基于增强对象学习和注意力网络的视频描述方法

蔡晓东，龙顺宏，梁焜峻

（桂林电子科技大学 信息与通信学院，广西 桂林 541004）

摘　要：在视频描述任务中，常见问题之一是对象描述不够具体，主要原因在于模型没有充分学习视频中的对象信息。同时，视频包含了丰富的特征信息，如对象信息、运动信息和上下文信息，这使得如何提升模型在生成描述时学习关键信息的能力成为一项具有挑战性的任务。为解决上述问题，本文提出了一种基于增强对象学习和注意力网络的方法。首先，设计了一种新的增强对象学习模块，旨在充分学习视频中的对象信息，从而实现对视频内容的准确描述；其次，构建了一种注意力网络，致力于有效关注不同类型的信息，以提升模型在生成描述时学习关键信息的能力。在MSVD和MSR-VTT数据集上的实验中，本文方法生成的描述展现出更高的具体性和准确性，同时在各项评价指标上均超过了目前的先进方法，有效验证了该方法的可行性。

关键词：深度学习；视频描述；增强对象学习；注意力网络

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1671-5497(2026)02-0516-07

DOI：10.13229/j.cnki.jdxbgxb.20240251

Video captioning method based on enhanced object learning and attention networks

CAI Xiao-dong，LONG Shun-hong，LIANG Kun-jun

（*School of Information and Communication*，*Guilin University of Electronic Technology*，*Guilin* 541004，*China*）

Abstract：In video captioning tasks，one of the common problems is that the object caption is not specific enough，mainly because the model does not fully learn the information of the objects in the video. Meanwhile，videos contain abundant feature information，such as object information，motion information，and contextual information，making it a challenging task to enhance the model's ability to learn key information when generating captions. To address the aforementioned problems，this paper proposes a method based on enhanced object learning and attention networks. Firstly，a new enhanced object learning module was designed to fully learn object information in videos，thereby achieving accurate caption of video content. Secondly，an attention network was constructed to dynamically adjust the weights of different types of information，thereby enhancing the model's ability to learn key information when generating captions. In the experiments on the MSVD and MSR-VTT datasets，the caption generated by the method

收稿日期：2024-03-12.

基金项目：广西创新驱动发展专项项目（AA20302001）.

作者简介：蔡晓东（1971-），男，教授，博士. 研究方向：数据挖掘. E-mail：caixiaodong@guet.edu.cn

proposed in this paper showed a higher level of specificity and accuracy, and exceeded the current advanced methods in various evaluation indicators, effectively verifying the feasibility of the method.

Key words: deep learning; video captioning; enhanced object learning; attention network

0　引　言

视频描述是将视频内容转换为文字形式,以传达视频所展现的对象、动作或事件。近年来,视频描述领域取得了显著进展,但仍面临挑战。首先,描述对象的多样性和复杂性导致其在被描述过程中有时过于模糊,缺乏清晰的指代。其次,不同类型特征之间的异构性和融合策略的选择是一个具有挑战性的难题。

在视频描述任务中,对象建模十分重要[1]。之前的许多研究工作[2,3]都使用预训练的对象检测模型来生成对象特征,然而,这些方法忽略了视频帧内和视频帧之间的对象之间的关系。特别是当存在多个对象时,模型难以准确描述这些对象的信息。为解决这一问题,一些研究团队[4,5]通过构建关系图以捕捉对象之间的动态关系,从而生成更为准确、丰富的描述。与此同时,Kanani等[6]使用全局对象信息提升了模型生成段落描述的性能。这些方法在实验中证明了充分学习对象信息及其交互关系的重要性。因此,受Transfomer[7-12]在视频描述领域取得的突出成绩的启发,本文设计一种新的增强对象学习模块(Enhanced object learning module, EOLM),该模块基于Transformer架构,引入门控机制,形成了Gated Transformer。EOLM利用注意力机制聚焦关键对象信息,并通过门控机制建模不同对象之间的关系,以获得精确的对象特征表示,从而实现对视频内容的准确描述。

此外,由于视频涵盖多种特征信息,因此采用单一类型特征的方法很难生成准确的视频描述[13]。在一些注重特征建模的视频描述生成任务中,简单的拼接方法虽然丰富了特征信息,但也带来了信息冗余和噪声干扰等问题。针对这一问题,一些研究方法[14,15]利用门控机制实现了特征的融合,有效提升了特征的表达能力。在借鉴了文献[16,17]的研究成果后,本文构建了一种有效的注意力网络(Attention networks, AN)。该网络通过注意力机制自适应地融合不同类型特征的特征,使模型能有效学习视频的关键信息,进而提升视频描述的质量。

1　EOLM-AN模型设计

EOLM-AN模型的整体框架如图1所示,由特征提取、特征学习、特征融合和特征解码等模块组成。首先,特征提取模块利用R-CNN、2D CNN和3D CNN对输入视频进行编码,分别得到对象特征$\boldsymbol{O}$、上下文特征$\boldsymbol{C}$和运动特征$\boldsymbol{A}$;其次,特征学习模块进一步处理以上3种特征:通过增强对象学习模块捕获对象之间的长期依赖关系,得到更准确的对象特征$\boldsymbol{H}_{\mathrm{o}}$;同时,利用两个BiLSTM分别编码上下文特征$\boldsymbol{C}$和运动特征$\boldsymbol{A}$,得到$\boldsymbol{H}_{\mathrm{c}}$和$\boldsymbol{H}_{\mathrm{a}}$。再次,特征融合模块利用注意力网络将特征学习模块输出的3种特征自适应地融合,得到融合特征$\boldsymbol{H}$。最后,将融合特征$\boldsymbol{H}$输入特征解码模块中的解码器以生成文本描述w_{t}。

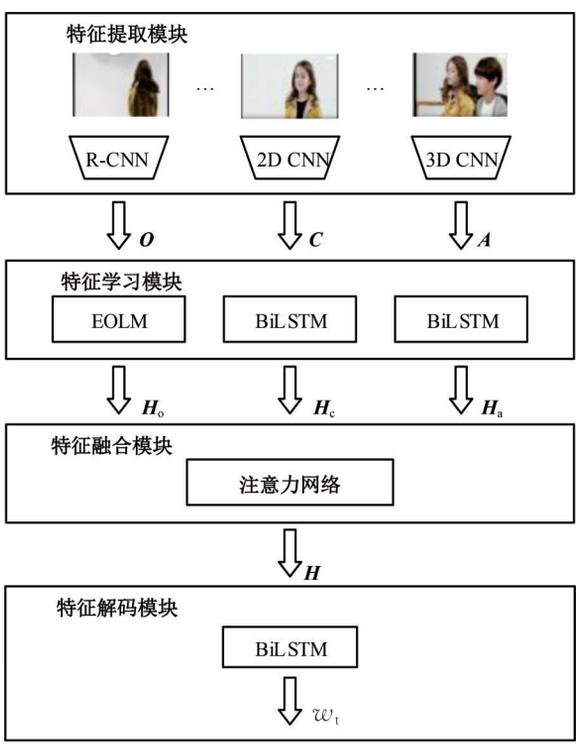


图1　EOLM-AN模型的整体框架

Fig. 1　Overall framework of the EOLM-AN model

1.1　EOLM设计

针对模型难以充分学习对象信息的问题,本文设计了一种基于Gated Transformer的增强对

象学习模块(如图2所示),旨在进一步编码初始对象特征O,以实现对视频中对象信息的深层理解。

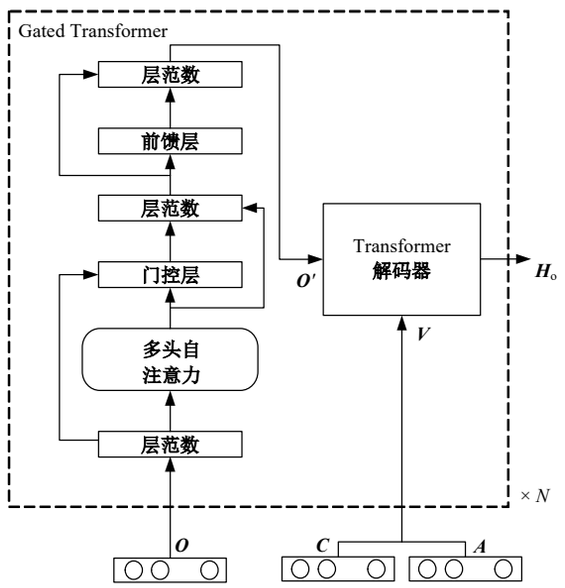


图2 EOLM结构图

Fig. 2 Structure of EOLM

首先,增强对象学习模块使用编码器对初始对象特征O进行处理,计算公式为:

$$O'=G\mathrm{TransEncoder}(O) \tag{1}$$

式中:GTransEncoder为Gated Transformer编码器,它能有效捕捉对象之间的关系和上下文信息。

具体来说,该编码器利用层范数和多头自注意力机制,根据不同对象之间的相关性自适应地进行加权处理,以产生更精确的特征表示H:

$$H=MultiHeadAtt(LayerNorm(O)) \tag{2}$$

其次,通过使用门控层、层范数和残差连接,对特征表示H和初始对象特征O进行处理,以得到隐藏状态H':

$$H'=Gating\,Layer(H,LayerNorm(O))+H \tag{3}$$

假设将输入视频的两个不同特征拼接得到的特征表示分别为x和y,则相关的公式为:

$$r=\mathrm{sigmoid}(\boldsymbol{W}_{r1}x+\boldsymbol{W}_{r2}y+\boldsymbol{b}_r) \tag{4}$$

$$z=\mathrm{sigmoid}(\boldsymbol{W}_{z1}x+\boldsymbol{W}_{z2}y+\boldsymbol{b}_z) \tag{5}$$

$$g=\tanh(\boldsymbol{W}_{g1}y+\boldsymbol{W}_{g2}(r\odot x)) \tag{6}$$

$$H=(1-z)\odot x+z\odot g \tag{7}$$

式中:sigmoid为一种S型激活函数;tanh为双曲正切函数;$\odot$为哈达玛积;$\boldsymbol{W}_{r1}$,$\boldsymbol{W}_{r2}$,$\boldsymbol{W}_{z1}$,$\boldsymbol{W}_{z2}$,$\boldsymbol{W}_{g1}$和$\boldsymbol{W}_{g2}$为可学习的参数;$\boldsymbol{b}_r$和$\boldsymbol{b}_z$为可学习的偏置。

然后,通过使用前馈层、层范数和残差连接,对隐藏状态H'进一步编码,形成输出O':

$$O'=FeedForward(LayerNorm(H'))+H' \tag{8}$$

最后,将C和A级联得到视频内容的特征V,并将其和O'馈送到Gated Transformer解码器,得到对象特征H_o:

$$H_o=G\mathrm{TransDecoder}(O',V) \tag{9}$$

1.2 AN设计

为了充分利用不同类型的特征,本文在解码器之前,构建了一种注意力网络。该结构能有效关注不同类型的特征,从而提升模型在生成描述时学习关键信息的能力。

如图3所示,注意力网络的输入包括$H_o=\{o_t\}_{t=1}^T$、$H_c=\{c_t\}_{t=1}^T$和$H_a=\{a_t\}_{t=1}^T$,其中T为视频帧数。该网络通过注意力机制关注了3种不同类型特征的关键信息,具体操作如下。

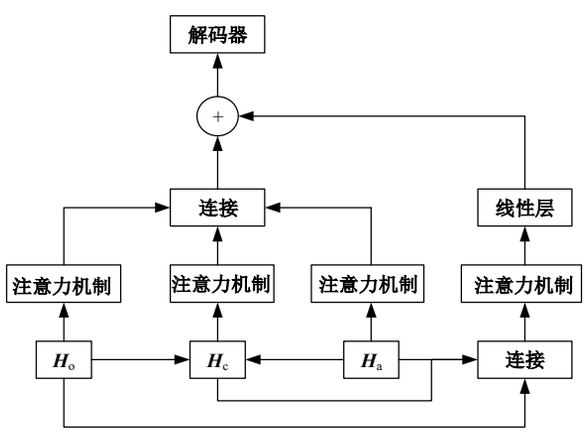


图3 AN结构图

Fig. 3 Structure of AN

首先,为了使网络可以关注到每一种特征的关键信息,本文使用注意力机制分别处理特征向量H_o、H_c和H_a,然后将它们级联以获得更丰富的特征表示H_x:

$$H_x=[Att(H_o);Att(H_c);Att(H_a)] \tag{10}$$

式中:Att为注意力机制。

以$Att(H_o)$为例,计算注意力权重β的过程为:

$$Att(H_o)=\{\bar{o}_t\}_{t=1}^T \tag{11}$$

$$\bar{o}_t=\sum_{i=1}^N\beta_{t,i}o_i \tag{12}$$

$$\beta_{t,i}=\exp(\bar{\beta}_{t,i})/\sum_{i=1}^N\exp(\bar{\beta}_{t,i}) \tag{13}$$

$$\bar{\beta}_{t,i}=\boldsymbol{W}_{o1}\tanh(\boldsymbol{W}_{o2}h_{t-1}+\boldsymbol{W}_{o3}o_i+\boldsymbol{b}_{o1}) \tag{14}$$

式中:$\beta_{t,i}$为对象特征H_o的第i个输入o_i在t时刻的权重;h_{t-1}为解码器在$t-1$时刻的隐藏状态;$\boldsymbol{W}_{o1}$,$\boldsymbol{W}_{o2}$,$\boldsymbol{W}_{o3}$为可学习的参数;$\boldsymbol{b}_{o1}$为可学习的

偏置。

另外，为了有针对性地学习不同类型特征之间的重要性权重，本文将 $\boldsymbol{H}_o$、$\boldsymbol{H}_c$ 和 $\boldsymbol{H}_a$ 级联，然后通过使用注意力机制和线性层对其进一步处理，得到 $\boldsymbol{H}_y$：

$$\boldsymbol{H}_y=\{\boldsymbol{W}_{y4}(\bar{y}_t)\}_t^T \tag{15}$$

$$\bar{y}_t=\sum_{i=1}^{M}\gamma_{t,i}\,y_i \tag{16}$$

$$\gamma_{t,i}=\exp(\bar{\gamma}_{t,i})/\sum_{i=1}^{M}\exp(\bar{\gamma}_{t,i}) \tag{17}$$

$$\bar{\gamma}_{t,i}=\boldsymbol{W}_{y1}\tanh(\boldsymbol{W}_{y2}h_{t-1}+\boldsymbol{W}_{y3}y_i+\boldsymbol{b}_{y1}) \tag{18}$$

式中：$\gamma_{t,i}$ 为第 i 个输入 y_i 在 t 时刻的权重；$\boldsymbol{W}_{y1}$，$\boldsymbol{W}_{y2}$，$\boldsymbol{W}_{y3}$ 和 $\boldsymbol{W}_{y4}$ 为可学习的参数；$\boldsymbol{b}_{y1}$ 为可学习的偏置。

最后，为了可以综合利用两者的信息，以实现更优化的特征效果，本文将特征 $\boldsymbol{H}_x$ 与特征 $\boldsymbol{H}_y$ 融合得到特征 $\boldsymbol{H}_z$：

$$\boldsymbol{H}_z=\boldsymbol{H}_x+\boldsymbol{H}_y \tag{19}$$

2　实验结果与分析

2.1　实验数据

本文采用了具有代表性的 MSVD 和 MSR-VTT 数据集进行实验，以验证 EOLM-AN 模型在视频描述任务上的有效性。这两个数据集都是广泛应用的视频描述数据集，涵盖了各种不同的对象、动作和场景，并且每个视频都有多个人工描述。这为本文的研究提供了丰富多样的数据资源，能够确保对 EOLM-AN 模型在不同情境下的表现进行全面评估。

本文的实验对 MSVD 和 MSR-VTT 数据集样本进行了分组。MSVD 数据集被划分为训练、验证和测试集，分别包含 1 200、100 和 670 个视频。MSR-VTT 数据集也被划分为训练、验证和测试集，分别包含 6 153、497 和 2 990 个视频。

2.2　评价指标

本文实验采用了 4 个被广泛使用的视频描述任务评估指标，包括 BLEU@4、METEOR、CIDEr 和 ROUGE-L。其中，BLEU@4 用于衡量生成的描述与参考描述之间的相似性，METEOR 适用于评估描述的多样性和准确性，CIDEr 重点关注描述的细节和丰富性，而 ROUGE-L 考虑了召回率，能够评估描述的准确性。这些指标的综合运用能够全面评估 EOLM-AN 模型在视频描述任务上的性能表现。

2.3　实验平台与参数设置

本文在 Ubuntu 20.04.6 LTS 操作系统下进行了实验，使用了配备有 Intel Core i9-10900X CPU @3.70 GHz×20 和 NVIDIA GeForce RTX2080 Ti 等硬件配置的计算机。实验中采用了 PyTorch 深度学习框架和 Python 编程语言进行模型的实现和验证。

针对 MSVD 数据集，实验采用了以下参数设置：增强对象学习模块的层数为 2，BiLSTM 的隐藏层维度为 512，学习率为 5e-5，dropout 为 0.2，batch size 为 32，epoch 为 30。

针对 MSR-VTT 数据集，实验采用了以下参数设置：增强对象学习模块的层数为 3，BiLSTM 的隐藏层维度为 512，学习率为 7e-5，dropout 为 0.1，batch size 为 64，epoch 为 25。

2.4　模型性能评估

为验证 EOLM-AN 模型的有效性，本文选择了几种具有代表性的视频描述方法进行了性能比较，包括 DMRM、ORG-TRL、POS-CG、LSRT、MA-LSTM、VADD、SwinBERT、TextKG、MAN、ViT/L14 和 HMN。具体的实验结果如表 1 所示，其中 B4、M、C 和 R 分别为 BLEU@4、METEOR、CIDEr 和 ROUGE-L 的简写。

一方面，相较于 DMRM 和 ORG-TRL 模型，EOLM-AN 模型更注重增强对象的学习，因此在生成描述时更能准确地反映视频内容，各项指标上都表现更优。与 POS-CG、LSRT、MA-LSTM 和 VADD 模型相比，EOLM-AN 模型更注重不同类型特征之间的关系，并通过注意力网络更有效地捕捉关键的特征信息，从而提升了模型对特征信息的学习能力，进而在各项指标上都取得了显著的性能提升。与基准模型 HMN 相比，EOLM-AN 模型在各项指标上均取得了显著提升，进一步证明了其在生成视频描述时充分学习对象信息、理解视频关键信息的能力，从而提高了描述的准确性和语义的丰富性。

另一方面，EOLM-AN 模型在 B4 和 R 指标上优于 SwinBERT、TextKG、MAN 等先进的方法，这归因于 EOLM-AN 模型生成的描述准确性更高，参考描述之间更相似。但是，在 M 和 C 指标上，EOLM-AN 模型低于 SwinBERT 和 ViT/L14，这体现了 EOLM-AN 模型在生成描述的丰

表1 在MSVD和MSR-VTT数据集上与最先进的方法进行比较

Table 1 Compare with state-of-the-art methods on the MSVD and MSR-VTT datasets %

方法	MSVD				MSR-VTT			
	B4	M	C	R	B4	M	C	R
DMRM[3]	51.1	33.6	74.8	—	—	—	—	—
ORG-TRL[5]	54.3	36.4	95.2	73.9	43.6	28.8	50.9	62.1
POS-CG[14]	52.5	34.1	88.7	71.3	42.0	28.2	48.7	61.6
LSRT[15]	55.6	37.1	98.5	73.5	42.6	28.3	49.5	61.0
MA-LSTM[16]	52.3	33.6	70.4	—	36.5	26.5	41.0	59.8
VADD[17]	51.5	34.8	72.1	91.5	42.4	28.2	49.7	61.7
SwinBERT[9]	58.2	41.3	120.6	77.5	41.9	29.9	53.8	62.1
TextKG[10]	60.8	38.5	105.2	75.1	43.7	29.6	52.4	62.4
MAN[11]	60.1	37.1	101.9	74.6	42.5	28.6	50.4	62.2
ViT/L14[12]	60.1	41.4	121.5	78.2	44.4	30.3	57.2	63.4
HMN[8]	59.2	37.7	104.0	75.1	43.5	29.0	51.5	62.7
EOLM-AN	61.5	39.0	106.7	77.5	44.2	29.4	52.1	63.6

富性和多样性方面还有待提升。

2.5 消融实验

为验证本文所提出的EOLM和AN对模型性能的影响,本文在MSVD和MSR-VTT数据集上进行了消融实验,并将实验结果总结在表2中。

表2 在MSVD和MSR-VTT数据集上的消融实验

Table 2 Ablation experiments on the MSVD and MSR-VTT datasets %

方法	MSVD				MSR-VTT			
	B4	M	C	R	B4	M	C	R
EOLM	61.3	38.1	105.3	77.2	44.1	29.1	51.7	63.4
AN	60.5	38.5	106.2	76.2	43.7	29.3	51.9	62.9
EOLM-AN	61.5	39.0	106.7	77.5	44.2	29.4	52.1	63.6

首先,本文对EOLM模块进行了单独验证,实验结果显示在表2的“EOLM”部分。相较于HMN模型,EOLM在B4和R指标上表现出更显著的提升。这表明EOLM在学习对象信息方面进行了优化,从而使其生成的描述不仅准确,且更符合参考文本的内容。

其次,本文对AN模块进行了单独验证,实验结果显示在表2的“AN”部分。与HMN模型相比,AN在所有评价指标上取得了显著的提升。这表明AN在探索特征信息方面做出的改进,使其能够更有效地学习视频的关键信息,并生成更流畅、更具语义丰富度的视频描述。

综上所述,EOLM和AN模块在视频描述生成任务上展现了卓越的优势,有效提升了模型的性能。

2.6 实例分析

为了直观展示本文所提出的EOLM-AN模型相较于基准模型HMN的改进效果,本文在MSVD和MSR-VTT数据集上选取了几个实例进行了可视化,并将结果展示在图4中,其中GT表示参考描述。

从图4(a)~(d)中可以看出,EOLM-AN模型在生成视频描述方面表现出几个方面的优势。首先,在图4(a)(b)中,相较于HMN模型将视频中倒入锅里的液体和女孩手里所拿的东西识别为“water”和“ball”,EOLM-AN模型能准确识别为“oil”和“egg”。这表明EOLM-AN模型能更有效地学习视频中的对象信息,并生成更具体和准确的描述。其次,在图4(c)(d)中,HMN模型生成的描述是“a monkey is fighting(一只猴子在打架)”和“two girls are playing with toys(两个女孩在玩玩具)”,而EOLM-AN模型生成的描述是“a monkey is doing karate(一只猴子在练空手道)”和“a girl is painting eggs(一个女孩正在画鸡蛋)”。这表明EOLM-AN模型能更准确地理解视频内容,并生成更符合视频内容的语义描述。

从图4(e)(f)中可以观察到,EOLM-AN模型在处理长时间跨度视频时,能够捕捉到每个动作的一些特征。然而,由于缺乏足够的上下文信息,该模型无法准确理解这些动作之间的关系。因此,它只能生成局部动作的描述,比如“a man is slicing a lemon(一个男人正在切柠檬)”和“a per-

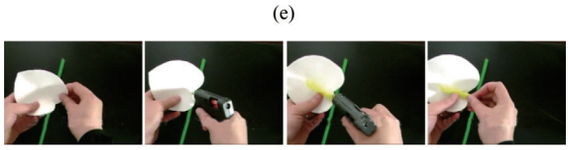
GT:a person pours oil in a pot.
HMN: a man is pouring water into a pan.
EOLM-AN: a person is pouring oil into a pan.
(a)

GT:a girl is playing with an egg.
HMN: a girl is playing with a ball.
EOLM-AN: a girl is playing with an egg.
(b)

GT:a monkey is practicing karate.
HMN: a monkey is fighting.
EOLM-AN: a monkey is doing karate.
(c)

GT:a woman is decorating eggs.
HMN: two girls are playing with toys.
EOLM-AN: a girl is painting eggs.
(d)

GT:a man is extracting lemon juice.
HMN: a man is slicing a vegetable.
EOLM-AN : a man is slicing a lemon.
(e)



GT:a person is making a craft.
HMN: a person is folding a piece of paper.
EOLM-AN: a person is folding a piece of paper.
(f)

图 4　EOLM-AN 模型在 MSVD 和 MSR-VTT 数据集上的可视化实例

Fig. 4　Visualization Examples of EOLM-AN Model on MSVD and MSR-VTT Datasets

son is folding a piece of paper(一个人正在折叠一张纸)”。这表明 EOLM-AN 模型没有实现对整个视频内容的整体理解。

3　结束语

本文提出了一种创新的视频描述方法,即基于增强对象学习和注意力网络的视频描述方法。首先,设计了一种新的增强对象学习模块,有效弥补了模型在学习对象信息方面的不足,从而显著提高了视频描述的准确性。其次,构建了一种注意力网络,能够自适应地关注不同类型的特征,有效增强了模型对关键信息的学习能力。在 MSVD 和 MSR-VTT 数据集上的实验中,相较于其他先进的方法,本文方法展现了优异的性能表现。未来的研究方向将继续聚焦视频描述任务,旨在探索更好的方法,以弥补模型在上下文建模方面的不足,进一步为该领域提供可行的方案。

参考文献:

[1] Zhang J, Peng Y. Video captioning with object-aware spatio-temporal correlation and aggregation[J]. IEEE Transactions on Image Processing, 2020, 29: 6209-6222.

[2] Zanfir M, Marinoiu E, Sminchisescu C. Spatio-temporal attention models for grounded video captioning [C]//Computer Vision-ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 2017: 104-119.

[3] Yang Z, Han Y, Wang Z. Catching the temporal regions-of-interest for video captioning[C]//Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, USA, 2017: 146-153.

[4] Zhang W, Wang X E, Tang S, et al. Relational graph learning for grounded video description generation[C]//Proceedings of the 28th ACM International Conference on Multimedia, Seattle, USA, 2020: 3807-3828.

[5] Zhang Z, Shi Y, Yuan C, et al. Object relational graph with teacher-recommended learning for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13278-13288.

[6] Kanani C S, Saha S, Bhattacharyya P. Global object proposals for improving multi-sentence video descriptions[C]//International Joint Conference on Neural Network, Montreal, Canada, 2021: 1-7.

[7] Parisotto E, Song F, Rae J, et al. Stabilizing transformers for reinforcement learning[C]//International Conference on Machine Learning, Vienna, Austria, 2020: 7487-7498.

[8] Ye H, Li G, Qi Y, et al. Hierarchical modular net-

work for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 17939-17948.

[9] Lin K, Li L, Lin C C, et al. SwinBERT: end-to-end transformers with sparse attention for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 17949-17958.

[10] Gu X, Chen G, Wang Y, et al. Text with knowledge graph augmented transformer for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, USA, 2023: 18941-18951.

[11] Jing S, Zhang H, Zeng P, et al. Memory-based augmentation network for video captioning[J]. IEEE Transactions on Multimedia, 2023, 26: 2367-2379.

[12] Shen Y, Gu X, Xu K, et al. Accurate and fast compressed video captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 15558-15567.

[13] Wang J, Jiang W, Ma L, et al. Bidirectional attentive fusion with context gating for dense video captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7190-7198.

[14] Wang B, Ma L, Zhang W, et al. Controllable video captioning with pos sequence guidance based on gated fusion network[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 2641-2650.

[15] Li L, Gao X, Deng J, et al. Long short-term relation transformer with global gating for video captioning[J]. IEEE Transactions on Image Processing, 2022, 31: 2726-2738.

[16] Xu J, Yao T, Zhang Y, et al. Learning multimodal attention LSTM networks for video captioning[C]//Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, USA, 2017: 537-545.

[17] Sun Z, Chen S, Zhong L. Visual-aware attention dual-stream decoder for video captioning[C]//IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 2022: 1-6.