

多模联动，语义贯通

——大模型跨模态信息融合的创新技术路径

李少华，张正

中国人民武装警察部队工程大学，陕西 西安 710000

DOI: 10.61369/ME.2024090014

摘要： 人类在现实社会通过视觉、听觉、触觉等多种感官获取信息，不同模态信息相互补充印证，帮助人们理解和认知世界。人工智能大模型的跨模态融合能够有效提升复杂信息理解处理能力，使得人工智能更加接近人类智能水平。此次研究综合运用文献资料法、案例分析法和对比研究法探索大模型跨模态信息融合的创新技术，提出多模数据的获取与预处理、多模联动的模型架构设计、多模联动的训练与推理机制、语义理解与分析技术、语言对齐与融合技术等技术路径，为大模型实现多模联动、语义贯通的模态信息融合提供新的思路与方向。

关键词： 大模型；人工智能；跨模态；信息融合；创新技术

Multi-modal Interaction and Semantic Integration -An Innovative Technical Path for Cross-modal Information Fusion in Large Models

Li Shaohua, Zhang Zheng

Engineering University of the People's Armed Police Force of China, Xi 'an, Shaanxi 710000

Abstract : In the real world, humans obtain information through multiple senses such as vision, hearing, and touch. Different modalities of information complement and verify each other, helping people understand and perceive the world. The cross-modal fusion of large models in artificial intelligence can effectively enhance the ability to understand and process complex information, bringing artificial intelligence closer to human intelligence levels. This research comprehensively employs methods such as literature review, case analysis, and comparative study to explore innovative technologies for cross-modal information fusion in large models. It proposes technical paths including the acquisition and preprocessing of multi-modal data, the design of multi-modal interaction model architectures, the training and inference mechanisms of multi-modal interaction, semantic understanding and analysis techniques, language alignment and fusion techniques, and knowledge graphs and semantic integration. These provide new ideas and directions for large models to achieve multi-modal interaction and semantic integration in cross-modal information fusion.

Keywords : large models; artificial intelligence; cross-modal; information fusion; innovative technologies

引言

人工智能在信息技术高速发展的当下对社会各个领域的运作模式产生了深远影响，大模型作为人工智能领域的核心技术已经成为推动科技创新和产业升级的关键力量。随着大数据、云计算和深度学习技术的持续发展，大模型的训练数据规模与模型参数数量呈现指数级增长态势，这使得模型学习的知识内容更加全面丰富，庞大的模型参与赋予模型极强的表达能力与泛化能力，大模型在处理复杂任务时更加精准高效，为人工智能的持续成长发展奠定了坚实基础^[1]。跨模态融合能够将来自不同模态的数据整合分析后获取全面准确信息提示，他模型借此充分利用不同模态数据的互补性实现信息深度理解与综合分析。跨模态信息融合对提升大模型在自然语言处理、计算机数据等领域性能具有显著作用，同时也为大模型在智能驾驶、智能医疗、智能服务等等领域的应用提供了重要支持。

作者简介：

李少华（1991.12-），男，汉族，山西河曲人，学历：硕士研究生，职称：助教，研究方向：云计算、大模型应用。

张正（1993.03-），男，汉族，山西太原人，本科，助理工程师，人工智能。

一、多模联动的创新技术路径

（一）多模数据的获取与预处理

大模型跨模态信息融合中多模数据来源广泛，覆盖了多种类型和多个领域，数据不仅仅可以来自于互联网公开数据，同时还可以通过与其他不同领域的不同类型机构合作获取各种数据。例如医疗领域的医学影像数据就是重要的图像数据来源，医疗机构的影像数据库存储了大量 X 光、CT、MRI 等影像，这些影响对疾病诊断和治疗具有重要价值。实际应用中还可以通过传感器获取多模数据，例如智能驾驶车辆上的摄像头、雷达、激光雷达等传感器能够实时采集车辆周围环境信息，这些数据对自动驾驶系统的控制和决策至关重要^[2]。

数据清洗是多模数据预处理的关键环节，它可以去除数据中的噪声、错误和不完整信息，提升数据质量与可用性。例如通过 Hunspell、语言模型等拼写检查工具检测纠正错别字和语法错误，使用哈希算法判断文本是否重复并删除重复文本数据，采用降噪算法去除音频的噪音和干扰等等。数据标准能够为数据赋予语义标签，这对跨模态信息融合的作用非常显著，文本标注可以根据不同任务需求开展词性标注、命名实体识别、情感分析等标注；图像标注可以采用目标检测标注、图像分类标注、语义分割标注等等；音频标准则可以采用语音转文本标准和音频分类标准。

数据增强能够扩充多模数据并提升模型性能，图像数据增强可以通过集合变化和颜色变化实现，例如颜色变化能够调整图像的色调、亮度、饱和度和对比度，生成不同颜色风格图像。文本数据增强的常用方法包括同义词替换、随机插入、随机删除和回译等，例如同义词替换就是将文本中某些词语替换为同义词，保持文本语义不变的前提下增加文本多样性，使模型学习到不同表达方式下的语义理解。音频数据增强可以通过改变音频音量、速度、音调等参数来实现，例如加快或减慢音频播放速度能够让模型学习到不同语速下的语音特性，提高模型对语速变化的鲁棒性。数据增强技术能够扩充多模数据，增加多模数据多样性和复杂性，使模型在训练过程中学习到丰富特征，提升模型泛化能力与性能^[3]。

（二）多模联动的模型架构设计

统一 Transformer 架构在多模态联动中展现出显著优势，逐渐成为跨模态大模型的核心架构，其核心是自注意力机制，该机制使得模型在处理序列数据时动态关注输入序列的不同序列信息，其在机器翻译、文本生成、问答系统等自然语言处理任务中取得显著效果。统一 Transformer 架构能够统一编码处理不同模态数据，将文本、图像和音频等不同模态数据转化为统一特征表示，充分利用不同模态数据的互补信息实现跨模态语义理解和关联^[4]。统一 Transformer 架构的并行计算能力非常突出，该架构通过共同处理整个序列，不需要根据顺序依次处理各个时间步，这提升模型训练和推理效率，减少了资源与时间消耗。统一 Transformer 架构还具有很好的可拓展性，它通过堆叠多层 Transformer 模块构建更深层次模型，提高了模型表达能力与学

习能力，这使得模型能够适应不同规模与复杂程度多模态数据，持续提升跨模态信息融合效果。

多模态编码器和解码器分别负责多模态数据的编码和解码，实现跨模态信息处理和生成。多模态编码器能够将不同模态数据专户为统一特征表示以便后续模型处理分析，卷积神经网络是常用的编码器，它能够提取图像局部特征和全局特征，通过多层卷积和池化操作得到图像特征表示。多模态编码器需要解决不同模态特征融合问题，这可以通过不同模态特征的拼接或加权融合实现，也可以通过注意力机制动态融合不同模型特征，使模型可以根据任务需求自动调整模态特征权重。多模态解码器负责根据编码器输出特征生成相应输出，解码器在图像描述任务中可以根据图像特征表示生成描述图像内容的文本，根据文本特征表示生成对应图像，常用的多模态解码器包括 Transformer 解码器或循环神经网络解码器，前者主要通过自注意力机制和交叉注意力机制实现连贯准确输出，后者则通过循环结构加好的完成序列生成任务^[5]。

多模联动模型在寻求你过程中需要持续优化改进以提升性能和效果，随机梯度下降及其变种是常用的优化算法，例如 Adam 算法能够自适应调整学习率，通过合理调整算法超参数可以使模型更快收敛以提升训练效率。模型正则化也是常用优化手段，L1 和 L2 正则则在顺势函数中添加正则化项能够方式模型过拟合，提高模型泛化能力。为了进一步提升模型性能，可以在 Transformer 架构基础上提出变体，处理图像、长序列数据或长文本等等。此外，数据处理的增强也是优化模型的有效手段，通过多模态数据的精细预处理和增强能够增加数据多样性与复杂性，保证模型能够学习到更加丰富特征。

（三）多模联动训练与推理机制

多模联动模型训练依靠先进算法和策略完成海量多模态数据的有效学习和模型性能优化，随机梯度下降机器变种在多模联动训练中占据了重要地位，它在每次迭代中随机选择小批量数据样本，计算样本的损失函数梯度并根据梯度更新模型参数，这种算法能够在大规模数据集上快速收敛，在处理多模态数据集时不断迭代调整模型参数，使得模型能够学习到不同模态数据的关联与特征^[6]。Adagrad 是 SGD 的变种，它能够自适应调整每个参数的学习率，根据每个参数在以往迭代中的梯度平方和调整学习率，增大不常更新参数的学习率，减少频繁更新参数的学习率，这种自适应学习率调整机制时期在多模态数据处理时能够更好地平衡不同参数更新速度，提升模型训练效果。多模联动模型训练还可以通过数据并行和模型并行提升训练效果，数据并行将数据集分为多个小批次并分别在不同计算设备上训练，汇总更新计算效果，这可以充分利用多个计算设备的计算能力加快训练效果；模型并行将模型的不同部分分配到不同计算设备计算来提升训练效率。

多模联动模型的推理过程能够将训练好的模型应用于实际任务，根据输入多模态数据生成相应输出。将图像特征向量和文本语义向量输入到多模联动模型后，模型能够利用自注意力机制对两种模态信息进行融合与交互，寻找图像和文本的语义关联，模

型根据融合后的信息推理生成对问题的回答，模型在推理过程中根据训练学习到的知识和模式进行逻辑判断和语义理解，最终得到正确答案。推理效率提升对多模联动杆模型的实际应用非常重要，模型压缩技术能够减少模型参数数量和计算量，模型剪枝是常用的模型压缩方法，它能够删除模型中不重要的连接或神经元，减少模型复杂程度以降低计算成本。优化推理算法也能够提升推理效率，通过对自注意力机制等核心算法优化能够减少计算量和内存消耗，合理设计融合算法可以避免重复计算与冗余操作，保证模型更快处理多模态信息。

二、语义贯通的创新技术路径

（一）语义理解与分析技术

词法分析是自然语言处理的基础，它将连续文本流切成独立单词或要素，识别单词的语言学信息，基于统计的分词方法利用大规模语料库，通过统计模型学习词语出现的概率及上下文关系实现文本切分。句法分析能够分析句子的语法结构，确定句子中各个词语之间的依存关系或成分及结构，常用分析方法主要包括基于规则方法和基于统计方法。语义角色标注在句法分析基础上标注句子各个谓词的寓意角色，为自然语言处理任务提供了丰富语义信息，它主要以来机器学习算法和语义知识库实现^[7]。

语音识别能够将语音信号转化为文本形式，使计算机“听懂”人类语言，语音识别技术广泛用于智能语音助手、语音输入法、智能客服等场景。语音识别从特征提取入手，常用梅尔频率倒谱系数和线性预测倒谱系数；声学模型和语言模型是语音识别系统的核心，常用隐马尔可夫模型和深度神经网络将语音特征转化为因素或音节的概率分布；语音合成能够将文本转化为语音，为用户提供自然便捷交互方式，常用技术包括参数合成和波形合成^[8]。

（二）语言对齐与融合技术

跨模态语义将不同模态的数据在语义层面匹配关联，使模型

能够理解不同模态数据的内部关联。深度学习方法在跨模态语义对齐中显现出巨大优势，它能够有效关联不同模态数据语义，通过构建对比学习模型将同一语义概念的不同模态数据作为正样本对，将不同语义概念的不同模态数据作为负样本对，将不同模态数据映射到一个能够反映语义关系的共同空间。生成对抗网络也被用于跨模态语义对齐，它主要由生成器和判别器组成，前者负责将一种模态数据转化为另一种模态数据，后者用于判断生成数据是否真实^[9]。知识图谱辅助的寓意对齐方法能够借助知识图谱中丰富语义知识和关系实现不同模态数据与知识图谱节点和知识的关联和匹配，进而实现语义对齐。

语义融合整合了不同模态语义信息，特征拼接是简单直观的语义融合策略，它将不同模态数据提取特征在维度上拼接形成统一特征向量^[10]。例如图像与文本跨模态任务中采用卷积神经网络提取图像视觉特征，使用 Transformer 提取文本语义特征，拼接后能够得到融合特征。注意力融合机制能够让模型根据任务需求自动关注不同模态数据的重要部分实现更有效语义融合，基于图神经网络的语义融合算法将不同模态数据表示为图结构，融合策略还可以根据任务特点和需求动态调整，例如在一些需要实时响应的任务中采用计算效率更高的融合策略，结合强化学习方法可以让模型根据任务反馈信息自动选择最有语义融合策略，进一步提升模型性能与效率。

三、结束语

本文围绕大模型跨模态信息融合展开研究，提出多模数据获取与预处理、模型架构设计、训练推理机制以及语义贯通等创新技术路径。这些成果为大模型实现多模联动、语义贯通提供了方向，未来我们将积极探索技术创新路径，应对更多复杂场景需求，让人工智能更好地服务于社会。

参考文献

- [1] 刘聪, 朱兰芹. 基于多模态交互的汽车人机交互设计研究 [J]. 汽车电器, 2022, (08): 3-5. DOI: 10.13273/j.cnki.qcdq.2022.08.003.
- [2] 潘昱辰. 基于提示优化和工具学习的大语言模型推理能力增强方法研究 [D]. 军事科学院, 2024.
- [3] 缪凯博. 基于多模态融合的电商数据分析系统设计与实现 [D]. 华东师范大学, 2024.
- [4] 贺菁希, 任昊炯, 夏以柠. 有图不一定有真相: AI 大模型应用的多模态虚假信息问题分析 [J]. 现代视听, 2024, (08): 36-41.
- [5] 姚奕, 陈朝阳, 杜晓明, 等. 多模态知识图谱构建技术及其在军事领域的应用综述 [J]. 计算机工程与应用, 2024, 60(22): 18-37.
- [6] 余德志. 基于度量学习与粗细粒度协同的多模态文档信息抽取算法研究 [D]. 杭州电子科技大学, 2024.
- [7] 刘学博, 户保田, 陈科海, 等. 大模型关键技术与未来发展方向——从 ChatGPT 谈起 [J]. 中国科学基金, 2023, 37(05): 758-766.
- [8] 李伟. 电厂热工过程控制中智能 PID 控制器的应用探讨 [J]. 应用能源技术, 2022, (06): 9-11.
- [9] 杨蓓莹, 王金桥. 多模态大模型的发展与思考 [J]. 财经智库, 2024, 9(02): 33-50+135-136.
- [10] 吴蔚. 人工智能多模态通用大模型数据合规技术应用风险动态规制 (英文) [J]. 科技与法律 (中英文), 2024, (02): 117-126.