ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 1,Jan. 2023,pp. 174－185
DOI:10. 16337/j. 1004-9037. 2023. 01. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

面向教学评价的课堂视频镜头边界检测新方法

谢从华[1]，罗德凤[2]，方雨洁[1]

(1. 常熟理工学院计算机科学与工程学院,苏州 215500;2. 重庆三峡学院外国语学院,万州 404100)

摘　要：课堂视频镜头边界检测对教学评价具有重要意义。针对教学视频视觉信息变化不明显、镜头边界信息不足、检测结果不利于教学评价等问题,引入注意力机制,提出了基于视觉和文本特征描述学习的课堂视频镜头边界检测方法。首先,提出了层次视觉Transformer模型学习教学评价关注的屏幕、教师和学生等感兴趣区域的视觉特征。其次,提出了层次文本Transformer模型从屏幕和语音文本中学习教学评价关注的文本特征。最后,构建基于二值交叉熵的镜头分类和边界检测损失函数。在数据集CLShots上的实验结果表明,本文方法在准确率、召回率、F_1分数和平均交并比等指标比当前先进的教学镜头检测方法SBLV分别提高了23.3%、22.4%、22%和35.7%,比通用领域深度学习方法TransNet V2分别提高了13.8%、14.5%、14.3%和21.3%。

关键词：教学评价;课堂视频分割;注意力机制模型;镜头边界检测

中图分类号：TP391　　**文献标志码**:A

A New Shot Boundary Detection Method of Lecture Video for Teaching Evaluation

XIE Conghua[1], LUO Defeng[2], FANG Yujie[1]

(1. School of Computer Science and Engineering, Changshu Institute of Technology, Suzhou 215500, China;2. School of Foreign Languages, Chongqing Three Gorges University, Wanzhou 404100, China)

Abstract: Shot boundary detection (SBD) of lecture video is of great significance to teaching evaluation (TE). This paper proposes a new SBD method to address the problems that the changes of visual information of lecture videos are subtle, only boundary information is insufficient and the detection results of current methods are not beneficial to TE. The proposed method is based on the vision and text representation learning features with attention mechanism. Firstly, the hierarchical vision transformer (HViT) model is proposed to learn the visual features from the regions of interest (ROI) such as screen projection, teacher and students. Secondly, the hierarchical text transformer (HTT) model is proposed to learn features concerned in teaching evaluation from the speech and screen text. Finally, the loss function is constructed with binary cross entropies of the shot classification and boundary detection jointly. Experimental results on CLShots dataset show that the average precision, recall, F_1-score and mean intersection over union of our method are higher by 23.3%, 22.4%, 22% and 35.7% compared with those of the state-of-art method of SBLV, while higher by 13.8%, 14.5%, 14.3% and 21.3% compared with those of the method of TransNet V2.

基金项目:教育部供需对接就业育人项目(20220102204);江苏省教育科学“十四五”规划课题(D/2021/01/110);江苏高校哲学社会科学研究项目(2020SJA1425);苏州市图书馆学会重点项目(21-A-02);常熟理工学院高等教育研究项目(GJ1905)。
收稿日期:2022-04-10;**修订日期**:2022-05-18

Key words: teaching evaluation; lecture video segmentation; attention mechanism model; shot boundary detection

引 言

随着视频采集设备和处理技术的快速进步,课堂视频为开展大规模教学评价提供了可能。基于课堂视频的教学评价对学校教学质量跟踪管理、教师专业成长、学生学业发展具有重要意义。但是由于视频的非线性结构,人工观察课堂视频的评价方法[1]不能快速访问感兴趣的内容,需要顺序浏览或拖动查看部分视频后评价教学,费时费力且容易存在遗漏。当打分或评价的时候,需要反复回溯视频,却难以定位或检索需要的镜头。基于人工智能的教学评价[2-6]需要以视频镜头为单位理解教学行为等高层语义。

镜头是视频的基本单位,镜头边界检测(Shot boundary detection, SBD),又称为镜头分割,是视频数据处理和分析的重要基础性工作,在国际学术界备受关注。目前针对教学视频的SBD研究比较少,主要有基于视觉信息和文本信息两类方法。基于视觉信息的SBD方法主要利用教学视频中的幻灯片或教师讲课等场景信息,困难在于视觉信息变化不明显。教学视频的局部帧序列有很强的相关性,从全局结构可以看成一个整体。据此,Zhang等[7]提出把教学视频的视觉过渡检测问题转换为低秩稀疏矩阵分解问题,通过提取每一帧的HSV直方图和水平投影特征构建特征矩阵,然后通过低秩稀疏矩阵分解识别视觉过渡边界。此外,利用教学视频中的幻灯片,Zhao等[8]提出了基于幻灯片边缘过渡信息的SBD方法。基于文本信息的SBD方法主要利用教学视频镜头往往与教学内容主题变化保持一致性原理。Tuna等[9]提出通过OCR(Optical character recognition)和ASR(Automatic speech recognition)工具分别从屏幕和语音中提取文本,基于文本相似性识别主题,然后按照主题渐变分割课堂视频。针对英文教学视频,王敏等[10]利用OCR提取关键帧的字幕,建立潜在狄利克雷分布(Latent Dirichlet allocation, LDA)模型计算主题概率分布,从语义层面分割镜头。此外,Soares等[11]和Dipesh等[12]提出了基于语音活动检测(Voice activity detection, VAD)的SBD方法,用ASR工具从语音中提取文本后用Word2Vec描述视频的语音脚本的文本特征,再提取底层的声学特征,合并文本和声学特征后用多目标函数的遗传算法(Genetic algorithm ,GA)识别主题,按照主题分割镜头。

现有教学视频的SBD方法取得了阶段性成果,但还存在3个方面的问题:(1)以视觉信息或文本主题的单模态信息为主,没有利用视觉和文本跨模态之间的互补信息;(2)镜头分割没有从教学评价关注的内容出发,导致SBD结果不能满足教学评价需要;(3)多数方法仅利用了镜头边界的视觉或文本信息,存在视觉变化不明显或文本信息不足的问题。课堂视频一般是由固定位置监控自动拍摄,镜头捕捉的场景有限,也没有后期视频编辑和制作过程,镜头以内容主题渐变为主,存在视觉变化少和镜头边界信息不足等特性。尽管现在有多种通用领域的SBD技术,但很少有适合课堂视频数据的方法。为此,本文从面向教学评价关注的教学事件或环节出发,设计基于注意力机制的Transformer模型描述学习视觉和文本特征,联合视频镜头分类和边界检测的新方法,记为HTLV-SBD(Hierarchical transformers for lecture videos-SBD)。

本文主要贡献有:(1)针对单个视觉注意力模型ViT (Video transformer) 难以学习边缘低层特征和计算量大而难以训练问题,提出按照教学评价关注的内容针对屏幕、教师和学生等区域建立多组层次ViT模型——HViT (Hierarchical video transformers)学习视觉特征;(2)针对单个文本注意力模型Transformer不利于学习以长文本和文本长度变化较大的教学视频文本内容问题,提出按照文本的来源分为OCR和ASR文本,并依据标点符号和静音时长分割句子建立多组层次文本Transformer模型——

HTT(Hierarchical text transformers)学习教学评价关注的文本特征;(3)针对教学视频边界信息不足问题,提出了融合课堂视频镜头和边界的视觉和文本特征的SBD方法。

1 相关研究

SBD研究主要针对新闻、互联网社交、纪录片、艺术表演、广告、电影及电子商务等视频数据集[13-14],采用基于视频序列相邻帧的视觉相关性,依据镜头切换方式(突变和渐变两类)和编辑方式(淡入淡出、叠化、定格、划像、翻转翻页、慢转换和扫换等)分割镜头任务。传统方法主要通过人工定义视频运动光流[15]、分块直方图全局特征[16-17]、SURF和SIFT[18]等局部特征,综合多种特征的Walsh-Hadamard核变化[19],综合局部描述算子和全局上下文背景的图结构[20]。

当前主流方法是基于3D卷积神经网络[21-23]、C3D模型[24-25]、扩展卷积网络[26-27]模型和SCTSNet模型[28]等以卷积滤波器的深度学习方法。因为卷积只能利用局部信息来计算目标像素会带来一些偏差,缺乏全局信息。虽然可以使用更大或更深的卷积滤波器网络,然而计算开销越来越大,结果并没有得到显著改善。

基于自注意力机制的描述学习方法在模型训练和预测过程中实现全局参考,具有良好的偏差和方差平衡性。基于自注意力的文本Transformer模型[29]是自然语言处理领域的主干网络结构,Transformer模型采用固定步长把所有的文本分割成片段学习,容易导致上下文碎片化和优化低效[30],不利于学习以长文本和文本长度变化较大的教学视频文本数据。受TT模型的启发,计算机视觉领域提出了类似的注意力模型ViT[31],在性能和算力权衡中显著优于基于卷积的主干网络ResNet。但是单个ViT模型直接在所有图像分块线性嵌入学习中难以学习图像边缘等底层特征,且计算复杂度,与token序列长度的平方相关难以训练[32]。

2 面向教学评价的SBD模型设计

教学视频镜头边界存在的本质原因不是视觉信息变化,而是教学设计的教学事件或环节决定了镜头边界。一般地,教师在备课时精心设计了课堂教学环节,课堂视频存在相应的教学结构,即在一定教育思想、教学理论和学习理论指导下,在一定环境中教学活动进程的稳定结构形式。例如,教学设计学科的开拓者之一Gagne提出了9大教学事件(引起学生注意、提示教学目标、复习先前经验、呈现教学内容、指导学生学习、展现学习行为、学习反馈、评定学习效果、加强记忆与学习迁移),BOPPPS教学模式有6大教学环节(导入、学习目标、前测、参与学习、后测、总结)。孙众等[33]提出了以教学事件为主的人工智能支持课堂教学分析框架TESTII。因此,本文从教学评价关心的教学事件或环节出发寻找决定教学视频镜头边界的关键因素,构建了基于注意力机制的HViT和HTT模型分层学习视觉和文本特征,提出了如图1所示的SBD模型框架,图中b为镜头边界标记。

2.1 基于HViT模型的视觉特征描述学习

通过大量观察发现教学评价关注的典型教学事件或环节的视觉特征主要体现在屏幕、教师和学生等区域。直接利用ViT模型[31]从所有课堂视频的分块中学习到屏幕、教师和学生的视觉特征比较困难,且分块数多导致计算量大而难以训练。为了更有效学习教学评价所关注的事件或环节中教师、学生和屏幕区域的视觉特征,同时减少视觉和文本信息在网络模型中的差距,提出了如图2所示的HViT模型学习视觉特征。构建多组ViT模型有针对性地学习屏幕、教师和学生区域的视觉特征,减少无关区域的学习,提高模型训练效果和速度。使用Faster R-CNN[34]模型从课堂视频帧中检测屏幕区、教师区和k个学生区,第1层$\mathrm{ViT}_{L=1}$分别建立不同个ViT模型学习屏幕、教师和学生的视觉特征。

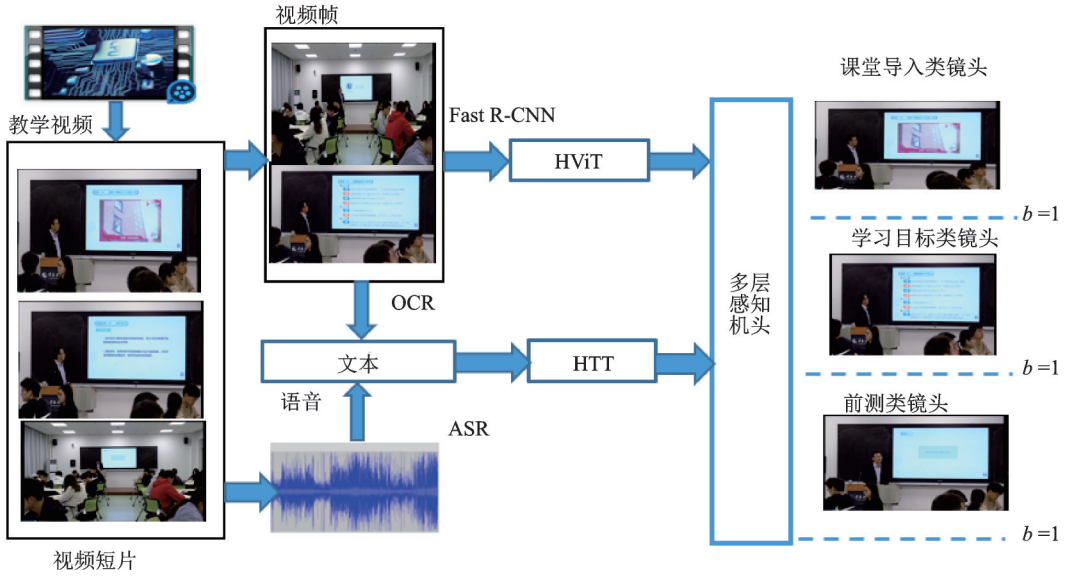

图1 教学视频镜头分类和边界检测框架

Fig.1 Framework for lecture video shot classification and SBD

第1层的$k+2$个$\mathrm{ViT}_{L=1}$模型输出$\{e_B, e_T, e_{S_1}, e_{S_2}, \cdots, e_{S_k}\}$作为第2层模型的输入,第2层采用TT模型[29],其输出e_v作为视觉特征。通过观察发现多数教学镜头至少持续1~2 s,这个时间段内视觉信息不会发生较大变化。因此,本文设置1 s间隔时间密集采样以学习视频短片中心帧的视觉特征。假设从教学视频中抽取了N_1个视频帧$F=\{f_1, f_2, \cdots, f_{N_1}\}$,每帧用Faster R-CNN模型提取屏幕、教师和学生等感兴趣区域。Transformer模型只接受1维的Token嵌入序列,为了处理2维图像数据,需要把2维的图像块转换为线性嵌入序列作为ViT模型的输入。Faster R-CNN提取的感兴趣区域大小不一致,为了简化位置编码,将视频帧中目标区域大小统一缩放为$P\times P$,并按$m\times m$大小划分为N_2个图像块,位置信息按照图像块的序号编码。

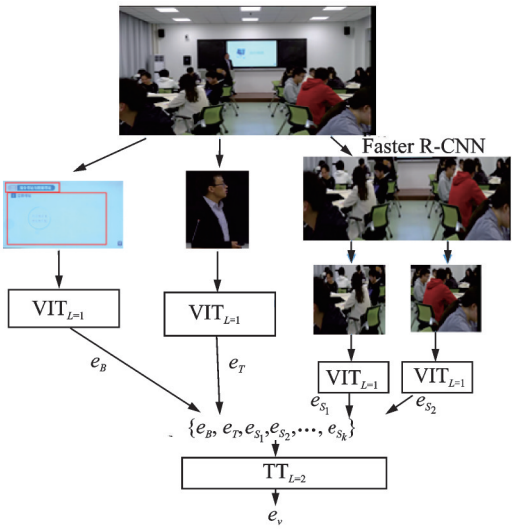

图2 HViT模型的网络结构

Fig.2 HViT architecture

假设颜色通道数为C,将第$i(1\leqslant i\leqslant N_2)$个图像块$\boldsymbol{x}^i\in\mathbf{R}^{m\times m\times C}$按照列拉长投影到1维的Token向量$\boldsymbol{x}_p^i\in\mathbf{R}^{1\times(m^2C)}$作为ViT模型的输入层。目标区域图像块的线性投影函数为

$$\boldsymbol{z}_0=[x_{\text{class}};\ \boldsymbol{x}_p^1\boldsymbol{E};\ \boldsymbol{x}_p^2\boldsymbol{E};\ \cdots;\ \boldsymbol{x}_p^{N_2}\boldsymbol{E}]+\boldsymbol{E}_{\text{pos}} \tag{1}$$

式中:x_{class}为类别信息;$\boldsymbol{x}_p$为二维图像块$\boldsymbol{x}$平展为一维的数组;$\boldsymbol{E}\in\mathbf{R}^{(m^2C)\times d}$,位置信息$\boldsymbol{E}_{\text{pos}}\in\mathbf{R}^{(N_2+1)\times d}$,其中$d$为投影后的维数。

假设Patch嵌入序列的初始状态为$z_0^0=x_{\text{class}}$,ViT模型的堆叠层数为N_3,每层网络包括多头自注意力(Multi-headed self-attention, MSA)和多层感知机(Multi-layer perceptron,MLP)两个子层,计算公式

分别如下

$$z'_l = \mathrm{MSA}(\mathrm{LN}(z_{l-1})) + z_{l-1} \quad 0 \leqslant l \leqslant L \tag{2}$$

$$z_l = \mathrm{MLP}(\mathrm{LN}(z'_l)) + z'_l \quad 0 \leqslant l \leqslant L \tag{3}$$

式中：LN(•)为Norm函数；L为隐含层的大小，最后一层输出的归一函数值为视觉特征，即

$$e = \mathrm{LN}(z_L^0) \tag{4}$$

模型$\mathrm{ViT}_{L=1}$输出屏幕、教师和k个学生区的视觉特征分别分为e_B、e_T和$\{e_{S_1}, e_{S_2}, \cdots, e_{S_k}\}$。

2.2 基于HTT模型的文本特征描述学习

教学视频的语音和屏幕文本信息是弥补视觉信息变化不明显的重要内容。采用OCR和ASR工具可以分别从屏幕和语音中识别文本。课堂教学文本的长短变化较大，且长文本较多，直接使用固定步长的Transformer模型容易导致上下文碎片化，且计算量与文本长度的平方成正比，难以训练优化。

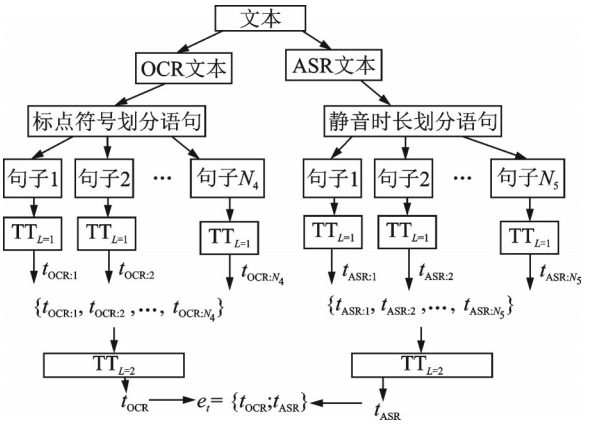

图3 HTT模型的网络结构

Fig.3 HTT Architecture

为了解决这些问题，本文把文本按照来源分为屏幕OCR和语音ASR两类文本。OCR识别文本根据标点符号划分为N_4条语句。根据超过1 000 ms的静音点把ASR识别的文本拆分为N_5个句子。依据文本的分层信息，构建了如图3所示的层次文本Transformers模型HTT。第1层分别为OCR和ASR识别的文本句子建立TT模型，记为$\mathrm{TT}_{L=1}$。上下文关系是文本内容的重要信息，类似语句内单词位置的编码方法，分别对OCR和ASR识别的奇数句采用正弦函数编码，偶数句采用余弦函数编码。OCR识别的第j_1条语句上下文顺序编码($1 \leqslant j_1 \leqslant N_4$)为

$$\mathrm{SE}(j_1, 2i) = \sin(j_1 / 10000^{2i/dm}) \tag{5}$$

$$\mathrm{SE}(j_1, 2i+1) = \cos(j_1 / 10000^{2i/dm}) \tag{6}$$

式中dm表示单词向量维度。ASR识别的第j_2条语句上下文顺序编码($1 \leqslant j_2 \leqslant N_5$)为

$$\mathrm{SE}(j_2, 2i) = \sin(j_2 / 10000^{2i/dm}) \tag{7}$$

$$\mathrm{SE}(j_2, 2i+1) = \cos(j_2 / 10000^{2i/dm}) \tag{8}$$

假设语句的词语序列为$(\omega^0, \omega^1, \cdots, \omega^{N_6})$，词向量和基于正余弦编码的位置信息构成$\mathrm{TT}_{L=1}$模型的网络输入层，$\mathrm{TT}_{L=1}$模型包括$N_7$个堆叠层。每个堆叠层包含2个子层：第1个子层包含4个头的注意力子层，残差连接后用Norm函数归一化处理；第2个子层向前反馈层和残差连接后用Norm函数归一化处理。最后输出$\{y_0, y_1, \cdots, y_{N_6}\}$的均值作为句子文本特征。

学习OCR文本输出的文本特征$\{t_{\mathrm{OCR},1}, t_{\mathrm{OCR},2}, \cdots, t_{\mathrm{OCR},N_4}\}$作为第2层模型$\mathrm{TT}_{L=2}$的输入，学习ASR文本输出的文本特征$\{t_{\mathrm{ASR},1}, t_{\mathrm{ASR},2}, \cdots, t_{\mathrm{ASR},N_5}\}$作为第2层模型$\mathrm{TT}_{L=2}$的输入，最后合并OCR和ASR两个$\mathrm{TT}_{L=2}$模型的输出$e_t = \{t_{\mathrm{OCR}}; t_{\mathrm{ASR}}\}$作为文本特征。

2.3 融合视觉和文本特征的镜头分类和SBD

教学镜头边界的视觉变化有限，且文本信息也较少。仅用镜头边界信息不足以学习教学视频的镜

头属性。为此,本文提出联合镜头分类和镜头边界信息学习网络参数,网络结构如图4所示。通过全连接层融合视觉和文本共享特征,视频镜头按照教学事件或环节分为课堂导入、学习目标、测试、参与式学习过程和课程总结等类别。如果前后两个镜头类别相同,则镜头边界标记$b=0$,否则镜头边界标记$b=1$。

通过全连接层融合视觉和文本特征,视频特征e_v映射到共享子空间的过程如下

$$v^*=f(W_3f(W_1e_v+b_1)+b_3) \tag{9}$$

式中:W_1和W_3为权重;b_1和b_3为偏差;f为激活函数;v^*为视觉特征e_v在共享子空间的映射。文本特征e_t映射到共享子空间的过程如下

$$e_t^*=f(W_3f(W_2e_t+b_2)+b_3) \tag{10}$$

式中:W_2和b_2分别为权重和偏差;e_t^*为文本特征e_t在共享子空间的映射。

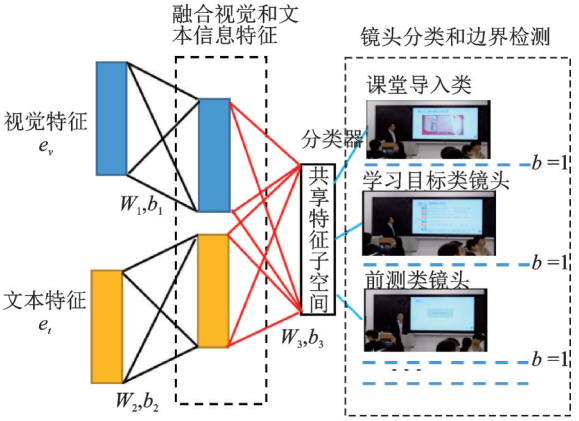

图4 基于MLP的教学视频镜头分类和边界检测

Fig.4 Lecture video shot classification and SBD using MLP

联合教学视频镜头分类和边界检测的损失函数定义为

$$L=\alpha L_{\text{classification}}+\beta L_{\text{SBD}} \tag{11}$$

式中α和β为分类损失函数和边界检测损失函数的权重,且$\alpha+\beta=1$。

基于二值交叉熵的镜头分类损失函数定义为

$$L_{\text{classification}}=-(l\log(q)+(1-l)\log(1-q)) \tag{12}$$

式中:l为真实的镜头分类二值函数值,如果视频类别标签分类正确则为$l=1$,否则为$l=0$;q为教学视频镜头分类的预测概率。

基于二值交叉熵的镜头边界检测损失函数定义为

$$L_{\text{SBD}}=-(l'\log(q')+(1-l')\log(1-q')) \tag{13}$$

式中:l'为真实的镜头边界二值函数值;q'为教学视频镜头边界的预测概率。如果是镜头边界则为$l'=1$,否则为$l'=0$。

2.4 时间和空间复杂度分析

由于自注意力机制的特征,TT模型[29]和ViT模型[31]在推理过程中需要计算和保存$n\times n$大小的注意力矩阵,其计算时间和内存空间的复杂度为$O(n^2)$,其中n为token序列长度。对于教学视频中的文本数据,普遍存在长文本,所以时间复杂度较高。本文的HTT模型通过建立N_4个层次TT模型学习OCR文本特征,时间和空间复杂度可以降低为$O(n^2/N_4)$。HTT模型通过建立N_5个层次TT模型学习ASR文本,时间和空间复杂度降低为$O(n^2/N_5)$。二维视频帧的token序列较长,所以ViT模型的计算量和内存空间较大。本文HViT模型使用Fast R-CNN模型分割出屏幕、教师和学生区域大小比视频帧小,HTT模型通过建立$k+2$个层次ViT模型学习视觉特征,减少了视觉token的序列长度,时间和空间复杂度降低为$O(n^2/(k+2))$。

3 实验研究

3.1 数据集和预处理

为了研究面向教学评价的课堂视频SBD方法,本文建立了数据集CLShots。由人工根据上下文场

景标注教学镜头的类别和边界，大规模教学视频中的SBD挑战在于镜头标注。教学视频一般长达1～2 h，较长时间的教学视频镜头标注比较费时、容易出现疲倦而导致标注质量下降。为此，本文采取了文献[28]的策略，按照教学设计的脚本把教学视频划分为时长为几分钟的视频短片，然后在短片上进行标注。

CLShots有160个教学视频，视频时长50～110 min左右，共有6 500个镜头，每个镜头有视觉和文本信息，部分教学事件或环节镜头类别实例如图5所示。课堂导入、学习目标、测试、参与式学习过程以及课程总结等5类镜头数分别有500、500、500、4 500和500个，分别选取300、300、300、4 300和300个镜头作为训练集，每类分别选择100个镜头作为验证集，每类分别选择100个镜头作为测试集。

采用OCR工具MMMT[35]从屏幕中识别文本，ASR工具Freeswitch从语音中识别文本。因为较重的口音、技术词汇、教室讲课环境噪声等原因，尽管现在OCR和ASR工具识别的正确率越来越高，但还是不可避免地存在错误识别，对教学视频镜头分割有一定影响。为了减少这种影响，文献[9]采用了人工纠错的办法，但效率比较低。本文采用了基于深度学习模型的中文文本纠错工具Pycorrector（https://github.com/shibing624/ pycorrector），可以有效降低OCR和ASR的识别错误。使用中文分词工具jieba完成了停用词清洗、分词和分句等文本预处理。

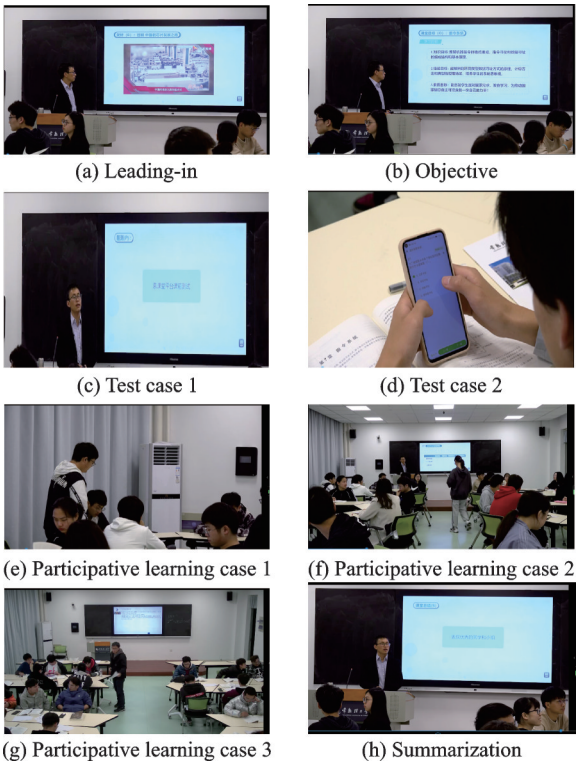

图5 教学视频镜头实例

Fig.5 Lecture video shot instances

3.2 评价指标

本文定义精确率、召回率、F_1分数和平均交并比4个指标评测算法。

(1)精确率P表示预测为正样本中有多少是真正的正样本，即

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{14}$$

式中TP、FP和FN分别表示分类正确的正样本数、分类错误的正样本数和分类错误的负样本数。

(2)召回率R表示样本中的正例有多少被预测正确，即

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{15}$$

(3)F_1分数是P和R的调和平均值，即

$$F_1=\frac{2PR}{P+R} \tag{16}$$

(4)平均交并比M_{iou}是真实值和预测值2个集合之间交集和并集的比例，即

$$M_{\text{iou}}=\frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{17}$$

3.3 实验环境

实验所用服务器配置:操作系统Ubuntu Linux、内存512 GB、GPU Tesla P100 16 GB。开发环境为Cuda 11.0+Pytorch 1.6.0+Pycharm 2021。训练模型优化器为AdamW,批大小为64,初始学习率为0.001,每30个epoch按10%的速率递减。

3.4 实验内容和结果分析

3.4.1 损失函数的参数实验

教学视频镜头分类和边界检测的联合损失函数参数α和β对模型的性能具有一定影响。如果α过大,则会导致镜头边界的优化被淹没。反之,如果α过小,则镜头分类的优化被淹没。为此,本文通过实验选出了最优的α和β参数值。采用相同的视觉和文本特征,参数权重范围在0~1之间,并以0.1的步长变化,比较损失函数的F_1分数和M_{iou}指标性能见表1。实验数据表明,镜头分类和边界检测对损失函数均具有较大的影响,如果仅仅利用边界信息或镜头分类不能得到最优的损失函数值,结合镜头分类和边界检测的准确率更高。从表1可知,当参数α=0.3和β=0.7时,F_1和M_{iou}指标的平均值最优。

表1 不同参数下SBD性能指标

Table 1 SBD performance index with different parameters

指标	参数									
	α=0.0 β=1.0	α=0.1 β=0.9	α=0.2 β=0.8	α=0.3 β=0.7	α=0.4 β=0.6	α=0.5 β=0.5	α=0.6 β=0.4	α=0.7 β=0.3	α=0.8 β=0.2	α=0.9 β=0.1	α=1.0 β=0.0
F_1	0.52	0.65	0.71	**0.72**	0.70	0.68	0.64	0.58	0.52	0.49	0.36
M_{iou}	0.48	0.36	0.48	**0.57**	0.49	0.43	0.37	0.31	0.25	0.21	0.15

3.4.2 消融实验

消融实验中,一方面比较HViT模型和ViT[31]模型性能以及HTT模型和TT模型[29]性能。另一方面比较视觉特征、文本特征和二者融合后的3类特征有效性。本文设计了基于2类模型和3类特征的6种组合比较实验。

HTT和TT模型设置:TT模型中堆叠层N_7=6,多头数为4,词向量维度dm=512;HTT模型的第1层$\text{TT}_{L=1}$和第2层$\text{TT}_{L=2}$与TT模型设置一致。

HViT和ViT模型设置:ViT模型中堆叠层N_3=12,隐含层大小d=768,MLP大小L= 3 072,多头数为12,图像块大小参数m=16。HViT模型基于ViT和TT模型,感兴趣目标区域大小参数P=112,HViT模型第1层$\text{ViT}_{L=1}$的参数与ViT模型一致,第2层中$\text{TT}_{L=2}$与TT模型参数一致。

各类镜头4个指标的平均值如表2所示,从特征视角比较模型性能。从视觉特征看,HViT模型的P、R、F_1和M_{iou}指标比ViT模型的指标分别提高了36%、27.8%、28.6%和36.4%,主要原因是HViT模型分组、分层学习了教学视频中屏幕、教师和学生区域更精细的视觉特征。从文本特征

表2 视觉、文本和融合特征的性能指标

Table 2 Indexes of vision, text and fused features

模型	特征	P	R	F_1	M_{iou}
ViT	视觉	0.25	0.18	0.21	0.11
TT	文本	0.47	0.39	0.43	0.32
ViT+ TT	融合	0.61	0.52	0.56	0.43
HViT	视觉	0.34	0.23	0.27	0.15
HTT	文本	0.68	0.51	0.58	0.42
HViT+HTT	融合	**0.74**	**0.71**	**0.72**	**0.57**

看，HTT模型的P、R、F_1和M_{iou}比TT模型分别提高了44.7%、30.8%、34.9%和31.3%，主要原因是HHT模型分组分层学习了OCR和ASR不同来源文本句子的精细文本特征。融合视觉和文本特征，HViT＋HTT模型比ViT＋TT模型的P、R、F_1和M_{iou}分别提高了21.3%、36.5%、28.6%和32.6%。综上可见，本文的HViT和HTT模型比传统的ViT和TT模型学习的视觉和文本特征更有效。

从模型视角比较视觉、文本和融合特征的性能。ViT模型的视觉特征的P、R、F_1和M_{iou}指标分别占比TT文本特征指标的53.2%、46.2%、48.8%和34.4%，分别占比ViT＋TT融合特征指标的41%、34.6%、37.5%和25.6%。类似地，HViT模型的视觉特征的P、R、F_1和M_{iou}指标分别占比HTT文本特征指标的50%、45.1%、46.6%和35.7%，分别占比HViT＋HTT融合特征指标的45.9%、32.4%、37.1%和26.3%。结果表明，教学视频镜头的视觉特征区分效果最差，文本特征的区分能力比视觉特征更好，而视觉和文本的融合特征性能最好，教学视频中的文本和视觉信息具有很好的互补性。融合视觉和文本的深度学习特征获得了更为丰富的镜头内容和边界特征，提升了镜头分类和边界检测的性能指标。

3.4.3 对比实验

为了定量评价方法的有效性，选择了专门针对教学视频数据镜头分割方法和本文的HTLV-SBD方法进行了对比实验：Zhang等的低秩稀疏矩阵奇异值分解法(SVD)[7]，Zhao等的幻灯片边缘过渡信息(SBLV)[8]、Tuna等的主题分割法(Topic)[9]，王敏等的字幕LDA主题法(S4VM)[10]。由于上述比较方法以传统浅层机器学习为主，为了增加对比公平性，选择了通用领域的几个深度学习方法与本方法比较：Hassanien提出的时空卷积方法(DeepSBD)[23]、Gygli提出的全连接方法(FCNNSBD)[25]，Souček提出的扩展卷积网络方法(TransNet V2)[27]。不同方法在5类镜头的指标平均值如表3所示。可以看出，SVD方法仅仅利用了视觉信息，4个指标的均值最低，表明教学事件或主要环节的视觉信息变化不明显；S4VM方法仅利用了OCR工具从字幕中提取的文本信息建立主题分割，尽管比SVD方法的指标好，但是镜头边界特征的区分度还不充分；Topic方法从语音和屏幕中提取文本信息，比S4VM方法提取了更为全面的文本信息，4个指标优于S4VM和SVD方法；SBLV方法通过检测鼠标指针或激光笔的点、幻灯片的过渡、动画显示等幻灯片进度等3个主要事件，识别幻灯片的Canny边缘特征和文本信息等镜头内容，比Topic、S4VM、SVD等方法更好；DeepSBD方法使用了卷积神经网络学习特征并输入SVM分类，比传统的SVD和S4VM性能要好，但是仅仅利用了图像特征，缺乏文本特征补充，性能不及SBLV方法。此外，由于没有直接用于端到端分类器，性能在深度学习方法中最差；FCNNSBD方法使用了网络较小的3D卷积网络，但指标比DeepSBD方法好；TransNet V2方法使用扩展卷积网络取得了比FCNNSBD方法更好的结果，但仅仅利用了图像信息，缺乏文本信息的补充。HTLV-SBD方法结果最好，比SBLV方法的P、R、F_1和M_{iou}指标分别提高了23.3%、22.4%、22%和35.7%，比TransNet V2方法的P、R、F_1和M_{iou}指标分别提高了13.8%，14.5%、14.3%和21.3%。主要原因在于注意力机制Transformer模型学习的视觉和文本特征信息比其他方法的特征更精准，且融合了视觉和文本特征的互补性，联合了镜头分类为镜头边界特征补充了更多信息。

表3 不同方法的性能指标

Table 3 Performance comparison of different methods

方法	P	R	F_1	M_{iou}
SVD	0.35	0.29	0.32	0.19
SBLV	0.60	0.58	0.59	0.42
Topic	0.58	0.52	0.55	0.38
S4VM	0.45	0.41	0.43	0.27
DeepSBD	0.55	0.51	0.53	0.36
FCNNSBD	0.61	0.55	0.58	0.41
TransNet V2	0.65	0.62	0.63	0.47
HTLV-SBD	**0.74**	**0.71**	**0.72**	**0.57**

为了更全面地展示教学视频镜头分类精度，比较了面向教学视频性能较好的传统方法SBLV和通

用领域深度学习方法TransNet V2与HTLV-SBD方法的混淆矩阵(Confusion matrix, CM)。假设课堂导入、学习目标、测试、参与式学习过程、课程总结等5类分别记为c_1、c_2、c_3、c_4和c_5。SBLV、TransNet V2与HTLV-SBD方法对应的混淆矩阵分别见表4~6。结果表明,HTLV-SBD对5类镜头分类结果优于SBLV和TransNet V2方法,课程导入和测试类的分类效果较好,学习的特征效果比较明显。参与式学习过程类的分类效果比较差,因为不同课程不同师生的参与式教学差别比较大。

表4 SBLV方法的混淆矩阵

Table 4 CM of SBLV method

真实	预测类别				
	c_1	c_2	c_3	c_4	c_5
c_1	55	8	7	20	10
c_2	8	60	5	18	9
c_3	4	3	65	19	8
c_4	9	12	7	53	19
c_5	10	6	4	24	56

表5 TransNet V2方法的混淆矩阵

Table 5 CM of TransNet V2 method

真实	预测类别				
	c_1	c_2	c_3	c_4	c_5
c_1	61	5	2	23	9
c_2	8	59	5	21	7
c_3	6	5	71	13	5
c_4	10	9	3	58	20
c_5	9	6	4	18	63

表6 HTLV-SBD方法的混淆矩阵

Table 6 CM of HTLV-SBD method

真实	预测类别				
	c_1	c_2	c_3	c_4	c_5
c_1	**78**	1	0	16	5
c_2	4	**72**	4	13	7
c_3	0	2	**77**	14	7
c_4	7	8	1	**60**	24
c_5	5	4	0	23	**68**

4 结束语

面向教学评价需求,本文提出了层次视觉和文本注意力机制模型HViT和HTT学习教学评价重点关注的视觉和文本信息特征,利用互补性融合视觉和文本2种模态特征克服了教学视频视觉信息变化少的问题,联合视频镜头分类和边界检测克服了镜头边界信息不足的问题。数据集CLShots的实验结果表明,本文的HTLV-SBD方法能有效地分割教学视频镜头以满足教学评价的需要。但是,本文仅考虑了视频的视觉和文本信息,没有利用重要的语音信号特征,后续研究将综合利用视觉、语音和文本信息实现教学SBD,推动基于视频的大规模教学质量的常态化监测发展。

参考文献:

[1] 郑太年,仝玉婷.课堂视频分析:理论进路、方法与应用[J].华东师范大学学报(教育科学版),2017,35(3):126-132.
ZHENG Tainian, TONG Yuting. Classroom video analysis: Theoretical approaches, methods and applications [J]. Journal of East China Normal University Educational Sciences, 2017, 35(3): 126-132.

[2] 刘清堂,何皓怡,吴林静,等.基于人工智能的课堂教学行为分析方法及其应用[J].中国电化教育,2019,392(9):13-21.
LIU Qingtang, HE Haoyi, WU Linjing, et al. Classroom teaching behavior analysis method based on artificial intelligence and its application[J].China Educational Technology, 2019, 392(9): 13-21.

[3] 周鹏霄,邓伟,郭培育,等.课堂教学视频中的S-T行为智能识别研究[J].现代教育技术,2018,28(6):54-59.
ZHOU Pengxiao, DENG Wei, GUO Peiyu, et al. research on intelligent recognition of S-T behavior in classroom teaching video[J].Modern Educational Technology, 2018, 28(6): 54-59.

[4] 胡钦太,伍文燕,冯广,等.人工智能时代高等教育教学评价的关键技术与实践研究[J].开放教育研究,2021,27(5):15-23.
HU Qintai, WU Wenyan, FENG Guang, et al. Research on the key technology and practice of higher education teaching evaluation in the AI era[J]. Open Education Research, 2021, 27(5): 15-23.

[5] 吴立宝, 曹雅楠, 曹一鸣.人工智能赋能课堂教学评价改革与技术实现的框架构建[J].中国电化教育, 2021,412(5):94-100.
WU Libao, CAO Yanan, CAO Yiming. Reform and practical paths of classroom teaching evaluation under artificial intelligence [J]. China Educational Technology, 2021, 412(5): 94-100.

[6] 牟智佳, 刘珊珊, 陈明选.循证教学评价:数智化时代下高校教师教学评价的新取向[J].中国电化教育, 2021,416(9):104-111.
MOU Zhijia, LIU Shanshan, CHEN Mingxuan. The evidence-based teaching evaluation: A new orientation of teaching evaluation in colleges and universities in the era of digital intelligence[J] China Educational Technology, 2021, 416(9): 104-111.

[7] ZHANG Xiangrong, LI Chen, LI Shangwen, et al. Automated segmentation of MOOC lectures towards customized learning [C]//Proceedings of the IEEE International Conference on Advanced Learning Technologies, Austin, TX, USA:IEEE, 2016: 20-22.

[8] ZHAO Baoquan, XU Songhua, LIN Shujin, et. al. A new visual interface for searching and navigating slide-based lecture videos [C]// Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019: 928-933.

[9] TUNA T, JOSHI M, VARGHESE V, et al. Topic based segmentation of classroom videos[C]//Proceedings of the Frontiers in Education Conference. EI Paso, TX, USA: IEEE, 2015. DOI:10.1109/FIE.2015.7344336.

[10] 王敏,王斌,沈钧戈,等.教学视频的文本语义镜头分割和标注[J].数据采集与处理,2016,31(6): 1171-1177.
WANG Min, WANG Bin, SHEN Junge, et al. Lecture video text semantic shot segmentation and annotation[J].Journal of Data Acquisition and Processing, 2016,31(6): 1171-1177.

[11] SOARES E R, BARRÉRE E. An optimization model for temporal video lecture segmentation using word2vec and acoustic features[C]//Proceedings of the 25th Brazillian Symposium on Multimedia and the Web. Rio De Janeiro, Brazil: Association for Computing Machinery, 2019: 513-520.

[12] DIPESH C, HASAN O. A framework for lecture video segmentation from extracted speech content[C]//Proceedings of the 19th World Symposium on Applied Machine Intelligence and Informatics. Herl′any, Slovakia: IEEE, 2021: 299-304.

[13] SINGH A, SINGH TD, BANDYOPADHYAY S. V2T: Video to text framework using a novel automatic shot boundary detection algorithm[J]. Multimedia Tools and Applications, 2022, 81: 17989-18009.

[14] 鲁雨佳,陈实,帅世辉,等.基于剪辑元素属性约束的可计算产品展示视频自动剪辑框架[J].计算机辅助设计与图形学学报, 2020,32(7): 1101-1110.
LU Yujia, CHEN Shi, SHUAI Shihui, et al. Computational product presentation video editing framework based on editing attribute constraints[J].Journal of Computer-Aided Design & Computer Graphics, 2020,32(7): 1101-1110.

[15] 汪荣贵, 胡健根, 杨娟,等. 光流的镜头边界检测[J]. 光电工程, 2016, 43(11): 38-45.
WANG Ronggui, HU Jiangen, YANG Juan, et al. Shot boundary detection based on optical flow[J].Opto-Electronic Engineering, 2016, 43(11): 38-45.

[16] CHAKRABORTY S, THOUNAOJAM D M, SINHA N. A shot boundary detection technique based on visual colour information[J]. Multimedia Tools and Applications, 2021, 80(4): 1-16.

[17] CHAKRABORTY S, SINGH A, THOUNAOJAM D M. A novel bifold-stage shot boundary detection algorithm: Invariant to motion and illumination[J]. The Visual Computer, 2022, 38(2): 445-456.

[18] ZHOU Shangbo, WU Xia, QI Ying, et al. Video shot boundary detection based on multi-level features collaboration[J]. Signal, Image and Video Processing, 2021,15: 627-635.

[19] LAKSHMI P G, DOMNIC S. Walsh-Hadamard transform kernel-based feature vector for shot boundary dtection [J]. IEEE Transactions on Image Processing, 2014, 23(12): 5187-5197.

[20] BISWAS S, MILANFAR P. One shot detection with Laplacian object and fast matrix cosine similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(3): 546-562.

[21] DU T, LUBOMIR B, ROB F, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile: IEEE, 2015: 4489-4497.

[22] XU Jingwei, SONG Li, XIE Rong. Shot boundary detection using convolutional neural networks[C]//Proceedings of 2016 Visual Communications and Image Processing. Chengdu, China: IEEE, 2016. DOI: 10.1109/VCIP.2016.7805554.

[23] HASSANIEN A, ELGHARIB M, SELIM A, et al. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks[[EB/OL]. (2017-7-27)[2022-03-15]. https://arxiv.org/abs/1705.03281.

[24] WU Lifang, ZHANG Shuai, JIAN Meng, et al. Two stage shot boundary detection via feature fusion and spatial-temporal convolutional neural networks[J]. IEEE Access, 2019(7): 77268-77276.

[25] GYGLI M. Ridiculously fast shot boundary detection with fully convolutional neural networks[C]// Proceedings of the 16th International Conference on Content-Based Multimedia Indexing. La Rochelle, France: IEEE Computer Society, 2018. DOI: 10.1109/CBMI.2018.8516556.

[26] SOUČEK T, MORAVEC J, LOKOČ J. TransNet: A deep network for fast detection of common shot transitions[EB/OL]. (2019-6-8)[2022-03-15]. https://arxiv.org/abs/1906.03363.

[27] SOUČEK T, LOKOČ J. TransNet V2: An effective deep network architecture for fast shot transition detection [EB/OL]. (2020-8-11)[2022-3-15]. https://arxiv.org/abs/2008.04838.

[28] JIANG Xuekun, JIN Libiao, RAO Anyi, et al. Jointly learning the attributes and composition of shots for boundary detection in videos[J]. IEEE Transactions on Multimedia, 2021. DOI:10.1109/TMM.2021.3092143.

[29] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[C]//Proceedings of the Conference Advances in Neural Information Processing Systems. [S.l.]: ACM, 2017: 5999-6009.

[30] DAI Zihang, YANG Zhilin, YANG Yiming, et al. Transformer-XL: Attentive language models beyond a fixed-length context [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2020. DOI:10.18653/v1/ P19-1285.

[31] ALEXEY D, LUCAS B, ALEXANDER K, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations. Virtual Event, Austria:[s.n.], 2021. DOI: 10.48550/arXiv.2010.11929.

[32] YUAN Li, CHEN Yunpeng, WANG Tao, et al. Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet[C]//Proceedings of the IEEE International Conference on Computer Vision 2021. [S.l.]:IEEE, 2021.DOI:10.48550/arXiv.2101.11986.

[33] 孙众, 吕恺悦, 施智平, 等. TESTII框架:人工智能支持课堂教学分析的发展走向[J]. 电化教育研究, 2021, 42(2): 33-39,77.
SUN Zhong, LV Kaiyue, SHI Zhiping, et al. TESTII framework: The tendency of artificial intelligence to support classroom teaching analysis[J]. e-Education Research, 2021, 42(2): 33-39,77.

[34] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[35] HUANG Jing, PANG Guan, KOVVURI R, et al. A multiplexed network for end-to-end multilingual OCR[C]//Proceedings of the Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE Computer Society, 2021: 4545-4555.

作者简介:

谢从华(1978-),**通信作者**,男,博士,副教授,硕士生导师,研究方向:图像处理和机器学习,E-mail: xiech@aliyun.com。

罗德凤(2000-),女,本科生,研究方向:英语教育数字化,E-mail: 1327459171@qq.com。

方雨洁(2001-),女,本科生,研究方向:教育人工智能,E-mail: 2781862725@qq.com。

(编辑:张黄群)