[文章编号] 1003－4684(2020)02-0061-06

基于多尺度特征的图像描述生成模型

周星光，靳华中，徐雨东，李晴晴，胡 满
（湖北工业大学计算机学院，湖北 武汉 430068）

[摘 要] 针对现有基于深度学习图像描述生成模型，在图像特征编码阶段，由于编码器提取的图像特征较为单一，图像信息利用不充分，造成文字对图片内容描述得不够准确、语义较模糊的问题，在 VGG19 基础上，改进现有模型对图像特征的编码形式，通过提取和融合图像多尺度特征的方法，获取更丰富的图像信息。在 MSCOCO 数据集上进行训练和测试，实验结果表明，提出的模型能够生成更加准确、完整，更有意义的图像描述语句。

[关键词] 图像描述生成；深度学习；多尺度；图像特征
[中图分类号] TP3-0 [文献标识码] A

随着深度学习的出现，计算能力的提升，人工智能技术得到了飞速发展。图像描述生成涉及计算机视觉和自然语言翻译技术，目的是将图像视觉信息和语言文字信息联系起来，经过对图像视觉信息的特征提取，自动生成关于图像内容的语言描述。图像描述生成对于计算机实现快速检索和分析图像数据具有非常重要的意义。自动生成准确的图像描述文字存在着诸多挑战和困难，是目前人工智能领域研究的难点和热点。

现有主流的图像描述生成模型通常采用编码器—解码器结构，其中编码器用来提取图像特征，解码器作为语言模型用来生成描述性语言。近年来，依靠深度学习的快速发展，特别是卷积神经网络[1]（Convolution Neural Network，CNN）计算机视觉领域已经取得了诸多颠覆性成果，其中发展迅猛的目标检测与识别技术在 ImageNet，MSCOCO 等公开的数据集上面都取得了突破性的进展。计算机视觉利用 CNN 提取图像特征特性。相对于传统的图像特征提取方法，CNN 可以更好地提取图像特征。

自然语言处理是研究如何使机器“读”和“说”，是实现人和机器之间用人类使用的自然语言进行更加有效沟通的关键技术。自然语言的飞速发展带来了人机交互形式的改革与创新。近年来，自然语言处理领域也在进行着飞速的发展。例如，在斯坦福大学发起的文本理解挑战赛（Stanford Question Answering Dataset，SquAD）中，微软亚洲研究院提交的模型在精准匹配指标上首次超越人类的水平，IBM 在自然对话环境中的语言识别错误率达到了接近人类的水平，基于神经网络的机器翻译的准确率和速度都实现显著的提升。

图像描述生成技术具有非常广阔的实际应用场景。图像描述生成可以应用到图像检索、机器人问答、辅助儿童教育及导盲等多个方面，对图像描述生成的研究具有重要的现实意义。图像描述生成对于人工智能的发展同样具有重要的作用，相当于建立了计算机视觉和自然语言处理的桥梁。

1 相关工作

对于一张图片，图像描述生成方法能够让计算机自动地生成描述图片内容的语句。根据图像描述生成模型的不同，图像描述的方法主要分为三类：第一类是基于模板[2]方法，首先对图片中的物体、场景等信息进行识别，然后将对应的词汇填入到句子模板中。该方法生成的句子较为呆板，形式较为单一，准确率不高；第二类是基于检索[3-4]的方法，首先在训练数据库中检索和测试样本相似的图像，在将检索到的图像描述转移到待测试图像上，进而生成图像描述。该方法严重依赖训练数据库中的图像，无法生成比较新颖的图像描述内容。第三类是基于深度学习的方法，卷积神经网络作为编码器提取图像特征，循环神经网络（Recurrent Neural Network，RNN[5]）作为解码器生成图像描述。通过将二者优

[收稿日期] 2019－10－12
[基金项目] 大学生创新创业训练计划项目（S201910500074）
[第一作者] 周星光（1993－），男，湖北孝昌人，湖北工业大学硕士研究生，研究方向为图像描述生成
[通信作者] 靳华中（1973－），男，湖北洪湖人，湖北工业大学副教授，研究方向为图像处理

势结合形成端对端的方法,共同指导图像的描述生成。该方法能够生成描述更加准确的句子。基于深度学习的图像描述生成研究以来,Mao 等在文献[6]中提出的多模态循环神经网络(multimodal RNN,m-RNN)的方法广泛应用。m-RNN 将图像描述的工作分成两个任务:利用 CNN 提取图像特征,RNN 建立语言生成模型将图像特征转化成文本信息。m-RNN 中 CNN 使用 AlexNet[7] 网络结构,RNN 使用两层嵌入层将文本信息编码成 One-hot 向量表示,然后输入到循环层中,最后通过 Softmax 层得到输出。虽然 m-RNN 将 CNN 作为编码器引入到图像描述任务中,但因 RNN 网络结构限制,对于较长的网络系列易出现梯度消息的问题。Vinvals 等[8] 使用长短期记忆网络 LSTM 代替一般的 RNN,并且使用带有批标准层的 CNN 提取图像特征,图像描述准确率和速度均有提升。

从注意力模型命名方式看,很明显借鉴了人类的注意力机制。视觉注意力机制是人类视觉所特有的大脑信号处理机制。人类视觉通过快速扫描全局图像,获得需要关注的目标区域,也就是注意力焦点,然后对注意力焦点区域投入更多注意力资源,以获得更多所需要关注目标的细节信息,而抑制其他无用信息。这是人类利用有限的注意力资源从大量信息中快速筛选出高质量信息的方法,是人类在长期进化中形成的一种生存机制,人类视觉注意力机制极大地提高了视觉信息处理的效率和准确性。深度学习中的注意力机制从本质上和人类的选择性视觉注意力机制类似,核心目标也是从众多信息中选取出对当前任务目标更关键的信息。文献[9]将文献[10]注意力机制引入到图像描述生成,提出 hard-attention 与 soft-attention 模型,提高了模型的性能。文献[11]使用基于注意力的翻译模型,可以并行训练模型,提升了翻译性能。文献[12]提出了一种自上而下和自下而上相结合的注意力机制,提升了模型在视觉问答和图像描述生成的性能。

尺度是计算机视觉与图像处理领域的一个非常重要概念。任何一个视觉问题的答案都依赖于其所在的尺度。Lin 等[13] 将多尺度图像作为输入,产生了不同尺度的特征图,提高了语义分割的精度。文献[14-15]的图像表示方法可以在突出对象内容的同时刻画对象特征之间的空间关系,但是都没有考虑到不同尺度下物体的意义。文献[16-17]提出的空间金字塔池化方法,此方法通过不同尺度 bin 的采样,将局部特征进行聚合,bin 越大采样的范围越广,因此可以为图像表示提供不同尺度的空间信息。

生成图像描述句子的准确率主要受以下两个方面的影响:一是对图片中的物体及场景特征提取能力;二是对物体间相互关系等信息的提取。以上的文献都是基于 CNN 提取图像特征,但是 CNN 决定了提取特征尺度单一,提取的图像特征处理较为单一,没有考虑到提取的图像特征利用不充分的问题。本文提出在图像编码阶段,编码器随着网络的深度不断加深,图像特征层的尺度在不断减小,提取不同层的特征作为多尺度特征,融合不同层的特征得到多尺度特征,获得更丰富的图像特征。将多尺度融合特征和 CNN 最后一层的特征输入到循环神经网络中;在图像解码阶段,利用自适应注意力机制 LSTM 语言模型生成描述语句。

2 基于多尺度特征的图像描述生成模型

2.1 本文模型结构

本文采用编码器-解码器的图像描述生成模型结构,其中编码器利用卷积神经网络(VGG19)来提取图像特征信息,解码器利用循环神经网络(LSTM)生成描述性语言。本文提出改进后的模型总体结构见图 1。

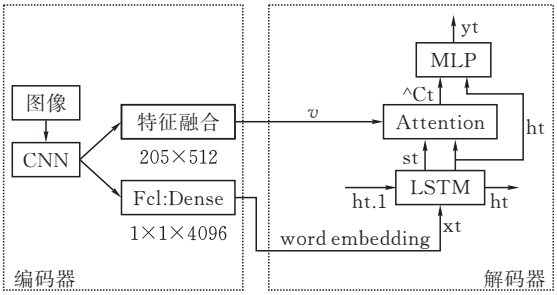


图 1 总体结构

2.2 基于多尺度图像特征提取的编码器

一幅图像中,只有在一定的尺度范围内,一个物体才有意义。例如,要观察一棵树,所选取的尺度应该是“米”级,重点关注树的形状而忽略树叶;如果要观察树叶,所选取的尺度应该是“厘米”,重点关注树叶而忽略树的形状;如果需要观察树叶的细胞结构,恐怕就需要“毫米”甚至“微米”级是必须的。图像中存在不同尺寸大小的对象目标,需要不同的尺度来提取图像特征。

本文基于 VGG19 提取不同层的特征图,从而提取到不同尺度的图像特征图进行特征融合,以增强对图像中不同尺度信息的提取。随着层数的增加,CNN 提取的图像特征具有更好的高层语义信息,因而选择提取靠近最后层的特征层。提取 Block5_conv2 层的 14X14X512 的特征向量和 Block5-pool 层的 7X7X512 特征向量,将这两层提

取到的特征向量进行 Concat,得到不同尺度融合特征向量。最后提取 FC1 层的 1X1X4096 特征向量。基于 VGG19 的不同层特征图提取的结构见图 2。将已训练好的包含 4096 维的特征和 205X512 不同层融合的特征作为图像描述模型的输入,导入到循环神经网络进行解码。

本文的模型关注图像的全局信息和多尺度融合信息,因而将卷积神经网络提取的 4096 维向量作为图像的全局特征,但是 4096 维的高维数据构成的特征在向量空间中表示,易造成数据稀疏的风险,因而对 4096 维向量进行降维处理。在模型的卷积神经网络输出阶段分别将 4096 维向量和 205x512 维向量映射到和文本相同的 256 维空间中。

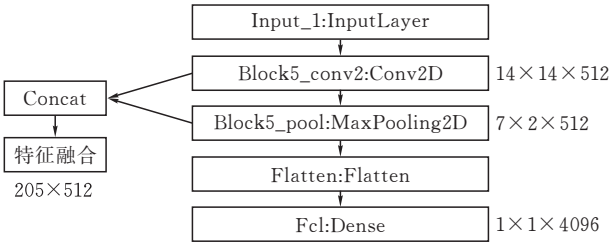


图 2　基于 VGG19 多尺度特征提取

2.3　基于自适应注意力 LSTM 的图像描述生成的解码阶段

使用 CNN+LSTM 网络进行图像内容描述的过程是一种编码-解码的过程。编码是使用 CNN 将图像映射为向量表示的过程,而解码是根据图像的特征,使用 LSTM 将特征转换为描述性语句的过程。

给定图像 I 和其对应的图像描述语句 X。首先使用 CNN 提取图像特征 $v(I)$。图像描述语句 $X=\{x_1,x_2,\cdots,x_L\}$,x_t 是语句中单词的表达形式,表示为 1-of-V(one-hot),其中,V 是训练字典库大小。在模型训练过程中,训练的目的是使图像特征与描述语义句子之间的映射关系最大化,即

$$\arg\max_{\theta}\sum_{i}\log p(X\mid I,\theta) \tag{1}$$

其中,θ 为模型参数,该参数是网络自学习的。由于每个图像的语义描述语句是由一系列单词组成,因此可以使用链式法则将其分解为

$$\log p(X\mid I,\theta)=\log p(x_1\mid I,\theta)+\sum_{t=2}^{L}\log p(x_t\mid I,x_{1,t-1},\theta) \tag{2}$$

可以使用 LSTM 求得 $t+1$ 时刻生成单词的概率分布,即

$$p_{t+1}=s(h_t) \tag{3}$$

$$h_t=L(\boldsymbol{W}x_t,\boldsymbol{U}h_{t-1},\mu\boldsymbol{C}v) \tag{4}$$

其中,$s(\cdot)$ 为 softmax 函数;$L(\cdot)$ 表示为 LSTM 网络;ht 为 LSTM 的隐藏层状态;$\boldsymbol{W},\boldsymbol{U},\boldsymbol{C}$ 为模型自学习的参数矩阵;x_t,h_{t-1} 分别为 LSTM 当前时刻的输入和上一时刻的隐藏层状态。

自适应注意力机制的主要功能是模型在生成句子描述时,模型可以自动选择关注图像的全局特征(4096)还是关注图像多尺度融合特征。自适应注意力在原有的 LSTM 基础上增加了两个公式:

$$g_t=\sigma(\boldsymbol{W}_x x_t+\boldsymbol{W}_h h_{t-1}) \tag{5}$$

$$s_t=g_t\odot\tanh(m_t) \tag{6}$$

其中 x_t 是 LSTM 的输入,m_t 是 memory cell。这里的 g_t 叫'哨兵'门,公式形式类似于 LSTM 中的输入门、遗忘门、输出门,决定了模型到底关注图像还是 visual sentinel;而 s_t 公式的构造与 LSTM 中的 ht=ot⊙tanh(ct)类似。

$$\hat{c}=\beta_t s_t+(1-\beta_t)c_t \tag{7}$$

自适应中的 Context Vector$\hat{c}$,βt∈[0,1]可以视为真正意义上的 sentinel gate,控制模型关注 visual sentinel 和 ct 的程度。与此同时,Spatial Attention 部分 k 个区域的 attention 分布 α_t 也被扩展成了 $\hat{\alpha}t$,做法是在 zt 后面拼接上一个元素:

$$\hat{\alpha}_t=\mathrm{softmax}([z_t;w_h^T\tanh(W_s s_t+(W_g h_t))]) \tag{8}$$

扩展后的 $\hat{\alpha}_t$ 有 k+1 个元素,而 βt=$\hat{\alpha}_t$[k+1]。最后生成单词的概率是:

$$p_t=\mathrm{softmax}(W_p(\hat{c}+h_t)) \tag{9}$$

本文选择 VGG19 最后一层卷积层的特征(1x1x4096)与 word embedding 拼接在一起成为 LSTM 的输入,多尺度的融合特征作为 attention 部分。

3　实验结果与分析

3.1　数据集与实验环境

本文数据集采用 MSCOCO2014019 数据集。数据集中包含了图像中所包含物体的类别、物体的轮廓坐标、边界框坐标以及对该图像内容的描述,其中每张图像的描述均至少有 5 种。本文的训练集、验证集、测试集,分别包含 113287、5000 和 5000 张图像。

实验环境为 Win10 环境下安装 tensorflow 1.60深度学习框架,配置 32 G 内存 AMD Ryzen 5 2600X Six-Core Processor 3.6GHz CPU,NVIDIA2070 GPU,NVIDIACUDA9.0 和 cuDNN7.5 深度学习库加速模型训练和测试,Python 环境为 Python3.7。

本文在图像编码阶段使用 VGG19 提取最后一层的全局信息(1×1×4096),将提取后 Block5_conv2 与 Block5-pool 的特征进行融合得到多尺度

图像特征。在编码阶段，采用自适应注意力机制 LSTM 网络生成自然语言。在模型训练阶段，采用 Adam 优化算法和 Dropout 方法，将 LSTM 中的单元按照一定的概率进行屏蔽来防止过拟合，实验中 Dropout 设置为 0.5，学习率为 0.01，batch 大小为 128。

3.2 评价指标与实验结果

现有的图像描述生成的评测标准包括人工主观抽检评价和客观量化评分。主观评价即人工观测输出图像，评定图像描述的质量。目前最普遍的客观量化评分方法包括：BLEU [18]、ROUGE_L、METEOR [19]、CIDEr [20]。本文实验也采用 BLEU，METEOR，CIDEr 进行评价。

针对上面的三个评价标准，在 MSCOCO 数据集上分别评估 BLEU，METEOR，CIDEr，评估结果见表 1。

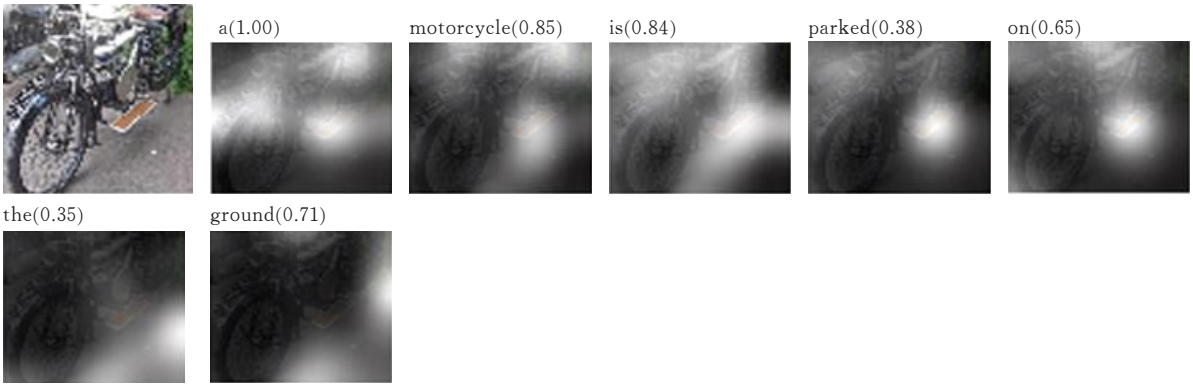
表 1 不同模型在 MSCOCO 数据集上的得分

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Google NIC	64.2	45.1	30.4	24.6	—	—
mRNN	67.0	47.6	32.3	25.8	23.7	—
Hard-Attentiom	70.2	48.9	34.1	26.5	24.1	—
VGG－LSTM	71.3	51.3	35.4	26.9	24.4	0.895
Our model	73.2	53.4	36.3	28.2	25.2	0.919

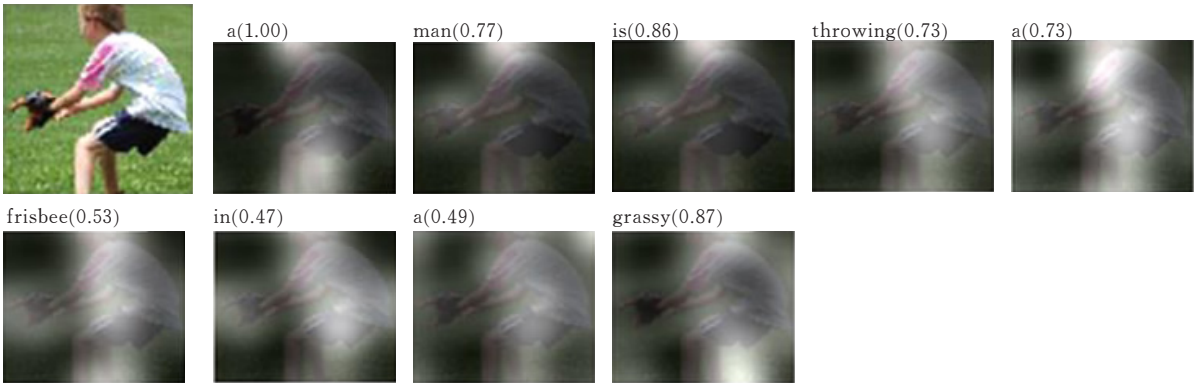
从评价结果来看，本文模型的各个指标均优于 Google NIC，mRNN，Hard-Attentiom 和 VGG－LSTM 模型。

3.3 结果分析

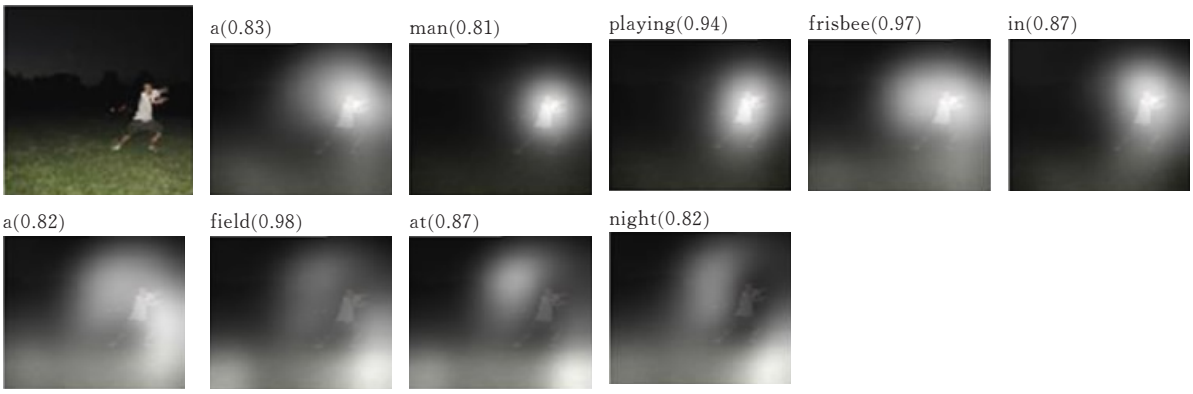
在测试结果中，图 3a、3b 和图 3d 中，本文模型相对于 Google NIC、mRNN 和 VGG-LSTM，模型更好的提取到目标特征信息；在图 3c 中，本文模型更好提取到图像背景信息，生成较为完整、准确的图像描述语句。从评价结果和测试结果来看，本文模型在各个评价指标上都有一定的提高，表明本文提出的模型对图像描述生成任务的有效性，同样表明本文多尺度融合特征更好的提取到图像信息。在图像编码阶段使用 VGG19 提取不同层的特征，得到不同尺度下图像中物体图像特征，融合得到多尺度融合特征，获取更丰富的图像信息以增强循环神经网络输入端的信息量。在图像描述生成阶段，语言描述模型自适应选择关注多尺度融合特征还是全局特征，从而生成更加完整、准确的句子。本文模型增加多尺度融合信息，可以更好地识别图中的对象，但是比较复杂的场景还没法达到较为准确的结果。



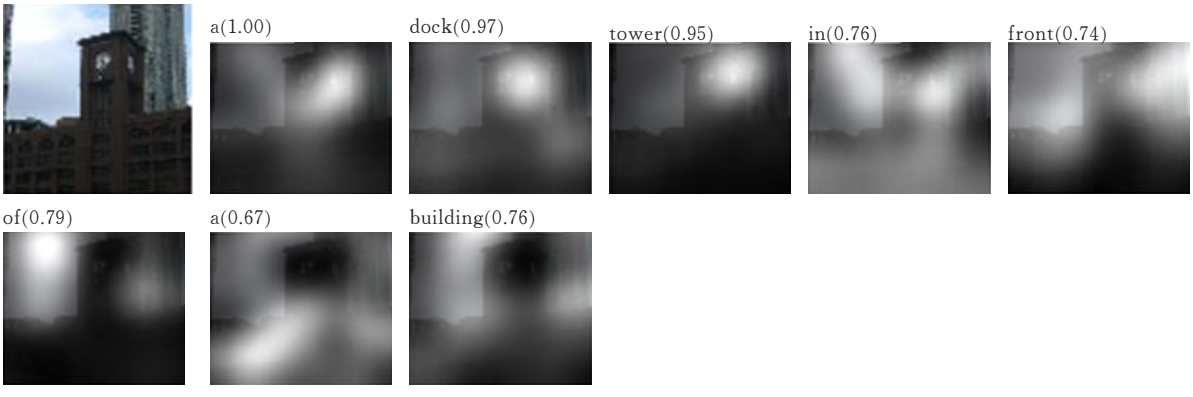
(a)(Google NIC) a motorcy parked on the ground(our model) a motorcycle is parked on the ground



(b)(mRNN) Man is throwing frisbee in grassy(our model) a man is throwing a frisbee in a grassy



(c)(Hard attention) man playing frisbee in a field(our model) a man playing frisbbee in a field at night



(d)(VGG-LSTM) dock tower in front of a building(our model) a dock tower in front of a building

图 3　测试结果

4　结束语

本文采用编码器—解码器结构的图像描述生成方法。针对现有图像描述生成中卷积神经网络提取单一尺度图像特征的不足，图像信息利用不充分，造成文字对图片内容描述的不够准确、语义较模糊。本文改进现有模型对图像特征的编码形式，提出了基于 VGG19 网络提取不同层的特征进行融合得到多尺度特征，获取更丰富的图像信息。在解码器阶段，基于自适应注意力机制 LSTM 网络生成图像描述语句。本文提出的模型在 MSCOCO 数据集上进行模型训练和测试，实验结果表明：本文模型很好融合了 CNN 不同层的特征，获取更丰富的图像信息，增强了语言模型输入的信息，自适应注意力 LSTM 网络模型生成更准确完整，更有意义的图像描述语句。

[参 考 文 献]

[1] Lecun Y , Bengio Y , Hinton G . Deep learning[J]. Nature, 2015, 521(7553):436.

[2] Fang H , Gupta S , Iandola F, et al. From captions to visual concepts and back[C]// 2015 IEEE Conference on Computer Vision And Pattern Recognition (CVPR). IEEE, 2015.

[3] Kuznetsova P, Ordonez V, Berg A C, et al. Collective generation of natural image descriptions [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 359-368.

[4] Kuznetsova P, Ordonez V, Berg T L, et al. Treetalk: composition and compression of trees for image descriptions [J]. Transactions of the Association for Computational Linguistics, 2014(2): 351-362.

[5] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities [J]. Proceedings of the national academy of sciences, 1982, 79(8): 2554-2558.

[6] Mao J, Xu W, Yang Y, et al. Explain images with multimodal recurrent neural networks [EB/OL]. [2018-6-10]https://arxiv.org/pdf/1410.1090v1.pdf

[7] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]//NIPS2012: Proceedings of the 2012 International Conference on Neural Information Processing Systems. Nevada, USA: Curran Associates Inc. 2012: 1097-1105.

[8] Vinyals O, Toshev A, Bengio S, et al. Show and tell:

A neural image caption generator[C]//CVPR2015: Proceedings of the 2015 International Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 3156-3164.

[9] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [J]. Computer Science, 2015: 2048-2057.

[10] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2018-06-10]https://arxiv.org/pdf/1409.0473.pdf.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// NIPS2017: Proceedings of the 2012 International Conference on Neural Information Processing Systems. Long Beach, USA. 2017: 6000-6010.

[12] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6077-6086.

[13] Lin G S, Shen C H, van den Hengel, et al. Efficient piecewise training of deep structured models for semantic segmentation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA, New York: IEEE, 2016: 3194-3203

[14] Kalantidis Y, Mellina C, Osindero S. Cross-dimensional weighting for aggregated deep convolutional features [C]// Proc of European Conference on Computer Vision. Amsterdam: IEEE press, 2016: 685-701.

[15] Pan Xingang, Shi Jianping, Luo Ping, et al. Spatial as deep: Spatial cnn for traffic scene understanding [C]// The AAAI Conference on Artificial Intelligence. New Orleans: AAAI press, 2018: 7276-7683.

[16] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9): 1904-16.

[17] Jose A, Lopez R D, Heisterklaus I, et al. Pyramid Pooling of Convolutional Feature Maps for Image Retrieval [C]// IEEE International Conference on Image Processing. Athens: IEEE press, 2018: 480-484.

[18] Papineni K, Roukos S, Ward T, Zhu W J. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2012: 311-318.

[19] Banerjee S, Lavie A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [C]//Proceedings of the aclWorkshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2015: 65-72.

[20] Lin C Y. Rouge: a package for automatic evaluation of summaries[C]//Proceedings of the ACL-04 Workshop on Text Summarization Branches Out, Barcelona, 2004: 74-81.

An Image Description Generation Model Based on Multi-scale

ZHOU Xingguang, JIN Huazhong, XU Yudong, LI Qingqing, HU Man

(School of Computer Science, Hubei Univ. of Tech., Wuhan 430068, China)

Abstract: Aiming at the existing model based on deep learning image description, in the image feature encoding stage, the image features extracted by the encoder are relatively simple and the image information is not fully utilized, which causes inaccuracy in describing the content of the image of the text and fuzziness of the semantics. Based on VGG19, this paper improves the coding pattern of image features of existing models, and extracts and fuses image multi-scale feature methods to obtain more abundant image information. The method in this paper is trained and tested on the MSCOCO dataset. The experimental results show that the proposed model can generate more accurate, complete and meaningful image description statements.

Keywords: image description generation; deep learning; multi-scale; image features

[责任编校: 张岩芳]