

文章编号:1007-5321(2019)06-0155-07

DOI:10.13190/j.jbupt.2019-057

全卷积神经结构的段落式图像描述算法

李睿凡^{1,2}, 梁昊雨¹, 冯方向¹, 张光卫^{2,3}, 王小捷^{1,2}

(1. 北京邮电大学 计算机学院, 北京 100876; 2. 教育部信息网络工程研究中心, 北京 100876;

3. 北京邮电大学 网络技术研究院, 北京 100876)

摘要: 针对段落式图像描述生成研究中提升描述语句之间的连贯性问题, 提出了一种基于全卷积结构的图像段落描述算法. 采用基于卷积网络的区域检测器获取图像表示, 结合段落语言学角度的层次性, 构建一种层次性的深度卷积解码器对图像表示解码, 自动生成段落式文本描述. 同时将门控机制嵌入卷积解码器网络中, 以提升模型的记忆能力. 实验结果表明, 相比于基于循环神经网络等传统段落图像的描述方法, 新算法能够为图像生成更为连贯的段落式文本描述, 在评测指标上取得较好的结果.

关键词: 卷积网络; 深度学习; 图像描述; 连贯性

中图分类号: TN309.2

文献标志码: A



OSID 码:

Paragraph Image Captioning with Deep Fully Convolutional Neural Networks

LI Rui-fan^{1,2}, LIANG Hao-yu¹, FENG Fang-xiang¹, ZHANG Guang-wei^{2,3},
WANG Xiao-jie^{1,2}

(1. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. Engineering Research Center of Information Networks, Ministry of Education, Beijing 100876, China;

3. Institute of Network Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: How to improve the coherence among descriptive sentences for the paragraph image captioning is paid attention currently. A fully convolutional neural architecture for paragraph image captioning was proposed. An image representation is first obtained using a region detector based on a convolutional network. Then a hierarchical deep convolutional decoder is constructed to translate the image representation, automatically generating a paragraph text description. In addition, the gating mechanism is embedded in the convolutional decoder network to improve memory capacity of the model. Experiments demonstrate that compared with those traditional methods based on recurrent neural networks, the proposed algorithm can generate more coherent paragraph text descriptions for images, achieving better results on evaluation metrics.

Key words: convolutional networks; deep learning; image captioning; coherence

段落式图像描述(paragraph image captioning)的目标是为给定的图像生成描述性的自然语言段落.

收稿日期: 2019-04-14

基金项目: 国家重点研发计划项目(2019YFF0303302); 国家自然科学基金项目(61906018); 国家电网公司总部科技项目(5200-201918255-A-0-0-00)

作者简介: 李睿凡(1975—), 男, 副教授, E-mail: rli@bupt.edu.cn.

该任务一方面连接着计算机视觉和自然语言处理两个领域,是跨模态智能的重要研究方向. 另一方面,它是盲人导航以及幼儿早期教育等前沿应用的核心技术. 因而开展段落式图像描述的研究有着十分重要的意义.

当前,段落式图像描述算法主要延伸了编码器与解码器组合的端到端结构^[1]. 具体地,基于卷积神经网络(CNN, convolutional neural networks)的编码器首先将图像表示为较低维的视觉向量. 随后采用基于循环神经网络(RNN, recurrent neural network)的解码器将该视觉向量解码为自然语言段落. 其中,循环神经网络解码器在时间序列建模上有一定优势. 但是其长时记忆能力有限,且在训练过程中易遭遇梯度消失问题. 因而导致其建模较长段落的能力较差,生成段落的连贯性不能令人满意.

为了提升生成描述段落的连贯性,笔者提出了一种全卷积神经结构的段落式图像描述算法. 该神经网络结构的解码器由双层的门控卷积网络构成. 其中,句子CNN解码器捕捉段落内的句子之间关系以强调句子间的连贯性,词CNN解码器负责生成段落内的单词. 相较于RNN解码器,CNN解码器具有更好的“长时”视野. 通过门控机制,增强了解码器的长时记忆能力. 实验结果表明,该解码器在段落式描述任务上具有更好的效果.

1 相关工作

早期的图像描述任务研究集中于单句式描述上,即为给定图像生成一个描述性句子. 近年来,随着数据时代的到来,且得益于硬件计算能力的提升,深度学习成为解决图像描述问题的主流方法. Vinyals等^[1]在2015年提出了基于神经网络的编码器—解码器框架,此后的大多数研究都基于该框架展开. 得益于RNN解码器在短文本建模上的优良能力,单句式描述研究已经取得了瞩目的进展^[2-7].

随着单句描述研究逐渐成熟,研究人员将注意力转向更具挑战性的段落式描述任务中. 该任务由Krause等^[8]于2017年提出,目前已成为深度学习、多模态智能领域热门的研究方向之一. 作为图像描述的深化任务,图像段落描述的解决方案同样基于编码器—解码器结构. 而相比于单句描述任务,段落式描述需要更细粒度的图像内容理解和更强刻画能力的语言模型.

Krause等^[8]根据段落的层次性,提出了层次性

的RNN语言解码器,表明了层次模型在段落生成任务上的优越性,但其对句子间连贯性的监督较为粗糙,生成的段落连贯性有待提高. 随后有几个显著的研究工作. Liang等^[9]采用对抗学习^[10]方法来增强具有3个层次的RNN解码器的段落建模能力,以提高生成段落的连贯性,但生成对抗网络结合3层RNN解码器的网络训练复杂,模型收敛慢. Chatterjee等^[11]通过更丰富多样的监督信息指导层次RNN生成更为连贯的段落. 而Che等^[12]和Wang等^[13]通过增加视觉侧的监督信息,如图像深度估计图、对象关系等,以期丰富段落的描述内容. 然而,这些方法都延伸于RNN解码器的框架之下,无法避免RNN的固有问题. 当建模长文本时,随着时间序列的向后推移,RNN隐藏信息的衰减极大地削弱了解码器关注“上文”的视野大小,由此带来的长时记忆能力不足造成生成段落连贯性的下降.

针对现有方法中采用RNN解码器造成的不足,启发于卷积网络的非时间序列特性,提出了一种基于卷积结构的段落语言解码器,扩大了解码器的视野大小,增强了解码器的“长时”关注能力. 论文将卷积结构应用于图像段落描述生成任务上,并结合门控机制,增强了解码器的长时记忆能力,改进了现有方法生成段落连贯性较差的问题.

2 全卷积段落解码器

2.1 模型框架

编码器由目标检测器和卷积神经网络组成,将输入图像表示为压缩的图像特征. 解码器由句子CNN解码器和词CNN解码器构成. 句子CNN接收图像特征,并分析上下文语义,为段落中的每句话生成多模态语义向量. 词CNN接收每句话的语义向量,生成句子中的所有单词. 将所有句子按序排列,得到输出段落. 模型总体框架图如图1所示.

2.2 卷积解码器

模型中的句子CNN解码器和词CNN解码器具有相似的结构. 区别在于,句子CNN以图像特征为原始输入,词CNN以每句话的多模态语义向量为原始输入. 且句子CNN分析句子层的上下文语义,词CNN分析词层面的上下文语义.

对于词解码器,在图像描述任务的解码过程中,每个时间步需要生成一个单词. 因而解码器的任务就是在每个时间步接收过去时间步的“上文”信息,并综合图像语义信息,生成当前时间步相应的单词.

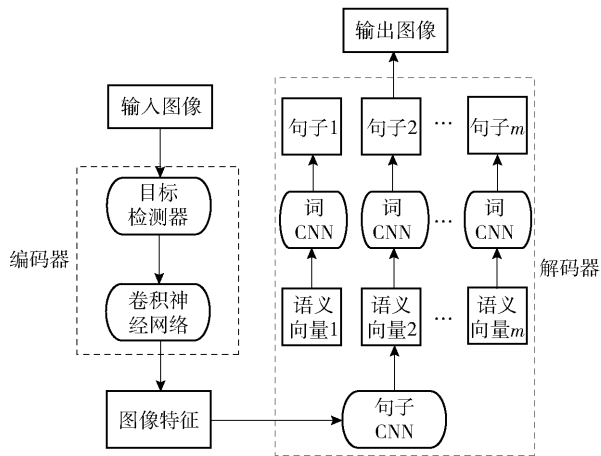


图1 模型框架图

传统的循环网络解码器通常以长短时记忆网络 (LSTM, long-short term memory) 单元为基本元素,“上文”信息依赖于隐藏单元存储。在第 t 个时间步,根据此时隐藏单元信息 h_t 生成单词 w_t ,即

$$w_t = \text{softmax}(\text{fc}(h_t)) \quad (1)$$

$$h_t = \text{LSTM}(h_{t-1}, w_{t-1}, v) \quad (2)$$

其中: v 表示图像信息, fc 表示全连接神经网络层。随着时间步的增长,隐藏单元中的“上文”信息将发生衰减,解码器的“长时”能力减弱,导致生成段落的质量较差。

为赋予解码器更强的长时记忆能力,对解码器的基本组成单元进行改进,将 LSTM 改进为带有门控的卷积网络。2 种结构的对比如图 2 所示。在 t 时刻,共有 $t-1$ 个“上文”信息。由于信息衰减,LSTM 解码器视野小于 $t-1$,而 CNN 解码器视野大小始终保持为 $t-1$ 。卷积解码器直接以上文所有单词作为输入,具有更强的“长时”能力。同时,当生成一个长度为 n 的句子时,LSTM 解码器的序列操作数为 $O(n)$,而 CNN 解码器仅需 $O(1)$ 级别的序列操作。

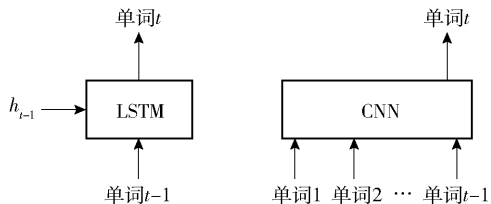


图2 2 种结构生成单词的差异比较

门控卷积解码器结构如图 3 所示。该结构包含 3 个网络层:嵌入层、门控卷积层以及输出层。嵌入层将输入单词映射为低维向量。门控卷积层接收所有已生成单词的向量并输出预测向量。输出层则将

预测向量映射成词表上的概率分布。

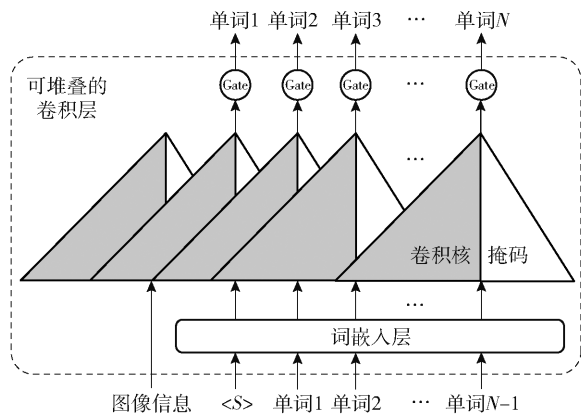


图3 门控卷积解码器结构

在生成每个单词时,卷积解码器直接将所有已生成单词的嵌入表示向量作为输入,无需通过隐藏层向量记忆上下文信息,其单词生成过程可表示为

$$w_t = \text{CNN}(v, w_1, w_2, \dots, w_{t-1}) \quad (3)$$

具体而言,使用一维卷积核对图像信息和输入单词序列进行操作,单词嵌入表示向量的每一维可被视作一个通道,卷积核的大小决定了视野的大小。通过堆叠多层卷积核,可达到在每个时间步能观察到已生成的所有单词的目的。

需要指出,传统 CNN 的卷积层和池化层的组合并不能给解码器带来记忆能力。为赋予所提解码器“记忆”能力,在卷积层后加入门控模块^[14]。卷积层的输出为 2 个相同维度的向量 U_t 和 G_t 。其中 U_t 蕴含已生成单词的语义信息。而 G_t 则充当控制门的角色,将 U_t 中的信息选择性地传递到输出层中:

$$U_t = W_U * X + b_U \quad (4)$$

$$G_t = W_G * X + b_G \quad (5)$$

$$h_t^c = U_t \odot \sigma(G_t) \quad (6)$$

其中: X 表示卷积层的输入,即上文单词嵌入序列,符号 $*$ 表示卷积运算, W_U, W_G, b_U, b_G 表示可学习参数, \odot 表示向量按元素乘法, σ 表示 sigmoid 函数,即 $\sigma(x) = 1/(1 + e^{-x})$ 。

进一步,输出层将 h_t^c 映射为在词表上的概率分布,即

$$p_t = \text{softmax}(W_p h_t^c) \quad (7)$$

最后,根据概率分布 p_t 对单词进行采样,得到 t 位置的单词。

3 段落生成算法

基于卷积神经结构的段落式图像描述算法分为

2个主要过程:一个是利用基于卷积网络的目标检测器对图像进行编码;另一个是通过卷积解码器对图像特征进行层次性的解码,得到描述性段落.算法实现步骤可总结如下.

算法1 基于全卷积神经结构的段落式图像描述算法 PCIC

输入:图像 I

输出:描述段落 P , 包含 m 个语句, 其中第 i 句话包含 n_i 个单词.

步骤1 利用目标检测器(region proposal network)提取图像中 K 个感兴趣区域.

步骤2 通过预训练的卷积神经网络,对每个感兴趣区域提取出维度为 4 096 的特征,并使用全连接前馈神经网络将其压缩为 1 024 维的向量表示,从而得到图像区域向量集 $\{v_1, v_2, \dots, v_K\}$.

步骤3 对区域向量集进行按位最大池化操作,得到该图像的全局向量表示 v .

步骤4 在 $t=1$ 时刻,通过词嵌入层获取句子开始标志的嵌入向量 S_0 ,将 S_0 和 v 拼接,输入到句子 CNN 解码器中,获得指导第一句话生成的多模态语义向量 I_1 .

步骤5 将 I_1 输入到词 CNN 解码器中,解码得到第一个句子 $y_1 = \{w_{11}, w_{12}, \dots, w_{1n_1}\}$.

步骤6 在 $t=2, \dots, m$ 时刻,通过词嵌入层获取第 $t-1$ 句话的综合嵌入特征 S_{t-1} ,该特征为句子中所有词嵌入向量的按位平均向量.将 S_{t-1} 和 v 拼接,输入到句子 CNN 解码器中,获得指导第 t 句话生成的多模态语义向量 I_t .

步骤7 将 I_t 作为 t 时刻的图像语义信息,输入到词 CNN 解码器中,解码得到第 t 个句子 $y_t = \{w_{t1}, w_{t2}, \dots, w_{tn_t}\}$.

步骤8 将 m 个句子 y_1, y_2, \dots, y_m 按序排列,即可得到描述段落 P .

在以上图像段落描述算法中,解码生成句子的步骤7可采取2种常用的单词采样算法,即最大概率采样和集束搜索方法.简言之,最大概率采样基于一种贪心策略.它在每个时间步选取当前概率最大的单词.该算法具有较小的时间和空间复杂度,但容易忽视全局最优解.而集束搜索是一种启发式搜索算法.它在每个时间步保留含有概率最大的若干词,词的数量由束大小确定.相比于最大概率采样,集束搜索的时间和空间消耗稍大,但更容易获得较优解.而当束的大小设定为1时,集束搜索方法

退化为最大概率采样方法.

4 实验

4.1 实验设置

为验证算法有效性,采用斯坦福大学最新建立的图像段落描述公开数据集^[8].该数据集包含从 Visual Genome^[15] 和 MS COCO^[16] 两个图像数据集中选取的 19 551 张图片,每张图片对应一个描述文本段落.总体而言,每个段落平均包含 5.7 个句子,且每个句子包含 11.9 个词.为了与基线方法进行公平比较,遵循其他文献将数据集划分为 3 个子集:训练集、验证集以及测试集.它们分别包含 14 757、2 487 以及 2 489 个图像—文本段落描述对的样本.

整个实验所使用的服务器操作系统环境为 Linux.该服务器配置了英伟达 GeForce GTX 1080Ti 显卡.软件环境为采用 Python 编程语言的开源框架 PyTorch.一些实验参数设置如下.区域检测器所检测的区域个数设定为 50,且每个区域向量的编码维度为 1 024.而词嵌入的维度同样设定为 1 024.经过对比实验,采用集束搜索方法生成段落.其中束的大小设置为 2.段落中最大句子数目设定为 6,同时每句话的最大单词数设定为 30.整个模型使用 Adam 优化器进行训练,其中的学习率设置为 10^{-4} .实验中依据算法在验证集上的表现确定超参数.

4.2 性能评估

采用 5 个客观评价指标评价算法生成描述段落的质量,包括 BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), BLEU-4 (B-4)^[17] 以及 CIDEr^[18].与 BLEU 指标相比,CIDEr 指标更贴近人的主观评价,因而在图像描述任务上具有更好的评价意义,更能衡量描述的连贯性,因而为众多研究者采用.

为了验证提出方法的有效性,实验将所提方法 PCIC 和 3 种基线方法进行对比,包括 Sentence-Concat^[2], Image-Flat^[2] 以及 Hierarchical-RNN^[3].其中,方法 Sentence-concat 将 5 个独立的单句描述拼接起来合成段落.方法 Image-Flat 通过单层的 RNN 解码器生成段落.而方法 Hierarchical-RNN 通过层次性的 RNN 解码器生成段落.表 1 所示为所提算法和基线方法的客观评价指标对比.此外,为了说明人类描述段落和机器生成段落的差异,该表的最后一行展示了人类描述段落在 5 个评价指标上的得

分. 这些图像描述的文字段落来自于斯坦福图像段落数据集中随机抽取的 500 个段落.

表 1 不同模型以及人评测指标结果

模型	CIDEr	B-1	B-2	B-3	B-4
Sentence-Concat	6.8	31.1	15.1	7.6	4.0
Image-Flat	11.1	34.0	20.0	12.2	7.7
Hierarchical-RNN	13.5	41.9	24.1	14.2	8.7
PCIC	15.9	41.3	23.9	14.1	8.2
Human	28.6	42.9	25.7	15.6	9.7

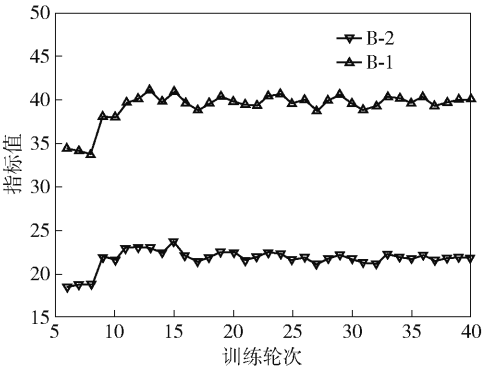
仔细观察表 1 可以看出,人类描述段落和机器生成段落的 BLEU 得分较为接近,而 CIDEr 得分相差巨大. 这说明 CIDEr 评价指标能够更好地说明机器描述方法和人类描述之间的显著差异. 相较于 BLEU 指标仅考虑 n 元组的匹配程度而忽略了语义,基于共识的 CIDEr 指标更好地反映生成段落的连贯性. 在 CIDEr 指标上,所提方法较 Sentence-Concat 方法高出 133.8%. 这显示出段落描述任务和单句描述任务间的巨大差异. 而且提出的方法比 Image-Flat 方法高出 43.2%. 这验证了所提解码器的层次性结构的有效性. 进一步,所提方法较 Hierarchical-RNN 方法高出 17.8%,说明了所提解码器卷积结构的优势. 对比 Sentence-Concat、Image-Flat、Hierarchical-RNN 3 种方法,所提方法在 CIDEr 指标上取得了更好的评测结果,提高了生成段落的质量,有效地弥补了传统方法建模段落描述能力不足的缺点.

如前所述,束大小是影响所提算法性能的一个重要参数,因而笔者评估了该参数对指标影响的结果. 表 2 显示了集束搜索中不同束大小参数对评测结果的影响. 参数的设置从 1~4. 其中,束大小为 1 的集束搜索等价于最大概率采样,即在每个时刻取当前概率最大的单词. 从表 2 可看出,当束大小为 2 时,评测结果达到最优. 当束大小为 1 时,由于单词搜索空间过小,丢失了较多解码信息,生成的段落非较优解. 当束大小逐渐增大时,单词搜索空间也不断增大,更容易获得较优解. 然而,当束大小大于 2 时,搜索空间的增大会使段落间句子重复度增加,损害了段落的多样性,造成评价指标的下降. 同时,束大小过大会导致解码时间复杂度大幅增加. 因此,束大小取 2 是平衡生成段落质量和解码时间复杂度的较好选择.

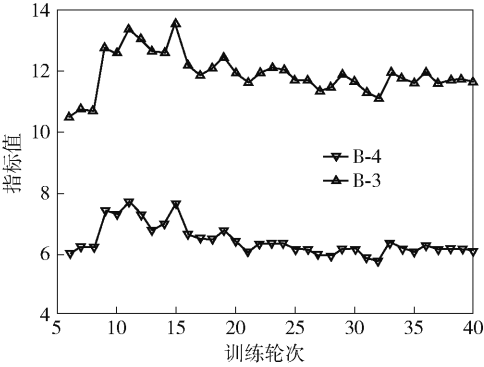
表 2 不同束大小参数的评测指标结果

束大小	CIDEr	B-1	B-2	B-3	B-4
1	14.8	40.9	23.1	13.6	7.7
2	15.9	41.3	23.9	14.1	8.2
3	15.1	41.5	23.7	14.0	7.8
4	13.7	40.4	22.3	12.8	7.5

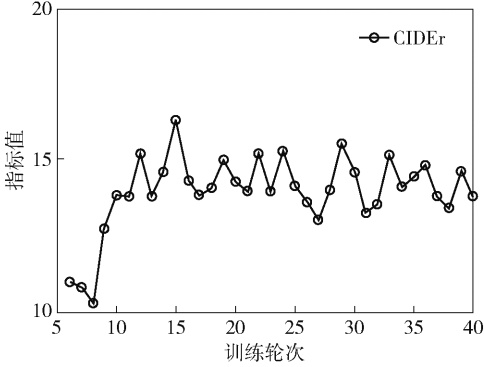
进一步,笔者考察迭代过程中各个指标的变化,表明各指标结果的一致性. 图 4 所示为不同迭代轮次时, BLEU-1 和 BLEU-2、BLEU-3 和 BLEU-4 以及 CIDEr 等 5 个指标的评测结果变化. 总体而言,这些指标随着迭代轮次的变化趋势基本保持一致. 在第 5~15 个轮次之间,评测指标呈上升趋势. 在第 15 个轮次左右时达到最优性能. 此后,模型出现了一



(a) BLEU-1和BLEU-2随训练轮次的变化



(b) BLEU-3和BLEU-4随训练轮次的变化



(c) CIDEr随训练轮次的变化

图 4 评测指标随训练轮次变化

定程度的过拟合,指标有一定的下降趋势.

4.3 主观评价

为进一步展示所提方法的有效性,笔者将细致考察生成段落描述的细节,随机选取测试集中的图

片以及对应的标签段落,并分别使用所提方法和 Hierarchical-RNN 方法生成描述段落. 图 5 展示了图像以及标注段落、所提方法、Hierarchical-RNN 方法产生的内容.





PCIC	Hierarchical-RNN	Ground Truth	
	There are two giraffes in the picture. The giraffes are very tall. The giraffe on the right is standing straight up and looking at the camera. The giraffe has long horns. The giraffe has a long neck. There is a pathway to the left of the giraffe.	The main focus of the picture is a tall giraffe. There are two giraffes in the photo. It is cream with brown spots. There are trees in the background. The giraffe is taller than the trees and <unk> to see it. The grass is very green and <unk> looking.	The are two giraffes in the forest. One giraffe is smaller than the other. The giraffes have horns. Their ears are pointy and white. They have brown spots. Their necks are long. Their legs are long. The end of their tail is black. Behind them are trees and bushes. On the ground is grass. the grass are dry. The face of the giraffes are brown.
	This picture is taken outside on a sunny day. There is a tall building with many windows on the side of the street. There is a white bus with a red stripe walking on the street. There are many tall buildings on the side of the road. There are many cars on the street .	There are two buses driving down the street . There is a bus going down the road. There are a lot of trees on the other side of the bus. There are a bunch of pedestrians walking down the street. There are several buildings along the side of the road. On the other side of the bus, there is a man walking down a road.	A double decker bus is driving down the street. The top half of the bus has many windows and the sides are blue. The bottom half of the bus is white. Under the bus is a paved area with many yellow lines painted on it. Behind the bus are tall grey skyscrapers. Next to the bus is a white van with silver hubcaps. There are a few green trees.
	A man is riding his surfboard on the water. He is riding a wave. The water has small waves. The surfboard is white with a black design on it. There is a shadow of the man on the surfboard. The water is a greenish color .	A man is surfing in the water. The man is wearing a black wet suit. There is a large wave behind the surfer. There is a person surfing on the wave in front of the person. There is a person sitting on the other side of the water looking.	A male is surfing in the blue ocean. There is a wave behind him with white foam surfacing on the water. The male surfer is laying on the surfboard in preparation to ride a big wave. He is wearing a black wet suit that covers his entire body. The surfer has dark brown hair and the board is red.
	A train is on the train tracks. The train is black and white. The train has a number on the front of it. There is a large metal building on the other side of the train near the train. There are many trees on the side of the train.	There is a train on the track. The train has a large headlight. There are wires above the tracks. The platform is covered in snow and there is snow on it. There is a platform on the other side of the train with a brown roof on it. It is a sunny day .	This is a large, black and red train. A jet of steam comes out the side and hits the ground. The train has three headlights on the front as well as an identification number. The door to the front car is slid open and a man stands at the rear of the car doing something. There are trees in the background behind he train.

图 5 部分段落生成结果

针对第 1 幅图,所提方法生成的段落用第 1 句话首先描述了图像中的最显著信息:两只长颈鹿(There are two giraffes). 随后的第 2~第 5 句话从不同的角度描述了长颈鹿的细节特征(tall、horn、neck 等). 最后一句话描述了图片中的非显著内容. 其他 3 幅图的段落描述结果也有类型现象. 通过与 Hierarchical-RNN 方法对比可看出,所提方法的生成段落具有更强的上下文连贯性和语言逻辑性. 相比于 Hierarchical-RNN 的生成段落存在大量冗余信息,一些句子存在重复,所提方法减少了信息的重复表达. 而且,这种金字塔式的描述方法和人类的认知系统非常贴近.

5 结束语

笔者提出一种基于全卷积神经网络结构的段落式图像描述生成模型. 用基于卷积网络的区域检测器获取图像表示. 针对语言段落的层次,构建一种层次性的深度卷积解码器对图像表示解码,并引入门控机制提升模型的记忆能力,生成更具连贯的段落式图像描述. 实验结果表明,该算法能够在评测指标上取得较好的结果,生成更为连贯的段落式图像文本描述.

参考文献:

[1] Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator[C] //2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2015: 3156-3164.

[2] Lu Jiasen, Xiong Caiming, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning[C] //2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2017: 375-383.

[3] Mao Yuzhao, Zhou Chang, Wang Xiaojie, et al. Show and tell more: topic-oriented multi-sentence image captioning[C] //Proceedings of the 27th International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 4258-4264.

[4] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention[C] // International Conference on Machine Learning. Lille, France: ACM, 2015: 2048-2057.

[5] You Quanzeng, Jin Hailin, Wang Zhaowen, et al. Image captioning with semantic attention[C] //2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2016: 4651-4659.

- [6] Karpathy A, Li Feifei. Deep visual-semantic alignments for generating image descriptions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2015: 3128-3137.
- [7] Anderson P, He Xiaodong, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 6077-6086.
- [8] Krause J, Johnson J, Krishna R, et al. A hierarchical approach for generating descriptive image paragraphs[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2017: 317-325.
- [9] Liang Xiaodan, Hu Zhiting, Zhang Hao, et al. Recurrent topic-transition GAN for visual paragraph generation[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE Press, 2017: 3362-3371.
- [10] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. Cambridge: MA, MIT Press, 2014: 2672-2680.
- [11] Chatterjee M, Schwing A G. Diverse and coherent paragraph generation from images[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 747-763.
- [12] Wang Z, Luo Y, Li Y, et al. Look deeper see richer: depth-aware image paragraph captioning [C] // 2018 ACM Multimedia Conference. Association for Computing Machinery. New York: ACM Press, 2018: 672-680.
- [13] Che Wenbin, Fan Xiaopeng, Xiong Ruiqin, et al. Paragraph generation network with visual relationship detection[C]//2018 ACM Multimedia Conference on Multimedia-MM' 18. New York: ACM Press, 2018: 1435-1443.
- [14] Dauphin Y N, Fan A, Auli M, et al. Language modeling with gated convolutional networks[C]//The 34th International Conference on Machine Learning-Volume 70. Sydney, Australia: ACM Press, 2017: 933-941.
- [15] Krishna R, Zhu Yuke, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.
- [16] Chen X, Fang H, Lin T Y, et al. Microsoft COCO captions: data collection and evaluation server[J]. arXiv preprint arXiv: 1504. 00325, 2015.
- [17] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//The 40th Annual Meeting on Association for Computational Linguistics (ACL). PA, USA: ACL, 2002: 311-318.
- [18] Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE Press, 2015: 4566-4575.