·专题:ChatGPT 与人工智能技术应用·

从 ChatGPT 到多模态大模型:现状与未来

李 耕[1,2] 王梓烁[1,2] 何相腾[1,2] 彭宇新[1,2*]

1. 北京大学 王选计算机研究所,北京 100871
2. 北京大学 多媒体信息处理全国重点实验室,北京 100871

[摘 要] 2022 年底,OpenAI 发布的 ChatGPT 聊天机器人将人工智能对通用自然语言任务的理解与生成能力提升到新的高度,引发各界广泛关注。当前 ChatGPT 仅支持文本模态的交互,而真实世界的感知则依赖于图像、文本、视频、音频等多个模态的协同处理。如何借鉴人脑的跨模态处理特性,跨越视觉、语言、听觉等不同感官信息实现对真实世界的感知和认知,是提升模型通用感知和交互能力、实现通用人工智能的关键。本文从 ChatGPT 的核心技术出发,分析 ChatGPT 在文本单模态限制下所面临的问题,并介绍 ChatGPT 与多模态分析技术结合的部分代表性工作,最后从多模态预训练、数据—知识双轮驱动等角度对 ChatGPT 多模态化的未来研究方向进行展望。

[关键词] ChatGPT;多模态分析;大语言模型;通用人工智能;多模态预训练

2022 年 11 月 30 日,OpenAI 发布了全新对话式通用人工智能工具 ChatGPT①,仅 5 天活跃用户数就高达 100 万,2 个月活跃用户数达到了 1 亿,成为历史上增长最快的消费者应用程序。ChatGPT 表现出惊人的语言理解、生成和知识推理能力,能够有效地理解用户意图,通过多轮对话提供内容完整、重点清晰、概括明了、逻辑有序、条理分明的文本回答。其关键技术大规模语言模型(Large Language Model,LLM),不仅在多项自然语言处理任务中达到了目前最先进水平,而且可迁移应用于计算机视觉、多模态分析等其他领域,正快速成为推动社会和经济发展的重要技术之一。

人类感知真实世界的方式是通过对视觉、语言、听觉等多模态信息进行综合分析与理解[1]。然而,目前以 ChatGPT 为代表的大规模语言模型技术多局限于文本单模态,限制了模型对真实世界的综合感知能力。因此,如何将 ChatGPT 与多模态技术结合,是当前面临的重要问题之一。本文主要介绍了 ChatGPT 核心技术、单模态局限性、多模态大模型和 ChatGPT 多模态化技术,并对未来研究方向进行了展望。

彭宇新 北京大学二级教授、博雅特聘教授,研究跨媒体分析、计算机视觉、人工智能。国家杰出青年科学基金获得者、863 项目首席专家、中国人工智能产业创新联盟专家委员会主任、中国工程院“人工智能 2.0”规划专家委员会专家、中国电子学会会士、中国人工智能学会会士、中国图象图形学学会会士、副秘书长。以第一完成人获 2016 年北京市科学技术奖一等奖和 2020 年中国电子学会科技进步奖一等奖。发表论文 200 多篇,包括 ACM/IEEE 汇刊和 CCF A 类论文 100 多篇。担任 *IEEE Transactions on Multimedia*、*IEEE Transactions on Circuits and Systems for Video Technology* 等期刊编委。

李耕 北京大学在读博士生,主要研究方向为跨媒体分析、计算机视觉。

1 ChatGPT 核心技术及主要问题

1.1 ChatGPT 模型训练

ChatGPT 训练的核心技术主要包括:(1) 预训

收稿日期:2023-07-18;修回日期:2023-09-29

* 通信作者,Email: pengyuxin@pku.edu.cn

本文受到国家自然科学基金项目(61925201, 62132001, 62272013)的资助。

① https://openai.com/blog/chatgpt

练语言模型；(2) 有监督微调；(3) 基于人类反馈的强化学习(Reinforcement Learning from Human Feedback，RLHF)。首先，通过自监督预训练使语言模型从大规模语料库中学习语言规律，具备基础理解和生成能力；然后，通过构造指令微调数据集并对模型进行有监督微调，提升模型对人类意图的理解能力，从而使模型按要求执行多种任务；最后，通过基于 RLHF，根据人类偏好进一步提升模型性能。

1.1.1　预训练语言模型

ChatGPT 基于 OpenAI 的 GPT3.5 自回归大规模语言模型，通过在大规模无标注语料数据上进行自监督预训练，使模型具备基础的语言理解和生成能力。目前主流自然语言预训练任务包括自回归语言建模(Auto-regressive Language Modeling)、掩码语言建模(Masked Language Modeling)、下一句预测(Next Sentence Prediction)、句子顺序预测(Sentence Order Prediction)等，如图 1 所示。代表性工作为 GPT 系列模型(包括 GPT-1[2]、GPT-2[3]、GPT-3[4]和 ChatGPT 等)和 BERT[5]。GPT 系列采用自回归语言建模预训练，即根据语料中前$(i-1)$个单词预测第 i 个单词。自回归任务天然符合生成式任务的特点，因此 GPT 系列模型具有较强的文本生成能力。BERT 采用掩码语言建模和下一句预测任务进行预训练。掩码语言建模任务主要内容是通过随机掩盖文本中的部分词语，使模型还原原始文本的内容。下一句预测任务的主要内容则是使模型预测符合上下文的下一句文本，提升了模型句子层面的理解能力。相比 GPT 系列，ChatGPT 更贴近生成式的自回归预训练任务，BERT 使用双向预测预训练任务，能够学习到综合上下文的文本表示，具有更高的文本理解能力，但在生成能力上表现一般。

1.1.2　有监督微调

预训练结束后，当前模型具有较强的语言生成能力，但难以理解人类输入不同指令的意图。因此，通过有监督微调(Supervised Fine-tuning，SFT)引导模型按照人类意图进行答案生成。具体而言，首先选取部分输入 Prompt，并人工为其构造符合人类意图的高质量答案。然后以上述数据中的〈Prompt，Answer〉模板对预训练模型进行精调，使模型初步具有理解人类意图、生成高质量答案的能力。

1.1.3　基于人类反馈的强化学习

最后，ChatGPT 采用基于 RLHF，根据人类偏好进行强化学习微调。RLHF 在 ChatGPT、GPT-4[6]、Claude[7]等模型中均有应用。为实现对人类反馈信息的利用，RLHF 使用大规模人类偏好数据训练可学习的奖励模型(Reward Model)。在预训练生成模型中，为保障奖励模型和生成模型具备相近的理解能力，现有方法[8, 9]的奖励与生成模型通常采用相同架构，但不固定模型规模。奖励模型在训练过程中，为避免不同人工评分导致的个体标准差异，通过对多种决策进行排序评估，鼓励模型生成排序靠前的答案。RLHF 能够有效利用人类反馈提升模型性能，大幅度提高了 ChatGPT 在各类自然语言理解生成任务中的用户体验。

1.2　ChatGPT 模型应用

经过上述训练后的 ChatGPT 模型已具备理解人类意图并生成高质量答案的能力，在实际应用中，可通过上下文学习[4]、思维链推理[10]等技术，进一步提高 ChatGPT 的回答质量。

上下文学习(In-context Learning)是指收集特定任务的部分案例并将其作为基础模型的上下文提示信息，采用〈example query，example answer；true query，?〉模板作为大规模语言预训练模型的

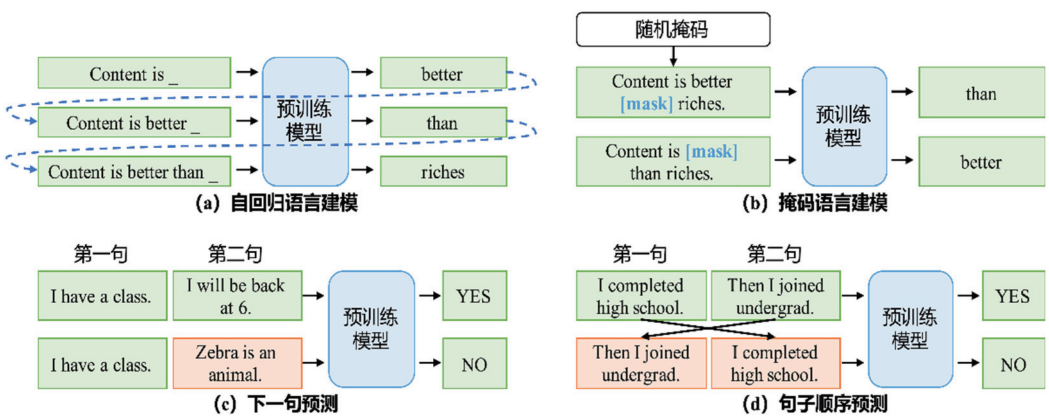

图 1　主流自然语言预训练方法

输入。通过多个案例集成，上下文学习赋予ChatGPT在零样本或少样本任务中进行推理的能力[11]。

思维链推理(Chain of Thought)主要用于解决算术推理、逻辑思考等复杂任务，通过采取类似人类的多步思维方式，拆解复杂任务为多个具体简单步骤，利用案例引导模型生成具体步骤及最终答案。

1.3 ChatGPT主要问题

ChatGPT目前存在以下主要问题：(1) 存在事实错误：生成的回答尽管逻辑正确，但存在事实性错误，称为“幻觉”现象。例如，ChatGPT会生成“鲁迅和周树人不是同一个人”的答案。(2) 数据需求量高：训练ChatGPT等大规模预训练语言模型，需要千万单词量级的预训练数据和数万条人工构造的指令，使得模型的训练成本极高。(3) 存在安全隐患：生成的回答中可能包含违反伦理道德、危害社会安全或侵犯知识产权的内容，例如ChatGPT可能被诱导生成网络攻击代码或表现出对特定事物的偏见。(4) 难以实时更新：若想添加更新的训练数据，需要重新训练整个模型，导致更新成本过高，使得模型包含的信息容易过时，造成ChatGPT回答出现错误。

除上述问题外，ChatGPT当前最大的局限性在于仅支持文本单模态的交互，限制了ChatGPT对真实世界的全面感知、通用理解和生成能力。因此，本文后续章节重点关注ChatGPT多模态化方面的研究，为构建图像、文本、视频、音频等多模态统一的交互式生成模型提供参考。

2 ChatGPT多模态化相关研究

围绕ChatGPT的多模态化问题，本文首先围绕多模态模型预训练和多模态模型生成两种范式介绍多模态的相关研究进展，然后介绍ChatGPT在多模态技术上的应用与拓展工作，并指出其典型应用场景与价值。

2.1 多模态模型预训练范式

现有多模态大模型在理解能力和生成能力发展过程中往往包含较为复杂的预训练策略，其中主要采用的基础策略可以归纳为以下两类：多模态动态对齐式预训练(Multi-modal Dynamic Alignment Pre-training)、单模态固定引导式预训练(Uni-Modal Frozen Guidance Pre-training)，如图2所示。

多模态动态对齐式预训练(Multi-modal Dynamic Alignment Pre-training)是指多模态大模型的不同模态编码器、解码器在预训练过程中采用动态对齐的更新策略，其特点包括：(1) 不同模态的表征和生成对应模块参数在训练中长期处于同步更新的平衡状态，覆盖模型中所有模块的参数；(2) 训练损失函数中包含对齐损失，并将其视作不同模态间统一的学习目标；(3) 该方法适用于具备一定规模的训练数据，对多种训练任务表现出较好的适应性和泛化效果。模态对齐预训练中各模态编码器和解码器既可以采用从零开始(From Scratch)的包含随机参数的单一模态编码器，也可以采用已经过预训练初始化后的参数作为对齐过程中的预训练初始参数，后者往往具备加速训练和提高模型整体性能的效果，已逐渐成为主流方案。Radford等人[12]提出的CLIP采用模态对齐式预训练策略，通过图像—文本对比损失函数，训练过程中从零开始同步更新图像编码器以及文本编码器参数，在高维表征空间中对齐两类模态编码器输出，借助在网络收集的4亿图像—文本数据，得到具备一定开放场景适应性的图像、文本两种模态内容理解能力的多模态预训练大模型。在此基础上，Xu等人[13]提出了面向视频模态的VideoCLIP以及Fan等人[14]提出用于游戏模拟中提供跨视觉文本奖励的MINECLIP，采用类似的训练方式在已有预训练视频、文本编码器的参数上沿用模态对齐式预训练策略，实现了视频模态以及文本模态理解能力的多模态预训练大模型。Wang等人[15]提出的BEIT-3，采用基于BERT架构的多路Transformer模型实现多种下游视觉文本任务如目标检测、语义分割、图像识别、视觉推理和视觉问答，其视觉、文本编码器及解码器均采用从零开始的初始化策略，并采用1.5千万张图像和2.1千万对图像文本作为预训练数据。Li等人[16]提出了VisualBERT，通过注意力层将文本与图像中内容进行对齐，在基于视觉信息的语言模型学习目标函数基础上，实现了视觉问答和视觉常识推理任务性能的提升。模态对齐式预训练在训练过程中对多种模

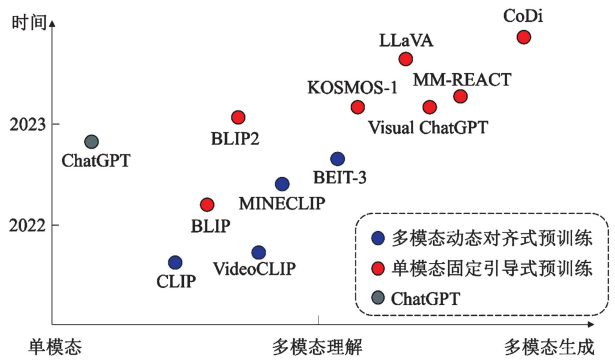

图2 多模态预训练模型的发展历程

态模块进行动态对齐，使其在多种任务上具备较好的适应性以及性能表现，但是同步对齐对采用预训练参数作为起始状态的多模态模块具有一定受灾难遗忘影响的风险，且实际中对数据规模往往有一定要求。

单模态固定引导式预训练（Uni-Modal Frozen Guidance Pre-training）是指多模态大模型训练过程中存在固定参数的预训练单模态编码器，其余编码器以及解码器在固定编码器输出的表征引导下进行参数更新，其特点包括：（1）存在至少一种模态下编码器为预训练编码器，并在训练过程中保持参数固定，训练策略仅覆盖模型其他模块参数更新；（2）训练损失函数中包含由固定预训练编码器作为引导项的损失函数，引导其他模态或该模态解码器完成编码内容理解与同步；（3）该方法适用于规模受限的预训练数据集，通过固定编码器参数对微调带来的灾难遗忘具备一定抵抗能力。Huang 等人[17]提出了 KOSMOS-1 多模态大语言模型，能够理解多模态信息并遵循指令完成文本生成、多模态对话、视觉问答、图片描述等多模态任务，在训练过程中，其采用 CLIP 的视觉编码器 ViT-L/14 作为图像编码器并固定其参数，通过多项包含视觉编码信息作为引导内容的训练损失更新文本编码及解码器，使其具备了一定的多模态理解与生成能力。Li 等人[18, 19]提出的 BLIP 系列工作，则通过固定预训练的图片编码器与大规模语言模型（LLM）参数，仅采用包含文本、视觉引导的多模态任务损失函数学习小规模注意力模块，实现图像查询信息提取与文本信息共享并支持视觉问答、图片描述等任务。Liu 等人[20]提出的 LLaVA 模型采用全连接层作为跨模态辅助模块，通过构造大量的指令遵循任务（Instruction-following Task）将固定视觉编码器产生的图像表征映射至 LLM 提示区间生成对应文本答案，其回答案例如图 3 所示。单模态固定引导式预训练通过借助预训练固定参数的模态编码器对整体多模态大模型进行引导训练，实现了在不受灾难遗忘影响下的对部分模态理解能力的迁移，但是该训练策略也可能导致固定的模态编码器缺少对其他模态理解的能力。

2.2　多模态模型生成范式

根据生成范式的不同，现有多模态生成模型可以分为：（1）基于单一跨模态解码器的生成模型，根据输入特定模态内容生成对应单一跨模态结果；（2）基于端到端多模态解码器的生成模型，通过端到端训练多种并行模态解码器同时生成对应多种模态内容。

在基于单一跨模态解码器的生成模型方面，Kirillov 等人[21]提出的 SAM 是一种能针对图像输入给出对应物体语义分割结果的开放域视觉大模型，生成案例效果如图 4 所示。Ramesh 等人[22]提出的 DALL-E 2 是一种由文本到图像的单一跨模态生成模型，通过扩散模型对文本信息逐步解码，支持由输入文本生成对应内容的图片创作，生成结果如图 5 所示。Deng 等人[23]提出的 LFR-GAN 挖掘输入文本细粒度属性，逐步引导对抗生成网络迭代生成图片，实现文本到图片的跨模态生成功能。Poole 等人[24]提出的 DreamFusion 借助文本到二维图像的预训练扩散模型作为 3D 建模的先验模型，利用梯度下降优化 3D 建模在二维投影与文本距离损失，实现了文本到 3D 模型的跨模态生成功能。Villegas 等人[25]提出的 Phenaki 则借助因果模型将视频压缩到离散单元，通过文本生成离散单元与离散单元解码过程，实现了文本到视频的跨模态生成。Radford 等人[26]提出的 Whisper 是一种语音到文本的生成大模型，通过在 68 万小时语音数据上训练，具备了英语语音识别能力。上述工作主要实现了文

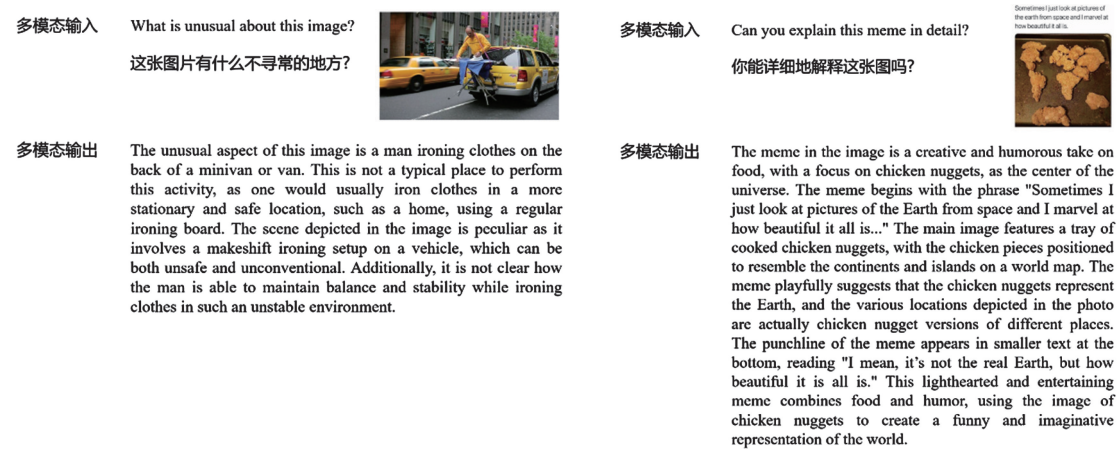

图 3　多模态大模型 LLaVA 的生成案例

本模态到其他单一模态的生成，分别打通了各类模态间一对一的生成关系，但无法解决多种模态同时生成的需求。

在基于端到端多模态解码器的生成模型方面，Tang 等人[27]提出了 CoDi 模型，在训练阶段通过图像、文本、视频、音频编码器与文本编码器对齐，生成阶段采用扩散模型进行不同模态图像、文本、视频、音频内容的生成，实现了多种模态端到端的统一理解与生成。CoDi 虽然具备了不同模态互通的生成功能，但是受限于现有部分模态解码器生成能力，生成内容的连贯性与效果仍有较大提升空间。近期中国科学院自动化研究所发布的“紫东太初 2.0”多模态大模型①具备全面模态的理解和生成能力，在图像、文本、语音、视频四种模态的基础上，融入 3D 点云、信号等其他模态数据，支持多轮问答、文本创作、图像生成、3D 理解、信号分析等多种任务。基于端到端多模态解码器生成模型，具备同时理解与生成多模态内容的优势，相对单一跨模态生成模型具备更广的应用范围。

2.3 ChatGPT 多模态化研究进展

为完善 ChatGPT 多模态功能，有必要对其进行多模态化研究。2022 年 3 月 OpenAI 和微软分别发布了 GPT-4 与 Visual ChatGPT[28] 两款基于 ChatGPT 的多模态预训练模型。GPT-4 作为目前 OpenAI 综合测评能力最强的多模态预训练模型，具备了一定的图片理解能力，能根据输入图片进行视觉问答，不仅能理解人类才能理解的图片笑话，还能根据图片进行相关网站代码生成。但受公司策略

图 4 视觉大模型 SAM 的生成案例

图 5 多模态生成模型 DALL-E 2 的生成案例

① 紫东太初 2.0：http://finance.people.com.cn/n1/2023/0616/c1004-40015792.html

与商业等因素影响，目前 GPT-4 相关技术内容尚未公开。Visual ChatGPT 相对 GPT-4 具备视觉内容生成能力，通过调用相关视觉模型插件接口引入相关工具，辅助 ChatGPT 完成特定多模态生成任务，例如引入 Stable Diffusion[29] 和 ControlNet[30] 等工具在与 ChatGPT 的循环协作中按要求逐步生成指定内容。除了 OpenAI 与微软在 ChatGPT 的后续工作外，国内外互联网公司如 Google、Meta、百度、阿里巴巴等分别在多模态化大语言模型上进行了相关研究，例如 Google 推出了 PaLM-E[31] 支持图片、机器人状态、神经场景表征等多模态任务，Meta 推出的 ImageBind① 支持 6 种模态下任意输入模型的跨模态检索功能，百度推出的“文心一言”② 多模态预训练模型具备文本生成图片和视频能力，阿里巴巴推出“通义千问”③ 预训练大模型具备图片—文本组合的跨模态输入与图片—文本的多模态生成能力。多模态预训练模型以及现有大语言模型多模态化已经成为工业界与学术界共同关注的热点。

表 1 列举了近年来主要的模块化跨模态预训练模型性能、模态支持、参数量等指标。其中 MME[32] 数据集通过 14 种粗粒度感知、细粒度感知、认知任务分别对多模态大模型感知和认知两方面任务进行测评，通过判断题的正确率来评价多模态大模型在多种粒度的多模态感知和认知任务中的表现效果。OwlEval[33] 是一种考察大模型图片理解能力和复杂对话能力的测试基准，具体考察任务包括自然图片理解、表格图片理解、流程图理解、OCR、知识密集对话、引用型对话等，主要通过人工打分排序的方式进行多模态大模型能力的评测。在表 1 中，每项性能指标保留原有测试基准的度量单位，分数越高代表测试结果越好，我们主要对比 BLIP2、MiniGPT-4[34]、LLaVA、mPLUG-Owl[33] 等模型在 MME 中感知、认知任务，OwlEval 中单论和多轮对话任务的表现性能。相关模型语言和视觉编码器部分模型参数结果如图 6 所示。上述工作在一定程度上弥补了 ChatGPT 在多模态任务能力上的缺失，逐步解决了多种模态间信息传递和生成的部分问题，为后续大语言模型在多模态领域的发展提供了参考方向。

表 1　多模态预训练大模型对比表

模型名称	支持模态	参数量 (M)	MME 测评		OwlEval 对话测评	
			感知	认知	单轮	多轮
BLIP2	文本、图片	12.1B	1 293.84	290.00	13	13
MiniGPT-4	文本、图片	14.2B*	866.58	292.14	35	25
LLaVA	文本、图片	13B	502.82	214.64	23	16
mPLUG-Owl	文本、图片	7.2B	967.35	276.07	32	31

* 官方论文暂未发布准确参数量，仅提供评估结果供参考。

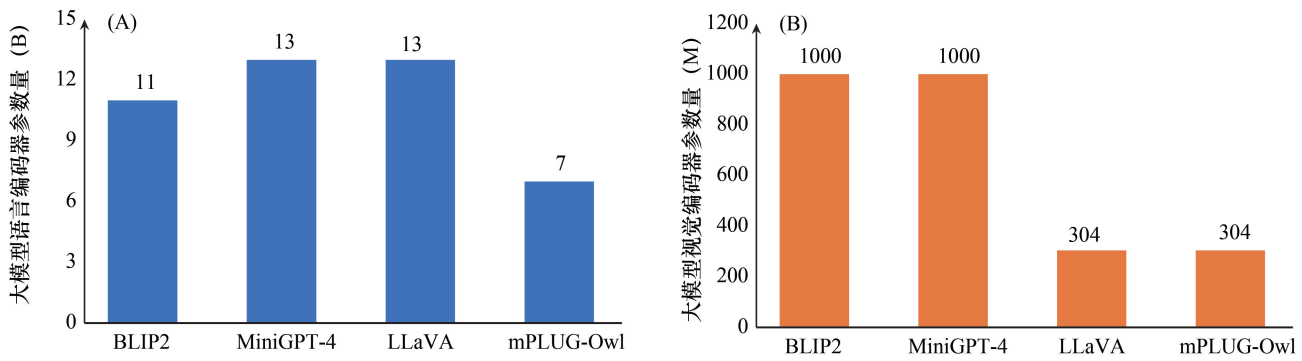

图 6　多模态预训练对比结果：A. 大模型语言编码器参数量，B. 大模型视觉编码器参数量

① Meta imagebind：https://imagebind.metademolab.com/
② 百度“文心一言”：https://yiyan.baidu.com/
③ 阿里“通义千问”：https://tongyi.aliyun.com/

2.4 多模态预训练模型应用场景

多模态预训练模型在推动通用人工智能技术应用中起到了关键作用：(1) 在政务服务领域，可以用于提高政务咨询系统的智能问答水平，增强多场景下的服务能力，辅助市民服务热线高效回应市民诉求，推进政务办事精准指引和高效审批。(2) 在医疗领域，根据患者的症状描述、医学图像、实时监测数据等，借助专业领域知识驱动推理和多模态大模型信息分析能力，提供专家级的科学诊断建议和个性化药物治疗推荐，提高医疗资源利用效率。(3) 在科学研究领域，可用于加速人工智能技术赋能新材料和创新药物等领域的科学研究，缩短科研实验周期，促进新材料、新蛋白质分子结构化序列等的发现。(4) 在金融领域，能根据历史的金融数据和市场图文情报构建用于决策和评估的思维树，并在金融和情报数据上对模型进行精调，帮助大模型进行风险评估和决策，提高金融机构对投资项目、贷款申请、交易风险等方面的分析能力。(5) 在自动驾驶领域，可用于研发多模态融合感知技术，提高自动驾驶模型的多维感知预测性能，有效解决复杂场景的长尾问题，辅助提高车载自动驾驶模型的泛化能力。(6) 在城市治理领域，可用于在城市大脑建设中应用大模型技术，加快多维感知系统融合处理技术的研发，实现智慧城市底层业务的统一感知、关联分析和态势预测，为城市治理决策提供全面综合的技术支撑。

3 ChatGPT 多模态化研究方向展望

ChatGPT 背后的核心技术，即大规模预训练模型，在自然语言处理领域的进展已领先于计算机视觉和多模态分析领域，原因是当前缺少视觉和多模态数据的通用预训练模型。同时，现有多模态预训练模型的多模态生成能力有限，难以同时生成文本、图像、视频等更多模态。另外，数据—知识双轮驱动作为新一轮人工智能的创新方向之一，在多模态大模型和人工智能生成内容领域有重要的研究潜力。本章从多模态预训练任务、多模态生成能力、数据—知识双轮驱动等方面对 ChatGPT 多模态化的研究方向进行展望。

3.1 通用多模态预训练任务

自监督预训练是预训练模型的核心方法，自然语言处理领域目前已有较为成熟的预训练方法，可以将多种任务统一为包含掩码的生成任务，经过掩码语言建模预训练的模型，通过提示学习等方法适配于不同下游任务。而在计算机视觉及多模态分析领域，尚缺少类似的通用预训练任务。目前主流自监督预训练任务的图像掩码建模（代表性工作为 BEiT[35]、MAE[36]）和图文对比学习（代表性工作为 CLIP、BLIP），与下游任务有较大差别，难以采用统一范式包含所有自监督预训练任务。上述自监督预训练方法局限于提供视觉表征，对下游任务的适配则依赖标注数据的精调，导致预训练的增益相比自然语言较为有限。这一方面是由于计算机视觉和多模态分析领域的下游任务种类更加繁多，包括图像分类、图像分割、视觉问答等，其输入和输出的模态均不相同，难以使用一套统一的框架；另一方面，自然语言数据实际上是经人类编码后的信息载体，是离散、信息精炼的，而视觉数据是直接取自现实世界的信息载体，是连续、信息冗余的。因此，从视觉数据中学习规律知识比从自然语言中学习需要更多的训练数据，且难度更高。

由于视觉等其他模态缺少有效的预训练任务，目前的多模态大模型大多以大语言模型为基础，利用已有文本模型的推理、理解、生成等能力，通过预训练视觉编码器实现视觉、文本等多模态下游任务，代表性工作有微软的 LLaVA、谷歌的 PaLM-E 等。以 LLaVA 为例，网络框架如图 7 所示。LLaVA 以预训练大语言模型（LLM）为基础，利用预训练视觉编码器处理图像输入，再通过矩阵映射对图像和文本特征进行对齐，然后共同输入 LLM。其中，视觉编码器采用多模态对比学习预训练的 CLIP ViT-L/14 模型，因此视觉特征已经与 CLIP 编码的文本特征在语义空间中对齐，再通过矩阵映射与 LLM 的文本特征进行对齐。最后利用 LLM 生成高质量的文本回答。训练时，首先冻结 LLM 和视觉编码器，仅训练特征对齐矩阵；然后通过多模态指令数据对整个模型进行有监督微调。

针对计算机视觉和多模态分析领域下游任务的多样性难点，Chen 等人[37]提出了视觉任务的统一框架 Pix2Seq，将视觉任务的像素输入转换为离散的序列输入，并通过提示词指定不同任务使模型输出不同的数据形式，能够处理目标检测、实例分割、关键点检测、图像描述四项视觉任务。Kirillov 等人提出可输入多模态提示词（Prompt）的图像分割方法 SAM，支持根据文本、点坐标、边界框、图像掩码等多种模态的输入信息。但上述方法对于多模态输入不够灵活，难以适应输入更多模态。如何对不同模态数据进行统一并设计有效的自监督预训练模型，

仍是目前亟待解决的问题。

3.2　多模态生成能力

现有多模态预训练模型如 GPT-4、LLaVA、PaLM-E 等，通过向预训练语言模型中添加图像编码器等方式，实现了图像和文本两种模态的共同输入，但目前仅能实现文本单模态的输出，无法同时输出图像、视频等更多模态。

在探索预训练语言模型的多模态生成能力方面，Yu 等人[38]提出语义金字塔自编码器 SPAE，使冻结参数的预训练语言模型能够执行图像、视频等更多模态的理解和生成任务。核心思想为将图像、视频中的视觉信息转换为预训练语言模型能够理解的语言。具体而言，采用类似离散自编码器的方式，将视觉信息嵌入离散空间，而该离散空间由预训练语言模型能够理解的语义单元（例如单词）组成。SPAE 能够实现输入图像和文本，输出文本；输入图像，输出图像和文本；输入文本，输出图像、文本等多模态任务。

然而，SPAE 生成的图像质量和多样性与当前主流的基于扩散模型的图像生成模型相比仍存在较大差距。另外，在有监督微调阶段，构造包含图像结果的指令数据难度远高于纯文本输出的指令数据。因此，如何利用预训练语言模型的生成能力，进行图像、视频等多模态生成，是目前亟待解决的问题之一。

3.3　多模态数据—知识双轮驱动

目前的大模型研究以数据驱动为主，主要面临以下挑战：数据需求高，依赖大量数据以监督或自监督进行参数拟合；可解释性弱，以黑盒方式优化模型，难以进行解释；鲁棒性弱，基于数据驱动的模型难以应对样本攻击[39]。此外，目前以 ChatGPT 为代表的问答模型存在“幻觉”现象，即生成的答案虽然逻辑正确但存在事实错误的问题。现有工作[40]表明，在单模态文本生成以外，多模态大模型同样存在严重的幻觉现象，主要体现在对图像内容的理解上。以微软 LLaVA 模型为例，向模型输入一张北京大学的图片，模型会回答这是“中国南方大学”，而“中国南方大学”实际上并不存在。因此，多模态预训练模型中的幻觉现象同样具有研究的必要性。上述问题可通过引入外部知识，实现数据—知识双轮驱动进行优化。

目前主流的知识组织形式为知识图谱，通过构建多模态知识图谱，从中提取相关知识条目，可对模型生成的回答进行约束，从而在生成可信度高的回答同时，为用户提供回答的依据，解决生成内容事实错误的问题。目前已有相关研究构建特定领域的知识图谱，包括医学[40]、数学[41]、物理[42]、金融[43]、历史[44]、电商[45]等。另外，通用领域方面，Wang 等人[46]提出了图文知识图谱 Richpedia，从维基百科和谷歌搜索引擎等来源获取数据，包含近 3 万个实体、300 万张图像以及 1 亿个三元组。然而目前知识图谱同样存在模态限制的问题，大多数多模态知识图谱仅包含图像、文本两种模态，对于真实世界中的多模态知识建模能力有限。如何构建包含视频、音频等多种模态的知识图谱仍是值得研究的问题。

目前将知识图谱用于增强预训练模型的方式主要包括在模型的中间层引入知识信息[47]、从知识图谱中提取相关概念补充原始数据的潜在关联[48]、将知识作为单独的模态进行对比学习[49]、通过图卷积网络聚焦多模态关系以增强图文嵌入[50]等。如图 8 所示，以知识作为单独模态的 K3M 为例，数据—知识双轮驱动往往采用数据与知识结合的方式得到任务层前特有的数据特征，保留知识图谱中结构信息以及数据中物体表征信息，有助于实现复杂推理。然而现有的知识图谱预训练结合方法主要关注特征表示阶段而非模型输出阶段，如何利用外部知识对 ChatGPT 等问答模型的输出进行约束，也是值得研究的问题。此外，在知识图谱增强预训练模型的同时，预训练模型也可促进知识图谱的构建和补全。

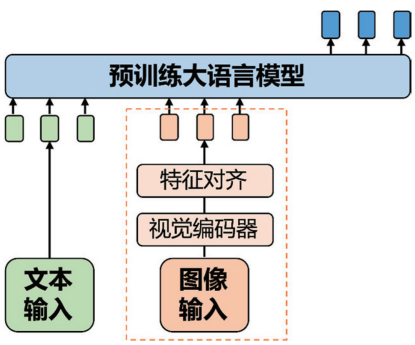

图 7　多模态大语言模型 LLaVA 网络结构

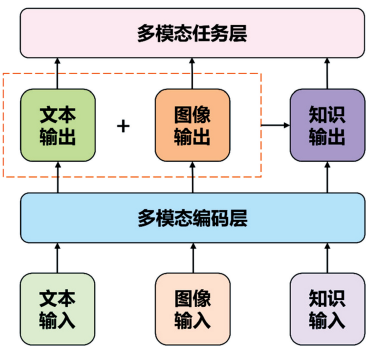

图 8　多模态数据—知识双轮驱动模型 K3M 网络结构

Lv 等人[51]提出基于预训练语言模型的知识图谱补全方法 PKGC，利用预训练模型中的隐含知识和知识图谱中的结构化知识共同推断新知识，实现对知识图谱中缺失条目的补全。知识图谱可增强预训练模型的推理能力，预训练模型也可用于知识图谱补全，两者相互促进，是实现数据—知识双轮驱动的关键技术。

3.4 多模态预训练大模型研究的着力点

综合上述研究方向展望，针对我国当前相关研究和业界应用现状，本节从模型架构、模型应用和模型部署三方面出发，讨论多模态预训练大模型研究的着力点包括如下三个方面：

在模型架构方面，应着力探索具备多种模态综合理解与生成能力的预训练模型架构。我国当前的主流多模态预训练模型支持文本、图像输入和文本输出，缺少对更多模态的支持。一方面，现有模型难以处理图文以外的其他模态输入；另一方面，大多数现有模型仅能输出文本，或采用一个单独的图像生成模型实现图像输出，导致图像生成结果与原问题匹配程度较低，目前未能实现同时生成图像、文本等多模态信息。

在模型应用方面，应着力结合领域知识开发专业、可靠的特定领域大模型。我国目前已具备多个领域的专业知识库基础，可结合领域专业知识，通过对通用领域的预训练大模型进行微调等方式，构建特定领域专用的大模型，相比通用大模型在各领域场景中具备更广泛的应用场景。同时，医学、电商等领域依赖图像、文本等多模态数据的协同分析，因此更需要领域专用的多模态预训练大模型。

在模型部署方面，应着力研究如何降低预训练大模型的计算成本。我国乃至全球目前的预训练大模型均依赖大量的训练数据和计算资源，这对大模型的开发和部署使用造成了难以克服的障碍。因此，研究如何降低预训练大模型的计算成本，包括训练数据量、模型参数量等方面，具有重要的研究和应用价值。本章讨论的数据—知识双轮驱动作为路线之一，同时也有其他路线尚待进一步探索。

4 总结与展望

本文讨论了 ChatGPT 及其多模态化方面的相关研究。首先介绍了 ChatGPT 的核心技术，包括预训练语言模型、有监督微调、以及如何从人类反馈中进行增强学习等，并分析了 ChatGPT 的主要问题。然后，详细介绍了多模态模型预训练范式和多模态模型生成范式，并讨论了 ChatGPT 在多模态化方面的研究进展以及应用前景。最后，展望了 ChatGPT 在多模态化方面的未来研究方向，包括探索多模态预训练任务、以及多模态数据—知识双轮驱动等研究方向。目前，ChatGPT 等大规模语言模型在多模态化方向的研究仍处于探索阶段，距离真实世界的多模态理解与生成任务仍存在很大挑战，希望本文的介绍和总结能为未来的相关研究提供有价值的参考。

参 考 文 献

[1] 彭宇新，綦金玮，黄鑫. 多媒体内容理解的研究现状与展望. 计算机研究与发展，2019，56(1)：183—208.

[2] Radford A，Narasimhan K，Salimans T，et al. Improving language understanding by generative pre-training.（2018-06-11）/[2023-06-20]. https://openai.com/research/language-unsupervised.

[3] Radford A，Wu J，Child R，et al. Language models are unsupervised multitask learners. OpenAI blog，2019，1(8)：9.

[4] Brown T，Mann B，Ryder N，et al. Language models are few-shot learners. Advances in Neural Information Processing Systems，2020：1877—1901.

[5] Devlin J，Chang MW，Lee K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Minneapolis：Association for Computational Linguistics，2019：4171—4186.

[6] OpenAI. GPT-4 Technical Report.（2023-03-27）/[2023-06-20]. https://arxiv.org/abs/2303.08774.

[7] Anthropic. Introducing claude.（2023-03-14）/[2023-06-20]. https://www.anthropic.com/index/introducing-claude.

[8] Christiano PF，Leike J，Brown T，et al. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems，2017：4299—4307.

[9] Bai Y，Jones A，Ndousse K，et al. Training a helpful and harmless assistant with reinforcement learning from human feedback.（2023-04-12）/[2023-06-20]. https://arxiv.org/abs/2204.05862.

[10] Wei J，Wang X，Schuurmans D，et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems，2022：24824—24837.

[11] Dong Q，Li L，Dai D，et al. A survey on in-context learning.（2022-12-31）/[2023-06-20]. https://arxiv.org/abs/2301.00234.

[12] Radford A，Kim JW，Hallacy C，et al. Learning transferable visual models from natural language supervision.（2021-02-26）/[2023-06-20]. https://arxiv.org/abs/2103.00020.

[13] Xu H, Ghosh G, Huang PY, et al. VideoCLIP: contrastive pre-training for zero-shot video-text understanding// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6787—6800.

[14] Fan LX, Wang GZ, Jiang YF, et al. MineDojo: building open-ended embodied agents with internet-scale knowledge. (2022-06-17)/[2023-06-20]. https://arxiv.org/abs/2206.08853.

[15] Wang W, Bao H, Dong L, et al. Image as a foreign language: beit pretraining for all vision and vision-language tasks. (2022-08-22)/[2023-06-20]. https://arxiv.org/abs/2208.10442.

[16] Li LH, Yatskar M, Yin D, et al. Visualbert: a simple and performant baseline for vision and language. (2019-08-09)/[2023-06-20]. https://arxiv.org/abs/1908.03557.

[17] Huang S, Dong L, Wang W, et al. Language is not all you need: aligning perception with language models. (2023-02-27)/[2023-06-20]. https://arxiv.org/abs/2302.14045.

[18] Li J, Li D, Xiong C, et al. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation// Proceedings of the 39th International Conference on Machine Learning. Virtual: PMLR, 2022: 12888—12900.

[19] Li J, Li D, Savarese S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. (2023-01-30)/[2023-06-20]. https://arxiv.org/abs/2301.12597.

[20] Liu H, Li C, Wu Q, et al. Visual instruction tuning. (2023-04-17)/[2023-06-20]. https://arxiv.org/abs/2304.08485.

[21] Kirillove A, Mintun E, Ravi N, et al. Segment anything. (2023-04-05)/[2023-06-20]. https://arxiv.org/abs/2304.02643.

[22] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation// Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR, 2021: 8821—8831.

[23] Deng ZJ, He XT, Peng YX. LFR-GAN: local feature refinement based generative adversarial network for text-to-image generation. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(6): 1—18.

[24] Poole B, Jain A, Barron JT, et al. Dreamfusion: text-to-3d using 2d diffusion. (2022-09-29)/[2023-06-20]. https://arxiv.org/abs/2209.14988.pdf.

[25] Villegas R, Babaeizadeh M, Kindermans PJ, et al. Phenaki: Variable length video generation from open domain textual description. (2022-10-05)/[2023-06-20]. https://arxiv.org/abs/2210.02399.

[26] Radford A, Kim JW, Xu T, et al. Robust speech recognition via large-scale weak supervision// Proceedings of the 39th International Conference on Machine Learning. Virtual: PMLR, 2023: 28492—28518.

[27] Tang Z, Yang Z, Zhu C, et al. Any-to-any generation via composable diffusion. (2023-05-19)/[2023-06-20]. https://arxiv.org/abs/2305.11846.

[28] Wu C, Yin S, Qi W, et al. Visual chatgpt: talking, drawing and editing with visual foundation models. (2023-03-08)/[2023-06-20]. https://arxiv.org/abs/2303.04671.

[29] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 10684—10695.

[30] Zhang L, Agrawala M. Adding conditional control to text-to-image diffusion models. (2023-02-10)/[2023-06-20]. https://arxiv.org/abs/2302.05543.

[31] Driess D, Xia F, Sajjadi MSM, et al. Palm-e: an embodied multimodal language model. (2023-03-06)/[2023-06-20]. https://arxiv.org/abs/2303.03378.

[32] Fu C, Chen P, Shen Y, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. (2023-04-20)/[2023-06-20]. https://arxiv.org/abs/2304.10592.

[33] Zhu D, Chen J, Shen X, et al. Minigpt-4: enhancing vision-language understanding with advanced large language models. (2023-04-20)/[2023-06-20]. https://arxiv.org/abs/2304.10592.

[34] Ye Q, Xu H, Xu G, et al. Mplug-owl: modularization empowers large language models with multimodality. (2023-04-27)/[2023-06-20]. https://arxiv.org/abs/2304.14178.pdf.

[35] Bao H, Dong L, Piao S, et al. BeiT: BERT pre-training of image transformers. (2021-06-15)/[2023-06-20]. https://arxiv.org/abs/202106.08254.

[36] He KM, Chen XL, Xie SN, et al. Masked autoencoders are scalable vision learners// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 15979—15988.

[37] Chen T, Saxena S, Li LL, et al. Pix2seq: a language modeling framework for object detection. (2021-09-22)/[2023-06-20]. https://arxiv.org/abs/2109.10852.

[38] Yu L, Cheng Y, Wang Z, et al. SPAE: semantic pyramid autoencoder for multimodal generation with frozen LLMs. (2023-06-30)/[2023-09-28]. https://arxiv.org/abs/2306.17842.

[39] 金哲，张引，吴飞，等. 数据驱动与知识引导结合下人工智能算法模型. 电子与信息学报，2023，45(7)：2580—2594.

[40] Yang X, Wu CK, Nenadic G, et al. Mining a stroke knowledge graph from literature. BMC Bioinformatics, 2021, 22(10): 1—19.

[41] Wang JN. Math-KG: construction and applications of mathematical knowledge graph. (2022-05-08)/[2023-06-20]. https://arxiv.org/abs/2205.03772.

[42] Shang JL, Huang JY, Zeng SH, et al. Representation and extraction of physics knowledge based on knowledge graph and embedding-combined text classification for cooperative learning// Proceedings of the 25th International Conference on Computer Supported Cooperative Work in Design. New York: IEEE, 2022: 1053—1058.

[43] Elhammadi S, Lakshmanan LVS, Ng R, et al. A high precision pipeline for financial knowledge graph construction// Proceedings of the 28th International Conference on Computational Linguistics. Stroudsburg: International Committee on Computational Linguistics, 2020: 967—977.

[44] Liu SA, Yang H, Li JY, et al. Preliminary study on the knowledge graph construction of Chinese ancient history and culture. Information, 2020, 11(4): 186.

[45] Xu GH, Chen HH, Li FL, et al. AliMe MKG: a multi-modal knowledge graph for live-streaming e-commerce// Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: ACM, 2021: 4808—4812.

[46] Wang M, Wang HF, Qi GL, et al. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. Big Data Research, 2020, 22: 100159.

[47] Wang RZ, Tang DY, Duan N, et al. K-Adapter: infusing knowledge into pre-trained models with adapters// Findings of the Association for Computational Linguistics ACL 2021. Stroudsburg: Association for Computational Linguistics, 2021: 1405—1418.

[48] Ye ZD, He XT, Peng YX. Unsupervised cross-media hashing learning via knowledge graph. Chinese Journal of Electronics, 2022, 31(6): 1081—1091.

[49] Zhu YS, Zhao HX, Zhang W, et al. Knowledge perceived multi-modal pretraining in e-commerce// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 2744—2752.

[50] Feng DD, He XT, Peng YX. MKVSE: multimodal knowledge enhanced visual-semantic embedding for image-text retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022, 19(5): 1—21.

[51] Lv X, Lin YK, Cao YX, et al. Do pre-trained models benefit knowledge graph completion? A reliable evaluation and a reasonable approach// Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg: Association for Computational Linguistics, 2022: 3570—3581.

From ChatGPT to Large Multimodal Model: Present and Future

Geng Li[1,2] Zishuo Wang[1,2] Xiangteng He[1,2] Yuxin Peng[1,2*]

1. *Wangxuan Institute of Computer Technology, Peking University, Beijing* 100871

2. *National Key Laboratory for Multimedia Information Processing, Peking University, Beijing* 100871

Abstract At the end of 2022, OpenAI released ChatGPT chatbot that takes AI to new heights of understanding and generation for general-purpose natural language processing tasks, attracting widespread interest. Currently, ChatGPT only supports text-based interactions, while real-world perception relies on the collaborative interaction of multiple modalities such as image, text, video, and audio. How to achieve real-world perception and cognition across multisensory information such as vision, language, and hearing by imitating the cross-modal processing characteristics of the human brain is the key to improving the generic perception and interaction capabilities of models, as well as realizing general artificial intelligence. This paper analyzes the core technology and the problems caused by the limitation of text unimodality of ChatGPT, then introduces some representative current works on the combination of ChatGPT and multimodal analysis technology, and finally provides an outlook on the future research directions of multimodal ChatGPT from the perspectives of multimodal pre-training and data-knowledge driven.

Keywords ChatGPT; multimodal analysis; large language model; general artificial intelligence; multimodal pre-training.

（责任编辑 崔国增 姜钧译）

* Corresponding Author, Email: pengyuxin@pku.edu.cn