基于双注意模型的图像描述生成方法研究

卓亚琦[1],魏家辉[2],李志欣[2]
(1. 桂林理工大学理学院,广西桂林 541004; 2. 广西师范大学广西多源信息挖掘与安全重点实验室,广西桂林 541004)

摘 要: 现有图像描述生成方法的注意模型通常采用单词级注意,从图像中提取局部特征作为生成当前单词的视觉信息输入,缺乏准确的图像全局信息指导. 针对这个问题,提出基于语句级注意的图像描述生成方法,通过自注意机制从图像中提取语句级的注意信息,来表示生成语句所需的图像全局信息. 在此基础上,结合语句级注意和单词级注意进一步提出了双注意模型,以此来生成更准确的图像描述. 通过在模型的中间阶段实施监督和优化,以解决信息间的干扰问题. 此外,将强化学习应用于两阶段的训练来优化模型的评估度量. 通过在MSCOCO和Flickr30K两个基准数据集上的实验评估,结果表明本文提出的方法能够生成更加准确和丰富的描述语句,并且在各项评价指标上优于现有的多种基于注意机制的方法.

关键词: 图像描述生成; 编码器-解码器架构; 单词级注意; 语句级注意; 双注意模型; 强化学习
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)05-1123-08
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210696

Research on Image Captioning Based on Double Attention Model

ZHUO Ya-qi[1], WEI Jia-hui[2], LI Zhi-xin[2]
(1. *College of Science, Guilin University of Technology, Guilin, Guangxi* 541004, *China*;
2. *Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin, Guangxi* 541004, *China*)

Abstract: The attention model of existing image captioning approaches usually adopt word-level attention, which extracts local features from images. The local features are used as the visual information input to generate the current word, lacking accurate image global information guidance. To solve this problem, this paper proposed image captioning approach based on sentence-level attention. The approach employs the self-attention mechanism to extract the sentence-level attention information from the image, which is used to represent the global image information needed to generate sentences. On this basis, we further proposes a double attention model which combines sentence-level attention with word-level attention to generate more accurate description. We implement supervision and optimization in the intermediate stage of the model to solve the problem of information interference. In addition, reinforcement learning is applied in two-stage training to optimize the evaluation metric of the model. Finally, we evaluated our approach on two baseline datasets, i.e. MSCOCO and Flickr30K. Experimental results show that the proposed approach can generate more accurate and richer captions. Hence it outperforms many state-of-the-art image captioning approaches based on attention mechanism in various evaluation metrics.

Key words: image captioning; encoder-decoder architecture; word-level attention; sentence-level attention; double attention model; reinforcement learning

1 引言

图像描述生成旨在为图像生成准确的自然语言描述,包括获取一幅图像、分析其视觉内容、生成文本描述以说明图像中的显著目标和行为等步骤[1]. 传统的视觉任务(如物体检测[2]或图像自动标注[3])都是在有限个类别上测试检测器或分类器的精度. 相比之下,图像描述生成的任务结合了计算机视觉和自然语言处理两个领域,更具综合性和挑战性.

图像描述系统通常采用卷积神经网络(Convolutional Neural Network, CNN)作为编码器从输入图像中提取视觉特征,同时采用循环神经网络(Recurrent Neu-

收稿日期:2021-05-31;修回日期:2021-09-22;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61966004, No.61866004);广西自然科学基金(No.2019GXNSFDA245018);广西研究生教育创新计划(No.XYCBZ2021002)

ral Network, RNN)作为解码器将编码信息逐词地解码为自然语言描述[4-6]. 为了更好地将图像信息整合到语言模型中,Xu等[7]首先在图像描述系统中引入了软、硬两种视觉注意机制. Chen等[8]在此基础上提出了一种基于空间和通道的双注意图像描述生成方法. Lu等[9]发现在语句生成过程中有些单词是不需要参考视觉信息的,提出了一种基于视觉哨兵的自适应注意模型. You等[10]提出了基于语义注意的图像描述方法,从图像中提取一些视觉语义概念来实现视觉信息的增强. 近年来,自注意机制也引起了越来越多研究人员的关注. Vaswani等[11]提出Transformer模型,完全抛弃了RNN,仅通过自注意机制就实现了机器翻译的最新成果. Yu等[12]提出用于机器阅读理解的QANet,使用CNN和自注意机制相结合构建网络模型. 此外,许多研究人员尝试利用强化学习技术进一步优化图像描述生成模型,通过梯度估计等方法直接优化评估度量,从而生成更符合人类感知的描述语句[13,14].

综上所述,目前图像描述生成的主流方法是基于编码器-解码器架构并结合注意机制的方法[7-10],但仍然存在一些问题. 首先,注意模型大多是单词级的局部注意,缺乏对图像整体的掌控. 其次,模型在训练和测试之间存在"暴露偏差"(exposure bias)[13]. 最后,存在训练损失和评估指标不匹配的问题. 模型在训练时通常使用交叉熵损失,而在测试时一般使用BLEU[15]、METEOR[16]、ROUGE-L[17]、CIDEr[18]这样的指标来评估生成语句的质量.

针对以上问题,提出了一种新的基于自注意的图像描述生成方法来探讨图像全局注意信息对于描述生成的有效性. 自注意机制可以自动地从图像中提取语句级的注意信息,具有比单词级注意特征更完整的视觉表示. 在训练时,通过在解码器的每个时间步中对语句级注意信息的调整,可以使它满足生成整个语句的需求. 语句级注意图和单词级注意图的示例如图1所示. 此外,为了能结合两种注意的优势,生成更加准确的描述,进一步提出了双注意模型,通过集成两种注意模型来生成最终的描述语句. 最后,为了解决"暴露偏差"和损失-评估不匹配问题[13,14],将强化学习应用于两阶段训练方法来优化模型的评估度量.

图1 视觉注意图示例

2 提出方法

本方法使用统一的编码器-解码器架构生成图像描述. 首先使用CNN将输入的图像编码为空间特征 $\boldsymbol{I}$,然后利用RNN的变体长短期记忆网络(Long Short-Term Memory, LSTM)将特征向量解码为单词序列 $\boldsymbol{Y}=(y_1, y_2, \cdots, y_T)$,其中 $y_i \in D$ 是预测生成的单词,D 是包含所有单词的词典,T 是语句的最大长度. 单词由词嵌入向量表示,每个语句的开头用一个特殊的<start>标记,结尾用一个特殊的<end>标记. 在模型解码过程中,上一时间步生成的单词会被反馈到LSTM中,结合图像的注意信息和上一时间步的隐状态 $\boldsymbol{h}_{t-1}$,生成当前时间步的隐状态 $\boldsymbol{h}_t$,然后根据 $\boldsymbol{h}_t$ 预测生成当前单词 y_t.

2.1 语句级注意模型

单词级注意是传统的上下文注意[7],根据视觉上下文信息的变化动态地从图像中提取注意信息,在不同的时间步注意信息是不同的,定义如下:

$$\begin{cases} \boldsymbol{Q}_t=\boldsymbol{W}_h\boldsymbol{h}_{t-1}, \quad \boldsymbol{K}=\boldsymbol{W}_w\boldsymbol{I}+\boldsymbol{b}_w, \quad \boldsymbol{V}=\boldsymbol{I} \\ \boldsymbol{a}_t=f(\boldsymbol{Q},\boldsymbol{K})=\boldsymbol{W}_a^{\mathrm{T}}(\mathrm{relu}(\boldsymbol{Q}\oplus\boldsymbol{K}))+\boldsymbol{b}_a \\ \boldsymbol{\alpha}_t=\mathrm{softmax}(\boldsymbol{a}_t^{\mathrm{T}}) \end{cases} \tag{1}$$

其中 $\boldsymbol{W}_h \in \mathbb{R}^{C'\times C'}$ 和 $\boldsymbol{W}_w \in \mathbb{R}^{C'\times C}$ 是将隐状态 $\boldsymbol{h}_{t-1}$ 和图像特征 $\boldsymbol{I}$ 映射到同一维度的变换矩阵,$\boldsymbol{W}_a \in \mathbb{R}^{C'}$ 是变换行向量,$\boldsymbol{b}_w$ 和 $\boldsymbol{b}_a$ 是偏置项. $\oplus$ 运算表示向量 $\boldsymbol{Q}_t$ 与矩阵 $\boldsymbol{K}$ 的加运算,将向量加到矩阵的每一列中. 计算出 $\boldsymbol{a}_t$ 是一个行向量,而 $\boldsymbol{\alpha}_t$ 为最终的权重向量,表示空间注意权重的分布,其维数与 $\boldsymbol{V}$ 中的通道特征向量个数相同. 由此可以得到加权后的单词级注意特征 $\boldsymbol{V}_t^w=\boldsymbol{V}\boldsymbol{\alpha}_t$.

图像的各个区域之间存在非常重要的语义关系,对象间的语义关系对图像描述语句的生成至关重要,但是单词级注意只能关注到少量的局部特征,很难有完整的视觉表示. 因此,提出一种具有语句级注意的图像描述生成方法. 通过对整幅图像的观察,从图像中挑选出对于整个生成语句最有用的信息,通过自注意机制在图像空间区域上产生注意权重,得到图像的注意特征. 在自注意机制中,$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 表示经过变换后的图像空间特征. 图2给出了语句级注意模型的结构,首先使用1×1卷积层将原始图像空间特征 $\boldsymbol{I}$ 转换为三个功能空间特征,然后使用单层的神经网络来融合 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 两个特征向量,最后通过softmax函数在 $\boldsymbol{V}$ 上生成注意权重分布. 具体表示如下:

$$\begin{cases} \boldsymbol{Q}=\boldsymbol{W}_q\boldsymbol{I}, \quad \boldsymbol{K}=\boldsymbol{W}_k\boldsymbol{I}, \quad \boldsymbol{V}=\boldsymbol{W}_v\boldsymbol{I} \\ \boldsymbol{a}=f(\boldsymbol{Q},\boldsymbol{K})=\boldsymbol{W}_s^{\mathrm{T}}(\mathrm{relu}(\boldsymbol{Q}+\boldsymbol{K}))+\boldsymbol{b}_s \\ \boldsymbol{\alpha}=\mathrm{softmax}(\boldsymbol{a}^{\mathrm{T}}) \end{cases} \tag{2}$$

其中 $\boldsymbol{W}_q \in \mathbb{R}^{C''\times C}$、$\boldsymbol{W}_k \in \mathbb{R}^{C''\times C}$ 和 $\boldsymbol{W}_v \in \mathbb{R}^{C''\times C}$ 是1×1卷积层的参数矩阵. 1×1卷积可以通过对每个通道的信息进行计算来实现通道间的信息整合,同时还可以有效地降低模型的计算复杂度. $\boldsymbol{W}_s \in \mathbb{R}^{C''}$ 是一个融合 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的变换行向量,$\boldsymbol{b}_s$ 是偏置项. $\boldsymbol{\alpha}$ 表示得到的空间注意权重分布,$\boldsymbol{\alpha}$

和 $\boldsymbol{V}$ 具有相同的空间大小,即 $L=k\times k$,由此可得到加权后的语句级图像特征 $\boldsymbol{V}^s=\boldsymbol{V}\boldsymbol{\alpha}$.

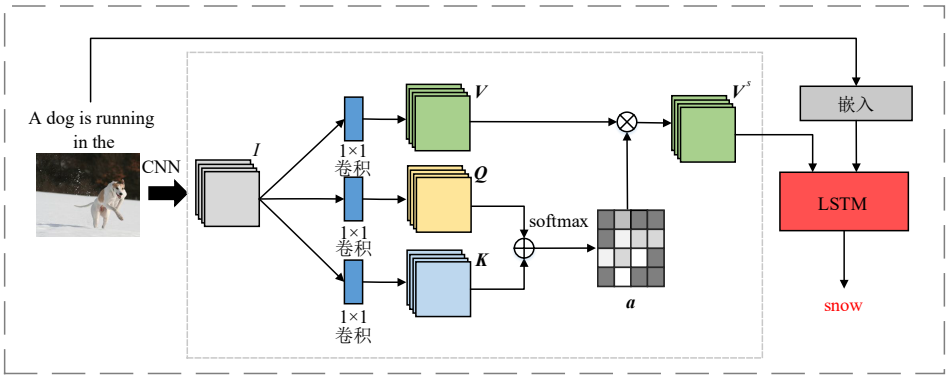

图2　语句级注意模型结构图

此外,本文方法在模型的损失中加入了双随机注意正则化. 在上下文注意模型中,视觉注意随着上下文信息的变化而不断变化. 为了防止一些区域的特征丢失,通常在模型的训练损失中加入双随机注意正则化[7],即:

$$\eta^w=\lambda\sum_{i=1}^{L}\left(1-\sum_{t=1}^{T}\alpha_{t,i}^2\right)^2 \tag{3}$$

其中 λ 是双随机正则化系数,$\alpha_{t,i}$ 是在 t 时间步第 i 个图像区域的注意权重.

而在语句级注意模型中,α_i 在每个时间步都是相同的,并不需要关注到图像的每个区域. 相反,模型可以选择性地只关注对整个语句最有用的部分. 因此,语句级注意模型的双随机注意正则化表示为:

$$\eta^s=\lambda\sum_{i=1}^{L}(T\alpha_i^2)^2 \tag{4}$$

2.2　双注意模型

本文集成了两种注意模型进一步提出双注意模型,并尝试采用两种不同的方式来构建双注意模型. 两种构建方式分别为:

(1) 串联结合方式:图像首先经过语句级注意模型进行初步筛选,得到图像的语句级特征表示,然后通过单词级注意模型进一步根据每个单词提取更加精细化的图像信息,最终将两个单注意模型生成的预测结果进行加权求和来预测最终输出的单词,结构如图3所示. 计算公式为:

$$\begin{cases}\boldsymbol{V}_t^s=\text{Attention}^s(\boldsymbol{I}),\ \boldsymbol{V}_t^w=\text{Attention}^w(\boldsymbol{W}\boldsymbol{I}+\boldsymbol{V}_t^s)\\ \boldsymbol{h}_t^s=\text{LSTM}^s([\boldsymbol{x}_t;\ \boldsymbol{V}_t^s],\boldsymbol{h}_{t-1}^s),\ \boldsymbol{h}_t^w=\text{LSTM}^w([\boldsymbol{x}_t;\ \boldsymbol{V}_t^w],\boldsymbol{h}_{t-1}^w)\\ \boldsymbol{p}_t^s=\text{softmax}(\boldsymbol{W}_p^s\boldsymbol{h}_t^s),\ \boldsymbol{p}_t^w=\text{softmax}(\boldsymbol{W}_p^w\boldsymbol{h}_t^w)\\ \boldsymbol{p}_t=\beta\boldsymbol{p}_t^w+(1-\beta)\boldsymbol{p}_t^s\end{cases} \tag{5}$$

其中 β 为可学习参数.

(2) 并联结合方式:直接将两个单注意模型生成的预测向量进行加权求和来预测模型最终输出的单词,结构如图4所示. 计算公式为:

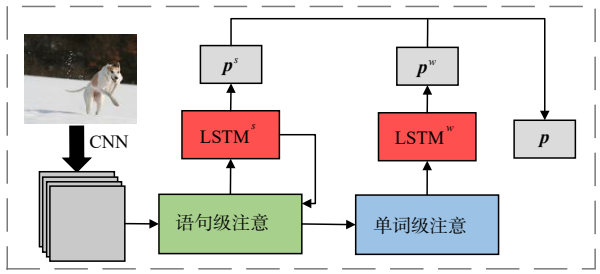

图3　串联型双注意模型

$$\boldsymbol{p}_t=\beta\boldsymbol{p}_t^w+(1-\beta)\boldsymbol{p}_t^s \tag{6}$$

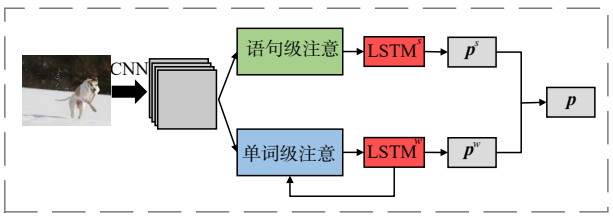

图4　并联型双注意模型

但是,当两种注意结合在一起时,不可避免地会导致信息间的相互干扰. 因此,本文在模型的中间阶段实施监督和优化,使用多任务学习方法将两个单注意模型的损失添加到最终的训练损失中,以调节两个单注意模型在每个时间步的输出,以此来解决两个模型信息间的干扰问题. 模型最终的训练损失可表示如下:

$$\begin{cases}L^s=-\sum_{t=1}^{T}\log p_t^s(y_t)+\eta^s\\ L^w=-\sum_{t=1}^{T}\log p_t^w(y_t)+\eta^w\\ L^d=-\sum_{t=1}^{T}\log(p_t(y_t))\end{cases} \tag{7}$$

$$L=L^s+L^w+L^d \tag{8}$$

其中L^s、L^w和L^d分别是语句级注意、单词级注意和双注意模型的损失,最终模型的训练损失是三个损失之和,η^s和η^w是双随机注意正则化项.

2.3 基于强化学习的训练方法

传统的图像描述模型在训练和测试之间存在差异.此外,模型使用交叉熵损失来进行训练,这与测试时使用的BLEU[15]和CIDEr[18]等评估度量并不相关.为了解决这些问题,可以通过强化学习方法直接在评估度量上优化模型语句的生成.其中LSTM解码器可以看作是与外部环境进行交互的代理,而策略p_θ由模型的参数θ来定义.该模型接收各种状态,并通过策略来确定下一步动作$p_\theta: y_t \sim p_\theta$,即预测下一个生成的单词.当语句生成结束时,可以通过计算所生成语句的评估度量分数来获得语句级的奖励$r(\boldsymbol{Y})$,LSTM更新其内部"状态".基于强化学习方法的训练目标是最大限度地减小负奖励期望:

$$L(\theta)=-E_{\boldsymbol{Y}\sim p_\theta}[r(\boldsymbol{Y})]\approx -r(\boldsymbol{Y}) \tag{9}$$

梯度$\nabla_\theta L(\theta)$通过REINFORCE算法[19]来计算.实际上,对于每个小批量的训练样本,可以使用来自p_θ的蒙特卡罗样本$\boldsymbol{Y}=\left(y_1,y_2,\cdots,y_T\right)$来估计梯度:

$$\begin{aligned}\nabla_\theta L(\theta)&=-E_{\boldsymbol{Y}\sim p_\theta}[r(\boldsymbol{Y})\nabla_\theta \log p_\theta(\boldsymbol{Y})]\\&\approx -r(\boldsymbol{Y})\nabla_\theta \log p_\theta(\boldsymbol{Y})\end{aligned} \tag{10}$$

然而,蒙特卡罗采样方法通常具有较大的方差.为了减小方差,遵循SCST(Self-Critical Sequence Training)方法[14],将模型通过蒙特卡罗采样生成的语句的CIDEr得分作为奖励$r(\boldsymbol{Y})$,使用模型贪心搜索生成的语句的CIDEr得分作为基线奖励$r(\hat{\boldsymbol{Y}})$.不同之处在于,训练时将强化学习方法应用于两阶段的训练,一阶段是优化单注意模型生成的描述,使用单注意模型自身生成的语句来作为基线,二阶段是优化模型最终生成的描述,不仅使用模型最终生成的语句来作为基线,而且还将一阶段中两个单注意模型生成的语句也作为基线.可表示为:

$$\begin{aligned}\nabla_\theta L(\theta)\approx -(&\Delta r(\boldsymbol{Y}_{1,s})\nabla_\theta \log p_\theta(\boldsymbol{Y}_{1,s})\\&+\Delta r(\boldsymbol{Y}_{1,w})\nabla_\theta \log p_\theta(\boldsymbol{Y}_{1,w})\\&+\Delta r(\boldsymbol{Y}_2)\nabla_\theta \log p_\theta(\boldsymbol{Y}_2))\end{aligned} \tag{11}$$

其中,

$$\begin{cases}\Delta r(\boldsymbol{Y}_{1,s})=r(\boldsymbol{Y}_{1,s})-r(\hat{\boldsymbol{Y}}_{1,s})\\ \Delta r(\boldsymbol{Y}_{1,w})=r(\boldsymbol{Y}_{1,w})-r(\hat{\boldsymbol{Y}}_{1,w})\\ \Delta r(\boldsymbol{Y}_2)=\dfrac{1}{3}[(r(\boldsymbol{Y}_2)-r(\hat{\boldsymbol{Y}}_2))+(r(\boldsymbol{Y}_2)-r(\hat{\boldsymbol{Y}}_{1,s}))+(r(\boldsymbol{Y}_2)-r(\hat{\boldsymbol{Y}}_{1,w}))]\end{cases} \tag{12}$$

其中$\boldsymbol{Y}_{1,s}$,$\boldsymbol{Y}_{1,w}$,$\boldsymbol{Y}_2$分别是模型通过蒙特卡罗抽样在一阶段生成的两个语句和最终生成的语句,$\hat{\boldsymbol{Y}}_{1,s}$,$\hat{\boldsymbol{Y}}_{1,w}$,$\hat{\boldsymbol{Y}}_2$是通过贪心搜索获得的语句.

3 实验结果分析

为了证明提出方法的有效性,在MSCOCO数据集上进行了充分的实验评估,并与当前先进的模型进行了对比,从定量和定性两个方面进行了结果分析.

3.1 数据集与实施细节

本文选择最常用的MSCOCO数据集来进行实验验证.MSCOCO数据集中共113287幅图像作为训练集,但测试集中没有标注语句.根据Karpathy的划分[5],从验证集中挑选出5000幅图像用于验证,5000幅图像用于测试.实验使用BLEU(1-4)[15],METEOR[16],ROUGE-L[17]和CIDEr[18]作为评估指标来衡量生成语句的质量.

实验实现了三种基于注意的图像描述模型:单词级注意模型、语句级注意模型和双注意模型,并且在双注意模型中对提出的两种不同构建方式分别进行了验证.这些实验基本遵循相同的参数设置.首先使用在ImageNet上预训练过的ResNet-101[20]对图像进行编码,然后在目标数据集上对参数进行微调.将CNN最后一个卷积层的输出作为图像视觉编码特征,并使用空间自适应平均池化将它们调整为2048×14×14的固定大小输出.同时,通过目标检测器来提取图像区域特征,获得图像显著区域的特征表示,并使用加权方法融合两种特征.实验中LSTM的神经元数量统一设置为512,语句级注意模型中三个1×1卷积层的神经元分别设置为64、64、512,单词级注意层中的神经元数量设置为512.双随机注意正则化系数设置为1.除了CNN部分之外,其他部分的参数采用随机初始化.

在训练过程中,首先关闭CNN的微调功能,使用Adam优化器[21]在交叉熵损失下训练模型.初始学习率为4×10^{-4},动量参数为0.9,批量大小为100,共训练25轮.然后打开CNN微调功能,在之前的基础上继续训练,此时批量大小调整为32,训练10轮.最后,在训练好的模型上运行基于强化学习的训练方法来优化模型的CIDEr评估指标.在这个阶段,学习率设置为5×10^{-5},批量大小为64,训练15轮.

3.2 微调CNN效果分析

在训练过程中,首先固定CNN的参数,在交叉熵损失下训练模型达到最好的效果,然后在先前训练的模型上对CNN进行微调训练.表1展示了本文模型在MSCOCO数据集上CNN微调前和微调后结果的对比(FT表示微调后的结果),其中SAIC表示语句级注意模型,WAIC表示单词级注意模型,DAIC表示双注意模型(其中DAICs表示串联方式构建的双注意模型,DAICp表示并联方式构建的双注意模型).从表1中可以看出,微调CNN对所有的模型都有非常大的提升.

表1　在MSCOCO数据集上微调CNN对结果的影响

APPROACH	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER
SAIC	73.6	56.8	42.6	32.7	25.9	54.1	100.4
SAIC(FT)	75.5	58.9	44.8	34.8	27.2	55.7	107.2
WAIC	73.3	56.9	42.8	33.0	26.3	54.3	101.7
WAIC(FT)	75.8	59.1	45.4	35.6	27.5	55.9	108.6
DAICs	74.2	58.0	44.1	34.2	26.5	54.6	105.4
DAICs(FT)	75.5	59.6	45.9	36.1	27.6	56.0	109.7
DAICp	74.8	58.1	44.4	34.7	26.5	54.9	105.7
DAICp(FT)	**76.1**	**60.2**	**46.7**	**36.5**	**27.9**	**56.4**	**111.3**

3.3　生成结果定量分析

表2列出了本文模型在MSCOCO数据集上的性能表现以及与其他先进模型的对比,其中XE表示使用交叉熵损失训练后的结果,RL表示使用强化学习优化后的结果.

从表中实验数据可以看出,本文模型相比于其他先进模型有更好的性能表现. 首先,对于在交叉熵损失训练下的结果,虽然语句级注意模型SAIC只有语句级的特征,但依然取得了很好的效果,这是由于语句级注意特征能很好地表达图像中最主要的信息,并且通过微调CNN可以更好地发挥语句级注意的特点,得到更好的语句级特征表示. 但是从相比于单词级注意模型,语句级注意模型的整体效果还是会稍差,这是由于语句级注意特征由图像本身计算,不能随着上下文信息的变动来分析图像,而且在每个时间步都输入整个语句级的注意特征,不可避免的会引入更多的噪声. 但是,可以看到将其与单词级注意模型协同,可以生成更好的描述. 与单注意模型相比,双注意模型DAIC在所有指标上都取得了更好的表现,充分表明语句级和单词级注意模型的结合可以显著提高图像描述系统的性能.

表2　在MSCOCO数据集上的性能比较

APPROACH	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER
GOOGLE NIC[4]	66.6	46.1	32.9	24.6	–	–	–
SOFT-ATTENTION[7]	70.7	49.2	34.4	24.3	23.9	–	–
GLSTM[22]	67.0	49.1	35.8	26.4	22.74	–	81.25
BI-LSTM[23]	67.2	49.2	35.2	24.4	–	–	–
RA+SS[24]	72.4	55.5	41.8	31.3	24.8	53.2	95.5
ATT[10]	70.9	53.7	40.2	30.4	24.3	–	–
SCA-CNN[8]	71.9	54.8	41.1	31.1	25.0	–	–
ADAPTIVE[9]	74.2	58.0	43.9	33.2	26.6	–	108.5
ARNET[25]	74.0	57.6	44.0	33.5	26.1	54.6	103.4
PG-BCMR[26]	75.4	59.1	44.5	33.2	25.7	55.0	101.3
SCST:ATT2ALL[14]	–	–	–	34.2	26.7	55.7	114.0
UP-DOWN[27]	79.8	–	–	36.3	27.7	56.9	120.1
HEK[28]	79.3	63.8	49.0	37.3	27.3	57.4	121.2
SAIC(XE)	75.5	58.9	44.8	34.8	27.2	55.7	107.2
WAIC(XE)	75.8	59.1	45.4	35.6	27.5	55.9	108.6
DAIC(XE)	76.1	60.2	46.7	36.5	27.9	56.4	111.3
SAIC(RL)	78.9	62.8	47.6	36.1	27.7	57.0	117.7
WAIC(RL)	79.6	63.5	48.5	36.9	28.0	57.3	119.6
DAIC(RL)	**80.4**	**64.6**	**49.3**	**37.8**	**28.2**	**57.9**	**121.8**

3.4　生成结果定性分析

为了更加直观地证明本文方法可以捕获到准确的视觉注意特征,将方法的注意权重进行了可视化. 首先将注意权重扩张24倍,并使用高斯滤波器将其调整到与输入图像相同的尺寸,然后将注意分布图叠加在原始输入图像上. 如图5所示,W表示单词级注意模型的生成结果,S表示语句级注意模型的生成结果,D表示双注意模型的生成结果. 从图中可以看出,语句级注意

可以针对生成的语句关注到对图像中最有用的部分，而不包含语义信息的图像边缘等区域通常不会分配注意，而单词级注意可以准确地关注到图像中的各个局部实体．此外，也可以看到比于单词级和语句级注意模型，双注意模型生成的描述相对会更好．例如，在图5的前三组图像中，生成的“with wine”，“on a runway”和“red and white”这些短语都使描述语句更加丰富和准确．但同时由于双注意模型结合了两个单注意模型，有时也会因为单个模型的错误，而导致最终生成的语句不好，例如在第四组图像中生成的“frisbee”，但从定量分析的各种评价指标上看，整体上双注意模型有更好的生成效果．

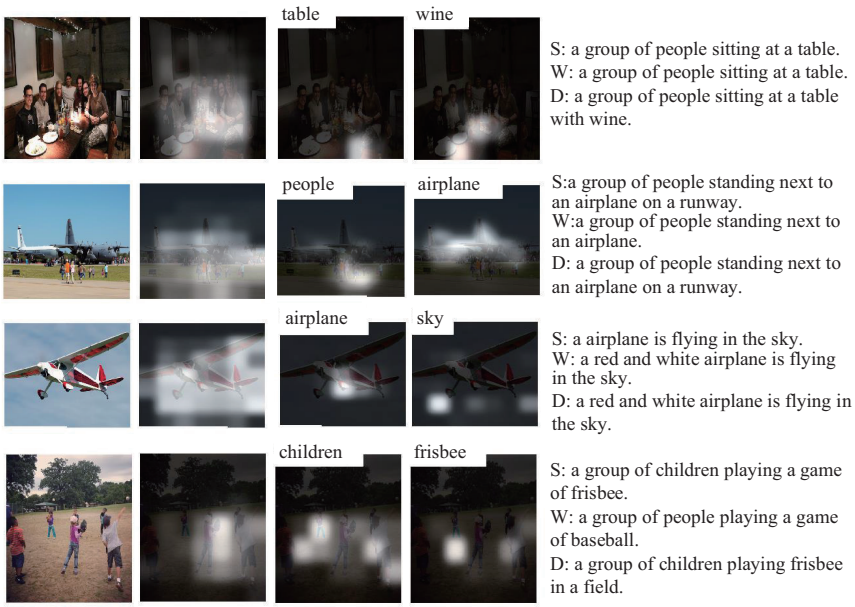

图5 生成结果的可视化

4 结束语

本文首先提出一种基于语句级注意的图像描述生成方法，引入自注意机制提取具有更完整视觉表示的语句级图像特征．在此基础上，进一步提出了双注意模型，通过对语句级注意和单词级注意的集成使最终模型有更好的生成效果．此外，本文将强化学习方法运用到两阶段的训练来优化模型的评估度量，大大提升了模型的整体性能．实验结果表明，双注意模型可以生成更好的描述语句，并且在各项评估指标上优于目前许多先进的模型．当然，也要看到本文方法需要大量的标注样本，在小数据集上的性能不是很令人满意．未来打算引入半监督学习和因果推理等技术进一步提升图像描述生成的性能．

参考文献

[1] 李志欣, 魏海洋, 张灿龙, 等. 图像描述生成研究进展[J]. 计算机研究与发展, 2021, 58(9): 1951-1974.
LI Zhi-xin, WEI Hai-yang, ZHANG Can-long, et al. Research progress on image captioning[J]. Journal of Computer Research and Development, 2021, 58(9): 1951-1974. (in Chinese)

[2] DAI J, LI Y, HE K, et al. R-FCN: Object detection via region-based fully convolutional networks[C]//Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2016: 379-387.

[3] LI Zhi-xin, LIN Lan, ZHANG Can-long, et al. A semi-supervised learning approach based on adaptive weighted fusion for automatic image annotation [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(1): article37.

[4] 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. Los Alamitos, USA: IEEE Computer Society, 2015: 3156-3164.

[5] 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. Los Alamitos, USA: IEEE Computer Society, 2015: 3128-3137.

[6] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks(m-RNN)[EB/OL]. [2021-09-22]. https://arxiv.org/abs/1412.6632.

[7] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//Proceedings of International Conference on Machine Learning. Cambridge, USA: MIT Press, 2015: 2048-2057.

[8] 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 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA: IEEE Computer Society, 2017: 6298-6306.

[9] 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 IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA: IEEE Computer Society, 2017: 3242-3250.

[10] YOU Q, JIN H, WANG Z, et al. Image captioning with semantic attention[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA: IEEE Computer Society, 2016: 4651-4659.

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

[12] YU A W, DOHAN D, LUONG M T, et al. QANet: Combining local convolution with global self-attention for reading comprehension[EB/OL]. [2021-09-22]. https://arxiv.org/abs/1804.09541.

[13] RANZATO M A, CHOPRA S, AULI M, et al. Sequence level training with recurrent neural networks[EB/OL]. [2021-09-22]. https://arxiv.org/abs/1511.06732.

[14] 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. Los Alamitos, USA: IEEE Computer Society, 2017: 1179-1195.

[15] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA: ACL, 2002: 311-318.

[16] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Stroudsburg, USA: ACL, 2005: 65-72.

[17] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of the ACL Workshop on Text Summarization Branches Out. Stroudsburg, USA: ACL, 2004: 74-81.

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

[19] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8(3-4): 229-256.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA: IEEE Computer Society, 2016: 770-778.

[21] DIEDERIK K, JIMMY B. ADAM: A method for stochastic optimization[EB/OL]. [2021-09-22]. https://arxiv. org/abs/1412.6980.

[22] JIA X, 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. Piscataway, USA: IEEE, 2015: 2407-2415.

[23] WANG C, YANG H, BARTZ C, et al. Image captioning with deep bidirectional LSTMs[C]//Proceedings of the 24th ACM International Conference on Multimedia. New York, USA: ACM, 2016: 988-997.

[24] FU K, JIN J, CUI R, et al. Aligning where to see and what to tell: Image caption with region-based attention and scene-specific contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2321-2334.

[25] CHEN X, MA L, JIANG W, et al. Regularizing RNNs for caption generation by reconstructing the past with the present[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos, USA: IEEE Computer Society, 2018: 7995-8003.

[26] LIU S, ZHU Z, YE N, et al. Improved image captioning via policy gradient optimization of SPIDEr[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2017: 873-881.

[27] 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. Los

Alamitos, USA: IEEE Computer Society, 2018: 6077-6086.

[28] HUANG Fei-cheng, LI Zhi-xin, WEI Hai-yang, et al. Boost image captioning with knowledge reasoning [J]. Machine Learning, 2020, 109(12): 2313-2332.

作者简介

卓亚琦 女,1976年6月出生,陕西咸阳人. 桂林理工大学理学院讲师. 研究方向为图像理解与机器学习.

E-mail: zhuoyaqi@126.com

魏家辉 男,1993年11月出生,山东东营人. 广西师范大学计算机科学与工程学院博士研究生. 研究方向为图像理解与机器学习.

E-mail: weijh@stu.gxnu.edu.cn

李志欣(通讯作者) 男,1971年10月出生,广西桂林人. 现为广西师范大学计算机科学与工程学院教授、博士生导师. 研究领域为图像理解、机器学习与跨媒体计算.

E-mail: lizx@gxnu.edu.cn