计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(5):273−282 [doi: 10.15888/j.cnki.csa.009100]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于特征强化与知识补充的视频描述方法①

王　林, 白云帆

(西安理工大学 自动化与信息工程学院, 西安 710048)
通信作者: 白云帆, E-mail: 354992168@qq.com

摘　要: 针对视频描述生成的文本质量不高与不够新颖的问题, 本文提出一种基于特征强化与文本知识补充的编解码模型. 在编码阶段, 该模型通过局部与全局特征强化增强模型对视频中静态物体的细粒度特征提取, 提高了对物体相似语义的分辨, 并融合视觉语义与视频特征于长短期记忆网络 (long short-term memory, LSTM); 在解码阶段, 为挖掘视频中不易被机器发现的隐含信息, 截取视频部分帧并检测其中视觉目标, 利用得到的视觉目标从外部知识语库提取知识用来补充描述文本的生成, 以此产生出更新颖更自然的文本描述. 在 MSVD 与 MSR-VTT 数据集上的实验结果表明, 本文方法展现出良好的性能, 并且生成的内容信息在一定程度上能够表现出新颖的隐含信息.

关键词: 视频描述; 编解码模型; 特征强化; 视觉目标; 知识补充; 人工智能; 自然语言处理

引用格式: 王林,白云帆.基于特征强化与知识补充的视频描述方法.计算机系统应用,2023,32(5):273–282. http://www.c-s-a.org.cn/1003-3254/9100.html

Video Description Method Combining Feature Reinforcement and Knowledge Supplementation

WANG Lin, BAI Yun-Fan

(School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: As texts generated by video descriptions are of low quality and not novel, this study proposes a codec model based on feature reinforcement and text knowledge supplementation. In the coding stage, the model enhances the fine-grained feature extraction of static objects in a video by strengthening local and global features, thus improving the resolution of similar semantics of objects. Then, it integrates visual semantics and video features into a long short-term memory (LSTM) network. In the decoding stage, to mine the hidden information that can hardly be discovered by machines in the video, the model intercepts partial video frames and detects the visual goals in them. Then, the obtained visual goals are used to extract knowledge from the external knowledge base to supplement the generation of descriptive texts and thus produce more novel and natural text descriptions. The experimental results on datasets MSVD and MSR-VTT demonstrate that the proposed method shows good performance, and the generated content can show novel implicit information to a certain extent.

Key words: video description; codec model; feature reinforcement; visual goals; knowledge supplementation; artificial intelligence; natural language processing (NLP)

伴随着互联网的飞速发展以及电子拍摄设备在人群中的高范围普及, 视频数量呈现爆炸式增长. 如何将繁琐的视频内容转为有效的文本信息供人们参考便成为一项研究热点[1,2]. 视频描述任务便是通过计算机将一段视频的内容信息转化为文本信息. 视频描述来自于图像描述, 其本质是实现计算机视觉以及自然语言处理两个模态融合, 将一段视频分解为一帧一帧的图像, 通过对每一张图像地逐帧理解处理, 生成一段能够

① 基金项目: 陕西省科技计划重点项目 (2017ZDCXL-GY-05-03)
收稿时间: 2022-11-07; 修改时间: 2022-12-10; 采用时间: 2022-12-23; csa 在线出版时间: 2023-03-24
CNKI 网络首发时间: 2023-03-27

简单描述该视频的语言文本[3]. 视频描述在视频检索、人机交互、辅助网检人员筛选视频、帮助视力缺陷人群导航等多方面有着宏远的应用市场.

传统的视频描述技术为基于模板的方法[4,5] 和基于检索的方法[6]. 模板方法通过检测视频中物体、状态以及物体之间的关系等, 将检测到的视觉目标与语义信息按照事先设定好的语言模板进行映射, 如主语-动词-宾语等. 通常这类方法生成的文本能够满足句子的基本语法规则, 但过度依赖于设定好的语言模板, 生成的文本描述受到模板的限制而过于单一, 句式缺少了灵活性. 检索方法通过检索大量人工标注的视频样本与给定视频进行映射, 根据相似的文本描述映射得到给定视频的文本描述. 这类方法通常能够得到更加符合人类语言表达的语句, 但过度依赖于样本集的大小以及给定视频与样本集内视频的相似度. 样本集过小或给定视频与样本集内视频不相似都会导致生成的描述语句与原视频内容不符.

如今主流视频描述技术大部分是基于深度学习的编解码框架[7]. 早期工作中 Venugopalan 等[8] 提出 S2VT 模型, 通过端到端训练模型来映射帧序列到单词序列, 该模型未考虑动态相关性. Yao 等[9] 针对视频动态相关性, 提出了一种结合时空 3D-CNN 的短时间动态表示, 引入注意力机制结合递归编码器, 利用局部和全局时间结构生成视频描述. Pei 等[10] 提出 MARN 模型, 在解码器中嵌入记忆体块增强描述质量. Gan 等[11] 利用视觉特征与人工视频标注, 训练一种语义检测网络来得到语义信息增强描述. Chen 等[12] 利用语义检测网络与计划采样辅助模型生成描述. 近年来随着多模态概念的提出, Chen 等[13] 利用双 Transformer 形成框架生成视频描述. 丁恩杰等[14] 提出 MMI 模型从视频的多维度和多模态提取信息进行特征融合, 从视频的静态、动态、地点等多方面维度提取特征传入模型中, 为描述提供更加丰富的视觉特征. 李铭兴等[15] 提出 MABVC 将视频的音频和视觉特征传入到 Transformer 框架中实现多模态信息融合的方式描述视频. Zhang 等[16] 引入外部语言模型来解决描述的长尾问题. 以上方法多关注于如何增多编码解码框架的输入来辅助模型, 却忽略物体相似时, 提取的视觉特征之间存在差异, 导致生成的部分单词不够精确, 使文本质量不高. 并且每个视频的标注语句有限, 想要获得视频中更加新颖的信息就需要从数据集外的知识进行补充. 针对以上两个问题, 本文通过添加特征强化模块使模型提高对视频中静态物体细粒度特征的提取能力, 解决相似物体之间识别的差异问题. 添加知识补充模块在预测单词时通过提高某些新颖单词的概率, 为模型提供更多线索辅助描述生成. 本文的主要工作有: (1) 融入特征强化模块于静态特征网络中, 提高模型对相似物体间的分辨能力; (2) 将知识语库引入编解码器中, 利用训练集中部分没有包含的外部知识来补充模型知识储备, 以此生成更有含义与新颖的语句; (3) 在公共数据集 MSR-VTT 与 MSVD 与上进行模型验证, 并使用评估指标对生成的描述语句进行评分.

1 相关工作

1.1 网络整体框架

本文整体流程如图 1 所示. 将需要生成文本的视频进行处理转化为视频帧序列, 将特征强化模块融入 2D-CNN 中提取视频的静态特征, 再利用 3D-CNN 提取视频的动态特征. 将提取的静态与动态特征通过拼接的方式传入到语义检测网络中得到视觉语义信息. 把视频中的真实标注转换为词向量, 将视频特征和词向量融合语义信息共同传入到 LSTM 进行预测. 同时利用 Faster-R-CNN 检测视频部分帧的视觉目标, 将视觉目标传入到知识补充模块中得到相关联的知识语库, 当预测的单词与检测出的视觉目标一致时, 检索语库中的关联知识并增加其概率, 最终得到的单词由当前预测的单词概率与检索语库中的检索概率共同得到. 通过给每个可能相关联的单词增添概率, 模型就能挖掘一些隐含的信息, 以此得到更加新颖和有含义的描述. 最终输出的所有单词构成整个文本描述.

1.2 视频特征提取

1.2.1 静态特征提取

本文采用 Inception-ResNet-V2 模型提取视频中的静态特征. Inception-ResNet-V2[17] 网络框架主要由 Stem、5 个 Inception-ResNet-A、10 个 Inception-ResNet-B、5 个 Inception-ResNet-C、Reduction-A 和 Reduction-B 几种模块堆叠而成, 如图 2 所示, 其中 keep 0.8 表示神经元保留比例为 0.8. Inception-ResNet-V2 对上一层的输入进行不同尺寸卷积核的卷积操作, 扩增了网络的宽度与深度, 同时使模型对不同尺度大小的图像有更强的适应能力.

1.2.2 特征强化模块

对于视频中物体相似可能存在语义不同的情况, 导致模型根据特征生成的单词可能与真实标注单词并

不相同, 生成的描述准确性降低. 如“boy”与“girl”之间外观相似, 视觉特征如果提取不够精确就会导致生成文本发生混乱, 并且频繁的维度切换可能会导致网络丢失掉某些较为重要的细粒度信息, 使得描述出现偏差.

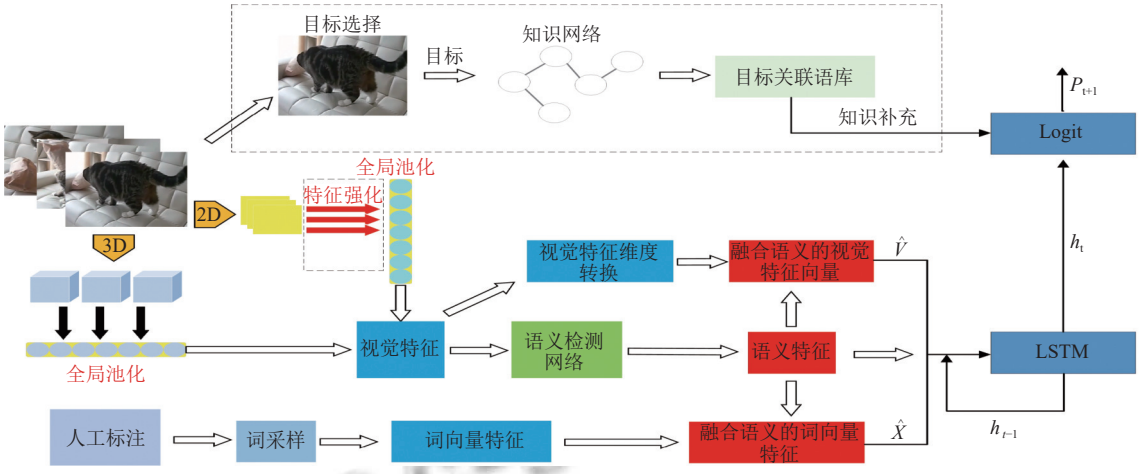


图 1 总体流程结构

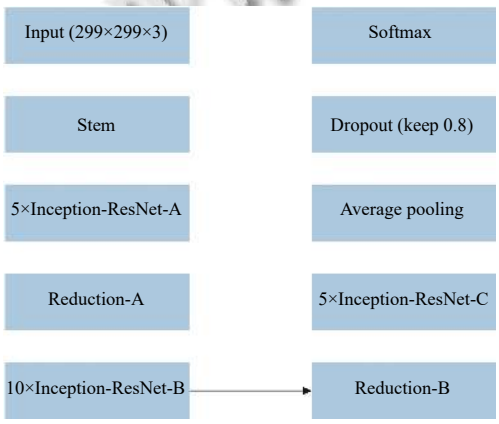


图 2 Inception-ResNet-V2 网络框架

针对上述情况, 本文设计一种特征强化模块 (feature enhancement module, FEM). FEM 具体如图 3 所示, 其中包括 3 条支路: (1) 通道支路 (左侧), 针对特征中的局部信息进行处理. (2) 压缩支路 (中间), 通过压缩通道来降低 FEM 的计算成本. (3) 空间支路 (右侧), 针对特征的全局信息进行处理, 利用空洞卷积对原有特征的感受野进行扩张. 其次将压缩支路的特征与其他两条支路进行融合并相加, 得到的特征经过尾部的空洞卷积进行第 2 次的感受野扩张, 获得经过局部和全局特征强化的融合特征. 强化后的特征 Y 计算公式如式 (1) 所示:

$$Y = C_{3,3}[C_1 \otimes \delta(C_1(C_{3,2}(X))) \oplus \delta(P(C_1(X)))] \quad (1)$$

其中, X 为上一层的输出特征, $C_{3,3}$和$C_{3,2}$为空洞数为 3 和 2 的 3×3 空洞卷积, C_1为 1×1 卷积, $\otimes$为元素相乘, $\oplus$为元素相加, δ为 Sigmoid 激活函数.

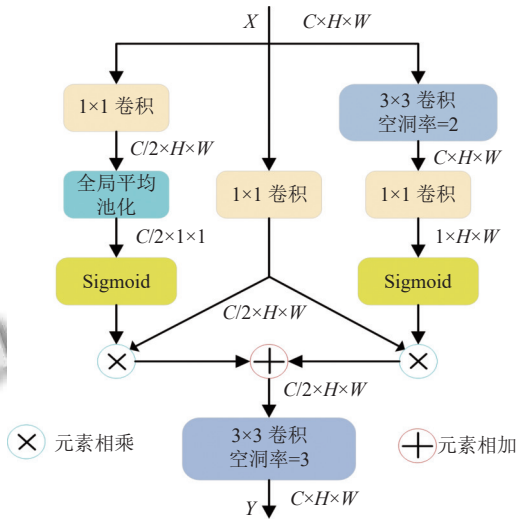


图 3 特征强化 FEM 模块

1.2.3 改进静态特征提取网络

将 FEM 添加于图 2 中 5 个 Inception-ResNet-C 堆叠之后, 如图 4 所示. 通过添加 FEM 模块, 将原网络提取的特征利用空洞卷积使感受野扩张获取全局强化型特征, 并通过 1×1 卷积实现局部通道间的信息交互获取局部强化型特征. 融合全局与局部强化后的特征并再次利用空洞卷积使感受野扩展得到整体强化特征. 强化后的特征相较于原有特征在细粒度方面更加精确,

使模型提高相似物体特征之间的分辨能力, 引导模型生成更加准确的视频静态特征.

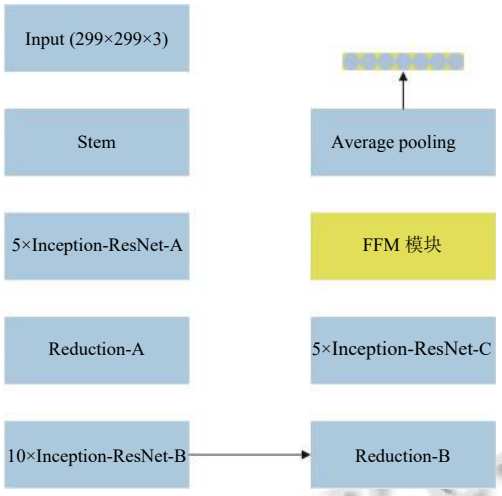


图 4 增添 FEM 的 Inception-ResNet-V2 网络框架

为了提取视频的静态特征, 首先对视频经过预处理, 每个视频提取 32 帧得到视频序列帧$I=(I_1,I_2,\cdots,I_m)$, 其中 m 为提取的有效帧的个数, 输入尺寸大小设置为 299×299, 每个图片的通道数为 3, 高和宽为 299. 通过改进后的 Inception-ResNet-V2 网络提取视频帧序列 I 的静态视觉特征, 加载在 ImageNet[18] 数据集上预训练的模型, 调用 Inception-ResNet-C 层之前的权重, 取经过平均池化层后得到的 1 792 维张量作为静态特征V_s.

1.2.4 动态特征提取

针对视频中时间尺度上的动态特征, 选择高效卷积神经网络 ECO (full)[19] 提取动态特征. 如图 5 所示. 该网络采用随机取样的方式整合各个时间段的特征信息, 利用 2D 卷积提取视频采样帧的特征并在时间维度上拼接, 拼接后的特征传入 3D 卷积处理在时间尺度上的信息, 增添 2D 卷积进一步利用 2D 部分的信息. ECO 网络通过 2D 加 3D 的组合方式更充分地进行特征提取, 在视频分类上达到较好精度的同时, 速度也更快.

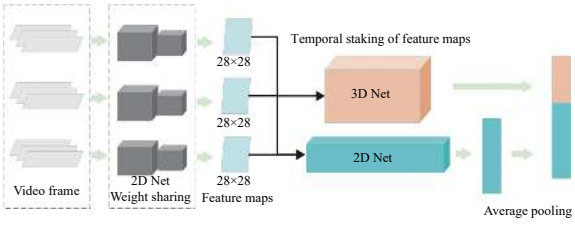


图 5 高效卷积神经网络 ECO (full) 网络框架

在编码阶段, 将视频按照每 32 帧为一组的方式得到视频帧序列, 将帧序列作为输入传到 ECO 网络中进行特征提取, 载入在 Kinetics-400 数据集[20] 上的预训练模型, 去除全连接层, 在平均池化层之后的 1 536 维张量作为动态特征V_d.

1.3 语义检测网络

本文采用基准模型[12] 中的方法, 人工地从数据集的训练集与验证集中选择出现次数较高的K_V个具有确切意思的词语作为语义词, 其中包含名词、动词、形容词等, 不包含“a”“with”等无意义词语. 语义检测类似于为将视频进行多标签分类, 输入为视频的整体特征V_c, 其中, V_c为静态特征V_s与动态特征V_d的拼接, 即$V_c=Concat(V_s,V_d)$. 输出为K_V维的语义向量S_i, 每个维度上的值对应该位置上所代表语义词的概率值, 此概率值反应视频属性语义的大小, 均属于 [0, 1] 之间. 为了训练语义检测网络, 需要对每个视频标注真实的语义标签, 引导语义检测网络生成与其接近的语义信息. 具体实现方式如下: 为每个视频生成一个K_V维的零向量, 检索视频的所有标注, 发现某个语义词出现在标注中, 便将该位置对应元素标注为 1, 反之为 0, 最终得到的向量作为该视频的真实语义标签$\tilde{S}_i$. S_i是第 i 个视频的语义信息, 其中$S_i=\sigma(f(v_i))\in(0,1)^K$, 其中 $f(\cdot)$ 为多层前馈神经网络, $\sigma(\cdot)$ 为 Sigmoid 函数, 它们与损失函数$L(S_i,\tilde{S}_i)$共同组建了语义检测网络. 其中损失函数为式 (2), 本质上为多标签多分类网络的交叉熵损失函数, 表示视频训练出的语义信息与真实语义信息之间的损失值.

$$L\left(S_i,\tilde{S}_i\right)=\frac{1}{N}\sum_{i=0}^{N-1}\sum_{j=0}^{K-1}\left(\tilde{S}_{i,j}\log S_{i,j}+\left(1-\tilde{S}_{i,j}\right)\log\left(1-S_{i,j}\right)\right) \tag{2}$$

从训练集中选取出现频率较高的 300 个单词, 以此 300 个词作为标签分类通过语义检测将每个视频得到所对应的真实语义标签, 再将整体特诊输入到语义检测网络中获取每个视频所对应的语义特征向量S_i, 其中$S_i=[s_{i1},s_{i2},\cdots,s_{iK_V}]$, 其中$K_V$=300.

采用 GloVe 的预训练模型将视频的人工标注语句转化为单词库, 选取出现概率前K_W个单词作为模型生成文本所需要的标注文本并进行编码, 得到K_W个单词的特征编码$\{k_1,k_2,\cdots,k_{K_W}\}$, 每个单词为 300 维的词向量.

1.4 文本解码生成

本文使用基准模型[12]中的 S-LSTM 作为本文的解码模型, 其能借助语义信息来辅助生成更加准确的文本. 因为视觉特征要和词向量特征同一纬度, 所以通过转换矩阵得到输入视觉特征V, 即$V = wV_c$. S-LSTM 的输入为视频的视觉特征V、语义信息S、t 时刻的输入词向量X_t以及上一时刻隐藏状态的输出h_{t-1}. 将语义信息S融入到视觉特征V、输入词向量X_t、上一时刻隐藏状态h_{t-1}, 得到融合语义信息的视频特征V'、t 时刻的输入向量X_t'以及上一时刻隐藏状态h_{t-1}', 如式 (3) 所示:

$$\begin{cases} V_j' = ((V \times W_{j,a}) \odot (S \otimes W_{j,b})) \times W_{j,c}, \ j \in \{i,c,f,o\} \\ X_{j,t}' = ((X_t \times P_{j,a}) \odot (S \otimes P_{j,b})) \times P_{j,c}, \ j \in \{i,c,f,o\} \\ h_{j,t-1}' = ((h_{t-1} \times U_{j,a}) \odot (S \otimes U_{j,b})) \times U_{j,c}, \ j \in \{i,c,f,o\} \end{cases} \tag{3}$$

其中, i,c,f,o分别代表输入门、细胞状态、遗忘门和输出门, W, P, U为网络训练参数, n_v是视觉特征的维度大小, n_h是视频语义特征的维度大小, n_h为 S-LSTM 输出维度的大小, $\odot$ 为哈达玛积运算, $\times$为矩阵相乘. 将融合语义信息后的视觉特征V', t 时刻单词输入向量X_t', h_{t-1}'传入到解码网络当中, S-LSTM 的输入门i_t、遗忘门f_t、输出门o_t、细胞状态c_t'和隐藏状态h_t的计算方式与标准 LSTM 类似, 如式 (4) 所示:

$$\begin{cases} i_t = \sigma(X_{i,t}' + h_{i,t-1}' + V_t' + z_i) \\ f_t = \sigma(X_{f,t}' + h_{f,t-1}' + V_f' + z_f) \\ o_t = \sigma(X_{o,t}' + h_{o,t-1}' + V_o' + z_o) \\ c_t' = \tanh(X_{c,t}' + h_{c,t}' + V_c' + z_c) \\ c_t = f_t \odot c_{t-1} + i_t \odot c_t' \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{4}$$

其中, σ为 Sigmoid 函数, z_j为偏置项, $j \in \{i,c,f,o\}$, tanh 为双曲正切函数. 当得到 t 时刻隐藏状态时, 可通过式 (5) 预测下一个单词, M_g为转移矩阵.

$$p_{t+1} = w_{t+1}^{\mathrm{T}} M_g h_t \tag{5}$$

1.5 知识补充模块

在视频描述的任务中, 描述语句的丰富性也是十分重要. 描述语句的主要来自于目前已有数据集中每个视频所对应的真实标签语句, 这些真实标签是人类为机器自动生成文本描述所提供的知识, 可以称作为内部知识. 但是随着视频越来越多, 其中内容难免有在一些数据集中不存在的知识, 这导致视频描述出的语句不够新颖. 因此, 本文考虑从外部资源中补充知识来辅助视频描述生成, 以此来增强模型的泛化性. 本文采用 ConceptNet[21]来辅助模型进一步地理解视频图像中的隐藏含义. ConcepNet 是一个包含多信息语义知识的开放式网络, 容纳了许多与人类日常生活紧密相关的常识性知识.

图片中每一个视觉目标均可以表现为此目标与另一目标在现实世界之间的关系. 为了得到视频中可能隐含的知识关系, 从预处理的视频中等间隔提取 5 帧的图片传入 Faster-R-CNN[22]中检测相关的视觉目标, 将得到的视觉目标传入到 ConceptNet 中得到在语义上有关联的知识. 因为检测出的目标受限于标签类别, 本文选择检测到的频率为前 3 的目标作为视觉目标, 数量不够 3 时以 0 填充. 图 6 表示外部知识补充过程. 如图 6 所示, 检索的每个知识后会得到一些相关语义知识的关联概率值, 这些概率值将成为模型利用知识的重要依据. 考虑到目标与不同常识性知识间有权重, 本文选择前 4 个且 weight≥1.0 的知识作为语库, 不够 4 时以 0 补充. 当检测到“cat”后, 将目标传入知识网络中得到“walk”“sleep”等关联知识, 每条关联知识都有相应的权重, 选择与目标相关的知识作为一个小型知识语库W_L以词嵌入的方式用来辅助模型生产单词.

如果直接将知识语库作为输入传入到 LSTM 中进行训练可能会产生不必要的噪声, 从而使得模型的性能降低. 因此, 本文并不将这些知识信息作为输入传入到 LSTM 中, 而是在模型预测单词时, 增加知识语库的词汇表中某些单词的概率, 使模型能够发现一些隐含信息, 为此将第 1.4 节中生成单词的式 (5) 的生成机制修改为式 (6):

$$p_{t+1} = \begin{cases} w_{t+1}^{\mathrm{T}} M_g h_t + \lambda p_l(w_{t+1}), \ w_{t+1} \in W_L \\ w_{t+1}^{\mathrm{T}} M_g h_t, \ \text{其他} \end{cases} \tag{6}$$

其中, λ为超参数, 用来调控外部知识的影响程度, 可自行设置. 当生成的单词w_{t+1}出现在构建的知识语库W_L中时, 则预测单词的生成概率由生成机制的概率和检索出的外部知识概率$p_l(w_{t+1})$共同决定. 应用 Softmax 函数得到一个归一化的单词概率分布, 在预测当前的单词时通过贪心搜索从单词的概率分布中选出合适的单词. 通过给每个可能的单词增添一个额外的概率, 能够使模型可能生成一些新颖更加有意义的描述.

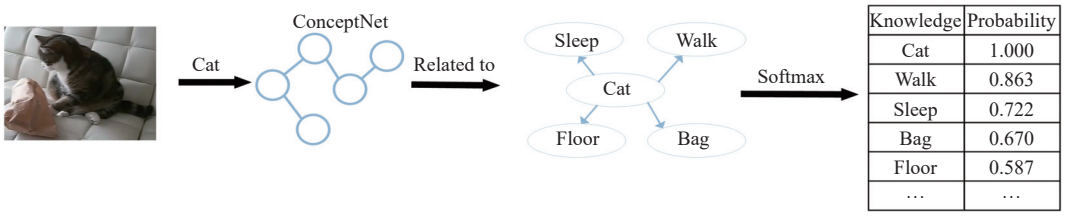


图 6 外部知识补充过程

2 实验设置

2.1 数据集和参数设置

模型在 MSVD[23] 和 MSR-VTT[24] 这两个公开数据集上进行实验验证. MSVD 数据集有 1 970 个短视频, 每个视频平均时长在 9 s, 且有 40 个人工标注的英文语句. MSR-VTT 数据集由多个视频剪辑而成, 共有 10 000 个短视频, 每个视频剪辑在 10–20 s, 且有 20 个人工标注的英文语句. 构建词典来自于两个数据集人工标注语句组成的语库, 删除语库中的生僻词和标点符号, 无法识别的词转为<UNK>. MSVD 和 MSR-VTT 构建的字典库分别为 12 954 和 13 794. 将数据集按照标准拆分为如表 1.

表 1 数据集拆分

数据集	训练集	验证集	测试集
MSVD	1 200	100	670
MSR-VTT	6 513	497	2 990

本文实验环境如表 2 所示. 其中训练隐藏层状态维度设置为 512, 训练 batch_size 大小设置为 64, 初始学习率为 0.000 1, 模型生成最大长度为 20, 最小为 1, 损失函数采用稀疏交叉熵损失函数. 采用 TensorFlow 框架中的 Adam 优化器优化参数, 训练 epoch 设置为 50.

表 2 实验环境配置

名称	环境配置
操作系统	CentOS 7.0
处理器	Intel(R) Xeon(R) CPU E5-2640 v4 @2.20 GHz
显卡	4块GeForce GTX 1080Ti , 11 GB
深度学习框架	TensorFlow 1.15
集成开发环境	Anaconda
内存	128 GB

2.2 训练策略

本文采用计划采样 (scheduled sampling)[12] 的训练策略, 即以采样率 p 选择模型自身的输出作为输入传到下一时刻, $1-p$ 选择真实标签作为输入传到下一时刻. 在训练过程中, 以概率的大小决定下一时刻的输入, 具体公式见式 (7):

$$p = p + epoch \times ratio \tag{7}$$

其中, *epoch* 为训练周期, *ratio* 为采样增长率, 设置为 0.008, 刚开始设置 p=0, 以真实标签作为输入, 随着训练周期增加, 模型训练得到足够的真实标签输入后便学会了上下文信息. 随着采样率 p 的不断增大, 最终会使训练模型与预测模型一样, 抹平两者之间差异.

2.3 评估指标

模型性能评估选择标准指标 *BLEU*[25]、*METEOR*[26]、*Rouge_L*[27] 和 *CIDEr*[28]. *BLEU* 用来判别模型生成的句子与实际标签句子之间的差异, 取值范围在 0.0 到 1.0 之间, 数值越大表示两个句子越相似. *BLEU* 指标是基于 *n*-gram (相邻单词片段) 的, *n* 的取值从 1 到 4, 对应 *BLEU*-1 到 *BLEU*-4, 每个 *n*-gram 的计算为式 (8):

$$P_n = \frac{\sum_i^E \sum_k^K \min(h_k(C_i), \min_{j \in M} h_k(s_{i,j}))}{\sum_i^E \sum_k^K \min(h_k(C_i))} \tag{8}$$

其中, $s_{i,j}$ 为第 i 个视频中有 j 条参考文本, C_i 为第 i 个视频生成的文本长度, $h_k(C_i)$ 和 $h_k(s_{i,j})$ 为第 k 个 *n*-gram 在生成译文和参考译文中出现的次数, M 为参考答案的个数, E 为生成文本的个数, min 表示第 k 个词组在 $s_{i,j}$ 中出现的最小次数, 即找到最相似的参考文本. 最终评估得分采用几何平均方式求平均并加权, 再乘以长度惩罚因子 *BP* 得到式 (9):

$$BLEU\text{-}N = BP \times \exp\left(\sum_{n=1}^{N} W_n \log(P_n)\right) \tag{9}$$

METEOR 是一种基于召回率和精确指标结合一起的评估方法, 与 *BLEU* 最大的不同在于 *METEOR* 能够利用单词字母组合与同义词之间的映射关系, 参考同义词来扩展词库, 计算结果对应模型生成文本与实际标签句子之间的准确率和召回率的调和平均值, 取值范围在 0.0 到 1.0 之间, 计算公式为式 (10):

$$METEOR = (1 - Penalty)F \tag{10}$$

其中, *Penalty* 为惩罚因子, F 为参考文本和模型生成文

本之间的准确率和召回率的调和平均值.

Rouge_L 是一种计算生成文本与参考文本之间最长公共子序列的评估方法, 计算结果对应最长子序列所占比例, 见式 (11):

$$Rouge\_L = \frac{(1+\beta^2)R_l P_l}{R_l + \beta^2 P_l} \tag{11}$$

其中, R_l与P_l为召回率与准确率, 分别对应最长子序列与参考文本长度和生成文本长度的比值.

CIDEr 是一种结合 *BLEU* 与向量空间的评估指标, 通过计算生成文本与参考文本之间的余弦相似度的平均值来评判模型性能, 见式 (12):

$$CIDEr = \frac{1}{m}\sum_j \frac{g^n(c)^{\mathrm{T}} g^n(s)}{\|g^n(c)\| \cdot \|g^n(s)\|} \tag{12}$$

其中, $g^n(c)$与$g^n(s)$分别代表 *n*-gram 在生成文本与参考文本之间的出现权重. 通过对文本中不同词组的出现权重比例, 衡量两个文本间的相似度.

3 实验分析

3.1 消融实验对比

本文方法主要在 MSVD 与 MSR-VTT 数据集上进行验证评估, 分数单位为 %, 分数越高代表模型生成语句与标注语句越接近, 模型越好. 基准模型为文献 [12]. 由于λ为超参数影响外部知识的增强程度, 为选出合理的数值, 通过消融实验来选取理想模型. 表 3 和表 4 为在 MSR-VTT 与 MSVD 数据集上不同λ值控制外部补充知识的程度对模型性能评估的消融实验, 由 0 到 0.9 依次直观地增加 0.1. 设模型 2D 卷积+3D 卷积+LSTM 为 BM, 特征强化为 F, 语义信息为 S, 知识引入程度λ, 对比加入特征强化与知识补充的结果来证明改进模型的有效性.

表 3 基于 MSR-VTT 数据集上模型的消融实验 (%)

模型名称	*BLEU*-4	*METEOR*	*Rouge_L*	*CIDEr*	模型名称	*BLEU*-4	*METEOR*	*Rouge_L*	*CIDEr*
BM	38.6	26.1	60.4	47.0	BM+F+S+λ(0.4)	44.3	29.0	62.0	51.7
BM+F	40.2	26.8	60.9	48.3	BM+F+S+λ(0.5)	43.7	28.8	61.6	50.5
BM+S	43.8	28.9	62.1	51.6	BM+F+S+λ(0.6)	43.2	28.3	61.5	49.3
BM+F+S	44.6	29.1	62.7	52.1	BM+F+S+λ(0.7)	42.9	27.9	61.2	49.2
BM+F+S+λ(0.1)	**44.8**	29.1	62.9	**52.4**	BM+F+S+λ(0.8)	39.7	26.6	60.7	48.6
BM+F+S+λ(0.2)	**44.8**	**29.2**	**63.1**	52.3	BM+F+S+λ(0.9)	37.9	25.7	59.4	47.5
BM+F+S+λ(0.3)	44.4	29.1	62.8	51.8	Baseline	43.8	28.9	62.4	51.4

表 4 基于 MSVD 数据集上模型的消融实验 (%)

模型名称	*BLEU*-4	*METEOR*	*Rouge_L*	*CIDEr*	模型名称	*BLEU*-4	*METEOR*	*Rouge_L*	*CIDEr*
BM	53.8	35.1	74.9	90.6	BM+F+S+λ(0.4)	61.9	37.9	78.1	104.5
BM+F	57.6	37.0	75.7	96.2	BM+F+S+λ(0.5)	60.7	37.3	77.6	101.7
BM+S	61.8	37.8	76.9	103.3	BM+F+S+λ(0.6)	57.6	36.7	75.3	95.9
BM+F+S	63.1	38.6	77.6	108.4	BM+F+S+λ(0.7)	53.7	35.5	74.7	92.7
BM+F+S+λ(0.1)	63.3	38.8	78.4	109.0	BM+F+S+λ(0.8)	48.6	34.8	71.9	88.2
BM+F+S+λ(0.2)	**63.4**	**38.9**	**78.6**	109.1	BM+F+S+λ(0.9)	44.2	32.9	71.2	79.3
BM+F+S+λ(0.3)	62.9	38.4	78.0	**109.3**	Baseline	61.8	37.8	76.8	103.0

表 3 为在 MSR-VTT 数据集上不同 λ 值的消融试验, 模型引入特征强化模块之后, 4 种评估指标相较于基准模型平均提升了 1.3%. 在引入知识补充模块 λ=0.1 时, *BLEU*-4 取得最大值, λ=0.2 时, *METEOR* 和 *Rouge_L* 取得最大值, 此时 4 种评估指标相较于基准模型平均提升了 1.8%. 表 4 为在 MSVD 数据集上不同λ值的消融试验, 模型引入特征强化模块之后, 4 种评估指标相较于基准模型平均提升了 2.6%. 当引入知识补充模块 λ=0.2 时, 除 *CIDEr* 外, 其他评估指标均取得最大值, 此时 4 种评估指标相较于基准模型平均提升了 3.4%. 随着 λ 的增大, 4 项指标得分均大幅度下跌, 当 λ=0.9 时, 模型评估得分很低, 表现出较差的性能. 由此分析得知, 当 λ 偏小时, 模型能根据视觉目标从语库中选择跟目标相关的语义来辅助模型生产文本. 当 λ 增大导致引入知识的概率过大, 难免会造成某些预测单词出现的概率过大, 从而造成生成的文本与原标注语句之间差异增大, 导致模型的稳定性降低, 尽管能够生产一些新颖的词, 但评估指标会有所降低. 因此引入外部知识的程度是有限的, 为了使模型能够生成新颖有意义的语句, 且不会降低模型的质量, 取得最大效益, 综合选取 BM+F+S+λ(0.2) 时的模型作为本文理想模型.

3.2 与其他算法对比

表 5 为基础模型引入特征强化模块与知识补充超参数设为 0.2 的理想模型与近年来其他方法在 MSVD 与 MSR-VTT 两个数据集上的对比结果. 表中各种方法的评估结果以文献中数据为准, 部分评估没有数据的以“—”代替.

其中 MARN[10] 在编解码框架中加入了记忆体快提高描述质量, SCN[11] 提出一个语义检测网络检测视频的语义信息, SBAT[13] 利用双视觉 Transformer 形成编解码框架生成视频描述, MMI[14] 通过利用视频中的场景, 静态动态特征, 图像描述形成的多模态辅助模型生产描述, MABVC[15] 通过融合视觉与音频传入到 Transformer 形成的多模态框架生成描述, ORG-TRL[16] 提出一种新的训练方式来解决人工标注中长尾分布的问题, ECO[19] 通过高效卷积神经网络捕获视频的动态特征生成描述, 本文在基准模型 SAM-SS[12] 上通过静态视觉的特征强化以及外部知识补充, 使模型对物体的细粒度特征提取更加全面, 并且在得到外部知识补充后, 既能辅助模型生成与视觉目标相关的文本, 也在一定程度上帮助模型发现视频中的隐含信息. 由表 5 实验结果可知, 本文方法在两个数据集上的表现相较于基础模型有了提升, 在 4 种评价指标上均取得了良好的效果, 并优于其他方法, 证明了模型的有效性.

图 7 和图 8 为本文加入特征强化 (FEM) 与知识补充 (λ) 的模型生成语句与人工标注的语句在 MSVD 与 MSR-VTT 数据集上的对比. 其中 GT 为人工标注的 3 个语句, Caption 为模型生成语句, Baseline 为基准模型, Ours-1 为加入特征强化后的模型, Ours-2 为加入特征强化与知识补充 (λ=0.2) 的理想模型, Ours-3 为加入特征强化与知识补充 (λ=0.6) 的模型. 图 7(a) 与图 8(a) 可见, Ours-1 相较于基准模型对“boy”与“girl”特征的有更高的辨识度, 并最终生成于文本中. Ours-3 在图 7 与图 (8) 中由于分别检测出“people”“cat”“toy”与“dog”, 过大补充了部分的常识性知识概率, 使生成文本过于偏离标注文本, 因此知识补充的程度是有限的. 图 7(b) 与图 8(b) 可见, Ours-2 为本文理想模型, 当检测到视觉目标“cat”或“dog”后, 通过知识补充, 分别增添了“walk”与“lying”常识性动作知识的概率, 并最终生成与文本中. 其中 Ours-2 在图 8(b) 中相较于基础模型, 发现了“lying”隐含信息, 使得描述更加符合视频中“dog”的行为. 由此可见在模型经过特征强化与定量知识补充后, 模型生成的描述语句能够区分更加细粒度的相似物体, 也可能得到一些隐含信息, 使描述更加符合人的感知, 真实地体现出方法的优势性.

表 5 本文模型与其他模型在 MSVD 和 MSR-VTT 数据集上对比 (%)

模型名称	MSVD				MSR-VTT			
	BLEU-4	*METEOR*	*Rouge_L*	*CIDEr*	*BLEU*-4	*METEOR*	*Rouge_L*	*CIDEr*
MARN[10]	48.6	35.1	71.9	92.2	40.4	28.1	28.1	47.1
SCN[11]	51.1	33.5	—	77.7	—	—	—	—
SBAT[13]	53.2	35.3	72.3	86.7	42.9	28.9	61.0	48.5
MMI[14]	46.7	33.6	65.0	76.8	39.3	28.5	61.2	44.6
MABVC[15]	54.6	36.7	73.6	90.4	43.8	29.6	61.5	52.5
ORG-TRL[16]	54.3	36.4	73.9	95.2	43.6	28.8	62.1	50.9
ECO[18]	53.5	35.0	—	85.8	—	—	—	—
SAM-SS (Baseline)[12]	61.8	37.8	76.8	103.0	43.8	28.9	62.4	51.4
Ours	**63.4**	**38.4**	**78.6**	**109.1**	**44.8**	**29.4**	**63.0**	**52.3**

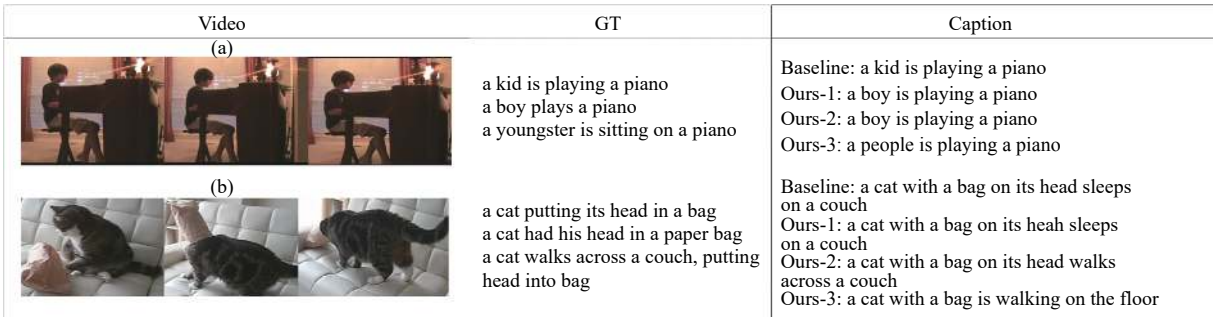


图 7 MSVD 数据集上模型描述对比

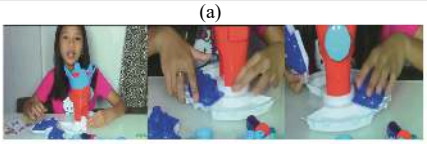
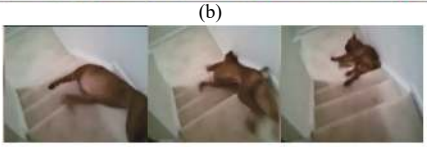
Video	GT	Caption
(a) 	a girl is playing with a toy a girl is arranging a doll a little girl playing with toys	Baseline: a woman is playing with toys Ours-1: a girl is playing with a toy Ours-2: a girl is playing with a toy Ours-3: a child is playing with a toy
(b) 	the dog slid down the stairs a dog is getting down the stairs dogs are crawling down stairs	Baseline: a dog is getting down the stirs Ours-1: a dog is getting down the stairs Ours-2: a dog is lying down the stairs Ours-3: a dog is lying on the floor

图 8 MSR-VTT 数据集上模型描述对比

4 结论与展望

本文探索了如何提升对静态视觉特征的提取能力和如何集成更多的知识信息来辅助模型生产更加新颖和符合人类感知的描述语句, 提出了一种基于特征强化与知识补充的视频描述方法. 在 MSVD 数据集与 MSR-VTT 数据集上的实验证明, 本文模型无论在评估指标还是直观的描述上都取得了较好的结果, 在提升评估指标的同时一定程度上能挖掘视频中的部分新颖内容. 未来考虑融入注意力机制与多模态输入使模型能获得更多有效信息来指导模型生成更加准确和更有质量的语句.

参考文献

1 Kojima A, Izumi M, Tamura T, *et al*. Generating natural language description of human behavior from video images. Proceedings of the 15th International Conference on Pattern Recognition. Barcelona: IEEE, 2000. 728–731.

2 Zhao B, Li XL, Lu XQ. CAM-RNN: Co-attention model based RNN for video captioning. IEEE Transactions on Image Processing, 2019, 28(11): 5552–5565. [doi: 10.1109/TIP.2019.2916757]

3 Liu CJ, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Transactions on Image Processing, 2002, 11(4): 467–476. [doi: 10.1109/TIP.2002.999679]

4 Song JK, Yang Y, Yang Y, *et al*. Inter-media hashing for large-scale retrieval from heterogeneous data sources. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013. 785–796.

5 Krishnamoorthy N, Malkarnenkar G, Mooney R, *et al*. Generating natural-language video descriptions using text-mined knowledge. Proceedings of the 27th AAAI Conference on Artificial Intelligence. Bellevue: AAAI Press, 2013. 541–547.

6 Odonez V, Kulkarni G, Berg TL. Im2Text: Describing images using 1 million captioned photographs. Proceedings of the 24th International Conference on Neural Information Processing Systems. Granada: Curran Associates Inc., 2011. 1143–1151.

7 Donahue J, Hendricks LA, Rohrbach M, *et al*. Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677–691. [doi: 10.1109/TPAMI.2016.2599174]

8 Venugopalan S, Rohrbach M, Donahue J, *et al*. Sequence to sequence-video to text. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 4534–4542.

9 Yao L, Torabi A, Cho K, *et al*. Describing videos by exploiting temporal structure. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 4507–4515.

10 Pei WJ, Zhang JY, Wang XR, *et al*. Memory-attended recurrent network for video captioning. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8347–8356.

11 Gan Z, Gan C, He XD, *et al*. Semantic compositional networks for visual captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1141–1150.

12 Chen HR, Lin K, Maye A, *et al*. A semantics-assisted video captioning model trained with scheduled sampling. Frontiers in Robotics and AI, 2020, 7: 475767. [doi: 10.3389/frobt.2020.475767]

13 Chen M, Li YM, Zhang ZF, *et al*. TVT: Two-view transformer network for video captioning. Proceedings of the 10th Asian Conference on Machine Learning. Beijing:

PMLR, 2018. 847–862.
14 丁恩杰, 刘忠育, 刘亚峰, 等. 基于多维度和多模态信息的视频描述方法. 通信学报, 2020, 41(2): 36–43. [doi: 10.11959/j.issn.1000-436x.2020037]
15 李铭兴, 徐成, 李学伟, 等. 基于多模态融合的城市道路场景视频描述模型研究. 计算机应用研究, 2022. [doi: 10.19734/j.issn.1001-3695.2022.06.0275]
16 Zhang ZQ, Shi YY, Yuan CF, *et al*. Object relational graph with teacher-recommended learning for video captioning. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13275–13285.
17 Szegedy C, Ioffe S, Vanhoucke V, *et al*. Inception-v4, Inception-ResNet and the impact of residual connections on learning. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017. 4278–4284.
18 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255.
19 Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 713–730.
20 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4724–4733.
21 Speer R, Chin J, Havasi C. ConceptNet 5.5: An open multilingual graph of general knowledge. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017. 4444–4451.
22 Girshick R. Fast R-CNN. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.
23 Chen DL, Dolan W B. Collecting highly parallel data for paraphrase evaluation. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland: ACL, 2011. 190–200.
24 Xu J, Mei T, Yao T, *et al*. MSR-VTT: A large video description dataset for bridging video and language. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 5288–5296.
25 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia: ACL, 2002. 311–318.
26 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor: ACL, 2005. 65–72.
27 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the 2004 Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.
28 Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575.

(校对责编: 孙君艳)