

多模态非结构化数据的智能预处理框架设计

李照璇

吉利学院, 中国·四川 成都 641423

摘要: 随着人工智能技术的不断演进与应用拓展, 非结构化数据成为推动模型性能提升的重要资源。尤其在医学、安防、社交媒体等领域, 图像、文本、语音、视频等多模态数据共同构成了复杂而庞大的信息体系。然而, 由于多模态数据来源异构、格式不统一、语义关联弱等问题, 其预处理过程成为当前系统建设的关键瓶颈。论文在深入剖析现有方法局限的基础上, 设计出一套面向多模态非结构化数据的智能预处理框架, 涵盖模态识别、数据清洗、标准转换、模态对齐、特征提取等核心功能。通过实际应用验证, 该框架在处理效率、对齐准确性及下游建模适应性方面均取得了显著提升, 展示出良好的通用性与扩展性。

关键词: 多模态数据; 非结构化数据; 数据预处理; 模态对齐

Design of Intelligent Preprocessing Framework for Multimodal Unstructured Data

Zhaoxuan Li

Geely College, Chengdu, Sichuan, 641423, China

Abstract: As artificial intelligence technology continues to evolve and its applications expand, unstructured data has become a crucial resource for enhancing model performance. In fields such as medicine, security, and social media, multimodal data, including images, text, voice, and video, form a complex and extensive information system. However, the heterogeneity of sources, lack of uniformity in formats, and weak semantic connections in multimodal data pose significant challenges during preprocessing, which is a critical bottleneck in current system development. This paper, after thoroughly analyzing the limitations of existing methods, proposes an intelligent preprocessing framework for multimodal unstructured data. The framework includes core functions such as modal recognition, data cleaning, standard conversion, modal alignment, and feature extraction. Practical application tests have shown that this framework has significantly improved processing efficiency, alignment accuracy, and downstream modeling adaptability, demonstrating excellent versatility and scalability.

Keywords: multimodal data; unstructured data; data preprocessing; modal alignment

0 前言

当今信息社会中, 数据的结构化程度正快速降低。传统以表格为主的结构化数据在整体数据总量中所占比例不断下降, 取而代之的是以自然语言、图像、语音、视频等形式存在的大量非结构化数据。据 IDC 发布的报告显示, 到 2025 年, 全球非结构化数据将占总数据量的 80% 以上。而多模态数据, 作为非结构化数据的高级形式, 在医疗诊断、无人驾驶、智能客服、公共安全等领域发挥着愈发关键的作用。

1 多模态非结构化数据的特征与预处理需求分析

1.1 多模态非结构化数据的异质性特征

多模态非结构化数据的首要特征是其高度的“异质性”, 既体现在数据形式的多样性上, 也体现在采集来源、时间分布与语义表达方式的差异上。与传统结构化数据以二维表格形式存在、字段清晰、主键明确不同, 多模态数据包括文本、图像、音频、视频等类型, 不同模态在内容表达、空间分布、

时间粒度以及语义浓度上具有本质区别。例如, 文本强调时序逻辑与语义层级, 图像突出空间结构与纹理分布, 音频则承载频谱与情感语调, 视频还叠加了时间动态性与空间连续性。这种模态间的异构性不仅体现在“表层格式”不同, 更深层地反映在信息维度的不等价性与语义解释的复杂性上。因此, 面对多模态数据, 传统“一刀切”式的数据处理策略难以适用。每一类模态都需定制化的预处理流程, 例如图像需处理分辨率与噪声, 文本需分词与实体识别, 音频需去除静音段与降噪。这就要求预处理框架具备“异构识别—模态解耦—语义协同”能力, 能够适应不同数据流的技术特性, 建立起多模态数据间协同处理的基础, 从而为后续融合建模打下坚实的语义基础。

1.2 多模态数据融合中的核心技术瓶颈

尽管当前多模态人工智能系统在语音识别、图像标注、跨模态检索等应用中表现出色, 但其基础环节—数据融合—依然面临一系列技术瓶颈, 而这背后往往是预处理层面的问题未能妥善解决。首先, 模态间的时空对齐困难是核心障碍之一。例如, 在视频问答任务中, 如何将一段文字问题精确

对齐到视频片段的关键帧，要求预处理层必须完成文本语义抽取、视频帧帧、特征匹配等一系列复杂操作，且须具备时间窗口控制与语义相似度度量能力。其次，模态间语义分布的不一致性也是融合难点。文本通常具备抽象表达，而图像和音频则更具感知性质，如何通过统一嵌入空间将其拉齐，严重依赖预处理阶段对高质量特征的提取与归一。最后，噪声传播问题也不容忽视。若某一模态数据质量较差，如视频卡顿、文本错别字、语音模糊，极易通过融合路径向其他模态传播误差，从而导致系统性能整体下降。因此，建立在高质量预处理之上的模态清洗、去噪、置信度评估机制，是打通多模态数据融合“任督二脉”的关键。归根结底，预处理阶段不仅是数据输入的清洁工，更是系统语义理解的筑基者，直接决定了多模态 AI 系统的下限与上限。

1.3 智能预处理的系统化设计需求

面对上述复杂特性与技术瓶颈，多模态非结构化数据的预处理必须转向系统化、智能化方向发展。从系统架构角度看，预处理不应是分散、割裂的流水线，而应构建统一入口、动态调度、模块解耦、可组合的智能化平台。具体而言，系统设计需满足五项核心能力：第一，具备自动模态识别能力，能够快速判断输入数据的类型及所需处理路径，减少人为干预；第二，具备质量检测与缺陷修复功能，自动识别数据噪声、缺失值、模糊区域等问题，并触发相应处理策略；第三，需构建标准化中间表示机制，使不同模态在处理后备具统一的结构输出，便于建模层调用；第四，支持多任务并发处理与异步调度，提升系统吞吐能力与实时性能；第五，框架须具备自学习能力，能根据反馈结果自动优化清洗参数、调整预处理顺序。在此基础上，构建一套可配置、可扩展、可嵌入的预处理微服务架构，将为医疗诊断、舆情分析、智慧交通等行业级多模态应用提供底层保障。

2 智能预处理框架的结构设计与模块实现

2.1 架构总览：多层次模块化设计理念

为了应对多模态非结构化数据预处理的复杂需求，论文提出的智能预处理框架采用多层次模块化设计理念，整体结构划分为四个核心层级：数据接入层、模态识别层、任务调度层与结构输出层。其中，数据接入层主要负责不同模态原始数据的自动识别与标准输入格式的封装；模态识别层负责智能判断数据的类型与质量等级，触发相应处理管道；任务调度层负责根据预定义策略与当前系统负载，分配计算资源并调用异构处理模块；结构输出层统一数据结构，生成标准化多模态特征表达，供下游建模模块调用。该架构不仅支持图像、文本、音频、视频等常见模态，还预留接口以接入 LiDAR 点云、遥感影像、ECG 信号等扩展模态。相较于传统线性处理流程，本框架强调“动态调度—模块解耦—并行执行”的系统架构优势，使其具备出色的横向扩展能力与纵向精细化调控能力。例如，在智慧医疗场景中，接入 CT 图

像+病历文本+心电图数据三类模态后，系统可自动识别数据类型、进行异构特征清洗、执行时间语义对齐，并以统一张量输出供诊断模型使用，极大提升数据准备效率。

2.2 智能模态识别与质量评估机制

智能预处理的第一关键是模态识别机制。传统方法常依赖文件名后缀或人工标记，不仅效率低下，且在复杂系统中容易出错。论文设计了一种结合 MIME 头部识别、数据特征统计与轻量卷积神经网络（CNN）分类器的三维识别机制，可实现图像、音频、视频、文本、混合流五大类的自动精准识别。识别准确率经在多模态数据集 MM-IMDb 和 AVE（Audio-Visual Event）上测试，分别达到了 97.1% 和 95.8%。识别完成后，框架将每条数据送入质量评估子模块，执行包括模糊度分析（图像）、乱码检测（文本）、静音段比率分析（音频）等技术操作，依据加权评分给出处理建议（保留/重处理/剔除）。如表 1 所示，在对 5000 条多模态数据进行自动评分后，系统精准识别出缺陷数据 1126 条，为后续处理节省大量人工成本。

表 1 不同模态类型下的识别准确率

模态类型	样本数量	检出缺陷率	识别准确率	处理建议
图像	1500	23.1%	97.6%	重采样 / 增强
文本	1000	18.4%	95.3%	分词 / 实体补全
音频	1000	29.2%	94.9%	降噪 / 裁剪
视频	1500	22.7%	96.1%	抽帧 / 去重

2.3 多模态数据清洗与格式转换模块

预处理框架的第二核心模块是数据清洗与格式转换机制。考虑到不同模态间处理方式的巨大差异，系统采用“模态解耦+任务模板”的设计策略。对于图像类数据，系统将执行三步操作：格式统一（如统一转为 PNG）、分辨率归一化（按模型需求调整为 256×256）、颜色通道标准化（转为 3 通道 RGB）；文本类数据则进行编码校验、分词、命名实体识别与标签修复；音频类数据会进行重采样（统一为 16kHz）、去静音、音量标准化等处理；视频则采用关键帧提取与抽帧算法，保留动作片段与场景过渡帧，并以时间戳打标。完成清洗后，所有模态数据统一封装为 HDF5 或 TFRecord 结构，保留原始模态标识、处理日志与数据完整性哈希值，确保后续可追溯性。系统在 OpenSubtitles 多模态字幕数据集上测试该模块，文本清洗精度达 94.5%，图像统一标准化耗时降低 41%，充分验证了模块性能。此外，模块支持异步处理和批处理机制，在多 GPU 平台上可实现 4 模态并行清洗，系统吞吐量达每分钟 1200 条数据，满足大规模应用需求。

2.4 多模态对齐与结构输出机制设计

实现多模态智能预处理的最终目标，是将原始异构数据转化为建模可用的统一特征表示，因此模态对齐与结构输出机制成为关键环节。论文框架采用双路径对齐策略：一方面通过“时间戳规整+滑动窗口”策略完成低层级数据对齐；

另一方面采用基于 Transformer 的跨模态注意力机制,实现高语义对齐。以视频—文本任务为例,系统可对字幕与视频关键帧进行双向嵌套配对,通过引入视觉—语言对齐损失函数(V-L Alignment Loss)动态校正语义偏移。此外,结构输出机制将所有模态映射到统一张量空间,最终输出为“多模态张量+元数据字典”的混合格式,结构如下:

图像特征: [B,3,256,256]

文本嵌入: [B,L,768]

音频特征: [B,T,128]

时间戳对齐矩阵: [B,M,N]

其中: B 为批大小, L 为文本长度, T 为音频帧长, M/N 为时间维度映射尺寸。实验显示,在社交媒体多模态数据任务中,结构输出准确率提升至 92.3%,对齐误差均值控制在 $\pm 0.5s$ 以内,有效提升下游模型的预测精度与稳定性。该模块不仅解决了模态差异融合难题,更为构建高可用、多模态 AI 平台提供了高适配性数据接口。

3 典型应用案例分析与发展前景展望

3.1 医疗数据融合中的智能预处理实践

在智慧医疗场景中,临床数据本身呈现出高度的模态多样性,包括影像(如 MRI、CT)、文本(电子病历、医生记录)、音频(问诊录音)、时序信号(如心电图、脑电图)等。在传统处理流程中,这些数据分别被孤立管理,缺乏统一的处理与建模接口,导致医疗 AI 系统训练样本不完整、预测性能不稳定。论文提出的智能预处理框架已在某三级甲等医院影像与病历数据处理平台上线测试。在实验中,我们选取包含 5000 例患者样本的多模态数据集(每例含 CT 图像、诊断文本和语音记录),通过该框架进行统一预处理。结果显示,图像清洗耗时从平均 7.8 秒降低至 4.1 秒,文本结构化准确率由 74% 提升至 91%,语音文本对齐误差下降至 ± 0.4 秒。进一步将处理后数据输入至肺部病灶智能识别模型中,相较传统方法,诊断准确率提升了 8.7%。这不仅体现了该预处理框架在提升模型精度方面的贡献,更关键的是,它显著提高了医疗数据处理的自动化程度,降低了人工参与率,为 AI 辅助诊断系统提供了高质量、高一致性的数据基础,从而有效推动智慧医疗向更精准、更高效、更低成本方向发展。

3.2 城市治理与智能安防中的多模态融合应用

在智能城市与公共安全领域,多模态数据的高效集成处理成为城市治理体系的重要支撑。监控视频、报警文字、通话录音、社交媒体信息、地理定位等数据源共同构成城市运行的数字“感知网络”。传统城管系统对这些数据处理割裂,难以实现实时响应与全景感知。论文所设计的智能预处理框架已在某市智慧安防项目中部署试点,主要任务为分析城市核心区域(覆盖面积约 30km²)内的视频监控与社交媒体舆情,识别潜在治安隐患与突发事件。在系统部署后的两个月内,框架每日处理视频数据约 1.6TB、文本评论约 18

万条、音频通话数据 1.2 万段,平均每日生成结构化警情线索 42 例。其中,智能预处理对视频中关键帧的提取效率提升了 55%,舆情数据的文本清洗准确率达 93%,语音转写出错率控制在 6% 以内,有效支撑后续多模态事件识别与空间风险预测模块。在一次重大交通事件中,系统在事发后 3 分钟内即完成“视频场景变化识别+市民社交评论聚集检测+语音报警内容解析”三模态数据融合,为指挥系统提供精准坐标与事件等级判断。

3.3 智能预处理技术的未来演化路径

尽管当前智能预处理框架已在多个实际场景中展现出良好表现,但从长远来看,仍存在诸多可扩展与升级的空间。首先,预处理机制需进一步引入自适应优化与元学习机制,实现对不同任务场景下处理策略的动态选择与调参能力,真正实现“数据驱动”的处理策略演进。其次,未来框架将更多整合联邦学习、差分隐私等安全计算技术,确保多源敏感数据(如医疗、司法、金融)的合法使用与信息保护,解决智能预处理在数据安全与合规方面的制度瓶颈。此外,随着边缘计算与物联网技术的发展,智能预处理不再局限于云端部署,而需拓展至“端—边—云”协同模式。在边缘设备侧,轻量化预处理模块须具备实时特征提取、事件触发、初级对齐等功能,实现数据“就地初步结构化”;在云端,则承担统一建模、语义融合与深度分析。最后,未来预处理系统将进一步融合大模型能力(如多模态预训练模型),实现跨模态信息的主动推理与知识补全。例如,系统可在缺失图像模态时,基于文本内容补全场景推理表示,提升模型鲁棒性与适应性。

4 结语

论文围绕“多模态非结构化数据”这一复杂而关键的对象,提出并详细设计了一套完整的智能预处理框架。通过模态识别、清洗、转换、对齐与特征统一五大模块,系统性解决了多模态数据预处理中的核心难题。在医疗、安防、社交等典型场景中的实践应用,进一步验证了该框架在提升数据处理效率与模型精度方面的显著优势。未来,随着数据规模与复杂度的持续提升,该框架仍有广阔的演化空间,有望成为支撑多模态 AI 系统落地的关键底层能力。

参考文献:

- [1] 谭宇辰,于重重,钱兆鹏,等.濒危土家语多模态语料库的构建[J].计算机仿真,2023,40(10):326-330.
- [2] 司斌斌,龚永昌,高龙.多模态数据在金融风控领域的应用[J].金融科技时代,2023(3):24-29.
- [3] 宗滕,吴松涛,周春华.基于多模态数据分析的典型智能化军事应用[J].信息安全与通信保密,2022(2):9-16.
- [4] 骆建豪.融合 EHR 非结构化数据的再入院预测方法研究[D].天津:天津大学,2021.
- [5] 吴蔚.基于大数据的医院电子档案资料管理方法[J].中国卫生产业,2018,15(36):57-59.