DOI:10.13811/j.cnki.eer.2023.06.005

多模态大模型的教育应用研究与展望

卢 宇[1]，余京蕾[2]，陈鹏鹤[1]，余胜泉[1]

（1.北京师范大学 未来教育高精尖创新中心，北京 100875；
2.北京师范大学 教育技术学院，北京 100875）

[摘 要] 多模态大模型逐渐成为人工智能领域研究的热点，目前已在通用领域有显著进展，但在教育领域仍处于起步阶段。文章提出可以构建教育领域通用大模型，并使其通过下游任务适配形成三类多模态教育大模型，从而形成三种典型教育应用，即教学资源自动生成、人机协同过程支持与教师教学智能辅助。在此基础上，文章以“多模态汉字学习系统”为例，利用多模态大模型实现跨模态释义生成，展示了多模态大模型在辅助语言学习方面的应用潜力。最后，文章针对教育领域通用大模型研究、多模态教育大模型的创新应用及其带来的潜在风险与可能触发的教育变革，提出针对性的建议与展望。

[关键词] 多模态大模型；人工智能教育应用；多模态汉字学习；教育大模型

[中图分类号] G434 [文献标志码] A

[作者简介] 卢宇（1982—），男，北京人。副教授，博士，主要从事人工智能及其教育应用研究。E-mail：luyu@bnu.edu.cn。

一、引 言

国务院《新一代人工智能发展规划》中提出，要充分利用人工智能等技术构建智能学习与交互式学习的新型教育体系[1]。人工智能技术也逐步被应用于教育环境建设、教学过程支持、教学精准评价与教育高效管理等关键环节与场景中。近年来，随着人工智能技术的快速演进，作为人工智能领域里程碑式的大模型被广泛应用于自然语言处理、计算机视觉、机器人等技术领域，其影响在各个行业逐步显现。

大模型又被称为基础模型（Foundation Model），指基于海量数据进行训练、具有超大规模参数且可以被应用于多种不同任务的人工智能模型[2]。大模型出现时间虽然不久，但已在多模态领域展现出卓越能力。本文将涵盖文本、音频、视频等多种模态的大模型称为多模态大模型。多模态大模型的相关研究源于自然语言处理领域的 Transformer 模型[3]。研究者基于这种具备高效计算能力与可扩展性的结构，逐渐扩大模型参数规模。谷歌于 2018 年发布了首个参数超过百万的单模态语言大模型 BERT[4]。其后，大模型的研究和应用进入快速发展阶段，模态也逐渐丰富。研究者开始基于海量文本与图像数据，构建图文模态融合的多模态大模型，实现跨模态理解与生成，如 Stable Diffusion[5]与 GPT-4[6]等。从技术角度看，多模态大模型属于基于深度神经网络的机器学习范畴，其理解、表达与学习能力相较于传统机器学习模型有显著提高，具备较强的通用性与泛化性。此类模型的训练数据量庞大，内部结构相对复杂且参数众多。例如：百度 ERNIE-ViLG 2.0 多模态大模型，其内部参数量已达到 240 亿，是目前最大的由文字生成图片类大模型，其训练过程需专业的分布式 GPU 集群完成[7]。

多模态大模型的研究和应用已在医疗、法律、金融、艺术等多个垂直领域取得显著进展，但在教育领域尚处于起步阶段，亟须相关基础性研究与应用型创新。当前，在教育领域，基于传统机器学习等算法模型的智能教育系统与平台，其智能性仍然难以充分满足

基金项目：北京市教育科学“十四五”规划 2021 年度重点课题“人工智能驱动的新一代智能导学系统构建研究”（课题编号：CHAA21036）

教师、学习者及教育管理者的实际需求。多模态大模型可以为解决这些技术瓶颈提供有效的途径与方法。

二、多模态大模型的构建与适配

多模态大模型的构建与适配可分为预训练与下游任务适配两个阶段,其基本过程如图 1 所示。其中,预训练阶段主要采用自监督学习方式,利用海量通用场景的多模态数据训练得到通用大模型。在构建的通用大模型基础上,下游任务适配阶段针对不同的具体任务,设计可直接应用在不同场景下的专用大模型。下游任务适配阶段的核心思想是迁移学习,其基本理念是将在先前任务或领域中学到的知识或经验,应用到新的任务或领域中。迁移学习可以实现基于相对较少的数据进行轻量且高效的下游任务适配,从而满足多种应用场景的需求。

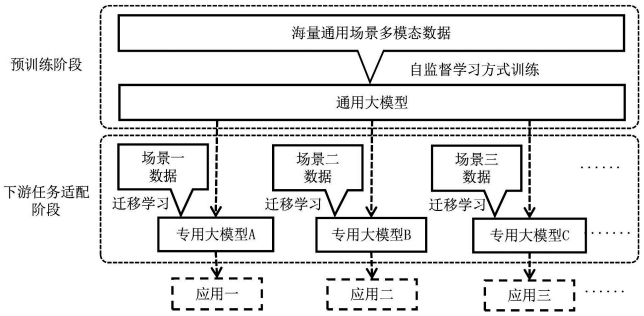

图 1　多模态大模型的构建与适配过程

通用大模型的预训练可以利用文本、图像、视频、音频等多种类型数据,这些数据可以来源于互联网等通用领域,也可以来源于多个专业领域。不同于传统的机器学习,通用大模型的训练更加青睐大量级与多模态数据。依据数据的不同模态,模型可采用不同结构的深度神经网络,以自监督学习方式对其网络内部参数进行持续调整和优化,直至完成预训练过程。基于上述方式得到的通用大模型具备通用领域知识,在通用场景中可以使用,但较难在特定场景和下游任务中应用及展现高性能。因此,通用大模型需要针对特定场景和任务,基于迁移学习的思想,学习下游任务中更深层次、更有价值的隐含规律与模式。

当前,下游任务适配可采用多种方法,包括微调方法(Fine-tuning)、提示学习方法(Prompt-based Learning)[8]和上下文学习方法(In-context Learning)[9]等。微调方法利用下游任务数据,对通用大模型整体参数进行再次训练,从而提升模型在下游任务中的适用能力。提示学习方法通过人工设计或自动生成离散或连续的提示模板,修改下游任务数据输入与输出的表达形式,对模型的局部参数进行调整,从而尽可能利用模型原有性能适配下游任务。上下文学习方法则充分利用模型自身的类比学习能力,仅利用少量下游任务提示示例或上下文提示信息与指令语句,直接对通用大模型进行适配,从而节省因调整模型参数带来的算力消耗。因此,上下文学习方法更为高效便捷,可以利用小样本、单一样本甚至零样本进行下游任务适配。

以语言大模型 GPT-3[10]为例,简述其构建与适配过程。GPT-3 采用自回归架构,其预训练数据由多个文本数据集组成,包括约一万亿文字量的网络爬虫数据集以及多个高质量图书、百科类文本数据集。GPT-3 采用自监督学习训练方式,对海量无标注数据进行训练,从而得到具备一定语言通用理解能力的大模型。当需要完成特定下游任务时,可以采用上下文学习方法进行模型适配。例如:针对英译法的下游翻译任务,可以设置提示指令为“将英语翻译为法语”,并设置提示示例为英文单词到法语单词的转换,如“Hello => Bonjour”。当 GPT-3 学到该翻译任务后,在实际应用中输入提示信息,已经适配好的模型便可以输出对应的法语单词,从而完成针对翻译任务的模型适配。

三、多模态大模型在教育中的应用

面向教育领域的多类迫切需求,可以首先构建教育领域通用大模型,并使其通过下游任务适配形成三类多模态教育大模型,从而形成其在教育领域的三类典型应用,即教学资源自动生成、人机协同过程支持与教师教学智能辅助。具体而言,如图 2 所示,首先,采集通用领域与教育领域的多模态海量数据与知识,作为构建教育领域通用大模型的信息基础,包括但不限于课堂音视频与作业试卷等教学场景数据,慕课与论坛等互联网数据,以及教学理论与学科知识。在此基础上,依据不同模态与模型间的相互组合构建模型框架,利用自监督式学习方式进行预训练,从而得到教育领域通用大模型。与传统模型相比,教育领域通用大模型可以充分利用其自身性能优势,深入理解教学资源、教学对象与教学过程三个教育要素。其中,模型需要重点理解教育资源的属性、关联与语义信息,教学对象的行为、语言与意图,以及教学过程的互动、活动与目标等。具备以上通用能力的大模型,可以为不同的下游教育任务进行适配,形成针对不同典型应用的三类多模态大模型,并分别为教学平台与系统、线上线下学习者、教师与教育管理者提供服务和支持。

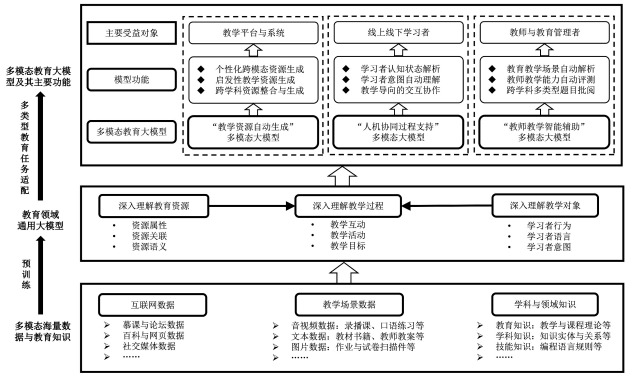

图 2　教育领域大模型构建及其多类型教育任务适配

(一)教学资源自动生成

基于学科资源特征与学习者学习风格等信息,可以实现个性化的资源推送,或利用知识图谱与智能获取的学习场景信息,通过动态提取网络资源,为学习者提供情境化学习资源[11]。然而,现有技术在资源自动构建和生成方面仍存在较多局限。首先,现有教学资源由检索机制得到,无法实现新颖独特的教学资源的自动生成与多模态转换;其次,学习者受既定推荐机制的限制,无法灵活自主地创造个性化的教学资源;最后,现有资源生成模型通用性差,难以利用单个模型实现跨学科的知识整合与资源生成。在现有多模态大模型的基础上,“教学资源自动生成”多模态大模型可望在内容自动生成方面的功能与性能不断取得进展,尤其在个性化跨模态资源生成、启发性教学资源生成、跨学科资源整合与生成等方面,突破和解决现有的局限与问题。

在教学资源自动生成方面,当前通用领域的多模态大模型已展现出一定的能力。Stable Diffusion 等图像生成模型,可以依据教学需求输入主体及其细节的文本描述,快速自动生成多种风格、高清逼真、蕴含美感的美育类教学资源,所生成的教学资源既具备显著的跨模态性,又具有新颖性与独特性。Open AI 提出 MuseNet 音频生成模型[12],可依据个人偏好输入部分音符及所期望的音乐风格,自动生成长达四分钟的音乐片段,支持多达十种乐器的呈现,并支持古典、乡村甚至披头士等多种音乐风格的融合。谷歌提出的 MusicLM 音频生成模型[13],可直接基于自然语言描述生成高质量的音乐片段,文本描述如“演奏一段平缓的小提琴曲并以吉他旋律为伴奏”。DeepMind 与斯坦福大学提出 Dramatron 文本生成模型[14],可生成包括标题、人物、故事节奏、地点描述与对话在内的连贯剧本内容,用于协助从业者共创戏剧或电影剧本。此类多模态大模型生成的创造性艺术资源,可作为智能化知识建构工具,在学习者与资源的双向交互过程中,帮助其探索与理解不同风格的美术、音乐、戏剧作品,启发其进行绘画、音乐与剧本创作。此外,在学科题目资源生成方面,美国莱斯大学等提出基于 GPT-3 生成多学科的高质量题目。用户可以基于学科需求输入指定科目及题目提示,模型即可生成能直接应用于教学的题目内容[15]。

(二)人机协同过程支持

随着智能技术的快速发展,人机协同学习逐渐成为教学活动开展的重要形式和组成部分,但仍然受限于人机交互的自然程度与专业化程度。现有的智能教育系统或平台大多缺乏准确理解学习者提问与意图的功能,也难以像人类教师一样用自然语言与学习者开展连贯的交流、问答与教学,因此,难以真正实现人机协同的学习过程。基于多模态大模型在跨模态信息理解与人机对话等方面的能力,教育领域可以进一步构建“人机协同过程支持”多模态大模型,重点关注学习者认知状态解析、学习者意图自动理解、教学导向的交互协作,以期真正实现人机协同的高效率学习过程。

在人机协同过程支持方面,当前通用领域的多模态大模型也已展现出良好的潜力。在知识问答方面,百度提出的 ERNIE 大模型[16]可以对领域实体知识与专业术语进行知识增强,并利用问答匹配任务进行模型训练,从而深入理解领域知识及其内在联系。此类模型可通过进一步增强教学与德育知识学习,在人机协同学习过程中,支持专业化学科知识点答疑与智能化育人咨询。在编程学习方面,OpenAI 等基于 GPT-3 针对计算机编程任务进行下游适配,开发 Codex 模型[17]。该模型可将自然语言描述的内容直接转化为计算机编程语言,且转换的语言种类涵盖 Python 等多种主要编程语言。基于该模型开发的编程语言学习助手 GitHub Copilot,已可以支持人机协同的编程语言学习。此外,由 OpenAI 提出的 GPT-4 多模态大模型具有里程碑式的多模态理解、推理、内容生成与问题解决能力。该类模型可作为百科全书为学习者提供便捷的资源检索,作为写作助手为学习者提供文章润色、思路启发等写作辅助服务,作为私人助教为不同学业水平的学生提供个性化辅导、引导式解决多学科的疑难问题,作为编程助手辅助学习者理解、修正和生成示例代码等。

(三)教师教学智能辅助

现有人工智能技术难以直接替代人类教师进行教学,但可以作为 AI 代理辅助教师完成部分机械重复的工作[18]。当前的通用领域多模态大模型已经具备较强的问题解决能力,可以为教师在课堂教学与备课

中提供辅助支持。在此基础上,教育领域可以进一步构建“教师教学智能辅助”多模态大模型,拓展人工智能技术辅助教师教学的范围和能力,尤其在教育教学场景自动解析、教师教学能力自动评测、跨学科多类型题目自动批阅等方面,多模态大模型可以发挥重要作用。

在利用大模型开展教师教学智能辅助方面,当前工业界和学术界也已开始进行积极的探索。好未来基于教师线上教学语音转写产生的约 2000 万条教育文本数据,构建了在线教学大模型TAL-EduBERT[19]。经过下游任务适配,该模型可以通过教师语言对其中细颗粒度的教学行为进行识别,类别包括“引导学生课后总结”“带领学生记笔记”“表扬学生”“提问学生”,从而在教师的教学反思与教学改进过程中提供有力的证据支持。MathBERT[20]基于 BERT,从多下游任务与多学段数学知识两个方面进一步训练和适配模型,从而对数学领域知识进行深入语义理解和知识融合,辅助教师进行自动批阅、题目知识点标注等具体工作。孟菲斯大学团队提出可以利用 T5 语言大模型[21]评估完形填空题目的难度及可读性等级,从而辅助教师自动评测学习者的阅读能力[22]。此外,研究者正在积极探索和建立具有更强逻辑推理能力的多模态大模型,从而自动解决物理、生物与数学等学科的定量科学问题。例如:哈佛大学与麻省理工学院联合研究团队基于 Codex 模型,将概率题目文本转换为计算机程序,并通过执行程序自动解决一系列概率与统计学问题,其准确率与人类表现相当[23]。GPT-4 凭借其多模态理解能力,可直接基于试卷图片及提示指令自动解答问题,并给出详细的解题步骤。微软团队在其评测报告中指出,GPT-4 可以解决数学、编程等学科中新颖且难度较大的任务,性能可接近人类水平[24]。

四、多模态大模型的教育应用案例

基于本团队的近期研究成果——“多模态汉字学习系统”,介绍将多模态大模型应用于教学资源自动生成的典型案例。该案例将多模态大模型应用于汉字的跨模态释义生成,体现了其在辅助语言学习方面的应用潜力。

(一)多模态汉字学习

汉字学习是汉语学习中一项重要的内容。字典是汉字学习过程中的有效工具,可查询汉字释义及组词等信息。但无论纸质字典或电子字典,往往只能提供单一模态的信息呈现。而在多模态信息呈现方面,研究者认为图片可以很好地表达复杂、抽象的场景或不常见、不熟悉的事物。由心理学家佩维奥提出的双重编码理论也强调了语言与视觉信息同时出现的重要性,且视觉信息比语言信息更易于记忆。因此,设计多模态信息辅助的汉字学习系统,将汉字与其对应的图片结合学习,将有助于辅助学习者记忆字义,提高汉字学习效果[25]。

(二)系统设计

“多模态汉字学习系统”的核心部分为跨模态释义生成模块,该模块可采用两种多模态大模型分别实现图文检索与图文生成功能。系统的基本框架与工作流程如图 3 所示。

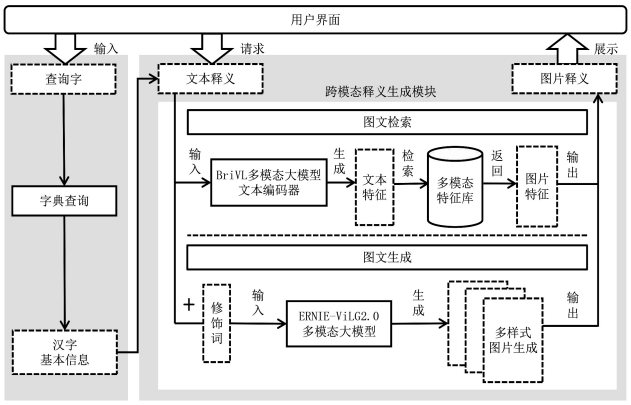

图 3 “多模态汉字学习系统”基本架构

1. 基于多模态大模型的跨模态图文检索

系统中跨模态释义生成模块的图文检索功能,采用文澜 BriVL 多模态大模型[26]加以实现。该模型参数量超过 10 亿个,通过特定接口实现云端计算和调用。文澜大模型将网络爬取的海量图文信息作为预训练数据,搭建文本编码器与图片编码器两个分支的双塔模型结构,并利用损失函数进行虚拟连接。

在预训练过程中,该模型基于“对比学习”算法框架[27],分别输入图文正负样本,以自监督学习方式训练文本与图片编码器,分别抽取图文特征并映射到同一多模态空间中。由于数据来自网络爬取而非人工标注,图文对应关系仅为弱相关,即文本信息不仅是图片具体内容的描述,更可能是图片背后的抽象释义。相比日常仅基于关键词的图文检索,该模型学到的图文关系更符合本场景需求,适用于找寻抽象文字描述对应的图片释义。在预训练结束后,双塔结构中的文本与图片编码器可拆分使用,支持离线的图文特征抽取与特征库构建。该模型在数据集 AIC-ICC 文本检索图片任务中,检索结果前十张为目标图片的概率为 65.26%。虽然模型在性能上仍有提升空间,但其基于图文弱相关的特征检索功能,可以为解释抽象文本提供图片支持,突破了现有依据关键词检索的局限,为构建

多模态、多语种语言学习系统提供了解决方案。

在使用系统时，用户可逐个点击文本释义，得到对应的图片释义。在具体实现中，系统首先将文本释义中的描述与各个组词切分为短语，然后利用文本编码器提取各候选短语文本特征，从而在跨模态特征库中检索与之最相近的图片特征，并将对应图片作为该文本释义的图片释义展示给学习者。其中，跨模态特征库为离线构建，图片特征由图片编码器抽取，并提前存储在跨模态特征库中。

2. 基于多模态大模型的跨模态图文生成

系统中跨模态释义生成模块的图文生成功能，采用 ERNIE-ViLG 2.0 多模态大模型加以实现。ERNIE-ViLG 2.0 基于扩散模型[28]，可以进一步增强对文本关键内容及图片关键区域的理解，从而提升图片的生成质量。目前，该模型在权威数据集 MS-COCO 文本生成图片任务上取得最好成绩，并在图片逼真度与图文一致性指标上以大比分超越同类模型。在模型应用过程中，用于下游任务适配的文本提示信息决定了生成图片的效果。提示信息可由生成内容的主体描述、细节描述及修饰词构成，其中，修饰词可以为艺术风格、艺术家、摄影词汇等。本系统中，根据汉字及其释义进行提示信息的构建。

在使用系统时，学习者同样可以逐个点击文本释义，得到模型生成的图片释义。与基于图文检索的功能不同，对于相同的提示信息，系统可生成多个不同样式的图片，并鼓励群体学习者为各自认为最匹配的图片点赞。系统将图片按照点赞量进行排序展示，新登录的学习者可以看到点赞量最高的图片。如果学习者对现有图片释义均不满意，可选择生成新图片并弹窗确认，否则将继续更换图片。在此基础上，系统将进一步设计开发学习者评论留言功能，在互动中促进协作学习与知识建构。

综上所述，多模态大模型可以针对包括汉字在内的多种文字进行跨模态图片释义的检索与生成，为抽象的文本描述提供直观的图片解释，从而辅助学生进行语言学习。此外，多模态大模型在语言学习中仍有很多潜在辅助功能待开发，如依据语言学习需求灵活检索或生成音频与视频资源、自主创设学习情境开展对话练习、提供多模态句子解释以辅助阅读理解等。

五、建议与展望

当前多模态大模型正处在快速演进和落地应用时期，本文提出以下建议和展望：

（一）推进教育领域通用大模型的深入研究与构建

目前，多模态大模型的构建研究多专注在通用或特定垂直领域，建议积极开展教育领域通用大模型研究和构建。此类模型可以充分利用教育领域的海量多模态与长周期数据，对学习者认知过程与教学交互过程等进行准确捕捉与深度理解，尝试利用模型输出帮助认知科学与学习科学更好地理解教学过程及其底层机制，并在此基础上构建和适配可用于多种类型教育任务的专用教育大模型和智能教育服务系统。

（二）拓展现有多模态大模型在教育中的创新性应用

目前，以 GPT-4 为代表的多模态大模型已逐渐展现出其在多领域与多任务上的泛化能力。建议充分利用人工智能领域已建立的多模态大模型及其各项能力，结合教育场景与教学需求，进行下游教育任务适配与创新应用，解决教育领域的实际问题。在上述的教学资源自动生成、人机协同过程支持与教师教学智能辅助的基础上，还应继续探索和解决其他典型教育任务，积极尝试组合使用多种模型，发挥各自的技术优势，促进教育领域的创新。例如：以多模态大模型作为控制器，自动解析教育任务或教师指令，选择并调用所需的技术或教育模型，解决典型教育场景下的复杂任务。

（三）重视多模态大模型可能带来的潜在风险

多模态大模型主要由海量无标注数据训练并构建，难以避免在资源生成等过程中存在数据偏见、知识产权、知识与计算准确性等原生性问题。例如：Stack Overflow 程序论坛已公开表示，由于大模型生成内容的准确性难以判定，将暂时禁止用户使用该模型生成内容作为论坛回答[29]。因此，在应用于教育领域时，需要从科学性、公平性、准确性与价值观等多个维度进行风险筛查。同时，在下游任务适配与应用过程中，需要教师或教育管理者监管，尤其是在人机协同与教学智能辅助方面，需对模型的使用范围有明确的功能限定，避免影响学习者的独立思考与认知过程。ChatGPT[30]等大模型一经推出便引起教育领域的广泛关注：学生可借助其完成文章或代码代写等作弊行为，且普通教师无法辨别，这直接影响了传统教育教学过程与制度。在学术界，为维护学术严谨性及作者责任制原则，《自然》等高水平期刊明确禁止大模型作为文章作者，如借助模型生成内容需特别注明[31]。

（四）拥抱多模态大模型触发的教育变革

多模态大模型对未来社会的影响已不可避免，相当一部分行业可能被以此类模型为代表的人工智能

技术冲击甚至取代,人才培养的需求也会由此发生根本变化。因此,教育领域需要积极适应这种变化,拥抱新技术所触发的教育变革。面对多模态大模型给教育带来的机遇与挑战,需要积极从教育治理、教学过程与教育评价等多个维度进行应对。教师使用高交互性人工智能工具开展教学,学习者使用高辅助性人工智能工具开展学习,应该会成为未来教育不可或缺的一部分。教育本身则更应该不断革新其理念和方式,重视培养学生的创造性、批判性与人机协作能力,从而满足未来智能化社会的需求。

[参考文献]

[1] 国务院.国务院关于印发新一代人工智能发展规划的通知[EB/OL].(2017-07-20)[2023-04-12]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[2] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models [J]. arXiv preprint,2021(1):1-212.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017,30:1-15.

[4] DEVLIN J, CHANG M, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics,2019:4171-4186.

[5] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society,2022:10684-10695.

[6] OpenAI. GPT-4[EB/OL]. (2023-03-14) [2023-04-12]. https://openai.com/research/gpt-4.

[7] FENG Z, ZHANG Z, YU X, et al. ERNIE-ViLG 2.0: improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts[J]. arXiv preprint,2022(1):1-19.

[8] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM computing surveys,2023,55(9):1-35.

[9] DONG Q, LI L, DAI D, et al. A survey for in-context learning[J]. arXiv preprint, 2022(1):1-16.

[10] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [J]. Advances in neural information processing systems,2020,33:1877-1901.

[11] 赵刚,初洁,朱文娟,尹江华,杨丽君,蒋姝凡,吴林静.基于知识图谱的户外动态学习资源智能生成与服务模型研究[J].电化教育研究,2022,43(4):55-62.

[12] OpenAI. MuseNet[EB/OL].(2019-04-25)[2023-04-12].http://openai.com/blog/musenet.

[13] AGOSTINELLI A, DENK T, BORSOS Z, et al. MusicLM: generating music from text[J]. arXiv preprint, 2023(1):1-15.

[14] MIROWSKI P, MATHEWSON K, PITTMAN J, et al. Co-writing screenplays and theatre scripts with language models: an evaluation by industry professionals[J]. arXiv preprint, 2022(1):1-102.

[15] WANG Z, VALDEZ J, BASU MALLICK D, et al. Towards human-like educational question generation with large language models[C]//Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27-31,2022, Proceedings, Part I. Berlin: Springer International Publishing, 2022:153-166.

[16] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics,2019:1441-1451.

[17] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[J]. arXiv preprint,2021(1):1-35.

[18] 余胜泉,王琦."AI+教师"的协作路径发展分析[J].电化教育研究,2019,40(4):14-22.

[19] 好未来.好未来开源教育领域首个在线教学中文预训练模型 TAL-EduBERT[CP/OL].(2021-01-25)[2023-04-12]. https://github.com/tal-tech/edu-bert.

[20] SHEN J T, YAMASHITA M, PRIHAR E, et al. MathBERT: a pre-trained language model for general NLP tasks in mathematics education[C]//NeurIPS 2021 Math AI for Education Workshop. Cambridge, Mass: MIT Press,2021:1-10.

[21] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The journal of machine learning research, 2020,21(1):5485-5551.

[22] OLNEY A M. Assessing readability by filling cloze items with transformers [C]// Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27-31,2022, Proceedings, Part I. Berlin: Springer International Publishing, 2022:307-318.

[23] TANG L, KE E, SINGH N, et al. Solving probability and statistics problems by probabilistic program synthesis at human level and predicting solvability [C]//Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II. Berlin: Springer International Publishing, 2022: 612-615.

[24] BUECK S, CHANDRASEKARAN V, ELDANl R, et al. Sparks of artificial general intelligence: early experiments with gpt-4[J]. arXiv preprint, 2023(1):1-155.

[25] YU J, SONG J, CHEN P, et al. An intelligent multimodal dictionary for Chinese character learning [C]// Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II. Berlin: Springer International Publishing, 2022:79-83.

[26] FEI N, LU Z, GAO Y, et al. Towards artificial general intelligence via a multimodal foundation model [J]. Nature communications, 2022,13:1-13.

[27] HE K, FAN H, WU Y, et al. Momentum contrast for unsupervised visual representation learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society,2020:9729-9738.

[28] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models [J]. Advances in neural information processing systems, 2020,33: 6840-6851.

[29] Stack Overflow. Temporary policy: ChatGPT is banned [EB/OL]. (2023-04-12)[2023-04-26].https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned.

[30] OpenAI. ChatGPT: optimizing language models for dialogue [EB/OL]. (2022-11-30) [2023-04-12]. https://openai.com/blog/chatgpt/.

[31] Nature. Initial submission [EB/OL]. [2023-04-26].https://www.nature.com/nature/for-authors/initial-submission.

Study and Prospect of the Applications of Large Multimodal Models in Education

LU Yu[1], YU Jinglei[2], CHEN Penghe[1], YU Shengquan[1]

(1.Advanced Innovation Center for Future Education, Beijing Normal University, Beijing 100875;
2.School of Educational Technology, Beijing Normal University, Beijing 100875)

[Abstract] Large multimodal models have gradually become a hot topic of research in artificial intelligence, and have significant progress in the general field. But they are still in the initial stage in the education field. This paper proposes to build a general large model in education and adapt it to three types of educational large multimodal models through downstream tasks, which constitutes three typical applications in education, namely, automatic generation of learning resources, human-AI collaboration, and intelligent teacher teaching assistance. Based on that, this paper takes "multimodal Chinese character learning system" as an example and uses large multimodal model to realize cross-modal interpretation generation, demonstrating the potentials of large multimodal model in assisting language learning. Finally, this paper proposes suggestions and prospects on the research of general large models in education, the innovative applications of educational large multimodal models, and the potential risks and changes in education that may be triggered by them, respectively.

[Keywords] Large Multimodal Model; Artificial Intelligence Applications in Education; Multimodal Chinese Character Learning; Large Models in Education