计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于改进注意力机制的图像描述生成算法
作者：李文惠，曾上游，王金金
收稿日期：2020-07-23
网络首发日期：2020-10-20
引用格式：李文惠，曾上游，王金金．基于改进注意力机制的图像描述生成算法．计算机应用. https://kns.cnki.net/kcms/detail/51.1307.TP.20201020.1101.006.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

doi:10.11772/j.issn.1001-9081.2020071078

基于改进注意力机制的图像描述生成算法

李文惠，曾上游*，王金金

(广西师范大学 电子工程学院，广西 桂林 541004)

(*通信作者电子邮箱 zsy@mailbox.gxnu.edu.cn)

摘 要: 图像描述是将图像所包含的全局信息用语句来表示。它要求图像描述生成模型既能提取出图像信息，又能将提取出来的图像信息用语句表达出来。传统的模型是基于卷积神经网络(CNN)和循环神经网络(RNN)搭建的，在一定程度上可以实现图像转语句的功能，CNN-RNN 模型在提取图像关键信息时精度不高且训练速度缓慢。针对这一问题，提出了一种基于卷积神经网络和长短期记忆网络(LSTM)改进的注意力机制图像描述生成模型，采用 VGG19 和 RESNET101 作为特征提取网络，在注意力机制中引入分组卷积替代传统的全连接操作，提高评价值指标。该模型使用了公共的数据集 Flickr8K，Flickr30K 来训练，采用多种评价指标(BLEU、ROUGE_L、CIDEr、METEOR)对模型进行验证，实验结果表明，与引入传统的注意力机制相比，提出的改进注意力机制图像描述生成模型对图像描述任务的准确性有所提升，其在 5 种评价指标中均优于传统的模型。

关键词: 图像描述；自然语言处理；卷积神经网络；长短期记忆神经网络；注意力机制

中图分类号:TP391 **文献标志码：A**

Image description generation algorithm based on improved attention mechanism

LI Wenhui, ZENG Shangyou*, WANG Jinjin

(School of Electronic Engineering, Guangxi Normal University, Guilin Guangxi 541004, China)

Abstract: Image description was to express the global information contained in the image in sentences. It required that the image description generation model can not only extracted image information, but also expressed the extracted image information in sentences. The traditional model was based on Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), which can realize the function of image-to-sentence translation to a certain extent. The CNN-RNN model has low accuracy and training speed when extracting key information of the image slow. To solve this problem, an improved attention mechanism image description generation model based on convolutional neural network and long-short-term memory network (LSTM) was proposed. VGG19 and RESNET101 were used as the feature extraction network, and grouped convolution was introduced in the attention mechanism to replace the traditional fully connected operation and improve the evaluation index .The model was trained by public datasets Flickr8K and Flickr30K and validated by a variety of evaluation indicators (BLEU, ROUGE_L, CIDEr, METEOR). Compared with the introduction of the traditional attention mechanism, the proposed improved attention mechanism image description generation model has improved the accuracy of the image description task, and it is better than the traditional model in the four evaluation indicators.

Keywords: Image description; Natural language processing; Convolutional neural network; Long-short-term memory neural network; Attention mechanism

0 引言

图像描述是将一张图像用自然语言句子表达出来，它是计算机视觉的主要研究任务之一。图像描述对于计算机而言不仅需要识别图像中的对象，而且还要理解图像中的内容以及对象之间存在的关系，最后计算机还要用自然语言句子去将图像内容正确地表达出来。因此，图像描述任务对于计算机视觉领域的研究来说还是存在一定的难度。目前图像描述存在的问题主要包括图像分类问题和自然语言处理问题。针对图像分类问题，卷积神经网络通过自动提取图像特征，使图像分类的准确率达到甚至超过了人类肉眼对图像分类识别的标准；针对自然语言处理问题，循环神经网络通过记住句子中词的相对关系，去处理自然语言句子。然而对于上述两者问题的结合而言，虽然目前存在相关网络能够在一定程度上简单的描述图像，但没有在各自领域研究的那么深入。实现图像描述的方法主要分三种：基于模板的图像描述生成方法，基于检索的图像描述生成方法和基于深度学习的图像描述生成方法。近年来，图像描述主流方法是深度学习。深度学习模型的训练方式是端到端，其优点是它可以自己学习特

收稿日期: 2020-07-23; 修回日期:2020-10-06; 录用日期: 2020-10-14。
基金项目: 国家自然科学基金资助项目(11465004)。
作者简介: 李文惠（1997—），女，湖南衡阳人，硕士研究生，主要研究方向：深度学习、人工智能； 曾上游(1974—),男，湖南双峰人，教授，博士，主要研究方向：神经网络与人工智能、复杂网络、生物信息处理和生物芯片； 王金金（1995—），女，安徽安庆人，硕士研究生，主要研究方向：深度学习、人工智能。

征，避免了人为的去设计参数。对于图像描述生成模型，整体大致分为两个部分：编码(ENCODER)和解码(DECODER[1])。在图像编码中，通过多层深度卷积神经网络[2,3,4]针对图像中的物体特征建立起模型；在图像解码中，通过循环神经网络针对文本信息建立起模型。运用循环神经网络[5,6]将文本信息与图像信息映射在同一个空间中，利用图像信息引导文本句子生成。随着深度学习研究的不断深入，强化学习[7,8]和基于注意力机制[9,10]的研究方法相继涌现。该方法对模板、规则的约束少，能自动推断出测试图像及其相对应的文本，自动地从大量的训练集中学习图像和文本信息，生成更灵活、更丰富的图像描述句子，还能描述从未见过的图像内容特征。本文中引入改进的注意力机制，不仅可以减少模型参数，而且能更准确地生成描述图像的自然语言句子和提升图像描述生成模型的评价指标。

1 相关工作及本文方法

首先简单介绍有关图像描述生成和注意力机制先前工作的背景。2014 年 Vinyals 等人在文献[1]提出了一个基本的卷积神经网络(Convolutional Neural Networks, CNN)联合循环神经网络(Recurrent Neural Networks, RNN)的图像描述框架，在图像描述的领域中，取得了巨大的突破，同时也提出了评价图像描述生成模型性能的指标，但是依然没有考虑到词对应图像位置这一缺陷，基于此问题，2016 年 Xu 等人[11]从人的视觉上受到启发，在上述的框架中引入了注意力机制，使得计算机描述图像更加符合人类的描述机制，在指标上也得到相应的提升，同时也验证了注意力机制的可行性。上述所说的基于深度学习的描述算法虽能产生描述图像的自然语言句子，但总体上有一定的局限性，如参数过多，注意力还有很大的提升空间。

在本文中，提出了一种基于卷积神经网络和长短期记忆元的图像描述生成，并引入改进的注意力机制的模型。改进的注意力机制是在文献[11]的基础模型上改进的，改进的点是将原全连接层替换成了文中注意力机制(ATTENTION)，全连接层不仅参数多而且关注很多无用的信息，造成信息冗余，文中引入注意力机制的结构能有效地避开这些问题。本文提取图像特征采用了两种卷积神经网络，分别是 VGG(Visual Geometry Group)和 RESNET(Residual Network)，解码采用长短期记忆(Long Short-Term Memory, LSTM)网络[12]，同时引入改进的注意力机制，最终生成图像描述的自然语言句子，能够有效提升图像中的内容与句子描述的相关联度，同时图像描述的相关评估指标有所提升，生成更接近人类语言的图像描述自然语言句子。

2 模型

模型分为两个模块，ENCODER 模块和 DECODER 模块。ENCODER 模块采用卷积神经网络，其功能在于提取图像的特征，对图像进行编码，将图像编码为特征向量，DECODER 模块是将编码后的图像解码成自然语句，它主要通过长短期记忆网络解码图像信息，其功能是提取句子单词之间的句法特征，依据选择的图像特征生成图像描述的自然语言句子。本文使用 CNN+LSTM+ATTENTION 的基本框架来完成[13]。将图像输入到卷积神经网络中，得到网络输出的特征向量，文本的词通过嵌入(EMBEDDING)层将词转成词向量，将特征向量和词向量拼接后输入到长短期记忆单元，产生新的预测词，通过集束搜索(Beam Search)的方式产生预测的句子。模型整体结构如图 1 所示。

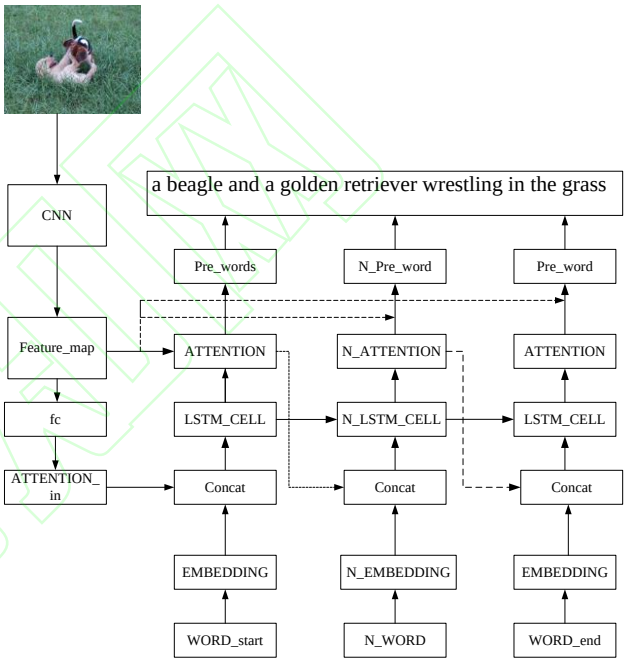


图 1 模型整体结构

Fig. 1 Overall model structure

ENCODER 模块采用的是 VGG19 网络和 RESNET101 网络，VGG19 网络是使用 3×3 卷积核的卷积层堆叠并交替最大池化层， VGG 网络的一大优点是简化了神经网络结构，本文选取 VGG19 网络中最后一个最大池化层的输出特征图，再加一个 1×1 卷积使得 VGG19 和 RESNET101 的输出特征图维度相同，1×1 卷积输出的特征图经自适应池化层后，得到的自适应特征图作为整个网络中的 ENCODER 模块输出特征图。VGG19 只有 19 层， ResNet101 有 101 层，它们在网络深度上完全不是一个量级，RESNET101 可以使用一个称为残差模块的标准网络组件来组成更复杂的网络，网络加深的同时也保持了网络的性能，解决了深度网络的退化问题，本文选取 RESNET 101 网络平均池化层的输入特征图，将经自适应池化层后的特征图作为整个网络中的 ENCODER 模块输出特征图。 DECODER 模块采用的是长短期记忆网络，该网络可以连接先前的信息到当前的信息上，语句的预测是和词

的先前信息有一定的关联的，而长短期记忆网络适合处理这类时间序列问题[14]。

本文引入分组注意力机制，结构如图 2 所示，Encoder_out 是卷积神经网络输出的特征图，大小为 2048×14×14，Decoder_hidden 是 LSTM 的隐藏输出，大小为 512×1×1。本文设计的是分组卷积注意力，通过 1×1 卷积(Conv_1×1)分别整合图片特征和词特征，用激活函数(ReLU)将整合的特征引入非线性，得到激活特征并将其分成两组卷积，分别是 3×3 卷积(Conv_3×3)和 1×1 卷积(Conv_1×1)，且都使用激活函数(ReLU)引入非线性，再拼接输入到线性层(linear)中，通过 Softmax 函数得到图像和词的关联度，进而形成新的注意力分布。通过分组的特征注意力，可以更加合理地分布原图和词对应的注意力，新的注意力分布与输入的图像相乘，得到词对应图像的注意图(Attention_feature)。

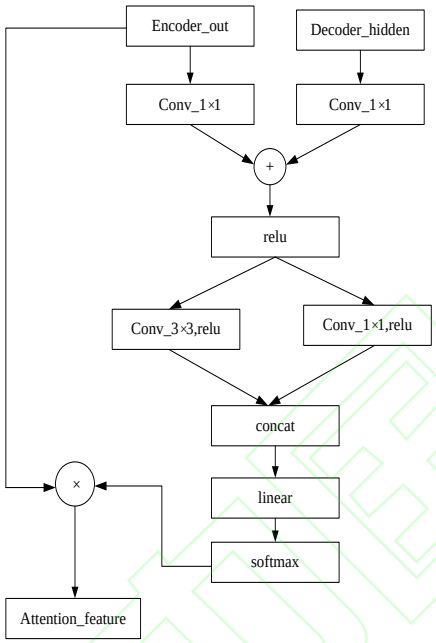


图 2 改进的注意力机制

Fig. 2 Improved attention mechanism

3 实验设置

实验环境：本实验使用 pytorch 作为深度学习底层框架，计算机内存为 32GB RAM 、英特尔 i7-6700K 四核八线程 CPU 以及 NVIDIA-GTX1080Ti GPU，操作系统为 windows 10 64 位。

3.1 评价指标

本文使用了多种评价指标，BLEU[15](Bilingual Evaluation Understudy)， CIDEr[16](Consensus-based Image Description Evaluation)， ROUGE-L[17](Recall-Oriented Understudy for Gisting Evaluation) 和 METEOR[18](Metric for Evaluation of Translation with Explicit Ordering)。与此同时，本文列出了上述评价指标的计算公式。

3.1.1 BLEU

$$p_n=\frac{\sum_i\sum_k \min(h_k(c_i),\max_{j\in m}h_k(s_{ij}))}{\sum_i\sum_k \min(h_k(c_i))} \tag{1}$$

BLEU 是比较候选译文和参考译文里的 *n-gram* 的重合程度，重合程度越高就认为译文质量越高。公式(1)中，$\max_{j\in m}h_k(s_{ij})$ 表示某 *n-gram* 在多条标准答案中出现最多的次数，$\sum_i\sum_k \min(h_k(c_i),\max_{j\in m}h_k(s_{ij}))$ 表示取 *n-gram* 在翻译译文和标准答案中出现的最小次数。

$$BP=\begin{cases}1, & lc>ls\\ \exp(1-ls/lc), & lc\le ls\end{cases} \tag{2}$$

公式(2)中，*BP* 表示长度惩罚因子，*lc* 表示翻译译文的长度，*ls* 表示参考答案的有效长度，当存在多个参考译文时，选取和翻译译文最接近的长度。当翻译译文长度大于参考译文的长度时，惩罚系数为 1，表示不惩罚，只有机器翻译译文长度小于参考答案才会计算惩罚因子。

$$BLEU=BP*\exp\left(\sum_{n=1}^{N}W_n\log P_n\right) \tag{3}$$

由于各 *N-gram* 统计量的精度随着 *gram* 阶数的升高而呈指数形式递减，所以为了平衡各阶统计量的作用，公式(3)中，对其采用几何平均形式求平均值然后加权，再乘以长度惩罚因子，得到最后的评价公式，*N* 的上限取值为 4，即最多只统计 *4-gram* 的精度。

3.1.2 ROUGE_L

$$ROUGE-L=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2P_{lcs}} \tag{4}$$

$$R_{lcs}=LCS(X,Y)/m \tag{5}$$

$$P_{lcs}=LCS(X,Y)/n \tag{6}$$

ROUGE-L 计算的是候选摘要与参考摘要的最长公共子序列长度，长度越长，得分越高。其中，*X* 表示候选摘要，*Y* 表示参考摘要，*LCS*(Longest Common Subsequence)表示候选摘要与参考摘要的最长公共子序列的长度，*m* 表示参考摘要的长度，*n* 表示候选摘要的长度，R_{lcs} 和 P_{lcs} 分别表示召回率和准确率。

3.1.3 CIDEr

$$CIDEr_n(c,S)=\frac{1}{M}\sum_{i=1}^{M}\frac{\boldsymbol{g}^n(c)\bullet\boldsymbol{g}^n(S_i)}{\|\boldsymbol{g}^n(c)\|*\|\boldsymbol{g}^n(S_i)\|} \tag{7}$$

公式(7)中，*c* 表示候选标题，*S* 表示参考标题集合，*n* 表示评估的是 *n-gram*，*M* 表示参考字幕的数量，$\boldsymbol{g}^n(\cdot)$ 表示基于 *n-gram* 的 TF-IDF(Term Frequency–Inverse Document

Frequency)向量。CIDEr 是把每个句子看成文档，然后计算其 TF-IDF 向量的余弦夹角，据此得到候选句子和参考句子的相似度。

3.1.4 METEOR

$$METEOR = (1 - pen) \times F_{means} \tag{8}$$

$$F_{means} = \frac{PR}{\alpha P + (1-\alpha) R} \tag{9}$$

$$P = m / c \tag{10}$$

$$P = m / \mathrm{r} \tag{11}$$

α 为可调控的参数，m 为候选翻译中能够被匹配的一元组的数量，c 为候选翻译的长度，r 为参考摘要的长度。公式(8)中，*pen* 为惩罚因子，惩罚的是候选翻译中的词序与参考翻译中的词序的不同。

3.2 实验数据集

本次实验采用了 Flickr8K[19]和 Flickr30K[20]数据集，两个数据集都是一张图片对应 5 句描述自然语言句子，Flickr8K 数据集约 8 000 幅图像和 Flickr30K 约 30 000 幅图像，这两个数据集中的图像都是针对特定对象和动作的。如图 3 所示。

图 3 图像对应的自然语言句子

Fig. 3 Natural language sentences corresponding to images

3.3 参数设置

本文设置的词嵌入维度是 512，LSTM 的输出维度为 512，输入数据的 batch size 为 32。微调卷积神经网络，训练网络时，卷积神经网络的学习率设置为 1e-4，长短期记忆网络学习率设置为 4e-4。整个网络采用 Adam 优化器训练，防止反向传播梯度爆炸，如果连续 8 个 epoch 评价指标都没有改善，则学习率降低为原来的 0.8，并在 20 个 epoch 后终止训练，实验时在反向传播中加入了梯度截断，可以有效地避免梯度爆炸。损失函数使用的是交叉熵损失函数。在测试中使用集束搜索(beam search)的方式，假设词汇表关联词汇 beam size 的大小为 5。

3.4 实验结果及分析

在 Flickr8K 和 Flickr30K 两个数据集的比对实验中，数据集使用的是公共划分标准[21]，使用数据集中的 1000 张图像进行验证，1000 张图像进行测试，其余用于训练。根据文献[21]可知数据集拆分的差异不会对整体性能产生实质性的影响。传统的 CNN+LSTM 网络和本文所使用的 CNN+LSTM+ATTENTION 网络在上述的两个数据集上做对比实验，对图像描述的各项指标如表 1 所示。

本文提出的注意力机制是通过对图像的特征和词的特征分组卷积，得到不同的注意力，再经过线性层整合这些不同的注意力，生成一个图像和词相关联的新注意力分布，将提出的注意力机制嵌入到传统的模型中，能更加准确地生成描述图像的自然语言句子。因此当选取的卷积神经网络为 VGG19 时，VGG19+LSTM+ATTENTION 比 VGG19+ LSTM 网络在指标上都有所提高，引入分组注意力的模型比传统模型的 BLEU_4 提升了 1.08%，ROUGE_L 提升了 0.91 %，CIDEr 提升了 3.06 %。

表 1 不同模型在 Flickr8K 数据集的 BLEU-1,4/METEOR/ROUGE_L/CIDEr 评价指标对比　　单位：%

Tab. 1 Comparison of BLEU-1, 4/METEOR/ROUGE_L/CIDEr evaluation indicators of different models in Flickr8K dataset　　unit: %

ENCODER	DECODER	BLEU_1	BLEU_4	ROUGE_L	CIDEr
VGG19	LSTM	61.39	20.03	45.32	47.89
VGG19	LSTM+ATTENTION	61.95	21.11	46.23	50.95
RESNET101	LSTM	63.01	21.12	46.36	52.23
RESNET101	LSTM+ATTENTION	64.68	23.06	48.15	58.36

从表 1 可知，当卷积神经网络为更深更复杂的 RESNET101 网络时，RESNET101+LSTM 网络在各评价指标已经高于 VGG19+LSTM+ATTENTION 和 VGG19+ LSTM 网络。在引入改进 ATTENTION 的 RESNET101+LSTM 网络之后，比 RESNET101+LSTM 网络的评价指标有更加明显的提高，特别是，BLEU_4 和 CIDEr 分别提升了 1.94%和 6.13 %。在 Flickr8K 数据集上引入注意力机制的 VGG 网络和 RESNET 网络，通过各项指标的比较，验证了本文提出的注意力机制的可行性和高效性。为了进一步验证改进的注意力机制的高效性，在数据集 Flickr30K 上做了相同的对比实验，实验结果如表 2 所示。传统的模型没有考虑到词和图片位置的关系，而本文提出的改进的注意力机制，使模型能够关注到词和图像的对应位置，更加符合人类的肉眼观察机制，在较大的 Flickr30K 数据集中，通过引入改进的注意力机制模

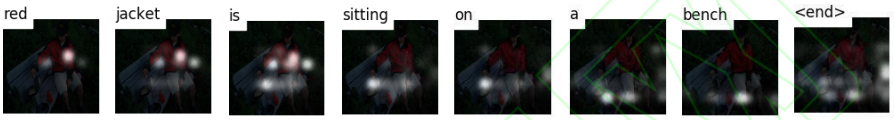
型和传统编解码模型这两种模型的对比，RESNET101 网络的各项指标比 VGG19 网络有更为突出的效果，在该数据集上，引入改进的注意力机制 RESNET101 网络和 VGG19 网络在评价指标 BLEU_4 上各提升了 4.91%和 4.71%。

表 2 不同模型在 Flickr30K 数据集的 BLEU-1,4/METEOR/ROUGE_L/CIDEr 评价指标对比　　单位：%

Tab. 2 Comparison of BLEU-1, 4/METEOR/ROUGE_L/CIDEr evaluation indicators of different models in Flickr30K dataset　　unit: %

ENCODER	DECODER	BLEU_1	BLEU_4	ROUGE_L	CIDEr
VGG19	LSTM	61.77	17.89	42.46	36.72
VGG19	LSTM+ATTENTION	64.95	22.60	45.11	44.69
RESNET101	LSTM	61.94	20.01	43.68	41.61
RESNET101	LSTM+ATTENTION	66.67	24.92	46.79	49.87

(a) 开始　(b) 一个　(c) 男人　(d) 穿着　(e) 一件

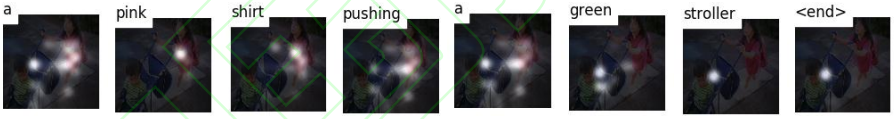


(f)红色　(g) 夹克　(h) 正　(i) 坐　(j) 在　(k) 一条 (l) 长椅上　(m) 结束

图 4 Flickr8K 中单词对应的注意力热力图

Fig. 4 Attention heat map corresponding to words in Flickr8K

(a) 开始　(b) 一个　(c) 小　(d) 女孩　(e) 穿着



(f) 一件 (g) 粉色的　(h) 短袖　(i) 推着　(j) 一辆　(k) 绿色的 (l) 婴儿车 (m) 结束

图 5 Flickr30K 中单词对应的注意力热力图

Fig. 5 Attention heat map corresponding to words in Flickr30K

在 Flickr8K 数据集和 Flickr30k 数据集中各自随机选取一张图像，并可视化描述语句对应该图片的注意力分布图，如图 4 和图 5 所示。

改进的注意力模型根据对语句中当前单词和图像关注到接下来需要描述的图像部分，将局部注意力映射到原图中，模型中分支的 3×3 卷积和 1×1 卷积可以分别关注词对应的不同局部特征，再连接分支的不同局部特征输入到全连接后，得到词对应多个存在关联的局部特征区域即注意力分布，不仅有效地减少特征的冗余，而且得到多个局部注意力特征。

表 3 本文算法与其他模型算法 BLEU_1,2,3,4/METEOR 评价指标对比　　单位：%

Tab. 3 Comparison of the algorithm in this paper with other model algorithms BLEU_1,2,3,4/METEOR evaluation index　　unit: %

模型	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR
Google NIC[1]	66.3	42.3	27.7	18.3	—
SCA-CNN-VGG[9]	64.6	45.3	31.7	21.8	18.8
Hard-Attention[11]	**66.9**	43.9	29.6	19.9	18.49
双向单注意力网络[22]	64.3	45.3	31.1	20.3	18.61
双向双注意力网络[22]	65.8	46.9	33.2	23.1	19.44
VGG+LSTM+ATTE	64.95	46.26	32.46	22.60	20.60
RESNET+LSTM+ATTE	66.67	**48.52**	**34.83**	**24.92**	**21.55**

表 3 中 Google NIC 模型是首次提出图像描述生成的编码-解码基本框架，图像描述生成任务中引入这样的架构已成为主流。注意力机制的基本思想是利用卷积层获取图像特征后，对图像特征进行注意力加权，之后再送入 RNN 中进行解码，表 3 中的 SCA-CNN-VGG(Spatial and Channel-wise Attention in Convolutional Neural Networks)模型是用通道注意力和空间注意力结合的方式来做图像描述生成，Hard-Attention 是即将图像中最大权重置为 1，而将其他区域权重置 0，以达到仅注意一个区域的目的，双向单注意力网络和双向双注意力网络都是近年对注意力较新的改进，Attention 机制已经成为一种主流的模型构件。

由表 3 可知，有注意力机制的模型相比 Google NIC 都有比较明显的提升，ENCODER 模块是使用相同的卷积神经网络，DECODER 模块是使用相同的长短记忆元网络，保证了实验的合理性和公平性。本文提出的改进注意力机制通过分组卷积注意力，更合理的分布原图和词对应的注意力。相对于其他的注意力模型，进一步的提升了准确率，说明本文改进的注意力机制能更有效的筛选有用特征作为长短记忆元网络的输入，表 3 中所有的模型都在 Flickr30k 数据集上验证，表明本文改进的模型有较好的泛化性。随机选取 Flickr8K 数据集和 Flickr30K 的示例图分别为图 6 和图 7，对比传统模型和改进模型对图像描述生成。

图像描述语句

语句1: A man in a red jacket is sitting on a bench whilst cooking a meal .
语句2: A man is sitting on a bench , cooking some food .
语句3: A man sits on a bench.
语句4: A man wearing a red jacket is sitting on a wooden bench and is cooking something in a small pot .
语句5: a man wearing a red jacket sitting on a bench next to various camping items .

图 6 Flickr8K 示例图片对应的自然语言句子

Fig. 6 Natural language sentences corresponding to Flickr8K sample pictures

传统模型(RESNET101+LSTM)生成的自然语言句子： a man in a blue jacket is sitting on a wooden bench.改进模型(RESNET101+LSTM+ATTENTION)生成的自然语言句子： a man in a red jacket is sitting on a bench. 传统模型将图片中的红色夹克信息生成了错误的蓝色夹克信息，而改进模型准确的生成了红色夹克信息。

图像描述语句

语句1: A little girl wearing a pink shirt and backpack is pushing a little boy wearing a green shirt in a blue stroller .
语句2: A little girl in a pink outfit pushing a little boy in a green outfit in a stroller .
语句3: A girl in a pink outfit pushing a blue stroller with a boy sitting in it .
语句4: Little girl with backpack pushing brother in stroller .
语句5: A child pushing their younger sibling in a stroller .

图 7 Flickr30K 示例图片对应的自然语言句子

Fig. 7 Natural language sentences corresponding to Flickr30K sample pictures

传统模型(RESNET101+LSTM)生成的自然语言句子： a little girl in a pink shirt is playing with a hula hoop.改进模型(RESNET101+LSTM+ATTENTION)生成的自然语言句子：a little girl in a pink shirt pushing a green stroller. 传统模型对 Flickr30K 示例图片中生成了错误的呼啦圈信息，而改进模型准确的生成绿色的婴儿推车信息。

在 Flickr8K 数据集和 Flickr30K 数据集中，RESNET101+LSTM 生成的语句中存在一些错误，翻译的不是很准确，而 RESNET101+LSTM+ATTENTION 模型能较为准确的翻译图片内容，且基本没有语法错误。

4 结语

本文提出了一种基于 CNN 和 LSTM 且引入了改进的注意力机制的网络模型。采用的经典 VGG19 网络以及具有更深层的 RESNET101 网络对图像进行特征编码，通过用 EMBEDDING 对自然语言句子的词进行词编码进而得到词向量，经 LSTM 将特征向量和词向量映射到同一空间中，在引入改进的注意力机制作用下，使图像信息引导生成与图像更加符合的自然语言句子，同时也提升了本文所提出的模型的鲁棒性。实验结果表明，本文提出的模型泛化能力明显更好一些，在图像描述生成的自然语言句子和评价指标上都优于传统的模型。

参考文献

[1] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 3156 -3164.

[2] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2016: 770-778.

[3] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2014.

[4] WEI Y, XIA W, LIN M, et al. HCP: A Flexible CNN Framework for Multi-Label Image Classification [J]. IEEE Transactions on Software Engineering, 2016, 38(9):1901-1907.
[5] SOCHER R, KARPATHY A, LE Q V, et al. Grounded compositional semantics for finding and describing images with sentences [J]. Transactions of the Association for Computational Linguistics, 2014, 2(1): 207-218.
[6] CHEN X, ZITNICK C L. Mind's eye: A recurrent visual representation for image caption generation [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 2422-2431.
[7] GAO J, WANG S, WANG S, et al. Self-Critical N-Step Training for Image Captioning [C]// Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2019: 6300-6308.
[8] LU J, XIONG C, PARIKH D, et al. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017: 3242-3250.
[9] CHEN L, ZHANG H, XIAO J, et al. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017: 6298-6306.
[10] 陈龙杰, 张钰, 张玉梅, 等. 基于多注意力多尺度特征融合的图像描述生成算法[J]. 计算机应用, 2017, 39(2): 354-359. (CHEN L J, ZHENG Y, ZHANG Y M, et al. Image caption algorithm based on multi-attention and multi-scale feature fusion [J]. Journal of Computer Applications, 2019, 39(2): 354-359.)
[11] 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.
[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[13] 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. (HUANG Y W, YOU Y D, ZHAO P. Image caption generation model with convolutional attention mechanism [J]. Journal of Computer Applications, 2020, 40(1): 23-27.)
[14] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1-6, 26. (YANG L, WU Y X, WANG J L, et al. Research on recurrent neural network [J]. Journal of Computer Applications, 2018, 38(S2): 1- 6, 26.)
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] HODOSH M, YOUNG P, HOCKENMAIER J, et al. Framing image description as a ranking task: data, models and evaluation metrics [J]. Journal of Artificial Intelligence Research, 2013, 47(1): 853-899.
[20] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions [J]. Transactions of the Association for Computational Linguistics, 2014, 2(1): 67-78.
[21] KARPATHY A, L. FEI-FEI. Deep visual-semantic alignments for generating image descriptions [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 3128-3137.
[22] 陶云松, 张丽红. 基于双向注意力机制图像描述方法研究[J]. 测试技术学报, 2019, 33(4): 347-350. (TAO Y S, ZHANG L H. Research on Image Description Method Based on Bidirectional Attentional Mechanism [J]. Journal of test and measurement technology, 2019, 33 (4):347-350.)

国家自然科学基金资助项目(11465004)

李文惠(1997-)，女，湖南衡阳人，硕士研究生，主要研究方向：深度学习、人工智能；曾上游(1974-)，男，湖南双峰人，教授，博士，主要研究方向：神经网络与人工智能、复杂网络、生物信息处理和生物芯片；王金金(1995-)，女，安徽安庆人，硕士研究生，主要研究方向：深度学习、人工智能;

This work is partially supported by the National Natural Science Foundation of China (11465004).

LI Wenhui, born in 1997, M. S. candidate. Her research interests include deep learning, artificial intelligence.

ZENG Shangyou, born in 1967, Ph. D., professor. His research interests include neural networks and artificial intelligence, complex networks, biological information processing and biochips.

WANG Jinjin, born in 1995, M. S. candidate. Her research interests include deep learning, artificial intelligence.