融合全局概念感知的多模态视频描述生成

汪凯[1,2],杨莉[2*],彭昱凯[3],陈淑琴[2],余良俊[2],巫世峰[4]

(1 湖北师范大学 人工智能与计算机学院,湖北 黄石 435002;2 湖北第二师范学院 计算机与人工智能学院,湖北 武汉 430205;3 武汉理工大学 计算机与人工智能学院,湖北 武汉 430070;4 中乾立源工程咨询有限公司,湖北 武汉 430071)

摘 要 随着人工智能技术的持续演进与突破,视频描述生成已然成为计算机视觉领域的核心任务之一,该任务在视频理解、智能检索等多个领域展现出重要的应用价值.但现有的视频描述方法仍然存在全局语义建模模糊、概念检测精度不足和计算效率低下等问题.提出了一种融合全局概念感知的视频描述生成模型(FC-GCA).该模型结合(CARE)框架的多模态概念检测能力与前向连接网络(FcNet)的全局语义提取优势,有效缓解了传统Transformer解码器的局部语义割裂与动态关联缺失问题.通过引入语义相似性矩阵显式建模词间依赖关系、解码器添加FcNet网络架构、利用混合损失函数优化全局与局部语义对齐,从而增强细粒度语义建模能力.实验在MSR-VTT与MSVD数据集上验证表明:FC-GCA模型能够生成更为准确、连贯的视频描述,具有较好的泛化性和鲁棒性.

关键词 视频描述;多模态融合;语义相似性矩阵;全局语义建模;FcNet网络

中图分类号 TP391.4 **文献标志码** A **文章编号** 1672-4321(2026)04-0532-08

doi:10.20056/j.cnki.ZNMDZK.20250863

Multimodal video captioning with global concept awareness

Wang Kai[1,2], Yang Li[2*], Peng Yukai[3], Chen Shuqin[2], Yu Liangjun[2], Wu Shifeng[4]

(1 College of Artificial Intelligence and Computer Science, Hubei Normal University, Huangshi 435002, Hubei China; 2 College of Computer and Artificial Intelligence, Hubei University of Education, Wuhan 430205, Hubei China; 3 College of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, Hubei China; 4 Zhongqian Liyuan Engineering Consulting Co., Ltd. Wuhan 430071, Hubei China)

Abstract With the continuous evolution and breakthroughs in artificial intelligence, video captioning has become one of the core tasks in the field of computer vision, demonstrating significant application value in video understanding, intelligent retrieval, and other related areas. However, existing video captioning methods still face challenges such as ambiguous global semantic modeling, insufficient concept detection accuracy, and low computational efficiency. To address these issues, a Fusion-based Global Concept-Aware Video Captioning Model (FC-GCA) is proposed. The model integrates the multimodal concept detection capability of the CARE framework with the global semantic extraction advantage of the Forward Connection Network (FcNet), effectively alleviating the problems of local semantic fragmentation and dynamic correlation deficiency that commonly exist in traditional Transformer-based decoders. Specifically, FC-GCA explicitly models inter-word dependencies by introducing a semantic similarity matrix, incorporates the FcNet architecture into the decoder to enhance global semantic reasoning, and employs a hybrid loss function to optimize global-local semantic alignment, thereby improving fine-grained semantic modeling. Experimental results on the MSR-VTT and MSVD datasets demonstrate that the FC-GCA model generates more accurate and coherent video descriptions, exhibiting superior generalization capability and robustness compared to existing approaches.

收稿日期 2025-09-26

* **通信作者** 杨莉(1969-),女,教授,博士,研究方向:人工智能、智能控制,E-mail:hbyangli@hue.edu.cn

基金项目 湖北省自然科学基金资助项目(2023AFB206);湖北第二师范学院人才引进科研启动经费资助项目(ESRC20230009);交通物联网技术湖北省重点实验室开放基金资助项目(WHUTIOT2023-006);基础教育信息技术服务湖北省协同创新中心开放研究基金资助项目(OFHUE202305)

Keywords　video captioning; multimodal fusion; semantic similarity matrix; global semantic modeling; FcNet

随着人工智能的飞速发展，如何让计算机理解视频内容并生成准确、连贯的自然语言描述，已成为计算机视觉与自然语言处理交叉领域的重要挑战.视频描述任务（Video Captioning）[1]旨在让计算机能够像人类一样描述视频内容，其不仅仅针对对象进行简单且泛化的叙述，它往往还包含各种潜在的空间关系、时间关系.由于建立视觉特征和语义特征的一致性被认为是复杂的，视频描述不仅在盲人导航、视频检索以及人机交互等诸多领域有着重要应用价值，也为探索多模态语义理解和生成机制提供了理论验证场景.

早期的视频描述生成用模板填充法，该方法预先定义好了固定的描述模板，而后将检测到的视频特征信息与模板响应位置进行匹配.但由于依赖预设模板和强规则驱动的逻辑，该方法难以生成丰富、精确的视频描述.

随着深度学习的快速发展[2]，Transformer架构在视频描述生成任务中的成功应用，基于视频描述的大规模预训练模型取得了显著的进展.熊等人[3]利用Transformer网络执行多分辨率图像特征的交叉融合.BEiT-3[4]提出在预训练任务中添加掩码数据建模.此外受到概念学习发展的启发，一些研究致力于从自然语言监督中学习抽象的概念语义表示.Yang等人[5]提出了基于概念感知的视频字幕生成框架CARE（Concep-t Aware Video Captioning），其利用有效的先验跨模态信息描述视频.以上这些方法在视频字幕生成任务中展现出了令人印象深刻的学习能力.然而，这些方法普遍依赖于千万级的视频-文本标注数据以及大量的算力支撑问题、复杂视频字幕生成精确性与连贯性不足问题、长序列视频语义挖掘不充分问题.

面对长序列视频语义挖掘不充分、复杂视频字幕生成不精确、生成文本语义不连贯等问题，本文融合了概念感知型视频字幕生成框架（CARE）的多模态概念检测能力以及前向连接网络-FcNet的全局语义融合优势，提出了一种改进的视频描述生成模型FC-GCA（FcNet Gloal Concept Aware），具体的说：CARE框架凭借多模态驱动概念检测以及全局 - 局部语义引导与混合注意力机制来实现高精度概念检测，但其解码器依赖传统Transformer结构，存在计算复杂度高、长序列依赖建模不足等问题，使其并未充分挖掘时序特征间的语义关联.针对此我们利用FcNet构建其语义相似性矩阵与前向连接网络，自适应提取全局上下文信息，提升描述生成的语义准确性，从而生成更准确且连贯性的视频描述.

1　相关工作

视频描述生成任务是结合自然语言处理和计算机视觉交叉学科的前沿领域.该任务主要是针对给定的视频内容进行生成式内容描述.

1.1　视频描述生成

在视频描述初级阶段，一部分工作致力于模板填充法研究，将检测到的视频特征信息，利用定义的固定或者半固定语言模板填充，最终生成视频描述.然而，模板填充法存在描述不准确、泛化性差、无法生成可变长度的句子描述等问题.现有的视频描述模型基于编码器-解码器（Encoder-Decoder）架构研究.在循环神经网络（RNN）中，尽管模型的输出会受到先前输入数据的影响，但其在实际训练的过程中，所采用的序列信息处理机制仍然表现出较高的局限性.具体地说，该机制存在固有的偏向性，如针对后序输入的序列信息，其对最终输出结果的影响权重较高；针对前序输入的序列信息，其对最终结果的影响权重较低，且无法动态调整该机制本身决定的信息衰减模式.

随着深度学习的发展，为了增强多模态特征表达能力，一部分工作致力于改进编码器研究.在视频分析领域，基于卷积神经网络（CNN）的编码器占据主导地位[6].此类模型通过层级化的感受域设计提取视频特征.而后以长短期记忆网络（LSTM）为代表的时序建模单元，构建了视频帧间隐藏状态向量的跨帧传递通道[7].该方法有效解决了传统CNN架构在动态场景中的表征局限性.Chen等人[8]引入了基于 Transformer 与LSTM的语言斟酌级联，自适应的构建全局语义信息，以解决模型预测出的单词仅与前文的隐藏状态相关联，而忽略了整体语境问题.

为了生成更精确且连贯的语句，一部分工作致力于改进解码器研究.早期的视频描述任务使用循环神经网络（RNN）作为解码器，但困扰已久的梯度消失与梯度爆炸问题始终无法完美权衡导致难以捕捉到长时序任务的依赖.为了解决此问题，LSTM作为解码器被引入至视频描述领域[9].其通过引入

门控机制,有效缓解了RNN梯度消失问题.Gao等人[10]提出了基于语法感知的解码器来动态度量视觉语义和语法相关词的贡献,实现了更细粒度更连贯的视频描述.

为了更多聚焦于视觉和文本模态的一致性,一部分工作致力于改进注意力机制研究.Yao等人[11]提出时空注意力机制,使用时间注意力权重来动态选择关键帧并加权融合特征,利用帧级和视频级特征的结合来增强对复杂动作的描述能力.Tu等人[12]提出了一种时空注意力机制从而利用视频中的空间时间结构关系;Deng等人[13]提出级联式注意力机制,有效解决了传统注意力机制中的不同模态特征交互性不足问题;尽管以上这些注意力机制都取得了不错的效果,但大多数方法任然存在概念检测精度有限和概念利用不充分的问题.

1.2 概念感知视频描述模型

概念是指与视频内容中的物体、动作及属性相关的具有意义的词汇的集合.概念不仅反映了视频内容的高级语义结构,还在视频和文本模态之间起到了重要的桥梁作用.

随着视频描述任务的持续发展,多模态驱动的概念检测被首次提出[6].首先,CA-baseline模型使用编码器-解码器结构进行训练,编码器部分利用单独的视觉特征提取模块进行处理,而后利用所获取的视觉特征,筛选出置信度最高的N个词语"概念1 概念2 概念3"作为概念序列,通过局部引导视频字幕生成发挥作用.

此外,Yang等人提出概念感知的视频字幕生成模型(CARE)使用编码器-解码器结构进行训练,编码器部分负责将视频内容(如视频帧、音频等)表示为紧凑特征.解码器负责基于生成的各模态特征和给定的语义向量生成字幕序列.其中,提出了多模态驱动的概念检测模块(MCD),利用视频获取的视觉、文本、音频特征,筛选出最频繁出现的N个词语"概念1 概念2 概念3"作为概念序列,再对于每个视频判断筛选出的概念词是否存在于视频的真实字幕中.随后,基于这些概念概率,通过Transformer注意力机制发挥作用.

然而,简单的利用Transformer架构进行建模存在无法很好的利用全局信息的问题.首先,Tansformer的全局注意力机制会平等的关注所有历史性的词汇,往往会导致局部语义割裂问题,致使在生成语义相互矛盾的句子时,无法有效分辨句子间的语义冲突,从而生成语句连贯性不足.其次,视频描述任务往往需要根据当前生成的词汇去动态调整上下文的权重,然而Transformer架构的固定位置编码无法达成这类动态聚焦.例如,在生成"射门"词汇时,无法提高与"足球""球门"等词汇的关联.

综合前文所探讨的各类理论模型、相关工作以及现存长序列视频语义挖掘不充分、复杂视频字幕生成不精确、生成文本语义不连贯等问题,本文的核心贡献如下:

(1)提出全局概念感知的多模态视频描述生成模型FC-GCA,通过前向连接网络-FcNet的全局语义提取优势,提高视频语义特征的潜在信息理解的同时提升细腻度语义关联性,引导模型生成更连贯、精确的视频描述.

(2)提出全局语义关联性融合,利用语义相似性矩阵的词间关联性质和词间的软注意力权重情况,提高全局概念语义连贯性.

(3)提出特定的多模态特征融合模块,利用语义相似性加权充分融合视觉-语义特征.在MSR-VTT和MSVD数据集上进行大量实验,结果表明本文方法能够取得更好的指标,具有更好的性能.

2 FC-GCA模型框架

本文提出一种新的融合全局概念感知的多模态视频描述生成模型-FC-GCA,采用编码器解码器架构.如图1所示,在编码阶段,级联与训练CLIP和3D-RX分别提取外观特征与动作特征.在解码阶段,利用Transformer中的注意力机制加深编码器阶段重要特征影响,而后利用FcNet网络学习细粒度语义特征联系,解决Transformer局部语义割裂问题,使得表征空间在学习中能够整合更为全面的信息.

2.1 视频描述基线模型

给定一个长度为N的采样帧序列,$V=\{v_1,v_2,\cdots,v_n\}$,以及一个预定义的模型维度d_h(例如512),编码器部分将序列V转化为具体的特征表示$F\in\mathbb{R}^{N\times d_h}$:

$$F=\mathrm{Encoder}(V)=\mathrm{LN}(\mathrm{Backbone}(V)W_{\mathrm{enc}}+b_{\mathrm{enc}}),\quad(1)$$

其中LN表示层归一化操作,$W_{\mathrm{enc}}\in\mathbb{R}^{d_h\times d_h}$与$b_{\mathrm{enc}}\in\mathbb{R}^{d_h}$为可训练参数值,$d_h$为与训练骨干网络的输出维度.在处理视觉模态的数据时,首先利用独立的编码器获取视觉特征$F^{(V)}\in\mathbb{R}^{N_v\times d_h}$.在编码阶段完成后,解码阶段生成一个长度为$T$的序列$Y=\{y_1,y_{2,\cdots},y_T\}$.其中,每个单词$y_t$的生成均取决于视频特征$F$和之前生成过的词序列$y_{<t}$(即$y_1-y_{<t-1}$).具体的说,在第$t$

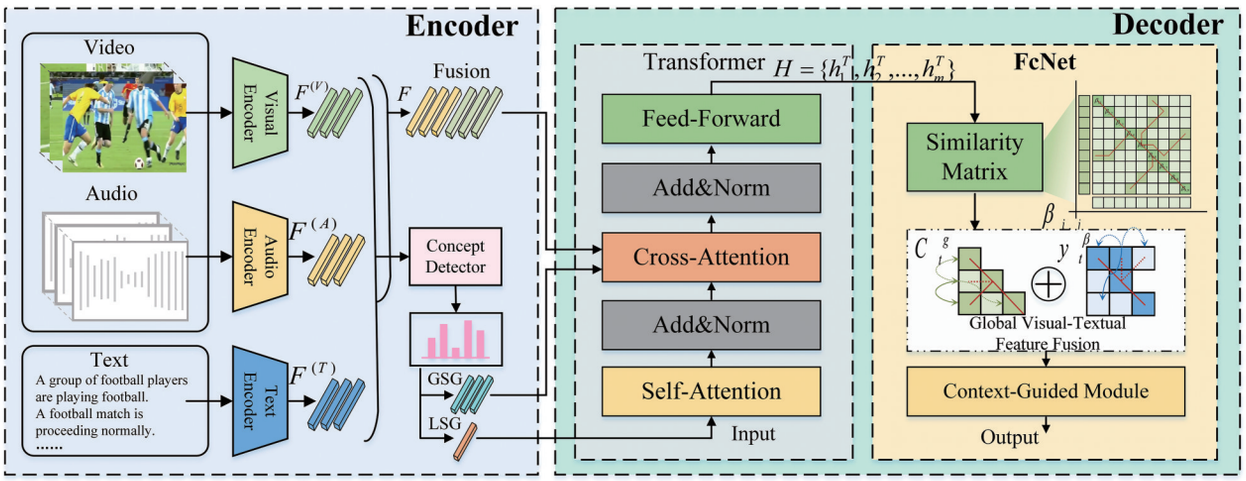


图 1　FC-GCA 模型整体框架图

个时间步通过以下公式计算输入嵌入：

$$e_t = \mathrm{Embedding}(y_t - 1) = \mathrm{LN}(W^{\mathrm{word}}_{[y_t-1]} + W^{\mathrm{pos}}_{[t]}), \quad (2)$$

其中 $W^{\mathrm{word}} \in \mathbb{R}^{|V| \times d_h}$ 与 $W^{\mathrm{pos}} \in \mathbb{R}^{T_{\max} \times d_h}$ 分别为词嵌入与位置嵌入的可训练矩阵；$|V|$ 表示词表大小；$T_{\max}$ 表示预定义最大序列长度. 而后将 $E_{\leqslant t} = \{e_1 e_2, \cdots, e_t\}$（即 t 时刻之前所有时间步的输入嵌入构成的集合）与 F 输入至解码器，生成隐藏状态 $h_t \in \mathbb{R}^{d_h}$，具体计算公式如下：

$$h_t = \mathrm{Decoder}(E_{\leqslant t}, F). \quad (3)$$

2.2　基于 FcNet 的解码器构建

2.2.1　语义相似性矩阵构建

在视频描述生成过程中，生成的词序列隐含了当前语句的上下文信息. FcNet 通过计算词向量之间的余弦相似度，量化词汇之间的语义关联强度. 设已生成词序列的隐藏状态集合为 $H = \{h_1, h_2, \cdots, h_t\}$，其中 $h_i \in R^d$ 为第 i 个词的隐藏表示（例如通过 LSTM 或 Transformer 编码得到）. 则语义相似性度 $\beta \in R^{t \times t}$ 的计算公式如下：

$$\beta_{ij} = \frac{\exp\left(h_i^{\top} h_j\right)}{\sum_{k=1}^{t} \exp\left(h_i^{\top} h_k\right)}, \quad (4)$$

语义相似性性矩阵充分关注词与词间的软注意力权重情况，例如：当生成“射门”词汇时，其与“足球”、“球员”等词汇的相似度相对较高，与“烹饪”、“情绪”等不相关词汇的相似度较低甚至为零，有助于模型在生成过程中判断词之间的语义联系度.

如图 2 所示，为了更好的契合生成过程中的动态演变情况，语义相似性矩阵往往以非静态的方式构建，从而根据生成步骤逐步给予调整. 当处于初始化阶段生成的首个单词时，相似性矩阵仅包含单独的一个元素 $\beta_{1,1} = 1$，此时并不存在上下文之间的关联，在进行迭代扩展的过程中，每生成一个新的词汇，矩阵的维度扩展为 $(t+1) \times (t+1)$. 为保证每行的权重总和为 1. 具体公式如下：

$$\sum_{j=1}^{t+1} \beta_{ij} = 1. \quad (5)$$

而后，对所有的相似性分数进行归一化处理. 该动态更新机制让模型在生成过程中自适应捕捉长距离语义依赖. 具体的说，当生成句子“球员-带球-射门-得分”时，“射门”与“得分”之间的相似度会高于“球员”和“得分”的相似度，使得模型在生成描述时更关注于紧邻动作间的因果关系，具体公式

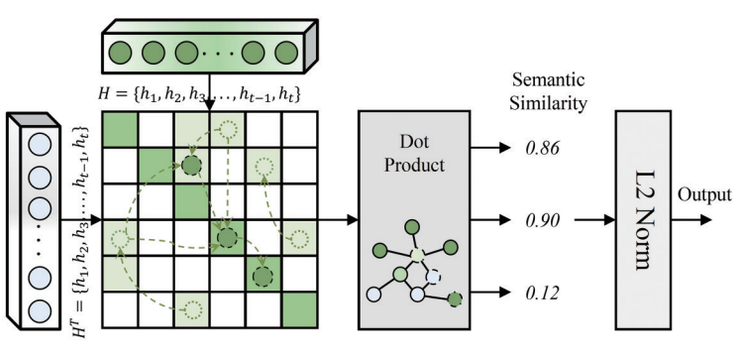


图 2　语义相似性矩阵构建

如下:

$$\beta = \mathrm{softmax}(H \cdot H^T). \tag{6}$$

2.2.2 视觉-语义融合模块

为了更好的将视频特征与语义相似性矩阵结合,提取与当前生成词相关的全局视觉信息.本文引入视觉-语义融合模块.在视频特征编码中使用3D CNN提取视频片段的时空特征,得到视觉特征集合$V = \{v_1, v_2, \cdots, v_n\}$,其中$V_i \in R^d$为第$i$个视频片段视觉特征.具体公式如下:

$$C_t^g = \sum_{j=1}^{t} \beta_{t,j} \cdot v_j, \tag{7}$$

词向量融合对已生成词汇的嵌入向量$E = \{e_1, e_2, \cdots, e_t\}$进行加权融合,得到融合后的语义信息,具体公式如下:

$$y_t^{\beta} = \sum_{j=1}^{t} \beta_{t,j} \cdot e_j, \tag{8}$$

其中,$\beta_{t,j}$表示第t个词与第j个视频片段的关联权重.例如,在生成"射门"时,模型会对包含"球员起脚"动作的视频片段赋予更高的权重($\beta_{\text{射门,起脚}} = 0.42$).相似性加权是通过加权求和对视频特征进行处理,以加强与当前生成词相关的视觉线索.具体如图3所示:

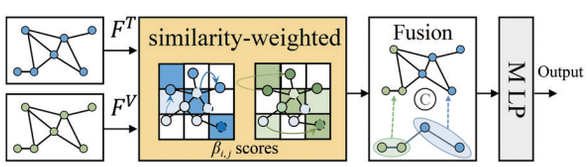


图3 视觉-语义融合模块框架

2.2.3 上下文引导模块

基于编码器生成的特征,解码器阶段旨在生成一个包含m个词的语句$C = \{y_1, y_2, \cdots, y_n\}$,为了进一步挖掘视频中的关键特征,本文引入上下文引导模块,其聚合历史生成词的语义信息,指导当前词的生成方向.

此外,在FcNet的双流LSTM结构中,局部LSTM接收当前词的嵌入e_t和视觉上下文c_t^g,输出局部隐藏状态h_t^l.全局LSTM接收融合后的语义向量y_t^g和视觉全局特征c_t^g,输出全局隐藏状态h_t^g.在联合预测中,将局部与全局隐藏状态拼接完成后,通过全连接层预测下一个单词,具体公式如下:

$$p_t = \mathrm{Softmax}(W_p[h_t^l, h_t^g]), \tag{9}$$

为平衡局部词预测与全局语义一致性,联合训练损失函数加权调整,损失函数包含以下两项:局部语义损失($\mathcal{L}_l$)、全局语义损失($\mathcal{L}_g$).具体公式如下:

$$\mathcal{L} = \lambda \mathcal{L}_l + (1 - \lambda)\mathcal{L}_g. \tag{10}$$

3 实验结果与分析

本文实验所用硬件环境:CPU为Inter(R)Core(TM)i7-13700、GPU为NVIDIA GTX 3090、内存为32GB.

其中软件环境:操作系统为Linux系统、训练与测试架构为PyTorch,编程语言为Python.

3.1 数据集与评价指标

本文使用微软开源视频字幕生成大规模数据集MSR-VTT(Microsoft Video Description)与MSVD(Microsoft Video Description),其中MSR-VTT数据集包含10000个视频片段,每条视频片段配置20条英文描述标注,其包含多类型真实场景,能从不同的角度描述视频核心内容.官方数据划分为6513个视频用于模型训练,497个视频用于验证,2990个视频用于测试.其中MSVD包含1970个视频片段,每条视频片段配置40条英文描述标注,其包含多类型真实场景,能从不同的角度描述视频核心内容.官方数据划分为1200个视频用于模型训练,100个视频用于验证,670个视频用于测试.

由于视频描述生成的多样性,生成精确且连贯的描述往往是具有挑战且复杂的,不同的语句在语法和语义上差距甚大.目前,视频描述生成任务的评估方法大多使用BLEU(Bilingual Evaluation Understudy)、ROUGE(Recall-Oriented Understudy for Gisting Evaluation)、METEOR(Metric for Evaluation of Translation with Explicit ORdering)、CIDEr(Consensus-based Image Description Evaluation)等机器翻译和自然语言处理的方法.因此,本文也将采用上述评估方法.

BLEU是机器翻译领域常用的自动评价指标,通过对比模型生成的候选文与人工参考译文的N元组重叠度计算得分,以量化候选译文接近人类译文的程度.具体公式如下:

$$\mathrm{BLEU} = \min\left(1, \exp\left(1 - \frac{\mathrm{reference - length}}{\mathrm{output - length}}\right)\right) \left(\prod_{i=1}^{4} \mathrm{precision}_i\right)^{1/4}, \tag{1}$$

ROUGE与BLEU类似,核心差异在于其以"召回率"为导向的相似性度量方法,更适用于文本摘要、问答等任务.具体来说,ROUGE通过量化候选文本与参考文本的信息覆盖度,从而衡量机器输出与人类标注的相似度.如表1所示,ROUGE包含ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-S四种.具体公式如下:

$$\mathrm{ROUGE}-N(x)=\frac{\sum_{k=1}^{n}\sum_{\omega\in W}\min(c_{\omega}(x),c_{\omega}(r_k))}{\sum_{k=1}^{n}\sum_{\omega\in W}c_{\omega}(r_k)}. \tag{2}$$

表1　ROUGE分类

分类	特点
ROUGE-N	基于连续N-Gram的召回率，粒度可调
ROUGE-L	基于最长公共子序列(LCS)，兼顾长句连贯
ROUGE-W	加权最长公共子序列(LCS)，强化连续匹配
ROUGE-S	基于调词二元组共现性精确度，捕捉非连续语义

METEOR核心是通过“精确+词干+同义词”的多维度匹配提升语义鲁棒性，评估文本词序与基准词序比对一致性从而构建惩罚减分机制.最后，计算精度与召回率的加权调和平均，并结合惩罚机制得分得到最终的METEOR值.具体公式如下：

$$\mathrm{METEOR}=(1-\mathrm{Pen})\times F_{\mathrm{mean}}, \tag{3}$$

CIDEr不同于上述的评价指标，CIDEr是图像描述生成的评估指标，其主要是通过“多参考共识”匹配候选描述与人类对同一图像的多个参考描述，通过余弦相似度衡量候选与基准语义的匹配度，从而得到最终评分.具体公式如下：

$$\mathrm{CIDEr}_n(c_i,S_i)=\sum_{n=1}^{N}\omega_n\mathrm{CIDEr}_n(c_i,S_i). \tag{4}$$

本文在数据集中对输入的视频序列已经提前利用卷积神经网络提取了视频特征，训练的batchsize大小设置为64，epoch大小设置为50次，模型尺寸配置选用base版本，提取视频特征所使用的方法设置为CLIP的Vision Transformer(ViT)，学习率设置为0.0001、学习率预热步数设置为10000、梯度裁剪值设置为1.0.

3.2　对比实验

为验证所提模型的有效性，本文在MSR-VTT数据集上对比了基线模型(baseline、CAbaseline、CARE)、两种常见视频描述生成模型BcNet、FcNet，以及6种不同改进的视频描述生成模型SGN、MGCMP、TVRD、RSFD、HNAT.并使用基于人工译文相似性的指标BLEU-4、基于计算精度与召回率的加权调和平均的指标METEOR、基于拥有最长公共子序列兼长句连贯性精度和召回率Fmeasure统计的指标ROUGE-L、基于余弦相似度衡量候选与基准语义匹配度指标CIDEr进行评价，最终结果如表2所示：

可以看出，本文FC-GCA模型与常见视频描述生成模型SGN相比，在BLEU-4、METEOR、ROUGE-L和CIDEr指标上分别提升了8.0、3.3、4.3、10.7；与FcNet相比，在各个指标上分别提升了7.7、3.3、4.0、10.9.值得注意的是，与6种不同改进的视频描述生成模型MGCMP、TVRD、BcNet、RSFD、HNAT、CARE相比，本所提模型在各项指标上均有一定竞争力.这说明FcNet网络的加入，在不丢失预测精度的基础上对全局信息进行了更深层的学习.

表2　MSR-VTT数据集上模型性能比较

方法	BLEU-4	METEOR	ROUGE-L	CIDEr
SGN[14]	40.8	28.3	60.8	49.5
MGCMP[15]	41.7	28.9	62.1	51.4
TVRD[16]	43.0	28.7	62.2	51.8
FcNet	41.1	28.3	61.1	49.3
BcNet	40.9	28.6	61.0	49.3
RSFD[17]	43.4	29.3	62.3	53.1
HRNAT[18]	42.1	28.0	61.6	48.2
FC-GCA(ours)	**48.8**	**31.6**	**65.1**	**60.2**

此外，表3也显示了MSVD数据集上指标对比，与在MSR-VTT数据集上的实验结果类似，本文FC-GCA模型在四个指标上均优于经典视频描述生成模型SGN和MGCMP.值得注意的是，与其他不同改进的视频描述生成模型相比，在4个指标上均有一定的竞争力.特别在最重要的评估语义一致性的CIDEr指标上分别有着不少的提升.这不难得出FC-GCA从解决局部语义割裂问题与词汇动态关联缺失问题出发，通过构建语义相似性矩阵进而引导上下文词汇生成，实现了更为准确且连贯的视频描述.

表3　MSVD数据集上模型性能比较

方法	BLEU-4	METEOR	ROUGE-L	CIDEr
SGN[14]	52.8	35.5	72.9	94.3
MGCMP[15]	55.8	36.9	74.5	98.5
TVRD[16]	50.5	34.5	71.7	84.3
FcNet	50.7	34.4	71.6	89.4
BcNet	51.3	34.7	71.9	91.1
RSFD[17]	51.2	35.7	72.9	96.7
HRNAT[18]	55.7	36.8	74.1	98.1
FC-GCA(ours)	**56.3**	**39.5**	**75.8**	**108.6**

3.3　消融实验

为比较不同模块对视频描述影响，本文在MSR-VTT与MSVD数据集上，利用BLUE@4、METEOR、ROUGE-L、CIDEr评估指标系统性的分析了视觉语义驱动、多模态语义驱动以及结合FcNet的全局语义引导不同模块结合后对模型性能的影响.消融实验如表4所示.

表4 消融实验测试结果

方法	MSR-VTT				MSVD			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
Baseline	47.5	31.1	64.4	57.6	53.1	37.9	74.1	102.1
CA-baseline	47.7	31.1	64.5	58.4	53.5	37.7	74.5	103.9
CARE	48.7	31.5	65.2	59.7	56.3	39.1	75.6	106.9
FC-GCA(ours)	**48.8**	**31.6**	65.1	**60.2**	56.3	**39.5**	**75.8**	**108.6**

相较于视频描述基线模型,在加入特定视觉特征提取模块后,在BLEU-4、ROUGE-L与CIDEr三个指标上数值有所提升.在加入多模态驱动的概念检测模块后,在BLEU-4、ROUGE-L与CIDEr三个指标上数值有所提升,这表明多模态的信息能够通过注意力机制来丰富上下文线索,以此来增强模型对视频内容的更深层次理解.在加入FcNet模块后,在BLEU-4、METEOR与CIDEr三个指标上数值有所提升,这表明全局语义引导能驱动模型具有更好的句子结构学习能力,以生成更连贯且精确的视频描述.

实验表明,融合FcNet模块能使模型在MSR-VTT与MSVD数据集上表现更好,有效的证明了改进方法的有效性和合理性.

3.4 案例结果分析

本文的视频内容描述生成案例结果分析如图4所示,针对第一个视频,基线模型由于无逻辑语义的文本生成,导致出现了"蔬菜(vegetables)"被认为成"胡萝卜(carrot)"的严重错误,而FC-GCA模型由于语义相似性矩阵对全局信息的关联处理,从而生成了更为精确且接近GT的描述"一位女士正在切蔬菜(a woman is cutting a vegetable)".在第二个视频中,基线模型生成了错误的目标语义信息"一只猴子正在行走(a monkey is walking)",而本文的FC-GCA模型通过更深层次的特征学习,生成了更为准确且更接近GT描述的内容"一群大象正在行走(a group of elephants are walking)".在第四个视频中,基线模型生成了模糊的目标语义信息"一个人正抱着一只小动物(a person is holding a small animal)",而本文的FC-GCA模型通过针对全局语义的特征学习,生成了更为拟人且精确的描述,模型判断出目标是"动物幼崽(baby animal)"而非"小动物(small animal)",更符合接近GT句子所表达含义.同时,其他例子也能证明全局语义结构学习对语义理解和文本内容生成具有积极意义.由此可得,本文FC-GCA模型能够有效的解决计算复杂度高、长序列依赖建模不足导致未充分挖掘时序特征间的语义关联问题,从而生成更为全面且精确的视频描述.

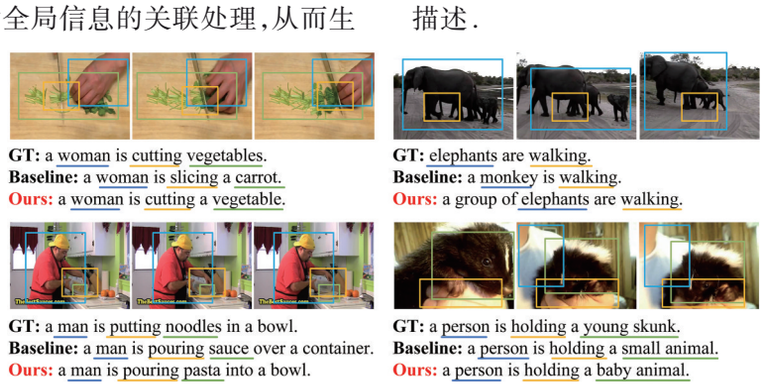


图4 视频内容描述生成对比

4 总结

本文在编码器阶段利用多模态驱动的概念检测机制对所获取到的视觉、文本、音频特征进行概念序列构建的方式有效地解决了概念检测精度低、依赖单一模态易歧义问题,加深了模型对潜在表征空间中的视频特征向量理解.同时还引入了概念概率分布向量门控,以构建全局-局部语义信息交互通道以加强后续传入潜在表征空间的视频语义更全面.在解码器阶段,本文创新性的融合了FcNet网络架构,通过构建语义相似性矩阵以显式建模单词间的依赖关系,并结合视觉-语义融合模块与上下文引导模块,构成高效特征处理机制.相较于传统Transformer解码器,FC-GCA具有更好的长序列依赖建模能力和句子结构学习能力,有效解决了视频描述生成中的局部语义割裂、动态关联缺失以及时

序特征间语义关联挖掘不充分问题，从而生成更为精确、连贯的视频描述.

在未来工作中，如何在动态复杂场景下生成连贯且精确的描述仍然是目前研究领域的巨大挑战，这对推动视频描述任务在盲人导航、智能监控等实际应用中的落地具有重要意义.同时，本文模型也存在一定不足，例如MSR-VTT与MSVD训练集其文本结构并不复杂，因此导致模型在处理更为复杂领域（如赛事解说、医疗事件）时文本生成优势提升不大，性能尚未完全挖掘，其泛化能力仍需在更为多样化的场景数据中进一步优化.

参 考 文 献

[1] Sun Z, Zhong X, Chen S, et al. Modeling context-guided visual and linguistic semantic feature for video captioning [C]//Artificial Neural Networks and Machine Learning – ICANN 2021. Cham: Springer, 2021: 677-689.

[2] Duan Y, Wang Z, Li Y, et al. Cross-domain multi-style merge for image captioning [J]. Computer Vision and Image Understanding, 2023, 228: 103617.

[3] 熊承义，马帅，高志荣，等. 基于Transformer多分辨率特征融合的图像压缩感知重构[J]. 中南民族大学学报（自然科学版），2025，44(3)：400-406.

[4] Wang W, Bao H, Dong L, et al. Image as a foreign language: BEIT pretraining for vision and vision-language tasks [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023: 19175-19186.

[5] Yang B, Cao M, Zou Y. Concept-aware video captioning: Describing videos with effective prior information [J]. IEEE Transactions on Image Processing, 2023, 32: 5366-5378.

[6] Zhang Z, Qi Z, Yuan C, et al. Open-book video captioning with retrieve-copy-generate network [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 9832-9841.

[7] Yuan Y, Ma L, Zhu W. Syntax customized video captioning by imitating exemplar sentences [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 10209-10221.

[8] Chen S, Sun Z, Yang L, et al. Refined linguistic deliberation for video captioning *via* cascade transformer and LSTM[J]. Multimedia Systems, 2025, 31(3): 200.

[9] Song P, Guo D, Cheng J, et al. Contextual attention network for emotional video captioning [J]. IEEE Transactions on Multimedia, 2023, 25: 1858-1867.

[10] Gao L, Lei Y, Zeng P, et al. Hierarchical representation network with auxiliary tasks for video captioning and video question answering [J]. IEEE Transactions on Image Processing, 2021, 31: 202-215.

[11] Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure [C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: ACM, 2015: 4507-4515.

[12] Tu Y, Zhang X, LIU B, et al. Video description with spatial-temporal attention [C]//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View: ACM, 2017: 1014-1022.

[13] Deng J, Li L, Zhang B, et al. Syntax-guided hierarchical attention network for video captioning [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(2): 880-892.

[14] Ryu H, Kang S, Kang H, et al. Semantic grouping network for video captioning [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2514-2522.

[15] Chen S, Jiang Y G. Motion guided region message passing for video captioning [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 1523-1532.

[16] Wu B, Niu G, Yu J, et al. Towards knowledge-aware video captioning *via* transitive visual relationship detection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6753-6765.

[17] Zhong Xian, Li Zipeng, Chen Shuqin, et al. Refined semantic enhancement towards frequency diffusion for video captioning [C]//Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence. Washington, D.C.: ACM, 2023: 3724-3732.

[18] Lei Y, He Z, Zeng P, et al. Hierarchical representation network with auxiliary tasks for video captioning [C]// 2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen: IEEE, 2021: 1-6.

（责编&校对　雷建云）