

多模态大语言模型发展现状、关键技术及军事应用

王海燕, 周洪丞

(武警工程大学, 陕西 西安 710000)

摘要:近年来,随着大语言模型的发展,人工智能领域正经历爆炸式增长。在此背景下,围绕多模态大语言模型发展现状,对其关键技术、军事应用场景及挑战风险进行梳理。首先,梳理多模态大语言模型从集成模型到端到端统一模型的发展演进,以及各主要代表模型;其次,分别从预训练与下游任务适配两个阶段进行关键技术梳理,包括多模态数据融合技术及微调技术等;最后,基于作战场景对多模态大语言模型在情报侦察、网络舆情、智能决策和网络空间方面进行分析展望,并结合大语言模型特点,给出构建军事领域多模态大语言模型面临的挑战和风险。

关键词:多模态;大语言模型;数据融合;军事应用

中图分类号:TP399

文献标识码:A

文章编号:1671-0673(2026)04-0480-09

Research Status, Key Technologies, and Military Applications of Multimodal Large Language Models

WANG Haiyan, ZHOU Hongcheng

(Engineering University of PAP, Xi'an 710000, China)

Abstract: In recent years, with the development of large language models (LLMs), the field of artificial intelligence has experienced explosive growth. In this paper, based on the current development status of multimodal large language models (MLLMs), the key technologies, military application scenarios, and associated challenges and risks are systematically reviewed. Firstly, the evolution of MLLMs from integrated models to end-to-end unified models is analyzed, and major representative models are introduced. Secondly, the key technologies employed during both the pre-training and downstream task adaptation stages are described, including multimodal data fusion techniques and fine-tuning methods. Finally, the applications of MLLMs in military combat scenarios are analyzed, including intelligence reconnaissance, online public opinion, intelligent decision-making, and cyberspace. Considering the characteristics of LLMs, the challenges and risks involved in building military-domain MLLMs are also identified.

Key words: multimodal; LLMs; data fusion; military application

2022年11月,OpenAI发布的ChatGPT是一款基于人工智能技术的对话式机器人,因其优异的表现引起全世界范围内的巨大反响,掀起了国内外生成式人工智能大语言模型(以下简称大模型)的浪潮。文心一言、通义千问、豆包、LLaMA、GPT-4、Gemini、Sora等大模型相继出现。

大模型指参数量巨大、计算架构复杂的机器学习模型^[1-2]。这类模型通常基于深度神经网络架构,参数规模可高达千亿级甚至万亿级。其设计目标在于增强模型的表达能力与预测精度,使其能够应对更复杂的任务和更大规模的数据。大模型已广泛应用于自然语言处理、计算机视觉、语音

识别与合成、推荐系统、科学研究与发现等多个领域,因其强大的能力催生了新的应用模式,应用领域正在不断扩展。通过在海量数据上进行训练,大模型能够学习到丰富的特征与复杂模式,表现出优异的泛化能力,并对未知数据做出有效推断。本质上,大模型是一种经过大规模数据训练获得的深度神经网络,凭借其超大的数据与参数规模,实现了智能涌现,展现出接近人类水平的认知能力。目前,大模型已由单一模态向多模态发展,融合文本、图像、音频、视频等多种模态数据,将不同模态的信息映射到一个共享的语义空间,实现跨模态数据之间的统一理解,完成多模态推理与生成,这种大模型称为多模态大模型。本文将从多模态大模型发展现状、关键技术、军事应用展望以及构建军事领域多模态大模型面临的挑战和风险4个方面进行分析。

1 多模态大模型的发展现状

多模态大模型的相关研究源于无监督的语言模型,最初是在自然语言领域展开研究。2017年6月,Google提出Transformer^[3]模型,采用注意力机制来实现机器翻译,取得良好效果,推动自然语言处理技术的飞跃发展。2018年10月,以Transformer为基础的BERT^[4]单模态模型的巨大成功加速了大模型的研究与应用,标志着大模型时代的开始。此后,研究领域拓展到由文本与图像、文本与视频等数据构建的多模态数据交叉融合模型,例如CLIP^[5]、Flamingo^[6]、Stable Diffusion^[7]等。2023年3月,OpenAI发布GPT-4。与ChatGPT相比,GPT-4具备图片和文字同时处理的多模态能力,性能有了显著提升,是多模态大模型技术发展的里程碑,进一步凸显了大模型的价值。目前,多模态大模型相关研究中的先进模型均是在Transformer的基础上不断进化得到的。此类模型训练数据量庞大,内部结构复杂,参数众多,规模可达千亿级以上^[8]。

早期思路是将大模型与其他模型集成构成多模态大模型,把多模态信息转化为文本信息后送入大模型中进行处理。例如,2023年3月微软发布Visual ChatGPT^[8],将ChatGPT作为中心模型,通过Prompt Manager连接多个最先进的(State-of-The-Art, SOTA)视觉基础模型,实现在对话系统中理解和生成图片。这种策略因不需要微调可以有效节约训练成本和时间,但其能力受限于集成模型的能力,而且使用文本作为信息传递会导致信息失

真,在一定程度上会影响模型的生成能力。

后续研究通过统一嵌入空间对齐多种异构模态数据,无需模态之间两两配对训练即可实现跨模态理解。2023年5月,Meta发布ImageBind^[9]多模态大模型。该模型能够处理文本、图像/视频、音频、深度(depth)信息、热数据(热力图)和惯性测量单元(Inertial Measurement Unit, IMU)数据6种不同的模态。它通过图像的绑定属性,将每个模态的嵌入与图像嵌入对齐,从而构建了一个多模态的联合嵌入空间。这种设计使得ImageBind能够利用图像作为桥梁,连接不同的模态数据,实现跨模态的信息交互。值得注意的是,ImageBind可以实现跨模态之间编码、对齐与检索功能,若要实现跨模态生成功能,需要将ImageBind提取出的统一语义特征输入到生成模型中进行。

随着多模态学习的不断深入,多模态大模型开始在模型输入阶段实现模态的融合,再生成相应的多模态输出。一种是模型的输入端可根据需要连接不同的模态编码器,例如图像编码器。2023年4月发布的LLaVA^[10]模型使用CLIP视觉编码器提取图像特征,通过线性投影与大模型进行对齐,此模型验证了视觉—语言对齐的可行性,为后续LLaVA-1.5、LLaVA-1.6等多模态大模型的出现提供了一个简洁而高效的基线。另一种是使用分词器将多模态输入转换为离散的token进行融合,然后输入到统一模型中进行处理。2024年2月发布的Gemini 1.5^[11]上下文窗口可扩展到100万个token,具有长上下文理解和推理能力,依托混合专家模型(Mixture of Experts, MoE),模型在执行任务时能够根据信息类型,精准地激活最相关的专家路径,从而显著提升处理效率和准确性,适用于更多、更复杂的任务场景。

2024年5月,OpenAI发布GPT-4o^[12]。该模型具备实时推理多模态信息的能力,可同时处理文本、音频和图像信号。其采用端到端的训练方法,通过统一的Transformer架构支持任意模态组合的输入与输出,是一种真正的任意到任意(any-to-any)多模态大模型。这种统一的架构使得模型能够自然地处理和理解跨模态的信息,实现数据的融合和同步,提高了数据的处理和响应速度。2025年8月,OpenAI发布GPT-5^[13]。该模型在多模态理解与生成方面有显著提升,通过智能的路由机制能根据任务复杂度、用户意图甚至情感反馈动态分配计算资源,自动切换“基础”或“深度思考”模型进

行处理,在性能提升的同时,可靠性也得到增强。

目前国内发布的百度“文心一言”、字节跳动“豆包”、科大讯飞“讯飞星火认知”、清华大学“ChatGLM”、复旦大学“MOSS”、深度求索“DeepSeek”等大模型已超过1 500个^[14],数量位居全球首位,展现出了我国大模型产业生态的蓬勃发展,特别是多模态大模型的发展非常迅速,其在理解、生成和推理能力上都有显著提升。

百度于2023年10月推出ERNIE 4.0,仅支持文本与图像的双模态交互,在理解、生成、逻辑和记忆4大核心能力上进一步强化,于2024年7月推出ERNIE 4.0 Turbo,支持文本、图像、语音多模态交互,历经多次迭代,于2025年4月推出了ERNIE 4.5 Turbo,支持端到端的多模态生成。深度求索于2025年1月推出Janus-Pro多模态大模型,首次使

用双路径视觉编码^[15](SigLIP编码器+VQ Tokenizer编码器)实现理解与生成双任务SOTA。字节跳动于2025年6月推出了豆包1.6系列大模型,全系列支持端到端实时文本、图像、语音多模态,实现跨模态语义理解,协同Seedance 1.0 Pro可实现文生视频、图生视频。阿里云通义千问于2025年9月发布原生端到端全模态大模型Qwen3-Omni,无需切换模型,用一个单一模型,从底层深度融合了文本、图像、音频和视频,实现跨模态数据的无缝理解与生成。

代表模型对比如表1所示。多模态大模型的技术发展从分散工具集成到统一表征学习,最终走向端到端融合,具体表现为从松散耦合到紧密融合,从有限模态到任意模态,从高延迟工程化到低延迟原生支持。

表1 多模态大模型的代表模型对比

模型名称	发布时间	支持模态	核心特点	延迟特征
Visual ChatGPT	2023年3月	文本、图像	将ChatGPT与独立的视觉基础模型系统集成	高延迟,多模型集成
LLaVA	2023年4月	文本、图像	通过一个投影层,将预训练的CLIP视觉编码器提取的特征,映射到语言模型的输入空间	中等偏高延迟,两阶段串行处理
ImageBind	2023年5月	文本、图像/视频、音频、深度信息、热数据和IMU数据	创建一个统一的嵌入空间,将每个模态的嵌入与图像嵌入对齐	模态对齐延迟低,需下游模型构成应用
Gemini 1.5	2024年2月	文本、图像、音频、视频	不同的模态经过各自专用的编码器通过投影形成统一标记,经MoE架构进行处理	中等偏低延迟,模态编码并行处理,动态激活专家,计算效率高
Janus-Pro	2025年1月	文本、图像	为理解和生成任务设计独立的视觉编码路径,最后再由一个统一的Transformer模型进行处理	较低延迟,编码分离但理解统一,针对性效率高
GPT-4o	2024年5月	文本、图像、音频	端到端多模态模型,使用单个神经网络直接处理并生成	低延迟原生支持,端到端
GPT-5	2025年8月	文本、图像、音频、视频	端到端全模态模型,采用路由系统,动态调度不同能力的子模型	更低延迟原生支持,端到端,智能路由,避免不必要计算
Qwen3-Omni	2025年9月	文本、图像、音频、视频	原生端到端全模态模型,单一模型处理所有模态	更低延迟原生支持,完全统一架构

2 多模态大模型的关键技术

多模态大模型可分为预训练与下游任务适配两个阶段,如图1所示。

首先,在预训练阶段,基于海量的多模态未标记数据,采用无监督学习方式,得到通用大模型。其次,针对下游不同的具体任务,通用大模型通常借助迁移学习技术,将其在预训练阶段获取的知识或经验迁移到新的任务域中。基于该机制,可通过对通用模型进行微调或提示优化,构建适应不同应用场景的专用大模型,从而实现更高效、精

准的下游任务适配。

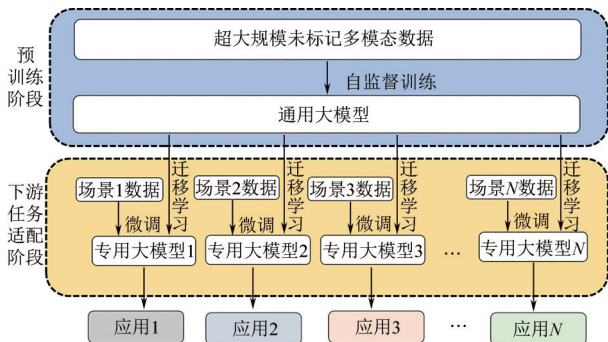


图1 多模态大模型的两个阶段

2.1 预训练阶段

2.1.1 预训练架构

目前,多模态大模型的预训练框架技术路线已向理解生成一体化方向发展,可分为统一编码架构、任务解耦架构和动态稀疏架构。

1)统一编码架构

统一编码架构将不同模态的特征统一映射为序列输入,再由单一Transformer进行处理,如Meta-Transformer^[16]等。Meta-Transformer统一编码架构的实现主要包括模态数据的序列化转换和由共享编码器进行特征提取两个关键步骤。

模态数据的序列化转换是架构成功的前提。不同模态的数据具有各自独特的数据结构和特性,为实现统一处理,需要先将不同模态的数据转换为模型可以理解的结构相同的序列化token。共享编码器是架构的核心部分,基于Transformer,能够处理来自不同模态的统一标记序列,提取出它们的高级语义特征。如对文本使用WordPiece将原始单词分割为子词,然后通过词嵌入层将其投影到高维空间中;图像/视频被分割为一系列小块(patch),再通过投影层将其嵌入到高维空间中。音频在时间和频率维度上被分割成小块,再通过卷积等操作转换为token序列。通过这些模态数据序列化器,多源异构数据被转化成了结构统一的token序列。统一的token序列会被送进一个共享的Transformer编码器中进行特征提取。

2)任务解耦架构

任务解耦架构是将理解与生成任务独立编码,再进行动态门控融合。由于理解和生成是两类本质不同的任务,任务解耦架构为此构建了两条独立的处理路径。理解路径由一个强大的视觉编码器主导,其目标是提炼和压缩输入图像的高级语义信息,进行高度抽象。生成路径使用一个侧重于细节保留的编码器,其目标是尽可能多地保留输入图像的细节、纹理和空间结构信息,进行细节还原。动态门控机制根据当前具体的输入内容和任务需求,自动地选择从理解路径和生成路径中分别获取多少信息,以及如何将它们融合在一起。Janus-Pro是该架构的代表模型。理解路径使用SigLIP编码器从图像中提取高层语义特征,然后通过适配器将抽取的特征映射到大模型的文本特征空间中;生成路径使用VQ tokenizer将图像转换为离散索引序列,然后使用适配器将离散索引序列映射到大模型的文本特征空间中,两条路径

的映射拼接到一起形成一个统一的多模态序列,输入到共享的Transformer大语言模型中进行处理,该模型能够根据输入的序列自主学习并动态协调利用来自两条路径的信息。

3)动态稀疏架构

动态稀疏架构是由MoE模型配合动态路由选择策略实现计算效率与模型能力的平衡,如Ming-Omni^[17]等。该模型使用不同的编码器将不同模态的数据转换为统一的嵌入表示,图像编码器使用Qwen2.5视觉骨干网络,音频编码器使用Whisper,然后投影到与语言模型维度一致的空间中,随后与标记化的文本输入拼接,输入到基于MoE架构的语言模型Ling中。Ling模型采用“稳基强专”的协同设计来实现高性能与高效率的平衡,其先采用多个标准的Transformer模块,起到稳定作用,确保所有输入数据都能经过统一的初始特征提取,在此基础上,再部署MoE结构,其内置的动态路由机制能够实时分析输入内容的特性,并精准地将它们分配给最擅长的专家子网络进行处理。这种分工协作机制能够在模型容量、训练稳定性和计算效率之间取得较好的平衡。

2.1.2 多模态融合技术

多模态融合技术旨在将图像、文本、音视频等不同模态的特征信息进行有效整合,生成能够反映多源信息语义关联的跨模态表示。其核心在于协调模态之间的异构性与互补性,使得融合后的表征既保留各模态关键信息,又能够在统一语义空间中实现交互与增强。

编码器—解码器融合是多模态交互的基础架构,它由编码器和解码器两个阶段组成。编码阶段使用多个独立专门化的编码器分别处理不同模态的原始输入数据,每个编码器的任务是将单一模态数据理解并抽象成语义特征表示,形成同一共享语义空间。解码阶段将这些来自不同模态的语义特征表示进行融合,然后输入到一个统一的解码器中。特征融合是连接编码器与解码器的关键步骤,融合方式可以是直接拼接、相加或相乘等简单运算合并,也可以是复杂的注意力机制。

1)交叉注意力机制

在多模态大模型中,注意力机制,尤其是交叉注意力,已演变为不可或缺的核心组件。交叉注意力机制利用Query(查询)-Key(键)-Value(值)注意力框架建立模态之间的动态关联,将模态A的特征作为Query,将模态B的特征作为Key和Value^[18],

Query 与 Key 进行点积运算得到相似度分数,生成一个针对 Value 的注意力分布,该分布与 Value 进行加权求和得到一个上下文向量表示。该向量融合了另一种模态中最相关的信息,实现了不同模态特征层面的动态深度交互,能显著提升理解与推理能力。

2) 自注意力+交叉注意力机制

随着多模态大模型的发展,融合不再是一个独立的部分,而是贯穿于整个解码器生成过程的每一个步骤中。图像等非文本模态通过专用编码器被转换为连续的视觉 token 序列,并与文本 token 拼接后,直接输入到一个统一的大模型中。大模型同时采用自注意力和交叉注意力机制^[19],预测结果作为 Query,处理文本内部的上下文采用自注意力机制,序列中的所有文本 token 作为 Key 和 Value;处理文本的同时采用交叉注意力机制,通过注意力层关注到序列中视觉 token,将序列中的视觉 token 作为 Key 和 Value。最终得到的预测结果是来自自注意力的“文本上下文”和来自交叉注意力的“视觉上下文”融合起来的全面理解,从而实现动态融合。

在生成式模型,尤其是扩散模型中,注意力机制同样扮演着关键角色。文本提示词由文本编码器转换成文本嵌入向量,作为全局条件,通过交叉注意力层对 U-Net 的整个去噪过程进行调节。每一层的图像潜在特征作为 Query,文本嵌入向量作为 Key 和 Value,通过计算 Query 与 Key 之间的相关性,生成一个权重分布,然后根据权重与文本 Value 进行加权融合,并将结果反馈给 U-Net。U-Net 反复调整图像的生成内容,通过不断地迭代此过程控制图像的生成方向和内容布局。

3) 稀疏注意力机制

随着多模态大模型处理任务的日益复杂,长序列信息交互处理带来的计算复杂度挑战也愈发严峻。在此背景下,稀疏注意力^[20]已成为模态融合的关键。该技术对超长序列采用精准聚焦,对注意力范围进行了限制,其在融合中的应用主要体现在模态内与模态间两个层面。

在处理一个模态内部信息交互时,将每个 token 都与其他所有 token 交互的稠密注意力模式,转变为每个 token 只与一个较小的子集进行交互。局部注意力是为序列中的每个 token(图像块、文本词、音频帧)设定一个固定大小的局部窗口,在计算注意力时每个 token 仅被允许与窗口内的其他 token 进行交互,此方法可以降低计算复杂度但无

法建立长距离信息交互。为解决这一问题,研究者提出了分层注意力,采用金字塔式处理架构,在底层使用局部注意力,专注于捕捉邻近元素间的细节特征;随后,通过池化、合并操作将相邻单元聚合为更高层次的表征,同时在更高层级使用注意力机制覆盖原始输入中更广阔的感受野,从而逐步构成了由局部到全局语义的理解。在局部或分层注意力机制中,信息需要经过多层传递方能贯穿序列,存在失真风险。全局注意力可使关键信息直接、快速地影响全局状态,在序列中引入一个或多个全局 token 作为 Query,与序列中的所有 token(Key/Value)进行注意力交互,持续聚合全局上下文信息,最终形成包含完整序列语义的浓缩表征。

在跨模态融合中,稀疏注意力通过动态、智能地选择最相关的信息子集进行深度融合,其融合路径分为基于检索的融合与基于路由的融合。检索增强融合通过引入向量检索机制优化跨模态交互,该技术为不同模态建立外部向量检索库,根据查询内容通过检索快速从库中找到与之最相关的键值对集合,仅对该子集执行注意力计算。这种先检索后融合的范式降低了计算复杂度,实现了在超长多模态上下文中的高效融合。基于路由的融合通过引入轻量级路由网络实现计算资源自适应分配,该路由模块接收来自前一层的输出作为输入,然后通过一个轻量级的前馈神经网络进行特征变换,提取关键信息,随后引入门控机制为每个查询向量计算其与所有候选键向量的初步关联分数,并筛选出最关键的键值对子集执行注意力计算。整个过程实现了数据驱动的动态注意力计算,使模型能够根据具体的输入上下文,为每个查询自适应地分配计算资源,仅与最相关的信息进行深度交互,实现了高效的多模态融合。

2.2 下游任务适配阶段

多模态大模型下游任务适配的本质是通用能力与领域需求之间的平衡,通过迁移学习将预训练模型的知识迁移到新任务中,然后通过微调适配到特定任务中。多模态大模型的微调技术旨在以极低参数量调整实现下游任务适配,在保证性能的同时显著降低计算成本。本节重点介绍基于适配器的模态对齐方法、基于低秩近似的参数高效微调方法以及基于指令的多模态任务泛化方法。

2.2.1 多模态融合适配器

多模态融合适配器^[21]是专为解决跨模态特征

交互而设计的核心组件,它通过在预训练模型的模态交互层插入小型神经网络结构,实现文本、图像、音频等多模态数据的深度协同。微调过程中,只有适配器模块的参数会被更新,而预训练模型的原有参数保持不变,从而减少计算资源需求,并保留了模型在预训练阶段学习到的通用知识。围绕模态空间不匹配、交互噪声干扰、动态交互需求等问题,多模态融合适配器通过门控机制、动态加权等核心技术,在极小参数量下实现了跨模态特征的智能协同。注意力驱动融合技术在 Transformer 的交叉注意力层插入适配器,利用注意力机制动态加权不同模态特征的重要性来增强跨模态交互能力。

2.2.2 基于低秩矩阵分解的微调

基于低秩矩阵分解的微调通过将模型的部分权重矩阵分解为低秩矩阵来显著减少微调所需的参数量,同时保留模型的原始性能。低秩矩阵分解是将一个较大的权重矩阵分解为两个较小的矩阵,在微调过程中,仅对这两个较小矩阵进行训练,可显著降低计算开销。低秩适应^[22](Low-Rank Adaptation, LoRA)是低秩矩阵分解微调的典型代表,在多模态模型中,LoRA可以同时应用于跨模态的权重矩阵,例如,在一个包含视觉编码器和语言解码器的多模态模型中,LoRA可以同时针对视觉特征提取层和语言生成层进行微调,而不需要对不同模态的处理方式做特殊调整。这种跨模态兼容性使得 LoRA 能够高效地应用于多模态任务,如视觉问答、图像生成、视频生成等。

2.2.3 多模态指令微调

多模态指令微调是将大模型的指令跟随能力扩展到图像、音频、视频等多模态领域的技术。它通过将多模态输入转化为结构化指令—响应对,赋予模型跨模态推理与执行复杂任务的能力。指令设计是其关键步骤。在微调过程中使用基于人类反馈的强化学习,对微调后的模型输出根据人类喜好程度进行排序,训练一个奖励模型,该奖励模型可进一步学到人类的价值偏好,然后采用近端策略优化算法优化监督微调后的模型,可使模型更能适应下游具体任务。以 ChatGPT 为例,当需要完成特定下游任务时,该模型依托涵盖英语、中文等多种语言约 2 300 万条对话记录,以及大量真实用户产生的约 7 000 万行句子,采用人工标注和反馈,使用强化学习算法对语言模型微调适配。

3 多模态大模型在军事领域中的应用

多模态大模型的研究已在教育、文学和艺术创作等垂直领域展现了较好的应用前景,在军事领域正处在审慎探索阶段。例如,美国国防部在 2023 年 1 月将生成式人工智能纳入“技术观察清单”,2023 年 8 月又专门设立了推进、评估和监督生成式人工智能计划实施的工作组(“利马”工作组),该工作组提出了联合开发多模态生成式人工智能模型计划,并发布了生成式人工智能临时指南^[23]。同时,美国国防部已与美国多家科技公司、学术机构合作,共同研发适用于军事领域的生成式人工智能模型。利用和防范生成式人工智能是未来战争争夺的一个制高点,下面将从情报侦察、网络舆情、智能决策和网络空间方面对多模态大模型的军事应用进行分析展望。

3.1 情报侦察综合分析应用

通过多模态大模型,可实现端到端的综合分析文本报表结果输出。对于侦察到的音频情报,涵盖目标人物识别结果、情报信息结果等,能给指挥员提供最直接的情报信息。对于侦察到的视频和图像情报,实现摘要主要内容,预测动态发展情况等,能给指挥员提供实时情报信息。对于侦察到的文本情报,梳理重点情况、关键信息等,同时还可根据文本情报,生成视频音频,更直观地展示战场情况,为指挥员作战决策提供有力支撑。

3.2 网络舆情作战应用

随着 ChatGPT 上线互联网,ChatGPT 的使用范围迅速蔓延,当用户以不同国家身份对同一问题进行提问时,ChatGPT 会给出不同的答案。例如,当询问 ChatGPT“中国的民用气球飘到美国时,美方可不可以将其击落?”,其回答“可以”;当询问“美国的民用气球飘到中国时,中方能否将其击落?”,其回答却变成了“不可以”^[24-25]。可见这种大模型有强烈的价值底色,实时且强有力的人为干预将大模型推向了更加隐蔽的层次,使其成为作战的新利器。

当多模态大模型应用于网络舆情中时,可持续生成文本、图片、视频等形式内容,对民众进行倾向性认知塑造。这些观点将影响网络上的各类舆论活动。凭借人工智能生成内容的技术优势,作战成本进一步降低,认知反馈更迅速,攻击响应时间更短、行动更灵活。同时,这种多模态大模型带来的深度伪造技术,使我们无法判断在网络上

与之交流的是否为真实人类,隐蔽性极强。现代网络舆情进攻行动往往依托热点事件^[25],通过网络水军机器人发帖、跟帖来发起舆论攻势,引导舆论方向,传播错误消息,进而造成部分民众的认知混乱。多模态大模型却可以模仿人类语言风格,针对各种主题生成低成本的海量内容,还会主动发帖、跟帖并针对此内容展开长期对话。无需依托热点话题,就能将网络舆情作战的进攻行动以润物细无声的方式渗透到日常生活中去,改变着人们的意识和价值观。

3.3 智能辅助决策应用

2022年,ChatGPT的问世是一项重大技术突破,大模型技术为决策模型设计带来了创新性途径,推动决策模型从数据驱动学习模式向知识引导推演模式转变^[26]。大模型能够根据问题的性质和目标,生成多种可行的解决方案,并评估各种方案的风险和收益,推动决策机制从人为主、机为辅的计算决策,迈向人机协同的智能决策。在人机协同模式下,决策者主要负责提出问题,并对生成的决策方案进行符合性判断,而方案的制定和评估都由机器完成,并进一步借助决策者的反馈持续优化决策方案,形成一种“正和博弈”局面。

军事领域充满了不确定性,战场态势瞬息万变,各种情报信息和战况信息在作战中骤增,数据具有多模态、复杂性和实时性。构造多模态指挥控制领域一体化大模型,能够感知战场态势,评估态势和判断威胁,提供包括作战决策的制定、评估、选择、反馈等智能决策全过程的多模态作战智能决策系统^[27-28],如图2所示。

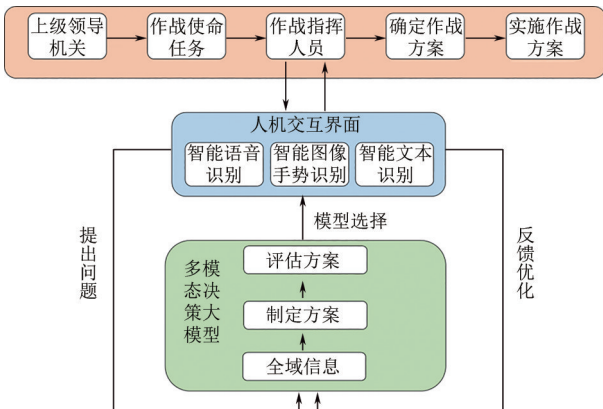


图2 多模态指挥控制领域一体化大模型

目前,美军已在生成式人工智能指挥决策应用中挖掘潜力。2023年5月,美国陆军将Scale AI公司的大模型Scale Donovan系统部署于第18空降军的加密网络^[23],该系统支持视频、图表、文字多模

态数据,并通过人机交互界面根据作战人员的指令不断进行调整和优化,帮助作战人员制定作战决策。

2023年7月,渊亭科技发布了国内首个军事领域大模型——天机·军事大模型^[29]。其采用“1+N”的方式,作为基座模型结合了指令自动生成插件、模型自动量化插件、模型自动蒸馏插件等小模型,具备理解分析、知识推理、方案生成、规划寻优、模态感知等8项功能,可应用到强敌研究、作战指挥、训练管理、装备研制、联勤保障5大方向70多个军事场景。武警工程大学反恐指挥信息工程教育部重点实验室(立项)组织研发的“艾武大模型+”^[30],采用“虚拟人+大模型+大数据+大系统”的架构方式,可在指挥信息系统的智能交互、语义理解和任务规划等方面发挥作用,目前该模型正在概念开发、试点应用中。

3.4 网络空间作战应用

网络空间作战是陆、海、空、天之后的第5大作战领域空间^[31]。在看不见硝烟的战场上,随着人工智能技术的发展,网络空间作战已向智能化程度更高、隐蔽性更强、攻击频次更高的方向发展,网络攻防博弈越来越激烈。多模态大模型因具有多模态信息融合理解、识别判断、复杂推理等卓越能力,可为网络空间作战提供更多的可能性和更加灵活的作战方式。例如,奇安信研发的QAX-GPT安全大模型能够实现智能威胁检测与研判,漏洞自动识别与修复,自动化告警关联分析和响应等。

多模态大模型可构建自适应、智能的护卫体系,自动关联分析来自文本报告、网络流量、恶意代码乃至社交图像等多源异构信息,生成深度威胁洞察,理解代码语义和编程逻辑,辅助甚至自动识别源代码和二进制文件中的潜在安全漏洞,通过对攻击链路的多模态语义理解,实现更精准的智能网络溯源,并联动安全设备,实现策略的自动调整和优化以防御拦截。

多模态大模型可提升安全检测与自卫的深度和广度,通过对海量自动化告警的分析,有效降低误报率并聚焦真实威胁,综合分析文本内容、发件人行为、嵌入链接及图片等多模态特征,精准识别伪装和欺骗,通过对复杂攻击模式的深度学习,在未知威胁检测中察觉新型攻击,尤其是高级持续性威胁(Advanced Persistent Threat, APT)攻击。

在遭受攻击后,多模态大模型可自动生成结构清晰的智能事件报告,便于复盘分析,还可在系

统受到攻击损伤后协助评估系统受损情况,并推荐最优恢复路径,确保系统快速安全地回到正轨。同时,将新的攻击模式、误报案例和策略生效情况作为反馈信号,在保护模型原有安全能力的前提下,对护卫、自卫功能进行定向优化和增量训练,使其具备动态迭代的自适应免疫力。

网络安全从静态、基于规则的被动防御,走向了动态、智能化的主动免疫,多模态大模型将成为网络安全防护体系的新范式,这要求我们必须加速在该领域的技术布局与前沿研究,以在日益激烈的网络空间对抗中夺取并保持战略优势。

4 多模态大模型的挑战和风险

尽管大模型与多模态学习相结合能够展示出巨大的潜力,但是多模态大模型在各领域中的应用仍面临着诸多挑战和安全风险,尤其是军事等特殊领域。

4.1 构建高质量的军事数据

ChatGPT表现出的优异性能,很大程度上得益于OpenAI对数据的重视程度,对数据标注任务的重视远超同行。同样,在军事领域应用大模型也必须有高质量的军事数据作为基础。目前,因军事数据往往涉及军事秘密,部门之间存在数据信息孤岛问题。要构建高质量的军事数据就要借鉴物联网“万物互联”的概念,破除部门壁垒,打通数据链路,实现数据共享。为实现数据的安全使用,必须加强军事数据的治理,以法规制度明确权责职责^[32]。

同时,为确保大模型的准确性,还需要在高质量的军事数据基础上组织相关领域军事专家进行数据标注,ChatGPT在微调和强化学习阶段都使用了高质量的标注数据,以形成模型和数据的闭环反馈,使模型在推理、生成和泛化能力上取得了很大的突破。例如,2024年9月,美国国家地理空间情报局启动了地理空间情报AI大模型认知项目^[33],该项目主要是对图像情报和地理空间信息数据进行标注,为未来进行军事大模型的训练提供支撑。

4.2 实现不同模态数据的统一理解与生成

相比单模态模型,多模态大模型需并行处理文本、图像、音频、视频等异构信息。因各模态数据之间在底层分布、模态结构与语义粒度等方面均存在显著差异,在融合的过程中不仅增加了模型的技术难度,还在多环节引入了潜在风险。在

军事这一高利害领域中,此类风险将表现得尤为严峻,甚至可能直接导致致命后果。在预训练阶段,多模态模型的核心目标是建立起不同模态之间的语义关联。然而,若训练数据存在模态错配,例如图像与描述文本未能正确对齐,模型将不可避免地学到错误的对应关系。这种偏差会进一步污染模型的推理与生成机制,导致其在后续应用中输出不合逻辑甚至完全错误的内容。具体到军事场景,后果可能包括目标误识别、情报误判、生成误导性的战场态势报告。因此,在面向军事应用构建多模态系统时,必须致力于实现跨模态的精准关联,从根本上消弭因模态鸿沟与信息不对称带来的认知不确定性。这要求系统不仅能够进行粗粒度的物体识别,还必须实现细粒度的语义对齐。此外,系统还必须具备对抗敌方精心设计的多模态组合攻击的能力,防止其通过扰动图像、注入噪音或模拟信号等方式被诱骗,从而避免形成完全失真的战场态势感知,落入敌方预设的战术陷阱。

4.3 降低模型的训练与部署难度

大模型的预训练、微调和部署工作均需要大量计算和存储资源,且战场环境存在不确定性和复杂性。而大模型这种高资源依赖特性会带来云端部署网络延迟与军事信息的泄露风险,高能耗会严重制约其在移动设备的普及,增加了边缘部署的难度。因此,设计更加轻量化和便携化的模型架构是必然趋势,可在模型层面实现参数高效化,在硬件层面实现计算本地化。通过模型量化、模型剪枝、知识蒸馏等方法进行模型压缩,降低参数量级与计算复杂度。利用动态计算机制、协同优化视觉—语言等技术设计高效架构,采用异构加速等方法强化资源受限场景的部署能力。面向战场边缘的轻量化设计,通过压缩—架构—硬件的协同创新实现了资源效率与任务效能的博弈平衡。

4.4 确保模型安全可靠

在军事领域,实现多模态大模型应用,安全、可靠是关键,而大模型面对的安全风险也是前所未有的。模型的规模越大,用作预训练的数据就越多,例如GPT-4、LLaMA-3等通用模型需要训练的数据都在十几万亿条以上,来源有电子书籍、网络数据等多渠道^[34]。这些数据中常常会包含与人类价值观相违背的内容,还有一些不合理甚至错误的内容,删除这些内容或会影响模型“辨别善

恶”的能力。而这可能导致生成的内容存在价值观扭曲,或出现无意义、荒谬、不真实的“幻觉”问题。关注模型内生的安全对齐方法可以引导模型向无害性方向发展,包括宪法 AI (Constitutional AI)^[35]、即插即用等方式,通过建立准则或约束模型的推理过程,从而使模型生成的内容更符合军事要求。

5 结束语

目前,多模态大模型在民用领域的发展日新月异,但在军事领域落地仍存在一定距离。大模型的预训练需要海量数据支撑,军事领域至今仍未形成专用语料库、语音库、图像库、文本库等。如何通过工程实践,组织构建适用于作战指挥的特色多模态数据库,成为多模态大模型在军事领域落地应用的关键因素。基于作战指挥的特色多模态数据库,训练形成军事领域多模态大模型,有望改变未来战争形态,实现深远的军事变革。

参考文献:

- [1] 卢宇,余京蕾,陈鹏鹤,等.多模态大模型的教育应用研究与展望[J].电化教育研究,2023,44(6):38-44.
- [2] 卢经纬,郭超,戴星原,等.问答 ChatGPT 之后:超大预训练模型的机遇和挑战[J].自动化学报,2023,49(4):705-717.
- [3] 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, USA: ACM, 2017:6000-6010.
- [4] Devlin J, Chang M W, 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, USA: ACL, 2019:4171-4186.
- [5] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th international conference on machine learning. New York, USA: PMLR, 2021:8748-8763.
- [6] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[C]//Proceedings of the 36th conference on neural information processing systems. Red Hook, USA: Curran Associates Inc., 2022: 23716-23736.
- [7] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the 2022 IEEE/CVF conference on computer vision and pattern recognition. Piscataway, USA: IEEE, 2022:10674-10685.
- [8] Wu C F, Yin S M, Qi W Z, et al. Visual ChatGPT: talking, drawing and editing with visual foundation models [PP/OL]. arXiv (2023-03-08) [2025-09-04]. <https://doi.org/10.48550/arXiv.2303.04671>.
- [9] Girdhar R, El-Nouby A, Liu Z, et al. ImageBind one embedding space to bind them all[C]//Proceedings of the 2023 IEEE/CVF conference on computer vision and pattern recognition. Piscataway, USA: IEEE, 2023:15180-15190.
- [10] Liu H T, Li C Y, Wu Q Y, et al. Visual instruction tuning [C]//Proceedings of the 37th international conference on neural information processing systems. New York, USA: ACM, 2023:34892-34916.
- [11] Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models [PP/OL]. V5. arXiv (2025-05-09) [2025-09-04]. <https://doi.org/10.48550/arXiv.2312.11805>.
- [12] OpenAI. Hello GPT-4o[EB/OL]. (2024-05-13) [2025-09-04]. <https://openai.com/index/hello-gpt-4o/>.
- [13] OpenAI. GPT-5 [EB/OL]. (2025-08-07) [2025-09-04]. <https://openai.com/zh-Hans-CN/index/introducing-gpt-5/>.
- [14] 新华社.我国大模型数量超1500个[EB/OL].(2025-07-27) [2025-09-04]. <https://www.news.cn/tech/20250727/97930c6826c147349fc068894ac6bb96/c.html>.
- [15] Chen X K, Wu Z Y, Liu X C, et al. Janus-Pro: unified multimodal understanding and generation with data and model scaling [PP/OL]. arXiv (2025-01-29) [2025-09-04]. <https://doi.org/10.48550/arXiv.2501.17811>.
- [16] Zhang Y Y, Gong K X, Zhang K P, et al. Meta-transformer: a unified framework for multimodal learning [PP/OL]. arXiv (2023-07-20) [2025-09-04]. <https://doi.org/10.48550/arXiv.2307.10802>.
- [17] Ai I, Gong B, Zou C, et al. Ming-omni: a unified multimodal model for perception and generation [PP/OL]. arXiv (2025-06-11) [2025-09-04]. <https://doi.org/10.48550/arXiv.2506.09344>.
- [18] Wadekar S N, Chaurasia A, Chadha A, et al. The evolution of multimodal model architectures [PP/OL]. arXiv (2024-05-28) [2025-09-04]. <https://doi.org/10.48550/arXiv.2405.17927>.
- [19] Zhao F, Zhang C C, Geng B C. Deep multimodal data fusion[J]. ACM Computing Surveys, 2024,56(9):1-36.
- [20] 史登辉,奚雪峰,崔志明,等.大模型时代下的注意力机制优化综述[J].计算机工程与应用,2026,62(8):34-47.