ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 4,Jul. 2024,pp. 922－932
DOI:10. 16337/j. 1004-9037. 2024. 04. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合多时间维度视觉与语义信息的图像描述方法

陈善学，王　程

(重庆邮电大学通信与信息工程学院,重庆 400065)

摘　要:传统的图像描述方法仅使用当前时刻的视觉信息和语义信息来生成预测词,而没有考虑过去时刻的视觉信息和语义信息,从而导致模型输出的信息在时间维度上比较单一,因此生成的描述语句在准确性上有所欠缺。针对此问题,提出一种融合多时间维度视觉与语义信息的图像描述方法,有效地融合了过去时刻的视觉信息和语义信息,并设计一种门控机制动态地对两种信息进行选择利用。在MSCOCO数据集上进行实验验证,结果表明该方法能够更准确地生成描述语句,和当前最主流的图像描述方法进行对比,性能在各项评价指标上都得到了可观的提升。

关键词:图像描述;视觉信息;语义信息;时间维度;门控机制

中图分类号:TP391　　**文献标志码**:A

Image Captioning Method for Fusing Multi-temporal Dimensional Visual and Semantic Information

CHEN Shanxue, WANG Cheng

(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Traditional image captioning methods use only the visual and semantic information of the current moment to generate prediction words without considering the visual and semantic information of the past moments, which leads to the output of the model to be relatively homogeneous in terms of temporal dimension. As a result, the generated captioning is lacking in terms of accuracy. To address this problem, an image captioning method that fuses multi-temporal dimensional visual and semantic information is proposed, which effectively fuses visual and semantic information of past moments and designs a gating mechanism to dynamically select both kinds of information. Experimental validation on the MSCOCO dataset shows that the method is able to generate captioning more accurately, and the performance is considerably improved in all evaluation metrics when compared with the most current state-of-the-art image captioning methods.

Key words: image captioning; visual information; semantic information; temporal dimension; gating mechanism

收稿日期:2023-05-24;**修订日期**:2023-06-26

引 言

图像描述[1]的任务是给定一张图像,计算机可以自动生成一句合理、通顺并且连贯的自然语言描述。它是计算机视觉[2]和自然语言处理[3]交叉领域中一个极具挑战性的任务。近年来,随着技术的不断发展,图像描述为图像检索[4]、人机交互[5]和视觉问答[6]等领域的应用带来了重大的意义。

早期的图像描述方法主要包括基于模版的方法和基于检索的方法。基于模版的方法首先使用相关算法提取图像的特征,然后根据图像特征检测图像中的目标词和属性词等重要单词,然后将这些单词填充到模版中。Farhadi等[7]首先检测出图像中目标词、关系词和属性词,最后形成对应的三元组,最后找到合适的单词填充到模版的插槽中。基于检索的方法类似于图像检索问题,它首先通过提取到的图像特征计算图像之间的相似度,然后找到数据集中相似图像的描述用来生成最终的描述。Kuznetsova等[8]首先在数据集中检索出和目标图像相似的图像,然后将其对应的描述语句通过随机树型结构算法提取出词组来生成自然语言描述。显而易见,上述方法对模型的优化方式有限,因此生成的描述语句质量并不高。

受益于深度学习的发展,研究者们发现使用深度学习的方法可以更好地解决这项任务。受机器翻译的启发,编码器-解码器的架构在图像描述领域中得到了广泛使用。Mao等[9]提出了一种基于编码器-解码器的图像描述模型,称为m-RNN(Multimodal recurrent neural network),该模型使用卷积神经网络(Convolutional neural network,CNN)提取图像的特征,然后将提取到的特征输入到循环神经网络(Recurrent neural network,RNN)中来生成预测词。Vinyals等[10]提出了一种NIC(Neural image caption)模型,采用CNN和长短期记忆网络(Long short-term memory,LSTM)作为编码器和解码器,LSTM可以有效地解决梯度消失和梯度爆炸的问题,因此LSTM也成为图像描述领域中主流的解码器。

后续的研究发现视觉的关注对于图像的时空连贯性非常重要,于是注意力机制被广泛应用到图像描述领域中。Xu等[11]提出将注意力机制引入到图像描述模型中,首先在编码端划分图像的区域,然后将LSTM输出的隐藏状态和图像特征输入到注意力机制中来决定图像中不同区域的权重,该方法可以动态地选择感兴趣的区域来指导预测词的生成。Chen等[12]提出了一种SCA(Spatial and channel-wise attention)模型,该模型使用了层级注意力机制动态地更新CNN提取的图像特征,并且融合了通道注意力机制和空间注意力机制。Lu等[13]提出了一种自适应注意力模型,该模型可以自动决定生成预测词时依赖语义信息还是图像信息。Anderson等[14]提出了一种Up-Down模型,该模型在编码端首先使用Faster R-CNN来提取图像的特征,Faster R-CNN可以提取局部的图像特征而不是全局的图像特征,因此可以避免一些不重要区域的干扰。

主流的图像描述模型一般通过计算交叉熵损失函数来训练模型,但是这种训练方式在模型训练阶段和测试阶段都存在曝光误差,即模型在训练和测试阶段的输入不同会造成误差积累,这样会导致生成的描述语句和图像不符,并且交叉熵损失函数不能直接对不可微的评价指标进行微分运算。为了解决这些问题,研究人员提出使用强化学习[15]的方法来解决图像描述问题。Rennie等[16]基于强化学习提出了一种自我评价序列的训练方式(Self-critical sequence training,SCST),该方法将推理阶段得到的描述语句作为基线,直接对CIDEr评价指标[17]进行优化。此外,基于Transformer的方法也逐渐应用于图像描述任务中。Zhang等[18]在解码端使用Transformer作为解码器,并使用自适应模块来自适应地选择视觉特征和非视觉特征以生成预测单词。Wang等[19]将Swin Transformer应用到图像描述工作中,通过融合当前时刻的预测词和全局视觉特征以生成更好的描述语句。

目前的图像描述方法仍然有不足的地方。在解码端仅使用当前时刻的视觉信息和语义信息来生成预测词,而没有考虑过去时刻的视觉信息和语义信息,从而导致生成的描述语句质量并不高。针对

此问题，本文提出了一种融合多时间维度视觉与语义信息的图像描述方法（Fusion of multi-temporal dimensional visual and semantic information，FMDVSI）。该方法提出了视觉注意力和语义注意力，有效地融合了过去时刻的视觉信息和语义信息，并设计了一种门控机制来对两种信息进行选择利用。

1 本文模型结构

本文以文献[14]提出的模型作为基准模型，提出了一种融合多时间维度视觉与语义信息的图像描述方法，模型的整体框图如图1所示。在编码端同时提取图像的全局场景特征和局部显著特征，并把两种特征进行融合后输入到解码端。在解码端首先通过视觉特征解码模块对图像的特征进行解码，然后通过视觉注意力模块和语义注意力模块分别对过去时刻的视觉信息和语义信息进行融合，最后通过语言模块输出描述语句。

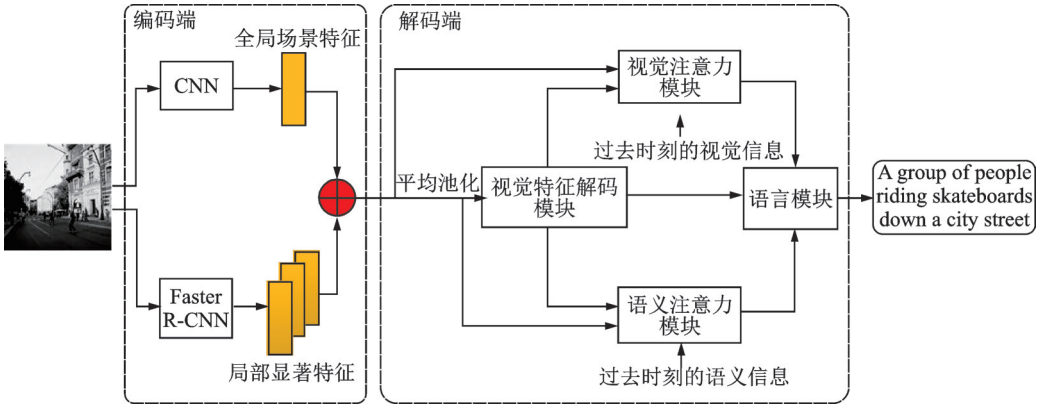

图1 本文模型整体结构

Fig.1 Overall structure of the proposed model

1.1 整体模型

图像描述模型的目标是输入一张给定的图像 I，输出一句能够表达图像内容的描述序列 $Y=\{y_1,y_2,\cdots,y_T\}$，$y_t\in\mathbf{R}^D$，其中 D 为词典集合的大小，T 为描述序列的最大长度。

在编码端，采用在 Visual Genome 数据集[20]上预训练的 Faster R-CNN 来提取一组区域特征作为图像的局部显著特征，可以表示为 $V=\{v_1,v_2,\cdots,v_L\}$，其中 L 为每张图像提取特征的数量，$v_i\in\mathbf{R}^d$ 表示每个区域的图像特征，d 为特征向量的维度大小。同时采用 ResNet-101 网络中最后一个卷积层的输出作为图像的全局场景特征 v_g。编码器的输出为局部显著特征与全局场景特征的融合，可以表示为 $V=\{v_1,v_2,\cdots,v_L,v_g\}$，即将局部显著特征向量和全局显著特征向量进行拼接操作。

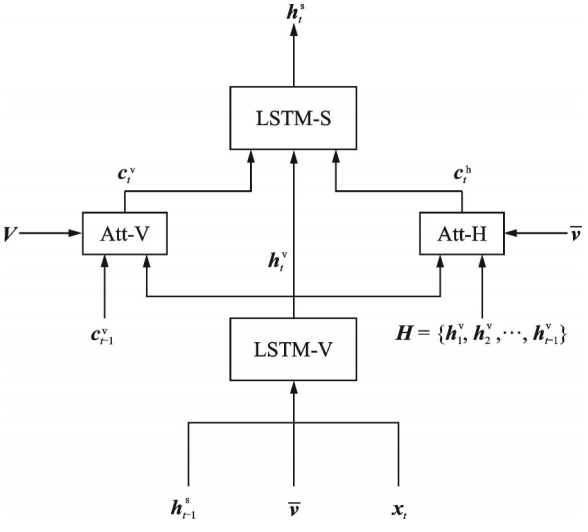

图2 解码端结构

Fig.2 Structure of decoding side

解码端结构如图2所示。t 时刻时，将编码端

提取到的图像特征 $\boldsymbol{V}$ 经过均值处理得到平均图像特征 $\bar{\boldsymbol{v}}$。并结合单词的嵌入向量 $\boldsymbol{x}_t$ 和上一时间步语言LSTM(LSTM-S)的隐藏状态 $\boldsymbol{h}_{t-1}^{\mathrm{s}}$ 共同输入到视觉解码LSTM(LSTM-V)中,输出的结果作为语义信息。其计算步骤如下

$$\bar{\boldsymbol{v}}=\frac{1}{L}\sum_{i=1}^{L}\boldsymbol{v}_i \tag{1}$$

$$\boldsymbol{h}_t^{\mathrm{v}}=\mathrm{LSTM}_{\mathrm{v}}\left(\left[\bar{\boldsymbol{v}},\boldsymbol{x}_t,\boldsymbol{h}_{t-1}^{\mathrm{s}}\right],\boldsymbol{h}_{t-1}^{\mathrm{v}}\right) \tag{2}$$

式中:$\boldsymbol{h}_t^{\mathrm{v}}\in\mathbf{R}^h$ 为视觉解码LSTM的隐藏状态,h 为隐藏状态的维度;$[\cdot]$ 表示向量之间的拼接操作;$\mathrm{LSTM}_{\mathrm{v}}(\cdot)$ 表示视觉解码LSTM的运算。

视觉注意力机制将视觉解码LSTM输出的隐藏状态 $\boldsymbol{h}_t^{\mathrm{v}}$、图像特征 $\boldsymbol{V}$ 以及上一个时间步的视觉上下文向量 $\boldsymbol{c}_{t-1}^{\mathrm{v}}$ 作为输入,得到当前时间步的视觉上下文向量 $\boldsymbol{c}_t^{\mathrm{v}}$。其计算过程可表示为

$$\boldsymbol{c}_t^{\mathrm{v}}=f_{\text{Att-V}}\left(\boldsymbol{h}_t^{\mathrm{v}},\boldsymbol{V},\boldsymbol{c}_{t-1}^{\mathrm{v}}\right) \tag{3}$$

式中 $f_{\text{Att-V}}(\cdot)$ 表示视觉注意力机制。

语义注意力机制将视觉解码LSTM输出的隐藏状态 $\boldsymbol{h}_t^{\mathrm{v}}$、平均图像特征 $\bar{\boldsymbol{v}}$ 以及过去时刻视觉解码LSTM输出的隐藏状态集合 $\boldsymbol{H}=\left\{\boldsymbol{h}_1^{\mathrm{v}},\boldsymbol{h}_2^{\mathrm{v}},\cdots,\boldsymbol{h}_{t-1}^{\mathrm{v}}\right\}$ 作为输入,得到当前时间步的语义上下文向量 $\boldsymbol{c}_t^{\mathrm{h}}$。其计算过程可表示为

$$\boldsymbol{c}_t^{\mathrm{h}}=f_{\text{Att-H}}\left(\boldsymbol{h}_t^{\mathrm{v}},\bar{\boldsymbol{v}},\boldsymbol{H}\right) \tag{4}$$

式中 $f_{\text{Att-H}}(\cdot)$ 表示语义注意力机制。

为了有效地融合视觉信息和语义信息,设计了一种门控机制来对视觉上下文向量 $\boldsymbol{c}_t^{\mathrm{v}}$ 和语义上下文向量 $\boldsymbol{c}_t^{\mathrm{h}}$ 进行选择利用。并将得到的结果和视觉解码LSTM输出的隐藏状态 $\boldsymbol{h}_t^{\mathrm{v}}$ 共同输入到语言LSTM中,以得到语言LSTM输出的隐藏状态 $\boldsymbol{h}_t^{\mathrm{s}}$。其计算过程可表示为

$$\hat{\boldsymbol{c}}_t=G_t\left(\boldsymbol{c}_t^{\mathrm{v}},\boldsymbol{c}_t^{\mathrm{h}}\right) \tag{5}$$

$$\boldsymbol{h}_t^{\mathrm{s}}=\mathrm{LSTM}_{\mathrm{s}}\left(\left[\hat{\boldsymbol{c}}_t,\boldsymbol{h}_t^{\mathrm{v}}\right],\boldsymbol{h}_{t-1}^{\mathrm{s}}\right) \tag{6}$$

式中:$G_t(\cdot)$ 表示门控机制;$\mathrm{LSTM}_{\mathrm{s}}(\cdot)$ 表示语言LSTM的运算。

最后通过语言LSTM输出的隐藏状态 $\boldsymbol{h}_t^{\mathrm{s}}$ 来生成预测词,表达式为

$$y_t\sim p_t=\mathrm{softmax}\left(\boldsymbol{W}_{\mathrm{p}}\boldsymbol{h}_t^{\mathrm{s}}\right) \tag{7}$$

式中:$\boldsymbol{W}_{\mathrm{p}}\in\mathbf{R}^{h\times h}$ 为权重矩阵;p_t 为预测单词的概率。

1.2 视觉注意力机制

为了有效融合过去时刻的视觉信息,本文在传统注意力机制的基础上提出了一种视觉注意力机制的方法,其结构如图3所示。该方法引入了自适应注意力[13],并且在每一时间步嵌入上一时间步的视觉上下文信息作为过去时刻的视觉信息,以确保图像信息在视觉上的连贯性。本节将详细介绍视觉注意力机制的计算过程。

在 t 时刻,将 $\boldsymbol{c}_{t-1}^{\mathrm{v}}$ 和 $\boldsymbol{h}_t^{\mathrm{v}}$ 进行拼接操作,并将其与图像的特征 $\boldsymbol{V}$ 进行融合,以得到图像的视觉信息分布 $\boldsymbol{c}_t$,该过程计算步骤如下

$$\boldsymbol{u}_{i,t}=\boldsymbol{w}_{\mathrm{u}}^{\mathrm{T}}\tanh\left(\boldsymbol{W}_{\mathrm{v}}\boldsymbol{v}_i+\boldsymbol{W}_{\mathrm{h}}\left[\boldsymbol{h}_t^{\mathrm{v}},\boldsymbol{c}_{t-1}^{\mathrm{v}}\right]\right) \tag{8}$$

$$\boldsymbol{\alpha}_{i,t}=\mathrm{softmax}\left(\boldsymbol{u}_{i,t}\right) \tag{9}$$

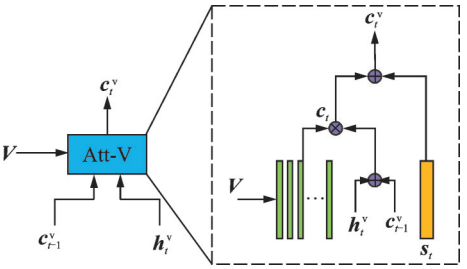

图3 视觉注意力结构

Fig.3 Structure of visual attention

$$\boldsymbol{c}_t=\sum_{i=1}^{L}\boldsymbol{\alpha}_{i,t}\boldsymbol{v}_i \tag{10}$$

式中：$\boldsymbol{w}_{\mathrm{u}}^{\mathrm{T}}\in\mathbf{R}^{1\times h}$，$\boldsymbol{W}_{\mathrm{v}}\in\mathbf{R}^{h\times d}$，$\boldsymbol{W}_{\mathrm{h}}\in\mathbf{R}^{h\times 2h}$为权重矩阵。

此过程同时引入一种“视觉哨兵”[13]的概念。视觉哨兵可以在注意力机制中决定对视觉信息的关注程度。在视觉解码LSTM中添加语义信息向量$\boldsymbol{s}_t$，其计算步骤如下

$$\boldsymbol{z}_t=\sigma\left(\boldsymbol{W}_{zx}\boldsymbol{x}_t+\boldsymbol{W}_{zh}\boldsymbol{h}_{t-1}^{\mathrm{v}}\right) \tag{11}$$

$$\boldsymbol{s}_t=\boldsymbol{z}_t\odot\tanh(\boldsymbol{m}_t) \tag{12}$$

式中：$\boldsymbol{W}_{zx}\in\mathbf{R}^{h\times h}$和$\boldsymbol{W}_{zh}\in\mathbf{R}^{h\times h}$为权重矩阵；$\odot$表示逐元素相乘；$\boldsymbol{m}_t$为视觉解码LSTM中的记忆单元；$\sigma(\cdot)$表示sigmoid激活函数。

视觉注意力机制输出的上下文向量$\boldsymbol{c}_t^{\mathrm{v}}$可以通过以下步骤获得

$$\boldsymbol{c}_t^{\mathrm{v}}=\beta_t\boldsymbol{s}_t+\left(1-\beta_t\right)\boldsymbol{c}_t \tag{13}$$

式中$\beta_t\in[0,1]$可以看做一个在t时刻的权衡变量，它随着时刻的变化而变化，并且可以自适应地选择关注图像的视觉信息向量$\boldsymbol{c}_t$还是关注语义信息向量$\boldsymbol{s}_t$。

为了计算β_t，需要对注意力变量$\boldsymbol{u}_{i,t}$进行更新得到一个新的注意力变量$\hat{\boldsymbol{u}}_{i,t}\in\mathbf{R}^{L+1}$。并取$\hat{\boldsymbol{u}}_{i,t}$中最后一个元素作为$\beta_t$的值，其计算步骤如下

$$\hat{\boldsymbol{u}}_{i,t}=\mathrm{softmax}\left(\left[\boldsymbol{u}_{i,t},\boldsymbol{w}_{\mathrm{u}}^{\mathrm{T}}\tanh\left(\boldsymbol{W}_{\mathrm{s}}\boldsymbol{s}_t+\boldsymbol{W}_{\mathrm{H}}\boldsymbol{h}_t^{\mathrm{v}}\right)\right]\right) \tag{14}$$

$$\beta_t=\hat{\boldsymbol{u}}_{i,t}[L+1] \tag{15}$$

式中：$\boldsymbol{W}_{\mathrm{s}}\in\mathbf{R}^{h\times h}$，$\boldsymbol{W}_{\mathrm{H}}\in\mathbf{R}^{h\times h}$为权重矩阵。

1.3 语义注意力机制

过去时刻的语义信息可以为模型在预测语义关系词时提供信息补充。为了有效融合过去时刻的语义信息，本文提出了一种语义注意力机制的方法，其结构如图4所示。该方法在每个时间步对过去时刻的语义信息进行关注，为模型的预测提供充足的语义信息，以确保预测单词的准确性。

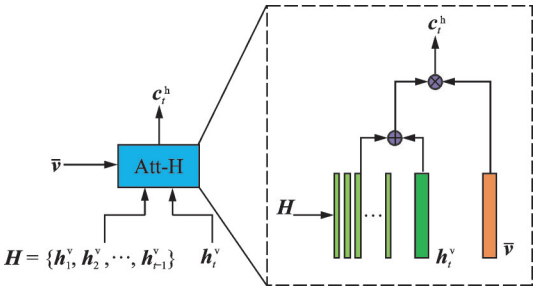

图4 语义注意力结构

Fig.4 Structure of semantic attention

在每个时间步，将当前时刻的语义信息$\boldsymbol{h}_t^{\mathrm{v}}$与过去时刻的语义信息$\boldsymbol{H}=\left\{\boldsymbol{h}_1^{\mathrm{v}},\boldsymbol{h}_2^{\mathrm{v}},\cdots,\boldsymbol{h}_{t-1}^{\mathrm{v}}\right\}$进行拼接得到一个新的语义信息集合$\left\{\boldsymbol{h}_1^{\mathrm{v}},\boldsymbol{h}_2^{\mathrm{v}},\cdots,\boldsymbol{h}_i^{\mathrm{v}}\right\}$。并通过图像的平均特征$\bar{\boldsymbol{v}}$对语义信息集合进行关注。其详细计算步骤如下

$$\boldsymbol{\gamma}_{i,t}=\boldsymbol{w}_{\gamma}^{\mathrm{T}}\tanh\left(\boldsymbol{W}_{\gamma\mathrm{v}}\bar{\boldsymbol{v}}+\boldsymbol{W}_{\gamma\mathrm{h}}\boldsymbol{h}_i^{\mathrm{v}}\right) \tag{16}$$

$$\boldsymbol{z}_{i,t}=\mathrm{softmax}\left(\boldsymbol{\gamma}_{i,t}\right) \tag{17}$$

$$\boldsymbol{c}_t^{\mathrm{h}}=\sum_{i=1}^{L}\boldsymbol{z}_{i,t}\boldsymbol{h}_i^{\mathrm{v}} \tag{18}$$

式中：$\boldsymbol{w}_{\gamma}^{\mathrm{T}}\in\mathbf{R}^{1\times h}$，$\boldsymbol{W}_{\gamma\mathrm{v}}\in\mathbf{R}^{h\times d}$，$\boldsymbol{W}_{\gamma\mathrm{h}}\in\mathbf{R}^{h\times h}$为权重矩阵。

1.4 门控机制

为了控制视觉注意力机制和语义注意力机制的输出，受文献[21,22]的启发，提出了一种门控机制。它可以在每个时间步动态地决定视觉信息和语义信息的权重。即当模型输出视觉目标词时，则更

多依赖于视觉信息;当模型输出语义关系词时,则更多依赖于语义信息。

门控机制以视觉上下文向量c_t^{v}和语义上下文向量c_t^{h}作为输入,并对其进行更新。其详细计算步骤如下

$$g_t=\sigma\left(W_{\mathrm{V}}c_t^{\mathrm{v}}+W_{\mathrm{H}}c_t^{\mathrm{h}}\right) \tag{19}$$

$$\hat{c}_t=\left[c_t^{\mathrm{v}}\odot\left(1-g_t\right),c_t^{\mathrm{h}}\odot g_t\right] \tag{20}$$

式中$W_{\mathrm{V}}\in\mathbf{R}^{h\times d}$为权重矩阵。

1.5 训练目标

本文分两个阶段对模型进行训练。在第1阶段,使用交叉熵损失对模型进行训练,通过训练图像对应的参考描述来使交叉熵损失达到最小,即

$$L_{\mathrm{XE}}(\theta)=-\sum_{t=1}^{T}\log\left(p_\theta\left(y_t^*|y_{1:t-1}^*\right)\right) \tag{21}$$

式中:y_t^*为图像对应的参考描述;θ为模型对应的参数;L_{XE}表示使用交叉熵损失的训练方式。

在第2阶段,使用基于强化学习的SCST[16]训练方法对CIDEr的分数进行优化。其训练目标为最小化负奖励期望的分数,即

$$L_R(\theta)=-E_{Y\sim p_\theta}\left[r(Y)\right] \tag{22}$$

式中:$r(Y)$表示生成描述Y的CIDEr得分;$L_R(\theta)$表示使用强化学习的训练方式。

SCST训练方法可以有效地解决模型在训练阶段和测试阶段存在的曝光误差问题,使得模型的性能得到了极大的提升。

2 实验结果与分析

2.1 数据集与评价指标

本文使用MSCOCO 2014数据集[23]来证明模型的有效性。该数据集一共含有123 287张图像,其中训练集和验证集分别含有82 783和40 504张图像,并且每张图像都含有5个人工标记的描述语句。根据文献[24]的方式对数据集进行划分,从数据集中随机选择5 000张图像作为测试集,5 000张图像作为验证集,其余的113 287张图像全部作为训练集。

采用BLEU[25]、METEOR[26]、ROUGE-L[27]、CIDEr和SPICE[28]五种评价指标来衡量模型的性能。其中,BLEU_n比较生成描述语句与参考描述语句之间n元组的重合程度;METEOR计算单词之间准确率和召回率的调和平均值;ROUGE-L通过计算最长公共子序列来衡量句子的准确性;CIDEr计算生成描述语句与参考描述语句之间的余弦相似度;SPICE通过语句间场景图的目标、属性和关系来衡量语义的相似性。

2.2 实验细节设置

在实验中,图像特征的维度设置为2 048,LSTM的隐藏维度设置为1 024,词嵌入向量的维度设置为2 048。在训练过程中,首先使用交叉熵损失的方式训练30轮,使用Adam优化器[29]对模型进行优化,初始学习率设置为0.000 5,动量参数设置为0.9,每3轮衰减一次,衰减率设置为0.8;然后使用SCST训练方法再训练20轮,初始学习率设置为0.000 05,同样为每3轮衰减一次,衰减率设置为0.8。在测试过程中,使用束搜索[30]的方式进行解码,束尺寸设置为3。

2.3 实验结果对比

为了验证本文方法FMDVSI模型的有效性,使用Up-Down模型[14]作为基线模型,在MSCOCO数据集上进行实验。同时,为了显示实验的公平性,基线模型设置和本文FMDVSI模型相同的参数。实

验结果如表1所示,其中XE[31]表示使用交叉熵损失的方式进行训练,RL[16]表示使用基于强化学习的SCST方法对CIDEr评价指标进行优化。

表1 本文方法与基线模型在MSCOCO数据集上的性能对比

Table 1 Performance comparison of the proposed method and the baseline model on the MSCOCO dataset %

模型	BLEU_1	BLEU_4	METEOR	ROUGE-L	CIDEr	SPICE
基线模型(XE)	75.8	35.7	27.5	56.3	111.7	20.5
基线模型(RL)	79.2	36.6	28.0	57.6	121.4	21.5
FMDVSI(XE)	76.6	36.9	27.8	57.6	116.9	20.7
FMDVSI(RL)	80.2	37.8	28.6	58.4	125.2	21.9

从表1结果得知,在使用交叉熵损失训练下,本文模型在各个评价指标上的得分都略高于基线模型。在通过使用SCST方法对CIDEr评价指标进行优化后,模型的整体性能有了明显的提升,在各个评价指标上都优于基线模型。其中,本文的FMDVSI(RL)模型相对于基线模型(RL)BLEU_1、BLEU_4、METEOR、ROUGE-L和SPICE评价指标上的得分分别提高了1.0%、1.2%、0.6%、0.8%和0.4%,而CIDEr评价指标的得分则提高了3.8%,达到了125.2%。实验结果表明了融合过去时刻的视觉和语义信息有利于生成更准确的描述语句,验证了本文方法的有效性。

为了进一步验证多时间维度视觉与语义信息的有效性,本文进行了消融实验,结果如表2所示。其中,FMDVSI-AttV表示在基线模型的基础上只使用视觉注意力机制的模型;FMDVSI-AttH表示在基线模型的基础上只使用语义注意力机制的模型。视觉注意力机制和语义注意力机制可以分别对过去时刻的视觉信息和语义信息进行关注,从而融合多时间维度的视觉和语义信息。从表2中数据可以看出,在基线模型的基础上单独使用视觉注意力机制或者语义注意力机制的情况下,模型的性能均有所提升。当本文模型同时使用视觉注意力机制和语义注意力机制时,模型的性能有了明显的提升。实验结果验证了多时间维度的视觉和语义信息有利于提升模型的性能。

表2 消融实验

Table 2 Ablation experiment %

方法	BLEU_1	BLEU_4	METEOR	ROUGE-L	CIDEr	SPICE
基线模型	75.8	35.7	27.5	56.3	111.7	20.5
FMDVSI-AttV	76.1	36.0	27.7	56.5	113.1	20.6
FMDVSI-AttH	76.0	35.9	27.6	56.4	112.8	20.6
FMDVSI	76.6	36.9	27.8	57.6	116.9	20.7

同时,为了验证门控机制的有效性,本文将使用门控机制的方法与直接拼接视觉和语义上下文向量的方法进行了对比实验,结果如表3所示。可以看出,在引入门控机制后,本文模型相较于传统的拼接方法在CIDEr评价指标上的得分提高了3.7%,证明了门控机制可以有效地对视觉信息和语义信息进行动态选择。

表3 门控机制方法与拼接方法性能对比

Table 3 Performance comparison of gating mechanism method and splicing method %

方法	BLEU_1	BLEU_4	METEOR	ROUGE-L	CIDEr	SPICE
拼接方法	76.1	36.0	27.8	56.5	113.2	20.6
门控机制	76.6	36.9	27.8	57.6	116.9	20.7

此外,本文将FMDVSI模型与其他主流模型在MSCOCO数据集的实验结果上进行对比,结果如表4所示。其中,SCA-CNN[12]为融合空间和通道注意力的方法。Adaptive-Attention[13]为自适应注意力模型。本文的FMDVSI(XE)模型相比于Adaptive-Attention模型在BLEU_1、BLEU_4、METEOR和CIDEr评价指标上的得分分别提升了2.4%、3.7%、1.2%和8.4%。Att2in[16]为使用基于强化学习的SCST方式进行训练的方法。Up-Down[14]为本文的基线模型。本文复现的结果与文献[14]基线模型略有不同,但是均在合理的波动范围内。ELMo-MCT[32]为基于Transformer的方法。可以看出,本文模型对比于当前流行的基于Transformer的方法仍具有一定的竞争力。A_R_L[33]为基于视觉关系和上下文感知注意力的方法。ASIA[34]为基于自适应空间特征注意力的方法。同时,本文与文献[35]进行对比,文献[35]侧重于对图像的视觉信息进行处理,而对于语义信息的处理有所不足。因此,本文模型具有一定的优势。

总体来看,本文模型拥有比较有竞争力的性能,进一步验证了融合多时间维度的视觉和语义信息的有效性。

表4 MSCOCO数据集上不同模型的性能对比

Table 4 Performance comparison of different models on the MSCOCO dataset %

模型	BLEU_1	BLEU_4	METEOR	ROUGE_L	CIDEr	SPICE
SCA-CNN	71.9	31.1	25.0	53.1	95.2	—
Adaptive-Attention	74.2	33.2	26.6	—	108.5	—
Att2in(XE)	—	31.3	26.0	54.3	101.3	—
Att2in(RL)	—	33.3	26.3	55.3	111.4	—
Up-Down(XE)	77.2	36.2	27.0	56.4	113.5	20.3
Up-Down(RL)	79.8	36.3	27.7	56.9	120.1	21.4
ELMo-MCT	76.2	34.2	—	56.0	111.5	—
A_R_L	75.9	35.8	27.8	56.4	113.7	—
ASIA	—	36.8	27.7	—	116.7	—
文献[35](XE)	—	36.6	28.1	57.1	115.9	—
文献[35](RL)	—	36.8	28.0	57.7	124.8	—
FMDVSI(XE)	76.6	36.9	27.8	57.6	116.9	20.7
FMDVSI(RL)	80.2	37.8	28.6	58.4	125.2	21.9

2.4 可视化分析

为了能够更加直观地证明本文模型能够生成质量更好的描述语句,表5展示了FMDVSI模型与Up-Down模型的可视化结果对比,其中Ground-truth表示图像标注的参考语句。相比于基线模型,本文模型生成的描述语句更加准确,能够提取到更加准确的目标物体,例如表5第2行图描述语句中的“ocean”,表5第3行图描述语句中的“cow”和“building”。同时生成的描述语句在语义上更加通顺、连贯,例如表5第1行图描述语句中的“A group of people”比基线模型生成的“A man”更加准确。

同时引入过去时刻的视觉和语义信息也可能会带来一些干扰信息,导致信息上的冗余,造成生成失败的案例,如表6所示。例如表6第1行图描述语句中的“A bunch of oranges and oranges”,表6第2行图描述语句中的“with a small house with a house”。

表5 可视化结果对比

Table 5 Comparison of visualization results

图像	描述语句
	**Ground-truth**: A group of people are riding bikes down the street in a bike lane. **Up-Down**: A man riding a bike down a street. **Ours**: A group of people riding bikes down a city street.
	**Ground-truth**: Young boy riding large breaking wave in open ocean. **Up-Down**: A man riding a wave on top of a surfboard. **Ours**: A man riding a wave on a surfboard in the ocean.
	**Ground-truth**: A cow standing in a grassy open field. **Up-Down**: A couple of animals that are in the grass. **Ours**: A cow standing in the grass in front of a building.

表6 失败案例

Table 6 Failure cases

图像	描述语句
	**Ground-truth**: A plate is piled high of orange slices while a bunch of bananas sits next to it. **Ours**: A bunch of oranges and oranges on a plate.
	**Ground-truth**: A large long train on a steel track near a barn. **Ours**: A model house with a small house with a house.

3 结束语

本文提出了一种融合多时间维度视觉与语义信息的图像描述方法,有效地融合了过去时刻的视觉信息和语义信息,并设计了一种门控机制对视觉信息和语义信息进行选择性利用。该方法使模型可以输出更加丰富的信息,确保了生成描述语句的准确性。另外,使用基于强化学习的SCST方法对模型的性能进行优化,使得模型的性能得到了极大的提升。最后在MSCOCO数据集上进行实验验证,实验结果表明本文方法的性能相比于其他主流方法在各种评价指标上都有明显的提升,生成的描述更加准确。

参考文献:

[1] YAN C, HAO Y, LI L, et al. Task-adaptive attention for image captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 43-51.

[2] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media,

2022, 8(3): 331-368.

[3] SUN T X, LIU X Y, QIU X P, et al. Paradigm shift in natural language processing[J]. Machine Intelligence Research, 2022, 19(3): 169-183.

[4] LI Y, MA J, ZHANG Y. Image retrieval from remote sensing big data: A survey[J]. Information Fusion, 2021, 67: 94-115.

[5] PUSTEJOVSKY J, KRISHNASWAMY N. Embodied human computer interaction[J]. KI-Künstliche Intelligenz, 2021, 35(3/4): 307-327.

[6] GUO W, ZHANG Y, YANG J, et al. Re-attention for visual question answering[J]. IEEE Transactions on Image Processing, 2021, 30: 6730-6743.

[7] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every picture tells a story: Generating sentences from images[C]//Proceedings of European Conference on Computer Vision. Berlin, German: Springer, 2010: 15-29.

[8] KUZNETSOVA P, ORDONEZ V, BERG A, et al. Collective generation of natural image descriptions[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Korea: ACL, 2012: 359-368.

[9] MAO J, XU W, YANG Y, et al. Explain images with multimodal recurrent neural networks[J]. Computer Science, 2014, 31(5): 182-190.

[10] 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, USA: IEEE, 2015: 3156-3164.

[11] 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. Lille, France: ICML, 2015: 2048-2057.

[12] 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. Boston, USA: IEEE, 2017: 5659-5667.

[13] 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. Boston, USA: IEEE, 2017: 375-383.

[14] 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. Boston, USA: IEEE, 2018: 6077-6086.

[15] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. USA: MIT Press, 2018.

[16] 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. Boston, USA: IEEE, 2017: 7008-7024.

[17] VEDANTAM R, LAWRENCE Z C, PARIKH D. Cider: Consensus-based image description evaluation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 4566-4575.

[18] ZHANG X, SUN X, LUO Y, et al. Rstnet: Captioning with adaptive attention on visual and non-visual words[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2021: 15465-15474.

[19] WANG Y, XU J, SUN Y. End-to-end transformer based model for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI, 2022: 2585-2594.

[20] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123: 32-73.

[21] WANG J, JIANG W, MA L, et al. Bidirectional attentive fusion with context gating for dense video captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2018: 7190-7198.

[22] YANG B, DENG X, SHI H, et al. Continual object detection via prototypical task correlation guided gating mechanism[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2022: 9255-9264.

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

[24] 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. Boston, USA: IEEE, 2015: 3128-3137.

[25] 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, Pennsylvania: ACL, 2002: 311-318.

[26] 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. Ann Arbor, USA: ACL, 2005: 65-72.

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

[28] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE: Semantic propositional image caption evaluation[C]// Proceedings of European Conference on Computer Vision. Berlin, German: Springer, 2016: 382-398.

[29] GARFINKEL S L, PAPADOURAKIS G. Adam: A method for stochastic optimization[C]// Proceedings of the 3rd International Conference for Learning Representations. San Diego, USA: ICLR, 2015: 1-15.

[30] WANG P, NG H T. A beam-search decoder for normalization of social media text with application to machine translation[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, USA: NAACL, 2013: 471-481.

[31] HE L, YUAN H. Improved cross-entropy research based on JS divergence: Iris flower data as an example[C]//Proceedings of 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications. Dalian, China: IEEE, 2022: 1455-1459.

[32] 杨文瑞, 沈韬, 朱艳, 等. 融合ELMo词嵌入的多模态Transformer的图像描述算法[J]. 计算机工程与应用, 2022, 58(21): 223-231.
YANG Wenrui, SHEN Tao, ZHU Yan, et al. Image caption with ELMo embedding and multimodal transformer[J]. Computer Engineering and Applications, 2022, 58(21): 223-231.

[33] WANG J, WANG W, WANG L, et al. Learning visual relationship and context-aware attention for image captioning[J]. Pattern Recognition, 2020, 98: 107075.

[34] ZHONG X, NIE G, HUANG W, et al. Attention-guided image captioning with adaptive global and local feature fusion[J]. Journal of Visual Communication and Image Representation, 2021, 78: 103138.

[35] 盛豪, 易尧华, 汤梓伟. 融合图像场景与目标显著性特征的图像描述生成方法[J]. 计算机应用研究, 2021, 38(12): 3776-3780.
SHENG Hao, YI Yaohua, TANG Ziwei. Image caption based on fusion of image scene and target saliency feature[J]. Application Research of Computers, 2021, 38(12): 3776-3780.

作者简介:

陈善学(1966-),男,博士,教授,研究方向:高光谱图像处理、图像描述,E-mail:chensx@cqupt.edu.cn。

王程(1999-),**通信作者**,男,硕士研究生,研究方向:图像描述,E-mail:717560396@qq.com。

(编辑:张黄群,王婕)