基于多模态融合的无监督视频摘要算法研究

潘　涛[1],陈　虎[1,2],黄　菊[3],吴长柯[2],邓　彪[3],吴志红[1,2]
(1. 四川大学 计算机学院,四川 成都 610065;
2. 四川大学 视觉合成图形图像技术重点学科实验室,四川 成都 610065;
3. 中国东方电气集团有限公司,四川 成都 610036)

摘　要:视频摘要生成算法通过选择视频内容中信息最丰富的部分来构建形成简洁而完整的概要,有利于快速了解视频内容并压缩存储空间。针对现有视频摘要方法存在的视频多模态信息利用不充分、特征表达能力弱等难题,该文提出了一种基于多模态融合及多尺度时序信息的无监督视频摘要生成算法。首先,基于视频图像、音频、文本特征,提出了一种两阶段特征融合模块,充分保留了模态间的非冗余信息,提升单帧特征表示能力;其次,采用自注意力和特征金字塔网络对融合特征进行全局及局部的依赖建模;最后,根据多尺度的上下文信息选择关键帧最终构成高质量的视频摘要。实验结果表明,与其他无监督视频摘要算法相比,该算法在 SumMe 数据集规范设置及增强设置中 F-Score 分别提升了 0.5 百分点和 1.4 百分点,在 TVSum 数据集上达到最佳 F-Score。

关键词:无监督视频摘要;多模态融合;自注意力网络;特征金字塔网络;特征编码
中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1673-629X(2024)11-0029-07
doi:10.20165/j.cnki.ISSN1673-629X.2024.0239

Research on Unsupervised Video Summarization Algorithm Based on Multimodal Fusion

PAN Tao[1],CHEN Hu[1,2],HUANG Ju[3],WU Chang-ke[2],DENG Biao[3],WU Zhi-hong[1,2]
(1. School of Computer Science,Sichuan University,Chengdu 610065,China;
2. State Key Laboratory of Fundamental Science on Synthetic Vision,Sichuan University,Chengdu 610065,China;
3. Dongfang Electric Corporation,Chengdu 610036,China)

Abstract:The aim of video summarization is to construct concise and comprehensive summaries by selecting the most important content of the video,facilitating a rapid understanding of the video and conserving storage space. Existing methods face challenges including inadequate utilization of multimodal information and weak feature expression capabilities. We propose an unsupervised video summarization algorithm based on multimodal fusion and multiscale temporal information. Firstly,we introduce a two-stage feature fusion module based on video images,audio,and text features,preserving non-redundant information between modalities and enhancing the representation capability of features. Then,we employ self-attention and feature pyramid networks to obtain global and local temporal dependencies,select keyframes based on multi-scale contextual information,and form a high-quality video summary. The experimental results demonstrate that compared to other unsupervised video summarization algorithms,the proposed algorithm has achieved an improvement of 0.5 percentage points and 1.4 percentage points in F-Score on the SumMe dataset under canonical and augmented settings,respectively. Moreover,it has achieved the highest F-Score on the TVSum dataset.

Key words:unsupervised video summarization;multimodal fusion;self-attention network;feature pyramid network;feature encoder

0　引　言

随着流媒体技术和存储技术的发展,互联网上的视频数量急剧增长,在海量视频中通过观看完整视频来检索感兴趣的内容非常耗时低效。视频摘要技术[1]能自动提取视频中的关键信息,生成简短摘要,快速了解视频内容,从而提高效率并节省时间和资源。该技术广泛应用于不同领域,包括电影预告片制作、体育视频剪辑、监控视频浓缩等。

收稿日期:2024-02-26　　**修回日期**:2024-06-27
基金项目:国家自然科学基金重点项目(U20A20162);四川省科技计划项目(2022JDJQ0045)
作者简介:潘　涛(1995-),男,硕士研究生,研究方向为多模态视频摘要;通信作者:黄　菊(1982-),女,高级工程师,研究方向为多模态信息处理、智能制造。

随着人工智能技术的进步,视频摘要领域已经提出了多种基于深度学习的自动摘要生成算法[2-5]。这些方法都遵从如下同一范式:(1)借助图像编码器获得特征向量来表示视频帧的视觉内容;(2)将特征向量输入摘要预测网络,得到帧级别重要性分数;(3)通过 Knapsack 背包算法选择关键帧转换为视频摘要。其中如何构建具有代表性的特征表示以及公平有效的评估帧的重要性分数是核心问题。

然而,现有方法大多基于视频帧的图像信息,仅利用图像单一模态构建特征表示,忽略了音频、文字等多模态信息,存在视频信息利用不充分的问题,导致生成摘要的质量较差。近年来,基于多模态特征融合在多种下游任务上的成功应用证明了多模态信息对于视觉信息理解具有重要意义,充分利用视频的多模态信息有利于视频摘要的高质量生成。

此外,已有算法基本属于基于数据驱动的监督预测范畴,对数据的依赖性非常强。监督学习方法虽然使用人工标记的重要性分数作为训练深度神经网络的监督信号,表现出显著性能,但需要获得大量的帧级别重要性评分,而获得这种评分是非常困难、费时和昂贵的。进一步,由于标注人员对帧的重要性评分存在主观性,可能导致标签不一致,从而影响摘要生成质量。为了减轻数据依赖性,研究人员更关注无监督方法,这种方法不需要费力费时的数据标注,更符合视频摘要任务的实际应用场景。Mahasseni[6] 率先提出了无监督深度学习方法,将两个 LSTM 网络分别用作变分自编码生成器和鉴别器,并通过生成对抗学习最小化重建的摘要视频和原始视频分布之间的距离。Li[7] 利用注意力机制来弥补 LSTM 对于长程依赖的不足,从视频的全局角度建立帧的时间依赖关系,生成更健壮的特征表示,提供全局系列帧信息,但其忽略了局部镜头帧之间的相互作用。从人工角度生成视频摘要,首先需要浏览整个视频,从全局的角度考虑每个片段的重要性,然后选择能够代表每个镜头的关键帧,并从这些关键帧中生成一个高质量的摘要。模拟人工的这种行为特点,Pang[8] 利用对比损失,定义了局部不相性和全局一致性属性,以直接量化每个候选帧的重要性分数,但这些属性涉及视频帧序列全局和局部时间依赖性,对于建模视频帧的时间依赖性提出了更高的要求。

基于上述视频摘要研究中存在的问题,该文提出了一种基于多模态融合及多尺度时序信息的无监督视频摘要生成算法。具体而言,首先利用图像、音频、文本编码器分别提取不同模态特征。其次,提出了一个两阶段融合模块,以充分促进模态间相互融合,利用不同模态中的互补和非冗余信息,增加单帧特征的信息丰富度,提升特征的表示能力。最后,受到人工视频摘要分别从全局和局部角度评估帧的重要性的启发,该文提出了并行自注意力网络和特征金字塔网络,用以构建全局和局部时间依赖关系,以确保候选关键帧不仅与视频的中心主题全局一致,而且与相邻区域中的其他帧局部不相似。

1 相关工作

1.1 视频摘要技术的发展

最早,视频摘要主要使用手工制作的特征进行处理,然后通过聚类来优化关键帧的选择[9]。随着帧级标注数据集 SumMe、TVSum 的出现,研究人员将自动摘要任务视为一项序列预测任务,以回归器评估帧的重要性,并将其与人工标注进行比较。有监督学习方法中,Zhang[10] 利用 LSTM 捕获连续帧间的时间依赖性,并使用多层感知器来预测帧的重要性分数。为了克服 RNNs 在远程依赖方面的不足和巨大的计算开销,VASNet[2] 使用注意力机制代替 LSTM。张喻恩[11] 提出混合注意力机制对于视频帧通道和空间维度进行深度相互依赖性建模。基于数据驱动的有监督的方法取得了显著成果,然而,获取大量的帧级人工标注是困难的,因此无监督的方法引起了研究者的关注。Mahasseni 等[6] 提出了基于无监督的视频摘要方法 SUM-GAN,该算法将两个 LSTM 网络分别用作变分自编码生成器和鉴别器,并通过对抗生成学习来最小化重建的摘要视频和原始视频分布之间的距离。后续基于该方法做了许多改进,SUM-GAN-AAE[12] 使用了确定性的注意力自动编码器代替了可变的自动编码器,而 AC-SUM-GAN[13] 将行动者-批评家(AC)模型嵌入到网络架构中。然而,对抗生成网络的训练是不稳定的,这使得这些网络难以收敛。Zhou 等[14] 巧妙设计多样性和代表性的奖励函数,使用一个简单的网络结构进行强化学习,克服了对抗生成网络难以训练的问题。孙浩然[15] 提出时间轴字幕文本作为语义奖励,通过时间轴摘要对图像摘要的过滤,获得综合的摘要结果。Pang[8] 通过对比度损失来定义全局一致性和局部不相似性,从而直接量化每个候选帧在最终摘要中的重要性分数。

1.2 特征金字塔网络应用

在卷积运算中,不同层级的特征图具有不同的感受野,浅层特征图的感受野较小,深层特征图的感受野较大。为了平衡处理不同尺寸对象的下游任务,需要生成在所有尺度上都具有强语义的多尺度特征表示。Lin[16] 提出了一种特征金字塔网络,通过采用自下而上和自上而下的水平连接路径,在所有尺度上融合高层语义信息和低层细节信息。这一结构直接应用于目标检测主干网络,如 R-CNN、Fast R-CNN、YOLO 系

列,表现出了显著的检测性能改进,证明了金字塔表示在解决多尺度问题中的重要性。Fu[17]利用双向并行多分支卷积特征金字塔网络构建特征金字塔,增强不同尺度特征层的特征表示,以解决群体无人机航拍图像中的小目标检测问题。Chen[18]提出 Info-FPN 网络以应对遥感图像中目标大范围变化的挑战。这些研究结果证明了特征金字塔网络在多尺度特征融合中的有效性。在视频摘要中,重要性分数的评估需要考虑候选帧的全局一致性和局部差异性,特征金字塔网络能够有效地整合多尺度的局部特征,生成有强语义表示的特征,具有生成高质量摘要的巨大潜力。

2　基于多模态融合及多尺度时序信息的无监督视频摘要生成算法

该文提出的基于多模态融合及多尺度时序信息的无监督视频摘要生成算法框架如图 1 所示,可以分为特征提取、两阶段特征融合模块、并行的全局及局部特征序列建模模块。其中特征编码器分别对图像、音频、文字模态编码,提取不同模态的特征表示。特征融合根据模态特征的特异性,以两阶段的融合构建具有代表性的融合特征。序列建模模块对融合特征进行多尺度的序列建模,结合上下文信息公平全面地预测帧的重要性分数。

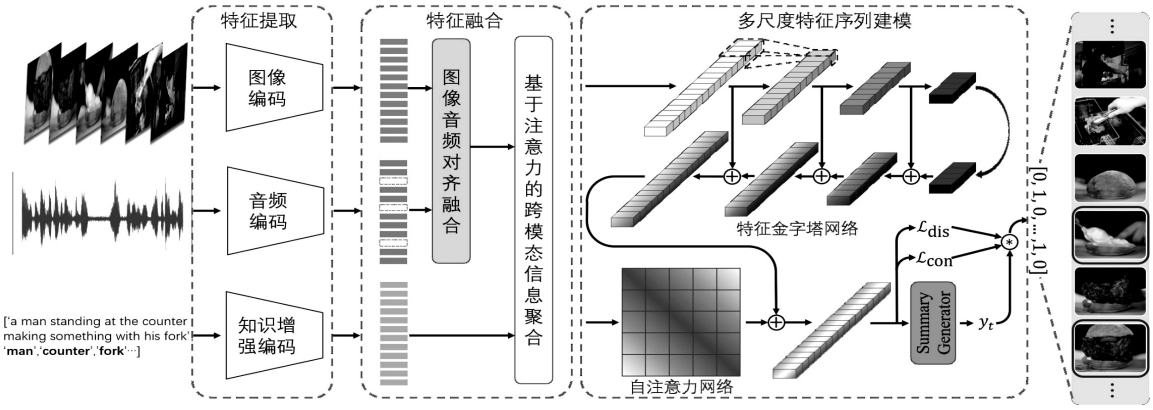

图 1　基于多模态融合及多尺度时序信息的无监督视频摘要生成算法

2.1　多模态特征提取

视频是一种高信息密度载体,一段视频包含图像帧、音频、字幕等多模态信息,通过不同特征编码器得到多种模态的高级特征表示。GoogLeNet 是一个有效的图像特征编码器,结合了不同大小的卷积核,有效地提取图像中的视觉内容。将视频 T 帧图像分别输入预训练的 GoogLeNet 模型,获取图像特征 $F_I=[f_{i,1},\cdots,f_{i,t},\cdots,f_{i,T}]\in R^{T\times M}$。音频信息对于视频内容理解提供额外的信息,节奏紧凑的音乐通常出现在视频高潮部分。将音频分为 960 ms 不重叠的音频帧,利用预训练的 VGGish[19]提取音频帧特征 $F_A=[f_{a,1},\cdots,f_{a,s},\cdots,f_{a,S}]\in R^{S\times N}$,$S$ 表示音频帧的数量 $S=\text{durartion}/0.96$。简单文本信息缺乏情感偏向,基于知识增强的文本编码器 KAENet[20],利用外部知识库 Conceptnet[21]来挖掘潜在情感关系以获得知识增强的文本特征 $F_K=[f_{k,1},\cdots,f_{k,t},\cdots,f_{k,T}]\in R^{T\times L}$。

2.2　多模态特征融合

多模态特征融合旨在融合模态间的非冗余信息,以构建更具表达能力的融合特征。由于不同模态特征呈现非常强的异源异质异构特性,一阶段融合方式将三个模态特征逐项相加或特征向量连接无法有效提取互补信息。该文提出了一种两阶段的多模态特征融合模块,通过探究三种模态特征的特异性及相关性,针对模态特征的不同来源特性采用与之相应的方法进行两阶段的特征融合。

在第一阶段,融合图像和音频特征。在音频特征提取过程中,为保证音频语义的完整,将其划分为 960 ms 不重叠的音频帧,因此一个音频段特征与多帧图像特征相对应。该文采用最邻近插值将音频特征与图像特征对齐。图像特征和音频特征均来源于 CNNs 架构网络,将对齐后的音频特征投影到图像特征的嵌入空间中,以进行特征融合,其过程表示为:

$$F_{IA}=F_I+(\text{interp}(F_A)\cdot W_A+b_A) \tag{1}$$

其中,$\text{interp}(\cdot)$ 表示最邻近插值,W_A,b_A 分别为可学习权重和偏差。

由于第一阶段融合特征 F_{IA} 与知识增强文本特征 F_K 两种模态语义特异性较大,使用简单线性投射不能充分捕捉两个模态之间的互补信息。该文采用基于注意力机制的跨模态信息融合,通过注意力权重对不同模态特征进行加权融合,多模态融合特征中 F_F 自动关注于特征表达更有效的部分。具体来说,将 F_{IA} 作为查询,F_K 作为键进行互相关运算,然后根据相关权重得到最终的融合特征。表达式为:

$$F_F=\Psi(F_{IA}W_Q,F_KW_K,F_KW_V)W_O \tag{2}$$

其中,Ψ 表示注意力跨模态融合,W_Q,W_K,W_V,W_O 分别为可学习权重。

2.3 多尺度特征序列建模

融合特征在单帧的信息表示上具有丰富性,但其缺乏对上下文信息的理解,关键帧的选择是一个序列回归问题,需建立序列帧之间的时间依赖关系,结合上下文信息来评估帧的重要性。现有方法多基于 LSTM、自注意力网络对整个序列特征进行依赖建模,但 LSTM 对长序列依赖性建模具有局限性,自注意力网络相比 LSTM 能有效建立长程依赖,但在长序列上存在权重分散的问题,缺乏对局部镜头内信息的关注。为了解决这个问题,该文设计了一个并行多尺度序列建模来捕获全局及局部范围内的时间依赖关系。总的来说,该模块由两个并行结构组成:用于捕获全局长期依赖关系的自注意力网络和用于捕获局部多尺度依赖关系的特征金字塔网络。

2.3.1 全局依赖建模

自注意力网络能够实现长时间跨度的上下文信息理解,为此该文将其对整个视频帧序列依赖建模,构建全局的时间跨度依赖。简而言之,将多模态融合特征 F_F 分别投影为查询、键和值,然后使用缩放点积进行权重计算,以获得全局上下文输出:

$$F_G = \Lambda(F_F W_Q', F_F W_K', F_F W_V') W_O' \tag{3}$$

其中,Λ 表示多头自注意力网络,W_Q', W_K', W_V', W_O' 分别为可学习权重。

2.3.2 局部依赖建模

自注意力网络在长序列上建模会出现权重分散的问题,局部范围帧差异无法得到有效关注。该文采用特征金字塔网络在时间维度上进行视频帧序列的多尺度局部特征依赖建模。特征金字塔网络可以分为自底向上和自顶向下两个路径。首先自底向上在时间维度上对融合特征 F_F 进行多层一维卷积,以得到多个不同时间跨度的局部特征 $\{F_{P,j}\}_{j=1}^{4}$。其次在自顶向下的路径中,先上采样顶层特征与底层特征对齐,再将底层特征通过卷积变换后与对齐后的顶层特征通过权重进行融合,这样获得相邻两层不同时间跨度的局部融合特征。重复上述操作以获得与输入特征具有相同维度的输出特征 F_L,包含多尺度局部上下文信息。上述过程表示为:

$$F_{P,j} = \mathrm{Conv1}(F_{P,j-1}) \tag{4}$$

$$F_{P,i-1} = \mathrm{Conv2}(F_{P,i-1}) \oplus U(F_{P,i}) \tag{5}$$

其中,$\mathrm{Conv}(\cdot)$ 表示不同设置的一维卷积,$\oplus$ 为按照权重相加,$U(\cdot)$ 为上采样。

最后,将上述两个并行网络输出 F_G 和 F_L 通过加权相加以构建包含全局和局部上下文信息的时间序列依赖特征 $F_{GL} = F_G \oplus F_L$。将包含全局及局部信息的特征通过两层全连接的摘要预测器将特征投影到关键帧的概率空间,获得帧级别的重要性分数 $\{y_t\}_{t=1}^{T}$,其中 T 为视频帧数,$y_t \in (0,1)$。

2.4 损失函数

该文提出一种联合损失函数,结合局部不相似性和全局一致性[8]来量化每个帧的重要性,并通过长度正则化损失来防止模型选择大量的关键帧,其公式为:

$$\mathcal{L} = \beta \bar{\mathcal{L}}_{\mathrm{dis}}(F_{GL}) + \lambda \bar{\mathcal{L}}_{\mathrm{con}}(F_{GL}) + \mathcal{L}_{\mathrm{reg}}(y_t) \tag{6}$$

其中,$\mathcal{L}_{\mathrm{dis}}, \mathcal{L}_{\mathrm{con}}, \mathcal{L}_{\mathrm{reg}}$ 分别表示局部不相似性、全局一致性和长度正则化,$\bar{\mathcal{L}}$ 表示计算平均值,β, λ 为联合权重参数。

2.4.1 局部不相似性

经过序列依赖建模的特征 $F_{GL} = \{f_{gl,i}\}_{i=1}^{T}$ 能够很好地表示帧的语义信息,使用余弦相似度来检索每帧的最相似的前 $\lfloor \alpha T \rfloor$ 相似域集合 Θ,其中 α 为超参数。对于每一帧,计算其在相似域 Θ 中的局部不相似性。

$$\mathcal{L}_{\mathrm{dis}}(f_{gl,t}) = \frac{1}{|\Theta|} \sum_{f_{gl} \in \Theta} \| f_{gl,t} - f_{gl} \|_2^2 \tag{7}$$

$\mathcal{L}_{\mathrm{dis}}(f_{gl,t})$ 值越大,表示该帧与其相似域中的其他帧具有不同的语义信息,是关键帧的理想候选帧。将 $\mathcal{L}_{\mathrm{dis}}(f_{gl,t})$ 变换到 (0,1) 区间可直接用于评估 $f_{gl,t}$ 的重要性分数。

2.4.2 全局一致性

全局一致性通过帧 $f_{gl,t}$ 与视频中其他帧之间的平均成对高斯势来计算。

$$\mathcal{L}_{\mathrm{con}}(f_{gl,t}) = \log\left(\frac{1}{T-1} \sum_{\substack{f_{gl,i} \neq f_{gl}, \\ f_{gl} \in F_{GL}}} \mathrm{e}^{-2\|f_{gl,i} - f_{gl}\|_2^2}\right) \tag{8}$$

$\mathcal{L}_{\mathrm{con}}(f_{gl,t})$ 值越大,表示该帧与整个视频的中心主题一致,也是关键帧的理想候选帧。

2.4.3 长度正则化损失

为了惩罚网络模型选择大量帧来产生摘要,该文使用长度正则化损失进行约束。

$$\mathcal{L}_{\mathrm{reg}}(y_t) = \left\| \frac{1}{T} \sum_{t=1}^{T} y_t - \sigma \right\| \tag{9}$$

其中,σ 为超参数。

2.4.4 重要性评分

由于局部不相似性和全局一致性可以直接量化帧的重要性分数,将两个指标变换到 (0,1) 区间,与预测头产生的概率分数共同组成最终的重要性分数。

$$p_t = \mathrm{Gause}[\mathcal{L}_{\mathrm{dis}}(f_{gl,t}) \mathcal{L}_{\mathrm{con}}(f_{gl,t}) y_t + \epsilon] \tag{10}$$

其中,$\mathrm{Gause}(\cdot)$ 为高斯平滑处理,ϵ 避免重要性分数为零。

3 实验与结果分析

3.1 实验设置

3.1.1 数据集

为了将文中方法与之前的工作进行公平比较,在

SumMe[22], TVSum[23], OVP[24] 和 YouTube[24] 四个数据集上进行实验。SumMe 数据集包含 25 个时长 1 ~ 6 分钟的视频,每个视频由 15 到 18 个用户以关键片段的形式标注。TVSum 包含 50 个时长为 1 ~ 11 分钟的视频,每个视频由 20 个用户以帧级重要性分数的形式进行标注。另外使用 OPV、YouTube 作为增强数据训练集,前者包括 50 个视频,后者包括 39 个视频,由 5 个用户生成的帧级别重要性标注。

做如下设置来验证文中方法的性能:①规范设置(C):训练数据和测试数据来自同一个数据集,按照 8∶2 的比例划分。②增强设置(A):用其他三个数据集来增强训练集(以 SumMe 为例,80% SumMe + TVSsum+YouTube+OVP 作为训练集,20% SumMe 作为测试集)。对于两种实验设置,使用五折交叉验证来评估模型的性能,报告的结果是五折交叉验证的平均值。

3.1.2　评价指标

F-Score 目前被大多数视频摘要方法用于生成摘要质量评估。它通过计算模型生成摘要(G)和用户定义的摘要(U)之间的时间重叠来评估两个摘要之间的相似性。其计算公式如下:

$$P = \frac{|G \cap U|}{|G|} \tag{11}$$

$$R = \frac{|G \cap U|}{|U|} \tag{12}$$

$$\text{F-Score} = \frac{2 * P * R}{P + R} * 100\% \tag{13}$$

其中, P,R 分别表示准确率和召回率, $\cap$ 表示重叠时间, $|\cdot|$ 表示持续时间。

3.1.3　实验环境

在 ubuntu18.04 环境下进行实验,Pytorch 版本为 1.7.1,使用 Nvidia RTX 2080Ti GPU 进行计算加速。采用的 Adam 优化器,学习率和权重衰减分别为 5×10^{-5} 和 10^{-5} ,训练周期为 400 次迭代。遵从之前的工作[2],生成的重要性分数通过 Knapsack 算法选择关键帧生成摘要,生成的摘要的长度不超过原始视频长度的 15%。

3.2　实验结果与对比

3.2.1　对比实验

为了证明文中算法的有效性,选择了多种具有代表性的无监督视频摘要算法与文中算法进行比较。主要可以分为几类:①SUM-GAN[6]、ACGAN[25]、SUM-GAN-AAE[12]、CAAN[26] 基于生成对抗网络进行摘要生成;②DR-DSN[14]、3DST-UNet[27] 基于代表性和多样性奖励函数进行强化学习;③SumGraph[28] 通过递归图构建帧节点间的关联性,CSNet+GL+RPE[29]、SUM-GDA[7] 使用注意力对帧进行时间序列依赖建模;④MCSF[30] 基于多源特征提取网络提供更多视觉信息,曾凡锋[31] 基于深浅层不同视觉特征融合来丰富特征信息;⑤CTVSUM[8] 通过对比学习损失直接量化帧的重要性。

表 1 列出了文中方法在 SumMe 和 TVSum 两个数据集上与上述方法的对比结果。文中算法在 SumMe 数据集上的两种设置中均获得了最好的 F-Score,相较于最好的方法分别提升了 0.5 百分点和 1.4 百分点。在 TVSum 数据集的规范设置和增强设置中分别取得了 59.6% 和 61.2% 的 F-Score,均与最好的方法相当,展现出文中方法的稳定性能。实验结果表明,在两个数据集上文中算法能够生成更高质量的视频摘要。

表 1　与其他无监督视频摘要算法对比结果(F-Score)　%

方法	SumMe		TVSum	
	C	A	C	A
SUM-GAN[6]	39.1	43.4	51.7	59.5
ACGAN[25]	46.0	47.0	58.5	58.9
SUM-GAN-AAE[12]	48.9	–	58.3	–
CAAN[26]	50.8	50.9	59.6	59.8
DR-DSN[14]	41.4	42.8	57.6	58.4
3DST-UNet[27]	44.6	49.5	58.3	58.4
SumGraph[28]	49.8	52.1	59.3	61.2
CSNet+GL+RPE[29]	50.2	–	59.1	–
SUM-GDA[7]	50.0	50.2	59.6	60.5
MCSF[30]	47.9	–	59.1	–
Zeng's Method[31]	48.6	–	59.4	–
CTVSUM[10]	46.8	45.5	59.5	59.9
文中算法	51.3	53.5	59.6	61.2

3.2.2　消融实验

如表 2 所示,进行了多项消融实验。序号 1-3 分别对单一的图像、音频、知识增强三个模态特征直接进行重要性分数预测,可以看出每个模态都保留一定生成摘要的信息,其中图像模态的性能表现更佳,这是因为图像特征间更具差异性,音频和知识增强特征在相邻帧之间的差距并不大。对比序号 1 和 4-5,在融合了音频特征、知识增强特征后,在 SumMe 和 TVSum 两个数据集上性能都有所提升,特别是 SumMe 规范设置中分别提升了 1.4 百分点及 4.1 百分点,表明了文中方法是有效的,融合音频及知识增强特征对视频摘要生成提供了额外的补充信息。对比序号 5 和 6,只增加特征金字塔局部序列建模网络,生成摘要的质量有所下降,这是因为特征金字塔网络过度关注局部

区域内的依赖关系,生成的摘要与整个视频中心主题存在偏差。对比序号 5 和序号 7 表明自注意力网络对全局序列建模,相较于完全缺乏上下文信息的融合特征直接预测性能有所提升。序号 8 结合全局及局部的依赖关系,文中模型达到了最好的效果,证明了文中算法的有效性。

表 2 消融实验结果(F-Score) %

序号	图像	音频	知识增强	金字塔网络	注意力网络	SumMe		TVSum	
						C	A	C	A
1	√					46.3	52.1	58.8	60.7
2		√				46.3	48.8	58.0	58.9
3			√			43.9	43.8	59.4	60.4
4	√	√				47.7	50.0	59.2	60.8
5	√	√	√			50.4	52.0	59.0	60.9
6	√	√	√	√		48.2	51.6	58.3	60.4
7	√	√	√		√	51.3	52.2	59.1	60.7
8	√	√	√	√	√	51.3	53.5	59.6	61.2

3.2.3 可视化分析

为了更加充分地验证文中算法的有效性,进行了生成摘要的可视化分析。图 2 展示了文中算法在 TVSum 数据集 Video-43 上所选择关键帧与人工标注的重要性分数的可视化分析。图中的浅色柱状图表示人工标注的帧重要性分数,而深色柱状图表示文中算法选择的关键帧,可以看出文中算法所选择的关键帧与人工标注的重要部分相匹配。从视频整体内容来看,原始视频主要讲述自行车骑行和维修保养,而生成的摘要准确地捕捉了这一主题内容,所选择的关键帧都包括了自行车这一目标,充分证明了文中算法在生成高质量摘要方面具有出色的性能。

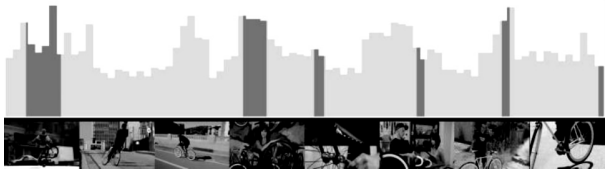

图 2 TVSum 数据集 Video43 生成摘要分布结果

4 结束语

该文提出了一种基于多模态融合及多尺度时序信息的无监督视频摘要生成算法。为了应对图像、音频和知识增强不同模态特征的特异性,设计了一个分阶段的特征融合模块,能够有效提取不同模态间的非冗余互补信息,生成更具代表性的单帧特征表示。同时,在重要性分数预测方面,采用了自注意力网络和特征金字塔网络,对帧序列进行全局和局部的时间依赖建模,通过上下文信息有效量化视频帧序列间的局部不相似性和全局一致性,选择关键帧生成高质量的视频摘要。实验结果表明,该算法在两个基准数据集上均优于现有的其他无监督方法,能够生成更高质量的视频摘要。此外,经过详尽的消融实验和可视化分析,提出的各个模块的有效性均得到了证明。

参考文献:

[1] APOSTOLIDIS E,ADAMANTIDOU E,METSAI A I,et al. Video summarization using deep neural networks:a survey [J]. Proceedings of the IEEE,2021,109(11):1838-1863.

[2] FAJTL J,SOKEH H S,ARGYRIOU V,et al. Summarizing videos with attention[C]//Proceedings of the Asian conference on computer vision. Perth:Springer,2019:39-54.

[3] LI Z,YANG L. Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward[C]//Proceedings of the IEEE winter conference on applications of computer vision. Waikoloa:IEEE,2021:3238-3246.

[4] ZHAO B,LI H,LU X,et al. Reconstructive sequence-graph network for video summarization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,44(5):2793-2801.

[5] XU W,JIANG H,LIANG X,et al. Learning subjective time-series data via utopia label distribution approximation[J]. arXiv:2307.07682,2023.

[6] MAHASSENI B,LAM M,TODOROVIC S. Unsupervised video summarization with adversarial LSTM networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu:IEEE,2017:2982-2991.

[7] LI P,YE Q,ZHANG L,et al. Exploring global diverse attention via pairwise temporal relation for video summarization [J]. Pattern Recognition,2021,111:107677.

[8] PANG Z,NAKASHIMA Y,OTANI M,et al. Contrastive losses are natural criteria for unsupervised video summarization[C]//Proceedings of the IEEE winter conference on applications of computer vision. Waikoloa:IEEE,2023:2010-2019.

[9] HADI Y,ESSANNOUNI F,THAMI R O H. Video summari-

zation by k-medoid clustering[C]//Proceedings of the 2006 ACM symposium on applied computing. Dijon:ACM,2006:1400-1401.

[10] ZHANG K,CHAO W L,SHA F,et al. Video summarization with long short-term memory[C]//Proceedings of the European conference on computer vision. Amsterdam:Springer,2016:766-782.

[11] 张喻恩,李泽平. 基于多尺度混合注意力机制的视频摘要算法[J]. 计算机工程与设计,2023,44(11):3305-3311.

[12] APOSTOLIDIS E,ADAMANTIDOU E,METSAI A I,et al. Unsupervised video summarization via attention-driven adversarial learning[C]//Proceedings of the international conference on multimedia modeling. Daejeon: Springer, 2020: 492-504.

[13] APOSTOLIDIS E,ADAMANTIDOU E,METSAI A I,et al. AC-SUM-GAN:connecting actor-critic and generative adversarial networks for unsupervised video summarization [J]. IEEE Transactions on Circuits and Systems for Video Technology,2020,31(8):3278-3292.

[14] ZHOU K,QIAO Y,XIANG T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward[C]//Proceedings of the AAAI conference on artificial intelligence. New Orleans:AAAI,2018.

[15] 孙浩然. 基于字幕语义的无监督视频摘要方法研究及应用[D]. 苏州:江苏大学,2023.

[16] LIN T Y,DOLLÁR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu:IEEE,2017:936-944.

[17] FU L,GU W,LI W,et al. Bidirectional parallel multi-branch convolution feature pyramid network for target detection in aerial images of swarm UAVs[J]. Defence Technology, 2021,17(4):1531-1541.

[18] CHEN S,ZHAO J,ZHOU Y,et al. Info-FPN:an informative feature pyramid network for object detection in remote sensing images[J]. Expert Systems with Applications, 2023, 214:119132.

[19] HERSHEY S,CHAUDHURI S,ELLIS D P W,et al. CNN architectures for large-scale audio classification[C]//Proceedings of the IEEE international conference on acoustics, speech and signal processing. New Orleans:IEEE,2017:131-135.

[20] XIE J,CHEN X,LU S P,et al. A knowledge augmented and multimodal-based framework for video summarization [C]//Proceedings of the ACM international conference on multimedia. Lisboa:ACM,2022:740-749.

[21] SPEER R,CHIN J,HAVASI C. Conceptnet 5.5:an open multilingual graph of general knowledge[C]//Proceedings of the AAAI conference on artificial intelligence. San Francisco:AAAI,2017.

[22] GYGLI M,GRABNER H,RIEMENSCHNEIDER H,et al. Creating summaries from user videos[C]//Proceedings of the European conference on computer vision. Zurich:Springer,2014:505-520.

[23] SONG Y,VALLMITJANA J,STENT A,et al. Tvsum:summarizing web videos using titles[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Boston:IEEE,2015:5179-5187.

[24] DE AVILA S E F,LOPES A P B,DA LUZ JR A,et al. VSUMM:a mechanism designed to produce static video summaries and a novel evaluation method[J]. Pattern Recognition Letters,2011,32(1):56-68.

[25] HE X,HUA Y,SONG T,et al. Unsupervised video summarization with attentive conditional generative adversarial networks[C]//Proceedings of the ACM international conference on multimedia. Nice:ACM,2019:2296-2304.

[26] LIANG G,LV Y,LI S,et al. Video summarization with a convolutional attentive adversarial network[J]. Pattern Recognition,2022,131:108840.

[27] LIU T,MENG Q,HUANG J J,et al. Video summarization through reinforcement learning with a 3D spatio-temporal u-net[J]. IEEE Transactions on Image Processing,2022,31: 1573-1586.

[28] PARK J,LEE J,KIM I J,et al. Sumgraph:video summarization via recursive graph modeling[C]//Proceedings of the European conference on computer vision. Glasgow:Springer, 2020:647-663.

[29] JUNG Y,CHO D,WOO S,et al. Global-and-local relative position embedding for unsupervised video summarization [C]//Proceedings of the European conference on computer vision. Glasgow:Springer,2020:167-183.

[30] KANAFANI H,GHAURI J A,HAKIMOV S,et al. Unsupervised video summarization via multi-source features[C]//Proceedings of the international conference on multimedia retrieval. New York:ACM,2021:466-470.

[31] 曾凡锋,王春真,李　琛. 基于深浅层特征融合的无监督视频摘要算法研究[J]. 计算机工程与科学,2023,45(9):1602-1610.