从视频到语言: 视频标题生成与描述研究综述

汤鹏杰 [1] 王瀚漓 [2, 3, 4]

摘 要 视频标题生成与描述是使用自然语言对视频进行总结与重新表达. 由于视频与语言之间存在异构特性, 其数据处理过程较为复杂. 本文主要对基于 "编码–解码" 架构的模型做了详细阐述, 以视频特征编码与使用方式为依据, 将其分为基于视觉特征均值/最大值的方法、基于视频序列记忆建模的方法、基于三维卷积特征的方法及混合方法, 并对各类模型进行了归纳与总结. 最后, 对当前存在的问题及可能趋势进行了总结与展望, 指出需要生成融合情感、逻辑等信息的结构化语段, 并在模型优化、数据集构建、评价指标等方面进行更为深入的研究.

关键词 视频描述, 卷积神经网络, 循环神经网络, 语段生成, 情感表达, 逻辑语义

引用格式 汤鹏杰, 王瀚漓. 从视频到语言: 视频标题生成与描述研究综述. 自动化学报, 2022, **48**(2): 375−397

DOI 10.16383/j.aas.c200662

From Video to Language: Survey of Video Captioning and Description

TANG Peng-Jie[1] WANG Han-Li[2, 3, 4]

Abstract The task of video captioning and description is to summarize and re-express the visual content of video with natural language/text. It is challenging because it involves the transformation of different modal information, and there exists heterogeneity between the visual data and language. In this work, the models based on the "encoder-decoder" pipeline are mainly elaborated in detail. According to the encoding and usage of visual features, the current models are classified into four types: the models based on mean/max pooling feature, the models based on video sequential memory, the models based on 3D CNN feature, and the models based on hybrid features. A number of popular works of each type are described and analyzed. Finally, the existing problems and possible trends worth studying are summarized. It is pointed out that the prior knowledge including emotion and logical semantics in complex videos should be further mined and embedded for the generation of logical paragraph description. Moreover, it is still desired to further investigate the techniques of model optimization, dataset construction and evaluation metrics for video captioning and description.

Key words Video description, convolutional neural network, recurrent neural network, paragraph generation, emotion expression, logical semantics

Citation Tang Peng-Jie, Wang Han-Li. From video to language: Survey of video captioning and description. *Acta Automatica Sinica*, 2022, **48**(2): 375−397

收稿日期 2020-08-17 录用日期 2020-12-14

Manuscript received August 17, 2020; accepted December 14, 2020

国家自然科学基金 (62062041, 61976159, 61962003), 上海市科技创新行动计划项目 (20511100700), 江西省自然科学基金 (20202BAB202017, 20202BABL202007), 井冈山大学博士启动基金 (JZB1923) 资助

Supported by National Natural Science Foundation of China (62062041, 61976159, 61962003), Shanghai Innovation Action Project of Science and Technology (20511100700), Natural Science Foundation of Jiangxi Province (20202BAB202017, 20202BABL202007), Ph. D. Research Project of Jinggangshan University (JZB1923)

本文责任编委 白翔

Recommended by Associate Editor BAI Xiang

1. 井冈山大学电子与信息工程学院 吉安 343009 2. 同济大学计算机科学与技术系 上海 201804 3. 嵌入式系统与服务计算教育部重点实验室 (同济大学) 上海 200092 4. 同济大学上海智能科学与技术研究院 上海 200092

1. College of Electronics and Information Engineering, Jinggangshan University, Ji′an 343009 2. Department of Computer Science and Technology, Tongji University, Shanghai 201804 3. Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 200092 4. Shanghai Institute of Intelligent Science and Technology, Tongji University, Shanghai 200092

视频标题生成与描述任务是对给定的视频进行特征抽象, 并将其转换为自然语言, 对视觉内容进行结构化总结与重新表达. 它与目前流行的图像描述任务一样, 同属于计算机视觉高层语义理解范畴, 但鉴于视频数据的时空特性与语义的多样性、复杂性, 其比图像描述更具挑战性.

如图 1 所示, 它不仅需要检测出空间域中的物体、场景、人物等静态要素, 还要能够识别时间域上的动作及事件, 反映各视觉语义对象的时空变化, 最后选择合适的词汇及句式结构将其组合在一起, 形成符合人们表达习惯的描述语句. 该任务对于自动解说、导航辅助、智能人机环境开发等领域应用前景广阔, 在推动旅游、教育及计算机学科本身发展等方面意义巨大. 但由于该任务涉及计算机视觉、自然语言处理, 甚至社会心理学等学科, 数据处理过程较为复杂, 具有很大的挑战性.

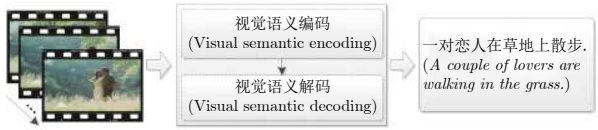


图 1 视频标题生成与描述任务示例

Fig. 1 Example of video captioning and description

视频标题生成与描述研究历史较为悠久. 在其发展早期, 人们主要借助于 SIFT 特征 (Scale-invariant feature transform, SIFT)[1]、方向梯度直方图特征 (Histogram of oriented gradient, HOG)[2] 等手工特征, 采用统计的方式对视频内容进行抽象, 提取视频中的语义特征, 然后运用机器学习、分类/识别、检索、检测等技术获取视觉语义对象, 并将其按照预定模板或规则填入相应位置, 组成可读的描述句子[3−6]. 后来, 人们借鉴机器翻译的流程, 设计出能够生成句式更为灵活、用词更为丰富的 "编码−解码" 框架结构, 提升了生成句子质量[7]. 但受限于手工特征的表达能力, 其生成的句子在准确性和语义丰富程度等方面与人工表达仍有较大差距, 难以满足人们的需求. 随着深度学习技术的发展, 研究人员使用大规模训练数据对深度卷积神经网络 (Deep convolutional neural networks, DCNN) 进行优化[8−11], 并将其应用于视频特征提取[12−14]. 深度特征更加抽象, 表达能力更强, 将其与循环神经网络 (Recurrent neural networks, RNN) 进行结合, 使得生成的句子中词汇更加准确、语义更为丰富. 目前, CNN-RNN 框架已成为视觉描述任务的基础架构. 在此基础上, 研究人员结合三维卷积神经网络 (3D CNN)[15−16]、门限循环单元 (Gated recurrent unit, GRU)[17]、注意力机制[18]、视觉概念/属性机制[19] 等, 设计了多种更为复杂的模型与算法, 进一步改善了视频标题与描述的生成质量.

除对简单视频进行高度总结与抽象, 为其生成简单描述之外, 人们也在寻求对更为复杂的视频进行精细化表达, 或以事件/场景变化为依据, 对其中的视觉语义片段进行更为细致的描述, 或者提取整个视频的逻辑语义, 将各片段描述组合为具有一定逻辑结构的描述语段等. 但由于视频数据的复杂性, 各视觉语义对象本身的变化、各对象之间的逻辑关联及其交互等仍存在建模困难、挖掘与利用不充分等弊端. 同时, 将其映射为更为抽象的词汇表达与逻辑语段也在准确性、连贯性及语义性等方面存在较大挑战, 生成的描述难以应用在实际场景中. 此外, 在复杂视频的情感挖掘与个性化表达方面, 目前尚无较为有效的方法与模型, 生成的描述缺乏生动性与吸引力, 且难以对隐含在视频内部的潜在语义及可能的外延信息进行推理显化与表述, 视觉信息与语言之间的语义鸿沟仍然较为明显.

目前已有部分工作对视频描述任务进行梳理与总结, 如 Aafaq 等总结了当前视频描述的主流方法、数据集和评价指标, 但他们侧重于从学习策略 (如序列学习、强化学习等) 上对各模型进行归类分析[20]. Li 等则从更大的视角出发, 系统总结了视觉 (包括图像和视频) 到语言的建模范式, 并从视觉特征编码方式的层面上对各视频描述主流工作进行了介绍[21]. 本文参考了他们的思路, 但为了更加详细而清晰地呈现视频标题与描述生成的研究脉络, 首先回顾了视频描述研究的发展历史, 对其中典型的算法和模型进行了分析和总结. 然后对目前流行的方法进行了梳理, 尤其是基于深度网络的模型框架, 以视频特征编码方式为依据, 按照不同的视觉特征提取与输入方式, 将各类模型分别归类到基于视觉均值/最大值特征的方法、基于 RNN 网络序列建模的方法、基于 3D 卷积网络的方法, 以及基于混合特征编码的方法. 在每类方法中, 首先对视频简单描述模型进行了举例与概括, 然后对视频密集描述、段落描述等精细化表达模型做了分析与总结. 此外, 还介绍了视频描述任务的各类常用验证数据集及其评价指标体系, 列举了部分典型模型的性能表现, 并对结果进行了对比分析. 最后对视频描述任务面临的问题及可能研究方向进行了阐述与说明.

1 基于模板/规则的视频描述

不同于静态图像, 视频中的视觉内容是动态可变的, 在静态的二维数据基础上, 增加了时间维度, 蕴含的视觉信息更为丰富, 但数据结构也更为复杂. 在为视频生成标题与描述时, 不仅需要考虑每帧上的视觉语义对象, 还需要兼顾对象随着时间的变化及其与环境、其他语义对象的交互. 同时还要考虑多尺度时空上的上下文信息, 对视觉信息进行高度抽象, 并将其表现在生成的描述语句中. 正是由于视频携带了更为丰富的视觉信息, 人们一般认为视频标题生成与描述更具有现实意义, 在自动解说、监控理解等方面具有巨大的应用价值, 因此其发展历史也更为悠久. 在具体方法方面, 早期研究者主要是结合基于模板或固定规则的框架, 设计手工特征从视频中获取视觉语义表达, 或使用识别检测技术检测出人物、动作、场景等, 将其填入预设的语句模板中, 或按照固定规则组合成描述语句. 其基本框架如图 2 所示.

20 世纪 90 年代, Nagel 根据车辆行驶轨迹及其运动类型, 使用计算机视觉技术, 检测车辆运动,

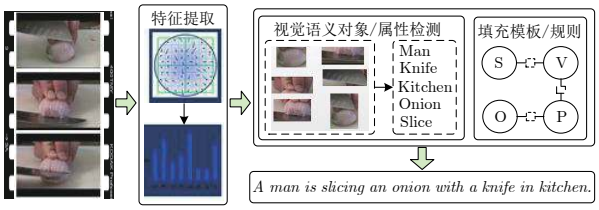


图 2　基于模板/规则的视频描述框架

Fig. 2　The template/rule based framework for video captioning and description

并将运动类型根据一定规则生成车辆行驶的简单自然语言描述, 如 “穿越道路”、“驶离”、“到达”等[3]. Kojima 等为缩小视频图像与文本描述之间的语义鸿沟, 建立视觉对象/动作与特定概念之间的映射关系, 并确定其对应的句法成分 (如谓语、宾语等), 进而组合成为可用的视频描述语句[4]. Gupta等则提出一种故事线模型, 以视频动作或事件为线索, 采用 “与–或图”模型模拟视频内容的运动与变化, 按固定规则为每个动作片段生成单条句子, 然后根据 “与–或图”中的关系, 将句子组合成具有一定逻辑关系的语段描述[5]. Guadarrama 等还通过构建层次化语义模型和小样本学习技术 (如零样本学习) 预测视频可能的动作类型, 并结合人物、对象及场景等视觉因素, 根据预先设定的句子模板元组, 为视频生成可用的标题与描述[6].

可以看出, 由于视频的运动特性, 以上研究工作对于动作语义的捕获更为关注, 它们突破了原有视频动作识别任务中只能根据视频内容输出有限动作类型的限制. 但同时也需注意到, 由于其更为关注动作语义, 忽视了描述语句的其他组成部分, 在句子结构、灵活性表达等方面都受到极大制约. 同时模型使用 HOG、光流方向直方图 (Histogram of oriented optical flow, HOF)、运动边界直方图 (Motion boundary histogram, MBH) 和可形变部件模型 (Deformable parts model, DPM) 等手工特征和检测框架, 借助于支持向量机 (Support vector machine, SVM) 等分类器, 过程较为复杂, 且其各步骤之间是离散的, 用词的准确性和整体语义也难以达到人们的要求.

除以上方法外, 研究人员还结合更为贴近人们表达习惯的流程框架及神经网络等技术方法, 进一步丰富了基于模板或规则的视频描述框架内容. Rohrbach 等首先提出了在视频描述任务中使用 “编码–解码”的流程框架, 将视频的特征表达作为 “源语言”, 待生成的描述语句作为 “目标语言”[7]. 他们使用条件随机场 (Conditional random field, CRF) 对检测到的视觉语义对象进行关系建模, 结合视频及其对应描述的先验知识, 模拟机器翻译流程, 生成更为灵活的描述语句. Xu 等使用深度神经网络特征对视频内容进行编码, 同时使用 Word2Vec 将相应的描述语句解析为具有一定结构的短语或简单句子并提取其特征, 然后将视觉特征与语言特征进行联合嵌入, 对模型进行优化, 最后在测试时可直接生成具有与训练数据相似结构的描述句子[22]. 这两种方法虽然也借用了基于模板的思路, 但已不再是单纯的 “检测–填充”模式, 而是引入了更为先进的思路和方法, 为视频描述研究的进一步发展提供了新的借鉴. 人们虽然对基于模板或规则的视频描述生成方法做了多次改进, 采用更为抽象的特征或性能更为优越的框架, 生成句子的质量也在不断提升, 但它不符合人们表达的习惯, 难以真正有效弥合视觉与语言之间的 “语义鸿沟”, 固定的模板与规则仍会限制视觉语义的高效表达.

2　基于神经网络的视频描述

基于模板或规则的视频描述方法其弊端较为明显, 生成的描述句子在语法结构、语义表达等方面都不够灵活. 目前, 随着深度学习技术的广泛应用, 人们也将其应用在视频描述领域中, 从视频特征编码, 到描述语句生成, 设计了多种有效的模型与方法, 大幅提升了模型性能, 有效改善了生成语句的质量. 具体表现在, 人们参考机器翻译与图像描述中流行的做法, 使用深度卷积神经网络及三维卷积神经网络等对视频进行特征编码, 然后使用 RNN 网络对视觉特征进行解码, 逐个生成词汇并组成句子. 其通用框架与图像描述类似, 是将视频作为 “源语言”, 将待生成句子作为 “目标语言”. 在整个过程中, 其语句的语法、句型结构等不再通过人为设定模板或规则进行干预, 而是直接从训练数据中进行自主学习并记忆. 目前, 基于神经网络的流程与框架, 研究者已开发出多种效果显著的模型与算法. 但不同方法之间差异巨大, 所结合的相关技术涵盖了时序特征编码、检索与定位、注意力机制、视觉属性、对抗学习、强化学习等. 本文主要从视觉特征编码的角度对相关工作进行归纳与梳理, 对各模型与方法的设计动机、原理及所使用的技术进行详细分析.

不同于二维的静态图像, 视频一般包含运动信息. 因此, 其视觉特征编码部分是视频描述过程中的重要一环, 视频特征的抽象程度与表达能力、特征利用的合理性及充分程度等因素, 都将直接影响后续的语言模型所生成句子的质量. 针对视频特征提取问题, 研究者已提出多种效果显著的方法 (如

帧特征均值、光流特征均值、3D 卷积、RNN 网络等), 不同的视频特征提取方法也决定了其语言模型使用特征的方式. 根据视频特征的提取与使用, 本文将现有主要工作划分为四种类型: 1) 视频帧特征均值/最大值方式; 2) RNN 网络序列特征建模方式; 3) 3D 卷积特征建模方式; 4) 混合方式.

2.1 基于视觉特征均值/最大值的视频描述

视频具有多帧特性, 每帧的内容可能互不相同, 但又相互关联. 若只使用其中一帧图像的特征, 对于较为简单的视频 (如单场景、单个动作等), 同样也能够生成可用的描述句子. 但这种做法使得该问题退化为图像描述生成, 没有合理地使用其他关联特征与运动特征, 其生成句子的准确性和语义性都会受到很大影响, 尤其对于复杂视频, 其场景、动作的变换难以准确地被抽象、总结并表达出来. 为充分利用视频数据, 研究者寻求将所有有效帧信息进行融合, 使得每帧上的视觉内容都能够参与模型决策. Venugopalan 等提出一种帧特征均值池化 (Mean pooling) 的方式对视觉特征加以充分利用[1][23]. 他们首先使用在大规模图像分类与识别数据集 ImageNet[24] 上预训练完毕的 AlexNet 模型[8] 提取视频帧的 CNN 特征, 然后将所有视频帧特征使用均值池化以获取最终的视频特征向量, 最后将其送入 RNN 网络 (如长短时记忆网络 (Long-short term memory, LSTM)) 中的每个时间步, 结合已生成的前续词汇, 预测当前时间步的词汇输出, 并组成句子. 该种方法的模型基本架构如图 3 所示.

这种方法在形式上利用了所有视频信息, 但由于采用了计算平均值的方式, 不仅损害了各帧中原有视觉语义的结构化特性, 也没有获取到任何运动信息, 且特征的稀疏性也受到一定的破坏, 因此虽然其性能比其他基于模板的方法有所改善, 但整体结果仍难以满足人们的期望. Pan 等也采取了类似的方法, 但他们不仅使用了视频帧的 CNN 均值特征, 还使用 3D 卷积网络提取不同片段上的三维特征, 然后计算多个三维特征向量的均值, 并将其与 CNN 均值特征结合在一起作为视频的特征表示[25]. 他们采用了视觉模型与语言模型联合优化的方式, 通过设计关联损失函数计算参考句子与视觉信息的误差, 并使用相干损失函数计算生成句子与参考句子之间的误差. 这种方法虽然也对特征进行了均值计算, 但由于 3D 卷积特征包含了视频的部分动态信息, 抑制了时序特征的破坏程度. 同时使用联合训练的方法, 避免了模型陷入局部最优状态, 较好

[1] https://github.com/vsubhashini/caffe/tree/recurrent/examples/youtube

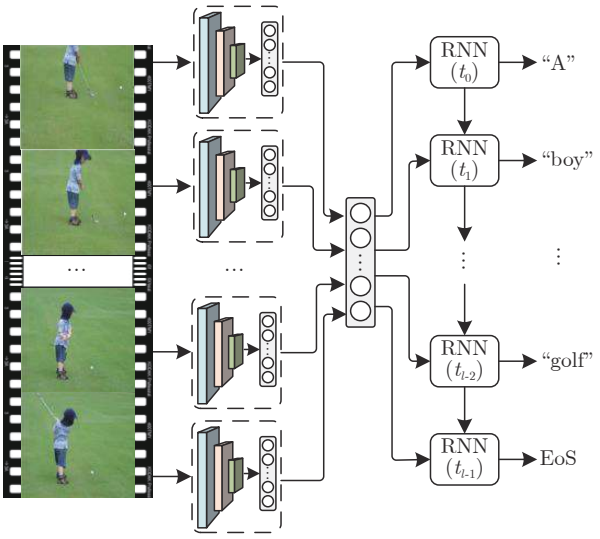


图 3 基于视觉均值/最大值特征的视频描述框架

Fig. 3 The mean/max pooling visual feature based framework for video captioning and description

地改善了模型性能. 此后, 他们还提出了另外一种使用 2D/3D 卷积特征均值的方法, 并结合多示例学习技术 (Multiple instance learning, MIL) 学习视频中的视觉概念, 并将其和均值特征一起送入语言模型, 进而生成句子[26]. 另外, 汤鹏杰等为解决视频特征间隔采样可能造成的视觉信息丢失问题, 提出一种基于密集帧率采样的视频描述模型[27]. 该模型舍弃间隔采样的方法, 而是将一段时间内的所有帧都利用起来, 使用最大值池化的方式获取视频局部特征表达, 然后将池化后的特征按顺序送入 LSTM 网络. 这种方法提升了特征的稀疏程度, 有助于改善模型的泛化能力, 同时由于在 LSTM 每个时间步上处理的是局部特征, 一定程度上抑制了池化操作对视觉语义中结构信息的破坏.

2.2 基于 RNN 序列特征建模的视频描述

连续帧特征均值池化或最大值池化的方式难以捕获视频片段内的时序特征, 造成动态信息的破坏与丢失. 其实, 为获取视频的时序特征, 研究人员设计了多种特征描述子及其模型, 如早期各种光流算法、HOF、MBH[28]、密集轨迹 (Dense trajectories, DT)[29]、改进的密集轨迹 (improved Dense trajectories, iDT) 框架[30], 以及目前常用的双流 CNN 框架 (Two-s)[31]、RNN 循环神经网络模型[17]、3D 卷积网络[15−16] 等. 这些方法已帮助视频动作与行为识别取得了突破性进展. 同样地, 随着对时序特征挖掘的深入研究, 研究者也将其嵌入到视频描述框架中, 进一步提升视频描述质量. 在多种方法中, 使用 RNN 网络对视频帧特征进行序列建模, 设计从视

频帧序列到语言词汇序列的模型框架为视频生成描述语句已逐渐成为一种新的潮流.

序列到序列的建模方式也是起源于机器翻译, 对源语言进行特征提取的编码器和生成目标语言的解码器使用同一个 RNN 网络, 在不同的时间步上实现各自的功能. 对于视频而言, 其视频帧与语言具有相似的表现形式, 都具有时序特性, 因此序列到序列的建模方式应同样适用于视频描述任务. 采用这种方法的一般模型框架如图 4 所示.

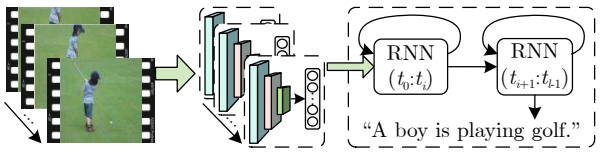


图 4　基于 RNN 序列建模的视频描述框架

Fig. 4　The RNN based framework for video captioning and description

2.2.1　基于 RNN 序列特征的视频简单描述

视频简单描述任务是指给定一段内容较为简单、变化相对较少的视频, 通过模型算法对其主要内容进行高度抽象与总结, 生成句子结构、用词及语义表达都较为简单的描述语句 (一般为一段视频只生成一句话). 本节针对视频简单描述任务, 按照对视频特征的处理方式将其分为基于视频全局特征的方法、基于视频特征选择与优化的方法及基于混合视频特征的方法, 对相关算法与模型进行了梳理与分析.

基于视频全局特征的方法是使用 RNN 网络对视频帧序列进行建模, 将提取到的各帧特征按顺序送入 RNN 网络中, 获取视频的时序动态特征; 然后将编码后的视觉特征送入语言模型进行解码, 在每个时间步上逐个生成词汇, 最终组成可读的描述语句. Venugopalan 等在其设计的 S2VT (Sequence to sequence: Video to text) 模型[2] 中, 首先使用 DCNN 模型提取视频帧特征和光流帧特征, 然后分别将其按顺序送入两条 LSTM 网络中, 对视频进行动态特征编码. 所有视频帧与光流帧特征编码结束后, 模型进入解码阶段, 在每个时间步上, 将处理视频帧 LSTM 的概率输出与处理光流帧的 LSTM 概率输出进行后融合 (Late fusion), 最后使用融合后的概率预测输出词汇[32]. 他们的工作将视觉动态特征编码与解码过程合二为一, 训练时采用端到端的方式, 避免模型陷入局部最优, 测试时只需输入视频帧与光流帧的 CNN 特征序列, 即可获得相应的视频描述. 该研究工作提出了使用序列到序列流程解决视频描述问题的思路, 不仅对视频的静态特征进行充分利用, 也提取并利用动态序列特征, 且模型较为简洁, 生成的句子在准确性和语义性方面都有了较大提升, 但采用光流帧的 CNN 特征对序列建模是冗余的. 首先, 光流本身即是对视频动态特征的发现与挖掘, 所含空域信息较少, 光流帧中的视觉信息也已较为抽象; 其次, 使用 LSTM 对其变换, 可能引起过拟合, 造成模型预测偏差较大, 与视频帧 LSTM 的概率进行融合后, 反而可能会降低整体性能.

以 S2VT 模型为基础, 研究人员也对其进行了多方面改进. Venugopalan 等在其后续的工作中, 使用大规模语料库, 充分挖掘语言先验知识, 在 S2VT 的基础上添加了一条语言挖掘分支, 辅助最终描述语句的生成[33]. Tang 等认识到 S2VT 框架中的弊端, 结合残差机制、多结构 LSTM 序列融合与视觉特征互补等思想, 提出一种 Res-F2F (Residual based fusion of imporved factored and unfactored model) 的视频描述框架[34], 摒弃了使用光流帧 CNN 特征的方法. 他们首先将用于视频帧特征提取的 DCNN 模型在图像描述数据集上使用端到端的方法进行预训练, 使得提取的视频帧 CNN 特征能够快速适应视频描述任务. 为避免单结构 DCNN 模型提取的视觉特征不够全面的问题, 使用多个 DCNN 模型 (GoogLeNet[10]、ResNet101[11]、ResNet152[11]) 分别提取视频帧特征, 然后将其融合以相互补充. 同时, 借鉴 ResNet 中的残差机制, 构建了更深的 LSTM 网络, 使得视频动态特征与语言特征都更为抽象, 增强了其表达能力. 此外, 他们还将因子分解与非因子分解的 LSTM 网络进行融合, 协同决策每个时间步上的词汇输出. 该模型基本框架如图 5 所示.

Bin 等设计了一种双向 LSTM 网络 (Bidirectional LSTM, BiLSTM), 从前、后两个方向上提取视频的时序特征, 在每个时间步上将两个方向的输出及帧序列 CNN 特征融合, 送入另一个 LSTM 网络中, 获取更为全面的时序特征, 最后将其送入语言模型进行解码[35]. Pasunuru 等也采用了双向 LSTM 对视频特征进行特征编码, 但他们认为单独的视频描述任务难以对视频中的时序信息与逻辑动态信息进行充分的提取. 为此, 他们提出一种多任务的学习方式, 使用视频预测任务学习更多的视频上下文知识, 同时使用一种视频蕴含语义推导任务学习视频中更多的语义信息, 最后通过三个任务的联合学习, 提升生成句子的准确性和语义性[36]. 这种方法通过多任务学习的方式改善生成句子的质

[2] https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt

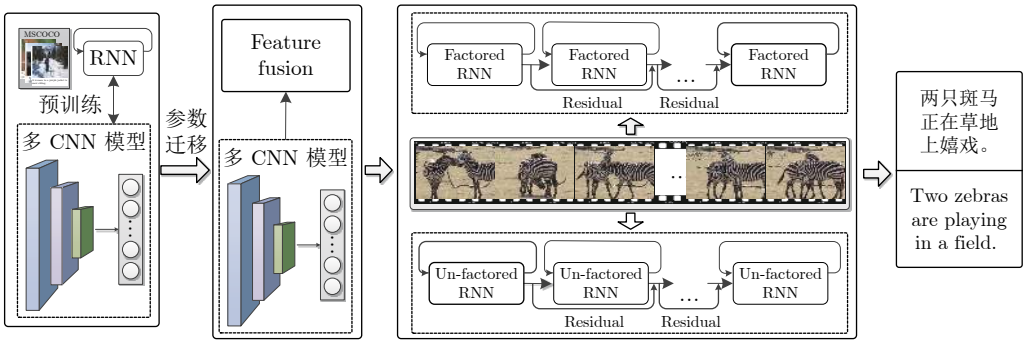


图 5 Res-F2F 视频描述生成流程

Fig. 5 The framework of Res-F2F for video captioning and description

量, 具有良好的借鉴意义. 但其使用的其他两种任务, 无论是视频预测还是蕴含语义推导都属于视频高层语义理解, 生成较为准确的视频帧或根据现有内容推理出其可能的隐含语义都具有极大的挑战性, 其简单的建模过程与粗糙的中间结果并不能真正缩小视觉数据与语言之间的语义鸿沟. 因此, 其模型与结果都有待进一步改善与优化. Li 等使用 LSTM 网络对视频进行序列建模, 也采用了多任务学习方式, 但其所涉及的任务较为简单, 通过属性预测、奖励计算及标题生成等任务构建了一个端到端联合优化的强化学习模型, 有效改善了句子质量[37]. 但在各任务完成过程中, 其中间结果仍较为粗糙, 如在属性预测中只使用帧级均值特征, 限制了模型性能.

除直接使用 LSTM 网络对视频特征进行重新建模外, 研究者也尝试挖掘视频内部的时序关联, 提升时序特征表达能力和语义丰富程度. Pan 等提出一种层次化 LSTM 网络, 底层 LSTM 接收视频帧 CNN 特征, 经过一定固定间隔的时间步后, 其输出送入高层 LSTM, 高层 LSTM 的最终输出作为视频的高层语义表达送入语言模型[38]. 这种方法既降低了模型的运算复杂度, 又能够获取更为抽象的时序特征. Baraldi 等也利用了同样的层次化思想, 但将视频帧的 CNN 特征输入一个特别设计的 LSTM 单元中, 该 LSTM 单元具有检测时序边界的功能; 当遇到场景、动作等发生变化时, 其当前时间步上的输出作为该片段的特征表达, 并将其送入另外一层 LSTM 中. 以此类推, 使用高层 LSTM 的输出作为整个视频的最终时序特征, 并送入 GRU 网络进行解码, 生成描述句子[39]. 与 Pan 等的方法[38]相比, 这种方式具有更好的可解释性, 它能够自动检测视频片段的边界, 而不是特别指定视频中各片段的长度, 输入到高层 LSTM 的特征其表达能力更强, 语义性也更为丰富.

基于视频全局特征的模型能够使得语言模块在解码时参考更多的视觉信息, 尤其是层次化模型, 既考虑了视频中的低层语义信息, 也兼顾了较为抽象的高层上下文信息, 有助于改善生成句子的整体语义. 但对于语言模型而言, 使用全局视觉特征可能会引入额外的视觉噪声, 在某些时刻上, 与该时间步输出无关的视觉信息可能会对模型造成一定的干扰, 影响词汇预测的准确性. 针对该问题, 研究人员借鉴机器翻译与图像描述任务中的注意力机制, 并将其引入到视频描述任务中. 具体而言, 在语言解码阶段, 在不同时间步上关注不同的视频特征 (可为不同的视觉区域、不同的帧或片段等), 根据训练集中的先验知识, 自适应地重点参考视频的局部特征, 提升词汇预测的准确性.

Xu 等将注意力机制引入到视频描述任务中, 首先将视频帧特征、3D 卷积特征以及音频特征使用 LSTM 网络进行序列建模, 然后通过一种自适应的融合单元将多模特征结合在一起, 送入语言解码模块, 并使用多级注意力机制对融合特征与各模态特征进行过滤, 在每个时间步上通过关注不同的视觉信息, 实现词汇的精准预测[40]. Song 等则将注意力机制引入到层次化 LSTM 网络中, 构建了一个包含注意力单元的双层 LSTM 网络 hLSTMatt[41]. 首先使用 DCNN 模型提取视频帧特征, 然后使用注意力机制在每个时间步上决定需要重点关注的视频帧, 并协调是否需要关联相关词汇. 这种方法虽然并没有直接使用 RNN 网络对视频的空域特征进行序列建模, 但使用了注意力机制在不同的时间步上关注不同的空域信息, 对特征进行选择与优化, 其实质上仍属于使用序列模型生成视频描述的范畴. 该方法将时域特征选择与语言生成过程有机融合, 改善了生成描述的准确性与语义性. Li 等为提取更具针对性的视频特征, 也采用了注意力机制, 使用预训练的 DCNN 模型提取视频帧的 CNN 特征后,

在每个时间步上通过注意力单元关注不同帧的特征, 并结合使用单独的记忆网络提取视频帧的序列特征, 为视频生成语句描述[42]. Chen 等认为人们在观看视频时, 对运动信息更为关注, 因此他们将光流帧作为注意力的关注对象, 使用特别设计的门限注意力 RNN 单元 (Gated attention recurrent unit, G-ARU), 对光流帧的 CNN 特征图进行特征选择, 引导语言模型对视频帧的 CNN 均值特征进行解码, 最终生成描述句子[43]. 这些工作在使用视频信息时, 在不同时间步上, 选择符合人们观察习惯的视觉内容, 对大量的冗余特征进行了筛选, 缓解了视觉噪声对语言模型的干扰, 有效提升了生成词汇的准确性. 尤其是结合注意力机制的层次化 RNN 架构, 其模型更为简洁, 对视频内容的利用更加符合人们的直觉性, 实验结果也证明其性能非常优越; 而且, 层次化 RNN 网络不仅在为视频生成单句描述中效果显著, 对于视频的密集描述及结构化表达也具有一定的借鉴意义.

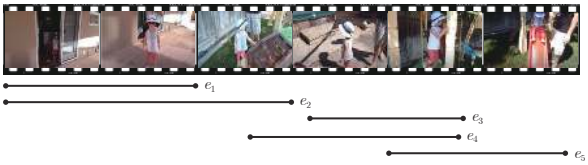
除直接使用 RNN 网络对视频特征进行建模之外, 人们也在寻求使用更为有效的序列特征建模方法. Zhang 等设计了一种双向时序图 (Bidirectional temporal graph) 模型, 对视频中的语义对象进行行为建模, 表征其动态演化轨迹; 然后使用一种卷积门限循环单元 (Convolutional gated recurrent unit, C-GRU) 对各语义对象进行特征聚类, 提取其共性特征, 增强特征的表达能力; 最后结合层次注意力机制 (时序注意力与目标注意力), 使用 GRU 网络对特征进行解码[44−45]. Wang 等采用了一种更为简洁的视觉特征使用方式[46−47], 在提取视频各帧的 CNN 特征后, 借鉴 Yao 等所提出的注意力权重分配方法[48], 直接进入解码阶段, 在不同的时间步上为不同帧的特征赋予不同的权重, 并进行加权融合, 以此作为解码依据. 不仅如此, 他们还设计了一个重构单元, 根据生成的词汇, 对视觉特征进行重构, 并计算重构损失, 以此优化整个模型.

由以上工作可以看出, 以 RNN 网络为基础, 对视频特征进行序列建模, 并为视频生成简单描述, 仍是研究热点之一. 在具体使用时, 则多结合注意力机制, 对视觉特征进行选择与优化, 在此过程中, 还引入知识图模型、重构优化思想等, 对视频中的语义对象及其动作做更为准确的预测与建模. 但在不断提升模型性能的同时, 视频中的情感、逻辑、个性化、隐含语义等却常被忽视, 句子较为呆板, 缺乏吸引力. 因此, 需要融合更多的先验知识对上述因素进行挖掘与表征, 在简单描述的语句中嵌入更多更具吸引力的词汇或短语, 增强句子的 “灵性”.

2.2.2　基于 RNN 序列特征的视频密集描述

对于内容较为复杂的视频, 生成简单的单句描述难以对其进行较为完整的表达, 尤其是在面向真实场景的应用中, 需为其生成更为全面而详细的多句描述. 为此, 研究人员将密集描述的概念引入到视频描述领域中, 并提出了多种性能优越的方法与模型.

具体而言, 视频密集描述是对于给定的内容更为复杂、变化更为丰富的视频, 使用模型算法为其中的多个语义片段分别生成语句描述, 这些描述之间可以是相互独立的, 也可以具有一定关联. 对于视频来说, 其密集描述任务不仅需要考虑空域上的视觉区域信息, 还需要考虑时域上的多粒度/多尺度事件信息. 其任务示例如图 6 所示.



e_1: A man is opening his back door to let a little girl outside.
e_2: The man is following the little girl.
e_3: A dog enters the scene and the little girl jumps with happiness.
e_4: The little girl walks around her slide.
e_5: The little girl starts to climb up the slide and the man helps her get to the top.

图 6　视频密集描述任务示例
Fig. 6　Example of dense video captioning and description

Shen 等提出一种基于弱监督多示例多标签学习的视频密集描述模型[49], 在每个视频帧中选取固定数量的视觉语义区域, 并将不同帧中的这些区域组成多条合理的视觉区域序列, 借鉴图像密集描述工作中所使用的全卷积网络 (Fully convolutional network, FCN)[50] 为各区域提取 CNN 特征, 然后将其送入双向 S2VT 进行时序动态特征编码和解码. 该方法可追踪不同视觉语义对象的变化, 并对其进行自然语言描述. 在此基础上, 用户可根据实际需求自主选择复杂视频中的特定对象 (如人物、物体等), 通过描述句子了解其在视频中的行为及其动态变化. 但由于视频区域序列的组合较多, 其使用次模函数最大化 (Submodular maximization) 的方法进行组合选择将会额外增加模型的复杂度. 此外, 该方法也忽视了不同视觉对象之间的关系与交互, 过于强调单独对象的时序链, 造成生成的语句与实际内容可能存在较大偏差.

Wang 等则从事件的角度出发, 首先使用 3D 卷积网络提取视频特征, 随后将其送入一种基于双向注意力机制的 LSTM 网络, 预测可能事件的边

界, 同时对事件内的视频帧重新进行时序特征提取, 并将其与视频上下文特征进行融合, 送入语言模型[51]. 这种方法根据视频的动态特性, 以事件为基本描述单位, 更为符合人们的观察与表达习惯. 但在待描述的视频场景更为复杂时 (如同一时间段内包含多条事件链), 则生成的密集描述句子面临描述不够精细、表达不够完整等问题. 同样地, Zhou 等也采用事件的概念, 以视频中事件语义的变换为切分点, 为每个视频片段生成单独描述[52]. 但他们为解决复杂视频时序特征的长期依赖问题, 将转换器 (Transformer)[53] 引入到视频序列特征建模中, 代替 RNN 网络对视频进行事件定位, 增强视觉序列特征的表达能力. 同时使用一种遮挡 (Mask) 机制, 用生成的句子对视频中事件的定位位置进行更新修正, 对编码与解码两部分进行端到端训练. 这种使用转换器代替传统 RNN 网络的模型仍属于使用序列建模机制对视频进行动态特征编码, 但由于转换器模型具有传统 RNN 网络难以实现的可堆叠、可并行等特性, 在未来工作中, 尤其是对于视觉高层语义理解任务, 值得进一步挖掘其功能, 探索新的使用方法. 此外, Zhou 等提出一种基于区域注意力的视频密集描述模型[54]. 首先使用快速 R-CNN (Faster region based CNN, Faster R-CNN) 检测各帧中的视觉语义区域, 然后使用带有注意力单元的 LSTM 对视觉区域进行序列建模, 并使用序列均值特征 (时序特征或空域特征) 作为全局特征, 并结合视频片段标记信息 (如片段索引、开始时间、结束时间等), 为视频生成更为准确贴切的描述语句.

2.2.3 基于 RNN 序列特征的视频结构化描述

为视频生成密集描述时, 虽然都是以事件作为主要依据, 提升了描述的可用性, 但生成的描述句子都是独立的, 即其假定各事件之间是离散的, 忽视了其内在的语义关联, 且由于检测到的推荐事件一般过多, 生成的描述冗余性较大. 为此, 研究者提出了视频结构化描述任务, 其在密集描述的基础上, 将各独立的语义片段描述重新整合为具有一定逻辑结构的描述段落. Mun 等认为同一个视频中各事件之间具有时序依赖关系, 人为割裂这种关系可能会造成描述不准确, 与实际内容产生一定偏差. 为此, 他们首先将视频进行等分, 采用 C3D 进行初次特征编码, 然后将其送入 GRU 网络, 搜索可能的事件边界. 然后通过融合注意力机制的 RNN 网络, 建立各事件之间的关联依赖, 并将其按顺序送入语言模型, 逐条生成语句, 最终组成具有一定时序关系的描述[55]. 这种方法已不是单纯地对视频进行密集描述, 而是为其形成了具有一定结构的描述段落, 虽然视频中的复杂逻辑还难以进行有效发现与挖掘, 但事件之间的简单依赖已可以通过部分时序词汇、指代词汇等体现出来, 增强了表达的连贯性、灵活性与吸引力. 实质上, 针对视频的结构化段落描述已吸引了很多研究人员的注意, 对视觉数据进行逻辑化整理与结构化重新表达, 是缩小视觉数据与自然语言表达之间语义鸿沟的重要途径.

Wang 等提出一种基于强化学习的层次化描述框架, 为视频生成细粒度的描述语段[56]. 首先将视频帧的 CNN 特征输入到一个双层 LSTM 网络中, 分别提取视频的低层和高层时序特征, 低层时序特征结合注意力机制输入到工作模块 (Worker), 根据管理模块 (Manager) 设置的目标选择相应操作 (如与环境交互、输出状态等), 而高层时序特征同样结合注意力输入到管理模块, 为工作模块设置目标. 然后使用内部评价模块 (Internal critic) 判别工作模块的目标是否已完成, 并与管理模块进行交互, 其模型框架如图 7 所示. 该工作将更为抽象的视频序列特征输入到管理模块, 使其能够关注粒度更大的全局动态特征; 而将更为具体的时序特征输入工作模块, 关注较小粒度的局部动态特征; 并使两者结合, 通过与周围环境 (上下文) 及评价模块的相互协同, 生成具有结构化的语段描述. 这种使用强化学习的思想符合人们对周围事物的认知规律, 且其将层次化方法较为合理地嵌入到强化学习的框架中, 对于开发出能够实用的视频描述系统具有很大的启发意义.

Xiong 等也采用了强化学习的思路, 为了生成相关性与连贯性强、且语言简洁的段落描述, 首先使用结构化视频片段检测网络 (Structured segment network, SSN) 对视频中的事件进行检测与定位; 然后使用时序片段网络 (Temporal segment network, TSN) 提取视频帧与光流帧特征, 并将其送入一个用于事件片段选择的 LSTM 网络, 逐步去除冗余, 对于选出的每个事件片段, 为其生成单独描述, 并最终组合成具有一定逻辑结构的描述语段[57]. 总体上看, 由于使用强化学习的方法, 其目标函数的设计更为接近评价机制, 以此为基础的模型性能一般要优于使用传统交叉熵的模型方法, 因此, 在解决视觉描述问题上, 这种方法值得进一步探索. 需要注意的是, 这种模型优化策略以评价指标为基础, 即模型认为用于优化的评价指标是正确的. 但是, 通过对各评价指标的分析可知 (见本文第 4.1 节), 每种指标的设计大多是侧重于对句子某一方面的评价, 并不能综合衡量句子的质量, 因此这种方法所带来的高性能还需要使用其他评价方法进行更

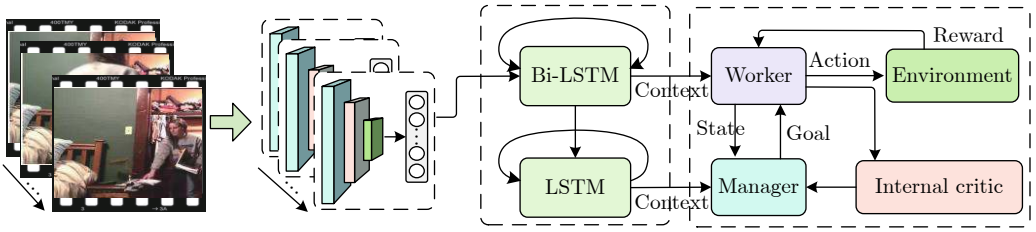


图 7　基于强化学习的层次化视频描述框架

Fig. 7　The reinforcement learning based framework for video captioning and description

为全面的验证.

以上对使用 RNN 网络进行时序特征编码的视频描述模型做了总结与梳理. 从多项研究工作可以看出, 目前使用 RNN 网络仍是视频序列建模的主流方法. 虽然可以使用 RNN 网络提取视频的动态特征, 但直接使用 RNN 对其进行建模也存在着有效信息利用不充分、效率不高等问题. 为此而引入的注意力机制, 能够有选择地对空域或时域特征进行选择与关注, 过滤冗余和无关信息, 改善生成句子的准确性和语义性. 此外, 面向更复杂的结构化描述任务, 研究人员又引入了层次化序列建模、强化学习等技术, 进一步对视频时序特征进行挖掘与利用.

2.3　基于 2D/3D 混合卷积特征的视频描述

使用 RNN 网络或其他同类方法对视频帧 CNN 特征进行编码的方法可以方便实现动态特征提取模块与语言模块的端到端训练, 但 CNN 模块的优化一般是单独进行的, 其与整个模型仍然是分离的, 而且 CNN 特征经过序列变换之后, 也可能导致视频帧中空域信息的丢失. 因此, 除上述方法外, 研究者还经常使用 3D 卷积的方式对视频的时空特征进行编码, 将空域特征提取与时域特征提取融为一体, 既能够提取视频各帧中的静态语义特征, 也能够挖掘时域动态特征.

Ji 等于 2013 年提出一种 3D 卷积网络, 并用于动作识别[15]. 他们将时间维度引入 CNN 网络, 相邻的多个视频帧作为多通道信息进行卷积与池化变换, 最终将所有通道特征合并在一起作为视频的特征表达. 但由于该模型使用的卷积核较大, 且深度不够, 其性能与其他方法相比优势并不明显. Tran 等基于该思想, 使用更小的卷积核, 设计了一种更深的 3D 网络 (C3D), 并在动作识别任务上获得了性能突破[16]. 此后, 人们将 C3D 模型作为视频特征提取的重要手段之一, 并将 C3D 特征应用在多种视觉任务上. 在视频描述领域, 研究人员使用在大型动作/行为识别数据集 (如 Sports 1M[3][58]、Activity Net[59] 等) 上预训练完毕的 C3D 模型提取视频的时空特征, 并结合 RNN、视觉属性、注意力机制等技术, 已取得一系列研究成果, 为该领域的发展提供了新的思路. 基于 3D 卷积特征的视频描述基本框架如图 8 所示, 对视频片段进行 3D 卷积操作之后, 一般还需要结合如均值/最大值融合 RNN 序列建模或注意力机制对 3D 卷积特征进行再次处理, 然后送入语言模型进行解码, 生成描述语句.

2.3.1　基于 3D 卷积特征的视频简单描述

与 RNN 序列建模不同, 3D 卷积网络能同时捕获视频中的空域与时域信息, 对其中的静态视觉语义对象特征和动态视觉事件特征具有较好的表达能力. Yao 等采用 3D 卷积网络提取视频特征, 将其应用于简单描述任务. 该方法首先将视频按时间维度分为多个时空立体网格, 并使用 HOG、HOF 和 MBH 对其进行表达, 然后将其送入优化完毕的 3D 卷积网络, 提取局部时序结构特征. 他们还引入了注意力机制, 在不同时间步为不同的 3D 时空特征分配不同的权重, 指导描述句子生成[48]. Shetty 等为获得更好的模型性能, 使用了视频帧 CNN 均值特征、C3D 特征及多种手工特征, 并使用不同的组合将其送入 LSTM 网络, 生成描述语句; 然后通过评价网络对生成的句子进行评估, 为每个视频选择出最佳的特征组合[60]. 此项工作在 2016 年的 MSR-VTT 视频描述大赛中获得了优异成绩, 这说明 C3D 特征不仅能提取表达能力较强的时空特征, 还可与其他表达能力较弱的特征进行互补, 进一步提升生成句子的质量. 其实, 在 Pan 等使用均值特征的工作[25−26] 及 Mun 等对视频生成密集描述的工作[55] 中, 也结合了 C3D 特征, 使其与其他特征与方法协同工作, 有效改善了模型性能.

Yu 等提出一种基于视线跟踪编码的注意力网络, 将人类的视觉跟踪机制融入到注意力模型中. 该模型设计了一种循环视线预测 (Recurrent gaze prediction, RGP) 模块, 在提取视频的 2D/3D 卷积

[3] https://github.com/gtoderici/sports-1m-dataset/blob/wiki/Project-Home.md

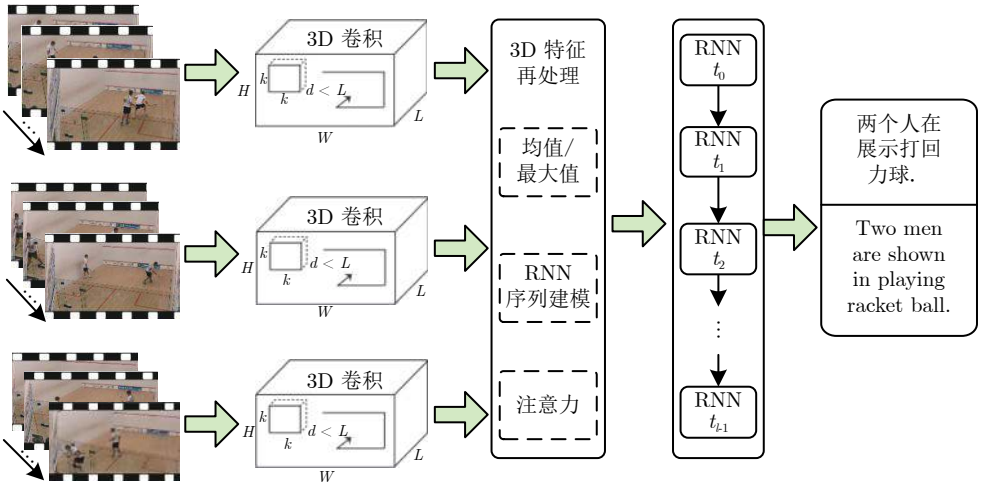


图 8 基于 3D 卷积特征的视频描述基本框架

Fig. 8 The 3D CNN based framework for video captioning and description

特征后, 将其送入该模块中, 得到空间域上的注意力区域, 并对各运动片段的区域特征进行池化表达, 同时使用 GRU 网络对每个运动片段特征进行时域注意力权重分配, 并结合空域注意力特征为视频生成描述语句[61]. Wang 等为了解决 LSTM 网络中多模态信息的长期依赖与语义错位问题, 提出了一种多模态记忆模型 (Multimodal memory model, M3). 首先提取视频帧的 2D (VGGNet[9]、GoogLeNet[10]) 和 3D (C3D[16]) 卷积特征, 然后结合注意力机制, 将其送入 M3 单元, 对视觉信息和语言信息共同建模并记忆其中的序列信息[62]. 这种方法在形式上仍然采用了对视频帧特征进行序列建模的方式, 且与图像描述中的多模 RNN 模型[63−64] 具有相似之处, 但它将视觉与语言信息更为紧密结合在一起, 使其共享记忆单元, 实现视觉与语言的语义对齐, 有效提升了生成句子中用词的准确性.

Pei 等认为当前的主流模型在训练时没有充分挖掘不同视频的共性特征 (如使用同一词汇), 导致生成的句子不能有效利用其他视觉数据的上下文信息. 为此, 他们提出了一种注意力记忆循环神经网络 (Memory-attended recurrent network, MARN), 增强词汇与视觉语义对象的关联性. 该方法首先提取视频的 2D 和 3D 卷积特征, 并通过注意力机制对不同特征进行融合, 同时使用一种记忆结构体记录词汇与视觉语义的映射关系. 最后构建 GRU 网络在每个时间步上输出预测词汇[65]. Li 等也采用了 2D/3D 卷积网络提取视频的静态和动态时空特征, 但他们侧重于不同层级注意力的协同, 从视觉区域注意力、帧级注意力及文本注意力等层面对多模态信息进行协同利用, 提升生成句子的准确性[66−67]. Chen 等采用了融合 2D/3D 卷积特征、MFCC (Mel frequency cepstrum coefficient) 音频特征的多模特征, 对视频进行更为全面的表达. 但他们不是直接将编码后的特征送入语言模型进行解码, 而是设计了一种基于主题引导的描述生成模型. 该方法首先结合视觉特征从参考句子中挖掘可能的隐含主题类型, 然后指导视觉模型根据视觉特征推理出其蕴含的视觉主题, 最后将发现的视觉主题采用因子分解的方式嵌入到语言模型中, 并使用时序注意力引导整条描述句子的生成[68]. 此种方法并没有对视觉特征进行过多处理, 而是通过从已有的描述句子中发现相关视觉内容可能蕴含的主题方向, 是从数据角度进行挖掘与分析; 同时通过对 LSTM 网络中的权重因子矩阵进行分解, 将主题作为其中一项自然地嵌入到每个时间步中, 这与图像风格化描述中 Gan 等的工作[69] 具有相通之处, 都是将某一方面的视觉内容视为矩阵的一部分, 通过矩阵分解实现该方面内容的嵌入与表达. Pan 等则从任务角度出发, 认为对时空语义对象的高层次理解是改善句子质量的关键[70]. 该方法提取视频的 2D/3D 卷积特征作为场景特征, 为模型提供上下文信息; 同时通过目标检测算法提取视频帧中的各语义对象特征, 并建立时空图 (Spatio-temporal graph), 使用图卷积网络提取各语义对象在时空域上的交互特征. 此外, 他们还提出了一种基于对象感知的知识萃取机制, 通过计算语义对象交互与场景上下文两个分支中词汇预测概率分布的 KL 散度进行模型优化, 去除噪声, 实现可用知识的统一表达. Hemalatha 等提出一种特定域语义引导的视频标题生成方法[71], 其首先使用 2D 和 3D 卷积网络提取视频特

征, 然后通过一种局部聚合描述子特征向量 (Vector of local aggregated descriptors, VLAD) 提取方法对特征进行聚合表达, 并根据特定知识领域中的标签 (Tag) 对视频进行描述. Cherian 等为实现时空之间的信息互补, 设计了一种时空/空时注意力模型 (Spatio-temporal and temporo-spatial attention model, STaTS)[72], 具体使用 I3D (Infalated 3D CNN) 模型[73] 及 Faster R-CNN 提取视频特征, 并在空间和时间两个层面加入注意力机制. 类似地, Wang 等也利用了其他知识库的语义标签, 使用 2D/3D 及 Faster R-CNN 对视频特征进行提取后, 结合标签信息, 引导描述句子生成[74]. 此外, Hou 等提出一种使用基于语法表达和视觉线索翻译的视频描述方法, 他们也是使用了 2D/3D 特征对视频进行编码, 然后通过词性标签 (Part-of-speech, POS) 组成的句子模板对句法结构进行表达, 结合视觉信息进行模型学习与优化[75].

此外, Zhang 等也利用了 2D/3D 特征对视频进行视觉特征编码. 他们通过构建一种视觉对象关系图学习帧内和帧间对象的视觉关系, 捕获更多的语义细节, 获得更加丰富的视觉特征, 并采用教师推荐学习 (Teacher recommended learning, TRL) 策略, 引入大量外部语言先验知识, 解决句子中的长尾问题 (Long-tailed problem)[76]. Zheng 等提出一种基于语法感知及动作引导的视频描述模型[77]. 他们认为视频标题生成的主要目的是使用自然语言描述视频中的对象及其交互, 而当前工作多聚焦于视觉对象的检测与使用, 对其中的交互关注较少. 为此, 他们通过同时使用检测到的语义对象和动态信息学习视频中的动作, 并采用多种优化策略对模型进行验证. Hou 等在使用 3D 模型提取视频特征后, 根据常识将视觉特征嵌入语义空间形成语义图, 并通过图神经网络对语义图进行编码与关系推理, 进而生成视频描述[78].

除了直接使用 3D 卷积特征外, 研究者还开发了具有相似功能的时序特征提取技术, 以期能够更加充分地挖掘利用视频中的时空信息. Chen 等为避免使用 RNN 网络对视频进行动态编码时的梯度消失/爆炸问题, 设计了一种时序可变形卷积网络, 通过对输入的 CNN 特征进行时域卷积运算得到新的特征序列. 然后计算所有特征的均值作为全局特征输入语言模型, 并引入时序注意力机制对不同的时域卷积特征进行区别关注, 通过转移卷积网络对输入的特征结合注意力权重进行解码[79]. 这种方法在形式上对 3D 卷积网络进行了再次展开, 将空域特征提取与时域特征提取过程重新做了分离, 通过使用时间长度可变 (可堆叠) 的卷积网络模块对空域特征进行变换, 克服了 3D 卷积网络与 RNN 网络长度固定的弊端, 在功能上扩展了时序特征的语义性. Liu 等也曾为了更加充分有效地利用视觉信息, 开发了一种时序卷积模块 (Temporal convolutional block, TCB) 代替 RNN 对视频进行特征编码, 基于该模块设计了视觉内容编码与 “视觉语义–语言内容”联合编码两条分支, 然后使用注意力机制对两条分支的不同输出部分进行区别关注, 并结合 RNN 网络生成内容描述[80]. 该方法采用了与对抗学习相似的思路, 在视觉内容编码部分使用结合了 TCB 的自动编码器, 通过原视频帧与重构帧之间的差异计算损失, 对该分支进行优化. 同时使用结合 TCB 的视觉–语言联合语义嵌入分支实现视觉内容与语言的语义对齐. 最后将两者的多个输出通过注意力机制进行融合, 提升句子质量. Aafaq 等也认为对视频进行更为有效的特征编码是提升性能的关键因素之一. 他们虽然也使用了 2D/3D 特征, 但并不是将其直接送入语言模型或者通过其他方式对 2D/3D 特征进行选择使用, 而是使用层次化短时傅里叶变换, 对特征进行再次抽象与压缩, 获取多尺度的时空特征; 然后结合目标检测、动作识别, 挖掘更多的视觉对象语义, 组成语义更为丰富的视频特征. 最后经过全连接变换, 其输出送入双层 GRU 网络进行解码, 进而生成描述句子[81]. 可以看出, 这些工作已不满足于 3D 卷积特征的简单使用, 而是从更本质的层次上对其进行变换, 更为全面而深入地挖掘视频中的有效时空特征, 并通过再次选择与优化, 将其有效注入到语言模型中, 为生成更高质量的视觉描述服务.

2.3.2 基于 3D 卷积特征的视频密集描述与结构化表达

与使用序列网络生成视频描述类似, 人们也探索使用 3D 卷积特征为视频生成密集描述语句或结构化描述语段. Yu 等为了弥补单条句子不能完整描述视频内容的缺陷, 设计了一种层次化 RNN 模型, 为同一视频生成多条内容可互补的描述句子[82]. 该方法首先使用 C3D[16] 与 iDT 模型[30] 提取视频特征, 同时与相应的句子词汇特征进行嵌合, 并输入到第一级 RNN 网络中 (具体使用 GRU). 该层输出与视频特征进行结合后, 使用两级注意力机制实现视觉语义与语言之间的对齐. 同时第一级 RNN 的输出也送入第二级 RNN 网络, 与句子级嵌入式特征向量一起, 判断段落的当前状态, 并将其返回第一级 RNN 网络中. 这种方法虽然通过两层 RNN 网络的级联实现密集描述, 但其只侧重于挖掘语言 (句子) 之间的关系, 忽视了视频内容之间的高层语

义关联, 同时也缺乏用于训练的参考段落, 只使用句子级的嵌入特征辅助语段生成, 其生成的多条句子看起来仍然是离散的, 缺少明显的结构化特征, 句子格式和用词也缺乏灵活性. Iashin 等则结合 I3D 特征与 VGGish 音频特征对视频进行编码表达, 并设计了一种双模态转换器 (Bi-modal transformer), 进而实现事件定位与密集描述生成的联合优化与测试[83], 其模型简洁, 性能优越. Park 等则提出一种基于身份感知的视频多条句子生成模型[84]. 他们利用 I3D 模型提取视频时空特征, 并结合脸部特征, 使用转换器模型将其解码为多条语句描述, 并通过人物身份体现多条语句之间的相互语义关联. 这种方式已初具结构化的特征, 但仍较为粗糙, 各语句之间不能体现事件之间的时序、因果等逻辑关系.

Krishna 等则提出了一种面向事件的密集描述方法[85], 使用 3D 卷积网络提取视频特征, 然后将其送入一种改进的深度动作推送 (Deep action proposals, DAP) 模块[86], 以获取不同事件尺度上可能的事件, 并将关于每个事件的隐层输出送入语言模型, 同时结合相邻事件的上下文信息生成事件的描述句子. 该方法不追求结构化的段落描述, 也不刻意挖掘各事件之间的语义关联, 而是以能够生成更加全面而详细的描述为目标, 通过引入多粒度的事件检测机制, 使得用户能够通过描述全方位地了解视频内容. Li 等则更进一步, 他们将事件定位与描述生成进行联合优化, 避免模型陷入局部最优[87]. 具体来说, 该方法首先采用 3D 卷积网络提取视频的片段级时空特征, 然后结合动作、背景等先验知识预测推送的事件, 并划分调整各事件边界. 对各推送事件排序后, 选择置信概率高的事件, 结合视觉属性, 送入语言模型进行解码 (语言模型可单独采用基于 METEOR 指标的强化学习框架进行优化). 整个过程中, 推送事件检测、边界划分、句子生成等环节采用端到端的方式进行训练, 模型较为简洁, 训练较为方便, 且在生成描述时, 去除了大量冗余句子, 具有较高的实用价值. Wang 等使用 C3D 模型提取视频特征后, 使用一种时序事件推送模块对视频事件进行推送, 然后构建一种基于事件的时序–语义关联模型, 为视频生成密集描述[88].

Park 等则针对视频结构化描述生成问题, 采用了对抗学习 (Adversarial learning) 的思路, 使用生成器为每个事件生成多条可用的候选句子, 并通过判别器对其进行最优选择[89]. 具体地, 他们首先提取视频的 2D/3D 卷积特征及视觉区域卷积特征, 并使用注意力机制对特征进行融合, 然后将其送入语言模型. 在训练时, 通过设计视觉判别器判断句子与相应事件的关联程度, 使用语言判别器评估句子结构与语义信息的准确程度, 并通过构建 “语句对”判别器计算不同句子之间的关联程度; 最后根据三者的误差对生成器中的参数进行优化更新.

2.4 混合方式

通过对现有主流模型进行梳理可以发现, 很多模型使用了多种视觉特征处理方法, 如基于 3D 卷积特征的模型, 其一般会结合序列均值特征或 RNN 序列特征建模的方法, 对视频信息进行更为充分的挖掘与利用; 而基于记忆网络序列特征的模型, 也会结合序列均值特征, 指导描述语句的生成.

除使用更复杂的方法对视频特征进行编码外, 研究者考虑在语言模态编码与解码, 以及视觉与语言两种模态信息的相互融合等方面提升模型性能. 随着自然语言处理技术的发展, 研究者已不满足于只从视觉特征的充分挖掘与利用等方面改善模型表现, 而是从语言的语义关联挖掘及其与视觉信息的关联协同出发, 进一步提升模型性能.

不同于传统框架中在语言编码时简单使用独热码 (One-hot) 或 Word2Vec 生成嵌入式向量的方法, 研究者开始研究使用更为复杂的语言处理模型, 以提取表达能力更强的语言特征. 如 Sun 等开发了 VideoBERT 模型[90], 将预训练的语言模型 BERT[91] 引入到了视频描述任务中. 他们使用基于端到端转换器的视频描述框架[52], 利用 VideoBERT 提取视频与语言关联语义特征, 并将其与 S3D (Separable 3D CNN) 特征[92] 结合在一起, 使用转换器为视频生成语句描述. 此外, 他们还提出了一种对比双向转换器 (Contrastive bidirectional transformer) 模型[93], 直接使用 S3D 提取视频特征, 使用预训练完毕的 BERT 模型提取文本特征, 然后将两种特征送入一个交叉的转换器, 并结合注意力机制, 进行多任务训练. Luo 等试图构建一个用于多模态理解与生成的统一视觉–语言预训练模型. 他们采用转换器作为骨干网络, 设计了包括语言和视觉单模编码器、视觉–语言交叉编码器及解码器等在内的多个组件, 并通过多个目标函数对各组件在大规模视频数据集上进行联合优化, 以获得更好的视觉和语言特征表达[94].

2.5 讨论

上述总结与分析说明, 对于视频标题生成与描述的研究目前已经取得巨大进展. 但无论是单独使用序列网络对视频特征进行建模、3D 卷积网络提取视频时空特征, 还是各种混合方法, 其模型对于

[4] https://github.com/google-research/bert

视频内部的情感与个性化信息挖掘都较为欠缺. 一方面是由于视频内容更为复杂多变, 尤其是对于复杂的长视频 (如图 9 所示), 其中可能包含多个需要表现情感的主体, 每个主体随着时间线的推进, 其情感也可能发生变化, 其情感信息的发现与表征都较为困难.

图 9 含有情感与动态时序信息的复杂视频示例
Fig. 9 Video with rich emotion and motion feature

另一方面, 当前也缺少相应的情感描述数据集与合理的评价方法, 而对于包含情感与个性化及其相关变化的数据样本在收集、标注等方面都较以往的其他任务数据集更为费时、费力, 且情感与个性化的评价存在较大的主观成分, 通过自动评价的情感或个性化评价指标设计在合理性解释、有效性证明等方面也存在很大障碍. 针对这些问题, 可以参考视频密集描述工作, 按动作或事件对视频进行划分; 并结合目标检测、表情识别等技术, 对情感主体进行追踪与表征. 同时, 建立时空语义拓扑图, 对其中各主体之间的交互与演化进行表示. 为解决数据集收集与标注困难问题, 一方面可结合小样本学习技术, 通过有限的学习样本, 对未知的视频数据进行预测. 另一方面, 可借鉴弱监督学习方法, 从已有的数据中学习大量的先验知识, 结合现有的训练样本, 辅助情感或个性化句子的生成. 而对于情感与个性化评价方法, 仍可参考现有的自动评测方法, 但需要设计专门的评价机制, 如情感用词准确性、情感句子语义准确性等, 并通过客观实验, 将其结果与传统指标进行对比, 同时也需要开展相关主观实验, 验证指标结果与人类认知共同体的整体契合程度, 对其合理性、可行性与有效性进行证明. 其实, 不仅对于视频标题生成与描述任务, 针对图像描述, 目前虽然已有部分相关方法、模型算法及数据集等[69, 95−100], 但上述问题及思路同样适用于融合情感的图像描述任务, 同样需要在方法、数据集及评价指标等方面进一步研究与探索.

此外, 当前工作难以对视频内部所蕴含的情感与逻辑语义进行挖掘与表达, 其原因还在于关于视频与语言之间的跨媒体/模态之间的转换缺乏更多合理的可解释性分析. 如在训练时, 语言的结构化逻辑信息如何体现在不同的视觉信息中, 如何建立其与各视觉实体之间的关联等; 视频中的情感信息通常是表现在具体的表情、动作或场景中, 判别表现强烈情感 (如高兴、悲伤等) 的属性较为容易 (如可通过大笑、跳舞等表情或动作), 但当情感表现较为微弱 (如内疚、遗憾等) 时, 则必须结合多种视觉信息 (如上下文环境、关系等) 对其进行综合推理. 无论是构建更强更合理的视觉与语言之间的逻辑关联, 还是根据各种先验知识对情感进行推理, 都需要对模型的可解释性进行更加深入的理解与分析, 进而明确其内部的运行机理, 并指导模型的设计与优化.

3 相关数据集与评价方法

视频标题生成与描述的验证与评价比其他传统的视觉任务 (如分类识别[8−11]、目标检测[101−103]、图像/视频检索[104−105] 等) 更加复杂. 在对生成的标题与描述进行统计分析时, 其评价指标不仅需要对词汇预测的精度、句子长度、连贯性进行评价, 还需要对句子的语义丰富程度进行衡量. 在验证数据集的构建方面, 不仅需要考虑视频的类型、复杂程度, 在标注时, 还需要兼顾用词的准确性、与视频内容的关联度, 以及整条句子的连贯性与语义性, 构建过程较为耗时、费力. 而对于更高层次的视频理解与描述任务, 如融合情感、个性化及隐含语义挖掘的视频描述, 其评价指标的设计与数据集构建更为困难. 目前, 针对视频简单描述、密集描述与结构化描述, 已出现多个公开的数据集; 同时, 人们也借鉴机器翻译中的 BLEU (Bilingual evaluation understudy)[106]、METEOR (Metric for evaluation of translation with explicit ordering)[107]、ROUGE-L (Recall-oriented understudy for gisting evaluation)[108] 等评价方法, 并将其引入到视觉描述任务中, 对生成的描述进行多方面的考量. 本节对目前常用的视频描述数据集、相关评价方法, 以及部分模型性能进行了梳理与总结.

3.1 视频描述常用评价方法与指标

由于视觉描述 (包括视频描述与图像描述) 任务与机器翻译具有相似的流程, 其评价也多是借用机器翻译中的思想与方法, 将测试集中的参考句子与生成句子进行对比分析, 统计准确用词或短语的数量, 计算参考句子与生成句子之间的相似程度等. 目前, 在多数视觉描述工作中, 人们一般使用 BLEU[106]、METEOR[107]、ROUGE-L[108] 与 CIDEr (Consensus-based image description evaluation)[109] 等指标对生成的描述句子进行综合评价. 对于 BLEU 指标, 又可分为四个子指标 BLEU-1 (B-1)、BLEU-2 (B-

2)、BLEU-3 (B-3) 与 BLEU-4 (B-4). 该方法主要通过计算生成句子与参考句子中 "n-元组 (n-gram)" 的匹配程度 (其中 $n\in\{1, 2, 3, 4\}$), 为生成句子进行统计评分, n 取值越大, 且 BLEU-n 分值越高, 说明句子的连贯性越好. 该指标中还设计了惩罚因子, 当生成句子长度小于参考句子时, 对句子进行惩罚, 降低相应的分值. BLEU 指标能够对生成句子质量进行较为直接的衡量, 但由于其重点考察生成句子中词汇/短语预测的准确率 (Precision), 未考虑召回率 (Recall), 因此难以反映生成句子的语义丰富程度.

METEOR 方法则同时兼顾了生成句子词汇/短语选择与使用的准确率与召回率. 它使用多种匹配对齐的方式 (精确匹配、同义词匹配、词根匹配) 生成对齐集合, 并以该集合大小与生成句子长度的比值为准确率, 以与参考句子长度的比值为召回率, 然后使用调和均值的方式, 计算生成句子的评价分值. 同样地, METEOR 方法也定义了相应的惩罚因子, 但其对于句子的连贯性更为关注, 当句子中的词汇/短语顺序与参考句子不一致时, 其惩罚因子将发挥作用, 降低对应分值. 该评价方法不仅对生成句子的准确性与连贯性进行较为合理的评价, 对其语义丰富程度也能够进行一定程度的衡量 (使用了同义词匹配与词根匹配), 因此其应用更为广泛. 如在密集描述与结构化描述任务中, 其更注重语义性的表达, METEOR 方法能够较为合理地反映出多条句子语义丰富程度. 除 METEOR 方法外, ROUGE-L 方法也同时考虑了准确率与召回率两个因素. 该方法定义了最长公共子串 (Longest common subsequence, LCS) 的概念, 将参考句子在生成句子中的最长公共子串长度与生成句子长度的比值作为准确率, 以与参考句子长度的比值作为召回率, 最后计算其调和均值作为评价分值. 相对而言, ROUGE-L 方法虽也兼顾了召回率, 但其更关注句子的连贯性, 评价较为单一.

无论是 BLEU、METEOR, 还是 ROUGE-L, 其设计初衷都是为机器翻译而服务, 但在机器翻译任务中, 其语言含义具有确定性, 不同的译者翻译出来的句子差别较小. 对于视觉数据而言, 不同的人由于知识、经验、习惯, 以及对于视觉内容的理解等可能有很大区别, 因此其标注的句子在句式结构、用词/短语、整体表达等方面也存在很大差异. 为此, Vedantam 等提出一种基于标注 "共识"的思想, 并设计了 CIDEr 指标[109], 实现更具针对性的视觉描述语句的语义性评价. 具体来说, CIDEr 将与待描述图像/视频对应的所有参考句子作为一个整体, 统计其中 "n-元组" 的分布, 并以此为依据, 为生成句子中的 "n-元组" 赋予不同的 TF-iDT (Term frequency-inverse document frequency) 权值; 然后以携带 TF-iDT 信息的 "n-元组" 为基础, 计算参考句子与生成句子之间的相似度, 得出评测分值. 此外, Anderson 等则从视觉语义对象准确性的角度出发, 设计了 SPICE (Semantic propositional image caption evaluation) 评价指标[110]. 该方法使用基于概率的上下文无关文法依赖方法, 将参考句子与生成句子都解析成为语义对象场景图的形式, 然后再分别将其转换为 "n-元组"集合, 以此为基础, 计算生成句子与参考句子中各视觉语义对象的匹配程度, 具体借鉴 METEOR 中的方法, 采用精确匹配、同义词匹配与词根匹配的方式统计对齐集合, 然后计算生成句子中语义对象的准确率与召回率, 采用调和均值的方法计算最终得分. SPICE 指标也能够较为合理地衡量生成句子的语义性, 但其较为关注静态视觉语义对象 (如物体、颜色、属性等), 对动态语义 (如动作、关系变化等) 的判断可能不够准确, 影响对整条句子的语义性判断. 在对模型进行具体评价时, 一般都是结合多种指标, 从多个侧面衡量生成句子的质量, 对模型进行更为客观的评价. (关于 BLEU、METEOR、ROUGE-L、CIDEr 等评价指标, 目前已有具体的代码实现[5]).

除以上自动评测方法外, 人们也常使用人工方法对句子进行打分评价. 如在微软举办的视频到文本 MSR-VTT 挑战赛[6][111] 中, 组织者对提交的生成句子不仅使用 BLEU、METEOR、CIDEr 等指标对结果进行评分, 还使用人工对生成句子的连贯性、相关性及可用性 (对盲人的可帮助程度) 等方面进行评比, 力求更全面地对模型性能进行评价. 但对于大规模的测试数据, 人工评价耗费巨大, 且受限于评判者的个人经验, 其结果具有一定的主观性, 在模型复现与对比时, 不易操作, 不同的评判者可能会产生不同的结果.

通过对以上视觉描述评价方法的总结与分析可以发现, 当前指标一般都是面向通用视频描述任务, 每个指标其衡量的侧重点可能有所不同, 但任何单独一类指标都难以真正对句子质量进行较为合理的评价. 尤其是针对如融合情感语义、个性化/风格、逻辑语义等方面的视觉描述任务, 当前方法难以对其进行有效评价. 如对于融合情感语义的视频描述任务, 即使其 BLEU 或者 CIDEr 的分值较高, 但句子中并不一定包含情感信息, 相反地, 生成句子中含有较为丰富的情感语义, 但其它评价指标的分值

[5] https://github.com/tylin/coco-caption

[6] http://ms-multimedia-challenge.com/2017/challenge

也可能较低, 因此只使用现有的评价指标难以对执行这些任务的模型进行较为合理、公平的对比与评价. 在解决这一问题时, 需要设计单独的用词准确性、词汇嵌入的合理性与语义性等更具针对性的评价指标, 同时也需结合现有的其他指标 (如 BLEU、METEOR、CIDEr 等) 对句子进行综合评价.

3.2　视频标题生成与描述数据集

对于视频而言, 其数据形式在二维静态结构的基础上增加了时间维度, 数据结构更加复杂, 对其进行语义抽象并通过自然语言进行表达也更为困难. 因此, 数据集的构建更加耗时费力. 目前用于视频描述的数据集多集中于传统的单句描述, 其描述的视频也多是单个场景或动作, 内容较为简单. 随着研究的深入, 人们提出视频密集描述与段落描述的任务, 由此也产生了用于这些任务的数据集. 本节主要对上述三种数据集进行阐述, 并给出各主流模型在其上的性能表现, 并分析当前存在的问题及面临的困难.

3.2.1　视频简单描述数据集及各模型性能

在传统视频描述领域, 模型提取视频特征, 并据此为视频生成高度概括的描述句子. 为验证模型性能, 研究人员已构建出多个规模不一的相关数据集. 目前流行的常用公开数据集主要包括 MSVD[7][6], 以及更大的 MSR-VTT2016[8][111]. MSVD 数据集由微软研究院发布, 共包含 1 970 个视频, 时长一般较短 (10 ~ 25 秒), 视频内容较为简单, 多是单一生活场景或动作 (如切菜、锻炼等). 该数据集含有多个语种的描述句子, 一般只使用其英文部分, 共有 80827 条句子, 每个视频对应的句子条数不一, 但多数都在 20 条以上. 按照常用的划分标准, 1 200 个视频及对应的 48 774 条句子用于模型训练, 100 个视频与对应的 4 290 条句子用于参数寻优, 其余 670 个视频及其 27 763 条句子用于模型测试. 该数据集应用较为广泛, 是常用的视频描述数据集之一. 其具体示例如图 10 所示.

基于不同视频特征处理方式的部分主流模型在该 MSVD 上的性能表现如表 1~表 4 所示. 由结果可知, 虽然不同模型所使用的视觉特征类型可能有所不同, 直接对比缺乏公平性, 但总体上, 对于视觉特征的处理方式并不是直接决定模型性能的主要因素; 即使是视觉语义结构可能被破坏的序列均值特征方式, 在模型后期对其进行合理的操作后, 仍能获得较为良好的性能表现 (如 RecNet 模型[47-48]). 但与采用强化学习的框架相比, 其性能则稍有落后,

R1: A squirrel is eating a peanut.
R2: A chipmunk is eating a nut.
R3: A squirrel eating a fruit.
R4: A rat is playing and eating something.
R5: A small animal is chewing on a nut.
…

R1: A woman is slicing a strawberry.
R2: She teaching strobery food making.
R3: A chef cuts the leaves off of strawberries.
R4: A person is cutting the stems off of strawberries.
R5: A woman is removing head of strawberry.
…

图 10　MSVD 数据集部分示例 (训练集)

Fig. 10　Examples from MSVD (training set)

表 1　部分基于视觉序列特征均值/最大值的模型在 MSVD 数据集上的性能表现 (%)

Table 1　Performance (%) of a few popular models based on visual sequential feature with mean/max pooling on MSVD

Methods (方法)	B-1	B-2	B-3	B-4	METEOR	CIDEr
LSTM-YT[23]	—	—	—	33.3	29.1	—
DFS-CM(Mean)[27]	80.0	67.4	56.8	46.5	33.6	—
DFS-CM(Max)[27]	79.8	67.3	57.1	47.1	34.1	—
LSTM-E[25]	78.8	66.0	55.4	45.3	31.0	—
LSTM-TSA$_{IV}$[26]	82.8	72.0	62.8	52.8	33.5	74.0
MS-RNN(R)[112]	82.9	72.6	63.5	53.3	33.8	74.8
RecNet$_{local}$(SA-LSTM)[47]	—	—	—	52.3	34.1	80.3

因为强化学习能使模型优化的目标与测试保持一致. 因此, 采用强化学习策略是突破当前性能瓶颈较为有效的技术手段之一 (如 SibNet 模型[80]). 此外, 将多种特征处理方法结合在一起, 进一步改进语言模型, 引入多种领域先验知识 (如构建对象关系图[78, 94] 等), 也可进一步提升词汇预测的准确性和整条句子的语义性.

对于 MSR-VTT2016 数据集[111], 也是由微软研究院收集并发布. 其采用了主题收集的方式, 使用 20 个类别, 包含 257 个常用主题搜索相关视频. 相比于 MSVD, 该数据集更大, 包含了 10 000 段视频

[7] http://www.cs.utexas.edu/users/ml/clamp/videoDescription/YouTubeClips.tar
[8] http://ms-multimedia-challenge.com/2017/dataset

表 2 部分基于序列 RNN 视觉特征建模的模型在 MSVD 数据集上的性能表现 (%)

Table 2 Performance (%) of a few popular models based on visual sequential feature with RNN on MSVD

Methods (方法)	B-1	B-2	B-3	B-4	METEOR	CIDEr
S2VT[32]	—	—	—	—	29.8	—
Res-F2F(G-R101-152)[34]	82.8	71.7	62.4	52.4	35.7	84.3
Joint-BiLSTM reinforced[35]	—	—	—	—	30.3	—
HRNE with attention[38]	79.2	66.3	55.1	43.8	33.1	—
Boundary-aware encoder[39]	—	—	—	42.5	32.4	63.5
hLSTMat[41]	82.9	72.2	63.0	53.0	33.6	—
Li et al[42]	—	—	—	48.0	31.6	68.8
MGSA(I+C)[43]	—	—	—	53.4	35.0	86.7
LSTM-GAN[113]	—	—	—	42.9	30.4	—
PickNet(V+L+C)[114]	—	—	—	52.3	33.3	76.5

表 3 部分基于 3D 卷积特征的模型在 MSVD 数据集上的性能表现 (%)

Table 3 Performance (%) of a few popular models based on 3D visual feature on MSVD

Methods (方法)	B-1	B-2	B-3	B-4	METEOR	CIDEr
ETS(Local+Global)[48]	—	—	—	41.9	29.6	51.7
M^3 -inv3[62]	81.6	71.4	62.3	52.0	32.2	—
SAAT[77]	—	—	—	46.5	33.5	81.0
Topic-guided[68]	—	—	—	49.3	33.9	83.0
ORG-TRL[76]	—	—	—	54.3	36.4	95.2

表 4 其他部分主流模型在 MSVD 上的性能表现 (%)

Table 4 Performance (%) of a few other popular models on MSVD

Methods (方法)	B-4	METEOR	CIDEr
FGM[115]	13.7	23.9	—
TDConvED I[79]	53.3	33.8	76.4
SibNet[80]	54.2	34.8	88.2
GRU-EVE$_{\text{hft+sem}}$(CI)[81]	47.9	35.0	78.1

(总时长约为 41.2 小时), 每段视频对应 20 条参考句子. 按照使用规则, 7 010 段视频及其对应句子用于模型训练和验证 (其中 6 513 段视频与参考句子用于训练, 497 段视频与参考句子用于参数寻优), 其余 2 990 段视频及其参考句子用于测试. 图 11 为该数据集的部分示例.

在该数据上, 目前常用方法的性能表现如表 5~表 8 所示. 由结果可以看出, 在该数据集上, 各模型的性能表现与在 MSVD 数据集上的性能趋势类似, 但整体而言, 采用序列均值/最大值的视觉特征处理方式的模型性能确已落后于 RNN 序列建模与 3D 卷积特征建模方法的模型. 图 12 中展示了部分

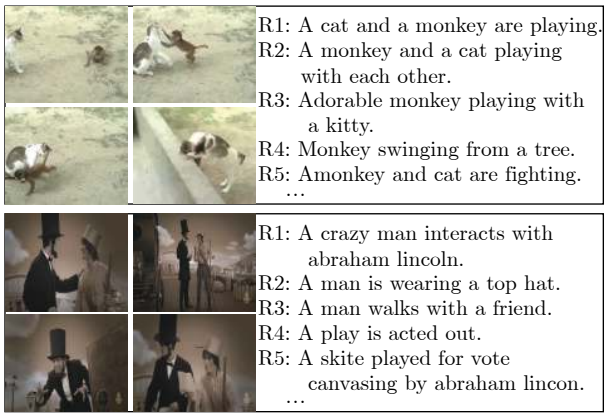


图 11 MSR-VTT2016 数据集部分示例 (训练集)

Fig. 11 Examples from MSR-VTT2016 (training set)

表 5 部分基于视觉序列均值/最大值的模型在 MSR-VTT2016 数据集上的性能表现 (%)

Table 5 Performance (%) of visual sequential feature based models with mean/max pooling on MSR-VTT2016

Methods (方法)	B-1	B-2	B-3	B-4	METEOR	CIDEr
LSTM-YT[23]	75.9	60.6	46.5	35.4	26.3	—
MS-RNN[112]	—	—	—	39.8	26.1	40.9
RecNet$_{\text{local}}$(SA-LSTM)[47]	—	—	—	39.1	26.6	42.7
ruc-uva[116]	—	—	—	38.7	26.9	45.9
Aalto[60]	—	—	—	41.1	27.7	46.4

表 6 部分基于 RNN 视觉序列特征建模的模型在 MSR-VTT2016 数据集上的性能表现 (%)

Table 6 Performance (%) of a few popular models based on visual sequential feature with RNN on MRT-VTT2016

Methods (方法)	B-1	B-2	B-3	B-4	METEOR	CIDEr
Res-F2F (G-R101-152)[34]	81.1	67.2	53.7	41.4	29.0	48.9
hLSTMat[41]	—	—	—	38.3	26.3	—
Li et al[42]	76.1	62.1	49.1	37.5	26.4	—
MGSA(I+A+C)[43]	—	—	—	45.4	28.6	50.1
LSTM-GAN[113]	—	—	—	36.0	26.1	—
aLSTM[117]	—	—	—	38.0	26.1	—
VideoLAB[118]	—	—	—	39.5	27.7	44.2
PickNet(V+L+C)[114]	—	—	—	41.3	27.7	44.1
DenseVidCap[49]	—	—	—	44.2	29.4	50.5
ETS(Local+Global)[48]	77.8	62.2	48.1	37.1	28.4	—

由基于 3D 卷积特征的 SAAT 模型[76] 所生成的句子示例.

3.2.2 视频密集/结构化描述数据集及各模型性能

视频中视觉内容更为复杂, 语义更为丰富, 单条句子更加难以对其进行较为完整的表述. 为此, 研究者提出了一系列视频密集描述或结构化描述方法, 同时构建了多个较为典型的验证数据集. 该类

表 7　部分基于 3D 卷积特征的模型在 MSR-VTT2016 数据集上的性能表现 (%)

Table 7　Performance (%) of a few popular models based on 3D visual sequential feature on MRT-VTT2016

Methods (方法)	B-1	B-2	B-3	B4	METEOR	CIDEr
ETS(C3D+VGG-19)[111]	81.5	65.0	52.5	40.5	29.9	--
M^3 –inv3[62]	—	—	—	38.1	26.6	—
Topic-guided[68]	—	—	—	44.1	29.3	49.8
ORG-TRL[76]	—	—	—	43.6	28.8	50.9
SAAT(RL)[77]	79.6	65.9	52.1	39.9	27.7	51.0

表 8　其他主流模型在 MSR-VTT2016 上的性能 (%)

Table 8　Performance (%) of other popular models on MRT-VTT2016

Methods (方法)	B-4	METEOR	CIDEr
TDConvED (R)[79]	39.5	27.5	42.8
SibNet[80]	41.2	27.8	48.6
GRU-EVE$_{\text{lft+sem}}$(CI)[81]	38.3	28.4	48.1
v2t navigator[119]	43.7	29.0	45.7

RF: {'A man explains how to solve a rubik s cube.'; 'A man points at a rubex cube.';'A person discussing how to solve square puzzle.'; 'A person is solving a rubik scube.'; 'A person showing how to solve a rubix cube.'}
SAAT: A person is solving a rubik s cube.

RF: {'A man is placing a cup into a microwave.'; 'A man using a microwave.'; 'A man heated a cup of coffee in the microwave.'; 'A man is operating a microwaveoven.'}
SAAT: A man is putting a container in the microwave.

图 12　SAAT 模型生成描述句子示例 ("RF"表示参考句子, "SAAT" 表示模型所生成的句子)

Fig. 12　Candidate sentence examples with SAAT model ("RF" stands for references, and "SAAT" denotes the generated sentences with SAAT)

型的数据集构建与标注比一般视频描述更为困难, 不仅需要考虑视频中的动作、事件及场景变换, 还要兼顾各视觉语义对象的不同粒度问题, 标注的工作量也更大. 目前用于视频密集描述或结构化描述的数据集主要包括 ActivityNet Captions[9][86]、YouCookII[10][120] 等.

对于 ActivityNet Captions 数据集, 共包含了约 20 000 个视频片段, 多数视频含有 3 个以上的事件, 每个事件被标注了开始时间和结束时间, 整个数据集约有 100 000 条描述语句. 对于每段视频的描述, 约 94.6 % 的视觉内容都能够被重新表达出来, 同时约有 10 % 的描述内容是重复的, 这也说明不同的事件定位存在相互重叠或覆盖情况. 按照一般的使用方法, 该数据集中 10 024 个视频与其对应的描述语句用于训练, 4 926 个视频及其描述用于验证, 其余的 5 044 个视频及其描述用于模型测试. 在该数据集的验证集上, 当前部分主流模型的性能表现如表 9 和表 10 所示. 其中 SDVC (Streamlined dense video captioning) 模型[55] 生成的部分描述示例如图 13 所示.

表 9　部分基于 RNN 视觉序列特征建模的模型在 ActivityNet captions 数据集 (验证集) 上的性能表现 (%)

Table 9　Performance (%) of a few popular models based on visual sequential feature with RNN on ActivityNet captions dataset (validation set)

Methods (方法)	B1	B2	B3	B4	METEOR	CIDEr
Masked transformer[53]	9.96	4.81	2.42	1.15	4.98	9.25
TDA-CG[51]	10.75	5.06	2.55	1.31	5.86	7.99
MFT[42]	13.31	6.13	2.82	1.24	7.08	21.00
SDVC[55]	17.92	7.99	2.94	0.93	8.82	30.68

表 10　部分基于 3D 卷积特征的模型在 ActivityNet captions 数据集 (验证集) 上的性能表现 (%)

Table 10　Performance (%) of a few popular models based on 3D visual sequential feature on ActivityNet captions dataset (validation set)

Methods (方法)	B1	B2	B3	B4	METEOR	CIDEr
DCE[86]	10.81	4.57	1.90	0.71	5.69	12.43
DVC[87]	12.22	5.72	2.27	0.73	6.93	12.61

从表中结果可以看出, 目前的主流模型在 BLEU 和 CIDEr 等指标上的性能并不优越, 这意味着所生成句子在词汇准确性、连贯性与语义性方面都还存在很大的提升空间. 而且, 当前的结构化描述模型也都是在密集描述数据集上进行验证, 对生成语段的整体连贯性与逻辑性等缺乏较有针对性的评估. 因此, 在视频的精细化描述方面, 包括密集描述、结构化语段描述, 以及融合情感、个性化/风格与逻辑语义的结构化描述等, 还存在大量问题亟待解决, 在模型设计、数据集构建, 以及更为合理、公平的评价指标设计等方面还留有很多空白, 值得进一步研究.

[9] https://cs.stanford.edu/people/ranjaykrishna/densevid/

[10] http://youcook2.eecs.umich.edu/

RF-$e1$: A man is seen speaking to the camera that leads into several clips of a gym.

RF-$e2$: Many people are seen performing gymnastics on a mat while the camera follows close behind.

RF-$e3$: People continue flipping around the gym while also stopping to speak to the camera.

SDVC-$e1$: A man is seen speaking to the camera while holding apole and speakingto the camera.

SDVC-$e2$: The man then jumps onto a mat and begins performing a routine.

SDVC-$e3$: The man continues to perform several more tricks and ends with him jumping down.

图 13 SDVC 模型生成的部分描述示例 ("RF-e"表示参考语句, "SDVC-e" 表示 SDVC 模型生成的句子)

Fig. 13 Description examples with SDVC model ("RF-e" stands for the references, and "SDVC-e" denotes the generated sentences with SDVC)

4 总结与展望

视频描述任务与图像描述类似, 都是将非结构化的视觉数据转换为结构化的自然语言, 其间使用中间语言 (视觉特征) 进行桥接, 以机器学习技术 (尤其是深度学习技术) 为支撑, 运用多种计算机视觉和自然语言处理技术, 为视频生成准确、连贯且语义丰富的描述句子. 目前, 针对图像标题生成与描述, 人们已开发出多种效果显著的模型与方法, 在图像简单描述[18−19, 63−64, 121−124]、图像密集描述[50, 125−127]、结构化段落描述[128−129], 以及情感及个性化描述[69, 95−90]等方面均开展了卓有成效的研究工作. 但由于视频在静态图像的基础上增加了时间维度, 其数据更为复杂, 信息更为丰富, 视觉语义提取与挖掘更加困难, 为其生成可靠且质量较高的描述语句的挑战性也更大. 目前人们已借鉴机器翻译任务的流程与框架, 结合图像标题与生成中的多种技术, 使用 3D 卷积网络、RNN 序列建模机制、注意力机制、视觉属性、视觉概念、层次化序列记忆网络、强化学习技术等, 设计出一系列效果显著的方法与模型, 已能为视频生成简单描述语句, 或为部分视频生成密集描述/结构化描述语句, 推进了视频标题生成与描述任务的进展.

本文系统回顾了视频标题与描述生成的研究历史, 分析了其研究现状与前沿动态, 对当前的主流模型进行了梳理、归纳, 并指出了部分模型的优越性及可能的局限性. 在未来的工作中, 以下几个方面值得进一步研究与探索:

1) 在含有多个场景、人物及事件的复杂视频中, 对其逻辑语义的发现、表征及嵌入的研究尚存在大量空白. 在具体研究中, 不仅要分析视觉信息中各物体、人物、事件之间的关系, 还需要将其映射为自然语言的具体成分, 合理地嵌入到生成的句子中, 实现视频的精细化、结构化表达与描述. 为解决该问题, 一方面可借助于视觉推理技术, 以目标识别与检测等方法完成视觉对象的感知与发现, 以关系检测、图网络等方法发现并构建相应的视觉关系及其演化拓扑, 完成视觉关系的知识图构建与关联推理; 另一方面, 研究视觉关系与语言逻辑之间的映射与转换, 合理使用视觉属性、视觉概念等先验知识, 设计更为鲁棒的层次化序列模型, 实现视觉关联语义到语言逻辑语义的自然嵌入.

2) 视频描述模型的学习代价比一般的分类、识别等任务更为高昂, 其训练数据的收集与标注常耗费大量的人力与物力, 且质量也难以管控. 针对这一问题, 可借鉴零样本与小样本学习技术, 通过样本中的概念与属性推理, 以较少的训练数据实现模型较为充分的优化, 生成较为流畅、语义较为丰富、质量较为可靠的描述句子. 同时也可结合迁移学习及强化学习策略, 引入域外知识, 对模型参数进行快速优化, 或通过不断试错, 增强模型对于正确解的敏感程度, 实现模型在样本受限情况下的自主学习. 除研究模型的优化策略外, 同样也需要构建更为完备的相关数据集, 对其构建方法、标注规则及其质量管控等方面作出更为有益的尝试, 以质量更优的训练数据推进视频描述任务走向实际应用.

3) 在各种复杂视频中, 尤其是包含人物的视频, 其内容常包含丰富的情感变化及隐含语义, 同时不同的视频内容对人们也会产生相应的情感影响或个人理解. 而目前人们在研究视频描述时, 往往只关注其中的事实表达, 对情感、个性化及隐含信息关注较少, 造成生成的句子趣味性、可读性不强. 为此, 需要结合人类的情感心理及视觉情感发现技术, 在表情、动作及上下文语义环境上建立其与情感的映射关系, 并通过视觉属性/概念、注意力机制等技术将情感及个性化信息有机嵌入到生成的句子中. 同时加强对视频描述可解释性的研究, 构建相应的知识图谱, 并结合零样本学习策略, 通过对现有知识的学习, 对视觉信息之外的隐含语义进行预测和推理, 进一步增强生成句子的可用性.

4) 视觉描述任务的评价内容及过程比其他视觉任务更加复杂, 不仅需要判断生成句子对于视频中物体、人物、动作及关系描述的准确性, 还需要对句子的连贯性、语义性及逻辑性进行衡量. 目前的策略多是借鉴机器翻译的评价指标, 评价内容较为单一. 当前虽然也有如 CIDEr、SPICE 等面向视觉

描述任务的评价方法, 但在一些更具针对性的评价任务中, 如对于情感、个性化及逻辑语义的判断与评价, 这些方法都难以对其进行有效的衡量. 因此, 需要结合现有的评价方法设计思路, 开发更为合理的具有针对性及综合性的指标体系, 为模型及其描述提供更为客观、公平的评价机制, 尤其是为强化学习的模型优化方法, 提供更为贴近人们描述与评价习惯的学习与反馈策略.

References

1 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

2 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA: IEEE, 2005. 886−893

3 Nagel H H. A vision of "vision and language" comprises action: An example from road traffic. *Artificial Intelligence Review*, 1994, **8**(2): 189−214

4 Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 2002, **50**(2): 171−184

5 Gupta A, Srinivasan P, Shi J B, Davis L S. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009. 2012−2019

6 Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, et al. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 2712−2719

7 Rohrbach M, Qiu W, Titov I, Thater S, Pinkal M, Schiele B. Translating video content to natural language descriptions. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 433−440

8 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc., 2012. 1097−1105

9 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA, 2015.

10 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 1−9

11 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 770−778

12 Hu Jian-Fang, Wang Xiong-Hui, Zheng Wei-Shi, Lai Jian-Huang. RGB-D action recognition: Recent advances and future perspectives. *Acta Automatica Sinica*, 2019, **45**(5): 829−840
(胡建芳, 王熊辉, 郑伟诗, 赖剑煌. RGB-D行为识别研究进展及展望. 自动化学报, 2019, **45**(5): 829−840)

13 Zhou Bo, Li Jun-Feng. Human action recognition combined with object detection. *Acta Automatica Sinica*, 2020, **46**(9): 1961−1970
(周波, 李俊峰. 结合目标检测的人体行为识别. 自动化学报, 2020, **46**(9): 1961−1970)

14 Wu J C, Wang L M, Wang L, Guo J, Wu G S. Learning actor relation graphs for group activity recognition. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 9956−9966

15 Ji S W, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(1): 221−231

16 Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4489−4497

17 Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: ACL, 2014. 1724−1734

18 Xu K, Ba J L, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015. 2048−2057

19 Yao T, Pan Y W, Li Y H, Qiu Z F, Mei T. Boosting image captioning with attributes. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4904−4912

20 Aafaq N, Mian A, Liu W, Gilani S Z, Shah M. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 2020, **52**(6): Article No. 115

21 Li S, Tao Z Q, Li K, Fu Y. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2019, **3**(4): 297−312

22 Xu R, Xiong C M, Chen W, Corso J J. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas: AAAI Press, 2015. 2346−2352

23 Venugopalan S, Xu H J, Donahue J, Rohrbach M, Mooney R, Saenko K. Translating videos to natural language using deep recurrent neural networks. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: ACL, 2015. 1494−1504

24 Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015, **115**(3): 211−252

25 Pan Y W, Mei T, Yao T, Li H Q, Rui Y. Jointly modeling embedding and translation to bridge video and language. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4594−4602

26 Pan Y W, Yao T, Li H Q, Mei T. Video captioning with transferred semantic attributes. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 984−992

27 Tang Peng-Jie, Tan Yun-Lan, Li Jin-Zhong, Tan Bin. Dense frame rate sampling based model for video caption generation. *Journal of Frontiers of Computer Science and Technology*, 2018, **12**(6): 981−993
(汤鹏杰, 谭云兰, 李金忠, 谭彬. 密集帧率采样的视频标题生成. 计算机科学与探索, 2018, **12**(6): 981−993)

28 Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance. In: Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: Springer, 2006. 428−441

29 Wang H, Kläser A, Schmid C, Liu C L. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 2013, **103**(1): 60−79

30 Wang H, Schmid C. Action recognition with improved trajectories. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 3551−3558

31 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 568−576

32 Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K. Sequence to sequence-video to text. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4534−4542

33 Venugopalan S, Hendricks L A, Mooney R, Saenko K. Improving lstm-based video description with linguistic knowledge mined from text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: ACL, 2016. 1961−1966

34 Tang P J, Wang H L, Li Q Y. Rich visual and language representation with complementary semantics for video captioning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2019, **15**(2): Article No. 31

35 Bin Y, Yang Y, Shen F M, Xu X, Shen H T. Bidirectional long-short term memory for video description. In: Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands: ACM, 2016. 436−440

36 Pasunuru R, Bansal M. Multi-task video captioning with video and entailment generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: ACL, 2017. 1273−1283

37 Li L J, Gong B Q. End-to-end video captioning with multitask reinforcement learning. In: Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, 2019. 339−348

38 Pan P B, Xu Z W, Yang Y, Wu F, Zhuang Y T. Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 1029−1038

39 Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 3185−3194

40 Xu J, Yao T, Zhang Y D, Mei T. Learning multimodal attention LSTM networks for video captioning. In: Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, California, USA: ACM, 2017. 537−545

41 Song J K, Gao L L, Guo Z, Liu W, Zhang D X, Shen H T. Hierarchical LSTM with adjusted temporal attention for video captioning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017. 2737−2743

42 Li W, Guo D S, Fang X Z. Multimodal architecture for video captioning with memory networks and an attention mechanism. *Pattern Recognition Letters*, 2018, **105**: 23−29

43 Chen S X, Jiang Y G. Motion guided spatial attention for video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, **33**(1): 8191−8198

44 Zhang J C, Peng Y X. Object-aware aggregation with bidirectional temporal graph for video captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 8319−8328

45 Zhang J C, Peng Y X. Video captioning with object-aware spatio-temporal correlation and aggregation. *IEEE Transactions on Image Processing*, 2020, **29**: 6209−6222

46 Wang B R, Ma L, Zhang W, Liu W. Reconstruction network for video captioning. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 7622−7631

47 Zhang W, Wang B R, Ma L, Liu W. Reconstruct and represent video contents for captioning via reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(12): 3088−3101

48 Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, et al. Describing videos by exploiting temporal structure. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 4507−4515

49 Shen Z Q, Li J G, Su Z, Li M J, Chen Y R, Jiang Y G, et al. Weakly supervised dense video captioning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 5159−5167

50 Johnson J, Karpathy A, Fei-Fei L. DenseCap: Fully convolutional localization networks for dense captioning. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4565−4574

51 Wang J W, Jiang W H, Ma L, Liu W, Xu Y. Bidirectional attentive fusion with context gating for dense video captioning. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 7190−7198

52 Zhou L W, Zhou Y B, Corso J J, Socher R, Xiong C M. End-to-end dense video captioning with masked transformer. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 8739−8748

53 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017. 6000−6010

54 Zhou L W, Kalantidis Y, Chen X L, Corso J J, Rohrbach M. Grounded video description. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6571−6580

55 Mun J, Yang L J, Zhou Z, Xu N, Han B. Streamlined dense video captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6581−6590

56 Wang X, Chen W H, Wu J W, Wang Y F, Wang W Y. Video captioning via hierarchical reinforcement learning. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 4213−4222

57 Xiong Y L, Dai B, Lin D H. Move forward and tell: A progressive generator of video descriptions. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 489−505

58 Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 1725−1732

59 Heilbron F C, Escorcia V, Ghanem B, Niebles J C. ActivityNet: A large-scale video benchmark for human activity understanding. In: Proceedings of the 2015 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 961−970

60 Shetty R, Laaksonen J. Frame- and segment-level features and candidate pool evaluation for video caption generation. In: Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands: ACM, 2016. 1073−1076

61 Yu Y J, Choi J, Kim Y, Yoo K, Lee S H, Kim G. Supervising neural attention models for video captioning by human gaze data. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 6119−6127

62 Wang J B, Wang W, Huang Y, Wang L, Tan T N. M3: Multimodal memory modelling for video captioning. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 7512−7520

63 Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 2625−2634

64 Tang P J, Wang H L, Kwong S. Deep sequential fusion LSTM network for image description. *Neurocomputing*, 2018, **312**: 154−164

65 Pei W J, Zhang J Y, Wang X R, Ke L, Shen X Y, Tai Y W. Memory-attended recurrent network for video captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 8339−8348

66 Li X L, Zhao B, Lu X Q. Mam-RNN: Multi-level attention model based RNN for video captioning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, 2017. 2208−2214

67 Zhao B, Li X L, Lu X Q. Cam-RNN: Co-attention model based RNN for video captioning. *IEEE Transactions on Image Processing*, 2019, **28**(11): 5552−5565

68 Chen S Z, Jin Q, Chen J, Hauptmann A G. Generating video descriptions with latent topic guidance. *IEEE Transactions on Multimedia*, 2019, **21**(9): 2407−2418

69 Gan C, Gan Z, He X D, Gao J F, Deng L. StyleNet: Generating attractive visual captions with styles. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 955−964

70 Pan B X, Cai H Y, Huang D A, Lee K H, Gaidon A, Adeli E, Niebles J C. Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 10867−10876

71 Hemalatha M, Sekhar C C. Domain-specific semantics guided approach to video captioning. In: Proceedings of the 2020 Winter Conference on Applications of Computer Vision. Snowmass, CO, USA: IEEE, 2020. 1576−1585

72 Cherian A, Wang J, Hori C, Marks T M. Spatio-temporal ranked-attention networks for video captioning. In: Proceedings of the 2020 Winter Conference on Applications of Computer Vision (WACV). Snowmass, CO, USA: IEEE, 2020. 1606−1615

73 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 4724−4733

74 Wang L X, Shang C, Qiu H Q, Zhao T J, Qiu B L, Li H L. Multi-stage tag guidance network in video caption. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, WA, USA: ACM, 2020. 4610−4614

75 Hou J Y, Wu X X, Zhao W T, Luo J B, Jia Y D. Joint syntax representation learning and visual cue translation for video captioning. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 8917−8926

76 Zhang Z Q, Shi Y Y, Yuan C F, Li B, Wang P J, Hu W M, et al. Object relational graph with teacher-recommended learning for video captioning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 13275−13285

77 Zheng Q, Wang C Y, Tao D C. Syntax-aware action targeting for video captioning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020. 13093−13102

78 Hou J Y, Wu X X, Zhang X X, Qi Y Y, Jia Y D, Luo J B. Joint commonsense and relation reasoning for image and video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**(7): 10973−10980

79 Chen J W, Pan Y W, Li Y H, Yao T, Chao H Y, Mei T. Temporal deformable convolutional encoder-decoder networks for video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, **33**(1): 8167−8174

80 Liu S, Ren Z, Yuan J S. SibNet: Sibling convolutional encoder for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(9): 3259−3272

81 Aafaq N, Akhtar N, Liu W, Gilani S Z, Mian A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 12479−12488

82 Yu H N, Wang J, Huang Z H, Yang Y, Xu W. Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 4584−4593

83 Iashin V, Rahtu E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In: Proceedings of the British Machine Vision Conference (BMVC). Online (Virtual): Springer, 2020. 1−13

84 Park J S, Darrell T, Rohrbach A. Identity-aware multi-sentence video description. In: Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 360−378

85 Krishna R, Hata K, Ren F, Li F F, Niebles J C. Dense-captioning events in videos. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 706−715

86 Escorcia V, Heilbron F C, Niebles J C, Ghanem B. DAPs: Deep action proposals for action understanding. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 768−784

87 Li Y H, Yao T, Pan Y W, Chao H Y, Mei T. Jointly localizing and describing events for dense video captioning. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018. 7492−7500

88 Wang T, Zheng H C, Yu M J, Tian Q, Hu H F. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, **31**(5): 1890−1900

89 Park J S, Rohrbach M, Darrell T, Rohrbach A. Adversarial inference for multi-sentence video description. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6591−6601

90 Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: A joint model for video and language representation learning. In: Proceedings of the 2019 IEEE/CVF International

Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 7463−7472

91 Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: ACL, 2019. 4171−4186

92 Xie S N, Sun C, Huang J, Tu Z W, Murphy K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 318−335

93 Sun C, Baradel F, Murphy K, Schmid C. Learning video representations using contrastive bidirectional transformer. arXiv: 1906.05743, 2019

94 Luo H S, Ji L, Shi B T, Huang H Y, Duan N, Li T R, et al. UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv: 2002.06353, 2020

95 Mathews A P, Xie L X, He X M. SentiCap: Generating image descriptions with sentiments. In: Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press, 2016. 3574−3580

96 Guo L T, Liu J, Yao P, Li J W, Lu H Q. MSCap: Multi-style image captioning with unpaired stylized text. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 4199−4208

97 Park C C, Kim B, Kim G. Attend to you: Personalized image captioning with context sequence memory networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 6432−6440

98 Shuster K, Humeau S, Hu H X, Bordes A, Weston J. Engaging image captioning via personality. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 12508−12518

99 Chen T L, Zhang Z P, You Q Z, Fang C, Wang Z W, Jin H L, et al. "Factual" or "Emotional": Stylized image captioning with adaptive learning and attention. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 527−543

100 Zhao W T, Wu X X, Zhang X X. MemCap: Memorizing style knowledge for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**(7): 12984−12992

101 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 580−587

102 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1440−1448

103 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137−1149

104 Yandex A B, Lempitsky V. Aggregating local deep features for image retrieval. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1269−1277

105 Kalantidis K, Mellina C, Osindero S. Cross-dimensional weighting for aggregated deep convolutional features. In: Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 685−701

106 Papineni K, Roukos S, Ward T, Zhu W J. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: ACL, 2002. 311−318

107 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: ACL, 2005. 65−72

108 Lin C Y, Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). Barcelona, Spain: ACL, 2004. 605−612

109 Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 4566−4575

110 Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic propositional image caption evaluation. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 382−398

111 Xu J, Mei T, Yao T, Rui Y. MSR-VTT: A large video description dataset for bridging video and language. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 5288−5296

112 Song J, Guo Y Y, Gao L L, Li X L, Hanjalic A, Shen H T. From deterministic to generative: Multimodal stochastic RNNs for video captioning. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(10): 3047−3058

113 Yang Y, Zhou J, Ai J B, Bin Y, Hanjalic A, Shen H T, et al. Video captioning by adversarial LSTM. *IEEE Transactions on Image Processing*, 2018, **27**(11): 5600−5611

114 Chen Y Y, Wang S H, Zhang W G, Huang Q M. Less is more: Picking informative frames for video captioning. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 367−384

115 Thomason J, Venugopalan S, Guadarrama S, Saenko K, Mooney R. Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. 1218−1227

116 Dong J F, Li X R, Lan W Y, Huo Y J, Snoek C G M. Early embedding and late reranking for video captioning. In: Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands: ACM, 2016. 1082−1086

117 Gao L L, Guo Z, Zhang H W, Xu X, Shen H T. Video captioning with attention-based LSTM and semantic consistency. *IEEE Transactions on Multimedia*, 2017, **19**(9): 2045−2055

118 Ramanishka V, Das A, Park D H, Venugopalan S, Hendricks L A, Rohrbach M, et al. Multimodal video description. In: Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands: ACM, 2016. 1092−1096

119 Jin Q, Chen J, Chen S Z, Xiong Y F, Hauptmann A. Describing videos using multi-modal fusion. In: Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands: ACM, 2016. 1087−1091

120 Zhou L W, Xu C L, Corso J J. Towards automatic learning of procedures from web instructional videos. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, Louisiana, USA: AAAI Press, 2018. 7590−7598

121 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 3156−3164

122 Zhang M X, Yang Y, Zhang H W, Ji Y L, Shen H T, Chua T S. More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing*, 2019, **28**(1): 32−44

123 Yang L Y, Wang H L, Tang P J, Li Q Y. CaptionNet: A tailor-made recurrent neural network for generating image descriptions. *IEEE Transactions on Multimedia*, 2020, **23**: 835−845

124 Tang Peng-Jie, Wang Han-Li, Xu Kai-Sheng. Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using LSTM. *Acta Automatica Sinica*, 2018, **44**(7): 1237−1249
(汤鹏杰, 王瀚漓, 许恺晟. LSTM逐层多目标优化及多层概率融合的图像描述. 自动化学报, 2018, **44**(7): 1237−1249)

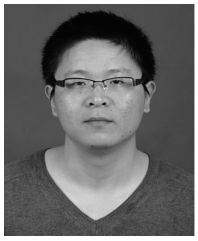
125 Li X Y, Jiang S Q, Han J G. Learning object context for dense captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, **33**(1): 8650−8657

126 Yin G J, Sheng L, Liu B, Yu N H, Wang X G, Shao J. Context and attribute grounded dense captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6234−6243

127 Kim D J, Choi J, Oh T H, Kweon I S. Dense relational captioning: Triple-stream networks for relationship-based captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6264−6273

128 Chatterjee M, Schwing A G. Diverse and coherent paragraph generation from images. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 747−763

129 Wang J, Pan Y W, Yao T, Tang J H, Mei T. Convolutional auto-encoding of sentence topics for image paragraph generation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: AAAI Press, 2019. 940−946



汤鹏杰 井冈山大学电子与信息工程学院副教授. 主要研究方向为机器学习, 计算机视觉, 多媒体智能计算.
E-mail: tangpengjie@jgsu.edu.cn
(**TANG Peng-Jie** Associate professor at the College of Electronics and Information Engineering, Jinggangshan University. His research interest covers machine learning, computer vision, and multimedia intelligent computing.)

王瀚漓 同济大学计算机科学与技术系教授. 主要研究方向为机器学习, 视频编码, 计算机视觉, 多媒体智能计算. 本文通信作者.
E-mail: hanliwang@tongji.edu.cn
(**WANG Han-Li** Professor in the Department of Computer Science and Technology, Tongji University. His research interest covers machine learning, video coding, computer vision, and multimedia intelligent computing. Corresponding author of this paper.)