DOI:10.19652/j.cnki.femt.2305361

基于增强全局—局部特征融合的视频描述生成方法*

黄飞燕 曾上游 邱泓语
(广西师范大学电子与信息工程学院/集成电路学院 桂林 541004)

摘 要:现有的视频描述生成方法提取的特征及特征组合的方式较为简单,导致模型丢失了部分与视频描述相关的重要语义信息,限制了对视频内容的准确描述和理解。分析存在的不足,提出了一种基于增强全局—局部特征融合的视频描述生成方法。首先采用不同特征提取器分别对视频片段提取局部特征和全局特征,为了建模不同级别特征(局部和全局)的相关性,利用特征融合增强网络进行特征融合,丰富模型的特征信息。解码器使用的双向长短期记忆网络,并在其后加入重构网络,重构经编码器处理得到的视频特征序列,最终经过长短期记忆网络生成视频的描述语句。在 MSVD 与 MSR-VTT 数据集上的实验结果表明,提出的模型可以显著提高生成的描述语句的准确性。

关键词:视频描述生成;增强特征融合网络;自然语言处理
中图分类号:TP391.4;TP183 **文献标识码**:A **国家标准学科分类代码**:520.20

Video description generation method based on enhanced global-local feature fusion

Huang Feiyan Zeng Shangyou Qiu Hongyu
(School of Electronic and Information Engineering/School of Integrated Circuits, Guangxi Normal University, Guilin 541004, China)

Abstract: Existing video description generation methods extract features and feature combinations in a simpler way, resulting in the model losing some of the important semantic information related to the video description, limiting the accurate description and understanding of the video content. Analysing the deficiencies, this paper proposes a video description generation method based on enhanced global-local feature fusion. Firstly, different feature extractors are used to extract local and global features for the video clips respectively, and in order to model the relevance of different levels of features (local and global), feature fusion is performed using a feature fusion enhancement network to enrich the feature information of the model. In this paper, the bi-directional long and short term memory network used by the decoder is followed by a reconstruction network, which reconstructs the video feature sequences obtained by the encoder processing, and finally generates the descriptive statements of the video through the long and short term memory network. Experimental results on MSVD and MSR-VTT datasets show that the model proposed in this paper can significantly improve the accuracy of the generated descriptive statements.

Keywords: video description generation; enhanced feature fusion network; natural language processing

0 引 言

人类可以轻易地识别他们的周围环境,并可以用他们的自然语言描述任何图像或视频场景,但对于机器来说,却很难像人类一样生成对图像和视频的描述。然而,机器可以在一定程度上从视频帧和图像中识别各种人类活动,但如何准确描述复杂和长期的人类活动的视觉场景仍然是一项具有挑战性的任务。虽然视频描述可能看起来是一项纯粹的数字任务,但是它对于人类进行物理交互并理解其周围环境的机器人系统具有启示意义。例如,在视频搜索、视频推荐、视频编辑和虚拟现实等领域都具有重要的应用价值。

收稿日期:2023-07-27
* **基金项目**:国家自然科学基金(61976063)项目资助

视频描述算法的研究面临着许多挑战和问题。首先，视频是一种时间序列数据，在时间和空间尺度上具有更复杂的结构和变化。因此，如何有效地建模视频的时空信息以及视频内容的演化过程是一个关键问题。其次，视频描述算法需要兼顾对视频中的动作、物体和场景等内容进行全面的理解和表达，这需要综合利用多模态信息，如视频图像、音频和文本等。最重要的是，视频描述算法需要具备一定的自适应能力，能够根据具体任务和用户需求生成相应的描述。

当前，现有的大多数视频描述模型主要采用的是基于深度学习的编码器－解码器结构，其中，编码器通常使用卷积神经网络（convolutional neural network，CNN），CNN 模型在视频的每一帧上进行卷积操作，提取出视频帧的视觉特征。2015 年，Venugopalan 等[1]提出了一种用于生成视频描述的端到端模型，该方法利用 VGG 网络提取视频各帧特征，但是后续发现这种 2D 卷积的做法只能捕捉每一帧图像的静态信息，对于动态特征的处理能力相对较弱。文献[2-4]则结合了 2D 卷积和 3D 卷积作为视觉特征提取器进行特征提取，提取视频的静态信息和动态信息。文献[5]提出了一种使用生成对抗网络（conditional generative adversarial network，CGAN）学习视频字幕分布的方法，与传统方法相比，他们的方法可以更好地学习到真实字幕的分布，并生成更准确和多样化的字幕。最近，由于基于 Transformer 的视觉预训练模型在计算机视觉领域取得了惊人的突破[6]，为视频描述领域提供了一种新的有效方法，文献[7]基于 Transformer 架构模型，通过将图像划分为图像块，并使用 Transformer 编码器来对这些图像块进行处理和编码，能够成功地应用在各个领域，包括但不限于机器翻译、文本分类和目标检测等。文献[8]将基于 Transformer 的视觉预训练模型和强化学习用于视频描述任务中，专注于提取视频内容的整体特征。文献[9]交互增强转换器（interaction augmented transformer，IAT）的新方法，该方法具有判别编码和解耦解码，对编码结果执行重构对比约束，能有效的提升描述的准确性。

而解码器则选择能够对时间进行建模的递归神经网络（recurrent neural network，RNN）以及变种，解码器的性能和效果很大程度上取决于编码器提供的上下文向量的质量和丰富程度[10]。然而，目前生成的描述往往只关注视频的局部特征，而忽视了整体的语义信息。例如，模型可能只描述视频中的某一个动作或场景，但无法正确地表明它们之间的因果关系或者时间顺序。其次，由于视频内容的复杂性，现有模型往往无法准确地描述视频中的全部信息。特别是对于长时间、包含多人或复杂场景的视频，现有模型的性能尤其不足[11]。

针对上述情况，本文提出了一种基于融合全局－局部特征的视频描述生成方法。局部特征和全局特征代表了图像的不同层次的信息，将局部和全局特征结合起来可以得到更强大的表征能力，有助于区分不同类别的样本或提取更具判别性的特征。此外，直接拼接局部特征和全局特征可能无法明确地表示局部特征和全局特征之间的关联关系，通过利用特征融合增强网络进行特征融合，建模不同级别特征（局部和全局）的相关性，提高特征的表达能力，减少无关特征的干扰，适应不同任务和场景的需求，并最终提升模型的性能和泛化能力。

1 相关工作

近年来，深度学习在视频处理和计算机视觉领域取得了很大的进步，进而推动了视频描述算法的研究，基于深度学习的视频描述方法也逐渐成为主流。基于深度学习的视频描述方法主要分为 2 种。基于序列到序列的视频描述方法和基于 Transformer 的视频描述方法。本文采取的是序列到序列视频描述方法。

1.1 基于序列到序列的视频描述方法

编码器－解码器架构是基于序列到序列的视频字幕模型的基础，该模型面临的挑战主要有两点：1）如何从视频中提取更有价值的特征；2）如何利用特征信息生成更准确的视频描述语句[12]。

在最初的编码器－解码器架构中，通常使用 2D-CNN 作为编码器提取视觉特征，解码器则使用的是 RNN。然而 2D-CNN 并没有考虑时间维度信息，为了更好的获取时间尺度上的动态信息，Zolfaghari 等[13]提出了高效卷积神经网络（efficient convolutional，ECO），将 2D-CNN 和 3D-CNN 结合能够更好地捕获时空信息。上述的方法都是使用给定的先前单词和视觉内容在本地生成单词，在整体上未能充分利用句子语义和视觉信息之间的关联。Pan 等[14]提出了一种联合建模嵌入和翻译的方法，用于将视频和语言之间建立联系和生成描述。为了生成更加准确的视频描述，Bilkhu 等[15]则引入了自注意力机制，通过对视频中的不同帧之间的关联进行建模。后来，Zhang 等[16]通过使用图卷积网络来对视频中的帧进行建模，并生成更准确和连贯的视频描述。Seo 等[17]提出了从未标记视频中学习并用于生成任务的多模式视频生成预训练模型（multimodal video generative pretraining，MV-GPT），该模型从原始像素生成字幕，并将生成的字幕直接转录成语音。为了更好的利用语义信息，Vaidya 等[18]则提出了一种基于共分割辅助的双流架构来进行视频字幕生成，通过引入额外的语义信息，模型能够在生成准确度和多样性方面取得良好的表现。此外，Chen 等[19]提出了检索增强卷积编码器－解码器网络，该网络将 RAM 集成到卷积编码器－解码器结构中，这有助于单词预测并提高视频字幕的性能。

1.2 基于 Transformer 的视频描述方法

最近，基于 Transformer 的方法在广泛的语言任务上取得了重大的突破。由于该方法在视频描述任务中具有长距离依赖建模、上下文感知能力、并行计算能力、多模态

处理和预训练能力等优点，越来越多的研者将 Transformer 用于视频描述任务中[20]。该方法的核心是自注意力机制，它能够同时考虑序列中的所有位置并捕捉全局的关联关系，从而提高序列建模的能力。为了更好的将基于 Transformer 的方法应用于视频描述的任务中，Sun 等[21]提出了一种结合了视频和文本数据进行学习的 BERT 扩展版本，它是基于预先训练和微调，可用于动作分类和视频字幕等众多任务。在 2020 年，Luo 等[22]提出了一种统一的视频和语言预训练模型，用于多模态理解和生成，它改进了视频文本相关的下游任务，如基于文本的视频检索和多模式视频字幕。上面讨论的方法不能完全使用语义概念，为了解决这个问题，使模型能够尽可能的利用语义信息，Zhong 等[23]提出了捕获从左到右的语义上下文，这种上下文感知能力有助于生成更加准确和一致的描述，从而使模型能够产生更好的字幕。为了更好地探索对象之间的关系，Li 等[24]在视频字幕和捕捉长期和短期依赖关系中提出了长短期关系转换器，通过这种方法缓解了过度平滑的问题，并加强了关系推理。为了解决视频数据与文本数据的语义鸿沟，Shi 等[25]提出了一种带有视频字幕检索单元的视频－文本对齐模块，该模型通过生成正确和独特的标题来减少视觉数据和文本数据之间的语义差距。

2 增强全局与局部特征融合的视频描述模型

本文提出了一种基于增强全局-局部特征融合的视频描述生成方法 EGLF，EGLF 是基于端对端编码器－解码器－重建器架构来构建视频描述模型，EGLF 模型框架图如图 1 所示。编码器部分包括局部特征提取和全局特征提取，局部特征又由静态特征和动态特征组成，编码器则使用的是注意力机制 Attention 和双向长短期记忆网络（bidirectional long short-term memory，Bi-LSTM）的方法，并且在编码器后加入了 Attention＋重建机制，通过利用重建机制，可以增强输入视频序列与描述语句的相关性，并再现视频序列的信息，最后生成更为准确的视频描述语句。

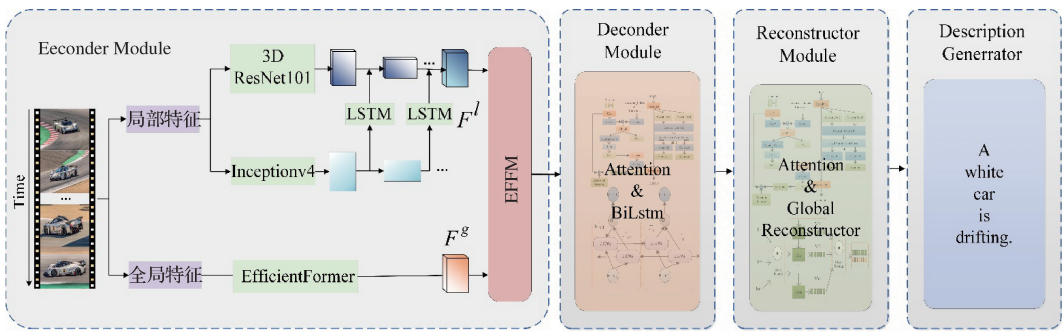

图 1 EGLF 模型框架图

2.1 视觉特征提取方法

1）局部特征提取

在 CNN 中，因为卷积运算擅长提取局部特征，因此 CNN 具有很强的局部上下文特征提取的能力。所以本文采用 CNN 提取视频的局部特征，考虑到静态特征和动态特征，本文使用了两种特征提取器进行局部特征提取。具体而言，给定视频序列帧 V 和对应的参考译文 $S=\{s_1,s_2,\cdots,s_L\}$，针对静态特征，首先采用在 ImageNet 数据集预训练的 inceptionV4 作为静态特征提取器来提取视频的静态特征 $C=\{c_1,c_2,\cdots,c_N\}$，使用 LSTM 建模时间的时序关系：

$$V^f=LSTM(F) \tag{1}$$

其次，采用在 Kinetics-400 数据集和 Moments in Time 视频数据集上预训练的 3D ResNet101 作为动态特征提取器来提取视频片段的动态特征 $M=\{m_1,m_2,\cdots,m_N\}$。最后，为了更好的将静态特征和动态特征进行结合，特征提取器每隔 24 帧提取一次特征，并将这个时刻提取的静态特征和动态特征用 cat 操作进行拼接，作为模型的局部特征，其计算公式为：

$$F^l=cat(V^{f_i},m_i) \tag{2}$$

式中：F^l 表示视频的局部特征；concat 表示拼接操作。

某个时刻的动作信息只是整个动作的一个小部分。因此，通过将同一时刻的静态特征和动态特征进行融合，可以将各个时刻微小的动作特征，整合成完整的动作特征。

2）全局特征提取

在 Vision Transformer 中，把视频帧切成小块然后编码成一个序列输入多头注意力中，去获取全局上下文信息，从而获得全局特征表示。由于传统的 Vision Transformer 计算复杂度高，计算效率较低，收敛慢。

因此，对于视频全局特征的提取，本文采用了轻量化模型 EfficientFormerV2 提取视频的全局特征。首先，视频先经过预处理，对每个视频提取 32 帧的有效视频帧，在视频帧分辨率上选择了 640×640，EfficientFormerV2 解码器的特征维度都是 512，每层有 4 个自注意头，通过加载在 ImageNet 数据集上预训练的 EfficientFormerV2 作为特征提取器来提取视频全局特征。通过提取全局特征可以提供关于整体样本分布的全面信息，能更好地描述样本的整体特性。

2.2 特征融合增强网络

全局特征主要表征视频中的整体内容和结构特征，而局部特征主要表征视频中的局部细节和特定目标，因此两者特征具有互补的关系。如果直接将二者进行拼接，则不能充分利用二者之间的互补关系，从而不能充分发挥各自的作用。

因此，为了建模不同级别特征（局部和全局）的相关性，利用特征融合增强网络 EFFM 对提取的特征进行融合，建模他们的语义关联性，过滤掉与任务无关或冗余的特征，使得两种语义交互过程中充分准确的表征出视频中的整体信息和细节信息。EFFM 具体如图 2 所示，首先，为了更好的融合局部特征和全局特征，对输入的全局特征 F^g 和局部特征 F^l 先做初始的特征融合：

$$X = (F^g \oplus F^l) \tag{3}$$

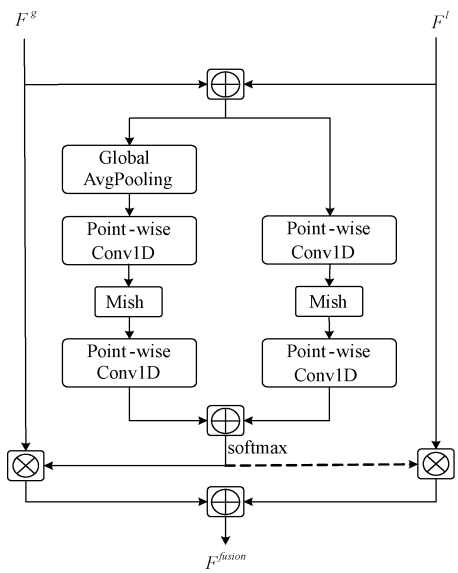

图 2 EFFM 模型框架图

再将融合后的特征分为左、右侧两路。右侧支路，针对特征中的局部特征进行处理。局部特征处理的计算公式为：

$$L(X) = PWConv_2(\delta(PWConv_1(X))) \tag{4}$$

式中：$PWConv_1$ 1×1 点卷积将输入 X 的特征通道数减少为原先的 $1/r$，δ 表示 mish 激活函数，通过 $PWConv_2$ 1×1 的卷积将通道数目恢复成与原输入通道数目相同。

左侧支路，针对特征中的全局特征进行处理，和局部特征处理不同的是，先对输入的特征 Z 进行一次全局平均池化操作，更注重关注全局上下文信息。全局特征处理的计算公式为：

$$G(X) = PWConv_4(\delta(PWConv_3(GAP(X)))) \tag{5}$$

式中：δ 表示 mish 激活函数；GAP 表示全局平均池化。

最后，将两条支路的特征作加权求和，获得局部特征和全局特征建模相关性后的融合特征。融合后的特征 F^{fusion} 计算公式为：

$$M = \delta(L(X) + G(X)) \tag{6}$$

$$F^{fusion} = M \otimes F^g \oplus (1-M) \otimes F^l \tag{7}$$

可分离卷积是文献提出的 MobileNets 模型中的高性能卷积结构[26]，能够在不损失卷积精度的情况下显著减少运算量。本文在 EFFM 中使用的卷积是一维深度可分离卷积神经网络（depthwise separable convolution neural network，1D-PCNN），一维卷积在时间序列和文本信息等序列数据上有强大的建模能力，捕捉到序列中的模式、趋势、周期性等信息，使用一维卷积可以更好的处理序列数据中的关系，这样的结构可以更好地捕捉序列数据的局部和全局信息，可以帮助解决序列相关的问题，从而提高模型的性能。在 EFFM 中使用 1D-PCNN 能够在保持良好性能的同时，也大大的压缩了模型的参数量。使用的激活函数是更为平滑 mish 激活函数，能够使信息更容易在神经网络中传递。

2.3 视频描述文本生成模型

1）解码器

在视频描述中，解码器的作用是将编码器生成的特征映射转化为人类可读的描述文本。本文使用的解码器是 Attention＋BiLSTM，其模型框架如图 3 所示。

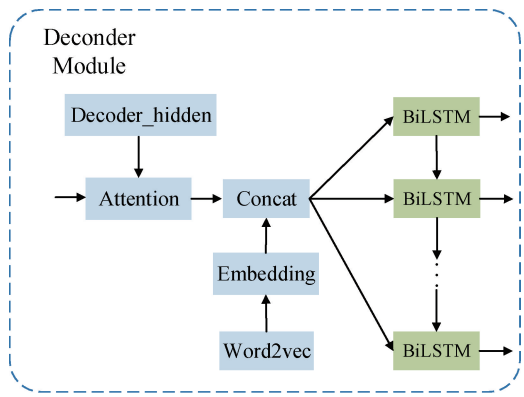

图 3 解码器模型框架图

对于每一个 LSTM 层，BiLSTM 模型的组成，分别包含了前向 LSTM 和后向 LSTM 两个部分，通过这两个部分，获得两个彼此独立的隐藏状态序列。假设输入序列为 $X = \{x_1, x_2, \cdots, x_n\}$，则有隐藏状态（前向）和隐藏状态（后向）的公式分别为：

$$h_t = LSTM\_forward(x_t, h_{t-1}) \tag{8}$$

$$\tilde{h}_t = LSTM\_backward(x_t, \tilde{h}_{t-1}) \tag{9}$$

式中：h_t 是前向 LSTM 在时间步 的隐藏状态；$\tilde{h}_t$ 是后向 LSTM 在时间步 t 的隐藏状态；$LSTM\_forward()$ 和 $LSTM\_backward()$ 表示前向 LSTM 层和后向 LSTM 层的前向传播过程。

通过利用 Attention 更新两个隐藏状态，可以得到更加准确的向量表示，即：

$$h_{j+1}, \tilde{h}_{j+1} = Attention[h_j, \tilde{h}_j] \tag{10}$$

通过结合注意力机制和 BiLSTM 作为解码器能够有

效的提升模型对上下文信息的准确理解和捕捉。这种结构在生成准确的、连贯的描述文本上具有优势，并且对于处理长序列和需要关注全局上下文的任务尤为有效。同时，这种结合还为模型提供了更稳健的建模能力和泛化能力，使其在处理多种时序数据和语义复杂性较高的任务时表现良好。

2)重构器

编码器只考虑了从源视频到目标句子的语义信息，但是未考虑从句子反生成视频的语义信息，为了解决从描述到视频的后向信息，受到编译码重构网络(reconstruction network，RecNet)[27]的启发，本文在解码器后加入了重构器，其模型框架如图 4 所示。重构器的组成是 Reconstructor 和 Attention，其中 Reconstructor 是由 LSTM 组成，Attention 使用的是坐标注意力机制。具体而言，首先用注意力策略选择解码器的关键隐藏状态中再现每个帧的特征表示：

$$\mu_t = \sum_{i=1}^{n} \beta_i^t h_i \tag{11}$$

式中：h_i 表示重构器在 t 时刻的隐藏状态；β_i^t 表示由注意力机制在时间步 t 为第 i 个隐藏状态计算的权重。重构器采取的是逐帧生成特征表示，重构器的损失函数为：

$$L_{rec} = \frac{1}{m} \sum_{j=1}^{m} \phi(v_j, z_j) \tag{12}$$

式中：m 为样本数量；z_j 表示重构器的隐藏状态；v_j 表示重建的到的视频特征。

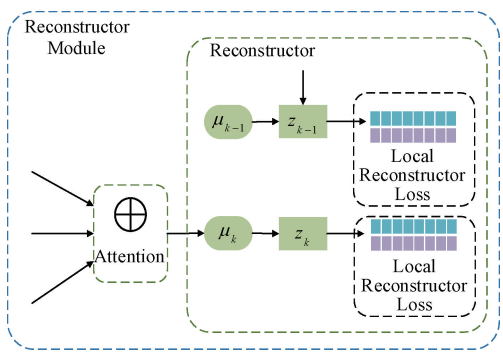

图 4 重构解码器模型框架图

在编码器后增加了重建器，通过使用重建网络和重建损失，可增强视频序列与描述语句间的关联性，使解码器嵌入更多的来自输入视频序列的信息，进而提高了视频序列与描述语句的关联性，促进生成更准确、更具有连贯性和语法准确性的描述。

3 实验结果与分析

3.1 数据集和评价标准

1)数据集

本文采用了 MSVD(microsoft research video description)数据集[28]和 MSR-VTT(microsoft research video-to-text)数据集[29]来验证模型的有效性。

MSVD 数据集是视频描述模型和算法的基准数据集之一，总共包含 1 970 个视频片段，每个视频片段均在 10～25 s 内描述任意领域的单个活动。其中，每个片段都附带了多个由人工生成的文字描述，将数据集进行划分，其中训练集有 1 200 张，验证集有 100 张，测试集有 670 张。

MSR-VTT 数据集。该数据集包含了 10 000 个多样化的 YouTube 视频片段，总计约 200 000 个视频描述句子。MSR-VTT 数据集视频时长更长，总计 41.2 h，内容涉及 20 多个领域，涵盖最全面的类别和多样化场景，包含多人场景和复杂场景。其中，使用数据集中的 497 个视频进行验证，2 970 个视频进行测试，6 513 个视频用于训练。

2)评价标准

为了验证视频描述任务中描述模型生成的语句质量，本文采用了 4 种评价指标进行验证，包括 BLEU[30]、METEOR[31]、ROUGE-L[32]、CIDEr[33]，这 4 种指标的分数越高，说明模型生成的描述语言更为准确。BLEU 用来比较候选译文和参考译文里的 n-gram 的差异，其计算公式为：

$$BLEU = BP \times \exp\left(\sum_{i=1}^{N} \frac{1}{2^n} \log P_n\right) \tag{13}$$

式中：BP 是为了惩罚较短的输出语句而设置的，若翻译结果与参考语句长度相等，则惩罚因子设为 1.0。P_n 是词数为 n 的子序列的精度，BLEU 的取值范围是[0，1]，BLEU 值越高表示两个句子越相似，模型的翻译能力更强。

METEOR 与 BLEU 相比，更全面地考虑译文与参考翻译之间的语义和结构相似性。其计算公式为：

$$F_{\text{mean}} = \frac{(\alpha^2 + 1) P_m}{R_m + \alpha P_m} \tag{14}$$

$$METEOR = F_{\text{mean}} \left(1 - \gamma \left(\frac{ch}{m}\right)^{\theta}\right) \tag{15}$$

通常 γ、θ、α 设置为默认参数，F_{mean} 为参考文本和模型生成文本之间的准确率 P_m 和召回率 R_m 的调和平均值，m 为 n-gram 对应的个数。ch 为 chunk 的个数，数目越少意味着每个 chunk 的平均长度越长，也就是说候选译文和参考译文的语序越一致。

本文使用 ROUGE-L 评价指标，衡量生成语句的流畅度，其计算公式为：

$$R_{lcs} = \frac{LCS(X, Y)}{m} \tag{16}$$

$$P_{lcs} = \frac{LCS(X, Y)}{n} \tag{17}$$

$$F_{lcs} = \frac{(1 + \beta) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{18}$$

式中：X 表示标准答案；Y 表示生成答案；m 和 n 分别表示 X 的 Y 的长度；$LCS(X, Y)$ 表示 X 的 Y 的最长公共序列；β 为超参数，需要自己设置；F_{lcs} 为 ROUGE-L 的得

分，F_{lcs} 越高说明模型的生成语句内容更为完整。

CIDEr 通过综合考虑词频、权重和相似度来评估视频描述的质量，其计算公式如下：

$$CIDEr_n(c_i,S_i)=\frac{1}{m}\sum_j\frac{g^n(c_i)\cdot g^n(s_{ij})}{\|g^n(c_i)\|\cdot\|g^n(s_{ij})\|} \tag{19}$$

式中：$g^n(c_i)$ 是由 n-gram 取长度 n 时 $g^n(c_i)$ 组成的向量，$\|g^n(c_i)\|$ 表示该向量的模；$\|g^n(s_{ij})\|$ 含义类似；m 对应视频 i 有多少条参考语句。$CIDEr_n$ 的值越高，说明模型生成的语句更为多样且更接近人类的描述。

3.2 实验环境及相关参数

训练的硬件平台是搭载了 Intel(R) Core(TM) i5-10400F CPU@2.90 GHz 和 NVIDIA GeForce GTX1080 Ti GPU，64 G 内存通的 PC。在训练阶段，采用 AdamW 作为优化器。其中，学习率设置为 5×10^{-5}，词向量和编码器的输入维度均为 468，隐藏层单元为 512，batch size 设置为 32。同时，epoch 设置为 50，如果连续 20 个 epoch，验证数据集上的 CIDEr 指标分数没有增加，训练停止。生成描述语句时，采取的是束宽大小为 5 的集束搜索方法。

3.3 消融实验结果分析

本文采用了 MSVD 和 MSR-VTT 这两种数据集在 4 种评价指标上验证模型的有效性，指标分数的单位为%，指标的分数越高，说明模型生成的描述语言更为准确。其中，假设模型局部特征＋解码器＋重构器为 LDR，模型全局特征＋解码器＋重构器为 GDR，模型全局-局部特征(直接将特征 cat 拼接)＋解码器＋重构器为 FDR，模型全局-局部特征＋EFFM＋解码器＋重构器为 EGLF。

表 1 和 2 分别为基于 MSVD 数据集和 MSR-VTT 数据集的消融实验。从表 1、2 可以看出，无论是使用了 EFFM 进行特征融合还是简单的将特征进行 cat 拼接，模型各项指标都明显比单独用提取局部特征或者全局特征的效果好，说明通过融合局部特征和全局特征，可以得到更强大的表征能力，使生成的描述语句更加准确。

表 1 基于 MSVD 数据集上模型的消融实验

模型	BLUE_1	BLUE_4	ROUGE_L	CIDEr	METEOR
LDR	83.52	52.73	71.48	83.28	36.39
GDR	81.60	50.31	70.20	79.61	35.30
FDR	84.71	53.10	72.66	85.95	36.83
EGLF	**85.33**	**54.87**	**73.92**	**87.69**	**36.71**

表 2 基于 MSR-VTT 数据集上模型的消融实验

模型	BLUE_1	BLUE_4	ROUGE_L	CIDEr	METEOR
LDR	76.28	41.69	61.43	50.79	29.20
GDR	73.03	41.25	55.90	48.51	28.11
FDR	76.65	42.40	61.97	51.10	29.64
EGLF	**78.65**	**43.83**	**62.20**	**51.12**	**29.87**

另外，加入了 EFFM 之后的 EGLF 模型，在常用的 METEOR、BLEU、ROUGE-L 和 CIDEr 评价指标上也都有了很大的提升，特别是 BLUE 和 ROUGE_L 这两个指标。其中，EGLF 模型在两个数据集上的 BLUE_4 分别提升了 1.77%和 1.43%，说明了加入 EFFM 可以建模不同级别特征(局部和全局)之间的相关性，能够丰富模型的特征信息，提升模型的性能和泛化能力，从而生成更加准确的视频描述语句。

3.4 对比实验结果分析

本文将提出的方法与近年主流算法在 MSVD 数据集和 MSR-VTT 数据集上进行比较，通过比较来验证所提方法的有效性和优越性。其中，进行比较的主要方法有 ORG-TRL[34]、MMI[35]、DSA-LSTM[36]、CSA-SR[37]、SGN(R101＋RN)[38]、MDT[39]、MABVC[40]、GSB-CoSB[18]、IAT[9]、C-SeqGAN[41]、SA-VA[42]、R-ConvED[19]、STG-KD[43]、SAAT[44]、NACF[45]、DVC-Net[46]、Transformer-LSTM-RL[8]、VMSG[47]。

表 3 和 4 分别代表 EGLF 与其他主流算法在 MSVD 数据集和 MSR-VTT 数据集上的性能对比，其中“—”代表没有该指标的数值。从表 3、4 可以看出，本文所提出的方法在两个数据集上其他近几年的视频描述模型相比，在各项评价指标中展现了所提方法优越的性能，该实验结果进一步证明了本文所提方法的有效性。

表 3 EGLF 与其他主流算法在 MSVD 数据集上的性能对比

模型	年份	BLUE_4	ROUGE_L	CIDEr	METEOR
ORG-TRL	2020	54.30	73.90	95.20	36.40
MMI	2020	46.70	65.00	76.80	33.60
DSA-LSTM	2021	49.53	68.93	78.31	34.35
CSA-SR	2021	52.20	72.70	83.40	35.60
SGN (R101＋RN)	2021	52.80	72.90	94.30	35.50
MDT	2021	52.00	72.80	92.60	36.50
MABVC	2022	54.60	73.60	95.20	36.70
GSB-CoSB	2022	50.7	72.10	**97.80**	35.30
IAT	2022	53.80	73.20	92.70	36.00
C-SeqGAN	2023	54.82	71.60	83.40	35.90
SA-VA	2023	53.50	—	83.30	—
R-ConvED	2023	54.82	71.6	83.40	35.9
EGLF(本文)	—	**54.87**	**73.92**	87.69	**36.71**

3.5 实验结果定性分析

为了更直观地证明本文的模型能够生成更准确的视频描述语句，通过对模型描述视频片段的环境区域、对象细节、句子丰富度和上下文连贯性等方面进行分析比较，来验证模型的有效性。EGLF、LDR、GDR、FDR 4 种模型在 MSVD 数据集和 MSR-VTT 数据集上生成的一些描述示例对比如图 5、6 所示。

表 4　EGLF 与其他主流算法在 MSR-VTT 数据集上的性能对比

模型	年份	BLUE_4	ROUGE_L	CIDEr	METEOR
MMI	2020	39.30	61.20	44.60	28.50
STG-KD	2020	40.50	60.90	47.10	28.30
SAAT	2020	40.05	60.90	49.10	28.2
NACF	2021	42.00	—	51.40	28.70
SGN	2021	40.08	60.80	49.50	28.30
DVC-Net	2021	41.80	60.10	—	—
MABVC	2022	43.80	60.20	52.50	29.60
Transformer-LSTM-RL	2022	42.00	62.00	**54.20**	—
VMSG	2023	41.00	60.80	49.00	—
SA-VA	2023	43.20	—	51.30	—
EGLF（本文）	—	**43.83**	**62.20**	51.12	**29.87**

视频片段	
参考语句	A young man is making a ghost costume out of a sheet.
LDR	A man is drawing on a sheet.
GDR	A person is drawing a face on a sheet.
FDR	A boy is drawing a picture on a cloth.
EGLF	A young man draws on cloth with sketch pen.

(a) 示例 1

视频片段	
参考语句	A fat man eating breads in his kitchen.
LDR	A man is eating bread.
GDR	A man is eating something.
FDR	A man is eating a piece of bread.
EGLF	A man is eating bread in his kitchen.

(b) 示例2

图 5　各模型在 MSVD 数据集上生成描述示例对比

从图 5 的对比结果可以看出，本文提出的 EGLF 模型，相比于其他的 3 个模型，可以更加精准的描述视频片段中的对象信息，图 5(a)中，EGLF 能够准确描述目标对象(“young man”)和(“pen”)。可以精准的描述目标对象使用“pen”这个工具进行绘画，而不是简单的描述在绘画，说明 EGLF 模型能够关视频片段注更多的细节信息。本文提出的 EGLF 模型，相比于其他的 3 个模型，可以更加准确描述视频片段中的目标对象所处的环境，图 5(b)中，可以准确描述目标对象的所处的环境(“kitchen”)，验证了本文提出的 EGLF 模型不仅能关注全局信息，还能关注局部

视频片段	
参考语句	two players are fighting on stage to win a wrestling match.
LDR	two men wrestle in a competition .
GDR	two men wrestle in a competition .
FDR	two competitors wrestle on the ground .
EGLF	two men are fighting on stage to win .

(a) 示例1

视频片段	
参考语句	A basketball game is going on between two teams wearing red and white .
LDR	Basketball players are playing basketball .
GDR	An nba basketball game .
FDR	A basketball team is playing a game .
EGLF	A basketball game is going on between two teams .

(b) 实例2

图 6　各模型在 MSR-VTT 数据集上生成描述示例对比

细节信息。

从图 6 可以看出，本文提出的 EGLF 模型，相比于其他的 3 个模型，能够更加关注视频片段的全局上下文信息。图 6(a)中，可以准确的描述“win”，而不是只关注前面的“wrestle”。说明 EGLF 模型与其他 3 个模型的描述语句相比，对全局上下文信息的捕获能力更强。本文提出的 EGLF 模型，可以生成更加丰富的描述语句。图 6(b)中，其他的模型只能描述“play basketball”，但是 EGLF 模型能够描述出是在“two teams”，而不是仅仅描述“play basketball”，说明 EGLF 模型与其他 3 个模型的描述语句相比，在描述视频片段的丰富度上要略胜一筹。

通过上述视频描述生成语句的示例对比中，可以得出本文提出的 EGLF 模型，无论是在单一的活动场景，还是在长时间、多人的复杂场景中，EGLF 模型生成的自然描述语句更为准确且更符合人类的描述习惯。

4　结　论

本文针对视频描述生成任务，提出了一种基于增强全局-局部特征融合的视频描述生成方法。该方法通过将局部特征与全局特征进行融合，充分利用局部和全局特征的互补性，帮助模型捕捉更多关键信息。为了建模不同级别特征的相关性，提高特征的表达能力，利用特征融合增强网络进行特征融合。最后，进行了一系列的实验和评估，验证了本文提出的方法在 MSVD 视频数据集和 MSR-VTT 数据集上的有效性和鲁棒性。

现有的视频描述模型大多都是主要关注视觉特征提取的问题，未来的研究可以进一步探索如何结合视觉和语

义信息，以及如何应用深度强化学习和无监督学习等方法来进一步提升视频描述的质量和效果。

参考文献

[1] VENUGOPALAN S, ROHRBACH M, DONAHUE J, et al. Sequenceto sequence-video to text [C]. Proceedings of theIEEE International Conference on Computer Vision, 2015: 4534-4542.

[2] BARALDI L, GRANA C, CUCCHIARA R. Hierarchical boundary aware neural encoder for video captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1657-1666.

[3] PAN B, CAI H, HUANG D A, et al. Spatio-temporal graph for video captioning with knowledge distillation [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10870-10879.

[4] RYU H, KANG S, KANG H, et al. Semantic grouping network for video captioning [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2514-2522.

[5] BABAVALIAN M R, KIANI K. Learning distribution of video captions using conditional GAN[J]. Multimedia Tools and Applications, 2023, DOI:10.1007/s11042-023-15933-6.

[6] 闫超，贾振堂. 基于 Transformer 与增强信息融合的双源情感识别[J]. 国外电子测量技术，2023，42(4)：187-193.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. Computer Science, 2020, DOI:10.48550/arXiv.2010.11929.

[8] ZHAO H, CHEN Z,GUO L, et al. Video captioning based on vision transformer and reinforcement learning[J]. PeerJ Computer Science, 2022, 8: e916.

[9] JIN T, ZHAO Z, WANG P, et al. Interaction augmented transformer with decoupled decoding for video captioning[J]. Neurocomputing, 2022, 492: 496-507.

[10] 戴俊，王俊，朱忠奎，等. 基于生成对抗网络和自动编码器的机械系统异常检测[J]. 仪器仪表学报，2019，40(9)：16-26.

[11] 曹磊，万旺根，侯丽. 基于多特征的视频描述生成算法研究[J]. 电子测量技术，2020，43(16)：99-103.

[12] 黄先开，张佳玉，王馨宇，等. 密集视频描述研究方法综述[J]. 计算机工程与应用，2023，59(12)：28-48.

[13] ZOLFAGHARI M, SINGH K, BROX T. Eco: Efficient convolutional network for online video understanding [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 695-712.

[14] PAN Y, MEI T, YAO T, et al. Jointly modeling embeding and translation to bridge video and language[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 4594-4602.

[15] BILKHU M, WANG S, DOBHAL T. Attention is all you need for videos: Self-attention based video summarization using universal transformers [J]. Computer Science, 2019, DOI: 10.48550/arXiv.1906.02792.

[16] 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.

[17] SEO P H, NAGRANI A, ARNAB A, et al. End-to-end generative pretraining for multimodal video captioning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 17959-17968.

[18] VAIDYA J, SUBRAMANIAM A, MITTAL A. Co-Segmentation aided two-stream architecture for video captioning[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2774-2784.

[19] CHEN J, PAN Y, LI Y, 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.

[20] 丁志江，李丹，马志程，等. 基于 Transformer 的车道线分割算法研究[J]. 电子测量与仪器学报，2022，36(10)：227-234.

[21] SUN C, MYERS A, VONDRICK C, et al. Video bert: A joint model for video and language representation learning[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7464-7473.

[22] LUO H, JI L, SHI B, et al. UniVL: A unified video and language pretraining model for multimodal understanding and generation[J]. Computer Science, 2020, DOI:10.48550/arXiv.2002.06353.

[23] ZHONG M, ZHANG H, WANG Y, et al. BiTransformer: Augmenting semantic context in video captioning via bidirectional decoder[J]. Machine Vision and Applications, 2022, 33(5): 77.

[24] 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.

[25] SHI Y, XU H, YUAN C, et al. Learning video-text aligned re-presentations for video captioning[J]. ACM Transactions on Multimedia Computing, Communica-

tions and Applications,2023, 19(2): 1-21.
[26] HAASE D, AMTHOR M. Rethinking depth wise separable convolutions: How intra-kernel correlations leadto improved mobilenets[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 14600-14609.
[27] WANG B, MA L, ZHANG W, et al. Reconstruction net-work for video captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018: 7622-7631.
[28] CHEN D, DOLAN W B. Collecting highly parallel data for paraphrase evaluation[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,2011:190-200.
[29] XU J, MEI T, YAO T, et al. MSR-VTT: A large video description dataset for bridging video and language[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 5288-5296.
[30] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: A method for automatic evaluation of machine translation [C]. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,2002:311-318.
[31] BANERJEE S, LAVIE A. METEOR: An automatic metricfor 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, 2005: 65-72.
[32] ROUGE L C Y. A package for automatic evaluation of summaries[C]. Proceedings of Workshop on Text Summarization of ACL, 2004: 74-81.
[33] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image description evaluation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4566-4575.
[34] ZHANG Z, SHI Y, YUAN C, et al. Object relationalgraph with teacher-recommended learning for videocaptioning [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 13278-13288.
[35] 丁恩杰,刘忠育,刘亚峰,等.基于多维度和多模态信息的视频描述方法[J].通信学报,2020,41(2):36-43.
[36] 王金金,曾上游,李文惠,等.基于扩张卷积的注意力机制视频描述模型[J].电子测量技术,2021,44(23):99-104.
[37] LEI Z, HUANG Y. Video captioning based on channelsof attention and semantic reconstructor[J]. Futureinternet,2021, 13(2): 55.
[38] RYU H, KANG S, KANG H, et al. Semantic grouping network for video captioning [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2514-2522.
[39] ZHAO W, WU X, LUO J. Multi-modal dependency tree for video captioning[J]. Advances in Neural Information Processing Systems, 2021, 34: 6634-6645.
[40] 李铭兴,徐成,李学伟,等.基于多模态融合的城市道路场景视频描述模型研究[J].计算机应用研究,2023,40(2):607-611,640.
[41] BABAVALIAN M R, KIANI K. Learning distribution of video captions using conditional GAN[J]. Multimedia Tools and Applications, 2023, DOI:10.1007/s11042-023-15933-6.
[42] FU Y, WANG M, YE O. Video captioning based on scenerepresentation object features syntax analysis[J]. Computer Engineering and Desing, 2023, 44(2): 488-493.
[43] PAN B, CAI H, HUANG D A, et al. Spatio-temporal graphfor video captioning with knowledge distillation [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10870-10879.
[44] ZHENG Q, WANG C, TAO D. Syntax-aware action targeting for video captioning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020:13093-13102.
[45] YANG B, ZOU Y, LIU F, et al. Non-autoregressive coarse-to-fine video captioning[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3119-3127.
[46] LEE S J, KIM I. DVC-Net: A deep neural network model for dense video captioning[J] IET Computer Vision,2021,15(1):12-23.
[47] YANG X, WANG X, YE X, et al. VMSG: A video caption network based on multimodal semantic grouping and semantic attention [J]. Multimedia Systems, 2023, DOI: 10.1007/s00530-023-01124-8.

作 者 简 介

黄飞燕,硕士研究生,主要研究方向为自然语言处理、计算机视觉等。
E-mail:1553329469@qq.com
曾上游(通信作者),硕士生导师,教授,主要研究方向为非线性动力学、人工神经网络、复杂网络等。
E-mail:zsy@mailbox.gxnu.edu.cn
邱泓语,硕士研究生,主要研究方向为自然语言处理、计算机视觉、人工智能等。
E-mail:qqj872377086@163.com