Computer Science and Application 计算机科学与应用, 2022, 12(10), 2225-2232
Published Online October 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1210226

Hans汉斯

基于多模态注意机制的全域视频描述生成技术研究

杜晓童

中国电子科技集团公司第十研究所，第四事业部，四川 成都

收稿日期：2022年9月2日；录用日期：2022年10月2日；发布日期：2022年10月9日

摘　要

基于多模态注意机制的深度神经网络模型，提出了一种针对全域视频的多语言描述自动生成技术。视频描述自动生成模型由端到端的卷积神经网络和双向循环神经网络组成，应用多模态注意机制，显著提升了模型的视频表征能力。通过构建双向循环神经网络编码器，对图像、光流、C3D以及音频等4种多模态视频特征进行融合编码，并引入基于注意机制的解码器，将编码获得的视频序列化特征最终解码为多语言描述序列。模型在开源视频描述数据集上进行了测试实验，实验结果表明了该方法的有效性，其中METEOR值提升了3.31%，为目前已公开的最佳结果。因此，该技术可作为相关领域研究的重要参考。

关键词

全域视频，卷积神经网络，双向循环神经网络，注意机制，多语言描述

Research of Multimodal Attention-Based Description Generation of Videos in Wide Domain

Xiaotong Du

The Fourth Division, The 10th Research Institute of China Electronics Technology Group Corporation, Chengdu Sichuan

Received: Sep. 2nd, 2022; accepted: Oct. 2nd, 2022; published: Oct. 9th, 2022

Abstract

Based on the deep neural network model of multimodal attention mechanism, this paper proposes

文章引用: 杜晓童. 基于多模态注意机制的全域视频描述生成技术研究[J]. 计算机科学与应用, 2022, 12(10): 2225-2232.
DOI: 10.12677/csa.2022.1210226

an automatic generation technology of multilingual description for global video. The automatic video description generation model is composed of an end-to-end convolutional neural network and a bidirectional cyclic neural network. The multi-modal attention mechanism is applied to significantly improve the video representation ability of the model. By constructing a bidirectional recurrent neural network encoder, four multimodal video features such as image, optical flow, C3d and audio are fused and encoded. And a decoder based on attention mechanism is introduced to decode the encoded video serialization features into a multilingual description sequence. The model has been tested on the open source video description dataset, and the experimental results show the effectiveness of the method, of which the meteor value has increased by 3.31%, which is the best result that has been published so far. Therefore, this technology can be used as an important reference for research in related fields.

Keywords

Wide Domain Videos, Convolutional Neural Networks, Bidirectional Recurrent Neural Networks, Attention Mechanism, Multilingual Description

1. 引言

作为对视频的一种高级语义理解，全域视频描述自动生成技术有着广泛的应用，例如为海量视频数据进行自动打标、分类管理，为视障人士提供描述视频服务(DVS)等。视频自然语言描述生成技术是人工智能领域中复杂度较高的任务，自 2012 年以来受到了计算机视觉和自然语言处理领域的极大关注[1] [2]。

随着深度学习模型的发展，很多优秀的方法被提出，如 Venugopalan [3]等人设计的使用循环神经网络(RNN)来训练图像和描述语句之间的双向映射关系，深度学习模型已成功应用于视频描述自动生成技术[4]。然而，即使是目前最优的模型生成的描述语句也存在不通顺、内容不符等问题，如何提高描述语句准确率仍然是一项困难的工作。除此之外，由于任务的复杂性，生成全域视频内容中文描述语句鲜有研究，难度更高。

2. 技术现状

视频的自然语言描述生成技术受到了图像描述生成技术[5]的启发，早期主要研究简单视频场景下固定动作的语言描述[6]，随着深度学习技术的发展，逐渐扩展为全域短视频中复杂和未知行为的描述生成[7] [8]。

深度学习方法大多遵循编码器－解码器架构，Xu 等人[9]将基础 CNNs 转换为多个完全 CNNs (FCN)，以形成用于生成全域短视频描述的多维度网络。Pasunuru 等人[10]提出了一种新的多任务学习模型，该模型基于注意机制在编码器和解码器之间共享参数。王金金等人[11]提出了基于扩张卷积的注意力机制视频描述模型，采用 Inception-v4 对视频特征进行编码，并在 MSVD 数据集上取得了之前的最佳结果。然而，以上模型缺乏对图像特征以外的其它重要视频特征的研究，结果的准确性有待提高。

一些关于全域视频描述自动生成的研究[12] [13] [14]考虑了其它模态的视频信息，如音频特征、C3D 特征等。然而，它们在编解码阶段使用的是基础 LSTM 模型，没有充分利用多模态信息，导致结果差强

人意。除此之外，以上所有的方法都是针对视频英文描述生成技术，而未扩展到其它语言。

针对上述问题，文本提出了一种通用高效的端到端短视频描述自动生成模型，主要优化如下：1) 利用包括 RGB 图像、光流、C3D 以及音频特征在内的多模态特征来提高视频的表征能力。2) 提出了一种新的语言模型，该模型集成了由三层 LSTM 计算单元组成的多个双向编码器和基于注意机制的多模态解码器。3) 模型适用于生成包括英文、中文在内的多语言视频描述，并给出了实验过程与结果。

3. 视频中文描述自动生成技术

完整的模型框架见图 1，整个模型由两部分组成，即多模态特征提取和自然语言生成，语言生成模型又由编码器和解码器组成。首先，将视频转换为并行多通道输入，并通过不同模型提取视频的不同模态特征。每个特征由序列向量表示，输入到由三层 LSTM 计算单元组成的双向编码器中。LSTM 的前两层(深色矩形)分别计算正向和反向特征序列的隐藏状态向量，第三层(浅色矩形)融合这两个方向的输出。最后，将每个模态特征的隐藏状态向量输入到解码器中，解码阶段由基于多模态融合的注意机制和一层 LSTM 组成，以生成序列描述语言。该模型在 MSVD 数据集上取得了目前最优的实验结果，可以表明这些优化方法的有效性。

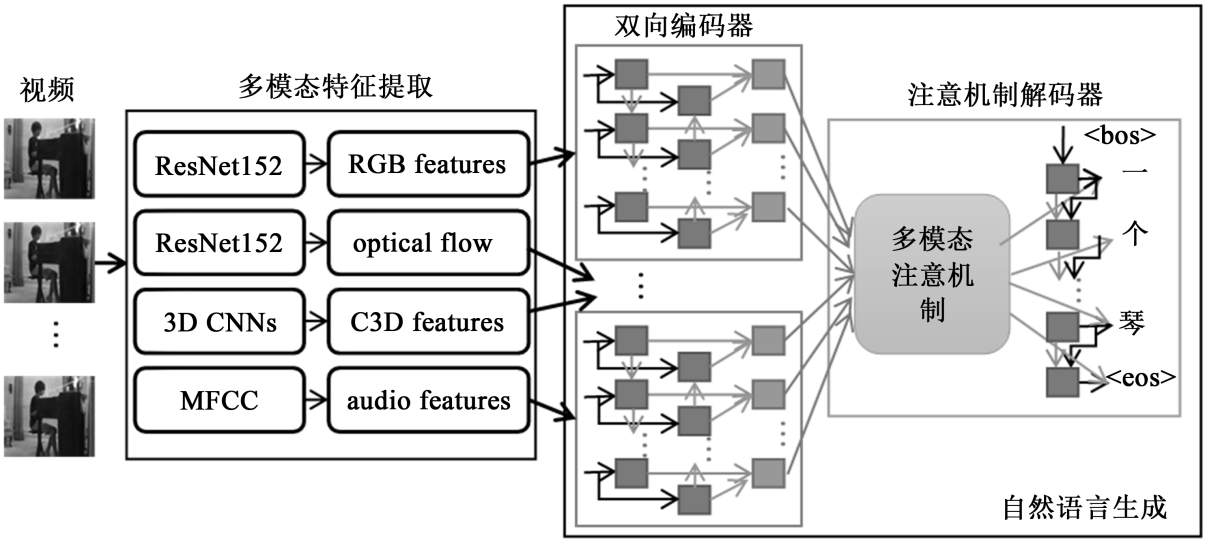

Figure 1. Automatic generation model of video Chinese description based on multimodal attention mechanism
图 1. 基于多模态注意机制的视频中文描述自动生成模型

下面将分别介绍多模态特征提取、双向编码器以及基于注意机制的解码器三个算法的详细步骤。

3.1. 多模态特征提取

特征提取模型将视频作为多通道输入，并通过不同的模型得到不同模态的视频表示，然后将这些特征分别输入到后续的双向编码器中。

图像特征：这是视频分析中最基本的特征，主要从深度卷积神经网络中提取。本文使用 ImageNet 数据集预训练 ResNet152 [15]，并提取最后 1000 维向量作为视频中每帧的 RGB 图像特征。

光流特征：该特征包含两个视频帧之间的动态信息，使用 UCF-101 数据集预训练 ResNet152，可以从视频的光流图像中提取特征，将不同方向的光流特征进行组合，便得到了包含 2000 维向量的视频帧的光流特征。

C3D 特征：光流特征只具有连续两帧之间的运动信息，而 C3D 特征[16]包含了更长时间序列连续帧上的动作信息。使用在 Sport-1M 数据集预训练深度三维卷积神经网络模型，并提取最后 4096 维向量作为视频的 C3D 特征。

音频特征：之前的研究中几乎没有针对该类特征的探讨，然而音频包含了与上述特征完全不同的活动物体的潜在信息，因此本文引入 MFCC 对该特征进行表示，最终音频特征由 68 维向量组成。

3.2. 双向编码器

自然语言生成模型由两个部分组成，即编码器和解码器。如图 2 所示，编码器由三层 LSTM 计算单元组成，前两层分别计算输入特征序列 $F\left(f_1,f_2,\cdots,f_n\right)$ 的正向 $H^f\left(h_1^f,h_2^f,\cdots,h_n^f\right)$ 和反向 $H^b\left(h_1^b,h_2^b,\cdots,h_n^b\right)$ 隐藏状态序列表示，见公式(1)和(2)。然后通过第三层拼接两个方向输出获得 $H^m\left(h_1^m,h_2^m,\cdots,h_n^m\right)$。

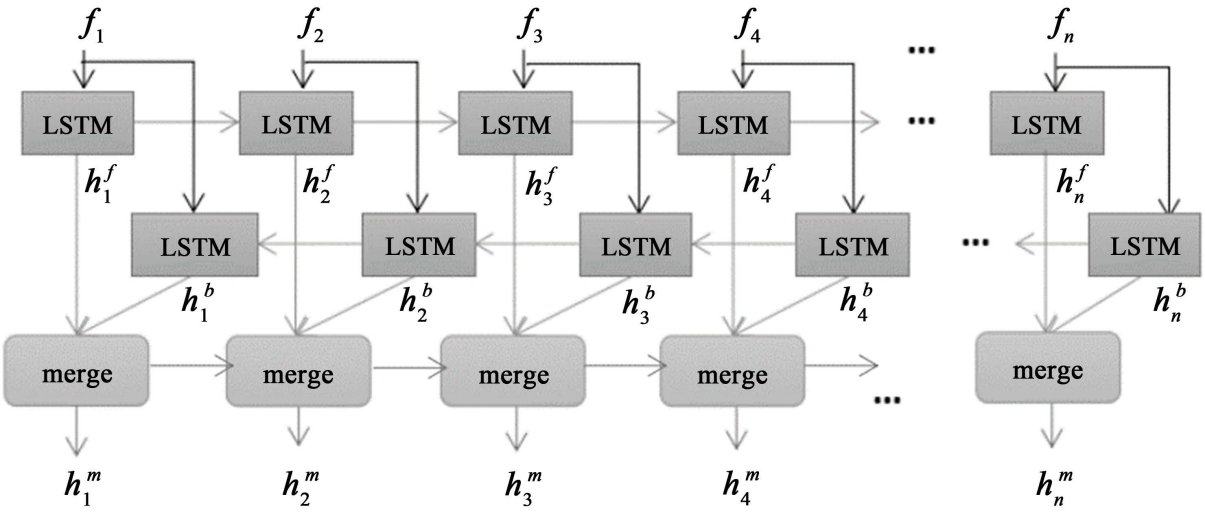

Figure 2. Bidirectional encoder
图 2. 双向编码器

$$h_t^f=\mathrm{LSTM}\left(h_{t-1}^f,f_t;\lambda^f\right) \tag{1}$$

$$h_t^b=\mathrm{LSTM}\left(h_{t+1}^b,f_t;\lambda^b\right) \tag{2}$$

以光流特征为例，假设一段视频的光流特征向量为 $F\left(f_1,f_2,\cdots,f_n\right)$，其中 f_i 是 2000 维的向量，n 是光流帧数量，则经过正反向编码后的融合向量为 $H\left(h_1,h_2,\cdots,h_n\right)$，其中 h_i 是 4000 维的向量，n 保持不变。其它模态特征的正反向编码方法与之相同，便可得到四种模态特征的编码向量表示，记为 H_1 到 H_4，分别输入到后续解码器中。

3.3. 基于多模态注意机制的解码器

自然语言生成模型的第二个部分是基于注意机制的解码器，它将多种模态特征的隐藏状态向量作为输入，如 $H_1\left(h_{11},h_{12},\cdots,h_{1n}\right)$，并输出由汉字构成的中文序列，即 $W\left(w_1,w_2,\cdots,w_m\right)$。

解码器的模型结构如图 3 所示，注意机制使得模型能够在考虑当前上下文的情况下关注特定时间或空间区域的隐藏状态，以便更准确地预测下一个词。本文使用多模态注意机制接收来自并行编码器的多个隐藏状态向量，然后将特征融合向量 $D\left(d_1,d_2,\cdots,d_m\right)$ 依次输入到最后一层 LSTM 中，以生成序列文字，以下是计算过程的详细信息。

多模态注意机制定义了整个输入序列中隐藏状态的注意力权重，对于第 i 个输出，每个模态特征由所有隐藏状态的加权和表示，见公式(3)，其中 $\alpha_{j,i,l}$ 是第 i 个输出字和第 j ($j \in [1,4]$)个模态特征的第 l 个隐藏状态之间的注意力权重。然后根据公式(4)将这些加权和组合成一个向量，即为特征融合，其中 W_{cj} 表示第 j 个模态特征的权重矩阵。在融合阶段，使用公式(5)作为激活函数，其中 b_s 是偏置值。

$$c_{j,i} = \sum_{t=1}^{N} \alpha_{j,i,l} h_{j,t} \tag{3}$$

$$d_i = W_{c1} c_{1,i} + W_{c2} c_{2,i} + W_{c3} c_{3,i} + W_{c4} c_{4,i} \tag{4}$$

$$g_i = \tanh\left(W_s s_{i-1} + d_i + b_s\right) \tag{5}$$

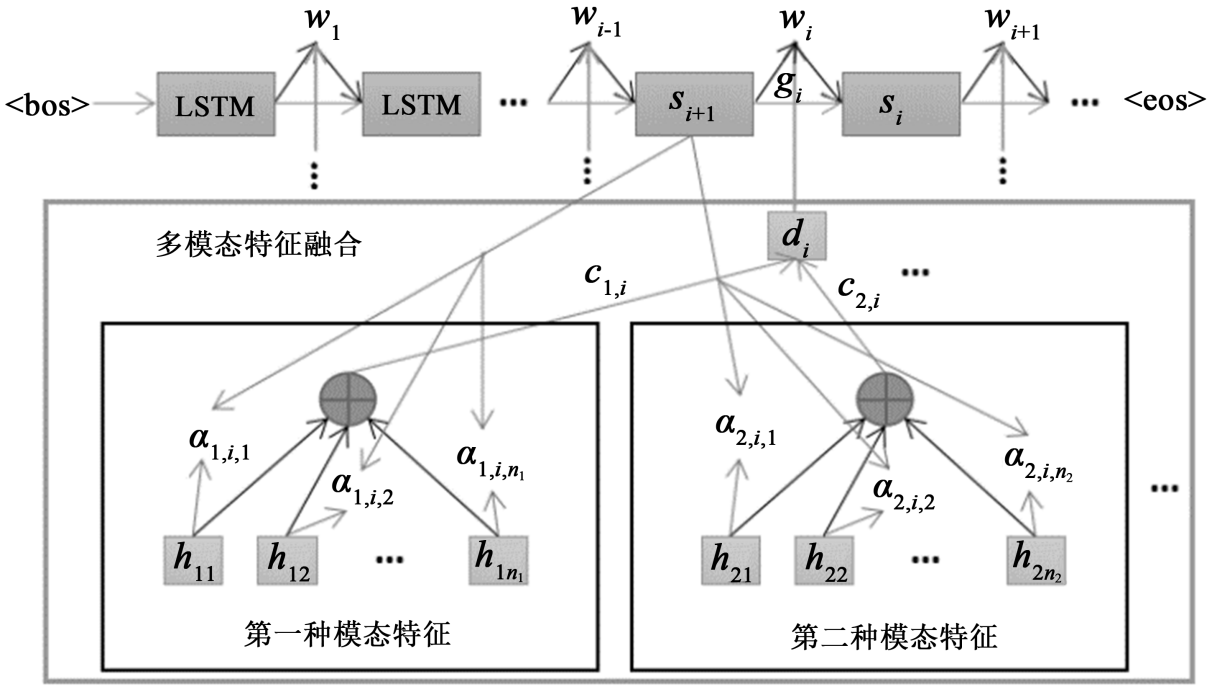

Figure 3. Decoder based on multimodal attention mechanism
图 3. 基于多模态注意机制的解码器

使用公式(6)计算输出的概率分布，并通过公式(7)生成序列单词 w_i，其中 V 表示从数据集的句子中生成的词典。<bos>标记指示 LSTM 开始生成单词，<eos>标记指示终止句子生成。

$$P\left(w \mid s_{i-1}, c_{1i}, c_{2i}, c_{3i}, c_{4i}\right) = \text{softmax}\left(W_g g_i + b_g\right) \tag{6}$$

$$w_i = \arg\max_{w \in V} P\left(w \mid s_{i-1}, c_{1i}, c_{2i}, c_{3i}, c_{4i}\right) \tag{7}$$

4. 实验

4.1. 数据集

MSVD 数据集是微软提供的视频描述生成比赛官方数据集，取自 YouTube，时长在 8 s 到 25 s 之间，它包含了 1970 段全域视频，对应 8 万 5 千条英文描述。为了添加音频特征，本文从 YouTube 网站上收集了总共 1600 段 MSVD 的现有视频，用于剪切音频信息。此外，MSVD 数据集还提供了 347 段全域短视频的 398 条中文描述语句，但这些数据量对于中文模型的评估而言有些不足，因此本文对 MSVD 进行扩充，将英文描述翻译为中文，共构建了 1600 段视频的近 6 万条中文描述语句。

4.2. 评估指标与参数设置

METEOR 最初用于评估机器翻译结果，它也同样适用于视频描述生成模型，METEOR 通过 WordNet 同义词精确比较标记匹配、词干标记、语义相似性匹配和释义匹配，保证了高可靠性。

双向编码器中 LSTM 的步长依据历史研究结果设置为 80，这个步长可在内存消耗与特征提取之间取得良好的平衡。在超出 80 帧的视频中采用平均间隔提取法，如果帧的总长度小于 80，则用零填充空白。最后，每种模态特征由矩阵 $F\left(f_1, f_2, \cdots, f_{80}\right)$ 表示。

对于训练集和测试集的数量，有 1300 段视频用于训练，300 段用于测试。除此之外，为了与其它模型方法进行对比，验证提出模型的有效性，文本还训练了全域短视频英文描述自动生成模型，除音频信息外，有 1570 段用于训练，400 段用于测试。随机选择 1200 段音频用于训练，其余 400 段用于测试，模型的所有训练参数如权重和偏置被随机初始化。

4.3. 实验结果与分析

表 1 给出了在基于多模态注意机制的全域视频描述生成模型上使用不同模态特征组合的比较结果，实验在 MSVD 中文测试集上进行。显然，所有模态特征的融合获得了最佳结果(METEOR: 20.5%)。

Table 1. Experimental results of multimodal feature combination
表 1. 多模态特征组合实验结果

特征组合	METEOR (%)
RGB + Optical Flow	18.2
RGB + C3D	17.9
RGB + Audio	18.5
RGB + Optical Flow + C3D	19.4
RGB + Optical Flow + C3D + Audio	**20.5**

上述结果表明，与光流特征和 C3D 特征(结果分别为 18.2%和 17.9%)相比，音频特征在与静态 RGB 图像特征融合时发挥了更为重要的作用，两者融合结果达到了 18.5%，这证明了在视频分析中使用音频信息的必要性。此外，三种特征的结合结果明显优于任意两种模态特征的结合结果，例如将 RGB 图像、光流和 C3D 特征作为整体输入的结果为 19.4%，这明显高于仅将 RGB 图像与光流或 C3D 结合得到的结果，这证明了输出质量与模态特征数量呈正相关。

本文针对中文描述生成模型的实验结果见表 2。原始数据集只有 347 段包含中文描述的短视频，我们将数据集扩展到 600 段，并使用其中的 100 段作为测试数据，评估每段视频包含的描述语句数量对于结果的影响。当每段视频有更多描述时，METEOR 值从 10.5%增加到 12.2%，然而再继续增加描述语句会导致过度拟合问题，使得结果下降到 10.9%。继续扩展数据集至 1300 段训练视频和 300 段测试视频，每段视频有五条描述语句，可以发现结果有明显的提升。最后，我们继续追加描述语句，结果从 19.3%上升到 20.5%，得到了 MSVD 数据集中文描述生成的最佳实验结果。

为了验证本文提出的模型优化方法对于提升全域视频描述生成结果的影响，在英文 MSVD 数据集上进行了模型训练与测试，并与近年来比较突出的其它研究成果进行了分析对比，结果见表 3。LSTM-YTcoco [2]使用具有卷积和递归结构的统一深度神经网络将视频直接翻译成句子，S2VT [3]第一次提出了包含 RGB 图像和光流特征的端到端视频描述生成网络模型，这两种方法都没有对编解码器做任何优化。

Joint-BiLSTM [17]中的 BiLSTM 模型深入捕捉视频中的全局时间信息，而 BLSTM [18]设计了基于软注意机制的卷积神经网络和双向循环神经网络的组合。多任务模型[19]在无监督视频预测和隐含生成任务的编码器和解码器之间共享参数，取得了不小的提升，扩张卷积模型[11]是之前最优的结果，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，但它们都没有考虑多模态视频特征。与其它先进方法相比，本文提出的基于多模态注意机制的全域视频描述生成技术获得了最佳结果(41.76%)，实验证明了该方法的有效性和优越性。

Table 2. Experimental results of video Chinese description generation
表 2. 视频中文描述生成实验结果

训练视频数	测试视频数	描述语句数量/视频	METEOR (%)
500	100	1-2	10.5
500	100	5	12.2
500	100	20	10.9
1300	300	5	19.3
1300	**300**	**20**	**20.5**

Table 3. Comparison with other models
表 3. 与其它模型的比较

模型方法	视频特征	双向编码	注意机制	METEOR
LSTM-YTcoco [2]	RGB	No	No	29.07
S2VT [3]	RGB + Optical Flow	No	No	29.8
Joint-BiLSTM [17]	RGB	Yes	No	30.3
BLSTM [18]	RGB	Yes	Yes	32.6
Multi-Task Model [19]	RGB	Yes	Yes	36.0
扩张卷积模型[11]	RGB	No	Yes	38.45
本文	**RGB + Optical Flow + C3D + Audio**	**Yes**	**Yes**	**41.76**

5. 结束语

本文提出的基于多模态注意机制的全域视频描述生成技术可以有效地从不同长度的视频中提取多模态特征，此外，基于多模态注意机制的双向语言模型使网络能够在整段视频中捕获更多的时间和行为信息。MSVD 数据集上的结果表明，本文提出的方法优于以往最先进的模型。除此之外，本文在 MSVD 数据集基础上扩展构建了视频中文描述生成数据集，其实验结果可以为今后的中文视频处理分析研究提供完整而详细的参考。并且本文提出的多模态视频特征提取模型可用于其它视频分析任务，如行为和对象识别等。

针对多语言描述生成任务后续还有很多可以继续研究优化的地方，如构建海量视频描述标注数据、视频其它模态特征的抽象与提取、中文自然语言模型的改进等等。

参考文献

[1] Barbu, A., Bridge, A., Burchill, Z., *et al*. (2012) Video in Sentences Out. 28*th Conference on Uncertainty in Artificial Intelligence*, Catalina Island, 14-18 August 2012, 274-283.

[2] Venugopalan, S., Xu, H., Donahue, J., *et al*. (2014) Translating Videos to Natural Language Using Deep Recurrent Neural Networks. *Proceedings of the* 2015 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Denver, May-June 2015, 1494-1504. https://doi.org/10.3115/v1/N15-1173

[3] Venugopalan, S., Rohrbach, M., Donahue, J., *et al*. (2015) Sequence to Sequence—Video to Text. *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 4534-4542. https://doi.org/10.1109/ICCV.2015.515

[4] 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[J]. 自动化学报, 2022, 48(2): 375-397.

[5] Chen, X. and Zitnick, C.L. (2015) Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 2422-2431. https://doi.org/10.1109/CVPR.2015.7298856

[6] Rohrbach, M., Qiu, W., Titov, I., *et al*. (2013) Translating Video Content to Natural Language Descriptions. *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 433-440. https://doi.org/10.1109/ICCV.2013.61

[7] 付燕, 马钰, 叶鸥. 融合深度学习和视觉文本的视频描述方法[J]. 科学技术与工程, 2021, 21(14): 5855-5861.

[8] 孙红莲, 李永刚, 季兴隆, 王霈烨, 吴小旭. 基于深度神经网络和自注意力的视频事件描述[J]. 电脑知识与技术: 学术版, 2020, 16(33): 187-189.

[9] Xu, H., Venugopalan, S., Ramanishka, V., *et al*. (2015) A Multi-Scale Multiple Instance Video Description Network. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 272-279.

[10] Pasunuru, R. and Bansal, M. (2017) Multi-Task Video Captioning with Video and Entailment Generation. *Meeting of the Association for Computational Linguistics*, Vancouver, 30 July-4 August 2017, 1273-1283. https://doi.org/10.18653/v1/P17-1117

[11] 王金金, 曾上游, 李文惠, 等. 基于扩张卷积的注意力机制视频描述模型[J]. 电子测量技术, 2021, 44(23): 99-104.

[12] Jin, Q., Chen, J., Chen, S., *et al*. (2016) Describing Videos Using Multi-Modal Fusion. *ACM on Multimedia Conference*, Amsterdam, 15-19 October 2016, 1087-1091. https://doi.org/10.1145/2964284.2984065

[13] Ramanishka, V., Das, A., Dong, H.P., *et al*. (2016) Multimodal Video Description. *ACM on Multimedia Conference*, Amsterdam, 15-19 October 2016, 1092-1096. https://doi.org/10.1145/2964284.2984066

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

[15] He, K., Zhang, X., Ren, S. and Sun, J. (2017) Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 770-778. https://doi.org/10.1109/CVPR.2016.90

[16] Tran, D., Bourdev, L., Fergus, R., *et al*. (2016) Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[17] Yi, B., Yang, Y., Shen, F., *et al*. (2018) Bidirectional Long-Short Term Memory for Video Description. *ACM on Multimedia Conference*, Seoul, 22-26 October 2018, 436-440.

[18] Peris, Á., Bolaños, M., Radeva, P. and Casacuberta, F. (2019) Video Description Using Bidirectional Recurrent Neural Networks. *International Conference on Artificial Neural Networks*, Munich, 17-19 September 2019, 3-11. https://doi.org/10.1007/978-3-319-44781-0_1

[19] Cho, K., Van Merrienboer, B., Bahdanau, D., *et al*. (2014) On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *Proceedings of SSST*-8, 8*th Workshop on Syntax*, *Semantics and Structure in Statistical Translation*, Doha, 25 October 2014, 103-111.