计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300146

联合自注意力和循环网络的图像标题生成

王 习[1] 张 凯[1] 李军辉[1] 孔 芳[1] 张熠天[2]
1 苏州大学计算机科学与技术学院 江苏 苏州 215006
2 国家工业信息安全发展研究中心 北京 100000
(20185427010@stu.suda.edu.cn)

摘 要 目前大多数图像标题生成模型都是由一个基于卷积神经网络(Convolutional Neural Network,CNN)的图像编码器和一个基于循环神经网络(Recurrent Neural Network,RNN)的标题解码器组成。其中图像编码器用于提取图像的视觉特征,标题解码器基于视觉特征通过注意力机制来生成标题。然而,使用基于注意力机制的RNN的问题在于,解码端虽然可以对图像特征和标题交互的部分进行注意力建模,但是却忽略了标题内部交互作用的自我注意。因此,针对图像标题生成任务,文中提出了一种能同时结合循环网络和自注意力网络优点的模型。该模型一方面能够通过自注意力模型在统一的注意力区域内同时捕获模态内和模态间的相互作用,另一方面又保持了循环网络固有的优点。在MSCOCO数据集上的实验结果表明,CIDEr值从1.135提高到了1.166,所提方法能够有效提升图像标题生成的性能。

关键词:图像标题;自注意力机制;循环神经网络
中图法分类号 TP391.1

Generation of Image Caption of Joint Self-attention and Recurrent Neural Network

WANG Xi[1],ZHANG Kai[1],LI Jun-hui[1],KONG Fang[1] and ZHANG Yi-tian[2]
1 School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China
2 China Industrial Control Systems Cyber Emergency Response Team,Beijing 100000,China

Abstract At present,most image caption generation models consist of an image encoder based on convolutional neural network (CNN) and a caption decoder based on recurrent neural network(RNN). The image encoder is used to extract visual features from images,while the caption decoder generates captions based on visual features with an attention mechanism. Although the decoder uses RNN with an attention mechanism to model the interaction between image features and captions,it ignores the self-attention of the internal interaction of images or captions. Therefore,this paper proposes a novel model that combines the advantages of RNN and self-attention network for image caption generation. On the one hand,this model can capture interactions within and between modalities in the unified attention area through the self-attention simultaneously. On the other hand,it maintains the inherent advantages of RNN. Experimental results on the MSCOCO dataset show that the proposed model outperforms baseline by improving the performance from 1.135 to 1.166 in CIDEr.

Keywords Image caption,Self-attention mechanism,Recurrent neural network

1 引言

自然语言处理(Natural Language Processing,NLP)和计算机视觉(Computer Vision,CV)是人工智能领域研究的两大热点。当前,跨领域研究已经成为了未来研究的一种趋势,引起了研究者们的极大兴趣。图像标题(Image Caption)正是结合计算机视觉和自然语言处理的一种跨领域研究,该技术最早由Farhadi等[1]提出,即给定二元组(I,S),其中I表示图像,S表示对该图像的描述,模型要完成$I\rightarrow S$的映射,因此模型首先需要学习到图片所对应的描述,然后用训练完成的模型实现输入一张图片就能够得到对这张图片的描述。“看图说话”对正常人来说非常简单,但对于计算机来说却是一项极大的挑战。计算机不仅仅要识别图片的内容,还要用人类的逻辑思维描述出人类可读的句子。

当前大多数图像标题生成模型采用编码器-解码器的框架,编码器用来提取图像的视觉特征,解码器基于视觉特征通过注意力机制来生成标题,但是这种方式忽略了图像和标题内部的相互作用,因此使用基于自注意力机制的Transformer[2]模型来同时捕获模态内和模态间的相互作用。为了能同时使用自注意力机制和循环网络,以进一步提高图像标题

到稿日期:2020-03-24 返修日期:2020-07-12 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61876120)
This work was supported by the National Natural Science Foundation of China(61876120).
通信作者:李军辉(jhli@suda.edu.cn)

生成的性能，本文使用 Dehghani 等[3]提出的 Universal Transformer(UT)模型，这是一种并行的自注意循环序列模型，可以重复和全局接收前馈序列模型。这种特性对于图像标题生成是很有必要的，因为图像生成标题不仅需要关注图像内部特征间、图像和标题多模态间、标题内部的相互作用，还需要利用循环序列模型在预测标题时的反复感应偏差。

因为强化学习的方法可以直接针对评价指标训练端到端的任务，所以我们考虑使用强化学习对图像生成标题系统进行优化，在训练时直接优化 CIDEr 指标，在预测时使用贪心解码算法。实验结果表明，本文方法在 MSCOCO 数据集上表现出了很好的性能，CIDEr 值从 1.166 提高到了 1.264，BLEU_4 从 0.364 提高到了 0.389。

2 相关工作

图像标题生成的传统方法是首先利用图像处理的一些算子来提取图像特征，经过支持向量机(Support Vector Machine，SVM)分类得到图像中可能存在的目标对象[4]；然后根据提取到的对象以及它们的属性，利用条件随机场(Conditional Random Field，CRF)或者是一些指定的规则来恢复对图像的描述。这种做法非常依赖图像特征的提取以及生成句子时所需要的规则，效果并不理想。

受神经机器翻译的启发，将机器翻译中编码源文字的循环神经网络替换为卷积神经网络来编码图像，进而将图像标题生成转化为机器翻译的问题。从翻译的角度，此处的源文字就是图像，目标文字就是生成的标题。图像标题采用的神经网络模型通常由编码器和解码器两部分组成，其中编码器使用 CNN 将图片转化为一个固定长度的向量，即图像的隐层表示；解码器使用 RNN 将编码器输出的固定长度的向量解析为目标语言句子。Vinyals 等[5]提出了神经图像描述(Neural Image Caption，NIC)模型，该模型将图像和单词投射到多模态空间，并使用长短时记忆网络(Long Short-Term Memory，LSTM)生成英文描述。Karpathy 等[6]提出利用片段图像来生成局部区域的描述。Mao 等[7]在基于传统 CNN 编码器-RNN 解码器的神经网络模型的基础上，提出使用多模态空间来为图像和文本建立联系。Xu 等[8]提出了 gLSTM 模型，该模型使用语义信息指导长短时记忆网络来生成标题，解决了图像仅在开始时传入 LSTM 的问题。在此基础上，Wu 等[9]提出了 att-LSTM 模型，该模型通过图像多标签分类来提取图像中可能存在的属性，这种方法解决了图像特征总是使用全局特征的问题。Xu 等[10]将注意力机制引入解码过程，使得标题生成网络能够捕捉图像的局部信息。然而这种加入注意力的方法存在一些缺点，比如虽然每个词都会对应一个图像区域，但是有些介词、动词等并不能对应实体；除此之外，注意力机制是基于卷积层的加权，因此将其对应到图像中会略显模糊，且其不能准确定位出图中对应的区域。为了解决这些问题，Lu 等[11]提出了一种自适应性的注意力机制，该机制使模型可以自己选择是根据先验知识(模板)还是图像中的区域来生成单词。上述工作均是基于解码器 RNN 的。然而 CNN 也是不可忽略的一个重点，Chen 等[12]提出利用卷积层的不同通道构建注意力，同时还利用了空间注意力机制。Li 等[13]构建了首个中文图像摘要数据集 Flickr8kCN，并提出中文摘要生成模型 CS-NIC，该模型使用 GoogleNet[14]对图像进行编码，并使用 LSTM 对图像描述生成过程进行建模。Rennie 等[15]提出 SCST 模型，该模型利用强化学习的方法，生成有区别度的标题。Anderson 等[16]提出了 Bottom-Up and Top-Down 模型，该模型结合了 Bottom-Up 和 Top-Down 视觉注意力机制，其中 Bottom-Up 机制用于提取视觉特征，Top-Down 机制用于关注词向量特征。

上述模型均采用了 CNN 编码器-RNN 解码器的框架。然而近年来，He 等[17]基于语义注意力机制提出了一个新颖的框架模型，该模型包含 DenseNet 和 Fully Convolutional 两个语义注意力模型。模型首先通过 DenseNet 来抽取图像特征，并重新编码，然后使用 Fully Convolutional 来提取图像标签，最后使用 LSTM 来生成图像标题。除此之外，Liu 等[18]还提出了带有双重注意力机制的图像标题生成的模型，该模型使用 CNN 来进行图像特征提取，RNN 作为解码器，其双重注意力机制包括图像特征的注意力以及文本的注意力，模型将两种注意力输入到 LSTM 模型中，用于图像标题生成。但是上述方法只考虑了图像和标题多模态间的相互作用，且在注意力机制中也只考虑了图像特征和句子描述，而没有考虑到图像特征以及标题内部间的交互作用。因此本文利用 Transformer 可以在统一的注意力区域内同时捕获模态内和模态间的相互作用的优点来弥补 RNN 的缺陷；同时考虑结合循环序列模型在预测标题时可以产生反复感应偏差的优点，提出使用 Universal Transformer 框架作为图像标题生成的模型，并在此模型的基础上做了一些改进；此外，本文还考虑加入强化学习策略以优化梯度。实验结果表明，本文提出的使用 Universal Transformer 的方法相比使用 Transformer 的方法，在不使用强化学习策略时，BLEU_4 提高了 0.9，使用强化学习策略时 BLEU_4 提高了 0.5。这说明本文提出的方法能有效地提升图像生成标题的性能。

3 自注意力机制模型

本节将简单地描述基于自注意力机制的 Transformer 模型，然后介绍如何将 Transformer 模型应用到图像标题生成的任务上。

3.1 自注意力机制模型的结构

基于自注意力机制的 Transformer 系统包括一个编码器和一个解码器，如图 1 所示。与基于循环神经网络的机器翻译不同，Transformer 编码器和解码器由一种新颖的注意力机制和一个前馈的神经网络构成。Transformer 的作者将这种新颖的注意力机制命名为“多头注意力机制”(Multi-head Attention，MHA)。同时，一个“多头注意力机制”又由多个点积注意力机制通过 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 计算得到，即：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中，d_k 表示向量 $\boldsymbol{K}$ 的维度。

最终，多个点积注意力机制的结果进行拼接后被送入到一个前馈神经网络中，即：

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{k},\boldsymbol{V}\boldsymbol{W}_i^{v}) \tag{2}$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_i)\boldsymbol{W}^o \tag{3}$$

其中，$\boldsymbol{W}_i^Q\in\boldsymbol{R}^{d_{model}\times d_k}$，$\boldsymbol{W}_i^k\in\boldsymbol{R}^{d_{model}\times d_k}$，$\boldsymbol{W}_i^v\in\boldsymbol{R}^{d_{model}\times d_k}$，$\boldsymbol{W}_i^o\in\boldsymbol{R}_i^{hd_v\times d_{model}}$，$h$ 为“多头注意力机制”中头的个数。

前馈神经网络为两层全连接层，使用“修正线性单元”(RELU)作为激活函数。该网络定义为：

$$FFN(x)=\max(0,x\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{4}$$

其中，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$，$\boldsymbol{b}_1$，$\boldsymbol{b}_2$ 表示模型参数。Vaswani 等[2]将输入和输出的维度均设置为 512，将隐层的维度设置为 2048。

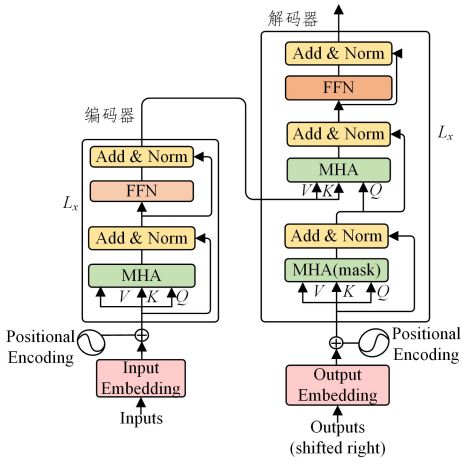

图 1　Transformer 模型的框架

Fig. 1　Framework of Transformer model

由于 Transformer 结构既没有使用 RNN，也没有使用 CNN，为了弥补输入序列顺序上的缺失，Transformer 使用了位置编码的方法在编码和解码词向量后增加了表示位置的信息，该信息是一种绝对的位置信息，被定义为：

$$\begin{aligned}PE(pos,2i)&=\sin\left(\frac{pos}{10\,000^{\frac{2i}{d_{model}}}}\right)\\PE(pos,2i+1)&=\cos\left(\frac{pos}{10\,000^{\frac{2i}{d_{model}}}}\right)\end{aligned} \tag{5}$$

其中，pos 指单词在输入的句子中的位置，i 表示向量的维度。

3.2　图像标题生成的自注意力机制模型

目前大多数图像标题生成任务都采用了编码器-解码器的框架，Transformer 的模型结构也采用了编码器-解码器的框架。从结构上，Transformer 很适合图像标题生成任务；从其他方面，Transformer 的注意力机制能同时捕捉图像特征内部间的相互作用、图像和标题间的相互作用以及标题序列内部的相互作用。因此使用 Transformer 结构来完成图像标题生成任务具有很多优点。

图像标题的 Transformer 框架如图 2 所示，整个网络结构由图像编码器和标题序列解码器组成。图像编码器首先将图像作为输入，使用 CNN 来提取视觉特征，然后将视觉特征经过线性层转换后输入编码器，最后通过自注意学习来获得所参与的视觉表示。解码器循环地接收参与的视觉特征，并利用前一个单词来预测下一个单词，同时通过自注意力机制来学习标题序列间的相互作用。为了使整个网络结构看上去更简洁，我们使用 Encoder 和 Decoder 来分别表示图 1 中的编码器和解码器。

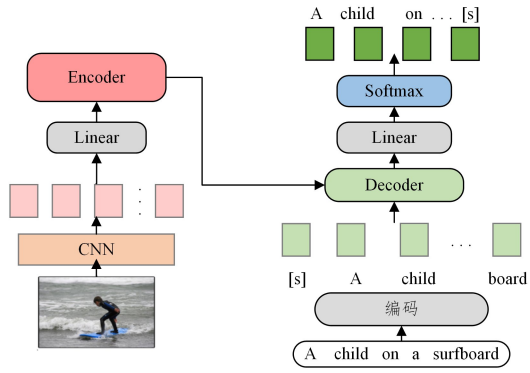

图 2　图像标题生成模型的框架图

Fig. 2　Framework of image caption generation model

4　联合自注意力和循环网络的图像标题生成

使用 Transformer 作为图像标题生成的模型尽管具有很多优点，但是其同时也失去了 RNN 的一些优点，比如在一些简单的任务中，RNN 可以轻松达到比较好的效果(例如在超出训练样本长度时拷贝字符串或者进行简单的逻辑推理)，因此 Dehghani 等[3]提出了 Universal Transformer(UT)模型。UT 是一种并行的自注意循环序列模型，可用于解决上述问题。实验结果表明 UT 在图像标题任务上优于 Transformer 模型。UT 整体模型结构和 Transformer 相似。如图 3 所示，UT 模型在 Encoder 端和 Decoder 端都使用了一个共享权重的 $Transition(\cdot)$ 函数来代替 Transformer 中的两层前馈神经网络 $FFN(\cdot)$ 函数，本文使用循环的方式来表示计算上的先后顺序，用 $\boldsymbol{H}_t$ 表示式(3)中第 t 层的 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 的映射，那么式(3)就可以表示为：

$$MultiHead(\boldsymbol{H}_t)=Concat(head_i)\boldsymbol{W}^o \tag{6}$$

然后在第 t 层，将所有的输入计算修正表示为：

$$\boldsymbol{H}_t=LayerNorm(\boldsymbol{A}^t+Transition(\boldsymbol{A}^t)) \tag{7}$$

其中 $\boldsymbol{A}^t=LayerNorm(\boldsymbol{H}^{t1}+\boldsymbol{PE}^t)+MultiHead(\boldsymbol{H}^{t-1}+\boldsymbol{PE}^t)$ $MultiHead(\boldsymbol{H}^t)=Concat(head_i)\boldsymbol{W}^0$。

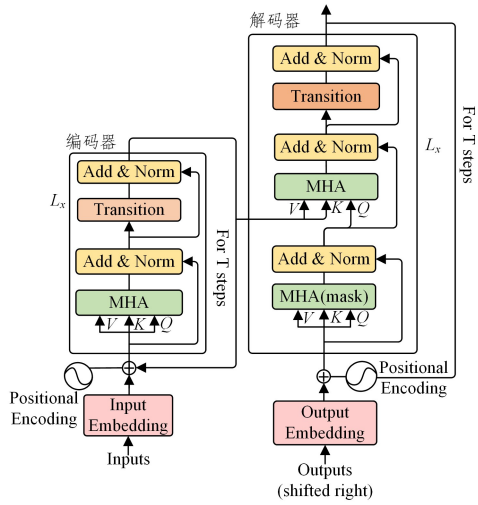

图 3　Universal Transformer 模型框架图

Fig. 3　Framework of Universal Transformer model

$LayerNorm(\cdot)$ 是由 Ba 等[19]提出的一种正则化项，$Transition(\cdot)$ 是 UT 相比 Transformer 所做的改动而添加

的一个函数，本文使用一个 1×1 的卷积作为 *Transition*(·)函数，***PE*** 表示输入序列添加位置编码信息。

本文对 UT 模型做了一些改变：考虑到两层的前馈神经网络可以学习到更多有用的参数，因此我们保留了前馈神经网络结构，只分别在 Encoder 端和 Decoder 端的前馈神经网络之后添加了 *Transition*(·)函数，模型其余结构保持不变。

无论使用 Transformer 模型还是 UT 模型，模型最终的输出均可表示为：

$$p(y_t \mid y_{1:t-1}) = \text{softmax}(\boldsymbol{W}_p h + \boldsymbol{b}) \tag{8}$$

其中，W_p 是模型要训练的参数。

标题生成的训练目标是最小化式(9)中的交叉损失函数，即：

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p(y_t^* \mid y_{1:t-1}^*, \boldsymbol{I})) \tag{9}$$

其中，$y_{1:T}^*$ 表示真实标签，θ 表示标题生成模型的参数。

5 使用强化学习优化图像标题生成模型

如上所述，图像标题通常使用交叉熵损失函数来进行训练优化。Rennie 等[20]的研究表明，使用强化学习策略梯度方法可以针对不可区分的指标来训练深度端到端的系统。因此本文使用强化学习策略对图像标题进行优化，通过使用 MSCOCO 任务的测试指标来对系统进行精心优化，可以显著提高模型性能。

将 Transformer 模型和 UT 模型作为“代理(agent)”与外部“环境(environment)”(图像特征和标题)进行交互，网络的参数 θ 定义了一个策略 p_θ，该策略导致了一个“动作(action)”，即对下一个单词进行预测。在执行完每个操作后，代理会更新其内部的“状态(state)”。当生成序列结束时，代理会观察到“奖励(reward)”，如生成句子的 CIDEr[21]得分(我们将其表示为奖励)，将生成序列和对应的序列进行比较，由评估算法计算奖励。训练的目的是最大程度地减少预期的负回报：

$$L(\theta) = -E_{\omega^s \sim p_\theta}[r(\omega^s)] \tag{10}$$

其中 $\omega^s = (\omega_1^s, \cdots, \omega_T^s)$，$\omega_t^s$ 表示第 t 时间步模型通过随机采样的方式得到的单词。

在实验中，$L(\theta)$ 可以近似表示为：

$$L(\theta) \approx -r(\omega^s), \omega^s \sim p_\theta \tag{11}$$

本文使用和 Rennie 等[20]相同的梯度处理策略，最终第 t 时间步的梯度导数可表示为：

$$\frac{\partial L(\theta)}{\partial s_t} = (r(\omega^s) - r(\hat{\omega}))(p_\theta(\omega_t \mid h_t) - l_{\omega_t^s}) \tag{12}$$

其中，$r(\omega^s)$ 表示模型随机采样得到的序列和真实序列的 CIDEr 得分，$r(\hat{\omega})$ 表示模型贪心策略得到的序列和真实序列的 CIDEr 得分。

6 实验与分析

6.1 数据集

本文使用的数据集为 MSCOCO2014[22]。我们使用 Karpathy 等[6]的方法将 MSCOCO 数据集分成训练集、验证集和测试集，其中训练集共有 113 287 张图片，验证集和测试集各有 5000 张图片，每张图片提供 5 句不同的英文标题。

本文的评测使用 BLEU-1, 2, 3, 4[23], METEOR[24], ROUGE_L[25], CIDEr 和 SPICE[26]这 5 种指标来评价图像生成的标题的质量。其中 BLEU 一般是用于测评机器翻译的翻译质量，其反映了生成结果与参考答案之间的 N 元文法准确率。METEOR 测量基于单精度加权调和平均数和单字召回率。ROUGE_L 与 BLEU 类似，是基于召回率的相似度衡量方法。CIDEr 是基于共识的评价方法，这个指标首先将每个句子都看作是“文档”，并将其用向量的形式进行表示，然后计算参考的标题与模型生成的标题的余弦相似度，最后根据余弦相似度进行打分。SPICE 是一种基于语义命题的图像标题评估方法，其针对有些句子虽然在 n-gram 上重叠很低，但是在所表达意思相近的情况下，会尽可能多地考虑每句话的语义命题。通过将候选标题和参考标题用基于规则的方法转换为场景图，将评估图的语义质量作为标题质量，只对图中的对象、属性和关系进行编码，在此过程中，SPICE 抽象出自然语言的大部分词汇和句法特性。

6.2 实验设置

6.2.1 视觉特征提取的网络设置

视觉特征提取网络 CNN(I)完成 $I \rightarrow V(\boldsymbol{I})$ 的特征映射，其中 I 为输入图像，V 输出为视觉特征。视觉特征提取使用 ResNet-101 结构[27]和 Faster-RCNN 结构[28]，其中 ResNet-101 结构在大规模单标签分类任务 ImageNet[29]上进行训练；Faster-RCNN 结构针对目标区域检测任务进行训练。本文使用已经训练好的结构来提取视觉特征，并将这种结构的最后一层卷积层的输出作为提取到的视觉特征。

首先将 ResNet-101 的最后一层卷积层的输出经过平均池化层，并将图像特征映射为(50, 14×14, 2048)的特征矩阵，然后经过隐藏单元数为 512 的全连接层，最后将视觉特征映射为(50, 14×14, 512)的特征矩阵作为最终的视觉特征，且在整个模型训练的过程中，ResNet-101 模型的参数不更新。对于 Faster-RCNN 结构，我们使用与 Anderson 等[16]相同的表示方式。

6.2.2 图像标题生成模型的设置

图像标题生成模型使用 Transformer 和 UT 模型，隐藏状态长度 H 均为 512，词向量长度 E 为 512。英文词汇表大小为 9487，未登录词用〈UNK〉表示。词向量和模型参数的初始值在区间[−0.1, 0.1]通过均匀分布得到，损失函数的优化使用自适应矩估计(Adam)算法[30]，学习率为 2×10^{-4}。训练时批处理大小为 50，测试时批处理大小为 10，并且在测试时使用大小为 4 的柱状搜索算法[31]。在强化学习策略阶段，测试时使用贪心算法。在训练过程中每一层使用 Dropout 正则化和归一化处理来提高模型的泛化能力[32]。

训练分为两个阶段：第一阶段使用交叉熵优化损失，采用最大步长为 25 个 epoch 的早停策略；第二阶段是在完成第一阶段训练的模型的基础上，使用强化学习策略优化损失，只用 5 个 epoch 就使模型能达到很好的性能。

6.3 结果与分析

本文首先和以前的方法进行比较，然后比较了 Transformer 模型、Universal Transformer 模型以及强化学习策略算法的实验结果，最后比较了不同的视觉特征提取网络对实验结果的影响。

表1列出了本文提出的UT模型使用强化学习策略和以前的方法对比的结果，本文使用了两种不同的模型来提取视觉特征，即表中的Up-Down：ResNet-101和Up-Down（Up-Down表示使用Faster-RCNN网络来提取图像的特征）。由表1可知，当选择Resnet-101网络来提取视觉特征时，UT模型相比Up-Down：ResNet-101模型，BLEU_4从0.334提高到了0.342，提高了0.008，CIDEr值从1.054提高到了1.086；Transformer模型相比Up-Down：ResNet-101模型，CIDEr值从1.054提高到了1.071。

由此可知，不论是UT还是Transformer模型都可以将图像特征和图像文本描述特征进行交互，并且进行注意力的计算。

表1　本文提出的英文图像标题方法和已有方法在MSCOCO Karpathy测试集上的比较

Table 1　Comparisons of the proposed English image captioning approach and existing methods on MSCOCO Karpathy test set

Method	CIDEr	BLEU_4	BLEU_3	BLEU_2	BLEU_1	ROUGE_L	METEOR	SPICE
DeepVS[5]	0.660	0.230	0.321	0.450	0.625	—	0.195	—
Hard-attention[9]	—	0.250	0.357	0.504	0.718	—	0.230	—
Adaptive[10]	1.085	0.332	0.439	0.580	0.742	—	0.257	—
SCST：Att2all[14]	1.140	0.342	—	—	—	0.557	0.267	—
Up-Down：ResNet-101[15]	1.054	0.334	—	—	0.754	—	0.261	0.192
Up-Down[15]	1.135	0.362	—	—	0.772	0.564	0.270	0.203
ResNet-101＋UT＋RL	1.178	0.364	0.477	0.618	0.776	0.566	0.275	0.212
Faster-RCNN＋UT＋RL	**1.264**	**0.389**	**0.506**	**0.648**	**0.799**	**0.587**	**0.286**	**0.223**

由表1可知，该模型还可以进一步提高预测图像标题的能力。当使用Faster-RCNN来提取视觉特征时，CIDEr值从1.135提高到了1.166；当使用强化学习策略时，模型达到了更好的性能，BLEU_4从0.364提高到了0.389，CIDEr值从1.166提高到了1.264。由此可知，强化学习策略可以进一步提升模型性能。综上，本文提出的方法和以前的方法相比能有效提升图像标题生成的性能。

表2列出了选择ResNet-101网络来提取视觉特征的实验结果，表中的tf表示Transformer模型。由表2可知，UT模型的BLEU_4值比tf模型高出0.009，这表明了UT模型的并行循环自注意力能够有效地弥补Transformer模型的不足。

表2　视觉特征使用ResNet-101网络的实验结果

Table 2　Experimental results of visual features using ResNet-101 network

Method	CIDEr	BLEU_4	BLEU_3	BLEU_2	BLEU_1	ROUGE_L	METEOR	SPICE
ResNet-101＋tf	1.071	0.333	0.439	0.576	0.743	0.547	0.267	0.198
ResNet-101＋UT-base	1.080	0.335	0.441	0.579	0.745	0.550	0.269	0.200
ResNet-101＋UT	1.086	0.342	0.446	0.581	0.747	0.551	0.269	0.202
ResNet-101＋tf＋RL	1.155	0.347	0.464	0.608	0.769	0.558	0.271	0.208
ResNet-101＋UT＋RL	**1.178**	**0.364**	**0.477**	**0.618**	**0.776**	**0.566**	**0.275**	**0.212**

表3列出了选择Faster-RCNN网络来提取视觉特征的实验结果。由表3可知，UT模型优于Transformer模型，且使用强化学习策略可以有效地提高图像标题生成的性能。由于修改后的UT模型保留了Transformer模型的两层前馈网络机制，由表2和表3可知，修改后的UT模型的性能优于UT-base模型。

表3　视觉特征使用Faster-RCNN网络的实验结果

Table 3　Experimental results of visual features using Faster-RCNN network

Method	CIDEr	BLEU_4	BLEU_3	BLEU_2	BLEU_1	ROUGE_L	METEOR	SPICE
Faster-RCNN＋tf	1.152	0.361	0.469	0.606	0.767	0.565	0.279	0.209
Faster-RCNN＋UT-base	1.161	0.363	0.470	0.609	0.770	0.568	0.280	0.211
Faster-RCNN＋UT	1.166	0.364	0.474	0.612	0.772	0.570	0.281	0.213
Faster-RCNN＋tf＋RL	1.252	0.384	0.501	0.645	0.797	0.582	0.284	0.223
Faster-RCNN＋UT＋RL	**1.264**	**0.389**	**0.506**	**0.648**	**0.799**	**0.587**	**0.286**	**0.223**

对比表2和表3可以发现，使用不同的网络来提取视觉特征会对实验结果产生很大的影响。

6.4　实例分析

图4给出了一张图片在不同的模型上得到的结果，从标题的生成结果上，ResNet-101＋UT＋RL模型预测的结果是‘a group of people standing on a beach holding surfboards’，Faster-RCNN＋UT＋RL模型预测的结果是‘a group of people walking across a beach holding surfboards’。模型预测的结果和人类描述的结果相比，ResNet-101＋UT＋RL模型把结果动词‘walking’预测成了‘standing’，语法使用不太准确；虽然ResNet-101＋tf＋RL模型可以将‘walking’预测成‘walking across’，但是整体的描述不太完整；Faster-RCNN＋UT＋RL模型不仅预测的结果全面，而且在表达上更加自然。虽然模型使用循环网络（UT）可以对图片进行更加完整的描述，但是因为其使用不同的视觉特征抽取模型，所以准确性不太稳定。

ResNet-101+tf: a people standing on top of a beach
ResNet-101+UT: a group of people standing on top of a beach
ResNet-101+tf+RL: a group of people walking across a beach
ResNet-101+UT+RL: a group of people standing on a beach holding surfboards

Faster-RCNN_tf: a group of people on a beach with surfboards
Faster-RCNN+UT: a group of people walking across a beach
Faster-RCNN+tf+RL: a group of people standing on a beach holding surfboards
Faster-RCNN+UT+RL: a group of people walking across a beach holding surfboards

图 4 模型预测结果

Fig. 4 Model prediction result

结束语 本文提出了使用 Transformer 模型、UT 模型以及强化学习策略的方法来提高图像标题生成的性能，并比较了不同的视觉特征提取网络对实验结果的影响。实验结果表明，本文提出的方法能有效提高图像标题生成的性能。

虽然本文提出的方法有效地提高了图像标题生成的性能，但是无论是 Transformer 模型还是 UT 模型，词向量都需要随着模型一起进行训练，因此随着 BERT 网络在 NLP 各项任务上的应用，未来会考虑使用 BERT[33] 加载预训练模型来进一步提高图像标题生成的性能，同时会考虑借助无监督学习的思想，实现在没有中文语料的情况下图像生成中文标题。

参 考 文 献

[1] FARHADI A, HEJRATI M M, SADEGHI M A, et al. Every Picture Tells a Story: Generating Sentences from Images[C]// Proceedings Part IV of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010: 15-29.

[2] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[3] DEHGHANI M, GOUWS S, VINYALS O, et al. Universal Transformers[J]. arXiv: 1807.03819, 2018.

[4] KULKARNI G, PREMRAJ V, ORDONEZ V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

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

[6] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.

[7] MAO J H, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv: 1412.6632, 2014.

[8] XU J, GAVVES E, FERNANDO B, et al. Guiding the long-short term memory model for image caption generation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2407-2415.

[9] WU Q, SHEN C H, LIU L Q, et al. What value do explicit high level concepts have in vision to language problems? [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 203-212.

[10] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015: 2048-2057.

[11] LU J, XIONG C M, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 375-383.

[12] CHEN L, ZHANG H W, XIAO J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5659-5667.

[13] LI X R, LANW Y, DONG J F, et al. Adding Chinese Captions to Images[C]// Proceedings of the 2016 Association for Computing Machinery (ACM) on International Conference on Multimedia Retrieval. New York, USA: ACM, 2016: 271-275.

[14] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR, 2015: 1-9.

[15] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7008-7024.

[16] ANDERSON P, HE X D, 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.

[17] HE X, YANG Y, SHI B, et al. VD-SAN: Visual-Densely Semantic Attention Network for Image Caption Generation[J]. Neurocomputing, 2019, 328: 48-55.

[18] LIU M, LI L, HU L, et al. Image caption generation with dual attention mechanism[J]. Information Processing and Management, 2020, 57(2): 102178.

[19] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. arXiv: 1607.06450, 2016.

[20] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical

Sequence Training for Image Captioning[J]. arXiv:1612.00563, 2016.

[21] VEDANTAM R,ZITNICK C L,PARIKH D,et al. CIDEr:Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566-4575.

[22] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft coco:Common objects in context[C]// European Conference on Computer Vision. Springer,Cham,2014:740-755.

[23] PAPINENI K,ROUKOS S,WARD T,et al. Bleu:a Method for Automatic Evaluation of Machine Translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia,PA,USA:ACL,2002:311-318.

[24] DENKOWSKI M,LAVI A. Meteor universal:Language specific translation evaluation for any target language[C]// Proceedings of the Ninth Workshop on Statistical Machine Translation. 2014:376-380.

[25] LIN C Y. Rouge:A package for automatic evaluation of summaries[C]// Text Summarization Branches Out, Post-conference Workshop of ACL 2004. Barcelona,Spain,2004:74-81.

[26] ANDERSON P,FERNANDO B,JOHNSON M,et al. Spice:Semantic propositional image caption evaluation[C]// European Conference on Computer Vision. Springer, Cham, 2016: 382-398.

[27] HE K M,ZHANG X Y,RENS Q,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[28] REN S,HE K,GIRSHICK R,et al. Faster RCNN:Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015:91-99.

[29] 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.

[30] KINGMA D P,BA J. Adam:A method for stochastic optimization[J]. arXiv:1412.6980,2014.

[31] WISEMAN S,RUSH A M. Sequence-to-sequence learning as beam-search optimization[J]. arXiv:1606.02960,2016.

[32] IOFFE S,SZEGEDY C. Batch Normalization. Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]// Proceedings of the 32nd International Conference on Machine Learning. Lille,France:JMLR.org,2015:448-456.

[33] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. Proceedings of NACL-HLT,2019,1:4171-4186.

WANG Xi, born in 1995, postgraduate, is a member of China Computer Federation. Her main research interests include natural language processing and image caption.

LI Jun-hui, born in 1983, associate professor. His main research interests include natural language processing and machine translation.