软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(12):5905−5920 [doi: 10.13328/j.cnki.jos.006736]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

语言结构引导的可解释视频语义描述*

李冠彬[1,2], 张锐斐[1], 刘梦梦[1], 刘 劲[1], 林 倞[1]

[1](中山大学 计算机学院, 广东 广州 510006)

[2](人工智能与数字经济广东省实验室 (广州), 广东 广州 510320)

通信作者: 林倞, E-mail: linliang@ieee.org

摘 要: 视频描述技术旨在为视频自动生成包含丰富内容的文字描述, 近年来吸引了广泛的研究兴趣. 一个准确而精细的视频描述生成方法, 不仅需要对视频有全局上的理解, 更离不开具体显著目标的局部空间和时序特征. 如何建模一个更优的视频特征表达, 一直是视频描述工作的研究重点和难点. 另一方面, 大多数现有工作都将句子视为一个链状结构, 并将视频描述任务视为一个生成单词序列的过程, 而忽略了句子的语义结构, 这使得算法难以应对和优化复杂的句子描述及长句子中易引起的逻辑错误. 为了解决上述问题, 提出一种新颖的语言结构引导的可解释视频语义描述生成方法, 通过设计一个基于注意力的结构化小管定位机制, 充分考虑局部对象信息和句子语义结构. 结合句子的语法分析树, 所提方法能够自适应地加入具有文本内容的相应时空特征, 进一步提升视频描述的生成效果. 在主流的视频描述任务基准数据集 MSVD 和 MSR-VTT 上的实验结果表明, 所提出方法在大多数评价指标上都达到了最先进的水平.

关键词: 视频描述; 编码器-解码器架构; 小管; 注意力机制; 依存分析

中图法分类号: TR391

中文引用格式: 李冠彬, 张锐斐, 刘梦梦, 刘劲, 林倞. 语言结构引导的可解释视频语义描述. 软件学报, 2023, 34(12): 5905–5920. http://www.jos.org.cn/1000-9825/6736.htm

英文引用格式: Li GB, Zhang RF, Liu MM, Liu J, Lin L. Interpretable Video Captioning Guided by Language Structure. Ruan Jian Xue Bao/Journal of Software, 2023, 34(12): 5905–5920 (in Chinese). http://www.jos.org.cn/1000-9825/6736.htm

Interpretable Video Captioning Guided by Language Structure

LI Guan-Bin[1,2], ZHANG Rui-Fei[1], LIU Meng-Meng[1], LIU Jing[1], LIN Liang[1]

[1](School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China)

[2](Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou 510320, China)

Abstract: Video description technology aims to automatically generate textual descriptions with rich content for videos, and it has attracted extensive research interest in recent years. An accurate and elaborate method of video description generation not only should have achieved a global understanding of the video but also depends heavily on the local spatial and time-series features of specific salient objects. How to model a better video feature representation has always been an important but difficult part of video description tasks. In addition, most of the existing work regards a sentence as a chain structure and views a video description task as a process of generating a sequence of words, ignoring the semantic structure of the sentence. Consequently, the currently available algorithms are unable to handle and optimize complex sentence descriptions or avoid logical errors commonly seen in the long sentences generated. To tackle the problems mentioned above, this study proposes a novel generation method for interpretable video descriptions guided by language structure. Due consideration is given to both local object information and the semantic structure of the sentence by designing an attention-based structured tubelet localization mechanism. When it is incorporated with the parse tree constructed from sentences, the proposed method can adaptively attend to corresponding spatial-temporalfeatures with textual contents and further improve the performance of video description

* 基金项目: 国家自然科学基金 (61976250, U1811463); 广东省基础与应用基础研究基金 (2020B1515020048)
收稿时间: 2021-06-24; 修改时间: 2021-11-08, 2022-02-17; 采用时间: 2022-04-20; jos 在线出版时间: 2023-05-18
CNKI 网络首发时间: 2023-05-19

generation. Experimental results on mainstream benchmark datasets of video description tasks, i.e., Microsoft research video captioning corpus (MSVD) and Microsoft research video to text (MSR-VTT), show that the proposed approach achieves state-of-the-art performance on most of the evaluation metrics.

Key words: video captioning; encoder-decoder framework; tubelet; attention machanism; dependency parsing

视频描述是指自动在给定视频中生成包含丰富内容的文字描述的任务. 该任务自然将视觉内容理解与自然语言处理结合在一起, 为各种实际应用 (如人机交互, 视频检索和仪表导航) 创造了可能性. 尽管过去几年来它在计算机视觉和人工智能领域受到越来越多的关注, 并且取得了长足的进步, 但视频描述仍然是一个具有挑战性的问题, 具有很大的改进空间.

早期的视频描述方法[1–3]通常是基于模板的, 这些方法在特定的语法规则中预定义了一系列的句子模板, 然后使用基于对象检测和动作识别的结果生成的语义实体填充模板. 由于固定句子模板的局限性, 这类方法不能进一步为涉及复杂对象关系和多种人与对象交互的视频生成各种准确的描述. 近些年来, 随着深度神经网络的发展, 特别是循环神经网络 (RNNs), 大量的端到端的基于深度学习的视频描述框架[4–8]应运而生. 在这些基于序列学习的方法中, 最流行的框架是编码器-解码器框架, 该框架采用由 CNN 和 RNN 组成的编码器, 以获取每个输入视频的时空表示, 然后输送到另一个基于 RNN 的解码器中逐字生成描述性句子. 借鉴于图像描述任务[9], 一种用于视频描述的软注意力机制[10]被提出并广泛用于其他研究中[11,12], 该模型可选择性地关注相关的视觉内容, 以生成特定的单词序列. 具体而言, 多层感知机 (MLP) 用于计算采样帧或整个视频帧的相等大小区域上的注意力权重. 然而, 这些注意力模块仅将关注目标限制在静态图像帧或帧的子集中的显著区域, 而忽略了物体的时空信息. 近年来, 陆续的相关工作[13,14]开始将研究重点放在捕获显著物体的局部空间特征和时序特征上. 本文认为, 有效考虑这些时空性提示对于更准确地生成视频描述至关重要. 例如, 假设看到一个人拿着盘子, 我们必须先观察手和盘子的运动轨迹序列, 然后才能得出结论, 他将要洗碗还是吃饭. 同时, 通过捕获详细的视觉信息获取充分的局部特征, 可以防止模型产生粗糙的描述, 例如生成“在碗中搅拌配料”而不是“烹饪”.

另一方面, 大多数现有工作都将句子视为一个链状结构, 并将视频描述任务建模为一个生成单词序列的过程, 而忽略了句子的语义结构, 这使得算法对解决复杂的句子描述生成和逻辑错误优化等问题上变得困难. 实际上, 句子结构暗示了句子中与单词相对应的对象之间的显著性和关联性, 并决定了在生成后续单词的过程中要使用哪些信息, 这自然地与注意力方案相匹配. 例如, 根据主语-谓语-宾语结构, 当使用运动特征生成动词时, 通常需要一个宾语特征, 并且在大多数情况下, 宾语将位于与主语和动词位置不同的区域中, 表明那些看不见的已识别对象值得关注. 然而, 将视频描述模型与句子结构信息相结合仍然是一个挑战, 因为在推理过程中整个句子都是未知的, 这使得准确地获取预定义的语义结构来引导注意力变得很困难.

为了解决上述问题, 本文通过在基于 LSTM 的主流编码器/解码器框架中结合树形结构的注意区域定位模块, 提出了一种语言树形结构引导的注意力小管编码器-解码器框架 (tree-structured attentive tubelet encoder-decoder model, TSAT). 它不仅为单词预测提供了更精确的语义结构化指导, 而且还通过更精确的视觉注意力赋予了生成的句子的可解释性. 在本文方法中, 一个预训练的基于状态的解析器被用于在生成的单词序列的基础上来挖掘句子结构信息. 解析器与常规的注意力模块结合在一起, 可以利用句子的语义结构来自适应地调整相关的局部特征以生成后续的单词. 为了对视频对象的外观和动向进行建模, 将序列上的候选对象边界框 (称为小管) 用作统一的局部时空表示, 并进一步设计了一个基于注意力的帧选择编码器来克服漂移问题. 具体来说, 本文提出的方法由 3 个阶段组成: 一个用于给定视频的全局和局部特征提取的预处理阶段, 一个基于注意力的帧选择编码器和一个句子结构感知解码器. 在预处理阶段, 从给定的视频中提取全局特征以及一组局部小管特征. 每个小管都会很好地捕获视频中某个对象或对象部分的外观和运动信息, 并将其送入一个基于注意力机制的编码器模块. 在编码阶段, 模型对提取的特征进行处理, 以提供紧凑和有代表性的视觉内容进行解码. 在解码阶段, 通过动态改变视频中候选时空区域的视觉特征向量以生成准确的句子描述. 其中, 基于从句子中解析的结构和对应的单词与相关小管的映射, 探索短语级的相应时空特征来模拟运动对象的整体结构, 作为预测后续单词的候选注意特征.

本文的主要贡献有以下 3 个方面.

• 本文提出了一种新颖的树形结构的注意力小管编码器-解码器框架, 通过设计一个基于注意力的结构化小管定位机制, 充分考虑了局部对象信息和句子语义结构. 结合句子的语法分析树, 所提出的方法能够自适应地加入具有文本内容的相应时空特征, 并为视频生成更精细和准确的描述.

• 注意力小管定位机制利用句子结构信息, 极大地提高了模型的可解释性. 而使用硬注意力机制的帧选择模块可以看作是解决视频内容漂移问题的有效方案, 可以很容易地扩展到其他视频任务中, 帮助获得更细粒度的对象特征, 增强视觉表示能力.

• 针对主流的视频描述任务基准数据集进行了大量实验. 实验结果表明, 本文提出的方法在大多数评价指标上都达到了最先进的水平. 另外, 在消融研究中还充分探索了各个模块的有效性.

本文第 1 节回顾了相关工作. 第 2 节介绍了本文提出的方法. 广泛的实验结果和比较结果在第 3 节中进行了介绍, 最后, 第 4 节对本文进行了总结.

1 相关工作

1.1 视频描述

现有的视频描述方法可以大致分为两大类, 包括基于模板的方法[1–3]和基于深度学习的方法[13–22].

基于模板的方法通常使用特定的语法规则预先定义一系列的句子模板, 然后通过视频分析 (例如视频对象检测, 动作识别等) 提取各种实体 (例如主语, 谓语, 宾语等) 来填充模板以获得生成的句子. 其中, 一种代表性的方法[23]首先提取人类动作的语义特征并使其与分层动作概念相对应, 然后使用这些元素确定句法成分并翻译成自然语言句子. 由于它严重依赖于手工设计的动作概念层次结构, 因此这项工作只能限于特定领域. 在文献 [24] 中作者提出了一种结合对象和活动检测器的整体数据驱动技术, 以选择最可能的主语-谓语-宾语三元组来描述视频. 另一种相关方法[1]学习了条件随机场 (CRF), 以对不同视觉组件之间的关系进行建模并生成视频描述. 但是, 由于句子模板缺乏灵活性, 因此该算法范式不能扩展为开放域视频生成更多样化和准确的描述.

随着深度学习的盛行, 视频描述领域也迎来革命性进步. 基于 CNN 和 RNN 的端到端无模板算法使生成的语句描述更加自然和详细. 长短期记忆网络 (LSTM) 作为 RNN 的一种变体, 用于克服梯度爆炸和梯度消失问题, 并且在句子生成中表现出更高的效率. 例如, 在文献 [5] 中提出将每个视频帧的平均池化特征输入 LSTM 以生成句子. 最近引入的编码器-解码器框架大幅改善了视频描述的性能表现. 该框架的基准[4]使用基于 CNN 的方法来提取视频帧特征, 由编码器 LSTM 对特征进行编码以生成视频的紧凑表示, 然后将其传递给解码器 LSTM 并生成单词序列. 受到注意力机制在图像描述中巨大成功的启发, 该框架与注意力建模进一步结合在一起, 可以有选择地对时空特征进行加权, 以提高针对显著目标对象生成的描述的准确性. 还有一些工作[7,10]借助软注意力机制聚焦视频中的特定帧, 来提升视频描述的生成效果. 然而, 这些模块仅在帧级别上进行注意力聚集, 而无法实现对局部显著物体时空信息的关注. 为了解决这一问题, 近年来, 许多工作开始考虑显著物体的局部特征提取来提升模型性能表现. OA-BTG[13]通过构建双向的时序图模型来捕获显著物体详细的时序动态变化, HTM[14]模型在编码层使用两个 LSTM 层依次在帧级别和物体级别构建时序结构, 在解码阶段使用相应的 LSTM 层通过多级注意力机制从模糊的全局特征到细粒度的局部特征动态提取关键信息. ORG-TRL[15]模型通过构建物体关系图融入视频中物体之间的交互信息来进一步增强物体级别的特征表达.

1.2 视觉特征表示

除了神经网络的设计之外, 视觉特征表示在计算机视觉任务 (包括视频描述) 中也起着不可替代的作用. 传统的视觉特征是手工制作的, 例如 HoG[25] , HoF[26]和 SIFT[27], 被广泛用于各种基本的计算机视觉应用如图像分类, 目标检测和语义分割. 在更丰富的数据集和更深层的网络体系结构的推动下, 越来越多的基于深度学习的视觉描述符被提出, 性能优于那些手工制作的特征. 在图像描述任务中, 通常采用深度卷积神经网络提取图像特征, 例如

VGG[28]和 ResNet[29]. 与静止图像描述不同, 外观和时间线索对于生成视频描述都是必不可少的, 需要视频描述算法在捕捉视觉上显著的内容的同时捕获目标动作, 并对其与其他对象单元或环境上下文的关系进行建模以产生相应的描述说明. 近年来, 部分工作[15−17]尝试学习更好的视觉特征表示来提升视频描述的效果, 然而, 表征学习上的困难依然是阻止这类问题被彻底研究的主要原因.

密集轨迹[30]及其改进版本: 改进的密集轨迹[31]描述了具有密集轨迹的视频, 这些轨迹是通过从视频中的每个帧采样密集点并根据密集光流场的位移信息对其进行跟踪的, 在视频识别和常规动作分类任务中取得了出色的性能. 另一项工作[32]提出了轨迹合并的深度卷积描述符 (TDD), 该方法结合了手工特征和深度学习特征的优点, 以实现视频动作识别中的有效描述符. 近年来, 一种名为“小管 (tubelet)”的新的视频对象表示得到了广泛使用, 并在各种视频任务中取得了很大的成功. 与静态图像目标检测任务中的边界框 (bounding box) 概念类似, 在视频任务上将其称之为小管, 即候选对象在视频中跨时间维度上的边界框组成的序列, 继而包含该对象完整的时空信息. T-CNN[33]对静止图像物体检测执行动作感知传播以生成小管. SiamMask[34]仅依靠单帧的边界框初始化, 通过学习相邻帧之间的相似性度量生成小管, 而不用在每一帧中检测目标, 表现出了更高的效率. 本文利用小管表示来捕捉视频中目标的外观特征和运动信息, 经验证明这可以提高视频描述的性能.

1.3 句子解析

除了单纯考虑视觉信息之外, 一些研究工作还尝试探索句子所承载的信息. 例如, 一种图像描述方法[35]将给定图像解析为关键的语义实体及其关系, 它们被视为描述视觉内容的重要线索, 并被用来提高描述性能. 另一个图像描述研究工作[36]试图为句子中的短语和图像中的显著区域构建表示形式, 并通过层次结构学习它们的对应关系. 所有这些现有方法证明, 可以利用句子解析来改善描述. 本文首次尝试将视频的时空结构与句子解析结合在一起, 以实现更加精细的视频内容和动作描述.

2 设计方法

本节将详细介绍提出的语言树结构引导的注意力小管编码器/解码器 (TSAT) 框架. 它被定义为一个基于 LSTM 的编码器-解码器架构, 并结合了一个动态区域管组库和一个用于小管选择和更新的结构化注意力机制, 其总体架构如图 1 所示, 它由 3 个阶段组成: 从视频中进行目标小管提取的预处理阶段, 用于获取全局帧序列和局部目标小管特征表示的编码器阶段, 以及后续的解码阶段, 通过在句子解析结构的引导下自适应地关注适当的区域小管来生成视频描述. 本节将详细介绍此框架中的所有模块.

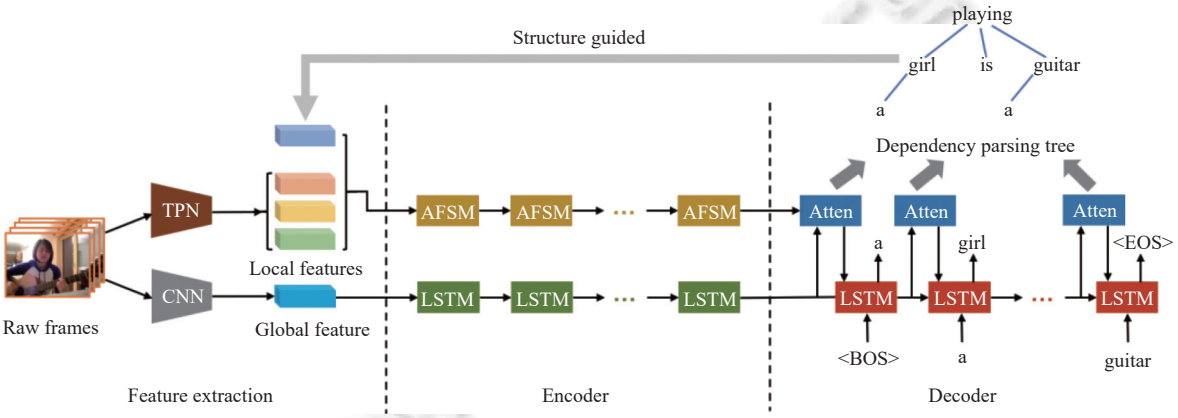

图 1 语言树结构指引的注意力小管编码器-解码器框架的图示

2.1 用于建模序列数据的 LSTM

作为前馈神经网络的扩展, 递归神经网络在序列数据建模方面取得了巨大的成功. 给定一个包含 n 个元素的

输入序列$\{x_1,x_2,\ldots,x_n\}$, 包含m个元素的输出序列$\{y_1,y_2,\ldots,y_m\}$由RNN通过以下递归公式生成:

$$h_t=\phi(Ux_t+Wh_{t-1}+b) \tag{1}$$

$$y_t=\phi(Vh_t+c) \tag{2}$$

其中, U、W和V代表权重矩阵, b和c表示偏置, t表示时间步, ϕ是激活函数, 如双曲正切函数. 由于存在爆炸和消失梯度问题, RNN难以处理时滞较大的视频.

长短期记忆网络(LSTM)[37]已证明能够解决上述问题. 在LSTM中, 有一个存储单元c_t和3个门: 输入门i_t, 忘记门f_t和输出门o_t. 存储单元c_t记录输入序列的累积历史信息. 输入门i_t决定当前时间步长需要考虑多少输入信息, 忘记门f_t决定当前时间要使用的前一个时间步长的存储单元中包含的信息的权重, 以及输出门O_t控制将多少信息从存储器c_t转移到隐藏状态h_t. 尽管LSTM有很多变体, 本文在提出的方法中使用最标准的LSTM[37], 可以用以下公式来表示:

$$i_t=\sigma(W_{ix}x_t+W_{ih}h_{t-1}+b_i) \tag{3}$$

$$f_t=\sigma\left(W_{fx}x_t+W_{fh}h_{t-1}+b_f\right) \tag{4}$$

$$o_t=\sigma(W_{ox}x_t+W_{oh}h_{t-1}+b_o) \tag{5}$$

$$g_t=\tanh\left(W_{gx}x_t+W_{gh}h_{t-1}+b_g\right) \tag{6}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot g_t \tag{7}$$

$$h_t=o_t\odot\tanh(c_t) \tag{8}$$

其中, $\odot$表示元素点积, 权重矩阵W和偏置b都是可训练的参数. σ和tanh是非线性激活函数Sigmoid和双曲正切.

为简单起见, 将LSTM在第t个时间步长的操作重写为如下符号:

$$h_t=LSTM(x_t,h_{t-1}) \tag{9}$$

其中, x_t是输入的向量, h_t是根据以上表示的LSTM的输出. 这里因为简单起见, 省略了对记忆细胞c_t的更新.

2.2 基础编码器-解码器框架

编码器-解码器框架已广泛用于视觉描述任务, 该框架由两个组件组成: 编码器和解码器. 具体地, 编码器旨在产生与视觉内容有关的代表性特征, 并且解码器利用这些特征来产生文本描述.

编码器和解码器的体系结构是多种多样的, 如果对它们的选择适合于任务, 则可以非常显著地提高性能. 例如, CNN是捕获空间视觉信息的自然选择, 同时, 如第2.1节中所述, RNN及其变体是建模序列数据的推荐选项. CNN和RNN的组合能够处理时空信息表示的问题. 本文用于视频描述任务的基础编码器-解码器框架中, 编码器是一个CNN, 其后跟随一个LSTM, 而解码器是单个的LSTM.

具体的, 给定一个由n个帧组成的视频X, 表示为$X=\{x_1,x_2,\ldots,x_n\}$, 编码器网络EN是一个把原始输入X到紧凑的特征空间V的函数:

$$V=\{v_1,v_2,\ldots,v_n\}=EN(X) \tag{10}$$

其中, $v_i\in R^c$表示输入的视频X中的第i个帧的c维的特征向量. 接着, 利用一个LSTM作为解码器DN来对视觉特征V进行建模, 生成一个包含m个单词的描述Y, 表示为$Y=\{y_1,y_2,\ldots,y_m\}$. 为了预测单词y_t和隐藏状态h_t, 解码器LSTM利用了视觉特征V, 前一个单词y_{t-1}及其前一个隐藏状态h_{t-1}, 因此该LSTM过程可以实例化为:

$$(h_t,y_t)=DN(V,y_{t-1},h_{t-1}) \tag{11}$$

解码器网络DN递归地更新其隐藏状态h_t, 并生成单词y_t, 直到所有的单词生成完毕.

本文提出的方法基于此编码器-解码器框架, 但补充和拓展了更有效的模块以提供更好的视觉特征和描述, 这些将在以下章节中进行介绍.

2.3 特征提取

为了提供足够的信息以进行描述生成, 全局特征和局部特征对于视频表示都是必不可少的. 通常, 从原始帧级图像中提取全局特征作为整个视频的概述. 与全局特征相比, 局部特征包含更多探索视觉内容的线索, 这些线索来自视频中的显著区域或片段.

2.3.1 全局特征

有一种通用的策略来表示具有全局特征的视频, 该策略在帧级别对从 CNN 生成的所有特征执行平均池化操作, 并为整个视频输出单个向量. 本文将视频的每个帧送入到 ResNet[29]中, 并按照先前的工作[4]从最终的卷积层中提取特征. 这些帧特征进一步与均值池化相结合.

给定一个由 n 个帧组成的视频 X, 本文的方法使用一个在 ImageNet[38]上预先训练的 ResNet-152 来处理每个帧, 对应输出一个特征图. 将第 i 个特征图表示为 $f_i \in R^{c\times h\times w}$, c, h 和 w 分别表示通道个数、高度、宽度, 通过将所有的帧特征沿着时间维度进行拼接, 从而形成一个完整的视频特征图 $f \in R^{n\times c\times h\times w}$. 此外, 将全局特征向量计算为所有帧特征图在高度和宽度维度上的平均池化结果:

$$\bar{f} = \frac{\sum^{h}\sum^{w} f}{h\times w} \tag{12}$$

其中, 输出 $\bar{f} \in R^{n\times c}$ 是平均池化全局特征向量.

2.3.2 局部特征

现有研究已证明小管是一种很好的视频物体的时间和上下文表示形式, 并已广泛用于对象检测和动作识别任务, 与此同时, 这种表示形式能够捕获视频中突出对象和动作的细粒度信息. 但是, 没有现有的工作来利用它来提高视频描述任务的性能. 与那些分别检测对象区域并识别动作的方法不同, 在本文提出的方法中使用小管作为统一的特征表示, 以将物体外观信息与运动信息集成在一起.

为了从给定的视频中提取小管, 本文采用了一种神经网络, 即候选小管提取网络 (TPN). 具体来说, 本文引用了一种称为 SiamMask[34]的快速而准确的方法, 该方法依赖于给定的初始边界框来跟踪对象. 为解决此问题, 在给定视频的第 1 帧上执行了候选区域提取网络 (RPN) 以在其中定位目标对象. 考虑到在开放域视频中, 显著对象不会总是在第 1 帧中出现这一事实, 本文按照文献 [39] 的工作将整个视频切分成等长的片段, 并在每个视频片段的第一帧上检测候选物体, 这也可以在一定程度上缓解漂移问题.

给定目标物体的初始边界框, SiamMask 可以通过在连续帧中建立目标物体与候选区域之间对应关系的策略快速生成小管, 即以滑动窗口方式学习目标对象与多个候选对象之间相似性的度量. 此任务输出一个密集的响应图, 可通过边界框回归对其进行细化. 按照这种方法, 可以在第 1 帧中获得每个对象候选的边界框序列, 并利用这些边界框对全局特征执行感兴趣区域 (RoI) 对齐操作[40], 然后通过平均池层, 提取得到最终的小管特征.

假设输入一个视频 X, 帧数为 n, 首先, 将整个视频按照每 l 帧进行分割, 得到 $p = \lfloor n/l \rfloor$ 的等长视频片段, $\lfloor\cdot\rfloor$ 为下截断操作. 每个片段的第 1 帧送入 Mask R-CNN[40]来生成初始边界框, 选择其中 k 个最高置信度的候选目标. 如上所述, 这些候选目标被进一步送入 SiamMask 模型以产生 k 个小管, 每个小管都可以表示为一系列的包围框. 记第 j 个小管为 $T_j = \{b_1, b_2, \ldots, b_l\}$, 将每个边界框及其对应的帧特征图 f_i 传递给一个 RoI 对齐层以提取一个小尺寸的调整大小的目标特征图 $m_{ij} \in R^{c\times h_o\times w_o}$ (例如, $c\times 2\times 2$), 其中 c, h_o, w_o 分别表示通道数, 目标特征图的高度和宽度. 和处理全局特征类似, 在第 j 个小管 T_j 中级联所有的目标特征图 m_{ij} 来生成一个小管特征图 $m_j \in R^{l\times c\times h_o\times w_o}$. 然后, 对每个小管特征图 m_j 使用一个核大小为 $h_o \times w_o$ 的平均池化, 计算出一个平均池化向量 $v_j \in R^{l\times c}$, 这些向量最终形成了一组局部视觉特征表示 $V = \{v_1, v_2, \ldots, v_k\}$.

值得注意的是, 局部特征是从全局帧特征图中提取的, 因此具有相同数量的通道数 c. 这些特征将在基于注意力的帧选择编码器中进一步处理.

2.4 帧选择编码器

与可以直接使用基本 LSTM 编码的部分选定帧中提取的全局特征不同, 局部特征会遇到漂移问题, 因此编码器应进行更加仔细的设计. 对小管的长度施加约束不足以解决此问题, 因为细分的小管中仍可能会发生漂移. 为了捕获更精确的特征表示, 本文借鉴了针对交通流量预测任务而提出的人流机器模块 (ACFM)[41]的思想, 该模块能够学习时变数据的动态表示, 我们对其结构进行了调整以适合小管编码, 命名为帧选择编码器. 通过研究小管帧之间的时间依赖性, 帧选择编码器被用于产生更加紧凑和具有代表性的特征.

如图 2 所示, 帧选择编码器由两个 LSTM 单元组成, 并与一个卷积层相连用于对每个帧进行权重预测, 以及一个多层感知机分支用于硬注意力. 底部的 LSTM 通过对与原始视频帧特征级联的小管特征对时间相关性进行建模, 输出隐藏状态与当前特征连接在一起送给卷积层以进行权重图推理, 并输送到 MLP 层以产生硬注意力图. 顶部 LSTM 与第 1 个 LSTM 具有相同的结构, 但在每个时间步均以重新加权的小管特征作为输入, 并循环编码整个小管以产生紧凑的时空表示.

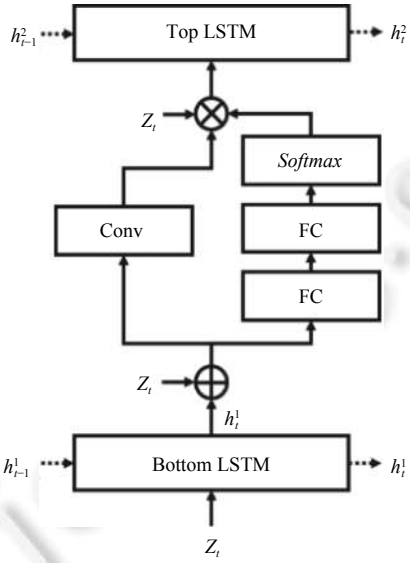

图 2 帧选择编码器的图示

具体地, 根据前面的假设, 给定一个视频 X 有 n 帧, 提取一个全局特征 $\bar{f} \in R^{n\times c}$, c 表示特征图的通道数, 以及 k 个局部小管特征 $V=\{v_1,v_2,\ldots,v_k\}$, 其中第 i 个局部特征记为 $v_i \in R^{l\times c}$, 其中 l 是分割的视频片段的长度. 作为重新加权的局部特征图 $\bar{f}$ 的指引, 全局特征被用于和 k 个局部特征级联. 但是, 全局特征的第 1 个维度 n 与局部特征的一个维度 l 不兼容, 其中 n 是整个视频的长度, 而 l 代表分割的视频片段的长度. 受文献 [42] 中提出的时间段网络的启发, 本文使用从整个视频中稀疏采样的一系列简短片段来进行近似视频特征表示, 这已在动作识别和其他视频理解任务中被证明是一种高效的视频特征表示. 因此, 通过对整个全局特征 $\bar{f}$ 划分成相同帧数的 l 段, 然后从每一段中随机采样一个帧特征图, 得到一个采样后的全局特征 $\overline{f}_s \in R^{l\times c}$. 这个过程使我们能够利用 k 个局部特征和全局特征的级联 $Z \in R^{q\times l\times c}$ $(q=1+k)$ 作为底部的 LSTM 的输入.

根据文献 [41], 底部 LSTM 在第 t 个时间步的过程可以表示为:

$$h_t^1 = LSTM\left(Z_t, h_{t-1}^1\right) \tag{13}$$

其中, $h_t^1 \in R^{q\times d}$ 是隐藏大小为 d 的隐藏状态元, $Z_t \in R^{q\times c}$ 是在级联 Z 内的第 t 个帧特征图. 在这里的隐藏状态 h_t^1 建模了之前 $t-1$ 个视频帧的动态时间信息.

为了对局部特征重新加权以处理权重偏移问题, 隐藏状态 h_t^1 和输入 Z_t 被连接起来然后送入一个核大小为 1×1 的卷积层, 以生成一个软注意力图 W_t^S, 可以被表示为:

$$W_t^S = \mathrm{conv}_{1\times 1}\left(h_t^1 \oplus Z_t, W_c\right) \tag{14}$$

其中, $\oplus$ 表示沿着特征通道的维度进行级联操作, W_c 是卷积层的参数矩阵. 软重新赋重图 $W_t^S \in R^1$ 表示在每个小管特征中的第 t 个帧的权重. 虽然这种软注意力图可能有助于减少漂移帧的权重, 并处理那些具有明显上下文依赖性的帧, 但更好的方案是删除所有携带不准确信息的漂移帧, 以防干扰生成的描述. 为此, 本文将特征 $h_t^1 \oplus Z_t$ 送入一个 MLP 分支, 其中包含两个线性变换层, 并跟随有一个非线性的 *Softmax* 函数, 然后执行一个 argmax 操作来生

成最终的硬选择图 W_t^H. 这个过程可以表示为:

$$\xi_t = \omega_m^{\mathrm{T}} \tanh\left(W_m^2 \tanh\left(W_m^1\left(h_t^1 \oplus Z_t\right) + b_m^1\right) + b_m^2\right) \tag{15}$$

$$\tau_t = Softmax(\xi_t) \tag{16}$$

$$W_t^H = \arg\max(\tau_t) \tag{17}$$

其中, $\omega_m \in R^2$, $W_m^1 \in R^{(d+c)\times r}$, r 是隐藏层的维度, $W_m^2 \in R^{r\times 2}$, $b_m^1 \in R^r$ 和 $b_m^2 \in R^2$ 是 MLP 层的所有可学习参数. 另外, $\xi_t \in R^2$ 表示线性变换层的输出, $\tau_t \in R^2$ 表示 $[0,1]$ 上的选择分布. 所以, 这个分支的输出 $W_t^H \in [0,1]$ 是第 t 帧的硬选择权重, 其中 0 和 1 分别表示丢弃帧和选择帧.

最后, 通过用软重新赋重图 W_t^S 和硬选择图 W_t^H 进行元素级相乘对第 t 个帧特征图 Z_t 重新赋权重, 然后对其进行归一化, 后将此标准化特征送入用于表示学习的顶部 LSTM, 可以表示为:

$$\widehat{Z}_t = \frac{W_t^H \otimes W_t^S \otimes Z_t}{\sum W_t^H} \tag{18}$$

$$h_t^2 = LSTM\left(\widehat{Z}_t, h_{t-1}^2\right) \tag{19}$$

其中, $\otimes$ 指的是元素级相乘操作. 在第 t 个时间步的隐藏状态 h_t^2 对当前输入的基于注意力的选择内容和之前 $t-1$ 个输入的上下文知识进行编码. 因此, 最后一个隐藏状态 h_l^2 编码了整个特征管道的信息, 在解码阶段作为一个紧凑、准确的时空视觉线索来生成自然语言句子.

2.5 依赖型解析树解码器

本文采用了一种解析树注意机制来探索视频中更细粒度的小管信息, 而不是直接应用简单 LSTM 作为解码器网络. 在介绍所提出的解析树解码器之前, 有必要回顾一下基于转换的依赖解析方法[43], 并描述如何将解析树中的语义关系与句子解码器结合起来. arc-standard 系统是最流行的转换系统之一, 它被用作解析器的基础, 其中有一个配置 C. 给定一个词堆栈 S, 一个词缓冲区 B 和一个依赖弧集 A, 配置 C 被定义为:

$$C = (S, B, A) \tag{20}$$

对一个句子 $Y = \{y_1, y_2, \ldots, y_n\}$ 来说, 初始配置 C_0 是 $S = [root]$, $B = [y_1, y_2, \ldots, y_n]$, $A = \varnothing$. 将 s_i $(i = 1,2,\ldots,n)$ 表示为栈顶的第 i 个元素, b_i $(i = 1,2,\ldots,n)$ 表示为缓冲区的第 i 个元素, arc-standard 系统定义了 3 种类型的转换.

• *Shift*: 将 b_1 从缓冲区移动到堆栈.

• *Left-arc*(l): 添加一个带有依赖标签 l 的弧 $s_1 \to s_2$, 并将 s_2 从栈中移除.

• *Right-arc*(l): 添加一个带有依赖标签 l 的弧 $s_2 \to s_1$, 并将 s_1 从栈中移除.

如果缓冲区为空, 并且堆栈包含单个节点 *root*, 而解析树由 A_C 给出, 则配置 C 终止.

为了更好地理解, 假设本文提出的方法的输入是视频 X 及其相应的描述性语句, 其中包含 n 个嵌入单词, 表示为 $Y = \{y_1, y_2, \ldots, y_n\}$, 以及编码器网络生成编码的全局特征 $e_{\bar{f}}$ 和 k 个编码的微管特征 $E = \{e_1, e_2, \ldots, e_k\}$. 另外, 还给出了由预训练的依赖解析器生成的解析转换序列 $P = \{p_1, p_2, \ldots, p_m\}$. 在解析树构造的开始, 配置 C 中的堆栈为 $S = [root]$, 缓冲区为 $B = [y_1, y_2, \ldots, y_n]$. 通过根据给定的转换序列 P 重复执行状态转换, 直到将配置 C 转换为最终状态为止, 生成输出语句. 在第 t 个时间步中, 此过程的详细信息如下.

(1) *Shift*: 如果转换 P 为 *Shift*, 则将顶部单词 y_t 从缓冲区 B 移动到堆栈 S. 如果将当前堆栈 S 视为要生成的句子, 这种转换类似于在描述任务中生成下一个单词 y_t. 在这种情况下, 利用软注意机制通过测量每个小管特征与历史信息之间的相关性来预测下一个单词 y_t. 特别地, 给定一组编码特征 $E = \{e_1, e_2, \ldots, e_k\}$ 和解码器网络的前一个隐藏状态 h_{t-1}, 将其送入一个线性变换层来得到一个未归一化的相关得分 ε_t, 然后更进一步使用一个 *Softmax* 函数处理以计算在 k 个特征向量上的注意力分布 α_t, 表示为:

$$\varepsilon_t = \omega^{\mathrm{T}} \tanh\left(W_\alpha E + U_\alpha h_{t-1} + b_\alpha\right) \tag{21}$$

$$\alpha_t = Softmax(\varepsilon_t) \tag{22}$$

其中, $\omega_m \in R^k$ 和 $W_\alpha, U_\alpha \in R^{c\times k}$ 是所有需要学习的参数矩阵, c 表示通道数. 另外, $b_\alpha \in R^k$ 是一个可以训练的偏置向量. $\alpha_t \in R^k$ 是量化 k 个小管特征和隐藏状态中记录的信息之间的依赖的注意力权重. 上下文视觉特征向量 c_t, 被用作产生单词的重要线索, 其计算过程可以表示为:

$$c_t = \sum_{i=1}^{k} \alpha_{ti} e_i \tag{23}$$

这种软注意力机制使解码器网络可以在不同的时间步长选择性地关注显着的小管特征 c_t, 该特征与先前的隐藏状态 h_{t-1} 以及先前的单词 y_{t-1} 被馈送到初始隐藏状态为 $e_{\bar{f}}$ 的解码器 LSTM, 以更新当前的隐藏状态 h_t, 公式表示为:

$$h_t = LSTM(c_t, y_{t-1}, h_{t-1}) \tag{24}$$

然后将当前的隐藏状态 $h_t \in R^r$ (r 是 RNN 隐藏大小的维度) 送入一个 MLP 层来预测在包含所有可能单词的词汇表上的一个概率分布 p_t:

$$p_t = Softmax\left(W_p h_t + b_p\right) \tag{25}$$

其中, $W_p \in R^{r\times s}$ 和 $b_p \in R^s$ 表示可训练的参数矩阵和偏置向量, 其中, s 表示词汇表大小. $p_t \in R^s$ 表示在词汇表上的归一化的概率, 可以理解为:

$$p_t = P\left(y_t | y_{1:t-1}, e_{\bar{f}}, E; \theta\right) \tag{26}$$

其中, $y_{1:t-1}$ 表示从第 1 个时间步到第 $t-1$ 个时间步的单词, θ 表示可训练的参数. 根据概率分布 p_t, 具有最大概率的单词被选中作为在第 t 个时间步的生成单词 y_t.

$$y_t = \arg\max(p_t) \tag{27}$$

其中, 单词 y_t 被认为是描述视觉显著内容的最相关的单词.

Left-arc: 当处理一个 *Left-arc* 的转换 P 时, 根据转换系统, 从栈 S 的第一个顶部元素 y_i 到第 2 个顶部元素 y_j 添加一个弧 $y_i \to y_j$, 同时从栈中移除 y_j. 同时, 对编码后的特征 E 进行动态修改, 增加更细粒度的视觉信息. 具体而言, 给定在依赖解析树中的顶部单词 y_i 有 n 个孩子节点 $\widehat{Y} = \left\{y_i^1, y_i^2, \ldots, y_i^n\right\}$, 映射编码特征 e_i 和 $\widehat{E} = \left\{e_i^1, e_i^2, \ldots, e_i^n\right\}$, 其原始小管为 T_i 和 $\widehat{T} = \left\{T_i^1, T_i^2, \ldots, T_i^n\right\}$, 包括 e_i 和 $\widehat{E}$ 在内的所有这些 $(n+1)$ 编码特征从小管特征集合 E 中去除, 用于计算一个添加到 E 中的新特征 $\tilde{e}$. 为了获得新特征 $\tilde{e}$, 本文尝试利用一个最小的小管 $\tilde{T}$ 来涵盖这些 $(n+1)$ 个特征所对应的在 $\tilde{T}$ 中的所有小管, 然后根据这个小管提取一个新的特征向量, 然后进一步通过编码网络编码以生成最终的紧凑特征 $\tilde{e}$. 这个新加入的特征 $\tilde{e}$ 携带有在 y_i 的孩子节点中存在的小管之间的关系信息, 有助于生成一个更精准的描述.

Right-arc: 处理一个 *Right-arc* 变换与 *Left-arc* 变换相似. 在这种情况下, 从栈 S 的第 2 个顶部元素 y_{t-1} 到第 1 个顶部元素 y_t 添加一个弧 $y_{t-1} \to y_t$, 这个过程的其余部分可以参考前面提到的 *left-arc* 变换过程.

2.6 损失函数

如公式 (26) 所示, 网络模型会在给定所有先前已看到的单词 y_{t-1}, 已编码的全局特征 $e_{\bar{f}}$ 以及承载输入视频中物体外观和运动信息的局部小管的隐藏表示集 E 的情况下, 学习在某个时间步预测单词 y_t. 单词 y_t 是从 MLP 层的输出生成的, 它是每个单词在词汇表上的概率分布, 因此, 本文自然地利用似然的负对数作为整体损失函数来指导参数的学习:

$$L(\theta) = -\sum_{t=1}^{T} \log P\left(y_t | y_{1:t-1}, e_{\bar{f}}, E; \theta\right) \tag{28}$$

其中, T 表示生成的句子中的单词总数, θ 表示模型参数. 使用随机梯度下降法来寻找最优解, 通过时间维度上的反向传播来计算梯度[38].

为了生成最后的句子, 本文选择集束搜索方法, 迭代地考虑到时间 t 之前的 k 个最佳句子集作为生成大小为 $t+1$ 的句子的候选句子, 并保留得到的最好的 k 个句子. 当 k 个句子完成后, 选择其中总体后验概率最大的句子作为最终的描述结果. 每步根据概率分布 p_t 对具有最大概率的单词进行采样的方法 (称为采样方法) 在大多数情况下都获得了次优的解决方案, 而集束搜索方法相较而言可以产生更好的结果, 本文实验使用集束搜索方法.

3 实验分析

本节对本文提出的方法进行全面的评估和分析. 首先介绍了在实验中使用的基准以及用于评估描述结果的指标. 接下来详细阐述了方法和实验的具体细节. 然后通过消融实验分析了提出的模块的有效性. 最后, 与最先进的方法进行了对比分析.

3.1 数据集

本文使用了两个在以前的工作中广泛使用的知名数据集来评估提出的方法.

● The Microsoft video description corpus (MSVD)[44]: 该数据集由 1 970 个视频组成, 每个视频带有由 Amazon Mechanical Turkers 标记的多个描述. 视频片段和描述对的总数约为 80 000. 本文在实验中使用文献 [4] 中提供的标准划分方式, 将原始数据集分为 1 200 个视频片段, 100 个视频片段和剩余部分, 分别作为训练集, 验证集和测试集, 以便与最先进的视频描述系统 (例如文献 [4,10]) 进行公平比较.

● MSR video to text (MSR-VTT)[45]: MSR-VTT 是最近的一个用于视频描述的大规模基准. 从一个商业视频搜索引擎 (如音乐、人类、人物、游戏、体育和电视节目) 中收集了 20 个类别的 1 万个视频片段. 每一段视频都有 20 个句子描述, 由 1 327 名标注工作者制作, 描述总数约为 20 万. 根据原论文的规定, 本文按索引号划分数据: 6 513 个用于训练, 497 个用于验证, 2 990 个用于测试.

3.2 评价指标

类似于传统的机器翻译, 可以通过将生成的视频描述语句与一组参考语句进行比较来评测其准确性. 机器翻译中有 4 个常用指标用于评估视觉描述结果: BLEU[46], METEOR[47], CIDEr[48]和 ROUGE-L[49].

● BLEU: BLEU 是用于评估描述的最常用的度量指标, 用于计算预测的候选句子与其若干对应的参考句子之间的 n 元组词精度. 例如, 1 元组会累加计算在预测句子和任何参考句子中都出现的单词数量, 然后除以候选单词中的单词总数, 生成一个精度为 1 元组的结果. 本文使用 BLEU-4, 通过几何平均将 1 元组, 2 元组, 3 元组和 4 元组的分数结合在一起.

● CIDEr: CIDEr 最初是为评估图像描述而提出的, 它是第 1 个基于共识的自动度量标准, 用于测量生成的句子与人类编写的一组真实句子的相似性. 为了计算该分数, 首先将所有单词映射到其词干或词根形式. 然后, 将每个句子表示为一组以 TF-IDF 加权的 n 元组. 最后, 针对每个 n 元组计算候选句子和参考句子之间的平均余弦相似度.

● METEOR: 考虑到召回率, METEOR 计算加权 F 均值分数, 该分数通过调和平均将精度和召回率相结合. 通过将候选句子与参考句子对齐, 以计算句子级别的相似性评分. 对于候选-参考对, METEOR 根据 WordNet 数据库比较其精确的单词匹配, 词干匹配, 释义匹配以及语义相似的匹配.

● ROUGE-L: ROUGE 系统被提出用于自动确定概要的质量. 该系统具有 ROUGE-N, ROUGE-L, ROUGE-W 和 ROUGE-S(U) 这 4 个版本, 其中 ROUGE-L 被广泛用于评估视频描述结果. ROUGE-L 指标将 F 均值应用于候选句子和参考句子之间的 1 元组的最长公共子序列.

遵循先前的工作[4,7], 本文利用 Microsoft COCO 评估工具包来获取本文中报告的所有这些评估结果, 以确保进行公正的评估.

3.3 实现细节

本节详细介绍了所提出的方法的具体实现, 详细说明了从给定数据集提取全局和局部特征的方法, 以及对句子的预处理, 并描述了在 PyTorch 框架中实现的模型的优化和推理.

● 预处理: 对于视频表示, 首先将给定的视频划分为 20 个片段, 从每个片段中随机采样一帧, 然后将采样的帧送入在 ImageNet[38]上进行预训练的 ResNet-152[29], 然后从该网络的 pool5 层提取一个 2 048 维的特征. 这些特征映射沿着时间长度连接, 作为全局视觉特征表示. 为了提取局部小管特征, 本文采用了 SiamMask[34], 它依赖于给定的初始边界框. 给定的视频每 20 帧被分割一次, 其中第 1 帧被馈送到 Mask R-CNN[40], 以产生初始目标建议.

SiamMask 将这些边界框以及连续的 19 帧作为输入并输出一些小管, 从中选择了 30 个置信度最高的小管并丢弃其他剩余的. 然后, 使用这些选定的小管对从 ResNet-152 的 conv5 层提取的 $2048\times7\times7$ 维的特征图执行 RoIAlign (感兴趣区域对齐), 以生成局部视觉表示. 对于那些没有足够小管的视频填充零特征. 对于句子预处理, 使用 Stanford CoreNLP 工具[50]中的 PTBTokenizer 将所有描述转换为小写, 对句子进行标记并删除所有标点符号. 参考中出现的所有单词都用于构建词汇表, 对于 MSVD 数据集, 该词汇表的大小为 12 593, 对于 MSR-VTT 数据集, 其词汇表的大小为 22 925. 句子中的每个单词都表示为"独热编码"向量 (词汇中的二进制索引向量), 使用 Stanford 解析器[43]来解析训练语句以获得相应的过渡序列.

● 训练: 在训练阶段, 将用于每个句子开头的句子开始标记〈BOS〉和用于结束每个句子的句子结束标记〈EOS〉添加到词汇表中以进行处理任意长度的句子. 根据之前的工作[39]的启发, 真值句的最大长度设置为 20 个单词, 长度超过限制的句子将被裁剪而较短的句子将被补零. 模型中所有 LSTM 单元的大小都设置为 1 000, 并按照经验将视觉特征嵌入大小以及单词嵌入大小都设置为 512. 所有全连接层和嵌入层的权重使用 Kaiming Uniform[51]初始化. 本文在全部训练视频句子对上对第 2.6 节中的目标函数方程进行优化, 在 MSVD 和 MSR-VTT 上的小批量大小设置为 50. 使用 Adam[52]优化器, 同时设置学习率 $\eta=1\times10^{-4}$, 衰变参数 $\beta_1=0.9$, $\beta_2=0.999$ 来优化目标损失函数. 对模型进行 50 次训练, 或者直到评估指标在验证集上没有改进就停止训练. 整个训练过程使用了一张 NVIDIA GTX 1080Ti GPU.

● 测试: 在测试阶段, 将句子开始标记〈BOS〉输入到训练好的模型中以触发视频描述生成过程, 该过程将一直持续到生成句子结束标记〈EOS〉. 与训练阶段不同, 由于没有提供参考语句, 因此没有真实值转换. 但是, 当在解码器中预测一个单词时, 会将其推入语法信息完整的堆栈. 遵照文献 [43] 的工作, 可以仅基于堆栈状态来训练依赖项解析器. 通过使用预训练的解析器, 本文的模型可以根据当前堆栈信息确定下一个转换以构造解析树, 并应用本文提出的树状结构注意力机制. 为了产生最终结果, 本文采用大小为 5 的集束搜索方法来同时以每个单词生成的最大概率考虑 5 个候选词, 最后在所有候选词都以〈EOS〉结尾时选择最佳的一个.

3.4 与最先进的方法的比较

在本节中, 我们在 MSVD 数据集和 MSR-VTT 数据集上, 对本文提出的树型结构注意力小管编码器-解码器模型 (TSAT) 与最先进的方法进行比较.

● MSVD 数据集的结果: 表 1 展示了 14 个最先进方法的定量性能表现, 包括二层 LSTM (LSTM-YT)[1], 基础编码器-解码器 (S2VT)[4], 基于软注意力的 LSTM 网络 (SA)[10], 基于层次处理的编码器 (HRNE)[53], 基于联合学习的 LSTM 嵌入网络 (LSTM-E)[6], 层次递归神经网络 (h-RNN)[7], 多模型注意力 LSTM 网络 (MA-LSTM)[8], 任务驱动动态融合网络 (TDDF)[54], 边界感知神经编码器 (BAE)[55], 基于注意力的语义一致性 LSTM (aLSTMs)[56], 基于强化学习的信息帧选择网络 (PickNet)[57], 时序注意网络 (STAT)[58], 运动引导空间注意网络 (MGSA)[20]以及语法感知动作目标网络 (SAAT)[18]. 最后一行报告了本文提出的方法的结果. 所有的值都以百分比 (%) 的形式报告, –表示未知的分数.

本文提出的 TSAT 在 BLEU-4, ROUGE-L 的指标上超过了所有的现有最先进方法, 同时获得了很高的 CIDEr 和 METEOR 得分. 更具体而言, TSAT 模型在 BLEU-4 和 ROUGE-L 上比目前最先进的方法相对提升了 0.19% 和 0.43%, 在 CIDEr 上也获得了较强的 75.4 的结果, 相对于 SA、h-RNN、aLSTMs、STAT 和 MGSA 分别提高了 45.84%、14.59%、0.80%、2.16% 和 1.62%. 这些结果验证了所提出方法的有效性.

● MSR-VTT 数据集的结果: 表 2 总结了 MSR-VTT 数据集上的性能比较. 在该数据集上, 本文对比了 TSAT 与二层 LSTM (LSTM-YT)[1], 基础编码器-解码器 (S2VT)[4], 基于软注意力的 LSTM 网络 (SA)[10], 基于联合学习的 LSTM 嵌入网络 (LSTM-E)[6], 多模型注意力 LSTM 网络 (MA-LSTM)[8]. 由表 2 可知, TSAT 模型在 4 个指标中的 3 个指标上表现最佳, BLEU-4 为 39.3%, METEOR 为 27.1%, CIDEr 为 43.4%, 并获得了具有可比性的 ROUGE-L 得分, 值为 59.1%. 与最接近的对比方法相比, 本文所提出的 TSAT 模型将 BLEU-4 的性能提高了 7.67%, 将 METEOR 的性能提高了 2.26%, 将 CIDEr 的性能提高了 5.85%.

表 1 在 MSVD 基准上的性能比较 (%)

模型	BLEU-4	METEOR	CIDEr	ROUGE-L
LSTM-YT[1]	33.3	29.1	—	—
S2VT[4]	—	29.8	—	—
SA[10]	41.9	29.6	51.7	—
HRNE[53]	43.8	33.1	—	—
LSTM-E[6]	45.3	31.0	—	—
h-RNN[7]	49.9	32.6	65.8	—
MA-LSTM[8]	52.3	33.6	70.4	—
TDDF[54]	45.8	33.3	73.0	69.7
BAE[55]	42.5	32.4	63.5	—
aLSTMs[56]	50.8	33.3	74.8	—
PickNet[57]	52.3	33.3	76.5	69.6
STAT[58]	52.0	33.3	73.8	—
MGSA[20]	49.5	32.2	74.2	—
SAAT[18]	46.5	33.5	81.0	69.4
TSAT	52.4	33.3	75.4	70.0

表 2 在 MSR-VTT 基准上的性能比较 (%)

模型	BLEU-4	METEOR	CIDEr	ROUGE-L
LSTM-YT[1]	35.7	25.6	38.1	58.2
S2VT[4]	36.0	26.0	39.1	58.4
SA[10]	34.8	25.1	36.7	57.1
LSTM-E[6]	36.1	25.8	38.5	58.6
MA-LSTM[8]	36.5	26.5	41.0	59.8
TSAT	39.3	27.1	43.4	59.1

3.5 定性分析

从定性的角度来看, 图 3 展示了 MSVD 数据集中的 3 个正面示例和 1 个负面示例. 每段视频示例用 4 帧代表, 并可视化关注目标的小管, 同时展示相应的描述语句, 包括由基本的编码器-解码器框架 (baseline) 和本文提出的 TSAT 生成的候选句子, 以及一个参照句子 (reference). 描述给定视频的准确单词用红色标记, 粗略的和不相关的单词用蓝色标记. 如图所示, 与基本的编码器-解码器框架相比, 本文提出的 TSAT 模型可以为给定的视频提供更准确更全面的描述. 例如, 在左上角的板块上, TSAT 模型可以很好地捕获对象“面食”和“碗”以及动作“倾倒”, 而不是将视频粗略地描述为“一个人正在做饭”. 从右上方和左下方版块的其他两个正面示例也可以得出类似的结论. 这些例子证明了 TSAT 模型所依赖的小管机制可以有效实现更加细粒度的具体对象的特征补充, 进而提高了生成描述的准确性. 同时, 本文借助注意力机制实现各个小管特征的自适应选择和融合, 并根据当前预测的单词信息动态调整小管结构及其对应编码特征, 一定程度上增加了模型对小管检测结果的鲁棒性. 当然, 在某些极端情况下, 小管检测的结果还是会影响模型的描述生成, 例如右下版块的负面示例, 由于小管检测失败, 模型捕获了“三个人”, 而不是准确的“两个男人”和“两个女人”, 但这从另一方面表明在生成描述时, 具有细粒度信息的基于小管的视觉表示被充分利用了. 因此本文认为, 随着候选小管提取方法的发展, TSAT 模型可能会得到进一步改善.

3.6 消融实验研究

为了证明 TSAT 中的基于小管的局部视觉表示, 帧选择编码器和树状结构注意力机制的有效性, 本文实现了完整模型的 3 个变体, 并在 MSVD 数据集上进行了实验以进行内部比较.

- 基本编码器-解码器模型 (ED): 仅使用全局特征作为视觉信息, 并用基本的 LSTM 进行编码和解码.
- 软注意力小管模型 (SAT): 使用了全局特征以及基于小管的局部特征. 在此模型中, 视觉特征由一个基本 LSTM 编码, 并由一个具有软注意力机制的 LSTM 解码.
- 树形结构管模型 (TST): 使用全局特性和基于小管的局部特征. 在这个模型中, 视觉特征由一个基本 LSTM 编码, 并由一个具有树形结构注意机制的 LSTM 解码.

如表 3 所示, 不具有小管特征的基本编码器-解码器网络 ED 在 BLEU-4, METEOR, CIDEr 和 ROUGE-L 的指标上仅分别达到 47.3%, 32.3%, 69.4% 和 67.4%, 相较于具有软注意力机制的小管模型 SAT 分别退化了 1.69%, 2.79%, 7.78% 和 2.67%, 证实了基于小管的视觉特征可有效提高视频描述性能. TST 采用树状结构的注意力机制, 使 BLEU-4 的性能提高了 5.6%, METEOR 的性能提高了 2.71%, CIDEr 的性能提高了 0.27%, ROUGE-L 的性能提

高了 0.87%, 这证明了树状结构注意力机制的有效性. TSAT 利用基于注意力的帧选择编码器解决漂移问题, 进一步将 BLEU-4, CIDEr 和 ROUGE-L 分别提高了 1.95%, 0.53% 和 0.29%.

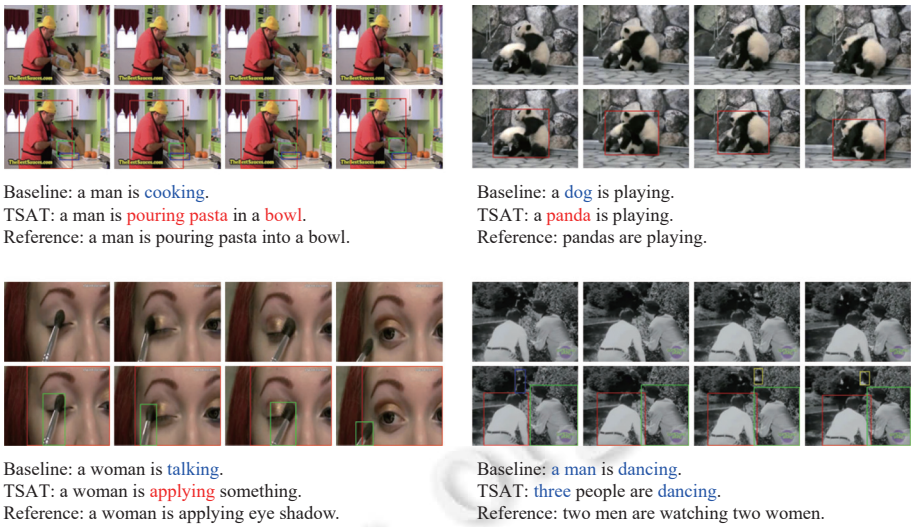

图 3 MSVD 数据集的示例结果

表 3 在 MSVD 数据集上对本文模型的不同变体进行定量比较 (%)

模型	BLEU-4	METEOR	CIDEr	ROUGE-L
ED	47.3	32.3	69.4	67.4
SAT	48.1	33.2	74.8	69.2
TST	51.4	34.1	75.0	69.8
TSAT	52.4	33.3	75.4	70.0

4 结 论

本文提出了一种树型结构的注意力小管编码器解码器框架, 其中小管被用作视频中物体外观和运动信息的集成时空视觉表示. 为了解决漂移问题, 引入了一种帧选择编码器来对每个小管中的帧重新加权, 丢弃掉漂移的帧, 并生成更具代表性的特征. 此外, 基于句子语义结构的注意力机制使解码器能够探索更多细粒度的信息, 并最终为视频内容生成更准确的描述. 广泛的实验结果证明, 提出的 TSAT 方法在 MSVD 和 MSR-VTT 数据集上均达到了最先进的性能, 每个组件对于在视频描述任务中生成准确的描述性句子都是必不可少的.

References:

[1] Rohrbach M, Qiu W, Titov I, Thater S, Pinkal M, Schiele B. Translating video content to natural language descriptions. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 433–440. [doi: 10.1109/ICCV.2013.61]

[2] Xu R, Xiong CM, Chen W, Corso JJ. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Proc. of the 29th AAAI Conf. on Artificial Intelligence. Austin: AAAI Press, 2015. 2346–2352.

[3] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Venugopalan S, Mooney R, Darrell T, Saenko S. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 2712–2719. [doi: 10.1109/ICCV.2013.337]

[4] Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K. Sequence to sequence-video to text. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4534–4542. [doi: 10.1109/ICCV.2015.515]

[5] Venugopalan S, Xu HJ, Donahue J, Rohrbach M, Mooney R, Saenko K. Translating videos to natural language using deep recurrent neural networks. In: Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies. Denver: Association for Computational Linguistics, 2015. 1494–1504. [doi: 10.3115/v1/N15-1173]

[6] Pan YW, Mei T, Yao T, Li HQ, Rui Y. Jointly modeling embedding and translation to bridge video and language. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4594–4602. [doi: 10.1109/CVPR.2016.497]

[7] Yu HN, Wang J, Huang ZH, Yang Y, Xu W. Video paragraph captioning using hierarchical recurrent neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4584–4593. [doi: 10.1109/CVPR.2016.496]

[8] Xu J, Yao T, Zhang YD, Mei T. Learning multimodal attention LSTM networks for video captioning. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain View: ACM, 2017. 537–545.

[9] Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the 32nd Int'l Conf. on Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

[10] Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville C. Describing videos by exploiting temporal structure. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 4507–4515. [doi: 10.1109/ICCV.2015.512]

[11] Long X, Gan C, De Melo G. Video captioning with multi-faceted attention. Trans. of the Association for Computational Linguistics, 2018, 6: 173–184. [doi: 10.1162/tacl_a_00013]

[12] Hori C, Hori T, Lee TY, Zhang ZM, Harsham B, Hershey JR, Marks TK, Sumi K. Attention-based multimodal fusion for video description. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 4203–4212. [doi: 10.1109/ICCV.2017.450]

[13] Zhang JC, Peng YX. Object-aware aggregation with bidirectional temporal graph for video captioning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8319–8328. [doi: 10.1109/CVPR.2019.00852]

[14] Hu YS, Chen ZZ, Zha ZJ, Wu F. Hierarchical global-local temporal modeling for video captioning. In: Proc. of the 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 774–783. [doi: 10.1145/3343031.3351072]

[15] Zhang ZQ, Shi YY, Yuan CF, Li B, Wang PJ, Hu WM, Zha ZJ. Object relational graph with teacher-recommended learning for video captioning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13275–13285. [doi: 10.1109/CVPR42600.2020.01329]

[16] Hou JY, Wu XX, Zhang XX, Qi YY, Jia YD, Luo JB. Joint commonsense and relation reasoning for image and video captioning. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(7): 10973–10980. [doi: 10.1609/aaai.v34i07.6731]

[17] Lei Y, He ZH, Zeng PP, Song JK, Gao LL. Hierarchical representation network with auxiliary tasks for video captioning. In: Proc. of the 2021 IEEE Int'l Conf. on Multimedia and Expo (ICME). Shenzhen: IEEE, 2021. 1–6. [doi: 10.1109/ICME51207.2021.9428461]

[18] Zheng Q, Wang CY, Tao DC. Syntax-aware action targeting for video captioning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13093–13102. [doi: 10.1109/CVPR42600.2020.01311]

[19] Tan GC, Liu DQ, Wang M, Zha ZJ. Learning to discretely compose reasoning module networks for video captioning. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI, 2020. 753–759.

[20] Chen SX, Jiang YG. Motion guided spatial attention for video captioning. Proc. of the AAAI Conf. on Artificial Intelligence, 2019, 33(1): 8191–8198. [doi: 10.1609/aaai.v33i01.33018191]

[21] Hou JY, Wu XX, Zhao WT, Luo JB, Jia YD. Joint syntax representation learning and visual cue translation for video captioning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 8917–8926. [doi: 10.1109/ICCV.2019.00901]

[22] Jin T, Huang SY, Chen M, Li YM, Zhang ZF. SBAT: Video captioning with sparse boundary-aware transformer. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI, 2020. 630–636.

[23] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. Int'l Journal of Computer Vision, 2002, 50(2): 171–184. [doi: 10.1023/A:1020346032608]

[24] Krishnamoorthy N, Malkarnenkar G, Mooney RJ, Saenko K, Guadarrama S. Generating natural-language video descriptions using text-mined knowledge. In: Proc. of the 27th AAAI Conf. on Artificial Intelligence. Bellevue: AAAI Press, 2013. 541–547.

[25] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proc. of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893. [doi: 10.1109/CVPR.2005.177]

[26] Chaudhry R, Ravichandran A, Hager G, Vidal R. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 1932–1939. [doi: 10.1109/CVPR.2009.5206821]

[27] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

[28] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[29] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision

and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[30] Wang H, Kläser A, Schmid C, Liu CL. Action recognition by dense trajectories. In: Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2011). Colorado Springs: IEEE, 2011. 3169–3176. [doi: 10.1109/CVPR.2011.5995407]

[31] Wang H, Schmid C. Action recognition with improved trajectories. In: Proc.of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 3551–3558. [doi: 10.1109/ICCV.2013.441]

[32] Wang LM, Qiao Y, Tang XO. Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proc.of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4305–4314. [doi: 10.1109/CVPR.2015.7299059]

[33] Kang K, Li HS, Yan JJ, Zeng XY, Yang B, Xiao T, Zhang C, Wang Z, Wang RH, Wang XG, Ouyang WL. T-CNN: Tubelets with convolutional neural networks for object detection from videos. IEEE Trans. on Circuits and Systems for Video Technology, 2018, 28(10): 2896–2907. [doi: 10.1109/TCSVT.2017.2736553]

[34] Wang Q, Zhang L, Bertinetto L, Hu WM, Torr PHS. Fast online object tracking and segmentation: A unifying approach. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1328–1338. [doi: 10.1109/CVPR.2019.00142]

[35] Chen FH, Ji RR, Su JS, Wu YJ, Wu YS. StructCap: Structured semantic embedding for image captioning. In: Proc. of the 25th ACM Int'l Conf. on Multimedia. Mountain View: ACM, 2017. 46–54. [doi: 10.1145/3123266.3123275]

[36] Niu ZX, Zhou M, Wang L, Gao XB, Hua G. Hierarchical multimodal LSTM for dense visual-semantic embedding. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1899–1907. [doi: 10.1109/ICCV.2017.208]

[37] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[38] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang ZH, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF. ImageNet large scale visual recognition challenge. Int'l Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

[39] Wu X, Li GB, Cao QX, Ji QG, Lin L. Interpretable video captioning via trajectory structured localization. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6829–6837. [doi: 10.1109/CVPR.2018.00714]

[40] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[41] Liu LB, Zhang RM, Peng JF, Li GB, Du BW, Lin L. Attentive crowd flow machines. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 1553–1561. [doi: 10.1145/3240508.3240681]

[42] Wang LM, Xiong YJ, Wang Z, Qiao Y, Lin DH, Tang XO, van Gool L. Temporal segment networks: Towards good practices for deep action recognition. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 20–36. [doi: 10.1007/978-3-319-46484-8_2]

[43] Chen DQ, Manning C. A fast and accurate dependency parser using neural networks. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 740–750. [doi: 10.3115/v1/D14-1082]

[44] Chen DL, Dolan WB. Collecting highly parallel data for paraphrase evaluation. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: Association for Computational Linguistics, 2011. 190–200.

[45] Xu J, Mei T, Yao T, Rui Y. MSR-VTT: A large video description dataset for bridging video and language. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5288–5296. [doi: 10.1109/CVPR.2016.571]

[46] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002. 311–318. [doi: 10.3115/1073083.1073135]

[47] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proc. of the 9th Workshop on Statistical Machine Translation. Baltimore: Association for Computational Linguistics, 2014. 376–380. [doi: 10.3115/v1/W14-3348]

[48] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

[49] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Proc. of Workshop on Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.

[50] Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. The stanford CoreNLP natural language processing toolkit. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore: ACL, 2014. 55–60. [doi: 10.3115/v1/P14-5010]

[51] He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1026–1034. [doi: 10.1109/ICCV.2015.123]

[52] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[53] Pan PB, Xu ZW, Yang Y, Wu F, Zhuang YT. Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1029–1038. [doi: 10.1109/CVPR.2016.117]

[54] Zhang XS, Gao K, Zhang YD, Zhang DM, Li JT, Tian Q. Task-driven dynamic fusion: Reducing ambiguity in video description. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6250–6258. [doi: 10.1109/CVPR.2017.662]

[55] Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3185–3194. [doi: 10.1109/CVPR.2017.339]

[56] Gao LL, Guo Z, Zhang HW, Xu X, Shen HT. Video captioning with attention-based LSTM and semantic consistency. IEEE Trans. on Multimedia, 2017, 19(9): 2045–2055. [doi: 10.1109/TMM.2017.2729019]

[57] Chen YY, Wang SH, Zhang WG, Huang QM. Less is more: Picking informative frames for video captioning. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 367–384. [doi: 10.1007/978-3-030-01261-8_22]

[58] Yan CG, Tu YB, Wang XZ, Zhang YB, Hao XH, Zhang YD, Dai QH. STAT: Spatial-temporal attention mechanism for video captioning. IEEE Trans. on Multimedia, 2020, 22(1): 229–241. [doi: 10.1109/TMM.2019.2924576]

李冠彬(1986－), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习.

刘劲(1995－), 男, 硕士生, 主要研究领域为计算机视觉.

张锐斐(1998－), 男, 硕士生, 主要研究领域为计算机视觉.

林倞(1981－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 机器学习.

刘梦梦(1997－), 女, 硕士生, 主要研究领域为计算机视觉.