DOI:10.19651/j.cnki.emt.2107570

基于扩张卷积的注意力机制视频描述模型*

王金金 曾上游 李文惠 张介滨
(广西师范大学 电子工程学院 桂林 541004)

摘 要: 针对视频描述过程中视觉特征和词特征关联度不足、训练效率低、生成的自然语言出现错误和指标分数不高的问题,提出了一种基于扩张卷积的注意力机制视频描述模型。在模型的编码阶段,采用 Inception-v4 对视频特征进行编码,然后将编码后的视觉特征和词特征输入到基于扩张卷积的注意力机制中,最后通过长短期记忆网络进行解码,生成视频的自然描述语句。在视频描述公共数据集 MSVD 上进行对比实验,通过评价指标(BLEU、ROUGE_L、CIDEr、METEOR)对模型进行验证,实验结果表明,基于扩张卷积的注意力机制视频描述模型在各个指标上都有明显提升,对比基线模型 SA-LSTM (Inception-v4),在 BLEU_4、ROUGE_L、CIDEr 和 METEOR 指标下分别提升了 4.23%、4.73%、2.11%和 2.45%。

关键词: 视频描述;Inception-v4;长短期记忆网络;扩张卷积;注意力机制
中图分类号: TP391.4;TP183 **文献标识码:** A **国家标准学科分类代码:** 520.20

Video description model of attention mechanism based on dilated convolution

Wang Jinjin Zeng Shangyou Li Wenhui Zhang Jiebin
(School of Electronic Engineering, Guangxi Normal University, Guilin 541004, China)

Abstract: In order to solve the problems of insufficient correlation between visual features and word features, low training efficiency, errors in generated natural language and low index scores in the process of video description, a video description model based on the attention mechanism of dilated convolution is proposed. In the encoding stage of the model, Inception-v4 is used to encode the video features, and then the encoded visual features and word features are input into the attention mechanism based on dilated convolution. Finally, the video is decoded through the long short-term memory network to generate the natural description statement of the video. A comparative experiment was conducted on the public video description data set MSVD, and the model was verified by evaluation indicators (BLEU, ROUGE_L, CIDEr, METEOR). The experimental results showed that the video description model based on the attention mechanism of dilated convolution has significantly improved in all indicators. Compared with the baseline model SA-LSTM (Inception-v4), the BLEU_4, ROUGE_L, CIDEr and METEOR indicators have increased by 4.23%, 4.73%, 2.11% and 2.45% respectively.

Keywords: video description;Inception-v4;long short-term memory network;dilated convolution;attention mechanism

0 引 言

当今移动互联网、大数据时代,个人智能终端数量的激增与各种短视频平台的快速发展,使得视频呈现爆发式增长,完全通过人力对数据进行标注和描述已成为一项不可能完成的任务,因此视频描述任务再次引起研究的高潮。视频描述即是用自然语言自动描述视频,其又称为视频字幕,是将视觉和自然语言连接起来的重要手段,也是计算机视觉中一个极具挑战性的问题。视频描述具有大量的实际应用,它可以帮助用户高效检索视频,也有利于视障者更好地理解视频内容,在智能安防、人机交互等应用中具有巨大的发展前景。然而,由于视觉内容与自然语言之间的语义鸿沟仍然是一个难题,如何弥合它们之间的语义鸿沟是视频描述技术的主要研究重点。

现主流的视频描述算法主要是基于编码器-解码器结构。编码器利用卷积神经网络 (convolutional neural

收稿日期:2021-08-13
* 基金项目:国家自然科学基金(61976063)项目资助

network，CNN)对视频的信息进行特征编码，解码器使用循环神经网络(recurrent neural network，RNN)根据其编码特征生成视频的自然描述语句[1-2]。2015 年 Venugopalan 等[3]提出了 S2VT，通过预先训练的 VGG[4]网络提取输入视频中每一帧图像的特征向量，所有帧或其采样帧的特征通过简单的平均池进行处理，得到整个视频片段的矢量表示，再将特征矢量输入到两层的长短期记忆网络(long short term memory，LSTM)[5]进行编解码，该模型在编码和解码器中均使用了 LSTM。然而，利用该模型提取的视觉特征比较简单，对视频内容生成的自然语言不够丰富。正如好的图像描述[6]通常对图像更显著的部分进行关注，一个好的视频描述模型也应该有选择地集中在视频序列更加显著的特征上，因此注意力机制被应用到视频描述中。近年来，随着注意力机制在目标检测和机器翻译等领域的广泛应用，Li 等[7]提出结合注意力机制的递归编码器模型，注意力机制为每一帧视频图像的特征分配权重，然后基于注意权重进行融合。但该注意力机制较为简单，不能很好的将视频中的视觉特征和词特征联系起来。

本文提出一种端到端的、基于扩张卷积的注意力机制视频描述方法，引入了一种新的基于扩张卷积的注意力模块，能在生成单词时更好地关注更加重要的部分视频帧，同时为了训练时能更好更快的收敛，采用 AMSGrad[8]优化器。在微软视频描述(microsoft video description corpus，MSVD)[9]公共数据集上对本文的方法进行验证，采用多种评价指标对生成的自然语言进行评价，其实验结果证明了本文所提方法在视频描述问题上的有效性和可靠性。

1　相关研究

视频描述主要分为 2 类。第 1 类是用自然语言描述一段视频的主要内容，该类的输入一般是一个短视频，输出则是一句或多句自然描述语言。第 2 类则是对视频内容的密集描述，通常情况下需要将视频中的人物、物体、场景、动作及其相互关系和变化的过程描述清楚。本文的方法主要解决视频描述的第 1 类问题。

1.1　传统的视频描述方法

传统的视频描述方法是基于模板的方法[10]和基于检索的方法[11]。基于模板的方法是预先定义句子的语言和语法规则，将句子分为主语、谓语和宾语等几个部分，然后利用视觉处理技术处理视频片段，将从视频数据里检测到的单词通过预先定义的模板进行填充，产生具有预定义模板的最终描述语句。由此过程可以清楚看到基于模板的视频描述方法的优缺点十分明显，优点是由于模板的存在，生成的描述性语句语法较为准确，但缺点也是因为固定模板的存在，使其描述性语句受到限制，让句子的生成变得极为不灵活且内容不丰富。基于检索策略的描述方法主要是在数据库中搜索与输入的视频数据相似的样本，然后将该样本的描述性语句作为输入视频的结果输出，这种描述方法虽然操作起来较为简单，但其结果受到样本库的很大限制。

1.2　基于深度学习的视频描述方法

近年来随着深度学习的兴起，基于深度学习的视频描述方法成为主流。如今的视频描述算法大多是基于编码器-解码器框架，通常是利用 CNN 对输入的视频进行特征提取和编码，利用 RNN 或其变体作为解码器生成最终的自然描述语句[12]。近几年里，研究人员发现，视觉注意力机制也有助于机器理解视觉内容。Xu 等[13]首先将注意力机制应用于图像描述中，即动态选择图像中特定的部分，比如主要对象或感兴趣部分，最后进行加权和融合。注意力机制使模型将计算资源更好地分配给视觉内容中更加重要的部分，实际上，注意力机制就像人类的行为，指的是一个人在某个时刻只关注其想要的部分，而忽略其他部分，在视频描述中，注意力机制也得到了广泛的应用。Yi 等[14]采用深度软注意机制网络，整合前向和后向传递的信息，并用双向循环的结构对时间序列进行全局分析和局部探索。Guo 等[15]提出了一种基于注意力机制的多模态特征融合方法，用在融合视听特征的两帧图像上生成最终字幕。

2　视频描述模型

2.1　模型整体结构

视频描述的目的是生成一个句子 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n\}$ 来描述输入视频 $\boldsymbol{V}$ 的内容，经典的编码器-解码器结构直接逐字模拟字幕生成概率如式(1)所示。

$$P(\boldsymbol{S}\mid\boldsymbol{V})=\prod_{i=1}^{n}P(\boldsymbol{s}_i\mid\boldsymbol{s}_{<i},\boldsymbol{V};\theta) \tag{1}$$

式中：θ 是视频描述模型的参数，n 表示句子的长度，$\boldsymbol{s}_{<i}$ 表示生成的部分描述。本文提出的视频描述整体框架如图 1 所示。

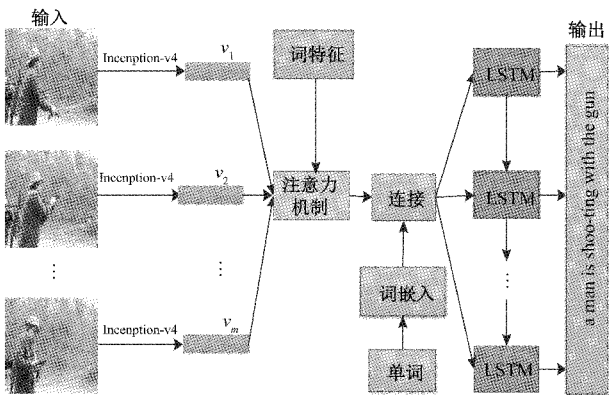

图 1　视频描述模型整体框架

首先对输入的视频进行预处理，利用编码器对视频进行编码，将给定的视频序列编码为序列表示 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m\}$，其中 m 表示提取的视频帧的总数。将视频编码成具有高级语义信息的固定长度后将其和编码器的隐藏输出，通过本文提出的基于扩张卷积的注意力机制进行关联，再与词向量拼接，最后输入到解码器 LSTM 中生成更加准确

的视频描述语句。

本文提出方法的关键在于基于扩张卷积的注意力机制的引入,它能更好地把握不同时刻不同帧的重要性,通过嵌入该注意力机制到编码器-解码器的框架中,使得模型在生成当前时刻单词时能够动态关注更相关的视频的帧子集,从而生成更加准确的视频描述语句。

2.2 视频帧特征提取模型

视频中存在大量的物体、人物及动作等多种信息,为了生成可靠的视频描述语句,需要更好地提取视频的视觉特征来获取视频的高层语义信息,以往的方法通常依赖于卷积神经网络 CNNs,如 VGG 和 AlexNet[16] 等。在本文中,考虑到更深层次的网络更适合于特征提取,采用了 ImageNet[17] 分类任务数据集上预训练的 Inception-v4[18] 作为编码器来提取视频帧的高级视觉特征。Inception 架构能以相对较低的计算成本实现良好的性能,Inception-v4 是通过更深入和更广泛来提高效率,它比 Inception-v3 具有更统一的简化体系和更多的 Inception 模块,在增加网络深度和宽度的同时,也增加了对图像尺度大小的适应性。

视频先经过预处理,本文对每个视频提取 60 帧的有效帧,尺寸设置为 299×299,f_{CNN} 表示对视频的特征提取函数,当输入视频的序列帧 $\boldsymbol{I}=(\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_m)$,其中 m 为提取的有效帧数,每帧图片的通道为 3,高和宽均为 299。对每一帧有效帧提取视觉特征如式(2)所示。

$$\boldsymbol{V}=f_{CNN}(\boldsymbol{I}) \tag{2}$$

可以得到视频序列编码后的特征序列表示 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m\}$。

2.3 基于扩张卷积的改进注意力模块

注意力机制是从模拟生物学的角度出发,使得神经网络能专注于其输入或特征子集的能力,考虑到某些视频片段对视频字幕的重要性,本文提出了一种新的注意力机制。该注意力机制是在 SA-LSTM (Inception-v4)[19] 模型中的注意力基础下进行改进,SA-LSTM (Inception-v4)中的注意力机制如图 2 所示。该注意力机制主要由全连接层组成,首先将经过卷积神经网络编码后的视频帧特征和解码器的隐藏层输出即词特征分别经过全连接层后相加,再将得到的特征输入到一个全连接层中,经过 Softmax 函数得到视频帧和词的关联度,最后与输入的视频帧特征相乘,获得词和视频帧的注意力图。

为了模型在生成单词时能更好地关注更相关的视频帧子集,本文改进了注意力机制,并进一步在改进注意力模块上引入扩张卷积,提出了基于扩张卷积的注意力机制,结构如图 3 所示。未引入扩张卷积时的改进注意力机制是将图 3 结构中的扩张卷积替换成相同卷积核的普通卷积。

视频帧特征和词特征分别经过卷积层进行特征处理和整合,并通过 ReLU 函数引入非线性,在将其分为两组,一组经过扩张卷积并使用 ReLU 激活函数引入非线性,对整合的帧和词进行局部特征筛选,其全连接层则对筛选后的

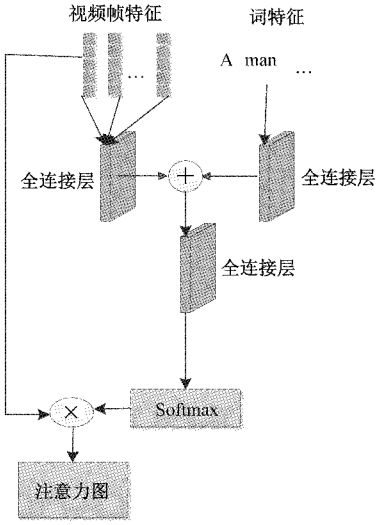

图 2 SA-LSTM 中的注意力机制

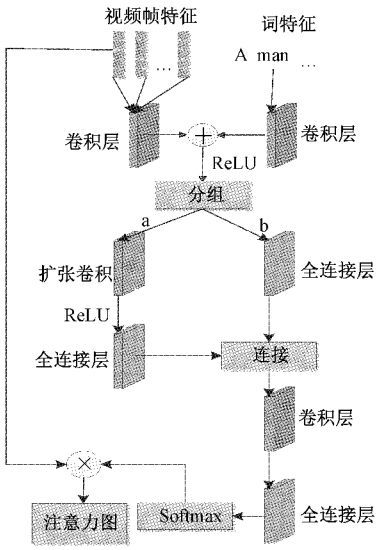

图 3 基于扩张卷积的注意力机制

特征进行整合,另一组中全连接层是只对特征分配全局的权重。融合两组的局部和全局特征,通过卷积层再次将局部和全局特征重新筛选出有效的信息,再经过全连接层整合有效特征,最后由 Softmax 函数得到新注意力分布,并将其与输入的视频帧特征相乘,得到词对应帧的注意力图。

基于扩张卷积的注意力机制中的卷积层均是 1×1 的卷积层,扩张卷积(dilated convolution),又称为空洞或膨胀卷积。它是在标准的卷积核中注入空洞,以此来增加感受野,同时保持特征图的尺寸不发生变化。二维扩张卷积的示意图如图 4 所示,本文采用的扩张卷积的扩张系数为 2,卷积核大小是 3×3。

2.4 视频描述语言生成模型

解码器的主要目的是逐个生成单词,由这些单词组成的句子描述视频的内容。本文视频描述语言生成模型采用 LSTM,LSTM 通过门控机制弥补了传统 RNN 信息丢失的

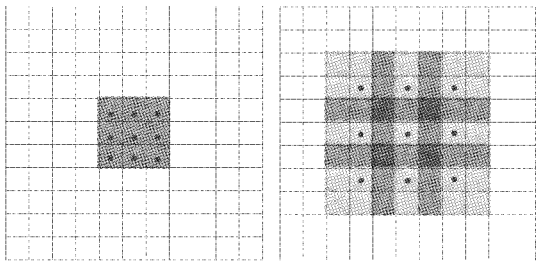

图 4　不同扩张系数的扩张卷积示意图

问题，其核心在它的细胞单元结构，如图 5 所示。主要是由遗忘门(forget gate)、输入门(input gate)、输出门(output gate)和记忆单元(memory cell)等组成。

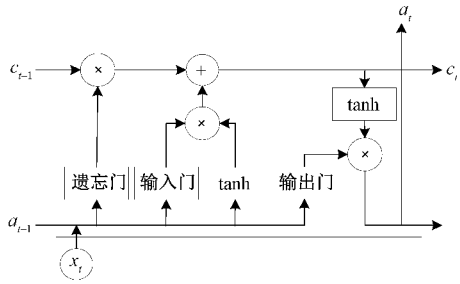

图 5　LSTM 细胞单元结构

假设在 t 时刻，输入的特征向量为 $\boldsymbol{x}_t$，对应的隐藏层特征为 $\boldsymbol{a}_{t-1}$，记忆单元特征为 $\boldsymbol{c}_t$，则 t 时刻 LSTM 单元的前进计算式为：

$$i_t = \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{ai}\boldsymbol{a}_{t-1} + \boldsymbol{b}_i) \tag{3}$$

$$f_t = \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{af}\boldsymbol{a}_{t-1} + \boldsymbol{b}_f) \tag{4}$$

$$o_t = \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ao}\boldsymbol{a}_{t-1} + \boldsymbol{b}_o) \tag{5}$$

$$g_t = \text{Ø}(\boldsymbol{W}_{xg}\boldsymbol{x}_t + \boldsymbol{W}_{ag}\boldsymbol{a}_{t-1} + \boldsymbol{b}_g) \tag{6}$$

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

$$a_t = o_t \odot \text{Ø}(c_t) \tag{8}$$

其中，i_t、f_t、o_t 分别表示输入门、遗忘门和输出门，$\boldsymbol{W}_{xj}$、$\boldsymbol{b}_j$ 均是可训练的权重矩阵和偏置向量，σ 表示 sigmoid 函数，Ø 表示 tanh 函数，$\odot$ 是哈达玛积运算，即向量的点乘。

2.5　AMSGrad 优化器

Adam 是在深度学习中用来替代随机梯度下降的优化算法，结合了 AdaGrad 和 RMSProp 中最优的部分，且能解决稀疏梯度和噪声问题，调参也较为简单。但 Adam 优化器在一些情况下不收敛且错失最优解，本文模型在训练时采用了 AMSGrad 优化器，该优化器保留了 Adam 的优势，同时通过添加额外的约束，达到更好的效果，其梯度更新公式为：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{9}$$

$$V_t = \beta_2 V_{t-1} + (1-\beta_2) g_t^2 \tag{10}$$

$$\hat{V}_t = \max(\hat{V}_{t-1}, V_t) \tag{11}$$

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{\hat{V}} + \varepsilon} \tag{12}$$

其中，m_t 为梯度的一阶矩估计；g_t 表示时间步为 t 时的梯度；β_1、β_2 为一阶矩和二阶矩的指数衰减率；V_t 表示对梯度的二阶矩估计；θ 为需要更新的参数；η、ε 分别为学习率和为了稳定而添加的常数。

3　实验结果与分析

3.1　实验数据集

本文选取了视频描述生成中常用的公共数据集 MSVD 来验证模型。MSVD 数据集收集了 1 970 个来自 YouTube 网站的短视频，每一个视频均在 10～25 s 内描绘一个单一的活动，通过土耳其机器人得到多种语言的视频描述，本文只使用英文描述。每个视频片段大约有 40 个英文描述，平均每个句子有 7 个单词。将数据集中划分成训练集、验证集和测试集，其中训练集有 1 200 个短视频，验证集有 100 个短视频，测试集有 670 个短视频。

3.2　评价指标

在评价指标中，本文采用微软官方提供的广泛使用的 4 种文本质量评价指标，即 METEOR[20]、BLEU(n)[21]、ROUGE-L[22]和 CIDEr[23]。METEOR 指标是基于词库的单精度加权调和平均数和单字召回率的标准，与人类的评价标准更加相关；BLEU 指标评估生成的句子和参考句子之间的 n-gram 相关性；ROUGE-L 指标测量生成的句子和参考句子之间最长公共子序列的 F-Score；CIDEr 指标通过计算每个 n-gram 的 TF-IDF (term frequency-inverse document frequency)权重来测量描述注释的一致性。

3.3　实验训练环境及参数设置

实验环境为 64 位 Windows10 操作系统下安装 pytorch 作为深度学习框架，计算机内存为 32 GB RAM、Intel i7-6700K 四核八线程 CPU 以及 NVIDIA-GTX1080Ti GPU。

本文实验中词向量和解码器的输入维度均为 468，隐藏层单元为 512，训练模型时 batch size 为 32，模型生成句子的有效长度最大设为 30，最小为 1，损失函数采用交叉熵损失函数。训练时采用 AMSGrad 优化器对模型的参数进行优化，epoch 设定为 50，当验证数据集上的 CIDEr 指标在 20 个连续 epoch 内停止增加时，训练停止。在测试中，使用大小为 5 的集束搜索来生成最终的字幕。

3.4　实验与分析

在 MSVD 数据集中，本文针对基础模型即 Inception-v4+LSTM，引入本文不包含扩张卷积的改进注意力机制以及进一步提出的基于扩张卷积的注意力机制进行对比实验，验证本文提出的两种改进注意力机制的有效性，结果如表 1 所示。

由表 1 可知，在基础模型 Inception-v4+LSTM 中引入本文改进的注意力机制后，各个指标均有很大地提升，BLEU_1、BLEU_2、BLEU_3 和 BLEU_4 指标分别提升了

表 1 不同注意力机制在 MSVD 上的评价指标对比 %

模型方法	BLEU_1	BLEU_2	BLEU_3	BLEU_4	CIDEr
基础框架	77.67	65.02	55.81	46.38	70.29
改进注意力	79.61	67.56	57.88	47.54	76.35
基于扩张卷积	80.23	68.53	59.34	49.53	78.31

1.94%、2.54%、2.07% 和 1.16%，CIDEr 指标提升了 6.06%，由此可见改进注意力机制的有效性；当引入本文进一步改进的基于扩张卷积的注意力机制后，各项指标在此基础上再次得到了明显的提升，验证了本文提出的基于扩张卷积的注意力机制的良好性能。

本文模型与近几年的视频描述模型在 MSVD 数据集的对比实验结果如表 2 所示，可知与基线模型 SA-LSTM (Inception-v4) 相比，本文模型在 BLEU_4、ROUGE_L、CIDEr 和 METEOR 指标下分别提升了 4.23%、4.73%、2.11% 和 2.45%，表明了本文基于扩张卷积的注意力机制视频描述模型的优越性能。同时相比其他模型也性能优良，证明了本文提出的基于扩张卷积的注意力机制是有效且可行的，能很好地应用于视频描述领域。在 MSVD 数据集中的测试集里选取部分视频进行测试，并将其视频可视化，结果如图 6 所示。

表 2 本文模型和其他模型在 MSVD 上的表现对比 %

模型	BLEU_4	ROUGE_L	CIDEr	METEOR
TDDF[24]	37.3	59.2	43.8	27.8
PickNet[25]	46.1	62.9	76.0	33.1
SA-LSTM (Inception-v4)[19]	45.3	64.2	76.2	31.9
GRU-EVE[26]	47.9	71.5	78.1	35.0
MMI[27]	46.7	65.0	76.8	33.6
本文模型	**49.53**	68.93	**78.31**	34.35

本文：a squirrel is dancing on a grass field

GT1：A squirrel is dancing

GT2：A squirrel is singing

GT3：The squirrel danced on the lawn

(a) 示例1

本文：a woman is cutting a piece of meat

GT1：a woman is putting meat on a tray

GT2：someone is breading meat

GT3：someone is breading meat

(b) 示例2

图 6 在 MSVD 数据集上的视频描述示例

图 6 中的 GT(ground truth)为该视频的真实标注，每个视频选取 3 个真实标注，由图 6 可知，与真实标注相比，本文模型能更准确地生成视频的自然语言描述，且语言更加丰富，证明了基于扩张卷积的注意力机制的引入使得模型具有更好的性能。

4 结 论

本文提出了一种基于扩张卷积的注意力机制视频描述模型。该模型使用了深度卷积神经网络 Inception-v4 来提取视频的帧图像特征，采用本文的注意力机制结构加强了模型中视频帧特征和词特征的相关性，然后利用 LSTM 进行解码生成视频描述的自然语言语句。经过在 MSVD 数据集上的实验证明，本文引入基于扩张卷积的注意力机制的视频描述生成模型性能优越。

在以后的研究中，在本文工作的基础上，可进一步改良注意力机制，使模型能更加关注有用信息，并探索 3D 卷积，让模型在编码阶段能更好地提取视频帧特征，进一步提高视频描述模型生成自然语言的质量。

参考文献

[1] ZHAO B, LI X, LU X. CAM-RNN: Co-attention model based RNN for video captioning[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5552-5564.

[2] WANG B, MA L, ZHANG W, et al. Reconstruction network for video captioning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7622-7631.

[3] VENUGOPALAN S, ROHRACH M, DONAHUE J, et al. Sequence to sequence-video to text[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, Santiago: IEEE, 2015:4534-4542.

[4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. CVPR 2014: 2014 Computer Vision and Pattern Recognition, Columbus: IEEE Computer Society, 2014.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] 黄友文，游亚东，赵朋. 融合卷积注意力机制的图像描述生成模型[J]. 计算机应用，2020，40(1)：23-27.

[7] LI Y, TORABI A, CHO K, et al. Describing videos by exploiting temporal structure[C]. Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2015: 4507-4515.

[8] SASHANK J R, SATYEN K, SANJIV K. On the convergence of Adam and beyond[C]. Proceedings of the 6th International Conference on Learning

Representations, ICLR, 2018.
[9] CHEN D L, DOLAN W B. Collecting highly parallel data for paraphrase evaluation[C]. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Stroudsburg, ACL, 2011: 190-200.
[10] ROHRBACH M, QIU W, TITOV I, et al. Translating video content to natural language descriptions[C]. Proceedings of the IEEE International Conference on Computer Vision, IEEE, 2013: 433-440.
[11] YU H N, SISKIND J M. Learning to describe video with weak supervision by exploiting negative sentential information[C]. Proceedings of the Association for the Advance of Artificial Intelligence, AAAI, 2015: 3855-3863.
[12] 曹磊，万旺根，候丽. 基于多特征的视频描述生成算法研究[J]. 电子测量技术，2020，43(16)：99-103.
[13] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]. Proceedings of the 2015 International Conference on Machine Learning, New York: ACM, 2015: 2048-2057.
[14] YI B, YANG Y, FUMIN S, et al. Describing video with attention-based bidirectional LSTM [J]. IEEE Transactions on Cybernetics, 2018, 49(7): 1-11.
[15] GUO N N, LIU H P, JIANG L H. Attention-based visual-audio fusion for video caption generation [C]. Proceedings of the IEEE International Conference on Advanced Robotics and Mechatronics, 2019: 839-844.
[16] ALEX K, ILYA S, GEOFFREY E H. ImageNet classification with deep convolutional neural networks[C]. Conference and Workshop on Neural Information Processing Systems, Lake Tahoe, NV, United States, 2012: 1097-1105.
[17] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[18] CHRISTIAN S, SERGEY I, VINCENT V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning [C]. Proceedings of the Thirty-First Conference on Artificial Intelligence, San Francisco, California, USA, 2017: 4278-4284.
[19] WANG B, MA L, ZHANG W. Reconstruction network for video captioning[C]. Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018: 7622-7631.
[20] BANERJEE S, LAVIE A. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments[C]. Proceedings of the 2005 Meeting of the Association for Computational Linguistics, Stroudsburg, PA: Association for Computational Linguistics, 2005: 65-72.
[21] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: A method for automatic evaluation of machine translation[C]. Proceedings of the 2002 Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 311-318.
[22] LIN C. ROUGE: A package for automatic evaluation of summaries[C]. Proceedings of the 2004 Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 74-81.
[23] VEDANTAM R, ZITNICK C L, PARIKH D, et al. CIDEr: Consensus-based image description evaluation[C]. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC: IEEE Computer Society, 2015: 4566-4575.
[24] ZHANG X, GAO K, ZHANG Y, et al. Task-driven dynamic fusion: Reducing ambiguity in video description[C]. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017: 3713-3721.
[25] CHEN Y, WANG S, ZHANG W, et al. Less is more: Picking informative frames for video captioning [C]. Proceedings of the European Conference on Computer Vision, ECCV, 2018: 358-373.
[26] AAFAQ N, AKHTAR N, LIU W, et al. Spatiotemporal dynamics and semantic attribute enriched visual encoding for video captioning[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12487-12496.
[27] 丁恩杰，刘忠育，刘亚峰，等. 基于多维度和多模态信息的视频描述方法[J]. 通信学报，2020，41(2)：36-43.

作者简介

王金金，硕士研究生，主要研究方向为自然语言处理、深度学习、图像和视频描述等。
E-mail：1516844863@qq.com

曾上游（通信作者），教授，博士，主要研究方向为神经网络与人工智能、复杂网络、生物信息处理和生物芯片等。
E-mail：zsy@mailbox.gxnu.edu.cn

李文惠，硕士研究生，主要研究方向为深度学习、人工智能等。
E-mail：935024086@qq.com

张介滨，硕士研究生，主要研究方向为深度学习、人工智能等。
E-mail：1140013224@qq.com