Journal of System Simulation

Volume 36 | Issue 5 Article 2

5-15-2024

Dense Video Description Method Based on Multi-modal Fusion in Transformer Network

Xiang Li
School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China, lixiang3278@163.com

Haifeng Sang
School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China, sanghaif@163.com

Dense Video Description Method Based on Multi-modal Fusion in Transformer Network

Abstract

Abstract: In order to solve the problems that most of the current dense video description models use twostage methods, which have low efficiency, ignore audio and semantic information, and have incomplete description results, a multi-modal and semantic information fusion dense video description method was proposed. An adaptive R(2+1)D network was proposed to extract visual features, a semantic detector was designed to generate semantic information, audio features were added to supplement it, a multi-scale deformable attention module was established, and a parallel prediction head was applied to accelerate the convergence rate and improve the accuracy of the model. The experimental results show that the model has good performance on the two benchmark datasets, and the evaluation index BLEU4 reaches 2.17.

Keywords

dense event description, Transformer network, semantic information, multi-modal fusion, deformable attention

Recommended Citation

Li Xiang, Sang Haifeng. Dense Video Description Method Based on Multi-modal Fusion in Transformer Network[J]. Journal of System Simulation, 2024, 36(5): 1061-1071.

第 36 卷第 5 期 系统仿真学报© Vol. 36 No. 5
2024 年 5 月 Journal of System Simulation May 2024

基于Transformer网络多模态融合的密集视频描述方法

李想，桑海峰*
(沈阳工业大学 信息科学与工程学院，辽宁 沈阳 110870)

摘要：针对目前的密集视频描述模型大多使用两阶段的方法存在效率较低、忽略音频及语义信息，描述结果不全面的问题。*提出了一种基于Transformer网络多模态和语义信息融合的密集视频描述方法。提取自适应R(2+1)D网络提取视觉特征，设计了语义探测器生成语义信息，加入音频特征进行补充，建立了多尺度可变形注意力模块，应用并行的预测头，加快模型收敛速度，提高模型精度。*实验结果表明：模型在2个基准数据集上性能均有很好的表现，评价指标BLEU4上达到了2.17。

关键词：密集事件描述；Transformer网络；语义信息；多模态融合；可变形注意力
中图分类号：TP391 文献标志码：A 文章编号：1004-731X(2024)05-1061-11
DOI: 10.16182/j.issn1004731x.joss.23-0017
引用格式: 李想, 桑海峰. 基于Transformer网络多模态融合的密集视频描述方法[J]. 系统仿真学报, 2024, 36(5): 1061-1071.
Reference format: Li Xiang, Sang Haifeng. Dense Video Description Method Based on Multi-modal Fusion in Transformer Network[J]. Journal of System Simulation, 2024, 36(5): 1061-1071.

Dense Video Description Method Based on Multi-modal Fusion in Transformer Network

Li Xiang, *Sang Haifeng**
(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: In order to solve the problems that most of the current dense video description models use two-stage methods, which have low efficiency, ignore audio and semantic information, and have incomplete description results, *a multi-modal and semantic information fusion dense video description method was proposed. An adaptive R(2+1)D network was proposed to extract visual features, a semantic detector was designed to generate semantic information, audio features were added to supplement it, a multi-scale deformable attention module was established, and a parallel prediction head was applied to accelerate the convergence rate and improve the accuracy of the model.* The experimental results show that the model has good performance on the two benchmark datasets, and the evaluation index BLEU4 reaches 2.17.

Keywords: dense event description; Transformer network; semantic information; multi-modal fusion; deformable attention

收稿日期：2023-01-04 修回日期：2023-03-24
基金项目：国家自然科学基金(62173078)；辽宁省自然科学基金(2022-MS-268)
第一作者：李想(1999-)，女，硕士生，研究方向为视频描述。E-mail：lixiang3278@163.com
通讯作者：桑海峰(1978-)，男，教授，博士，研究方向为视觉检测技术与图像处理。E-mail：sanghaif@163.com

0 引言

计算机视觉(computer vision, CV)、自然语言处理(natural language processing, NLP)等飞速发展。视频描述结合CV和NLP两个领域，通过CV的识别算法对视频进行识别，再使用NLP的生成器算法输出视频的描述。视频描述的研究基本都是用序列到序列的方法进行建模：通过将视觉特征信息和文本特征信息映射到同一个向量空间，并学习其分布，实现序列到序列的映射。文献[1]先使用卷积神经网络提取视觉图像特征，再通过长短时记忆网络学习向量空间中的分布，从而生成单词序列。然而，在现实生活中，视频通常很长，没有剪裁，由各种各样的事件组成，并且背景内容通常是无关的，所以单一描述往往比较平淡，信息量较少。

为了解决上述困境，研究人员提出了密集视频描述，对视频中的多个事件进行自动定位和描述，可以显示出详细的视觉内容，并生成连贯完整的描述。密集视频描述可以分为2个子任务，即事件定位和事件描述。因此，大多数现有的方法以两阶段的方式设计：用于事件生成的事件模块和用于为每个事件添加描述的描述模块。早期密集视频描述的方法主要集中在连接2个子任务上。文献[2]提出了密集视频描述模型，结合了多尺度事件模块和文本感知描述模块。文献[3]应用标注的描述对视频中的事件进行定位，根据所定位的事件生成描述来重建基准描述，使用最小化重建误差来训练整个模型。文献[4]提出了一种新的三维注意力模型，自动定位视频中的关键元素，无需任何额外的标注。注意力模型沿视频段的空间和时间维度生成局部区域的注意力权重，从而获得视频片段更有效的特征。文献[5]采用了双向循环神经网络，从2个方向解析自然语言描述，设计具有层次结构的网络，以联合建模语言描述和视频内容，利用多个粒度的视频内容，组合成具有一定逻辑结构的描述语段。但是由于网络复杂，导致训练速度及结果精度都不能得到很好的提升。

因Transformer网络[6]训练和测试速度快，所以越来越多的科研工作者开始使用Transformer网络解决问题，Transformer网络的成功也为视觉–文本的跨模态研究带了新的思路，图像描述和视频描述相继出现了应用Transformer网络的模型。文献[7]提出了具有可区分掩码的端到端框架，确保在学习过程中可以通过描述模型的梯度来细化事件位置。文献[8]提出了一种新的密集视频描述方法，利用任意数量的模态来描述事件，结合I3D特征与VGGish音频特征对视频进行编码，展示了音频语音模式可以改善密集视频描述模型，并利用Transformer架构将多模态输入特征转换为文本描述。文献[9]利用视觉音频特征来生成事件建议，通过捕获它们的时间和语义关系来增强事件级表示，并开发了一种动态融合和调节多模态信息的注意门控机制进行模态融合。现有的大多数研究都只关注视频的视觉特征，而忽略了其他模态带来的信息。不同的模态信息代表着看待事物的角度不同，因此，不同模态信息之间具有互补性。对于密集的视频描述任务，现有的研究[10-12]已经进行了一些多模态的尝试，如音频和语义。文献[13]使用C3D模型提取视频特征后，使用一种时序事件提取模块对视频事件进行提取，构建一种基于事件的时序–语义关联模型，为视频生成密集描述。文献[14]考虑到生成的句子包含对整个事件的丰富语义描述，将密集的视频字描述任务制定为视觉线索辅助的句子概括问题，借助语义特征将所有生成的句子概括为描述性句子。这种方法虽然做到了不同模态信息之间的交互，但是忽略了单一模态自身之间，以及语义信息对描述的影响。

针对上述问题，本文提出了一种基于Transformer网络多模态和语义信息融合的密集视频描述框架(dense video caption based on multi-modal and semantic information fusion in

DOI: 10.16182/j.issn1004731x.joss.23-0017

Transformer network，MSTVC)。主要贡献包括：

(1) 将密集视频描述任务视为集预测问题，使用定位头和描述头，将视频精确地分割为多个事件片段，提出语义探测器(semantic meaning detector, SMD)对视频片段的语义特征进行提取。

(2) 提出自适应R(2+1)D网络(self adaptive R(2+1)D，A-R(2+1)D)进行特征提取，使用前期融合，再将提取的视觉、音频以及语义特征直接解码为具有位置和描述的事件集。

(3) 在编码阶段，提出了将多尺度可变形时序注意力模块(multi-scale deformable attention, MSDT)应用到Transformer网络，生成能够互相关联，描述全面的结果。

1 基于Transformer网络多模态和语义信息融合的密集视频描述模型

本文采用PDVC[15]作为基线模型，PDVC是一种端到端的密集事件描述框架，与以往方法使用两阶段的模型不同，该模型直接生成一组特定于事件的句子，通过注意机制捕获帧间、事件间和事件帧交互，并产生一组事件查询特征，应用事件计数器从全局视图中预测事件数量，生成更加连贯的描述语句。

如图1所示，模型在PDVC的基础上，进一步增加了语义信息提取、多模态特征融合模块结合多尺度可变形注意力机制改善模型性能。模型遵循编码器—解码器结构，在编码阶段，输入一个视频帧序列，应用A-R(2+1)D网络、语义探测器及音频特征提取器对视频特征进行多模态提取。采用多模态特征融合模块对提取的特征进行融合。Transformer编码器融合了位置嵌入生成特征序列，并加入MSDT模块对网络进行优化。在解码阶段，Transformer解码器以事件查询序列和编码特征为输入，后面有3个并行头部。定位头和描述头分别预测每个查询的时间点和标题，事件计数器预测视频中事件的实际数量。

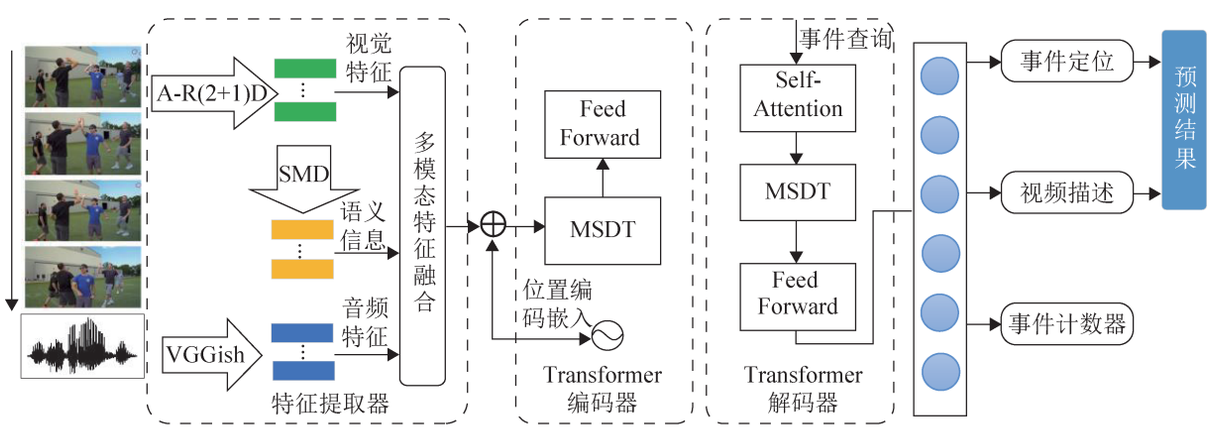


图1　网络结构图

Fig. 1　Network structure

1.1 密集事件描述

在密集事件描述任务中，输入是一个视频帧序列$\boldsymbol{v}=\{v_t\}$，其中，$t\in 0,1,\cdots,T-1$，按时间顺序对帧进行索引。输出为一组句子$s_i\in\boldsymbol{s}$，$\boldsymbol{s}=(t^{\text{start}},t^{\text{end}},\{w_i\})$，由每个句子的开始和结束时间组成，每个句子由一组单词$w_i\in\boldsymbol{W}$组成，每个句子的长度不同，$\boldsymbol{W}$为词汇集。模型首先通过事件分割得到一组事件$\boldsymbol{P}=\{(t_i^{\text{start}},t_i^{\text{end}},c_i,h_i)\}$，$c_i$为事件对应的分数；再将隐藏层$h_i$作为视频描述模块的输入，得到最终的描述结果。

1.2 特征提取器

特征提取器模块由多模态特征提取网络和多

模态特征融合网络构成，并应用多尺度特征提取器，能够更加全面的对特征进行提取。对于视频的视觉特征，提出了 A-R(2+1)D 网络进行提取；对其局部片段特征进行语义信息的提取；对于视频音频特征，使用 VGGigh 网络。

1.2.1 多尺度特征提取器

本文使用 R(2+1)D 网络作为视频提取器的主干。R(2+1)D 网络：将 3D 卷积分解为 2 个单独且连续的运算，即一个 2D 空间卷积和一个 1D 时间卷积，3D 卷积核大小为 $t\times d\times d$，分为 $1\times d\times d$ 空间卷积核和 $t\times 1\times 1$ 时间卷积核，如图 2 所示。该网络将 3D 卷积网络分解，不仅使 2 个运算之间增加的非线性修正单元提高了网络的非线性能力，而且分解操作有助于优化，既降低了训练损失，又降低了测试损失。由于网络复杂度的降低会削弱表示能力，提出了 A-R(2+1)D 网络，引入了超参数 T 将 R(2+1)D 网络的参数数量恢复为 3D 块：

$$T=\frac{m\times n\times t\times d^2}{n\times d^2+m\times t} \tag{1}$$

式中：m、n 分别为输入输出维度；t、d 为卷积核大小。

应用多尺度特征提取器提取视频帧序列，提取多模态的特征。每个多尺度特征提取器都由一个 A-R(2+1)D 网络组成。为了提取视频中丰富的时空特征，采用预先训练的动作识别网络提取帧级特征。通过插值将特征图的时间维度重新缩放到一个固定的数字 H，以方便批处理。由于 A-R(2+1)D 网络的特殊结构，拥有 1D 时间卷积层，能够更好地利用多尺度特征预测多尺度事件。

输出序列长度为

$$H'=\sum_{l=0}^{L}\frac{H}{2^l} \tag{2}$$

式中：L 为时间卷积层层数，即 A-R(2+1)D 网络层数。

1.2.2 语义探测器(SMD)

由于通用的特征提取网络提取出的视频特征与实际视频语义存在偏差，所以，目前视频描述的结果仍会出现与视频内容语义之间不匹配的情况。为了改善这个问题，将语义信息与视频特征进行融合，聚合到编码器–解码器网络中，从而改善视频描述的准确性，使描述结果更加接近人工标注的描述语句。

SMD 网络的输入为 A-R(2+1)D 编码的局部片段特征的平均池化，由 AlexNet 网络及多层感知机组成，建立 FC+Sigmoid 层，得到语义信息的预测概率，整体架构如图 3 所示。为了构建语义标签，从 ActivityNet caption 的基础真句中选取最常用的 K 个名词、动词、形容词、副词作为词汇，从词汇表中去掉停止词。假设有 N 个视频，则 $\boldsymbol{V}=\{V_i\}_{i=1}^{N}$，$\boldsymbol{Y}=\{\boldsymbol{y}_i\}_{i=1}^{N}$，$\boldsymbol{y}_i=[y_{i1},y_{i2},\cdots,y_{ik}]\in\{0,1\}^K$，$\boldsymbol{y}_i$ 为第 i 个视频的语义标签。

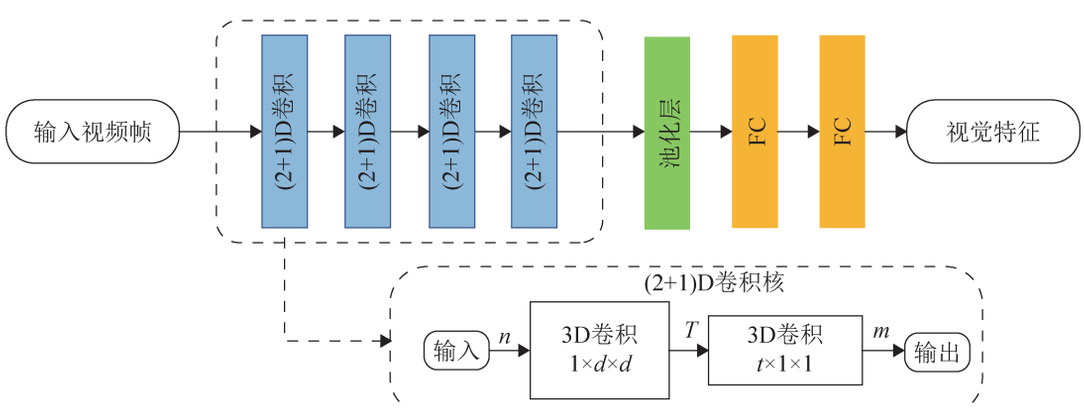


图 2 A-R(2+1)D 网络结构图
Fig. 2 A-R(2+1)D network structure

https://dc-china-simulation.researchcommons.org/journal/vol36/iss5/2
DOI: 10.16182/j.issn1004731x.joss.23-0017

$$y_{ik}=\mathrm{SMD}(V_i) \tag{3}$$

$$L_{\mathrm{SMD}}=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(y_{ik}\log_2(y_{ik})+(1-y_{ik})\log_2(1-y_{ik})) \tag{4}$$

1.2.3 音频特征提取器

音频特征广泛存在于视频数据中，无论是视频数据集还是真实世界，视频数据往往伴随着音频信号。单独将音频特征作为视频描述生成，会导致视频描述不全面，但是音频信息却是视觉信息特征很好的补充，对分辨视频中不同场景的事件十分有用。因此，本文将音频特征与视觉特征结合，采用VGGish网络，得到一个128维的数据，再经过PCA即得到最终的特征。该模型使用YouTube数据集进行预训练，网络模型基于VGG网络，网络结构如图4所示。

1.2.4 多模态特征融合

多模态特征融合模块融合了所有模态的特征，以及语义信息。将特征投影到嵌入空间中，再通过帧进行拼接，融合特征记为$\{f_j\}_{j=1}^{H'}$，对于特征提取阶段提取的视觉特征和音频特征需要经过处理。由于音频文件长度不一，最终分割后提取的音频特征长度也不一致。将音频特征经平均处理后，长度对齐，扩展成视觉特征相同维度。数据集中存在一些视频数据，没有包含语音信息，此时则用零向量代替，以确保最终的拼接融合向量维度相一致。

1.3 MSDT模块

本文使用具有编码器-解码器结构的Transformer网络，通过注意机制捕获帧间、事件间和事件帧之间的交互，并产生一组事件查询特征。然后，两个平行的预测头同时预测每个事件查询的定位和描述。事件计数器从全局视图预测事件编号。通过高置信度选择事件来获得最终结果，以保证故事的完整和连贯。

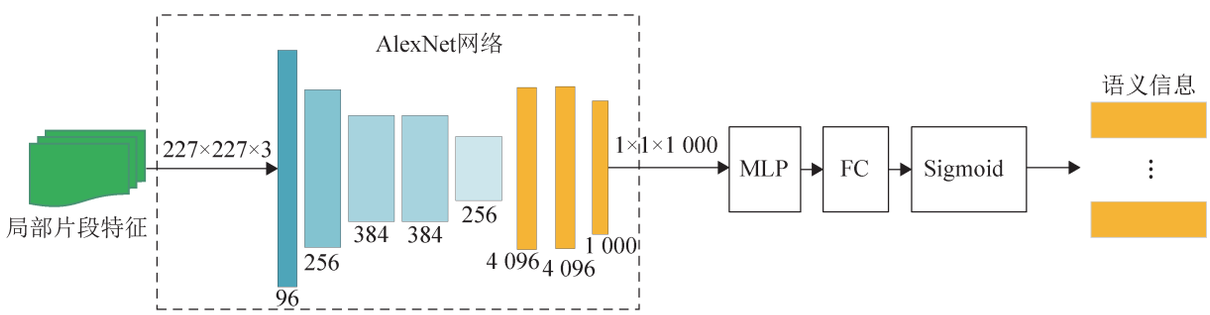


图3 语义探测器网络
Fig. 3 Semantic meaning detector network

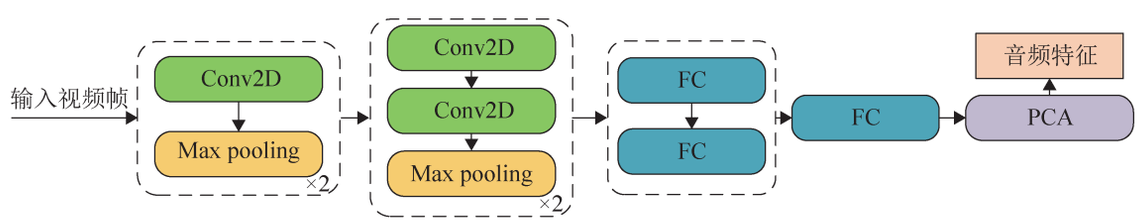


图4 音频特征特区网络
Fig. 4 Audio characteristic zone network

可变形 Transformer 用 MSDT 模块取代了 Transformer 编码器中的自注意模块和 Transformer 解码器中的交叉注意模块，使其具有较快的收敛速度和更好的目标检测表示能力，提高模型的局部信息提取能力，从而使生成的文本描述更具有逻辑性和可读性。给定多尺度特征映射 $\boldsymbol{X}=\left\{x^l\right\}_{l=1}^{L}$，$x^l\in\mathbb{R}^{C\times H'\times W}$，通过 L 尺度特征映射上 $K\times L$ 采样点的加权和输出上下文向量为

$$\mathrm{MSDT}(q_j,p_j,X)=\sum_{l=1}^{L}\sum_{k=1}^{K}A_{jlk}\boldsymbol{W}\boldsymbol{X}^l_{\tilde{p}_{jlk}} \tag{5}$$

$$\tilde{p}_{jlk}=\phi_l(p_j)+\Delta p_{jkl} \tag{6}$$

式中：q_j 为第 j 个事件查询；$p_j\in[0,1]^2$ 为归一化参考点；$\boldsymbol{W}$ 为 k 的投影矩阵；$\tilde{p}_{jkl}$ 和 A_{jkl} 分别为第 j 个事件查询和第 k 个采样键值在第 l 个尺度上的位置权重和注意力权重；ϕ_l 为 l 层上 p_j 到特征图的投影；Δp_{jkl} 为偏移量。MSDT 将融合的特征序列与位置嵌入一起，通过应用多尺度可变形注意力来产生最终的可视化表示。MSDT 有助于捕获多尺度帧间交互。

在激活函数阶段，本文使用了高斯误差线性单元(Gaussian error linear unit, GELU)激活函数，GELU 将非线性与随机正则化结合，是 Adaptive Dropout 的改进。令 $X\sim N(0,1)$ 为标准正态分布的累积分布函数。GELU 激活函数为

$$\mathrm{GELU}(x)=xP(X\leqslant x)=$$

$$x\phi(x)=0.5x\left(1+\mathrm{e}^{\frac{x}{2}}\right) \tag{7}$$

$$\mathrm{e}^x\approx\frac{2}{\sqrt{\pi}}\tanh(x) \tag{8}$$

1.4 视频描述模块

对于视频描述，使用长短期记忆网络(LSTM)将 Transformer 解码器预测的事件、上下文特征、前面的单词作为输入，最终得到预测句子 $\boldsymbol{S}_j=\{w_{j1},w_{j2},\ldots,w_{jM_j}\}$，其中，$M_j$ 为句子长度。

1.5 损失函数

在事件预测部分，如果事件查询与标注事件匹配，则标签目标设置为基础真值；如果不匹配，标签目标设为一个全零向量。在训练过程中，模型生成了一组 N 个事件及其定位和标题，预测的事件集必须与标注的事实相匹配，使用匈牙利算法得到最佳的二分匹配结果。匹配算法定义为

$$C=\alpha_{\mathrm{GIoU}}L_{\mathrm{GIoU}}+\alpha_{\mathrm{cls}}L_{\mathrm{cls}} \tag{9}$$

式中：L_{GIoU} 为预测时间段与标注段之间的广义 IoU；L_{cls} 为预测分类评分与标注标签之间的交叉熵。计算整体预测损失为 GIoU 损失、分类损失、描述损失的加权和：

$$L=\beta_{\mathrm{GIoU}}L_{\mathrm{GIoU}}+\beta_{\mathrm{cls}}L_{\mathrm{cls}}+\beta_{\mathrm{cap}}L_{\mathrm{cap}} \tag{10}$$

式中：L_{cap} 为预测单词概率和根据标题长度归一化的标注之间的交叉熵；L_{cap} 为动作 L_{action}、时序 L_{tep} 和语义损失 L_{SMD} 的加权和。

$$L_{\mathrm{cap}}=L_{\mathrm{action}}+\alpha L_{\mathrm{SMD}}+\beta L_{\mathrm{tep}} \tag{11}$$

2 实验结果与分析

2.1 数据集

本文使用大型基准数据集 ActivityNet caption 和 YouCook2。ActivityNet 数据集包含各种人类活动的 20 000 个未修剪的视频。平均每个视频时长为 120 s，并配有 3.65 句标注好的句子。本文应用 10 009/4 925/5 044 的标准分割用于训练、验证和测试的视频。

YouCook2 数据集有 2 000 个未经剪辑的烹饪过程视频，平均时长 320 s。每个视频有 7.7 个带标注的片段和相关的句子。本文应用官方拆分 1 333/457/210 视频进行训练、验证和测试。

2.2 评价指标

在事件定位性能方面，使用 IoU 为{0.3, 0.5, 0.7, 0.9}的平均精度和召回率；在密集描述的性能方面，本文遵循 ActivityNet Challenge 2018 提供的官方评估工具，计算生成描述与相应标注描述之间的平均精度，即评价指标 Bleu_4、METEOR[16] 和 CIDEr[17]。

然而，这些评价指标并没有考虑到描述的整

DOI: 10.16182/j.issn1004731x.joss.23-0017

体质量，而面向故事的密集视频描述评价框架(story oriented dense video captioning evaluation framework, SODA)[18]被用于衡量视频故事描述系统的性能，即生成的描述是否覆盖视频的整体。所以本文进一步采用SODA_c进行全面评估。

2.3 实验参数设置

使用PDVC模型作为基线模型。在编码阶段，对于ActivityNet caption数据集，使用预训练的A-R(2+1)D提取帧级运动和外观特征。对局部片段特征进行语义信息提取和并行计算，所有的特征序列在时间上调整为相同的长度。长度大于1 024的序列缩减为1 024的时间长度。长度小于1 024被填充为1 024。使用了一个具有多尺度(4级)变形注意的Transformer。在Transformer的MSDT层中，隐藏层维度为512，前馈层维度为2 048。在解码阶段，利用分类头对标注的区域标签进行预测。对于事件计数器，选择ActivityNet GIoU的最大计数为10。二分匹配权重比为α_{GIoU}:α_{cls}=2:1，损失权重比为：β_{GIoU}:β_{cls}:β_{cap}=2:1:1:1，应用Adam优化器，初始学习率为0.000 05，设置$\alpha=1$，$\beta=0.1$。

输入数据：使用带有时间标注的未修剪的视频对模型进行预训练。给定一个未修剪的视频，采样一个固定大小的输入片段X，大小为$3\times L\times H\times W$，其中，3是RGB通道，$L$是帧数，$H$和$W$是帧的高度和宽度。将每个视频名称映射到一个大小为$N\times512$的特征张量，$N$是特征的数量，512是特征的大小。使用16帧的剪辑，帧率为15帧/s，步幅为16帧。每1.067 s为一个特征向量。

2.4 实验结果

将本文模型与其他视频描述模型生成的结果进行对比，为了验证密集视频描述的描述性能，本文采用不同模型在ActivityNet数据集上进行实验，并与本文模型进行对比。在该数据集上，使用数据集内标注的事件，进行视频描述任务。实验结果如表1所示。从表1可以看出，本文各个指标上都具有较好的效果，尤其在Bleu_4得到了较高的分数。本文主要针对描述文本的逻辑性进行改进，Bleu_4评价指标是对候选语句与标注文本中的相匹配的4元组，主要考虑到描述语句的逻辑，说明本文使用的多模态特征及语义信息融合的方法能够使文本描述更加精细化，MSDT模块的加入使得模型能够更加关注局部信息，从而很好的改善视频描述结果的逻辑性。METEOR、CIDEr、SODA_c评价指标则是关于描述的召回率、一致性和密集事件匹配度的评价，而软注意力机制的加入会使模型更关注局部信息，使得模型对全局信息的应用会稍显不足，因此，在这3个指标上，本文模型虽优于大部分模型，但是略低于一些模型结果。

表1 ActivityNet数据集上应用标注事件的实验结果
Table 1 Experimental results of ground-truth proposals on ActivityNet dataset

模型	Bleu_4	METEOR	CIDEr	SODA_c
Transformer-XL[19]	1.93	10.03	40.32	5.21
SGN[20]	1.75	9.43	40.33	—
BMT[8]	1.99	8.78	39.12	5.45
MMT[21]	1.51	8.62	40.02	—
MART[22]	1.93	8.93	41.32	5.66
PDVC[15]	2.07	9.34	42.05	6.11
MSTVC	2.17	9.03	41.14	6.05

注：“—”代表该模型无此指标的评测结果。

对于密集事件描述，本文使用未标注的事件进行实验，从表2结果可以看出，模型依然具有较好的效果，本文使用的方法不同于其他模型，将密集视频描述分解为两阶段，而是直接进行端到端的事件描述，减少了两阶段模型数据的损失，提高了模型效率，而且生成的预测结果依然优于大部分模型。

表2 ActivityNet数据集上应用预测事件的实验结果
Table 2 Experimental results of predicted proposals on ActivityNet dataset

模型	Bleu_4	METEOR	CIDEr
BMT	1.75	7.55	25.33
MMT	1.70	7.41	26.66
Transformer-XL	1.65	8.03	28.52
MART	1.78	7.68	28.11
SGN	1.80	8.08	29.59
MSTVC	1.87	8.01	29.04

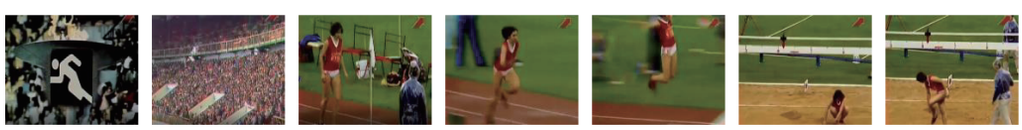
表 3 为模型在 YouCook2 数据集上使用标注事件进行的实验，在所有评价指标上都取得了较好的结果，说明本文模型具有很好的泛化性。

表 3　YouCook2 数据集上使用标注事件的实验结果
Table 3　Experimental results of ground-truth proposals on YouCook2 dataset

模型	Bleu_4	METEOR	CIDEr	SODA_c
BMT	0.81	3.75	21.01	3.95
Transformer-XL	0.76	3.43	20.33	3.55
PDVC	0.87	4.54	22.78	4.42
MSTVC	0.92	4.25	21.65	4.22

2.5 定性分析

图 5 为该视频在 BMT、PDVC、MSTVC 模型上所生成的事件描述，分割出的事件对应的开始时间、结束时间以及描述结果。红色字体为描述较好的位置，对于第 3 个事件，BMT 和 PDVC 模型都错误地预测了“woman”，这是因为在没有上文信息、特征提取不准确的情况下，模型难以预测到在此时间段的是“woman”，MSTVC 模型所应用的 A-R(2+1)D 网络可以更准确的提取视频特征，得到正确的预测；对于图中第 2 个事件，MSTVC 生成更加准确的描述事件区间，并且没有生成重复的第 1 个和第 4 个事件，可以看出，MSDT 模块的加入是有效的，模型生成了相对准确且没有重复的事件；绿色字体为 MSTVC 模型相对其他 2 个模型生成的新事件，通过全局的关联，将事件定位在 0~4.8 s 区间，加入了音频、语义信息，使模型可以习得背景及音频信息，生成“speaking”，得到更加准确连贯的描述。同时，从定性结果可以看出，即使模型 METEOR 等的评分不是目前最优的结果，但最终生成的语句依然是更符合人类语言逻辑的。



BMT	PDVC	MSTVC
0.0-4.9:We see the closing title screen	0.0-4.7:We see the black title screen	——
2.7-29.0:A woman is seen running down a track and down a track while others watch on the sides	2.5-28.5:A woman is seen running on a track and down a track while others watch on the sides	2.5-28.1:A woman is seen running on the track and others watch on the sides
19.6-33.3: The man runs down the track and jumps into a sand pit	19.4-33.6: The man runs down a track and jumps into a sand pit	19.6-33.6: The woman runs down the track and jumps into a sand pit
0.9-2.5:We see a title screen	——	——
0.0-1.6:A man is seen sitting on a table with a white words on the screen	0.0-1.6:A man is seen sitting on a table with a white words on the screen	0.0-4.8：We see a screen with a white pattern and a man is speaking

图 5　定性结果
Fig. 5　Qualitative results

3　消融实验

3.1 A-R(2+1)D 网络的对比分析

综合考虑数据集及网络参数数量，使用不同网络进行对比实验，如表 4 所示。R(2+1)D-18 网络处理视频特征会出现过拟合现象，R(2+1)D-152 网络参数量、计算量大幅增加，导致模型训练时间过长，且对实验精度没有很大提升，因此，模型最终应用 R(2+1)D-34 网络进行实验。

表 4　R（2+1）D 网络参数量及效果对比结果
Table 4　Comparison of the number and effect of R(2+1)D network parameters

网络	参数量	计算量	METEOR
C3D	98.32	59.25	8.09
R(2+1)-18	33.15	30.55	6.25
R(2+1)-34	45.26	60.56	8.89
R(2+1)-152	100.88	91.54	8.91

DOI: 10.16182/j.issn1004731x.joss.23-0017

针对A-R(2+1)D网络提取视觉特征对于实验结果的影响进行了消融实验，在ActivityNet caption数据集上，应用不同网络进行实验对比，由表5可以看出，A-R(2+1)D网络在各个评价指标上均有提升。A-R(2+1)D网络的分解结构使网络训练速度更快，非线性能力让网络更适应视频的多样性，本文对A-R(2+1)D网络进行改进，改善了由于网络结构复杂度降低带来的问题。

表5 不同特征提取网络进行实验的对比结果

Table 5 Comparison of experimental results of different feature extraction networks

Features	Recall	Precision	Bleu_4	METEOR	CIDEr	SODA_c
C3D	55.20	57.36	1.82	8.09	38.16	5.47
TSN	56.21	57.46	1.92	8.63	39.00	5.68
I3D	55.88	57.97	1.97	8.97	40.21	6.11
R(2+1)	55.76	57.88	1.99	8.89	41.01	6.08
A-R(2+1)	55.79	57.39	2.09	9.03	41.14	6.05

3.2 多模态融合的的对比分析

大多数视频描述方法都集中在视觉特征上，没有音频的帮助，而音频又会出现在大多数视频中。音频可以通过提供背景信息等来帮助视频描述；语义信息可以使视频描述的结果与视频内容语义之间更加一致。本文使用视觉、音频、语义信息多模态的特征提取，在ActivityNet caption数据集上，不同特征输入结果对比如表6所示，音频特征的加入使描述结果更加丰富，而语义信息的加入使结果更具有逻辑性。

表6 不同模态特征输入进行实验的对比结果

Table 6 Comparison of experimental results of different modal feature inputs

Features	Recall	Precision	Bleu_4	METEOR	CIDEr	SODA_c
A-R(2+1)	56.21	57.46	1.92	8.63	29.00	5.68
A-R(2+1)+VGGish	56.03	57.12	2.03	9.11	41.01	6.10
A-R(2+1)+VGGish+语义信息	55.79	57.39	2.17	9.03	41.14	6.05

3.3 MSDT模块的对比分析

加入MSDT模块对Transformer网络进行了改进，表7为ActivityNet caption数据集上的消融实验。可以看出，MSDT模块的加入使模型的2个指标均有明显提升，证明了MSDT模块的有效性。MSDT模块中可变形注意力使网络目标检测能力提高，使网络能够更快速地检测到正确的事件，同时，多尺度的加入使模型可以考虑到上下文的结果，避免生成重复事件，应用GELU激活函数的平滑输出，使模型得到更好的结果。

表7 不同Transformer网络进行实验的对比结果

Table 7 Comparison of experimental results of different Transformer networks

模型	Bleu_4	METEOR	CIDEr
Transformer-XL	1.87	9.15	40.61
DT	1.92	9.05	40.52
MSDT	2.17	9.03	41.14

4 结论

目前，密集事件描述的模型大部分都是两阶段的，而两阶段的方法会更加依赖事件分割的准确性，导致2个子任务不能互相促进；且大多模型都应用视频的单一模态，忽略了不同模态之间的互补性。针对以上问题，本文提出了基于Transformer网络多模态和语义信息融合的密集视频描述框架，其应用多尺度可变形Transformer网络对融合了语义信息的多模态特征进行处理计算，对损失函数进行加权计算，生成密集视频描述，实现了对描述的优化。加入语义信息，使模型生成结果与标注的描述语义上更加一致；使用A-R(2+1)D网络提高特征提取的精度；应用音频特征、视觉特征的多模态输入，提高了描述结果的准确度；对编码器–解码器网络的改进增强了模型的泛

化能力。实验结果表明，模型在 ActivityNet caption 数据集上拥有较好的效果。

模型在公开数据集上取得了不错的进展，但依然有一些需要改进的地方。由于模型复杂度的增加，导致训练时间的增加，这使得模型效率降低，因此需要进一步探索如何高效的生成视频描述；可变形注意力的加入虽然能够使模型很好的处理局部特征，但仍存在一些不足，在后期的研究工作中，将对其进行改良，进一步提升模型性能。

参考文献：

[1] Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to Sequence-video to Text[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2015: 4534-4542.

[2] Krishna R, Hata Kenji, Ren F, et al. Dense-captioning Events in Videos[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 706-715.

[3] Duan Xuguang, Huang Wenbing, Gan Chuang, et al. Weakly Supervised Dense Event Captioning in Videos [EB/OL]. (2018-12-10) [2022-07-12]. https://arxiv. org/abs/1812.03849.

[4] Jiao Yifan, Li Zhetao, Huang Shucheng, et al. Three-dimensional Attention-based Deep Ranking Model for Video Highlight Detection[J]. IEEE Transactions on Multimedia, 2018, 20(10): 2693-2705.

[5] Ning Ke, Cai Ming, Xie Di, et al. An Attentive Sequence to Sequence Translator for Localizing Video Clips by Natural Language[J]. IEEE Transactions on Multimedia, 2020, 22(9): 2434-2443.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[7] Yu Zhou, Han Nanjia. Accelerated Masked Transformer for Dense Video Captioning[J]. Neurocomputing, 2021, 445: 72-80.

[8] Vladimir Iashin, Esa Rahtu. A Better Use of Audio-visual Cues: Dense Video Captioning with Bi-modal Transformer[C]//The 31st British Machine Vision Conference. Durham: BMVC, 2020: 111.

[9] Chang Zhi, Zhao Dexin, Chen Huilin, et al. Event-centric Multi-modal Fusion Method for Dense Video Captioning [J]. Neural Networks, 2022, 146: 120-129.

[10] Xu Yuecong, Yang Jianfei, Mao Kezhi. Semantic-filtered Soft-split-aware Video Captioning with Audio-augmented Feature[J]. Neurocomputing, 2019, 357: 24-35.

[11] Wu Chunlei, Wei Yiwei, Chu Xiaoliang, et al. Hierarchical Attention-based Multimodal Fusion for Video Captioning[J]. Neurocomputing, 2018, 315: 362-370.

[12] Sujin Lee, Incheol Kim. Learning Semantic Features for Dense Video Captioning[J]. Journal of KIISE, 2019, 46 (8): 753-762.

[13] Wang Teng, Zheng Huicheng, Yu Mingjing, et al. Event-centric Hierarchical Representation for Dense Video Captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(5): 1890-1900.

[14] Zhang Zhiwang, Xu Dong, Ouyang Wanli, et al. Show, Tell and Summarize: Dense Video Captioning Using Visual Cue Aided Sentence Summarization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(9): 3130-3139.

[15] Wang Teng, Zhang Ruimao, Lu Zhichao, et al. End-to-end Dense Video Captioning with Parallel Decoding[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 6827-6837.

[16] 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. Stroudsburg, PA, USA: ACL, 2005: 65-72.

[17] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based Image Description Evaluation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 4566-4575.

[18] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, et al. SODA: Story Oriented Dense Video Captioning Evaluation Framework[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 517-531.

[19] 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. Stroudsburg, PA, USA: ACL, 2019: 2978-2988.

DOI: 10.16182/j.issn1004731x.joss.23-0017

[20] Hobin Ryu, Sunghun Kang, Haeyong Kang, et al. Semantic Grouping Network for Video Captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3), 2514-2522.

[21] Valentin Gabeur, Sun Chen, Karteek Alahari, et al. Multi-modal Transformer for Video Retrieval[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 214-229.

[22] Lei Jie, Wang Liwei, Shen Yelong, et al. MART: Memory-augmented Recurrent Transformer for Coherent Video Paragraph Captioning[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 2603-2614.