Research on multimodal information fusion technology in large model intelligent question answering

Senda Zhang Xin Cui Yangang Qu Bo Liu Jian Sun

Beijing CLP Puhua Information Technology Co., Ltd., Beijing, 102200, China

Abstract

This paper describes the principle of large model training based on large-scale data, analyzes the characteristics of self-attention and multi-attention mechanism in Transformer architecture, and introduces the fusion method of multi-modal information in data layer and feature layer, as well as multi-modal data preprocessing technology such as text and image. Finally, it discusses its application in many fields, such as intelligent question answering based on text and image, auxiliary power equipment diagnosis and recognition; Intelligent question and answer integrating text and voice, taking intelligent voice assistant as an example to show the function realization process; Multi-modal fusion is used to answer complex questions such as maintenance solution consultation, and provides users with accurate solutions by integrating multi-modal information.

Keywords

AI; Large model; Language model; Intelligent question answering; Multimodal information

大模型智能问答中的多模态信息融合技术研究

张森达 崔昕 曲延刚 刘博 孙健

北京中电普华信息技术有限公司, 中国·北京 102200

摘要

阐述大模型基于大规模数据训练的原理, 剖析Transformer架构自注意力和多头注意力机制的特点, 同时介绍多模态信息融合在数据层、特征层的融合方法, 以及文本、图像等多模态数据预处理技术。最后探讨了其在多个领域的应用, 基于文本和图像的智能问答, 辅助电力设备诊断识别; 文本与语音融合的智能问答, 以智能语音助手为例展示功能实现过程; 多模态融合用于检修方案咨询等复杂问题回答, 通过整合多模态信息为用户提供精准方案。

关键词

AI; 大模型; 语言模型; 智能问答; 多模态信息

1 引言

大模型智能问答系统正逐渐成为人们获取信息、解决问题的重要工具。从日常的生活咨询, 到专业领域的知识问答, 其应用场景不断拓展。然而, 随着用户需求日益复杂多样, 单一模态的智能问答系统在信息获取的全面性、准确性和交互体验上暴露出诸多不足。多模态信息融合技术为解决这些问题带来了新的契机。通过整合文本、图像、语音等多种模态的信息, 系统能够更全面地理解用户需求, 提供更丰富、更精准的回答, 而深入研究大模型智能问答中的多模态信息融合技术, 对提升智能问答系统性能、拓展应用边界具有重要的现实意义。

【作者简介】张森达(1993-), 男, 满族, 中国黑龙江人, 本科, 工程师, 从事大数据、数据中台、人工智能(数据应用方向)研究。

2 大模型相关技术概述

2.1 大模型技术

2.1.1 基本原理

大模型基于大规模数据训练, 通过对海量文本、图像、语音等多类型数据的学习, 挖掘数据中的潜在模式与规律。数据规模庞大且多样, 为模型提供了丰富的知识来源, 涵盖多种领域与场景, 在自然语言处理的大模型训练中, 会使用数十亿甚至数万亿字的文本数据, 这些数据包含新闻资讯、学术论文、小说故事等各类文本体裁^[1]。在训练过程中, 模型不断调整自身参数, 通过优化算法(如随机梯度下降及其变体)来最小化预测结果与真实标签之间的损失函数, 实现参数学习。随着参数不断更新, 模型逐渐掌握数据中的语义、语法以及语义关联等信息, 从而具备对各种输入进行理解和处理的能力。

2.1.2 Transformer 架构特点

自注意力机制是 Transformer 架构的核心创新之一。传统的循环神经网络(RNN)或卷积神经网络(CNN)在处

理长序列数据时存在一定局限性,如RNN在处理长距离依赖关系时容易出现梯度消失或梯度爆炸问题^[2]。自注意力机制允许模型在处理序列中的每个位置时,同时关注序列中的其他所有位置,直接计算输入序列中各个元素之间的关联程度。通过计算输入序列中每个元素与其他元素的注意力分数,模型能够动态地分配注意力权重,从而更有效地捕捉长距离依赖关系。

为进一步增强模型对不同特征和关系的捕捉能力,Transformer引入了多头注意力机制。它通过多个不同的线性投影将输入映射到多个低维子空间,在每个子空间中独立计算注意力,然后将这些子空间的结果拼接起来,如此模型便能够从不同角度捕捉输入数据的特征,学习到更丰富的语义信息。

2.2 多模态信息融合技术原理

2.2.1 数据层融合方法

直接拼接是数据层融合中最为直观的方法,它来自不同模态的原始数据直接按顺序或特定规则连接起来。例如,在处理图像和文本数据时,可以将图像的像素矩阵与文本的词向量序列直接拼接成一个长向量。这种方法实现简单,无需复杂的算法转换,能最大程度保留原始数据信息。其优点在于简单直接,对硬件和算法要求较低,在一些对计算资源有限且数据结构相对简单的场景中较为适用。

2.2.2 特征层融合方法

联合特征提取旨在从不同模态数据中同时提取具有代表性的特征,并将这些特征进行融合。以文本和语音模态为例,对于文本可采用词嵌入、句法分析等手段提取语义和结构特征,对于语音可利用梅尔频率倒谱系数(MFCC)等方法提取声学特征,然后通过特定的融合策略(如加权求和、串联等)将这些特征组合起来^[3]。这种方法的优点是能够充分利用不同模态数据的独特信息,通过融合互补特征提升模型性能。由于在特征层面进行融合,可减少原始数据的噪声干扰,提高数据的质量和稳定性。而且相较于数据层融合,特征层融合能在一定程度上降低维度,减少计算量。

2.3 多模态数据处理技术

2.3.1 文本预处理

文本预处理依靠的是分词、词性标注、词向量生成等技术。分词是将连续的文本序列分割成一个个独立的词语或词块的过程。在英文文本中,由于单词之间天然通过空格分隔,分词相对简单,但在中文等语言中,词语之间没有明显的空格界限,需要借助特定的分词算法。词性标注是为每个分词后的词语标注其对应的词性,如名词、动词、形容词、副词等。这一过程有助于进一步理解文本的语法结构和语义信息^[4]。

2.3.2 图像预处理

尺度不变特征变换(SIFT)是一种广泛应用的图像特征提取算法。SIFT算法的核心在于提取图像中的关键点及

其描述子。它首先通过高斯差分金字塔(DoG)来检测图像中的尺度空间极值点,这些极值点在不同尺度和旋转下具有较好的稳定性。然后,对于每个关键点,计算其周围邻域内的梯度方向和幅值,生成一个128维的描述子向量。SIFT特征具有尺度不变性、旋转不变性以及对抗光照变化、视角变化等具有一定的鲁棒性。在多模态智能问答中,当涉及基于图像内容的问题时,如识别图像中的物体、场景等,SIFT特征提取能够帮助模型从图像中提取出具有代表性和区分性的特征,为后续的图像分类、目标检测等任务提供基础。

3 多模态信息融合在大模型智能问答中的应用

3.1 基于文本和图像的智能问答

3.1.1 电力设备故障图像诊断

在电力设备故障诊断领域,多模态信息融合的智能问答系统发挥着重要作用。以高压输电线路故障诊断为例,系统借助各类传感器与监测设备,获取高压输电线路多模态数据。利用高清摄像头采集线路杆塔、绝缘子、导线等部件的图像,利用红外热像仪获取线路发热情况的热红外图像,同时通过电力监测设备收集线路的电流、电压等电气参数数据。大模型利用自身强大的特征提取能力处理这些数据。对于图像数据,大模型凭借在海量电力设备图像上的预训练,能够自动识别杆塔的倾斜、绝缘子的破损、导线的断股等外观异常,以及通过热红外图像检测线路连接点的过热区域;对于电气参数数据,大模型能够分析电流、电压波形的细微变化,判断线路是否存在过载、短路等潜在故障隐患。

与此同时,系统收集与输电线路相关的文本信息,如线路的设计文档、历史维护记录、故障报告等。大模型对这些文本进行深入理解与分析,提取出线路的建设时间、设计标准、过往故障类型及处理措施、维护周期等关键信息。

3.1.2 设备型号识别领域

当面对一张电力设备的现场图像时,系统首先运用图像识别技术提取设备的外观特征。以输电杆塔为例,从图像中提取杆塔的结构特点,如杆塔的高度、横担的数量和形状、绝缘子的类型和分布等。利用特征提取算法,将这些外观特征转化为计算机能够理解的数值向量表示。系统还会收集与该电力设备相关的文本信息,包括设备的采购合同文档、设备安装调试记录、运行维护手册等资料中的文字内容,这些文本信息涵盖了设备的型号、生产厂家、适用场景、技术参数等详细数据,当用户提出关于电力设备的问题,系统通过融合图像特征和文本信息来进行回答,帮助识别设备型号。

3.2 文本与语音融合的智能问答

3.2.1 语音识别与文本处理技术融合流程

①语音采集与预处理。在用户使用智能语音问答系统时,首先通过麦克风等设备采集语音信号。由于实际环境中存在各种噪声干扰,采集到的语音信号需要进行预处理。这一步骤通常包括降噪处理,运用滤波算法去除环境噪声、设

备底噪等,提升语音信号的清晰度^[5]。

②语音识别。预处理后的语音信号被送入基于大模型的语音识别模块。大模型凭借其在海量语音数据上的预训练,具备强大的语音特征学习能力。它不仅能够像传统深度学习模型那样学习语音信号的声学特征与文字之间的映射关系,还能通过对大量语音样本的学习,更好地适应不同口音、语速和语言习惯。大模型的多模态融合能力使其能结合语音的上下文信息,例如在一段对话中,根据前文的语义和语境,更准确地识别当前语音内容。

③文本理解与分析。经大模型语音识别得到的文本进入文本处理模块,进行文本预处理,利用预训练学习到的语言知识,进行分词、词性标注、命名实体识别等操作。分词时,大模型能准确切分连续文本,处理复杂句式和生僻词汇;词性标注和命名实体识别也因大模型对大量文本的学习,能更精准地确定每个词的词性和文本中的特定实体,如人名、地名、时间等。在句法分析和语义理解阶段,大模型通过对大规模文本数据的深度理解,能够分析句子结构,把握文本的深层语义和意图。

④答案生成与语音合成。根据大模型对文本的理解结果,系统借助大模型强大的知识储备和推理能力,在内部知识库或通过调用相关服务获取对应的答案。大模型会对答案进行语言组织优化,使其表达更清晰、逻辑更连贯。例如,在回答关于电力设备故障处理的问题时,大模型能根据故障类型、设备情况等因素,生成详细且有条件的解决方案。

3.2.2 智能语音助手应用案例

以具备智能体特性的手机智能语音助手为例,它展现出强大且智能的交互能力。在信息查询场景下,当用户下达“查询今天北京的天气”指令时,语音助手如同一个智能个体,迅速做出反应。它首先利用先进的语音识别技术,精准地将语音转化为文本,这一过程借助智能体对多种口音和语言习惯的学习,即使在嘈杂环境中也能准确识别。接着,智能体凭借其深度语义理解能力,分析文本意图,理解用户对北京当日天气信息的需求。

3.3 多模态融合用于复杂问题回答

3.3.1 电网运维方案制定案例

在电网运维场景中,运维人员常常会面临复杂且综合性的问题,例如“下个月要对某区域电网进行全面运维检修,预算在50万元左右,要确保重点设备稳定运行,同时最大程度减少对居民用电的影响,有哪些高效的运维方案?”为了制定出准确有效的方案,智能系统需要整合多种模态的信息。

3.3.2 多模态信息整合过程

①文本信息收集。系统首先从运维人员输入的文本中提取关键信息,如运维时间“下个月”、预算“50万元左右”、重点任务“确保重点设备稳定运行”以及特殊要求“最

大程度减少对居民用电的影响”。

②图像信息辅助。利用无人机搭载的高清摄像头和热成像仪等设备,获取电网设备的图像数据,如输电线路的杆塔、线缆外观图像,变电站设备的热成像图等。通过图像识别技术,分析图像中的设备状态特征,例如杆塔是否倾斜、线缆是否有破损、设备是否存在过热异常等情况,进一步辅助对电网设备状况的评估。

③语音信息补充。在日常巡检和工作沟通中,运维人员可能会通过语音记录设备对发现的问题进行描述,或者上级部门以语音形式传达运维工作重点和要求。系统对这些语音信息进行语音识别和内容分析,获取其中的关键信息和指示。

④整合决策。系统将收集到的文本、图像和语音等多模态信息进行融合分析。运用多模态融合算法,根据不同信息源的可靠性和相关性为其分配权重。例如,对于预算相关的信息,文本中的成本数据可能权重较高;而对于设备实际运行状态的判断,图像信息的权重更大。通过综合分析,系统帮助制定出符合要求的电网运维方案。

4 结语

围绕大模型智能问答中的多模态信息融合技术展开了全面且深入的探讨。从技术原理层面剖析了大模型技术、多模态融合方法以及数据处理技术,为后续研究奠定了坚实的理论基础。在应用研究中,通过设备图像诊断识别、智能语音助手、维护咨询等多个场景案例,充分展示了多模态信息融合技术在提升智能问答准确性、丰富性和交互性方面的显著优势。多模态信息融合技术在大模型智能问答领域仍面临诸多挑战,如模态间的语义对齐难题、融合算法的优化空间等。未来研究可聚焦于开发更高效的融合算法,探索跨模态知识图谱构建,以进一步提升多模态信息融合的深度和广度。

参考文献

- [1] 赵晓伟,王师晓,李情,等.苏格拉底式问题支架:促进学生向AI大模型提出高质量问题[J].现代远程教育研究,2025,37(01):102-112.
- [2] 牟智佳,岳婷,朱陶.人机协同视域下基于认知智能大模型的个性化学习设计研究[J].电化教育研究,2025,46(02):80-87.
- [3] 袁毓林.从三种复杂句看ChatGPT是不是随机鹦鹉?——语言大模型能不能理解语言意义的测试与讨论[J].语言教学与研究,2025,(01):35-49.
- [4] 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[J].科技与法律(中英文),2024,(02):117-126.
- [5] 赵朝阳,朱贵波,王金桥.ChatGPT给语言大模型带来的启示和多模态大模型新的发展思路[J].数据分析与知识发现,2023,7(03):26-35.