提供模型基础支撑，实现 AI 在如视频配音、语音播报、标题摘要、海报创作等更多元场景的应用。

今年 5 月，中科院自动化所表示正在打造紫东太初 2.0 全模态大模型，该模型可实现文本、图片、语音、视频、3D 点云、传感信号等不同模态的统一表征和学习，助推通用人工智能时代加速到来。在文本、图片、音频、视频的基础上，紫东太初 2.0 可融入 3D、视频、传感信号等更多模态数据，并优化语音、视频和文本的融合认知以及常识计算等功能，进一步突破感知、认知和决策的交互屏障，让人工智能从感知世界进化为认知世界。

大模型将成为近年研究的热点，从产业价值的角度看，预训练大模型带来了一系列可能性，让产学研各界看到了由弱人工智能走向强人工智能，走向工业化、集成化、智能化的路径。在这样的驱动背景下，大模型会有一些可预见的走向，例如以往在模型专注方面都是各自做出模型后再集成耦合，而未来模型的多模态融合是必然趋势。另外，尽管参数量已达到百亿级别，但未来的研究也会关注不过分追求参数量且实现效果好的模型。©

作者单位：中国电信研究院（北京）