ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 5,Sep. 2025,pp. 1165－1176
DOI:10. 16337/j. 1004-9037. 2025. 05. 005

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

基于细粒度视觉与音视双分支融合的情感视频字幕生成

龚禹轩，韩婷婷

(杭州电子科技大学计算机学院,杭州 310018)

摘　要：情感视频字幕生成作为融合视觉语义解析与情感感知的跨模态任务,其核心挑战在于精准捕捉视觉内容中蕴含的情感线索。现有方法存在两点显著不足:一是对视频中主体(人物、物体等)与其外观特征、动作特征间的细粒度语义关联挖掘不够充分,导致视觉内容理解缺乏精细化支撑;二是忽视了音频模态在情感判别与内容语义对齐中的辅助价值,限制了跨模态信息的综合利用。针对上述问题,本文提出细粒度视觉与音视双分支融合框架。其中,细粒度视觉特征融合模块通过视觉、物体、动作特征的两两交互与深度融合,有效建模视频实体与视觉上下文间的细粒度语义关联,实现对视频内容的精细化解析;音频-视觉双分支全局融合模块则构建跨模态交互通道,将整合后的视觉特征与音频特征进行深层融合,充分发挥音频信息在情感线索传递与语义约束上的补充作用。在公开基准数据集上对本文方法进行验证,其评价指标均优于CANet、EPAN等对比方法,情感指标比EPAN方法平均提高4%,语义指标平均提升0.5,综合指标平均提升0.7。实验结果表明本文方法能有效提升情感视频字幕生成的质量。

关键词：情感视频字幕生成;跨模态情感感知;细粒度特征融合;注意力机制;视频理解
中图分类号：TP391.4　　**文献标志码**:A

Emotional Video Captioning Based on Fine-Grained Visual and Audio-Visual Dual-Branch Fusion

GONG Yuxuan, HAN Tingting

(School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Emotional video captioning, as a cross-modal task integrating visual semantic parsing and emotional perception, faces the core challenge of accurately capturing the emotional cues embedded in visual content. Existing methods have two notable limitations: First, they insufficiently explore the fine-grained semantic correlations between video subjects (such as humans and objects) and their appearance and motion features, leading to a lack of refined support for visual content understanding; second, they neglect the auxiliary value of the audio modality in emotional discrimination and content semantic alignment, which restricts the comprehensive utilization of cross-modal information. To address these issues, this paper proposes a framework based on fine-grained visual and audio-visual dual-branch fusion. Specifically, the fine-grained visual feature fusion module effectively models the fine-grained semantic associations between video entities and visual contexts through pairwise interactions and deep integration of visual, object, and motion features, thereby achieving refined parsing of video content. The audio-visual dual-branch global fusion module constructs a cross-modal interaction channel to deeply fuse the integrated visual features with audio features, fully leveraging the supplementary role of audio information in

收稿日期:2025-06-15;**修订日期**:2025-08-30

emotional cue transmission and semantic constraint. Validation experiments on public benchmark datasets show that the proposed method outperforms comparative methods such as CANet and EPAN across evaluation metrics. It achieves an average improvement of 4% over EPAN method in emotional metrics, an average increase of 0.5 in semantic metrics, and an average boost of 0.7 in comprehensive metrics. Experimental results demonstrate that the proposed method can effectively enhance the quality of emotional video captioning.

Key words: emotional video captioning; cross-modal emotional perception; fine-grained feature fusion; attention mechanism; video understanding

引　言

随着社交网络的日益普及和风靡，人们借助图片或视频在社交平台上表达观点变得愈发轻松。其中，视频凭借更丰富的情感传递能力受到大众青睐，这也使得视觉情感分析的需求愈发迫切。近年来，细粒度图像检索[1]技术的提升和图像理解[2]任务的进步，已经推动计算机在图像的识别、感知及理解方面取得重大突破。然而，视频内容理解，尤其是情感感知类视频理解任务，涵盖视频情感识别[3-4]、情感音乐视频检索[5]、情感视频字幕生成[6-7]等，仍面临重大挑战，这些正逐渐成为当前的研究热点。作为计算机视觉领域新兴的热门课题，情感视频字幕生成任务不仅要求模型精准解读视频内容，还需识别出视频中潜藏的复杂情感，并结合情感语义生成与之匹配的情感字幕。

传统的视频字幕生成方法主要是依据遵循一定语法规则的模板来生成句子，Kojima等[8]从人体视线、手的位置、身体姿势以及和其他物体关系提取动作的语义特征和动作相关词对应，然后找到这些词语的关联并生成句子。Krishnamoorthy等[9]通过利用大型语料库中挖掘的知识识别最佳的主谓宾三元组，然后根据这个三元组来生成对应的句子。但上述两种方法对模板严重依赖，导致句子形式固定，缺乏灵活性和多样性。受益于深度学习的快速发展和循环神经网络(Recurrent neural network, RNN)与卷积神经网络(Convolution neural network, CNN)的成功，编码器-解码器结构被广泛应用于生成具有灵活语法结构的描述。Venugopalan等[10]将连续视频帧通过CNN并均值池化得到视觉特征，然后通过LSTM[11]学习视频帧序列和单词序列从而生成字幕。随着Faster-RCNN[12]的兴起，通过Faster-RCNN等提取的事实对象在生成字幕中发挥着重要作用。Zhang等[13]提出一种基于双向时空图捕获视频中显著对象的详细时间动态的方法，然后利用GRU[14]从时间轨迹中学习视频目标动态信息以生成字幕。佟国香等[15]使用图神经卷积网络和引导向量构建了图像字幕生成模型。

情感视频字幕生成任务作为传统视频字幕生成任务的扩展，首先要理解视觉信息中包含的情感线索[16-17]。Yang等[18]根据语义概念和视觉特征构建情感图，并利用场景特征指导通过图卷积网络(Graph convolution network, GCN)对情感图推理得到的情感增强对象特征和基于场景的注意力机制进行融合。Mittal等[19]使用基于注意力策略和格兰杰因果关系对时间因果关系进行建模以用于多媒体内容的时间序列预测。在这个基础上，Wang等[20]公开一个新的情感视频字幕数据集。Song等[21]提出了一个上下文注意力网络，通过注意力机制学习视觉和语义之间的关系来识别和描述视频中的事实与情感；并且提出一种树形结构的情绪学习模块，实现明确且细粒度的情绪感知[22]。Ye等[7]提出了双路协同生成网络，总结全局视觉情感线索，然后通过局部时刻动态增强或抑制不同粒度子空间的语义实现情感的进化。然而，上述方法存在明显局限：一方面，它们未能深入挖掘视频中人物、物品等主体与其外观及动作变化之间存在的细粒度视觉特征语义；另一方面，也忽视了视频音频在理解视频内容与情感时所起到的辅助作用。

本文提出一种基于细粒度视觉与音视双分支融合框架。首先针对上面提到的细粒度视觉特征语义挖掘不足问题,本文提出细粒度视觉特征融合模块,通过对视觉、物体、动作3种特征进行细粒度融合,深度挖掘视频中主体与其动态变化间的语义关联;针对先前工作对音频特征辅助作用忽视的问题,本文提出了音频-视频双分支全局融合模块,借助音频模态的局部与全局建模,充分发挥音频信息对视频理解及情感感知的补充作用,最终实现更生动贴切的视频字幕生成。对于细粒度视觉特征融合模块,融合外观特征和运动特征,以捕捉外观变化和运动轨迹的时空一致性;融合外观特征和物体特征,以过滤与视频中关键人物、物品无关的背景噪声信息;融合运动特征和物体特征,以过滤运动序列中的冗余信息同时增强符合物体运动规律的运动特征。最后将3种融合特征经过拼接和多层感知机(Multilayer perceptron,MLP)进一步整合为捕捉人物和物品等主体、外观细节和动作动态三者之间细微语义关联的统一视觉特征,实现对视频内容的精细化理解。对于音频-视觉双分支全局融合模块,将上述整合后的视觉特征与音频特征进行跨模态融合,既保留视觉内容的细节特征,又融入音频传递的情感线索,显著提升了对视频内容和情感的综合理解能力。

综上所述,本文的贡献总结如下:(1)提出细粒度视觉特征融合模块,通过多维度视觉特征的两两交互,有效挖掘视频主体、外观、动作间的细粒度语义,为视频内容理解提供更精准的视觉表征;(2)提出音频-视觉双分支全局融合模块,实现视觉和音频特征的深层融合,充分发挥音频信息对视觉语义对齐和情感理解的辅助作用;(3)通过实验对本文方法在EmVidCap-L数据集[7]的3大类评价指标进行性能评估,并通过模型指标对比和评价指标分析验证了本文方法的性能和有效性。

1 本文方法

本文提出的基于细粒度视觉与音视双分支融合框架如图1所示,主要包括细粒度视觉特征融合模块、音频-视觉双分支全局融合模块和情感线索感知模块3个部分。

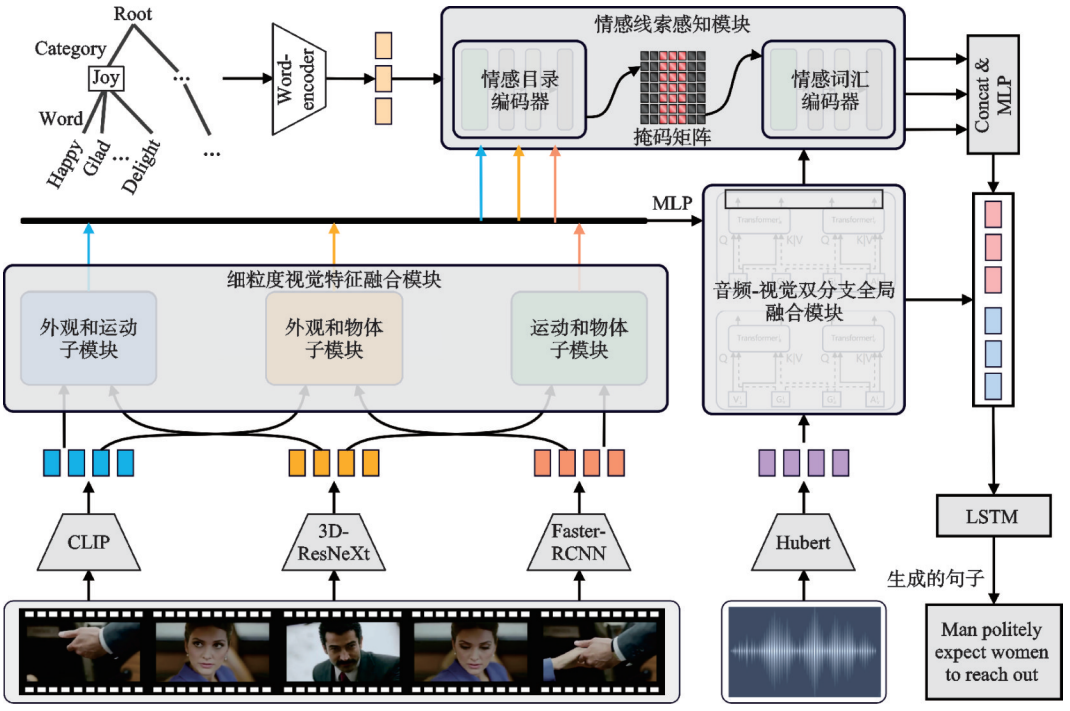

图1 基于细粒度视觉与音视双分支融合框架示意图

Fig.1 Diagram of framework based on fine-grained visual and audio-visual dual-branch fusion

1.1 视频和情感词特征提取模块

根据先前工作的设置，对给定视频进行下采样获得一组帧[7]，使用2D-CNN（在ImageNet数据集[23]上预训练的模型）或者CLIP（Contrastive language-image pre-training）提取外观特征和3D-CNN（在Kinetics数据集[24]上预训练的模型）提取运动特征。接下来，使用Transformer编码器对得到的两种特征进一步编码得到视觉特征 $\boldsymbol{V}_L \in \mathbf{R}^{N_v \times d_v}$，其中 $L \in \{\mathrm{a}, \mathrm{m}\}$ 分别表示外观和运动，N_v 表示视频帧数量，d_v 表示视频特征维度。随后用Faster-RCNN[12]捕获每个关键帧及其相邻视频帧组成的片段的目标区域，根据片段中边界框之间的IoU（Intersection over union）对这些区域进行聚类并均值池化，得到物体特征 $\boldsymbol{V}_o \in \mathbf{R}^{N_o \times d_o}$，其中 N_o 表示视频物体数量，d_o 表示物体对象特征维度。情感词特征通过GloVe[25]分别得到情感目录词的词向量特征 $\boldsymbol{D}_c \in \mathbf{R}^{N_c \times d_w}$，其中 N_c 表示情感目录词数量，d_w 表示情感词特征维度，以及情感词汇词的词向量特征 $\boldsymbol{D}_w \in \mathbf{R}^{N_w \times d_w}$，其中 N_w 表示情感词典词数量。对于音频特征首先将视频中原始音频按16 kHz采样率采样并保证单声道，随后调用Hubert模型[26]得到音频特征 $\boldsymbol{A} \in \mathbf{R}^{N_a \times d_a}$，其中 N_a 表示音频时间步数量，d_a 表示音频特征维度。

1.2 细粒度视觉特征融合模块

视频主体及其外观与动作是视频主要传达的内容，而之前的工作却忽视了细粒度视觉特征语义的挖掘。为此本文提出细粒度视觉特征融合模块，通过多维度视觉特征的两两交互融合，有效挖掘出视频中主体及其外观与动作之间的细粒度语义对齐，让模型对视频内容理解更加精准。

（1）外观与运动子模块

首先使用多头交叉注意力（Multi-head cross attention，MHCA），将外观特征作为Query，运动特征作为Key和Value，让外观特征关注时序上运动特征的位置（如视觉中手的外观特征与同一时序上挥手的动作语义对齐），反之亦然，得到关注运动的外观特征和关注外观的运动特征。随后采用类似GRU的门控机制的动态加权融合，得到捕捉时序交叉依赖又过滤冗余信息的融合视觉特征 $\boldsymbol{V}_{am} \in \mathbf{R}^{N_v \times d_v}$，具体计算过程如下

$$\alpha = \sigma(\boldsymbol{W}[\boldsymbol{V}_a; \boldsymbol{V}_m; (\mathrm{MHCA}(\boldsymbol{V}_a, \boldsymbol{V}_m) + \mathrm{MHCA}(\boldsymbol{V}_m, \boldsymbol{V}_a))/2]) \tag{1}$$

$$\boldsymbol{V}_{am} = \alpha \odot \boldsymbol{V}_a + (1 - \alpha) \odot \boldsymbol{V}_m \tag{2}$$

式中 $\boldsymbol{W}$ 为可学习参数。

（2）外观与物体子模块

从语义层面来看视频的外观特征本身已蕴含场景中各类物体的视觉表征信息，这为外观特征与物体特征的深度交互提供了天然的语义基础。为了充分挖掘这种内在关联，实现外观特征对物体信息的精准吸纳，使用多头交叉注意力机制进行特征融合。将外观特征作为Query，物体特征作为Key和Value，让每个时序位置自适应关注相关目标，从而得到融合物体信息的时序外观特征 $\boldsymbol{V}_{ao} \in \mathbf{R}^{N_v \times d_v}$，该特征既保留了原始外观特征的时序连贯性，又融入了物体层面的语义约束，其具体计算过程如下

$$\boldsymbol{V}_{ao} = \mathrm{MHCA}(\boldsymbol{V}_a, \boldsymbol{V}_o) \tag{3}$$

（3）运动与物体子模块

为突破单向特征融合中存在的信息传递偏差问题，即仅由单一模态特征主导信息交互可能导致的语义丢失或偏差，采用双向多头交叉注意力机制，构建运动特征与物体特征之间的双向语义关联，实现更全面的跨模态信息融合。先将运动特征作为Query，得到融合目标信息的运动特征。随后将物体特征作为Query，让每个目标关注其运动相关的时序位置，得到融合运动上下文信息的物体特征。最后将两者拼接并通过MLP进行非线性特征变换与维度整合，最终得到运动语义与物体语义高度对齐的融合特征 $\boldsymbol{V}_{mo} \in \mathbf{R}^{N_v \times d_v}$，其具体计算过程如下

$$\boldsymbol{V}_{\mathrm{mo}}=\mathrm{MLP}([\mathrm{MHCA}(\boldsymbol{V}_{\mathrm{m}},\boldsymbol{V}_{\mathrm{o}});\mathrm{MHCA}(\boldsymbol{V}_{\mathrm{o}},\boldsymbol{V}_{\mathrm{m}})]) \tag{4}$$

(4) 融合

随后将上述子模块得到的3个融合特征拼接起来,通过一个MLP得到捕捉人物和物品等主体、外观细节、动作信息三者之间细微语义关联的统一视觉特征 $\boldsymbol{V}\in\mathbf{R}^{N_{\mathrm{v}}\times d_{\mathrm{v}}}$。

1.3 音频-视觉双分支全局融合模块

音频信息在视频理解当中往往起着正面的积极作用,然而现有相关研究却普遍未对其加以利用。因此本文提出音频-视频双分支全局融合模块,通过实现视觉和音频特征的全局深层融合,充分发挥音频信息对视觉语义对齐和情感理解的辅助作用。

首先通过全局均值池化操作对输入的音频原始特征与视频原始特征进行处理,从而提取出能够表征整体语义的动态全局特征,分别记为音频全局特征 $\boldsymbol{G}_{\mathrm{A}}\in\mathbf{R}^{1\times d_{\mathrm{a}}}$ 和视觉全局特征 $\boldsymbol{G}_{\mathrm{v}}\in\mathbf{R}^{1\times d_{\mathrm{v}}}$。这一步骤的核心在于压缩局部细节干扰,保留模态内最具判别性的全局语义信息,为后续跨模态交互奠定基础。

随后在音频分支中,为实现音频特征与视觉全局信息的深度融合,采用Transformer架构作为跨模态交互的核心模块。具体来说,将经过局部-全局特征协同增强的音频特征(即同时包含局部细节与模态内全局信息的音频特征)作为Transformer层的Query,目的是引导模型聚焦于音频特征中需要与视觉信息交互的关键区域;与此同时,将融合了视觉全局特征的音频特征作为Transformer层的Key和Value,使得在计算注意力权重时,音频特征将参考视觉全局语义进行相似性度量,从而实现视觉全局信息对音频特征的语义引导。经过 k 层Transformer的迭代交互与特征精炼,最终得到充分整合了视频全局语义信息的音频特征 $\boldsymbol{A}_{\mathrm{v}}^{k+1}\in\mathbf{R}^{N_{\mathrm{a}}\times d_{\mathrm{a}}}$,该特征不仅保留了音频自身的模态特性,还蕴含了来自视觉模态的全局语义约束。类似地,在视觉分支中执行对称的跨模态交互操作:将局部-全局协同增强的视觉特征作为Transformer层的Query,而将融合了音频全局特征的视觉特征作为Key和Value。通过 k 层Transformer的注意力机制,视觉特征能够充分吸收音频全局语义信息,最终生成整合了音频全局语义的视觉特征 $\boldsymbol{V}_{\mathrm{A}}^{k+1}\in\mathbf{R}^{N_{\mathrm{v}}\times d_{\mathrm{v}}}$,其具体计算过程如下

$$\boldsymbol{A}_{\mathrm{v}}^{i+1},\boldsymbol{G}_{\mathrm{A}}^{i+1}=\mathrm{Transformer}_{\mathrm{A}}^{i}([\boldsymbol{G}_{\mathrm{A}}^{i};\boldsymbol{A}_{\mathrm{v}}^{i}],[\boldsymbol{G}_{\mathrm{v}}^{i};\boldsymbol{A}_{\mathrm{v}}^{i}])\quad i=1,2,\cdots,k \tag{5}$$

$$\boldsymbol{V}_{\mathrm{A}}^{i+1},\boldsymbol{G}_{\mathrm{v}}^{i+1}=\mathrm{Transformer}_{\mathrm{v}}^{i}([\boldsymbol{G}_{\mathrm{v}}^{i};\boldsymbol{V}_{\mathrm{A}}^{i}],[\boldsymbol{G}_{\mathrm{A}}^{i};\boldsymbol{V}_{\mathrm{A}}^{i}])\quad i=1,2,\cdots,k \tag{6}$$

1.4 情感线索感知模块

为了得到视频情感特征对视频字幕生成任务的引导,本文将音频-视觉双分支全局融合模块得到的融合全局视觉语义的音频特征 $\boldsymbol{A}_{\mathrm{v}}^{k+1}\in\mathbf{R}^{N_{\mathrm{a}}\times d_{\mathrm{a}}}$ 和融合全局音频语义的视觉特征 $\boldsymbol{V}_{\mathrm{A}}^{k+1}\in\mathbf{R}^{N_{\mathrm{v}}\times d_{\mathrm{v}}}$ 作为该模块的输入,并按照文献[22]的方法,生成分层情感特征。以音频融合特征为例,将其和情感目录特征 $\boldsymbol{D}_{\mathrm{c}}\in\mathbf{R}^{N_{\mathrm{c}}\times d_{\mathrm{w}}}$ 通过Transformer编码器(Encoder)得到 $\boldsymbol{E}_{\mathrm{c}}^{\mathrm{A}}\in\mathbf{R}^{N_{\mathrm{a}}\times N_{\mathrm{c}}}$,其计算过程如下

$$\boldsymbol{E}_{\mathrm{c}}^{\mathrm{A}}=\mathrm{Encoder}_{\mathrm{c}}(\boldsymbol{A}_{\mathrm{v}}^{k+1},\boldsymbol{D}_{\mathrm{c}})|_{\boldsymbol{Q}:\boldsymbol{A}_{\mathrm{v}}^{k+1},\{\boldsymbol{K},\boldsymbol{V}\}:\boldsymbol{D}_{\mathrm{c}}} \tag{7}$$

$\boldsymbol{E}_{\mathrm{c}}^{\mathrm{A}}\in\mathbf{R}^{N_{\mathrm{a}}\times N_{\mathrm{c}}}$ 通过均值池化层和线性层能得到情感目录词的分布概率 $\boldsymbol{P}_{\mathrm{c}}^{\mathrm{A}}\in\mathbf{R}^{1\times N_{\mathrm{c}}}$,从中选择概率最高的前 K 个,并根据目录词和词汇词的对应关系得到掩码矩阵 $\boldsymbol{M}\in\mathbf{R}^{N\times N_{\mathrm{c}}}$,再将掩码矩阵、音频特征、情感词汇词特征 $\boldsymbol{D}_{\mathrm{w}}\in\mathbf{R}^{N_{\mathrm{w}}\times d_{\mathrm{w}}}$ 通过编码器得到音频相关情感特征 $\boldsymbol{E}_{\mathrm{w}}^{\mathrm{A}}\in\mathbf{R}^{N_{\mathrm{a}}\times d_{\mathrm{e}}}$,其计算过程如下

$$\boldsymbol{E}_{\mathrm{w}}^{\mathrm{A}}=\mathrm{Encoder}_{\mathrm{c}}(\boldsymbol{A}_{\mathrm{v}}^{k+1},\boldsymbol{D}_{\mathrm{w}},\mathrm{mask}=\boldsymbol{M})|_{\boldsymbol{Q}:\boldsymbol{A}_{\mathrm{v}}^{k+1},\{\boldsymbol{K},\boldsymbol{V}\}:\boldsymbol{D}_{\mathrm{w}}} \tag{8}$$

$\boldsymbol{E}_{\mathrm{w}}^{\mathrm{A}}\in\mathbf{R}^{N_{\mathrm{a}}\times N_{\mathrm{w}}}$ 通过均值池化层和线性层得到情感词汇词的分布概率 $\boldsymbol{P}_{\mathrm{w}}^{\mathrm{A}}\in\mathbf{R}^{1\times N_{\mathrm{w}}}$,视觉特征 $\boldsymbol{V}_{\mathrm{A}}^{k+1}\in\mathbf{R}^{N_{\mathrm{v}}\times d_{\mathrm{v}}}$ 也采用相同方法得到视觉相关情感特征 $\boldsymbol{E}_{\mathrm{w}}^{\mathrm{v}}\in\mathbf{R}^{N_{\mathrm{v}}\times d_{\mathrm{e}}}$。随后将两个情感特征拼接通过一个全连接层得到最终情感特征 $\boldsymbol{E}\in\mathbf{R}^{N_{\mathrm{v}}\times d_{\mathrm{e}}}$。

1.5 情感字幕生成模块

本文沿用了文献[22]的情感字幕生成模块，首先通过GloVe[25]得到时间步 t 生成的文本特征 $\boldsymbol{Y}_t\in\mathbf{R}^{t\times d_w}$，然后将其和视觉特征 $\boldsymbol{V}_\Lambda^{k+1}\in\mathbf{R}^{N_v\times d_v}$ 一起送入到语义对齐模块中，得到视觉相关的文本特征 $\boldsymbol{\mathcal{T}}_t\in\mathbf{R}^{N_v\times d_w}$。随后将视觉特征、音频特征和文本特征拼接为 $\boldsymbol{C}_t\in\mathbf{R}^{N_v\times(d_v+d_a+d_w)}$，并将 $\boldsymbol{C}_t$ 与LSTM[11]上一层隐藏层状态 $\boldsymbol{h}_{t-1}\in\mathbf{R}^{t\times d_h}$（$d_h$表示隐藏层状态特征维度）通过归一化的加性注意力权重加权求和得到视觉、文本、情感上下文聚合的特征 $\boldsymbol{Z}_t$，最后使用LSTM得到下一个单词 $\boldsymbol{y}_{t+1}$ 的概率分布，其计算过程如下

$$\begin{cases}\boldsymbol{h}_{t+1}=\mathrm{LSTM}([\boldsymbol{Z}_t;\boldsymbol{y}_t],\boldsymbol{h}_t)\\ P(\boldsymbol{y}_{t+1})=\boldsymbol{U}_p\boldsymbol{h}_{t+1}+\boldsymbol{b}_p\end{cases}\tag{9}$$

式中 $\boldsymbol{U}_p$ 和 $\boldsymbol{b}_p$ 为可学习参数。

1.6 损失函数

由于情感词在生成文本中的重要性，本文使用了以情绪为中心的交叉熵损失，在情绪词上增加了一个惩罚项，以确保生成的情绪词的正确性，即

$$L_{ce}=\begin{cases}-(1+\beta)\sum_t\ln P(\boldsymbol{y}_t) & \boldsymbol{y}_t\in\boldsymbol{D}_w\\ -\sum_t\ln P(\boldsymbol{y}_t) & \text{其他}\end{cases}\tag{10}$$

除此之外，为了更准确地预测正确的心理类别和情感词，对情感线索感知模块中的目录词分布和情感词分布建立了层次情感分类损失，即

$$L_{cls}^X=-\sum_{x\in\hat{Y}\cap E_c}\ln P_c^X(x)-\sum_{x\in\hat{Y}\cap E_w}\ln P_w^X(x)\tag{11}$$

式中 X 代表细粒度视觉特征融合里的3个子模块以及音频-视觉双分支全局融合模块各自经情感线索感知模块处理后得到的词汇概率分布。

对于外观与运动子模块，为强化正样本与融合特征之间的关联强度，并抑制负样本的干扰，专门设计了对比损失 $\mathcal{L}_{nce}$。该损失函数通过构建正负样本对的判别性约束，强制正样本和融合特征之间的余弦相似度(sim)显著高于负样本，促使模型在特征空间中拉大正负样本与融合特征的距离差异，表达式为

$$L_{nce}=-\ln\frac{\exp(\mathrm{sim}(\boldsymbol{V}_{am},\boldsymbol{V}_a^+)/T)}{\exp(\mathrm{sim}(\boldsymbol{V}_{am},\boldsymbol{V}_a^-)/T)}\tag{12}$$

式中 T 表示温度。

考虑到外观与物体子模块和运动与物体子模块的核心目标是实现融合特征与对应物体特征的语义对齐，对两者均采用了余弦相似度损失进行优化。余弦相似度损失通过最大化融合特征与目标物体特征之间的余弦相似度，能够有效约束两类特征在高维空间中趋向于同一方向，从而强化它们在语义层面的一致性。具体来说，对于外观与物体子模块，该损失促使外观与物体的融合特征尽可能贴近对应的物体特征；而在运动与物体子模块中，它则推动运动与物体的融合特征向目标物体特征收敛，具体公式如下

$$L_{cos}=1-\mathrm{sim}(\boldsymbol{V}_{ao},\boldsymbol{V}_o)+1-\mathrm{sim}(\boldsymbol{V}_{mo},\boldsymbol{V}_o)\tag{13}$$

最终联合损失函数可以表示为

$$L=\lambda_{ce}L_{ce}+\lambda_{cls}L_{cls}^X+\lambda_{nce}L_{nce}+\lambda_{cos}L_{cos}\tag{14}$$

2 实验与分析

2.1 数据集与评价指标

本文在公共视频情感字幕数据集EmVidCap-L[7]上进行了实验。EmVidCap-L是基于情感预测数据集VideoEmotion-8[27],通过编写带有情绪表达的文本句来构建的。该数据集将1 141/382个视频和19 398/6 527个句子分别用于训练/测试。

本文采用视频字幕生成中常用的双语评估替代(Bilingual evaluation understudy, BLEU)[28]、显式排序翻译评估指标(Metric for evaluation of translation with explicit ordering, METEOR)[29]、面向召回的摘要评估指标(Recall-oriented understudy for gisting evaluation, ROUGE)[30]和基于共识的图像描述评估指标(Consensus-based image description evaluation, CIDEr)[31]作为评价指标。根据文献[20]的工作,使用情感词准确度Acc_{sw}和情感句子准确度Acc_c评估生成句子中情感的准确性。除此之外,本文在BLEU和CIDEr对文本和参考文本匹配度衡量的基础上,结合Acc_{sw}和Acc_c,用两个综合指标BFS和CFS对带情感的生成语句进行更综合的评估。

2.2 实验细节

本文利用ResNet101[32]来提取视频外观特征,并使用3DResNext-101[33]来提取视频运动特征,生成字幕的最大长度被设置为15;还构建了一个包含语料库中所有单词的整体词汇表,EmVidCap-L数据集的词汇量是13 980。对于情绪学习,情感特征生成模块中的情感目录词总数$N_c=34$,情感词汇词总数$N_w=179$,所有单词的特征都是通过GloVe[25]提取。视频、文本和情感特征的特征维度$d_v=d_w=d_e=300$,物体和音频特征的特征维度$d_o=d_a=768$,将情感特征编码层数和多头数均设置为2,将音频-视觉双分支全局融合模块层数k也设置为2,基于LSTM[11]的字幕解码器的隐藏层大小设置为512,损失函数的超参数设置为$\lambda_{ce}=0.3$,$\lambda_{cls}=0.3$,$\lambda_{nce}=0.6$,$\lambda_{cos}=1.0$,对于生成句子中情感的惩罚系数β设置为0.2。采用Adam优化器[34],学习率为7e−4,批次大小设置为128,训练轮数设置为30,在测试过程中使用束搜索[35]的方式进行解码,束尺寸设置为5。

2.3 实验结果对比

表1展示了本文模型和其他方法在EmVidCap-L数据集上的性能比较,其中R101表示ResNet101特征提取模型,RN表示3D-ResNeXt101特征提取模型。本文模型在所有指标上都优于基准模型EPAN,这证明了本文方法对性能提升有一定帮助,同时和最新方法DCGN[7]相比在情感指标上也取得了一定的提升。表2展示了本文模型和其他方法在EmVidCap-L数据集中对于CLIP特征的性能比较。结果也表明本文模型对性能有一定提升。

表1 在EmVidCap-L数据集上本文模型和其他方法的性能比较

Table 1 Performance comparison of the proposed model with other methods on EmVidCap-L dataset

特征	方法	Acc_{sw}	Acc_c	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	BFS	CFS
R101+RN	CANet[21]	41.9	39.7	66.9	44.8	29.3	19.3	18.2	37.9	23.3	33.9	26.8
	EPAN[22]	49.7	48.2	66.9	45.0	29.8	19.6	18.3	38.1	24.1	36.3	29.1
	DCGN[7]	51.3	50.6	**69.6**	**48.2**	**33.8**	**22.1**	**19.5**	**42.0**	**28.4**	**38.7**	**33.1**
	Ours	**53.4**	**52.5**	67.6	45.3	30.3	19.7	18.5	38.3	24.6	36.8	30.1

注:最佳结果用粗体显示,次优用下划线显示。

表2 本文模型与其他方法对于CLIP特征的性能对比

Table 2 Performance comparison of the proposed model with other methods for CLIP feature

特征	方法	Acc_{sw}	Acc_c	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	BFS	CFS
CLIP	SA-LSTM[36]	48.6	47.1	71.0	51.1	34.5	22.5	19.6	40.7	30.2	38.9	33.7
	EPAN[21]	64.3	62.1	72.5	52.8	37.2	27.0	20.8	41.7	33.5	44.2	39.4
	Ours	**65.7**	**64.2**	**74.4**	**54.0**	**38.1**	**27.8**	**21.9**	**42.6**	**35.1**	**45.4**	**40.8**

注:最佳结果用粗体显示。

从表1,2可以看出,首先本文方法在数据集上实现了情感指标Acc_{sw}和Acc_c对基线模型的全面提升,在R101+RN特征上比EPAN模型分别提高3.7%和4.3%,在CLIP特征上分别提高1.4%和2.1%。前者在情感上提升更大,得益于3D-ResNeXt-101提供了更丰富的运动表征,在细粒度视觉特征融合模块中融合了更加丰富的运动相关上下文信息,使模型能够更精准地捕捉视频中与情感倾向紧密关联的细微视觉线索,进而优化了情感语义的判别能力,从而使得生成句子中对于情感描述的准确性得到提升。其次,本文方法在语义一致性度量指标上和基线模型相比同样展现出显著的性能增益,特别是在CLIP特征上,对比BLEU-1、METEOR和CIDEr指标比EPAN模型分别提高1.9、1.1和1.6,相比于R101+RN特征提升更大。这一结果表明,CLIP特征凭借其预训练过程中视觉和文本的跨模态语义对齐能力,能够为模型提供更优的语义先验,从而更有效地引导模型生成语义准确性、表达生动性及内容多样性更优的字幕句子文本。此外,在混合度量BFS和CFS上,本文模型相比于EPAN模型在CLIP特征上分别提升了1.2和1.4,这一结果进一步验证了本文方法在情感视频字幕生成任务中的优越性,即模型能够同时精准感知视频的视觉语义信息与情感线索,并以此为约束引导生成更准确、更生动且与视频内容更匹配的描述文本,充分体现了本文方法在多维度语义融合与跨模态生成任务中的有效性。

2.4 消融实验

为深入探究本文方法中核心模块的具体贡献,本文针对细粒度视觉特征融合模块与音频-视觉双分支全局融合模块开展了消融实验,结果如表3所示,其中A表示音频-视觉双分支全局融合模块,B表示细粒度视觉特征融合模块。实验结果表明,音频-视觉双分支全局融合模块的使用对情感相关评价指标产生了更为显著的正向影响,在Acc_{sw}和Acc_c上比只使用细粒度视觉特征融合模块分别多提升0.6%和1.1%。这一现象充分印证了音频模态中蕴含的情感线索(如语调变化、背景音乐情绪倾向等)能够为模型提供额外的情感维度上的信息,可以有效辅助模型更精准地理解视频中的情感倾向,并促进情感特征在跨模态空间中的对齐。

表3 消融实验结果

Table 3 Ablation experiment results

A	B	Acc_{sw}	Acc_c	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	BFS	CFS
×	×	64.3	62.1	72.5	52.8	37.2	27.0	20.8	41.7	33.5	44.2	39.4
√	×	65.4	63.8	73.1	53.0	37.3	26.9	21.1	41.7	33.9	44.6	39.8
×	√	64.8	62.7	74.1	53.6	37.8	27.5	21.8	42.4	34.6	45.1	40.2
√	√	**65.7**	**62.4**	**74.4**	**54.0**	**38.1**	**27.8**	**21.9**	**42.6**	**35.1**	**45.4**	**40.8**

注:最佳结果用粗体显示。

与此同时,细粒度视觉特征融合模块的集成则显著提升了模型在语义度量指标上的表现,在BLEU-1、METEOR和CIDEr指标上比只使用音频-视觉双分支全局融合模块分别多提升1.0、0.7和0.7。这一结果表明该模块在视觉语义建模中的作用,即通过对视频中人物、物体、外观及动作等多维度视觉语义信息进行细粒度的关联与对齐,为模型提供了更丰富、更精准的视觉特征表征。

图2为消融实验可视化结果。综合图2结果来看,表3消融实验结果进一步验证了所提模型架构的有效性,通过细粒度视觉特征融合模块与音频-视觉双分支全局融合模块的协同作用,本文模型能够同时实现对视频视觉语义的深度理解与情感线索的精准感知,从而为生成更高质量的句子(兼具语义准确性与情感一致性)提供了有效的引导。

Fine-grained: the man looked at the dead man in angry.
Audio: The black man looked sadly at the dead man.
All: the white man is very sad because of the death of his family.
GT: The white man in black is sorrowfully looking at the dead.

Where there's death, there will always be death.

(a) Visualization example of 61 003 in EmVidCap-L

Fine-grained: the woman looked at the gift happily.
Audio: the family was surprised to see the gift on the table.
All: the woman was surprised when she saw the gift from her husband.
GT: A woman was surprised by the present her family gave her.

We want to make sure she's ready. Ready, set, go! You have to open that!Get over there with her. That said it's blindfold you, so I said I have my jacket. Hold on, Tristan you're in the way. Go around over there. And I said I have a shirt! Now Ashlyn's in the way.

(b) Visualization example of 70 824 in EmVidCap-L

图2 消融实验可视化结果

Fig.2 Visualization results of ablation experiment

2.5 可视化结果分析

为了从定性角度直观验证本文方法的优越性,本文在EmVidCap-L数据集上与EPAN模型进行了可视化对比实验,结果如图3所示,图中音频内容已通过第三方模型自动转写为英文文本。在图3(a)的例子中,EPAN模型将视频情感错误归类为“愤怒”,而本文方法能够精准识别出“害怕”这一核心情绪,且该结果与音频转写文本所传递的情感倾向高度一致。图3(b)的例子中,EPAN模型误将情绪判定为“悲伤”,而本文方法不仅准确识别出“愤怒”的情绪属性,还成功定位视频主体为“背后的男人”,体现出更优的语义理解能力。上述可视化结果表明所提模型架构的合理性:一方面,细粒度视觉特征融合模块通过对人物、动作、场景等多维度视觉信息的深度关联与整合,为模型提供了更丰富的视觉语义表征,显著提升了视觉内容建模的精准度;另一方面,音频-视觉双分支全局融合模块有效构建了跨模态信息交互通道,将音频模态中蕴含的情感线索与语义信息传递至模型决策过程,为视频内容理解与情感判别提供了关键的跨模态约束。两者的协同作用使模型能够更全面地解析视频中的视觉内容与情感线索,从而在复杂场景下实现更精准的语义理解与情感识别。

Ours: the man is so scared when he saw the insect in the box.
EPAN: a man is angrily trying to walk through a place.
GT: A man was so scared that he didn't dare to step into the place full of lobsters.

Oh, no, no, that's not happening. There's not, oh, for the love of God. Oh, no, Oh.

(a) Visualization example of 40 714 in EmVidCap-L

Ours: the man is very angry with the man who is talking about something.
EPAN: a little boy is crying with great sadness.
GT: The man is furious at his friend who is always talking in front of him.

Hey, what's in the show? Yeah, I'm good when I should do that type. You're watching a lot. It's them or summer. Hey, you need to do the best part.

(b)Visualization example of 11 510 in EmVidCap-L

图 3 可视化结果

Fig.3 Visualization results

3 结束语

本文针对情感视频字幕生成任务中视觉语义精细化解析不足与跨模态信息利用不充分的问题，提出了一种基于细粒度视觉与音视双分支融合的框架。细粒度视觉特征融合模块通过对视觉、物体、动作特征展开多维度交互与深度融合，成功捕捉视频主体及其动态变化间的细微语义关联，为视频内容的理解提供视觉表征信息；音频-视觉双分支全局融合模块搭建起高效的跨模态信息交互桥梁，将融合后的视觉特征与音频特征深度融合，让模型充分学习音频中蕴含的情感线索与语义信息，显著增强了对视频情感与内容的理解能力。最后，在公开基准数据集上的对比实验和消融实验结果均表明本文所提出的细粒度视觉与音视双分支融合的框架的有效性。本文研究不仅为情感视频字幕生成任务提供了新的技术思路，也为跨模态融合与细粒度视觉语义建模在相关领域的应用提供了参考。

参考文献：

[1] XIANG X, ZHANG Y, JIN L, et al. Sub-region localized hashing for fine-grained image retrieval[J]. IEEE Transactions on Image Processing, 2021, 31: 314-326.

[2] LI Z, TANG J, MEI T. Deep collaborative embedding for social image understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2070-2083.

[3] WANG Y, ZHANG W, LIU K, et al. Dynamic emotion-dependent network with relational subgraph interaction for multimodal emotion recognition[J]. IEEE Transactions on Affective Computing, 2025, 16(2): 712-725.

[4] LIU R, ZUO H, LIAN Z, et al. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities[J]. IEEE Transactions on Affective Computing, 2024, 15(4): 1856-1873.

[5] TSAI Y C, TSAI Y C, PAN T Y, et al. EMVGAN: Emotion-aware music-video common representation learning via generative adversarial networks[C]//Proceedings of the 2022 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia. Newark, NJ, USA: ACM, 2022: 13-18.

[6] SONG P, GUO D, YANG X, et al. Emotional video captioning with vision-based emotion interpretation network[J]. IEEE

Transactions on Image Processing, 2024, 33: 1122-1135.

[7] YE C, YE C, CHEN W, et al. Dual-path collaborative generation network for emotional video captioning[C]//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM, 2024: 496-505.

[8] KOJIMA A, TAMURA T, FUKUNAGA K. Natural language description of human activities from video images based on concept hierarchy of actions[J]. International Journal of Computer Vision, 2002, 50(2): 171-184.

[9] KRISHNAMOORTHY N, KRISHNAMOORTHY N, MALKARNENKAR G, et al. Generating natural-language video descriptions using text-mined knowledge[C]//Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. Atlanta, Georgia, USA: AAAI, 2013: 541-547.

[10] VENUGOPALAN S, ROHRBACH M, DONAHUE J, et al. Sequence to sequence: Video to text[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2016: 4534-4542.

[11] GRAVES A. Long short-term memory[M]//Supervised Sequence Labelling with Recurrent Neural Networks. Berlin, Germany: Springer, 2012: 37-45.

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

[13] ZHANG J, PENG Y. Object-aware aggregation with bidirectional temporal graph for video captioning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2020: 8319-8328.

[14] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. (2014-06-18). https://arxiv.org/abs/1406.1078.

[15] 佟国香,李乐阳.基于图神经网络和引导向量的图像字幕生成模型[J].数据采集与处理,2023,38(1): 209-219.
TONG Guoxiang, LI Leyang. Image caption generation model based on graph neural network and guidance vector[J]. Journal of Data Acquisition and Processing,2023, 38(1): 209-219.

[16] KHAN M U G, ZHANG L, GOTOH Y. Human focused video description[C]//Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona, Spain: IEEE, 2012: 1480-1487.

[17] SUN X, WANG J, REN F, et al. Dynamic emotional transition sampling and emotional guidance of individuals based on conversation[J]. IEEE Transactions on Computational Social Systems, 2024, 11(1): 1192-1204.

[18] YANG J, GAO X, LI L, et al. SOLVER: Scene-object interrelated visual emotion reasoning network[J]. IEEE Transactions on Image Processing, 2021, 30: 8686-8701.

[19] MITTAL T, MATHUR P, BERA A, et al. Affect2MM: Affective analysis of multimedia content using emotion causality [C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 5657-5667.

[20] WANG H, TANG P, LI Q, et al. Emotion expression with fact transfer for video description[J]. IEEE Transactions on Multimedia, 2022, 24: 715-727.

[21] SONG P, GUO D, CHENG J, et al. Contextual attention network for emotional video captioning[J]. IEEE Transactions on Multimedia, 2023, 25: 1858-1867.

[22] SONG P, SONG P, GUO D, et al. Emotion-prior awareness network for emotional video captioning[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, ON, Canada: ACM, 2023: 589-600.

[23] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248-255.

[24] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 4724-4733.

[25] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014: 1532-1543.

[26] HSU W N, BOLTE B, TSAI Y H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[27] JIANG Y G, XU B, XUE X. Predicting emotions in user-generated videos[C]//Proceedings of the AAAI Conference on

Artificial Intelligence. Washington, DC, USA: AAAI, 2014.

[28] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: ACL, 2001: 311.

[29] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. [S.l.]: ACL, 2005: 65-72.

[30] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. [S.l.]: ACL, 2004

[31] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: Consensus-based image description evaluation[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 4566-4575.

[32] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

[33] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6546-6555.

[34] KINGMA D P, BA J. ADAM: A method for stochastic optimization[EB/OL].(2014-12-03). https://arxiv.org/abs/1412.6980.

[35] WANG P, NG H T. A beam-search decoder for normalization of social media text with application to machine translation[C]//Proceedings of the 2013 Conference of the Association for Computational Linguistics: Human Language Technologies. [S.l.]: ACL, 2013: 471-481.

[36] WANG B, MA L, ZHANG W, et al. Reconstruction network for video captioning[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7622-7631.

作者简介:

龚禹轩(2001-),男,硕士研究生,研究方向:多媒体理解,E-mail: 231050029@hdu.edu.cn。

韩婷婷(1990-),**通信作者**,女,博士,副教授,研究方向:多媒体与多模态分析,E-mail: ttinghan@hdu.edu.cn。

(编辑:张黄群)