生成式人工智能的研究现状和发展趋势

车璐[1],张志强[2],周金佳[3*],李磊[3]

1. 西南科技大学环境与资源学院,绵阳 621000
2. 西南科技大学计算机科学与技术学院,绵阳 621000
3. 法政大学,东京 184-8584

摘要　随着ChatGPT的问世,生成式人工智能研究在文本、图像和视频等多模态信息处理领域取得了突破性的进展,引起了广泛的关注。梳理了生成式人工智能的研究进展,并探讨了其未来发展趋势。具体包含3个部分:一是从自然语言模型、图像与多模态模型回顾生成式人工智能的发展历程和研究现状;二是探讨生成式人工智能在不同领域的应用前景,主要聚焦内容交流、辅助设计、内容创作和个性化定制4个方面;三是分析了生成式人工智能面临的主要挑战及未来的发展趋势。

关键词　生成式人工智能;自然语言;多模态

1　生成式人工智能的历史和现状

生成式人工智能(artificial intelligence generated content,AIGC),其核心是利用人工智能技术生成和编辑各种类型的内容,如文字、语音、音乐、图像、视频等。在当前数字世界和物理世界加速融合的大背景下,AIGC重塑了数字内容的生产和消费模式。2018年,由人工智能(Artificial Intelligence,AI)创作的肖像画在纽约佳士得拍卖会上拍出了43.2万美元的高价,成为世界上首个出售的人工智能艺术品,引发各界关注。

AIGC的模型大致可分为2大类。一类是自然语言模型,即输入和输出的内容均为自然语言描述,例如,输入是一段文字,要求写一段故事或者是一个对话系统,输出也是一段文字,输出满足指令要求的一段文字,或者是和输入的文字进行对话。另一类是图像和多模态模型,即输入和输出是跨模态的,例如,输入文字输出视频,输入图片输出文字

收稿日期:2024-01-31;修回日期:2024-05-24
基金项目:西南科技大学研究生创新基金项目(24ycx3004)
作者简介:车璐,博士研究生,研究方向为人工智能多源数据融合技术,电子信箱:chelu1994@swust.edu.cn;周金佳(通信作者),副教授,研究方向为生成式人工智能,电子信箱:zhou@hosei.ac.jp
引用格式:车璐, 张志强, 周金佳, 等. 生成式人工智能的研究现状和发展趋势[J]. 科技导报, 2024, 42(12): 35-43;
doi:10.3981/j.issn.1000-7857.2024.01.00029

等。更进一步地,输入和输出都可以是多模态的,例如,输入文字加图片,输出一段视频序列和语音。这里的输出可以是重新生成的内容,也可以是对输入的编辑和修改。

1.1 自然语言模型

在AI生成内容的早期,不同领域,如自然语言处理(natural language processing,NLP)和图像生成领域之间,没有太多的重叠。在NLP领域,最初是使用N-gram朴素语言模型[1],学习单词分布,通过前一个字符来预测下一个字符。因为该模型记忆能力有限,所以无法生成超过一定长度的连贯文本。相比之下,基于神经网络语言模型能够生成较长的连贯文本。用于建模语言的第1类神经网络是循环神经网络(recurrent neural networks,RNN)[2],RNN逐个阅读单词,同时更新思维状态,使得该模型具备短期记忆。由于RNN存在着短期依赖瓶颈问题,长短期记忆网络(long-short term memory,LSTM)[3]被挖掘出来并用于长文本生成任务之中。在理论层面,LSTM可以实现长时间记忆。然而,在具体的实践中,经过几十到100个词后,该模型就开始偏离主题。为了解决这一问题,一种基于注意力机制的新型神经网络结构Transformer[4]在2017年被提出,同时受到了广泛关注。该架构的并行化处理使其能够充分利用图形处理单元(graphics processing unit,GPU)。此外,该结构在设计层面上允许不断地堆叠编码器或解码器结构,使得整个网络结构能够变得更为复杂,这为后续大语言模型的出现奠定了基础。

自2018年以来,基于Transformer架构,大语言模型开始逐步涌现,其中最著名的当属Google的来自Transformer的双向编码器表示(bidirectional encoder representations from transformer,BERT)模型[5]和OpenAI的生成式预训练(generative pre-training,GPT)系列模型,其包括OpenAI于2018年率先提出的GPT-1模型[6],约有1.2亿个参数。紧接着,Google于2019年提出了BERT模型,约有3.4亿个参数,其整体性能优于GPT-1。随后,OpenAI迅速提出了GPT-2模型[7],拥有的参数量高达15亿,并在40 GB的文本上进行了训练,实现了性能的进一步提升。之后,为实现更为优异的性能,OpenAI于2020年提出GPT-3模型[8],该模型具有1750亿个参数,其性能足以碾压之前的GPT-1、BERT和GPT-2模型。然而,由于缺乏有效的引导,GPT-3模型在生成文本内容时常常会出现一些不令人满意的结果。

为解决这一问题,OpenAI提出了InstructGPT模型[9],具体通过人类反馈强化学习(reinforcement learning from human feedback,RLHF)机制引导模型生成符合预期的内容结果。基于InstructGPT模型,OpenAI于2022年推出了ChatGPT[10],带来了AIGC面向大模型时代的浪潮。紧接着,OpenAI在2023年提出了GPT-4[11],其参数量高达到1.8万亿,整体性能令人惊叹。这些模型擅长文本理解,在文本分类、实体检测和问题回答等能力上具有卓越的表现。同时,其他新兴的大模型,如Sora,也为AIGC领域带来了新的视角,Sora模型通过其独特的架构和进阶的多模态处理能力,进一步拓宽了自然语言处理的应用范围。2024年,一些新的突破性研究工作进一步推动了NLP领域的发展。例如,Ding等[12]提出了新的高效微调方法,极大地减少了大模型的资源需求,并提高了大语言模型的适应性。Wu等[13]对持续学习在NLP中的应用进行了深入探讨,提出了自然语言处理未来可能的发展方向。NLP模型发展历程如图1所示。

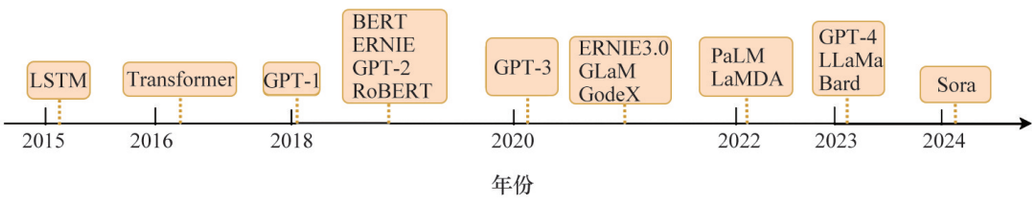

图1 NLP模型发展历程

从上述发展历程来看,AIGC在自然语言模型的发展已经正式进入大语言模型(large language model,LLM)时代。AIGC除了在自然语言模型上取得了突飞猛进的进展之外,在图像和多模态领域也取得了许多重大的突破性进展。

1.2　图像生成模型和多模态模型

在计算机视觉领域,传统的非深度学习图像生成算法大多使用纹理合成和纹理映射等技术。2012年,AlexNet的提出使得深度学习方法再一次回归到科研人员的视野之中[13]。基于深度学习,在图像生成领域中所有类型的图像生成模型都旨在学习训练集的真实数据分布,从而可以进一步产生具有一些变化的新数据。然而,由于无法学习到确切分布,所以现有的方法只能试图获得与真实数据分布尽可能相似的分布模型来生成图像。其中,一种常用和有效的生成图像的方法是变分自动编码器(variational autoencoder,VAE)[14],VAE旨在最大化数据对数似然下界来学习,它既能学习生成模型又可以推理模型。2014年12月,图像生成领域出现了一个具有里程碑意义的网络结构——生成对抗网络(generative adversarial networks,GAN)[15]。GAN包含一个生成器和一个判别器,生成器模型用于学习捕获数据分布,判别器模型用于判别样本是来自真实的数据分布还是生成器模型分布。GAN的核心是旨在实现生成器和判别器之间的对抗平衡,从而让生成器模型能够生成高质量的图像结果。GAN被提出之后,在各个领域都得到了广泛应用,其中最具代表性的结构是StyleGAN系列[16-18],其核心思想是风格调制,整个网络先将先验噪声映射到一个新的隐空间中,映射后的隐变量输入到生成器的多层次中,通过规范化层注入到生成过程,使模型在生成高质量图像基础上,做到层次特征可控。如生成人脸时,低层次控制是不同五官或人脸特征生成,高层次特征决定生成颜色。StyleGAN因其具有良好的可控性常被用于风格迁移或图像编辑任务之中。除了基于VAE和GAN结构之外,随着Transformer架构的出现,在图像生成领域涌现了一批基于Transformer的生成方法。2020年,Vision Transformer(ViT)[19]和Swin Transformer[20]通过将Transformer架构与视觉组件相结合,实现了高质量的图像生成效果。此外,在图像生成领域,扩散(Diffusion)模型[21]的引入实现了优质的图像生成效果,并开辟了图像生成的新方式。

受益于Transformer和扩散模型的出现,AIGC在多模态模型上也取得了许多重大突破。2021年1月,OpenAI发布文本合成图像模型DALL-E[22],其卓越的生成效果令人感到震惊。同年,对比语言图像预训练(contrastive language-image pre-training,CLIP)模型[23]问世。CLIP是一种结合了视觉语言模型ViT和Transformer的多模态模型。它通过接收大量文本和图像数据进行训练,在预训练过程中结合了视觉和语言知识,实现了文本作监督信号训练可迁移视觉模型。由于CLIP在图像和文本处理上的强大能力,后续的许多多模态模型均与CLIP模型进行结合,从而实现了优异的结果。之后,DALL-E2[24]和DALL-E3[25]分别于2022年4月和2023年10月发布。DALL-E2和DALL-E3只需要寥寥几句文本就可以生成超高质量的全新图像,将文本生成图像的逼真度和语言理解度提到了新的高度。除了DALL-E系列之外,Stable Diffusion[26]和Midjourney[27]也相继被推出,且生成效果广受好评。此外,基于Diffusion模型,AIGC在视频合成领域也取得了显著进步。Gen-2[28]和Pika[29]已经可以生成连贯的视频,但生成视频的质量和运动多样性还有待提高。在生成视频领域,目前最先进的模型是2023年12月发布的I2VGen-XL[30],它通过优化最初的600个去噪步从而实现了具有时间和空间一致性的高清视频生成结果,视频分辨率可以达到1280×720。图2总结了基于VAE、GAN和Diffusion生成模型的发展过程。

2　生成式人工智能的过程与应用前景

受益于计算资源和数据量的快速增长,AIGC算法在文本、图像和多模态信息处理方面均取得了令人瞩目的成就,这极大促进了AIGC在各个行业的应用落地。AIGC代表着人工智能领域的前沿技

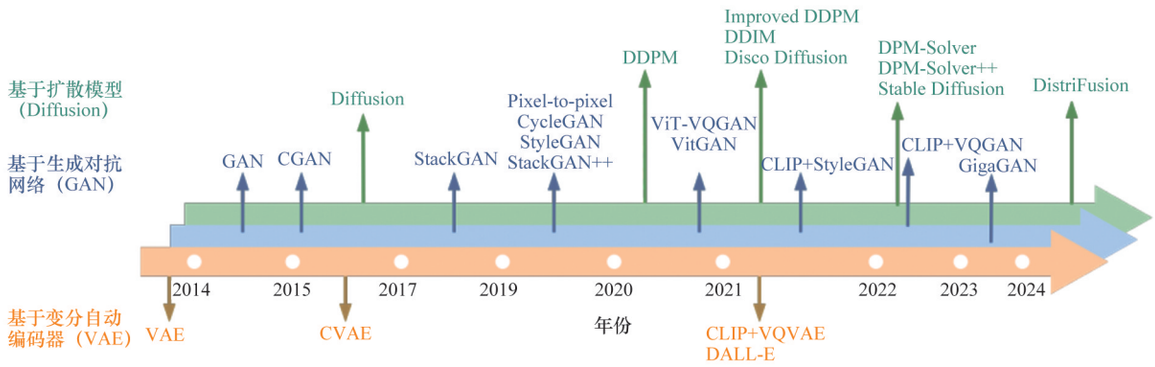

图2 基于VAE、GAN和Diffusion的生成模型发展历程

术，正在以前所未有的速度转变人们的工作方式、创意过程和日常生活。

2.1 AI的学习过程

2.1.1 AIGC的学习过程

AIGC的学习涉及到深度神经网络中的生成对抗网络（GAN）、变分自编码器（VAE）等技术。这些模型通过竞争学习过程，其中一个生成模型致力于创建越来越真实的数据，而另一个判别模型致力于更好地区分真实数据与生成数据。它们的共同目标是提高生成模型在未见数据上的泛化能力。此外，AIGC的训练不仅需注意数据量的大小，更需考量数据质量与多样性以及模型的训练策略和正则化技巧，以防模型过早地陷入局部最优而损失泛化性能。

2.1.2 传统AI的学习过程

传统AI的学习通常依赖于大量标记数据集，采用监督学习算法，通过反复迭代来降低模型预测与真实场景之间的误差。此学习方式在特定的任务上，如图像分类、语音识别等单一模态数据处理中表现突出。相比之下，AIGC则需要从数据中学习到潜在的分布，并根据这些学习到的分布来生成新的数据实例，这对模型在理解与把握数据内在特征方面提出了更高的要求。这一过程更像是“学习去生成”，而非“学习去预测”。

2.1.3 深度学习与迁移学习

AIGC的学习方法还与深度学习紧密相关，后者通过多层神经网络结构从复杂数据中自动学习到高层次特征。当这些深度学习模型配合迁移学习策略时，就可以将在一个领域学到的知识应用到另一个领域，从而在学习新任务时节省资源并提升效率。

2.2 AIGC与传统AI的区别

AIGC和传统AI在目标和方法、模型结构和训练方式、数据处理方式以及应用场景等多个方面展现出明显的差异性。

2.2.1 目标和方法的差异

AIGC的主旨在于创造，不仅是对现有信息的复现或者复制，而是在理解数据的基础上创造出全新的数据实例。这要求AIGC能够学习数据的内在分布和结构，进而生成与真实数据具有高度相似性，但又非完全相同的新数据。相反，传统AI，尤其是基于监督学习的模型，更多地关注于“预测”。它们通过从大量的输入—输出实例中学习，来预测给定输入所对应的输出。这些模型的主要目标是准确性和可靠性，而不直接关注创造性。

2.2.2 模型结构和训练方式

AIGC常用的生成对抗网络（GAN）和变分自编码器（VAE）在模型结构上具有独特性。例如，GAN通过引入生成器和鉴别器的对抗训练，模拟了一个迷你“博弈场”；而VAE则利用概率图模型来优化数据的潜在空间表示。传统AI模型，如回归模型、决策树、支持向量机（SVM）等，则通常具有更为直接的结构，它们通过最小化实际输出和预测输出之间的差异来进行训练。

2.2.3　数据处理方式

AIGC能够处理和生成的数据类型更为广泛,包括但不限于文本、图像、音频等。它们在处理数据时不仅关注特定的任务或标签,而且试图理解数据的全局属性和结构。而传统AI通常对特定的任务或问题进行优化,它们在数据处理上往往需要明确的标签信息,才能进行任务驱动的学习。

2.2.4　应用场景

AIGC因其独特的创造能力,在艺术创作、数据增强、虚拟环境模拟等领域展示了广泛的应用前景,能够帮助设计师创作新颖的设计方案,为研究人员提供丰富的训练数据,或为游戏和VR(虚拟现实)产业创造逼真的虚拟环境。传统AI则更多应用于数据分析、预测建模、自动化控制等领域,例如,在金融领域进行风险评估,在医疗领域进行疾病诊断,在制造业进行故障预测等。

理解AIGC与传统AI的差异不仅能帮助人们更合理地选用工具解决问题,也为AI的未来发展打开了新的视野和想象空间。

2.3　AIGC应用现状

当前,AIGC的应用可以分为以下4个方面。(1)内容交流。可以跟用户进行交流,并对用户提出的问题给予相应的解答。(2)辅助设计。可以辅助用户对相应的内容进行制作、修改和设计。(3)内容创作。可以基于用户的需求创作出全新的内容。(4)个性化定制。可以让用户根据需求对生成的内容进行个性化定制。

2.3.1　内容交流

AIGC当前最大的应用前景在于其能够与用户进行互动式交流,主要分为2个方面,一是聊天式交流,二是内容咨询式交流。聊天式交流主要可以用于情感聊天机器人,帮助患孤独症、抑郁症等精神疾病的人群缓解病情,辅助医生制定出对应的治疗方案。在内容咨询交流方面,可以面向各个行业领域构建AIGC内容咨询平台,如医疗咨询、法律咨询、生活常识咨询等。一方面,相比于传统的搜索平台,咨询平台能够通过交流的形式更好地明确用户的咨询需求,从而给出更有效的咨询结果。另一方面,构建的AIGC咨询平台能帮助不同领域的工作人员提高效率。例如,在医疗和法律咨询方面,用户可以根据自身需求在咨询平台中获得基本的建议,再去求助于医生或者律师。

2.3.2　辅助设计

AIGC在辅助设计层面具有广泛的应用前景。在教育行业,AIGC可以为教育工作者提供课程设计材料,通过自动创建和更新课程材料,教师只需要基于生成的课程材料进行进一步的修改即可,这大大地节省了教师的时间和精力。此外,AIGC可以帮助学生撰写日常报告等内容,并且AIGC能够对撰写的内容进行语法纠正、发现薄弱环节,并给出内容改进建议,帮助学生从错误中吸取教训,逐步提高写作能力。在媒体行业,AIGC可以辅助新闻工作者及时撰写紧急事件的新闻报道,并自动生成新闻标题,帮助新闻业提高效率和反应速度。此外,AIGC可以实现全天24小时的虚拟主持人新闻播报,能够减轻新闻工作者的负担。在电影行业,AIGC能够辅助进行剧本的加工工作,将老的剧本加工为精良的新剧本,之后再由导演和编剧进行进一步的修改。此外,它还可以提升电影的视觉效果,例如,改变电影画面的色彩化和分辨率等。除了上述行业之外,AIGC在计算机、医药和绘画行业都可以辅助工作者进行相关的设计和研究,例如,在计算机行业,它可以自动生成高质量的代码,并进行代码测试和重构工作;在医药行业,它能够辅助进行药物研发,进行蛋白质结构预测、蛋白质序列设计工作等;在绘画行业,AIGC可以辅助进行艺术品的保护和修复,能够将一些受损的艺术品恢复至初始状态。

2.3.3　内容创作

AIGC在内容创作层面具有良好的发展前景。在音乐行业,AIGC能够实现音乐的全过程创作,歌词、曲调、旋律等内容均可以由AIGC自动化生成。此外,在音乐创作过程中,它能够提供不同风格类型的音乐供用户选择。在绘画行业,AIGC可以制作出与众不同的复杂艺术作品,它可以通过分析图片来生成配色方案、图案和纹理信息,并创作出各种艺术形式的画作,如油墨画、抽象画、中国山水画和水墨画等。在广告行业,AIGC能够自动化生成

广告内容、海报以及设计徽标。在视频行业,AIGC能够生成具有创意的短视频内容,同时也可以生成电影场景内容。AIGC内容创作的优势在于它能够基于同样的内容自动化生成多种多样的结果供用户选择,能够较好地满足不同行业用户的需求。

2.3.4 个性化定制

AIGC在个性化定制层面具有极高的应用潜力。在教育行业,AIGC可以提供个性化的辅导,如可以生成独特幼儿外语教学产品,吸引儿童的注意力,调动其积极性,并提供一个有趣的学习环境;可以帮助高年级学生理解某些理论、概念和不同的语言文章,使其更有效地学习。在游戏行业,AIGC允许用户根据自身需要对游戏场景和故事情节进行个性化定制,使游戏体验更加身临其境。更进一步地,用户可以在游戏中举办大型活动,如演唱会、画展、毕业典礼等,使得全体参与人员具有独特的非凡体验。除了上述应用之外,AIGC在个性化定制层面最大的应用前景是实现数字永生。现阶段,利用AIGC技术已经能够实现人说话声音的改变、三维人像合成及内容交流。基于现有的AIGC技术及后续不断更新迭代的技术,只要大量收集整理某个人的语音、人像及交流模式这些数据,然后依靠这些数据就能够训练出此人的人物永生模型。该模型能够模拟此人说话的声音和方式,能够呈现此人的三维样貌,能够以此人的说话方式与他人沟通,如此便初步实现了此人的数字永生。即使在此人逝去之后,其他人也可以通过此人的人物永生模型与其进行交流。

3 生成式人工智能的潜在风险

在全面认识生成式人工智能应用潜力的同时,也必须正视伴随其发展出现的潜在风险。

3.1 知识产权的争议

AIGC技术能够创造出全新的艺术作品、音乐、文本等内容,这对于知识产权的定义提出了新的挑战。既存法律框架是建立在人类作者身上的,而AI创造出的作品并没有明确的“人类”作者。因此,谁拥有和控制由AI生成的作品的知识产权,以及这些作品是否应当被赋予知识产权保护,是目前亟待解答的问题。

3.2 数据隐私的威胁

AIGC通常需要大量的数据进行训练,这些数据不仅包括公开信息,而且可能包含个人敏感数据。如果不加以妥善管理,就可能导致未经授权的数据使用,进而侵犯个人隐私权益。此外,AIGC生成的虚假内容(如深度伪造)可能会用于社交工程攻击,给个人隐私带来更加直接的威胁。

3.3 道德使用的挑战

在没有充分监管的情况下,AIGC的输出可能会被用于不道德的目的。例如,制造虚假新闻、网络钓鱼信息,甚至是用于伪造历史证据等。这些行为不仅会给社会带来混乱,还可能威胁到社会制度和国家安全。

3.4 技术偏见的延续

AI系统通常会反映其训练数据的偏见。如果AIGC使用的数据集包含有性别、种族或其他形式的偏见,那么它生成的内容也可能会延续这些偏见,从而加剧现实世界中的不平等和不公正。

3.5 就业领域的冲击

类似于其他的自动化技术,AIGC在提高效率的同时,也可能导致某些工作领域能被机器取代,从而影响人类相应的就业机会。这不仅仅是对低技能劳动力市场的影响,也包括写作、设计等创意产业领域的专业工作。

AIGC作为一个强大的工具,其所带来的潜在风险是多方面的,涉及社会、法律、伦理等多个层面。因此,加强对AIGC应用的监管、确立道德使用原则以及制定相应的法律框架,将是人们面临的紧迫任务。只有这样,才能确保科技进步在不损害个人与社会利益的前提下,为人类带来更大的福祉。

4 生成式人工智能的挑战与发展趋势

4.1 主要挑战

尽管AIGC已经在各种生成式任务中展现了令人瞩目的成就,但AIGC目前仍存在诸多挑战,具体

有以下几个方面。

1）研究门槛过高。当前性能优异的AIGC算法均是基于“三超”(超大规模参数、超大规模数据和超大规模计算资源)环境实现的,使得AIGC算法研究的成本和门槛过高,让许多科研人员望而却步。这种情况极大限制了AIGC算法研究的进程。

2）生成内容不可控。尽管AIGC在文本、语音、图像、视频等多模态内容生成上取得了优质的生成效果,但内容生成的结果是不可控的。这种不可控主要体现在AIGC算法可能会生成带歧视性、暴力性、违法性等内容结果,这会带来法律和社会道德层面的问题。

3）生成性能不稳定。当前的AIGC算法在一些特定研究领域(如文本生成图像、文本生成视频、语音生成图像等)偶尔会生成一些特别差的结果,使得AIGC在这些领域的应用性较为一般。此外,一些特定的高风险领域(如医疗、金融服务、自动驾驶等)要求算法出错率极低或零错误,使得AIGC在这些领域的应用中只能起到一定的辅助作用。

4.2　发展趋势

当前的AIGC面临着上述的诸多挑战,整体上处于快速发展阶段。未来AIGC的发展趋势主要包含以下几个方面。

1）获取带标注的高质量数据。AIGC目前仍是以“暴力出奇迹”的方式实现了优异的性能,而要想实现“奇迹”,就需要基于“三超”环境进行研究,这又将大多数科研人员拒之门外。相比于“三超”环境,带标注的高质量数据能够在“三中”(中等规模参数、中等规模数据和中等规模计算资源)环境下实现优异性能。因此,未来需要在获取带标注的高质量数据上研究行之有效的方法,降低AIGC研究的门槛。

2）生成内容的检测和评估。AIGC现阶段面临的生成内容不可控问题的主要原因在于,在生成过程中没有对生成的内容进行检测评估导致了生成的带问题内容也被输出。因此,未来需要在生成内容的检测评估算法方面进行大量的研究,有效阻止有问题内容的输出。

3）面向特定领域进行研究。一方面,AIGC在某些特定领域的表现差强人意。另一方面,当前性能优异的AIGC模型大多是面向许多领域的,使得这些模型在特定领域上的表现仍有较大提升空间。因此,未来需要面向各个特定领域进行针对性的模型研究,在提高模型性能的同时,也使模型具有更好的可应用性。

5　结论

生成式人工智能毋庸置疑地成为了现代科技发展中的一大亮点,它像一把双刃剑,既有着改变游戏规则的潜力,也伴随着不容忽视的风险和挑战。未来生成式人工智能的发展需要合理利用其所带来好处的同时,也要规避其潜在风险,需要不仅关注技术本身的发展,还要着手制定相应的监管对策、法律框架及伦理准则。

未来,学术界、工业界和政策制定者需携手协作,通过跨领域合作与对话,不断完善对生成式人工智能的理解与应用,共同构建一个既能促进技术创新,又能确保社会公正与个人权利得到保护的生态环境,以此推动和实现生成式人工智能技术的健康发展,使其成为推动人类社会进步的正向力量。

参考文献(References)

[1] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Advances in Neural Information Processing Systems, 2000, 13: 932-938.

[2] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Interspeech. Baltimore: Johns Hopkins University, 2010, 2(3): 1045-1048.

[3] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[5] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2024-03-04]. http://arxiv.org/abs/1810.04805.

[6] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-06-11)[2024-03-04]. https://openai.com/blog/language-unsupervised.

[7] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL]. (2019-02-14)[2024-03-04]. https://openai. com/blog/better-language-models.

[8] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[EB/OL]. [2024-03-04]. http://arxiv.org/abs/2005.14165.

[9] Ouyang L, Wu J, Xu J, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, 2022: 27730-27744.

[10] OpenAI. ChatGPT[EB/OL]. (2022-11-30) [2024-03-04]. https://openai.com/index/chatgpt.

[11] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report[EB/OL]. (2023-03-15) [2024-03-04]. https://arxiv.org/abs/2303.08774.

[12] Ding N, Qin Y J, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models [J]. Nature Machine Intelligence, 2023, 5: 220-235.

[13] Wu T, Luo L, Li Y F, et al. Continual learning for large language models: A survey[EB/OL]. (2024-02-07)[2024-03-10]. https://arxiv.org/abs/2402.01364.

[14] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[15] Kingma D P, Welling M. Auto-encoding variational Bayes[EB/OL]. [2024-03-15]. http://arxiv.org/abs/1312.6114.

[16] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. New York: ACM, 2014: 2672-2680.

[17] Karras T, Laine S, Aila T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 4401-4410.

[18] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway, NJ: IEEE, 2020: 8110-8119.

[19] Karras T, Aittala M, Laine S, et al. Alias-free generative adversarial networks[EB/OL]. [2024-03-15]. http://arxiv.org/abs/2106.12423.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. [2024-04-11]. http://arxiv.org/abs/2010.11929.

[21] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[EB/OL]. [2024-04-13]. http://arxiv.org/abs/2103.14030.

[22] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[EB/OL]. [2024-05-01]. http://arxiv. org/abs/2006.11239.

[23] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation[EB/OL]. [2024-05-01]. http://arxiv.org/abs/2102.12092.

[24] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[EB/OL]. [2024-05-05]. http://arxiv.org/abs/2103.00020.

[25] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. [2024-05-06]. http://arxiv.org/abs/2204.06125.

[26] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[EB/OL]. [2023-12-21]. http://arxiv.org/abs/2112.10752.

[27] David H. MidJourney[EB/OL]. (2022-07-12)[2023-12-21]. https://www.midjourney.com/explore.

[28] Esser P, Chiu J, Atighehchian P, et al. Structure and content-guided video synthesis with diffusion models [EB/OL]. (2023-02-06) [2024-05-11]. https://arxiv. org/abs/2302.03011.

[29] Demi G. Pika[EB/OL]. (2023-11-29)[2024-05-11]. https://pika.art.

[30] Zhang S, Wang J, Zhang Y, et al. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models[EB/OL]. [2023-11-07]. https://arxiv.org/abs/2311.04145.

The research status and development trends of generative artificial intelligence

CHE Lu[1],ZHANG Zhiqiang[2],ZHOU Jinjia[3*],LI Lei[3]

1. School of Environment and Resource, Southwest University of Science and Technology, Mianyang 621000, China
2. School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang 621000, China
3. Faculty of Science and Engineering, Hosie University, Tokyo 184-8584, Japan

Abstract With the advent of ChatGPT, the research of generative artificial intelligence (GAI) has made a breakthrough in the field of multimodal information processing, such as text, image, and video, and has attracted broad attention. This paper aims to systematically review the research progress of GAI and to discuss its future development trend. Being divided into three parts, the paper first reviewed the development history and research status of GAI in terms of natural language models, image and multimodal models; secondly, it discussed the application prospects of GAI in different fields, mainly focusing on content communication, assisted design, content creation, personalized customization, and etc. In the third part, with an in-depth analysis of the main challenges facing GAI, the author summarized the development trends of GAI in the future.

Keywords artificial intelligence generated content; natural language; multimodal

（责任编辑　王微）