基于潜在特征增强网络的视频描述生成方法

李伟健，胡慧君

（武汉科技大学计算机科学与技术学院，湖北 武汉 430065）

摘　要：视频描述生成旨在用自然语言描述视频中的物体及其相互作用。现有方法未充分利用视频中的时空语义信息，限制了模型生成准确描述语句的能力。为此，提出一种用于视频描述生成的潜在特征增强网络（LFAN）模型。利用不同的特征提取器提取外观特征、运动特征和目标特征，将对象级的目标特征分别和帧级的外观特征与运动特征融合，同时对融合后的不同特征进行增强，在生成描述前利用图神经网络和长短时记忆网络推理对象之间的时空关系，从而得到具有时空信息和语义信息的潜在特征，同时使用长短时记忆网络和门控循环单元的解码器生成视频的描述语句。该网络模型能够准确地学习到对象特征，进而引导生成更准确的词汇及与对象之间的关系。在MSVD和MSR-VTT数据集上的实验结果表明，LFAN模型可以显著提高生成描述语句的准确性，并与视频中的内容呈现出更好的语义一致性，在MSVD数据集上的BLEU@4和ROUGE-L分数分别为57.0和74.1，在MSR-VTT数据集上分别为43.8和62.1。

关键词：视频描述生成；潜在特征增强网络；时空语义信息；图神经网络；特征融合

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0067206

Video Description Generation Method Based on Latent Feature Augmented Network

LI Weijian，HU Huijun

（School of Computer Science and Technology，Wuhan University of Science and Technology，Wuhan 430065，Hubei，China）

【Abstract】Video description generation aims to use natural language to describe objects and their interactions in videos. The existing methods do not fully utilize the spatio-temporal semantic information in videos，which limits the model's ability to generate accurate descriptive statements. To this end，a Latent Feature Augmented Network（LFAN）model is proposed for video description generation.Different feature extractors are used to extract appearance，motion，and target features，thereby fusing object level target features with frame level appearance and motion features. Concurrently，the fused different features are enhanced. Before generating descriptions，graph neural and long short-term memory networks are used to infer the spatio-temporal relationships between objects，thereby obtaining potential features with spatio-temporal and semantic information. Finally，a decoder using both a long short-term memory network and a gated loop unit is used to generate a description statement for the video. This network model can accurately learn object features and guide the generation of more accurate vocabulary and relationships with objects.The experimental results on MSVD and MSR-VTT datasets show that the LFAN model can significantly improve the accuracy of generating descriptive statements，exhibiting better semantic consistency with the content in the video. The BLEU@4 and ROUGE-L scores are 57.0 and 74.1 on MSVD，respectively，and 43.8 and 62.1 on the MSR-VTT dataset.

【Key words】video description generation；latent feature augmented network；spatio-temporal semantic information；graph neural networks；feature fusion

0　引言

视频描述生成旨在根据视频内容自动生成描述性的自然语言句子，视频中蕴含着丰富的信息[1]，以往的研究致力于理解静态的视觉信息，但是如何对视频中丰富的时空信息进行建模仍是一项具有挑战性的任务[2]。

视频中显著对象的检测是目前计算机视觉前沿领域必不可少的关键任务，该任务通常被称为目标检测任务。例如，文献[3]提出的OA-BTG通过构建双向序列图来提取视频中的重要目标，然后整合整个视频中的全局特征生成字幕。文献[4]提出的STG-KD采用图卷积网络（GCN）对检测到的对象进行关系推理，以增强对象级的关系表示。文献[5]提出的DETR首次采用Transformer方法进行视频描述生成中的对象检测，计算输出区域和真实区域的集合相似度，将对象检测视为1个直接的集合预测问题。因此，处理不同视频帧中对象之间的关系是视

收稿日期：2023-03-20　　**修回日期：**2023-05-13

基金项目：国家自然科学基金（62271359）。

通信作者 E-mail：liweijian@wust.edu.cn

频描述生成任务的关键。

现有方法往往是根据编码器-解码器的结构设计来生成视频描述。通过不同的特征提取器，如IncepResNetV2[6]、I3D[7]和 Faster R-CNN[8]，不同编码器可以从不同角度捕捉视频信息。显然，同时使用不同的特征进行连接可能会取得更优的性能，但是这种方法往往会忽略不同特征之间的上下文语义信息，而这些信息在具有时空信息的视频中起着重要作用。XU 等[9]和 WANG 等[10]通过对视频帧进行局部特征融合来学习判别性的特征，从而提高视频描述生成质量。例如，文献[11]提出的 SAAT 通过融合对象和时间特征来生成相应的动词。但是，只融合局部特征难以获得全局的时空语义视频信息。例如文献[12]提出的 POS+CG 设计 1 个交叉门控模块来融合外观和运动特征，并进行综合阐述，然而，仅通过预测的全局 POS 来表示生成的每个单词，而忽略了微妙的细节信息，从而难以捕获准确的对象。

为了解决上述问题，本文设计新的潜在特征增强网络（LFAN）模型。该模型融合不同的特征来生成具有更高维度的潜在特征，并且通过构建连接视频特征的动态图来获取时空信息，并利用 GNN 和长短时记忆（LSTM）网络推理对象间的时空关系，进一步丰富视频内容的特征表示，并结合 LSTM 和门控循环单元（GRU）设计一种新的解码方法来处理上下文信息和全局信息，从而生成准确、流畅的视频描述。

1 相关工作

1.1 视频描述生成

视频描述生成作为计算机视觉和自然语言处理的交叉领域，早期大多数方法都是基于特定的模板[13-15]，这些模板需要大量人工设计的语言规则，并且处理有限类别的对象、动作等，难以生成准确的语句描述。

随着深度神经网络的兴起，VENUGOPALAN 等[16]提出一种编码器-解码器框架来克服这些限制。当前，基于编解码框架的视频描述生成方法成为主流。YAO 等[17]提出一种动态总结视觉特征的时间注意力机制。CHEN 等[18]提出从视频中去除冗余帧，从而解码重要的视觉信息以生成视频描述。文献[19]提出的 M3 通过建立记忆网络来模拟长期的视觉文本依赖，以生成高质量的描述。文献[20]提出的 MARN 设计一种记忆结构来寻找候选词汇和包含它所有视频特征的关系。TAN 等[21]提出一种新的时空视觉推理模块 RMN，实现显式的、可解释的视频字幕处理。BAI 等[22]采用生成对抗网络（GAN）来保证生成描述的准确性。RYU 等[23]提出一种语义分组网络（SGN），通过语义组预测下 1 个生成的单词。Open-Book[24]从语料库中检索语句，作为生成描述性语句的指南。CHEN 等[25]提出 R-ConvED，从已注释的视频句子对中检索相关的视觉内容和句法结构，并利用这些上下文知识促进描述性语句的生成质量。

最新的研究挑战则是尝试构建大规模的端到端训练视频描述生成网络，如 LIN 等[26]在视频描述生成领域中使用 SwinBERT 进行端到端训练，以生成视频描述。但是这类网络模型通常使用 Transformer 进行编解码，训练参数量庞大，并且需要大量的计算资源。

1.2 潜在特征

不同的特征信息在生成视频描述中起着重要作用。文献[27]提出的 GRU-EVE 使用对象标签增强视觉特征的语义信息。文献[12]提出的 POS+CG 构建一种新颖的门控融合网络，对视频的外观和运动特征进行编码和融合。文献[4]提出的 STG-KD[通过 GCN 构建对象关系图，利用对象关系图推理视频对象之间的时空关系以获得潜在特征。文献[28]提出的 ORG-TRL 使用 GCN 实现关系推理从而获取视频中的潜在特征，丰富细节对象的表示。文献[11]提出的 SAAT 设计 1 个语法感知模型来增强动词的生成，使动作和目标之间的相关性更强。图 1 所示为 LFAN 生成描述的直观示例。本文使用基线模型 Baseline 作为对比，其中仅使用传统的编码器-解码器框架，没有使用图神经网络和改进的解码方式。从图 1 可以看出，基线模型缺乏时空语义信息，无法对视频上下文进行全面探索，从而产生较差的描述。SAAT 生成目标对象和相应的动词，但没有捕获完整的视频信息，从而生成不完整的描述语句。相反，本文模型通过捕捉突出的对象“man”“chicken”和“plastic”，学习它们之间的对应关系从而生成准确的动词“put”，并完整描绘出视频内容。

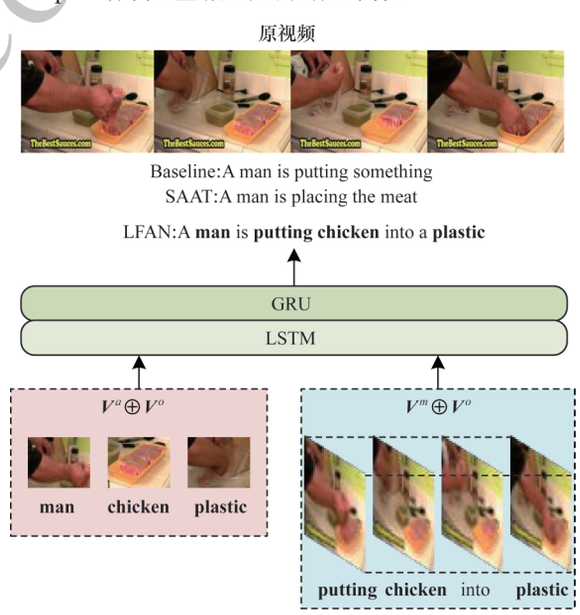


图 1 LFAN 生成描述的直观示例

Fig.1 An intuitive examples of LFAN generation descriptions

2 潜在特征增强网络

LFAN模型框架如图2所示。LFAN模型由编码层、潜在特征层和解码层组成。首先，利用空间GNN增强目标特征以获得更精确的目标区域；然后，利用语义GNN和LSTM融合外观特征、运动特征和对象特征，得到具有语义信息的潜在特征；最后，利用可以处理全局信息的解码器生成视频描述。

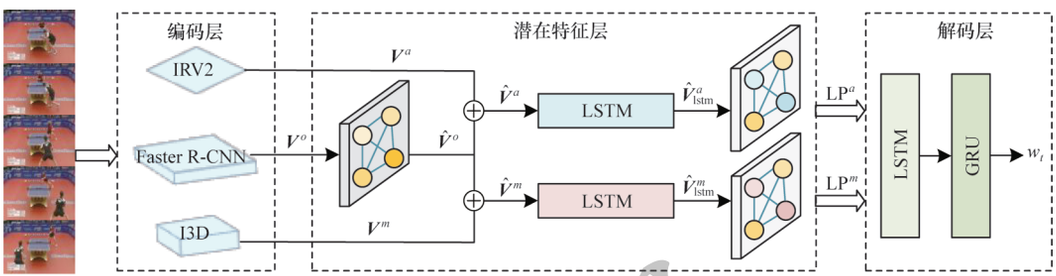


图2 LFAN模型框架

Fig.2 Framework of LFAN model

2.1 编码层

在编码阶段，本文使用3种预训练模型提取视频特征。对于给定的视频帧N，本文使用2D-CNNs和3D-CNNs分别提取外观特征$\boldsymbol{V}^a=\{\boldsymbol{v}_n^a\}_{n=1}^N$和运动特征$\boldsymbol{V}^m=\{\boldsymbol{v}_n^m\}_{n=1}^N$，然后使用R-CNNs提取区域目标特征$\boldsymbol{V}^o=\{\boldsymbol{v}_n^o\}_{n=1}^N$，区域目标特征包含空间上的额外维度。

2.2 潜在特征层

LFAN模型使用GNN融合不同的特征，得到潜在特征，利用GNN和LSTM实现潜在特征的增强，并得到更高维度的增强潜在特征，高维度的潜在特征蕴含着丰富的语义信息以生成更准确的视频描述。

对于先前生成的$\boldsymbol{V}^a$，$\boldsymbol{V}^m$和$\boldsymbol{V}^o$，本文首先利用动态显著区域图神经网络DyReg-GNN[29]对区域目标特征$\boldsymbol{V}^o$进行增强，DyReg-GNN可以通过学习发现与当前场景和目标相关的显著区域来改善视频的关系处理过程，增强后的$\hat{\boldsymbol{V}}^o$蕴含时空信息，如式(1)所示：

$$\hat{\boldsymbol{V}}^o=D_{\text{dyreg}}(\boldsymbol{V}^o) \tag{1}$$

其中：$D_{\text{dyreg}}()$表示DyReg-GNN中的图神经网络操作。

然后将增强后目标特征$\hat{\boldsymbol{V}}^o$的中间2维降为1维，再将特征$\boldsymbol{V}^i=\boldsymbol{V}^a,\boldsymbol{V}^m$分别和增强后的区域目标特征$\hat{\boldsymbol{V}}^o$由Softmax()函数计算关系矩阵权值：

$$\boldsymbol{W}_{\text{adj}}=\text{Softmax}(\hat{\boldsymbol{V}}^o/\hat{\boldsymbol{V}}^i) \tag{2}$$

其中：$\boldsymbol{W}_{\text{adj}}\in\mathbb{R}^d$表示可学习的参数；/代表矩阵点除。

为了让特征同时具有帧级的时间信息和对象级的空间信息，本文将特征$\boldsymbol{V}^i$和得到的关系矩阵相乘，相乘后的结果和区域目标特征拼接得到潜在特征：

$$\hat{\boldsymbol{V}}^a=\boldsymbol{V}^a\times\boldsymbol{W}_{\text{adj}} \tag{3}$$

$$\hat{\boldsymbol{V}}^m=\boldsymbol{V}^m\times\boldsymbol{W}_{\text{adj}} \tag{4}$$

$$\hat{\boldsymbol{V}}^i=\begin{cases}\text{add}(\hat{\boldsymbol{V}}^a,\hat{\boldsymbol{V}}^o)\\ \text{add}(\hat{\boldsymbol{V}}^m,\hat{\boldsymbol{V}}^o)\end{cases} \tag{5}$$

其中：$\hat{\boldsymbol{V}}^a$，$\hat{\boldsymbol{V}}^m$表示和图邻接矩阵相乘后的增强外观特征和运动特征；×代表矩阵乘法；$\hat{\boldsymbol{V}}^i$代表潜在特征；add()表示特征向量add拼接操作。

使用1个双向LSTM对潜在特征进行编码，将前一时刻的隐藏状态h_{t-1}作为输入：

$$\hat{\boldsymbol{V}}_{\text{lstm}}^i(h_t,c_t)=\text{LSTM}(\hat{\boldsymbol{V}}^i,(h_{t-1},c_{t-1})) \tag{6}$$

其中：$\hat{\boldsymbol{V}}_{\text{lstm}}^i$表示增强的潜在特征；$h_t$表示第$t$个时刻的隐藏状态；$c_t$表示第$t$个时刻的细胞状态。由于$h_t$具有丰富的历史信息，因此它对于增强潜在特征具有指导作用。

对增强的潜在特征$\hat{\boldsymbol{V}}_{\text{lstm}}^i$使用Transformer中的位置编码，保存特征之间的相对位置用于指导生成更流畅的描述语句，然后通过图神经网络将其融合为潜在特征并参与训练。外观和运动潜在特征如式(7)和式(8)所示：

$$\widehat{\text{LP}}^i=K^i(P_{\text{PE}}(\hat{\boldsymbol{V}}_{\text{lstm}}^i))\times P_{\text{PE}}(\hat{\boldsymbol{V}}_{\text{lstm}}^i) \tag{7}$$

$$\widehat{\text{LP}}^i=S_{\text{selfatt}}(\text{LP}^i) \tag{8}$$

其中：LP^i表示外观和运动潜在特征；$P_{\text{PE}}()$表示Transformer中的位置编码函数；$K()$表示kernel函数，里面是图神经网络模块，包含卷积和批量规范化操作以及GELU[30]激活函数；×表示矩阵乘法；$\widehat{\text{LP}}^i$表示最终的增强外观和运动潜在特征；$S_{\text{selfatt}}()$表示自注意力函数，后面还有1层LayerNorm函数。本文考虑到虽然ReLU[31]函数能够解决梯度消失，但是依然存在一些不可避免的问题，如无法避免梯度爆炸，神经网络无法调整学习率的值。因此，本文采用自然语言处理(NLP)领域最近表现较优的GELU作为激活函数，GELU在BERT和Transformer中也得到了很好的应用。

至此，LFAN模型完成帧级的外观特征和运动特

征同对象级目标特征的融合，从而生成具有时空动态信息的高级潜在特征。

2.3 解码层

本文参考 ORG-TRL[6] 并设计一种同时使用 LSTM 和 GRU 的解码方法。LFAN 模型通过注意力 LSTM 和 GRU 解码潜在特征层生成 $\widehat{\mathrm{LP}}_t^i$，从而逐渐生成最终的视频描述。

首先 LFAN 模型对生成的潜在外观特征和潜在运动特征进行均值操作，然后用 Cat 操作将它们拼接作为模型的全局视频特征：

$$\bar{\boldsymbol{V}}=\mathrm{Cat}(\mathrm{mean}(\widehat{\mathrm{LP}}^a),\mathrm{mean}(\widehat{\mathrm{LP}}^m)) \tag{9}$$

其中：$\bar{\boldsymbol{V}}$ 表示全局视频特征；Cat() 表示 Cat 拼接操作。

对于每个时间步长 t，LSTM 根据历史隐藏状态 h_{t-1}^{att}、历史细胞状态 c_{t-1}^{att} 与均值全局特征 $\bar{\boldsymbol{V}}$ 以及之前生成的单词 $\boldsymbol{w}_{t-1}$ 进行连接，历史隐藏状态和细胞状态的表达式如式（10）所示：

$$h_t^{\mathrm{att}},c_t^{\mathrm{att}}=\mathrm{LSTM}^{\mathrm{att}}([h_{t-1}^{\mathrm{lang}},\bar{\boldsymbol{V}},\boldsymbol{W}_e\boldsymbol{w}_{t-1}],(h_{t-1}^{\mathrm{att}},c_{t-1}^{\mathrm{att}})) \tag{10}$$

其中：h_t^{att} 表示 t 时刻的 LSTM 隐藏状态；h_{t-1}^{lang} 表示前一时刻语言 GRU 的隐藏状态；$\boldsymbol{W}_e$ 表示单词嵌入矩阵。

对于局部对象特征，LFAN 模型使用 DyReg-GNN 中的方法，首先将不同帧中的对象对齐并合并在一起，然后使用空间注意模块选择应该关注哪些对象，并提取局部上下文特征 $\boldsymbol{V}_t^c$。局部上下文特征的表达式如下：

$$\boldsymbol{V}_t^c=A_{\mathrm{ATT}}(\mathrm{Cat}(\widehat{\mathrm{LP}}^a,\widehat{\mathrm{LP}}^m)) \tag{11}$$

其中：$A_{\mathrm{ATT}}()$ 表示 DyReg-GNN 中空间注意模块。

最后，GRU 总结全局和局部上下文特征以生成当前隐藏状态 h_t^{lang}，这样本文生成描述时既有全局相关性也包含细粒度的上下文信息。在将单词概率 $\boldsymbol{P}_t$ 解码后是单层感知机和解码步骤 t 时刻的 Softmax() 运算。隐藏状态和单词概率的计算式如下：

$$h_t^{\mathrm{lang}}=\mathrm{GRU}^{\mathrm{lang}}(\mathrm{Cat}(\boldsymbol{V}_t^c,h_t^{\mathrm{att}}),h_{t-1}^{\mathrm{lang}}) \tag{12}$$

$$\boldsymbol{P}_t=\mathrm{Softmax}(\boldsymbol{W}_z h_t^{\mathrm{lang}}+b_z) \tag{13}$$

其中：$\boldsymbol{P}_t$ 表示词汇量的 D 维向量；$\boldsymbol{W}_z$ 表示权值矩阵；b_z 表示可学习的参数。

3 实验结果与分析

为合理评估该网络模型的有效性和先进性，本文在 2 个广泛使用的基准数据集 MSVD 和 MSR-VTT 上进行实验，并通过 4 个广泛使用的指标 BLUE@4、METEOR、ROUGE-L 和 CIDEr 进行评估，将该方法与最先进的方法进行比较，并进行消融实验。

3.1 数据集

MSVD 由 YouTube 收集的 1 970 个网络视频组成，平均视频长度为 10.2 s，每个视频大约有 41 个英文句子，每个描述平均长度约有 7 个单词。本文根据之前的工作[15]将数据集分为 1 200 个训练视频、100 个验证视频和 670 个测试视频。

MSR-VTT 数据集是开放领域视频字幕生成的大规模数据集，共包含 10 000 个视频，平均视频长度为 14.8 s，每个视频有 20 个人为标注的英文描述，每个描述的平均长度约为 9 个单词。本文采用标准分割将数据集分为 6 513 个训练视频、497 个验证视频和 2 990 个测试视频。

3.2 实验设置

本文在特征提取上使用预训练好的 Inception ResNetV2（IRV2）、I3D 和 Faster R-CNN 分别提取外观特征、动作特征和目标特征，每个视频采用 26 帧的均匀采样，Faster R-CNN 从固定的 26 帧中提取 36 个 proposal。对于语料库的预处理，本文将生成的所有描述转换为小写并去掉标点符号，最大词汇量设置为 26 个单词，对超过 26 个单词的描述进行零填充。本文将预训练 GloVe.6B.300d 词表引入到解码器参与词向量训练，词向量维度为 300。

本文用标准的交叉熵损失函数计算模型生成的描述和 Ground Truth 间的差异，采用 Adam 优化器优化 LFAN 模型，初始学习率设为 1×10^{-4}，动态调整学习率使其每 5 轮削减 50%。训练和测试批量大小分别设为 256 和 128，最大训练迭代轮次设为 60 次。在 MSVD 和 MSR-VTT 数据集上，所有 LSTM 模块隐藏状态大小分别设为 1 024 和 1 536，每个图卷积操作的特征大小为 1 024。在测试阶段本文分别使用大小为 4 和 5 的波束搜索来生成描述。

3.3 实验结果定量分析

为验证 LFAN 模型的有效性，本文选择使用 CNN 作为编码器和 LSTM 作为解码器，在 MSVD 和 MSR-VTT 2 个数据集上与最先进的方法进行比较。

在 MSVD 和 MSR-VTT 数据集上不同模型的实验结果如表 1 所示，其中，B@4、M、R、C 分别表示 BLUE@4、METEOR、ROUGE-L 和 CIDEr，加粗表示最优数据。从表 1 可以看出，LFAN 具有较强的竞争优势，在 MSVD 数据集上，反映描述准确性的 BLEU@4 分数为 57，反映描述丰富性的 CIDEr 分数达到了 100.1，在 MSR-VTT 数据集上，BLEU@4 分数为 43.8，CIDEr 分数为 50.2，在多个指标上都优于主流视频描述生成方法，证明 LFAN 模型的有效性。

表1 在MSVD和MSR-VTT数据集上不同模型的实验结果

Table 1 Experimental results among different models on MSVD and MSR-VTT datasets

方法	特征			MSVD数据集				MSR-VTT数据集			
	Context	Motion	Object	B@4	M	R	C	B@4	M	R	C
RecNet	InceptionV4	—	—	52.3	34.1	69.8	80.3	39.1	26.6	59.3	42.7
PickNet[18]	ResNet-152	—	—	52.3	33.3	69.6	76.5	41.3	27.7	59.8	44.1
MARN[20]	ResNet-101	C3D	—	48.6	35.1	71.9	92.2	40.4	28.1	60.7	47.1
OA-BTG[3]	ResNet-200	—	Mask R-CNN	56.9	36.2	—	90.6	41.4	28.2	—	46.9
GRU-EVE[27]	InceptionResNetV2	C3D	YOLO	47.9	35.0	71.5	78.1	38.3	28.4	60.7	48.1
POS+CG[12]	InceptionResNetV2	OpticalFlow	—	52.5	34.1	71.3	88.7	42.0	28.2	61.6	48.7
RMN[21]	InceptionResNetV2	I3D	Faster R-CNN	54.6	36.5	73.4	94.4	42.5	28.4	61.6	49.6
STG-KD[4]	ResNet-101	I3D	Faster R-CNN	52.2	36.9	73.9	93.0	40.5	28.3	60.9	47.1
SAAT[11]	InceptionResNetV2	C3D	Faster R-CNN	46.5	33.5	69.4	81.0	40.5	28.2	60.9	49.1
ORG-TRL[28]	InceptionResNetV2	C3D	Faster R-CNN	54.3	36.4	73.9	95.2	43.6	28.8	62.1	50.9
SGN[23]	ResNet-101	C3D	—	52.8	35.5	72.9	94.3	40.8	28.3	60.8	49.5
Open-Book[24]	InceptionResNetV2	C3D	—	—	—	—	—	42.8	**29.3**	61.7	**52.9**
LFAN	InceptionResNetV2	I3D	Faster R-CNN	**57.0**	**36.9**	**74.1**	**100.1**	**43.8**	28.8	**62.1**	50.2

在MSR-VTT数据集上，与不使用对象特征的RecNet、PickNet、MARN、SGN和Open-Book相比，LFAN仅略逊于Open-Book，其原因为Open-Book在生成关键词时从文本语料库中检索多个与视频内容相关的句子，生成的关键词与参考语句在生成关键词时，会从文本语料库中检索多个与视频内容相关的句子，因此生成的关键词与参考语句的相似度更高。这种方法在METEOR和CIDEr评价指标中会获得更高的得分。而在MSVD数据集上，LFAN在所有评价指标上都取得比其他方法更优的性能，表明对象特征在视频描述生成中发挥了重要作用，并且学到准确的对象特征。

此外，LFAN与使用对象特征的OA-BTG、GRU-EVE、RMN、STG-KD、SAAT和ORG-TRL进行比较。在MSR-VTT数据集上，当ORG-TRL引入TRL外部语言模块来指导模型生成描述语句时，ORG-TRL的CIDEr得分增加为50.9，当ORG-TRL去掉TRL外部语言模块后，在CIDEr上的表现不如本文模型，得分为50.1。本文提出的LFAN在BLUE@4和ROUGE-L中有更好的表现，表明LFAN生成的视频描述准确度和召回率更高。

3.4 消融实验

本文主要对潜在特征模块和解码模块进行改进。为了说明本文的改进措施能使模型学到更有效的信息以生成视频描述，本文在潜在特征模块上设计3个消融实验，分别是仅使用外观特征、运动特征和对象特征来生成视频描述的基线模型。

表2和表3所示为使用不同神经网络和不同解码方法的消融实验结果。LFAN-GNN表示使用图神经网络融合不同特征，LFAN-DG表示使用DyReg-GNN加强目标特征，LFAN-LSTM和LFAN-GRU分别是仅使用LSTM和GRU作为解码器。从表2可以看出，无论是使用图神经网络融合不同特征还是加入DyReg-GNN后，模型的各项指标都有所提升。相比LFAN-GNN，LFAN-DG在2个数据集上的BLUE@4分别提升了1.9和0.6，说明本文的改进方法使模型提取到更准确的对象信息。本文在MSVD数据集上的CIDEr分数比基线模型提高9.8，在MSR-VTT数据集上比基线模型提高了3.1的分数，进一步证明LFAN的有效性。

表2 使用不同图神经网络模块的消融实验结果

Table 2 Results of ablation experiments using different graph neural network modules

方法	MSVD				MSR-VTT			
	B@4	M	R	C	B@4	M	R	C
Baseline	50.6	35.3	72.1	90.3	41.1	27.9	60.5	47.1
LFAN-GNN	53.7	36.0	72.6	96.1	42.9	28.1	60.9	49.0
LFAN-DG	55.6	36.3	73.4	97.3	43.5	28.3	61.5	49.6
LFAN	57.0	36.9	74.1	100.1	43.8	28.8	62.1	50.2

表3 使用不同解码方法的消融实验结果

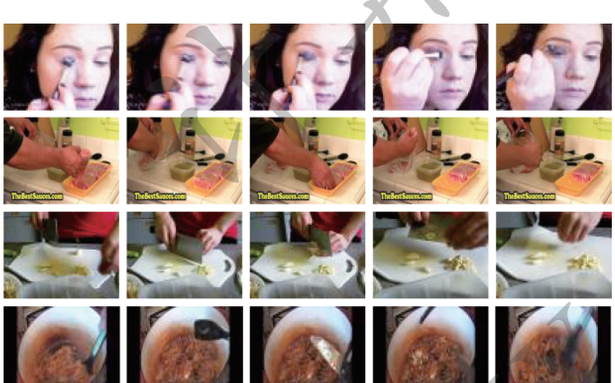
Table 3 Results of ablation experiments using different decoding methods

方法	MSVD				MSR-VTT			
	B@4	M	R	C	B@4	M	R	C
LFAN-LSTM	55.8	36.4	73.4	97.5	43.4	28.4	60.9	49.5
LFAN-GRU	53.6	35.2	73.2	98.9	43.5	28.3	61.4	49.3
LFAN	57.0	36.9	74.1	100.1	43.8	28.8	62.1	50.2

从表3可以看出，本文设计同时使用LSTM和GRU的LFAN显然比单独使用其中1个解码方法的性能更好，新的解码方法与LFAN-LSTM相比评估效率也得到了改善，这充分证明了本文改进方法的有效性。

3.5 实验结果定性分析

图 3 所示为 LFAN 生成的一些描述实例与参考描述(GT)的对比。图 3 中第 1 行参考视频描述:GT1“a woman is applying something on her eyelids”;GT2“a girl is applying eye makeup”;GT3“a girl is applying makeup to her eyelid”;LFAN“a woman is applying makeup on her eye”。第 2 行参考视频描述:GT1“the man is putting meat in the bag”;GT2“a man is adding chicken to a plastic cover”;GT3“a man puts chicken breasts into a bag”;LFAN“a man is putting chicken into a plastic”。第 3 行参考视频描述:GT1“a man is dicing food”;GT2“a man is slicing garlic”;GT3“a person is slicing garlic”;LFAN“a man is chopping garlic”。第 4 行参考视频描述:GT1“a woman is cooking”;GT2“a woman showing how to cut garlic cloves”;GT3“a woman is chopping garlic”;LFAN“a person is preparing some food in the kitchen”。LFAN 可以精准识别出“woman”在“applying makeup”,而不是“draw something”。在第 2 行的示例中,LFAN 成功地识别出主要对象信息“chicken”和“plastic”以及人物的动作“putting”,并且排除掉桌子上其他干扰对象信息,说明 LFAN 不仅可以识别出主要对象,并且可以精准地描述对象动作。



(a)原图 (b)GT1 (c)GT2 (d)GT3 (e)LFAN

图 3 LFAN 生成描述与参考描述实例分析

Fig.3 Example analysis of LFAN generation description and reference description

4 结束语

视频描述生成技术可以广泛应用于各种媒体软件,在视频推荐、辅助视觉、人机交互等领域也具有广泛应用前景[32]。本文提出一种基于潜在特征增强网络的视频描述生成模型 LFAN。该模型着重于增强视频特征的时空和语义信息,从而显著提升生成的视频描述质量。大量的定量、定性实验和消融实验结果都证明了 LFAN 的有效性,LFAN 模型能够精准地描述对象动作。由于在生成描述中一些视频的描述难以被模型正确地生成,这种情况尤其发生在一些罕见或复杂的场景或物体上,因此后续将基于多模态融合和 KL 散度对 LFAN 进行分析研究。

参考文献

[1] 付燕,马钰,叶鸥. 融合深度学习和视觉文本的视频描述方法[J]. 科学技术与工程,2021,21(14):5855-5861.
FU Y,MA Y,YE O. Video captioning method combining deep networks and visual text[J]. Science Technology and Engineering,2021,21(14):5855-5861. (in Chinese)

[2] 汤鹏杰,王瀚漓. 从视频到语言:视频标题生成与描述研究综述[J]. 自动化学报,2022,48(2):375-397.
TANG P J,WANG H L. From video to language:survey of video captioning and description[J]. Acta Automatica Sinica,2022,48(2):375-397. (in Chinese)

[3] ZHANG J C,PENG Y X. Object-aware aggregation with bidirectional temporal graph for video captioning[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2019:8319-8328.

[4] PAN B X,CAI H Y,HUANG D A,et al. Spatio-temporal graph for video captioning with knowledge distillation[C]// Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C. ,USA:IEEE Press,2020:1-10.

[5] CARION N,MASSA F,SYNNAEVE G,et al. End-to-end object detection with Transformers[C]//Proceedings of European Conference on Computer Vision. Berlin,Germany:Springer,2020:213-229.

[6] SZEGEDY C,IOFFE S,VANHOUCKE V,et al. Inception-V4,Inception ResNet and the impact of Residual connections on learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]:AAAI Press,2017:1-10.

[7] CARREIRA J,ZISSERMAN A. Quo vadis,action recognition? a new model and the kinetics dataset[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2017:6299-6308.

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

[9] XU Y J,HAN Y H,HONG R C,et al. Sequential video VLAD:training the aggregation locally and temporally[J]. IEEE Transactions on Image Processing,2018,27(10):4933-4944.

[10] WANG H Y,XU Y J,HAN Y H. Spotting and aggregating salient regions for video captioning[C]//Proceedings of the 26th International Conference on Multimedia. New York,USA:ACM Press,2018:1519-1526.

[11] ZHENG Q,WANG C Y,TAO D C. Syntax-aware action targeting for video captioning [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2020:13096-13105.

[12] WANG B R,MA L,ZHANG W,et al. Controllable video captioning with POS sequence guidance based on gated fusion network[C]//Proceedings of International Conference

on Computer Vision. Washington D. C. ,USA:IEEE Press, 2019:1-10.
[13] 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.
[14] BARBU A, BRIDGE A, BURCHILL Z, et al. Video in sentences out[EB/OL]. [2023-02-15]. https://arxiv. org/pdf/1204. 2742. pdf.
[15] DAS P, XU C L, DOELL R F, et al. A thousand frames in just a few words: lingual description of videos through latent topics and sparse object stitching[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press, 2013:2634-2641.
[16] VENUGOPALAN S, XU H, DONAHUE J, et al. Translating videos to natural language using deep recurrent neural networks [EB/OL]. [2023-02-15]. http://arXiv preprint arXiv:1412. 4729, 2014.
[17] YAO L, TORABI A, CHO K, et al. Describing videos by exploiting temporal structure[C]//Proceedings of International Conference on Computer Vision. Washington D. C. ,USA: IEEE Press, 2015:4507-4515.
[18] CHEN Y Y, WANG S H, ZHANG W G, et al. Less is more: picking informative frames for video captioning [C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2018:367-384.
[19] WANG J B, WANG W, HUANG Y, et al. M3: multimodal memory modelling for video captioning[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press, 2018: 7512-7520.
[20] PEI W J, ZHANG J Y, WANG X R, et al. Memory-attended recurrent network for video captioning[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press, 2019:8347-8356.
[21] TAN G C, LIU D Q, WANG M, et al. Learning to discretely compose reasoning module networks for video captioning [EB/OL]. [2023-02-15]. https://arxiv. org/abs/2007. 09049v1.
[22] BAI Y, WANG J Y, LONG Y, et al. Discriminative latent semantic graph for video captioning[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM Press, 2021:3556-3564.
[23] RYU H, KANG S, KANG H, et al. Semantic grouping network for video captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2021:2514-2522.
[24] ZHANG Z Q, QI Z A, YUAN C F, et al. Open-Book video captioning with retrieve-copy-generate network [C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press, 2021: 9837-9846.
[25] CHEN J W, PAN Y W, LI Y H, et al. Retrieval augmented convolutional encoder-decoder networks for video captioning [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(1s):1-24.
[26] LIN K, LI L J, LIN C C, et al. SwinBERT: end-to-end Transformers with sparse attention for video captioning[C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press, 2022: 17949-17958.
[27] AAFAQ N, AKHTAR N, LIU W, et al. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA:IEEE Press, 2019:12487-12496.
[28] ZHANG Z Q, SHI Y Y, YUAN C F, et al. Object relational graph with teacher-recommended learning for video captioning[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA: IEEE Press, 2020:13278-13288.
[29] DUTA I, NICOLICIOIU A L, LEORDEANU M. Discovering dynamic salient regions for spatio-temporal graph neural networks [C]//Proceedings of the 35th Conference on Neural Information Processing Systems. New York, USA: [s. n], 2021:1-10.
[30] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs)[EB/OL]. [2023-02-15]. https://arxiv. org/abs/1606. 08415v4.
[31] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press, 2016:770-778.
[32] 侯静怡,齐雅昀,吴心筱,等. 跨语言知识蒸馏的视频中文字幕生成[J]. 计算机学报, 2021, 44(9):1907-1921. HOU J Y, QI Y Y, WU X X, et al. Cross-lingual knowledge distillation for Chinese video captioning [J]. Chinese Journal of Computers, 2021, 44 (9): 1907-1921. (in Chinese)

编辑 薛晋栋