

基于大模型数据增强的课堂行为状态分析

曾建电¹, 邹文成², 王田¹

1. 北京师范大学自然科学高等研究院, 广东珠海;

2. 南京理工大学自动化学院, 江苏南京

摘要: 课堂行为状态分析是教育智能化研究的重要方向, 旨在通过自动化识别学生的学习行为, 揭示其学习状态和参与度, 为个性化教学和教育质量提升提供科学依据。然而, 传统单模态方法在处理低分辨率图像或复杂背景时存在信息不足和噪声干扰的问题, 难以全面捕捉学生行为特征。为此, 本研究提出了一种基于多模态大模型的数据增强方法, 结合图像和文本模态的联合建模, 对学生课堂行为状态进行深入分析。实验结果表明, 基于多模态大模型的数据增强方法在分类准确率和F1值指标上显著优于传统单模态方法, 尤其在处理复杂背景和低分辨率图像时表现突出。本研究通过结合多模态大模型和特征融合技术, 为课堂行为状态分析提供了一种高效、精准的解决方案, 为智能教育研究提供了重要参考。

关键词: 课堂行为; 大模型; 智能教育

Large Language Model Enhanced Data Augmentation for Classroom Behavior Analysis

Jiandian Zeng¹, Wencheng Zou², Tian Wang¹

1. Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong;

2. Nanjing University of Science and Technology, School of Automation, Nanjing, Jiangsu

Abstract: Classroom behavior analysis represents a critical frontier in intelligent education research, aiming to automatically identify student learning behaviors to reveal engagement patterns and cognitive states, thereby providing empirical foundations for personalized instruction and educational quality enhancement. However, traditional unimodal approaches exhibit inherent limitations in handling low-resolution images and complex classroom environments, suffering from insufficient feature representation and noise interference that impede comprehensive behavioral characterization. To address these challenges, this study proposes a novel data augmentation framework leveraging multimodal large language models. Our approach integrates cross-modal alignment between visual inputs and textual context through joint embedding spaces, enabling synergistic analysis of student postures, facial expressions, and environmental interactions. Experimental results demonstrate that the proposed multimodal framework achieves superior performance over traditional unimodal baselines in terms of accuracy and F1 metrics, especially when dealing with complex backgrounds and low-resolution images. This study integrates multimodal large language models with feature fusion techniques, delivering an efficient and precise analytical framework for classroom behavior analysis while providing valuable insights for intelligent education research.

Keywords: Classroom Behavior; Large Language Models; Intelligent Education

1 引言

近年来,随着教育信息化的不断推进,我国教育数字化阶段逐渐进入到以智慧学习为标志的智慧教育阶段[1]。智慧课堂作为智慧教育的重要组成部分,与师生之间、生生之间的联系紧密。同时,计算机视觉、大语言模型(Large Language Models, LLMs)等技术的突破,极大地推动了人工智能在教育领域的深入应用[2, 3]。

课堂行为状态分析作为教育场景中的重要研究方向,旨在通过对学生行为的自动化识别和分析,揭示学生的学习状态、参与度以及课堂互动情况。然而,传统的课堂行为分析方法通常依赖于人工观察或单一模态的数据处理方式,这不仅效率低下,还容易受到主观因素的干扰,难以全面、准确地反映学生的真实学习状态[4]。人工智能技术在教育领域的普及为智慧教学研究与实践提供了理论基础和技术支撑,两者的融合已成为必然趋势。在课堂学习上,学生的上课状态无疑对学习效率具有决定性的影响。同时,学生的行为状态以及教学互动等信息会产生大量的多模态数据:学生的表情、动作、语言以及与教师和其他同学的互动情况等,都蕴含着丰富的教学信息。通过实时分析学生的课堂状态,教师能够更全面地把握学生的行为、情感和认知过程,进而制定出更加贴合学生需求、个性化的教学计划和教学决策[5,6]。因此,如何利用人工智能技术对课堂行为进行高效、精准的分析,成为当前教育智能化研究中的重点[7,8]。

本研究以课堂行为状态分析为核心,提出了一种结合多模态大模型的数据增强方法,旨在通过视觉和文本模态的联合建模,全面分析学生的学习行为特征及其变化规律。一方面,借助多模态大模型的理解和交互能力对学生课堂行为进行解读,为教师提供更为精准的反馈;另一方面,探究学生学习行为规律,挖掘深层次的教育价值,以更好地设计课程和教学活动,推动教育的精准化实施。

2 研究设计

本研究通过多模态大模型捕捉学生外显行为特征,结合文本描述解析行为语义内涵,构建了“行

为表现-认知过程”的映射分析模型,有效克服了传统课堂观察的主观性局限,为教师理解课堂教学提供了客观依据。同时创新性地将多模态人工智能技术与教育认知分析深度融合:该“技术赋能-数据驱动-教育解释”的创新模式,一定程度上缓解了传统课堂中“仅观察”的实践困境。

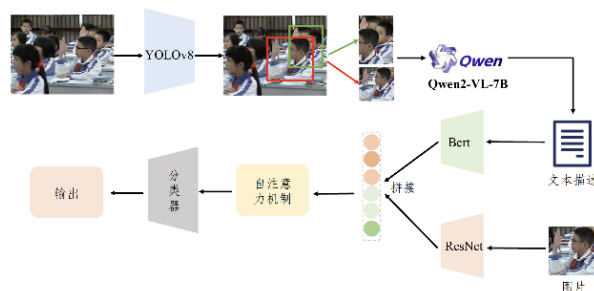


图1. 整体框架示意图

2.1 技术框架构建

具体而言,如图1所示,针对课堂场景中的图像数据,首先利用YOLOv8模型对输入的课堂图像进行目标检测与实例分割,提取出可能的状态实例图像。YOLOv8作为一种高效的目标检测算法,能够快速准确地定位图像中的目标区域,并生成对应的实例图像,为后续的多模态处理奠定了基础。在完成实例分割后,针对每个分割出的图像实例,进一步引入多模态大模型Qwen(Qwen2-VL-7B)对其进行描述增强。Qwen模型具备强大的视觉-语言理解能力,能够结合图像内容生成高质量的文本描述。这些文本描述不仅能够丰富图像的语义信息,还为后续的多模态特征提取提供了文本模态输入。通过该方式,每个图像实例都被赋予了一个与其内容高度相关的文本描述,从而实现了视觉模态与文本模态的初步关联。后续针对生成的图像模态和文本模态,分别采用不同的特征提取方法进行处理。在图像模态方面,使用ResNet模型从图像中提取出具有语义信息的高维特征表示,确保视觉特征的表达能力。对于文本模态,采用Bert模型对文本描述进行特征提取。Bert能够捕捉文本中的深层语义信息,为文本模态提供高质量的特征表示。在完成图像和文本模态的特征提取后,将两种模态的特征进行拼接,形成一个联合特征表示。最后,将经过自

注意力机制融合后的多模态特征输入到分类器中进行分类。分类器根据融合特征的语义信息，判断输入图像实例的类别。通过上述流程，整个框架实现了从图像实例分割、文本描述生成、多模态特征提取与融合到最终分类的完整处理链路。

2.2 多模态大模型的图片描述增强

考虑到样本中部分照片像素较低以及复杂背景可能导致的目标混淆问题，本研究引入了开源多模态大模型通义千问（Qwen2-VL-7B）对每个样本进行文本描述生成，以增强样本的特征表达能力。通过设计不同的提示词来引导多模态模型生成与样本内容高度相关的描述。例如，对于课堂场景中的学生行为样本，可以设置提示词如“描述学生的行为状态”或“学生当前的动作是什么”，以确保生成的文本描述能够准确反映图像中的关键信息。这种提示词引导策略不仅提升了模型输出的针对性，还增强了文本模态与视觉模态之间的关联性。生成的文本描述不仅丰富了样本的语义信息，还显著提高了样本的可解释性，使得模型能够更好地理解图像中的细节和语境。此外，文本模态的引入为后续的多模态特征提取和融合提供了重要支持。同时，模型能够有效提取图像中难以直接捕捉的语义信息，从而在一定程度上缓解了像素不足和背景干扰对特征提取的影响，不仅提升了样本的特征表达能力，还为后续的分析 and 模型训练提供了更为丰富和多样化的信息支持。

2.3 图片文本表征融合的行为识别

在通过多模态大模型生成图片的文本描述后，本研究进一步结合图片和文本模态信息进行行为识别。对于图片模态 x_{img} 和文本模态 x_{txt} ，分别采用ResNet和Bert模型提取其特征表示：

$$\begin{aligned} E_{img} &= \text{ResNet}(x_{img}); \\ E_{txt} &= \text{Bert}(x_{txt}). \end{aligned} \quad (1)$$

随后，将两种模态的特征进行拼接，形成联合特征表示：

$$E_c = [E_{img} || E_{txt}]. \quad (2)$$

拼接后的特征 E_c 被视为一个完整的序列输入，接下来通过自注意力机制进行编码：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (3)$$

其中的 Q, K, V 分别表示查询、键和值矩阵， d 为输入的向量维度。通过注意力机制，模型能够从不同的语义角度对图片和文本模态的特征进行融合，最终得到融合表征 H_c ：

$$H_c = \text{Attention}(E_c, E_c, E_c). \quad (4)$$

最后将融合后的表征 H_c 作为输入到 softmax 分类器，生成预测标签 $\hat{y}$ ：

$$\hat{y} = \text{argmax}(\text{softmax}(W_c H_c + b_c)), \quad (5)$$

其中 W_c 和 b_c 为对应的参数矩阵和偏置项。

3 结果分析

3.1 数据集与衡量指标

数据集：本研究基于SCB（Student Class Behavior, SCB）数据集，将学生的课堂学习状态作为分析对象，旨在通过多模态方法深入探究学生学习行为的特征及其变化规律。SCB数据集作为专注于课堂场景的行为分析数据集，涵盖了学生在学习过程中常见的行为状态。本研究聚焦于课堂中具有代表性的三种学习动作：举手、阅读和写作。这三种行为不仅是课堂学习活动的核心组成部分，还能够反映学生在不同学习阶段的参与度和专注度。例如，举手行为通常与学生的主动参与和互动意愿相关；阅读行为则体现了学生对学习内容的理解和吸收过程；写作行为则是学生知识输出和实践能力的直接表现。通过对上述行为的分析，可以更全面地了解学生的学习状态及其变化规律。具体分布如表1所示：

表1. SCB学生行为数据集分布

类别	举手	阅读	写作
训练集	7998	4555	1954
测试集	2080	1327	585
总和	10078	5882	2539

衡量指标：在本研究中，为了全面评估模型在学生行为分类任务中的性能，选取了准确率（Accuracy, ACC）和F1值（F1-Score）作为主要衡

量指标，以从不同角度反映模型的分类效果，确保对模型性能的全面评估。

3.2 提示词分析

提示词的设计对生成的文本描述质量具有显著影响。针对同一个样本，使用不同的提示词会导致生成的描述在内容、细节和相关性上存在明显差异。当提示词较为简洁时，生成的描述往往包含较多的细节信息，但其中可能掺杂一些与样本内容无关的冗余信息。该冗余信息虽然在某些情况下能够提供更多的上下文语义，但也可能干扰模型对样本核心特征的理解。例如，在描述学生的课堂行为时，使用简单提示词如“描述这张图片”可能会生成较为宽泛的描述，包含背景、环境或其他无关对象的信息。虽然具备丰富的语义信息，但可能会掩盖样本的关键特征，甚至引入噪声，降低分类的准确性。反之，当提示词设计得更加细致和具体时，生成的描述会更加聚焦于样本的核心语义。例如，使用提示词“描述学生的具体行为”或“学生正在做什么”时，生成的文本描述能够更准确地捕捉学生的动作特征，如“学生正在举手回答问题”或“学生正在低头阅读课本”。因此，在实际应用中，需要根据具体任务需求对提示词进行精细化设计，以在信息丰富性和相关性之间取得平衡。

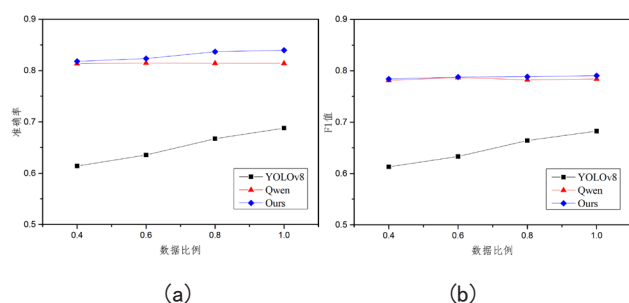


图2. 不同数据比例下的对比图：(a) 准确率；(b) F1值。

为了更全面地评估本研究提出的方法的有效性，本研究设计了对比实验，探究三种不同方法在学生课堂行为状态分类任务中的表现：1) 直接使用YOLOv8进行分类；2) 使用多模态大模型直接分类；3) 使用多模态大模型进行数据增强后分类（即本研究提出的方法）。通过对比这三种方法，可以更清晰地了解本研究方法在特征提取、数据增强

以及分类性能上的优势。

如图2所示，实验结果表明，在不同训练数据比例下，本研究提出的方法均显著优于其他两种方法。具体而言，直接使用YOLOv8进行分类的性能在训练数据较少时表现较差，分类准确率均较低，表明单一模态方法在数据不足或场景复杂时的局限性。使用多模态大模型直接分类的方法在一定程度上提升了分类性能，但由于未对数据进行增强，模型在特征表达上的潜力未能充分发挥。而本研究提出的方法通过多模态大模型生成高质量的文本描述，显著提升了数据的语义信息维度，并通过多模态特征融合进一步增强了分类性能，尤其在训练数据较少的情况下表现尤为突出，展现了较强的鲁棒性和泛化能力。

4 结语

本研究聚焦于学生课堂行为状态研究，提出了一种基于多模态大模型的数据增强方法，为课堂行为状态分析提供了一种高效、精准的解决方案，具有重要的教育改革意义。具体体现在：1) 通过实时、精准地分析学生课堂行为，教师能够及时掌握学生的学习状态，针对性地调整教学策略，实现因材施教；2) 多模态数据增强方法有效解决了传统方法的局限性，提升了课堂行为分析的准确性和鲁棒性，为智慧课堂建设提供了技术支撑；3) 探索了人工智能技术与课堂教学深度融合的新路径，为教育改革提供了理论依据和实践经验，推动了智慧教育的深入发展。未来研究可进一步探索更多模态数据的融合方法，提升模型鲁棒性和准确性。

致谢

本文由基金项目：国家自然科学基金青年项目“面向鲁棒性多模态情感分析关键技术研究”（62302049）；广东省本科高校教学质量与教学改革工程项目“基于多模态大模型的启发式学习探索”（jx2024309）；北京师范大学教学建设与改革项目（jx2024040, jx2024139）；广东省高等教育学会“十四五”规划2024年度高等教育研究课题（24GYB207）；教育部产学研合作协同育人项目

（240904497110437）资助。

参考文献

- [1] 别敦荣, 郭一蓉. 人工智能时代高等教育创新发展新趋势[J]. 中国高等教育, 2024, (Z1): 39-44.
- [2] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学:信息科学, 2023, 53(09): 1645-1687.
- [3] 王珩, 郭未来. 演绎与批判: 人工智能对教育数字化的介入[J]. 中国教育信息化, 2025, 31(02): 46-58.
- [4] 马利芳, 宁惠萍. 大学英语多模态课堂教学设计的原则和策略[J]. 中国高等教育, 2020, (20): 44-46.
- [5] 高山. 双师课堂提升人工智能创新人才培养质量[J]. 中国教育学刊, 2024, (S1): 177-179.
- [6] 李小娟, 刘清堂, 吴林静, 等. 混合同步课堂中师生多模态互动行为的动态协同分析[J]. 电化教育研究, 2022, 43(08): 43-50.
- [7] 颜天明. 基于云课堂的混合教学模式探索[J]. 中国高等教育, 2018, (17): 58-59.
- [8] 江婕, 王海燕, 于文婷. 智慧教室环境下多模态学习活动评价指标设计与应用[J]. 中国教育信息化, 2024, 30(11): 120-128.

Copyright © 2025 by author(s) and Global Science Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access