Journal of Computer Applications
计算机应用,2021,41(6):1640 - 1646
ISSN 1001-9081
CODEN JYIIDU
2021-06-10
http://www.joca.cn

文章编号:1001-9081(2021)06-1640-07
DOI:10.11772/j.issn.1001-9081.2020091439

基于多特征提取的图像语义描述算法

赵小虎[1,2],李　晓[1,2*]
(1.矿山互联网应用技术国家地方联合工程实验室(中国矿业大学),江苏 徐州 221008;
2.中国矿业大学 信息与控制工程学院,江苏 徐州 221008)
(*通信作者电子邮箱ts18060032a31@cumt.edu.cn)

摘　要:针对图像语义描述方法中存在的图像特征信息提取不完全以及循环神经网络(RNN)产生的梯度消失问题,提出了一种基于多特征提取的图像语义描述算法。所构建模型由三个部分组成:卷积神经网络(CNN)用于图像特征提取,属性提取模型(ATT)用于图像属性提取,而双向长短时记忆(Bi-LSTM)网络用于单词预测。该模型通过提取图像属性信息来增强图像表示,从而精确描述图中事物,并且使用Bi-LSTM捕捉双向语义依赖,从而进行长期的视觉语言交互学习。首先,使用CNN和ATT分别提取图像全局特征与图像属性特征;其次,将两种特征信息输入到Bi-LSTM中生成能够反映图像内容的句子;最后,在Microsoft COCO Caption、Flickr8k和Flickr30k数据集上验证了所提出算法的有效性。实验结果表明,与m-RNN方法相比,所提出的算法在描述性能方面提高了6.8~11.6个百分点。所提算法能够有效地提高模型对图像的语义描述性能。

关键词:图像语义描述;图像属性;双向长短时记忆网络;卷积神经网络;循环神经网络
中图分类号:TP391.41　　**文献标志码**:A

Image captioning algorithm based on multi-feature extraction

ZHAO Xiaohu[1,2], LI Xiao[1,2*]
(1. *National and Local Joint Engineering Laboratory of Internet Applied Technology on Mines (China University of Mining and Technology), Xuzhou Jiangsu* 221008, *China*;
2. *School of Information and Control Engineering, China University of Mining and Technology, Xuzhou Jiangsu* 221008, *China*)

Abstract: In image caption methods, image feature information is not completely extracted and the vanishing gradient is generated by the Recurrent Neural Network (RNN). In order to solve the problems, a new image captioning algorithm based on multi-feature extraction was proposed. The constructed model was consisted of three parts: Convolutional Neural Network (CNN) was used for image feature extraction, ATTribute extraction model (ATT) was used for image attribute extraction, and Bidirectional Long Short-Term Memory (Bi-LSTM) network was used for word prediction. In the constructed model, image representation was enhanced by extracting image attribute information, so as to accurately describe the things in the image, and Bi-LSTM was used to capture bidirectional semantic dependency, so that the long-term visual language interaction learning was carried out. Firstly, CNN and ATT were used to extract the global image features and image attribute features respectively. Then, the two kinds of feature information were input into Bi-LSTM to generate sentences that were able to reflect the image content. Finally, the effectiveness of the proposed method was validated on Microsoft COCO Caption, Flickr8k, and Flickr30k datasets. Experimental results show that, compared with the multimodal Recurrent Neural Network (m-RNN) method, the proposed algorithm has improved the description performance by 6.8-11.6 percentage points. The proposed algorithm can effectively improve the semantic description performance of the constructed model for images.

Key words: image captioning; image attribute; Bidirectional Long Short-Term Memory (Bi-LSTM) network; Convolutional Neural Network (CNN); Recurrent Neural Network (RNN)

0　引言

图像语义描述一直是人工智能领域中最重要的研究方向之一,是图像理解的高级任务。它首先需要识别图像中的对象和场景,描述目标类别、属性以及对象和其在场景中位置之间的关系,然后将描述信息转化为一个有一定的语法结构和语义的句子,这样人们可以在没有看到图像的情况下很快地理解图像内容。因此图像的语义描述设计了多种模型,根据句子生成方法的不同,可分为基于模板的方法[1-4]、基于检索的方法[5-8]和基于神经网络的方法。目前,基于深度神经网络

收稿日期:2020-09-16;**修回日期**:2020-12-05;**录用日期**:2020-12-08。
基金项目:国家重点研发计划项目(2017YFC0804400);徐州市重点研发科技项目(KC19112)。
作者简介:赵小虎(1976—),男,江苏徐州人,教授,博士,主要研究方向:矿山物联网、矿山通信、监视和控制、计算机网络、智能计算;李晓(1994—),女,安徽宿州人,硕士研究生,主要研究方向:计算机视觉处理、图像语义描述。

的图像语义描述方法在这一领域取得了重大突破,尤其是卷积神经网络(Convolutional Neural Network, CNN)与递归神经网络相结合的语义描述生成模型。此模型生成的句子与人工标注的句子非常接近,在多个数据集上都取得了良好的效果。

Mao 等[9]创造性地将卷积神经网络和递归神经网络相结合,解决了图像描述和句子检索等问题。自此之后,基于深度神经网络的图像语义描述方法得到了广泛的发展。Kiros 等[10]率先将编码-解码框架引入图像语义描述研究,利用深度卷积神经网络对视觉信息进行编码,同时利用长短时记忆网络对文本数据进行编码。Szegedy 等[11]提出了一种基于 GoogLeNet 和长短时记忆(Long Short-Term Memory, LSTM)网络的图像标语义描述模型,该模型只需要将整个图像特征放入 LSTM 的初始时间步长,以此降低模型的复杂度。为了生成与图像内容密切相关的图像标题,Jia 等[12]提出了一种扩展的 LSTM 模型 g-LSTM(guiding-LSTM),提取图像的语义信息来表示图像与描述之间的关系。在图像语义描述生成阶段,该模型利用语义信息指导 LSTM 的每一步生成。实验结果表明,语义信息可以显著提高描述性能。Xu 等[13]首先提出了一种基于注意力机制的图像标题方法,该方法将图像平均分成 14 × 14 图像块,利用"soft"和"hard"注意力机制对图像的突出区域进行自动搜索,生成图像标题。Li 等[14]提出了一种基于全局-局部注意机制(Global-Local Attention, GLA)的图像语义描述方法,该模型将注意机制分解为目标级局部表示和图像级全局表示,从而在保持图像全局上下文信息的同时,更准确地预测突出目标。Luo 等[15]利用从图像中检测出的语义概念实现图像语义描述。He 等[16]使用词性标注引导长短时记忆网络生成单词。Yang 等[17]提出了情感概念,用以增强文本描述的情感表达能力,通过将适当的情感概念组合到句子中来实现。

综上所述,现有方法主要是依靠卷积神经网络(CNN)获取固定的全局图像特征向量,并通过循环神经网络(Recurrent Neural Network, RNN)对其进行解码。然而全局图像特征不足以表示完整的图像信息,需要进一步获取图像中的场景类型、位置信息等以增强图像的语义表达。其次,在 RNN 中存在梯度消失问题,利用 LSTM 可消除梯度消失现象,但是 LSTM 只能够捕捉单向时序信息,未实现真正意义上的全局上下文。为了解决以上问题,本文提出了一种基于多特征提取的图像语义描述算法,在图像视觉特征提取时加入了图像属性提取,利用图像的全局特征和附加的高级语义信息来建立图像和句子之间的关系。此外,在解码阶段使用了双向长短时记忆(Bidirectional LSTM, Bi-LSTM)网络,使得模型能够捕捉双向语义依赖,有效地提高了模型对图像的语义描述性能。

1　本文模型

本文的模型是基于深度神经网络框架下的图像语义描述模型。如图 1 所示,该模型包括三个部分:图像特征提取[18]、图像属性提取以及用于单词生成的双向长短时记忆网络。其中,采用 Resnet-50 残差网络架构的卷积层与平均池化层提取图像的全局特征 $\boldsymbol{V}_{\text{img}}$。Resnet-50 残差网络在 ImageNet 分类数据集上预先训练。

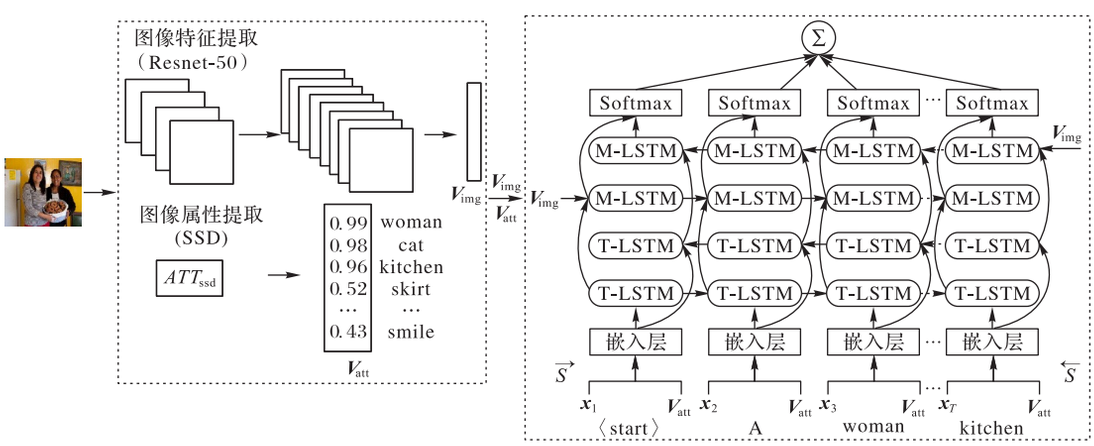

图 1　本文模型整体结构

Fig. 1　Overall structure of proposed model

1.1　图像属性提取

图像语义描述成功的关键在于良好的图像特征表示。根据人类对图像的直观描述可知,良好的图像特征能够准确地反映图像的内容和属性信息,如场景类型、位置信息等。但是只提取图像全局特征过于粗糙,会丢失很多重要信息,难以满足上述要求,因此在句子生成过程中会产生对象误差预测的问题。针对以上问题,本文采用 SSD(Single Shot multibox Detector)模型对图像的属性信息进行提取以增强图像特征表示,从而提高图像语义表达性能。和一般方法不同的是本文对 SSD 进行进一步的改进,使其不仅可以准确地描述目标细节,还可以更加准确地描述目标的行为及其所在场景,极大提高了其在真实场景中的应用。

SSD 网络[19]用于检测图像属性特征。如图 2 所示,选用 Resnet-50[20]残差结构为其前置网络,代替了原来的 VGG16 网络,解决目标尺度小、分辨率低等问题,并且相较于原来的网络增加了一层特征提取层,提高了网络的特征提取能力。选取的特征提取层为 Conv2_x、Conv3_x、Conv4_x、Conv5_x 以及 Conv7_x、Conv8_x、Conv9_x,共提取 7 个特征图。输入图像大小为 224 × 224。

为了能够检测到图像中不同尺寸的物体,此网络使用若干不同输出尺寸的特征图进行检测。位于不同层的特征图设置的先验框数目不同,其参数包括尺度和长宽比两个方面。

先验框的设置遵守线性递增规则：

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m - 1}(k - 1);\ k \in [1, n] \tag{1}$$

其中：n 为特征图个数；s_k 表示先验框相对于图像所占的比例，$s_{\min}$、$s_{\max}$ 分别为 0.2、0.9。对于先验框长宽比，一般选取 $a_r = \left\{\frac{1}{3}, \frac{1}{2}, 1, 2, 3\right\}$，则每个先验框的宽、高分别为：$W_k^a = S_k\sqrt{a_r}$，$h_k^a = \frac{S_k}{\sqrt{a_r}}$。先验框的中心点为 $\left(\frac{i + 0.5}{|f_k|}, \frac{j + 0.5}{|f_k|}\right)$，$i, j \in \left[0, |f_k|\right]$，$|f_k|$ 为第 k 个特征图的大小。

先验框的参数设置直接影响着模型检测不同尺度物体的性能，先验框长宽比的分布也同样影响模型检测目标的准确率。为了达到最高准确率，并且减少不必要的计算量，对先验框长宽比的选择进行了实验，结果如图 3 所示。

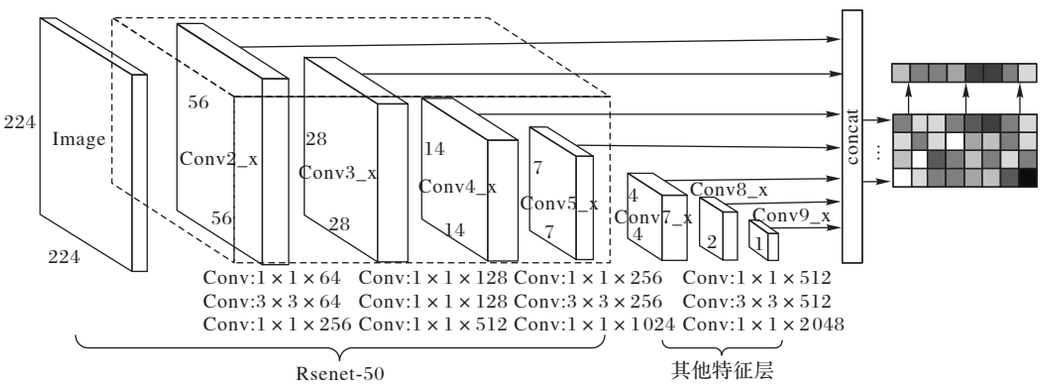

图 2 图像属性提取结构

Fig. 2 Structure of image attribute extraction

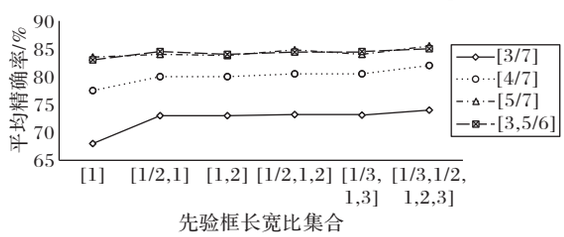

图 3 不同先验框长宽比的平均精确率对比

Fig. 3 Comparison of mean average precision of different anchor box length-width ratios

图 3 中，横坐标为使用的先验框长宽比的集合，例如，[1/2,1]表示针对当前图像采用的先验框的长宽比分别为 1/2 和 1。折线[3/7]表示 7 个卷积层的先验框都是 3 个，即先验框长宽比为[1/2,1,2]。从图 3 可知，折线[3,5/6]和[5/7]在此实验中的平均精确率(mean Average Precision，mAP)较高，分别为 85.2% 和 85.4%。相较[3,5/6]卷积层长宽比的分布，[5/7]的分布需要多计算 6 272 个先验框，增加了计算复杂度，但 mAP 值却只增加了 0.2 个百分点。因此本文选择的先验框参数如表 1 所示。

表 1 先验框参数

Tab. 1 Anchor box parameters

卷积层	先验框个数	先验框长宽比	先验框尺寸
Conv2_x	3	[1/2,1,2]	56 × 56
Conv3_x	5	[1/3,1/2,1,2,3]	28 × 28
Conv4_x	5	[1/3,1/2,1,2,3]	14 × 14
Conv5_x	5	[1/3,1/2,1,2,3]	7 × 7
Conv7_x	5	[1/3,1/2,1,2,3]	4 × 4
Conv8_x	5	[1/3,1/2,1,2,3]	2 × 2
Conv9_x	5	[1/3,1/2,1,2,3]	1 × 1

属性提取得到矩阵 $\boldsymbol{M}_{\text{att}}$，将输出矩阵通过式(2)计算得到图像属性 $\boldsymbol{V}_{\text{att}}$：

$$\boldsymbol{V}_{\text{att}} = \max_{i \in m} \boldsymbol{M}_{\text{att}}^{i,j};\ j \in 0, 1, \cdots, c \tag{2}$$

其中：m 为输入图片上的边界框个数，m=14 658；c 为检测类别数，在 Visual Genome 数据集上训练属性提取模型(ATTribute extraction model，ATT)，令 c=300。

1.2 双向长短时记忆网络

在传统的图像语义描述中，循环神经网络随着时间步长的增加，存在梯度消失的问题，缺乏前一时刻的指向性信息；并且该网络只能利用单向时序信息，未实现真正意义上的全局上下文。为了解决以上问题，本模型使用了双向长短时记忆网络，能够充分利用句子过去和将来的上下文信息预测语义，生成涵盖丰富的语义信息的语句，并且更加符合人类表达习惯。

Bi-LSTM[21]模型建立在 LSTM 单元上，LSTM 单元是传统递归神经网络的一种特殊形式。图 4 为长短时记忆网络单元示意图。读写存储单元 $\boldsymbol{c}$ 由一组 sigmoid 门控制，当时间步长为 t 时，LSTM 的输入来源有：当前输入 $\boldsymbol{x}_t$、所有 LSTM 单元之前的隐藏状态 $\boldsymbol{h}_{t-1}$ 以及记忆单元 $\boldsymbol{c}_{t-1}$。对于给定的输入向量 $\boldsymbol{x}_t$、$\boldsymbol{h}_{t-1}$ 和 $\boldsymbol{c}_{t-1}$，t 时刻门的更新如下：

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right) \tag{3}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right) \tag{4}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right) \tag{5}$$

$$\boldsymbol{g}_t = \sigma\left(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c\right) \tag{6}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \boldsymbol{g}_t \tag{7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \phi(\boldsymbol{c}_t) \tag{8}$$

其中：$\boldsymbol{W}$ 为网络的权重矩阵；$\boldsymbol{b}$ 为偏置向量；σ 是 sigmoid 激活函数，即 $\sigma(x) = \frac{1}{1 + \exp(-x)}$；$\phi$ 是双曲正切函数，即 $\phi(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}$；“*”表示门值计算。LSTM 隐藏层输出 $\boldsymbol{h}_t =$

$\left\{\boldsymbol{h}_{tk}\right\}_{k=0}^{K}$,$\boldsymbol{h}_t$和权重矩阵$\boldsymbol{W}_s$、偏置向量$\boldsymbol{b}_s$通过Softmax函数被用于预测下一单词的概率:

$$F\left(\boldsymbol{p}_{ti};\ \boldsymbol{W}_s,\boldsymbol{b}_s\right)=\frac{\exp\left(\boldsymbol{W}_s\boldsymbol{h}_{ti}+\boldsymbol{b}_s\right)}{\sum_{j=1}^{K}\exp\left(\boldsymbol{W}_s\boldsymbol{h}_{tj}+\boldsymbol{b}_s\right)} \tag{9}$$

其中$\boldsymbol{p}_{ti}$为预测词的概率分布。

如图5所示为Bi-LSTM模型,该模型由三部分组成:图像全局特征$\boldsymbol{V}_{\text{img}}$、用于编码句子输入的T-LSTM(Test LSTM)、用于将视觉和文本向量嵌入到公共语言空间的M-LSTM(Multimodal LSTM)。输入向量$\boldsymbol{x}_t$与图像属性$\boldsymbol{V}_{\text{att}}$共同作为Bi-LSTM的输入,因此T-LSTM层的输入$\boldsymbol{V}_t$可表示为:

$$\boldsymbol{V}_a=f_v\left(\boldsymbol{V}_{\text{att}};\ \boldsymbol{W}_v;\ \boldsymbol{b}_v\right) \tag{10}$$

$$\boldsymbol{V}_t=\tanh\left(\boldsymbol{x}_t\oplus\boldsymbol{V}_a\right);\ t\in\{1,2,\cdots,N-1\} \tag{11}$$

其中:$f_v(\cdot)$为全连接层;"⊕"表示级联运算。

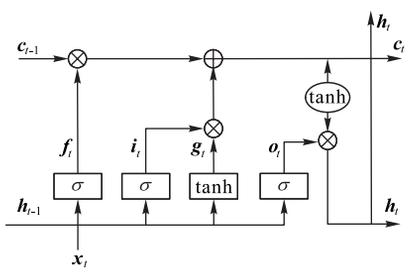

图4　长短时记忆网络单元

Fig. 4　LSTM network unit

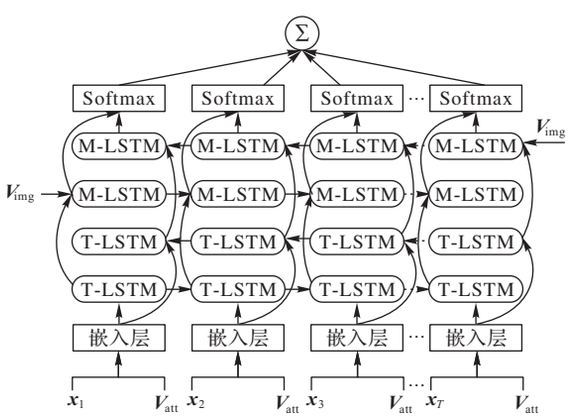

图5　Bi-LSTM结构

Fig. 5　Structure of Bi-LSTM

Bi-LSTM是由两层分开的LSTM组成,用于计算前向隐藏层序列$\vec{\boldsymbol{h}}$和后向隐藏层序列$\overleftarrow{\boldsymbol{h}}$。前向LSTM开始于$t$=1,后向LSTM开始于$t=T$。模型工作中,对前向句子$\vec{S}$和后向句子$\overleftarrow{S}$,编码表现为:

$$\vec{\boldsymbol{h}}_t^1=T\left(\vec{\boldsymbol{E}}\vec{S};\ \vec{\boldsymbol{\Theta}}_l\right) \tag{12}$$

$$\overleftarrow{\boldsymbol{h}}_t^1=T\left(\overleftarrow{\boldsymbol{E}}\overleftarrow{S};\ \overleftarrow{\boldsymbol{\Theta}}_l\right) \tag{13}$$

其中:T代表T-LSTM,$\boldsymbol{\Theta}_l$是它们相应的权重;$\vec{S}=\left\{\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_T\right\}$,$\overleftarrow{S}=\left\{\boldsymbol{x}_T,\boldsymbol{x}_{T-1},\cdots,\boldsymbol{x}_0\right\}$。$\vec{\boldsymbol{E}}$、$\overleftarrow{\boldsymbol{E}}$为从网络中学习得到的双向嵌入矩阵。之后将编码的视觉和文本表示通过以下计算式嵌入到M-LSTM:

$$\vec{\boldsymbol{h}}_t^2=M\left(\vec{\boldsymbol{h}}_t^1,\boldsymbol{V}_{\text{img}};\ \vec{\boldsymbol{\Theta}}_m\right) \tag{14}$$

$$\overleftarrow{\boldsymbol{h}}_t^2=M\left(\overleftarrow{\boldsymbol{h}}_t^1,\boldsymbol{V}_{\text{img}};\ \overleftarrow{\boldsymbol{\Theta}}_m\right) \tag{15}$$

其中:M为M-LSTM,$\boldsymbol{\Theta}_m$为其权重。目的是在不同的时间步长的情况下,捕捉视觉语境与词汇之间的相关性。在每个时间步长向模型中输入可视化向量$\boldsymbol{V}_{\text{img}}$,以捕获强大的可视化单词相关性。在M-LSTM的最上层是Softmax层,通过该层计算下一个预测词的概率分布:

$$\vec{\boldsymbol{p}}_{t+1}=F\left(\vec{\boldsymbol{h}}_t^2;\ \boldsymbol{W}_s,\boldsymbol{b}_s\right) \tag{16}$$

$$\overleftarrow{\boldsymbol{p}}_{t+1}=F\left(\overleftarrow{\boldsymbol{h}}_t^2;\ \boldsymbol{W}_s,\boldsymbol{b}_s\right) \tag{17}$$

其中,$\boldsymbol{p}\in\boldsymbol{R}^K$,$K$为字典大小。

1.3　损失函数

本文模型通过采用随机梯度下降(Stochastic Gradient Descent, SGD)的方法实现端到端的训练。在训练中,通过给定的图像上下文I和前向顺序$P(w_t|w_{1:t-1},I)$或者后向顺序$P(w_t|w_{t+1:T},I)$中的先前的单词上下文$w_{1:t-1}$预测单词w_t,分别设置$w_1=w_T=0$时为前向和后向的起点。联合损失函数$L=\vec{L}+\overleftarrow{L}$是累加前后向的Softmax损失得到。

$$\vec{L}(w,I)=-\sum_{1}^{T}\lg p(w_t|w_1,w_2,\cdots,w_{t-1},I) \tag{18}$$

$$\overleftarrow{L}(w,I)=-\sum_{1}^{T}\lg p(w_t|w_{t+1},w_{t+2},\cdots,w_T,I) \tag{19}$$

其中,T为生成序列的长度。本文的目标是最小化L,这就相当于最大化生成正确句子的概率。最后可从两个方向生成句子,本文根据句子的单词生成概率确定给定图像$p(w_{1:T}|I)$最终的句子。

$$p(w_{1:T}|I)=\max\left(\sum_{t=1}^{T}(\vec{p}(w_t|I))\sum_{t=1}^{T}(\overleftarrow{p}(w_t|I))\right) \tag{20}$$

$$\vec{p}(w_t|I)=\prod_{t=1}^{T}p(w_t|w_1,w_2,\cdots,w_{t-1},I) \tag{21}$$

$$\overleftarrow{p}\left(w_t|I\right)=\prod_{t=1}^{T}p(w_t|w_{t+1},w_{t+2},\cdots,w_T,I) \tag{22}$$

2　实验与结果分析

2.1　训练细节

在训练过程中,每张图像都有3个相关的注释。首先通过提取图像属性得到$\boldsymbol{V}_{\text{att}}$,$\boldsymbol{V}_{\text{att}}$被用于计算Bi-LSTM网络的每个时间步长。符号〈start〉是一个句子的开头,〈end〉是句子的结尾。本文使用了双层循环神经网络Bi-LSTM,隐藏单元数为512,权重衰减率为0.000 5。优化器学习速率为0.001,设置batch size为64、动量为0.9来训练本文的模型。

2.2　数据集

本文使用的数据集为Flickr8k、Flickr30k和MSCOCO数据集。Flickr8k数据集中有6 000张训练图像、1 000张测试图像和1 000张验证图像。Flickr30k数据集包含31 000张图像,随机将其中29 000张图像用于训练,1 000张图像用于测试,1 000张图像用于验证。两个数据集中的每个图像对应五个人工生成的描述。MSCOCO数据集包含82 738张用于训练的图像,40 504张用于验证的图像,每个图像有5个由AMT(Amazon Mechanical Turk)得到的句子。

2.3 结果分析

为了评估属性提取模型(ATT)和Bi-LSTM的有效性,分别在Microsoft COCO Caption数据集和Flickr8k、Flickr30k数据集进行实验,将本文的模型与当前流行的图像语义描述方法进行比较,结果如表2~3所示。所用的评估指标为BLEU(Bilingual Evaluation Understudy)、METEOR、ROUGE-L和CIDEr(Consensus-based Image Description Evaluation),具体内容如下:

1)BLEU是一种基于精确度的相似性度量机器翻译评价指标,用于分析候选译文和参考译文中n元组共同出现的程度。

2)METEOR是基于单精度的加权调和平均数和单字召回率,其目的是解决一些BLUE标准中的缺陷,与BLUE相比,其结果和人工判断的结果有较高的相关性。

3)ROUGE-L是基于最长公共子句的度量方法。参考译文与待评测译文的共现性精确度越高,则生成的句子质量越高。

4)CIDEr是专用于图像语义描述的度量标准,通过TF-IDF(Term Frequency-Inverse Document Frequency)计算每个n元组的权重来衡量图像语义描述的一致性。

表2为不同模型在MSCOCO(Microsoft COCO)数据集上的结果。评价指标有BLEU、METEOR、ROUGE-L和CIDEr。其中,模型Bi-LSTM+ATT+CNN_R以及Bi-LSTM+CNN_R在BLEU-1、BLEU-2、BLEU-3、BLEU-4指标上分别达到了74.5%、59.3%、45.5%、36.3%和74.0%、57.2%、44.5%、33.6%。由此可知,两种类型的图像语义信息能更准确描述图像信息。由表2还可以看出,本文模型的性能显著优于其他方法,表中用于对比的方法为最新的方法,分别为:NIC(Neural Image Caption)[22]、LRCN(Long-term Recurrent Convolutional Network)[23]、Deep-Vis(Deep Visual-semantic)[24]、m-RNN(multimodal Recurrent Neural Network)[9]、g-LSTM[12]、Hard-Attention[13]、Soft-Attention[13]、HMA(Hierarchical Multimodal Attention-based)[25]、VLM(Visual attention based on Long-short term Memory)[26]、GLA[14]和PSG(Part of Speech Guidance)[16]。这些方法使用了不同的特征提取方式[27],NIC和g-LSTM使用GoogLeNet获得图像特征;LRCN利用AlexNet来提取图像特征;Deep-Vis、m-RNN、Hard-Attention、Soft-Attention利用VGG16获得图像特征。不同的卷积神经网络提取特征的能力是不一样的,以上所有模型均未使用高级语义信息。实验结果表明,与仅用单一的语义信息特征的模型相比,两种类型的图像信息能够有更好的描述效果。递归神经网络的使用也是不同的:NIC、g-LSTM、Hard-Attention和Soft-Attention使用LSTM网络作为语言模型生成图像句子描述;m-RNN利用基本的RNN作为解码器;LRCN使用一个堆叠的两层LSTM将图像转换成句子描述;在本文的模型中,使用了Bi-LSTM得到图像语义描述。

不同模型在Flickr8k和Flickr30k数据集上的结果如表3所示。同样地,在Flickr8k和Flickr30k数据集上,本文的模型也获得了BLEU和METEOR指标的最佳性能。

表2 Microsoft COCO数据集上不同模型实验结果对比 单位:%

Tab. 2 Experimental result comparison of different models on Microsoft COCO dataset unit:%

模型	MSCOCO						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
NIC	—	—	—	27.70	23.70	—	85.50
LRCN	62.79	44.19	30.41	21.00	—	—	—
Deep-Vis	62.50	45.00	32.10	23.00	19.50	—	66.00
m-RNN	67.00	49.00	35.00	25.00	—	—	—
g-LSTM	67.00	49.10	35.80	26.40	22.70	—	81.30
Hard-Attention	70.70	49.20	34.40	24.30	23.90	—	—
Soft-Attention	71.80	50.40	35.70	25.00	23.00	—	—
HMA	71.00	51.30	37.20	27.10	23.30	—	—
VLM	72.30	52.20	37.10	25.20	—	—	—
GloLocAttEmb	72.50	55.60	41.70	31.20	24.90	53.30	96.40
PSG	72.80	56.00	42.00	31.30	24.90	—	94.20
Bi-LSTM+ CNN_R	**74.00**	**57.20**	**44.50**	**33.60**	**26.30**	**53.90**	**96.20**
Bi-LSTM+ATT+CNN_R	**74.50**	**59.30**	**45.50**	**36.30**	**28.70**	**55.80**	**99.70**

表3 Flickr8k、Flickr30k数据集上不同模型实验结果对比 单位:%

Tab. 3 Experimental result comparison of different models on Flickr8k, Flickr30k datasets unit:%

模型	Flickr8k				Flickr30k			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
NIC	63.0	41.0	27.2	—	66.6	42.3	27.7	18.3
LRCN	—	—	—	—	58.8	39.1	25.1	16.5
m-RNN	56.5	38.6	25.6	17.0	60.0	41.0	28.0	19.0
Hard-Attention	67.0	45.7	31.4	21.3	66.9	43.9	29.6	19.9
Soft-Attention	67.0	44.8	29.9	19.5	66.7	43.4	28.8	19.1
Bi-LSTM+CNN_R	**66.9**	**46.6**	**33.7**	**27.3**	**68.1**	**45.2**	**34.9**	**25.3**
Bi-LSTM+ATT+CNN_R	**68.1**	**48.6**	**35.9**	**28.0**	**69.7**	**47.8**	**36.2**	**27.4**

由表2~表3的结果可知,本文的模型在Microsoft COCO数据集和Flickr8k、Flickr30k数据集上是有效的,表明将Bi-LSTM网络与ATT相结合可以提高图像语义描述性能,使用两种语义信息能够有效地提高模型的表达能力,拥有较强

的竞争力。

2.4　描述效果对比

图6为图像在前向LSTM和后向LSTM生成的语义描述。从图6中可以发现两者的一些区别:1)在棒球和橘子汁图像中,一个描述静态场景,另一个描述在下一个时间可能发生的潜在动作或运动。2)生成的语句与标注语句有很高的相似度,例如在火车图像中,前向描述与标注语句"A passenger train that is pulling into a station."相似,后向描述与标注语句"A train is in a tunnel by a station."相似。由此可以看出,本文的模型具有较强的视觉语言关联学习能力和生成新句子的能力。

图7为不同图像在模型中生成的句子。虽然这三幅图像不同,但都可以通过模型生成与图像密切相关并且语言流畅的句子。以厨房图像为例,人物可以很容易识别,但人物的运动状态却很难被识别。与m-RNN模型生成的句子相比,本文的模型生成的"Two cooks cooking with pans in a restaurant kitchen."可以准确地描述图片中人物的动作。此外,在食物和沙滩图像中,本文的模型能准确地识别出"doughnuts"和"seagull",而不仅仅只是描述成"baked goods"和"bird"。结果表明,本文的模型可以更准确地识别出图像中的物体,并且能够准确表述出图像中各个物体之间的关系,生成更相关、更连贯的自然语言句子来描述图像。

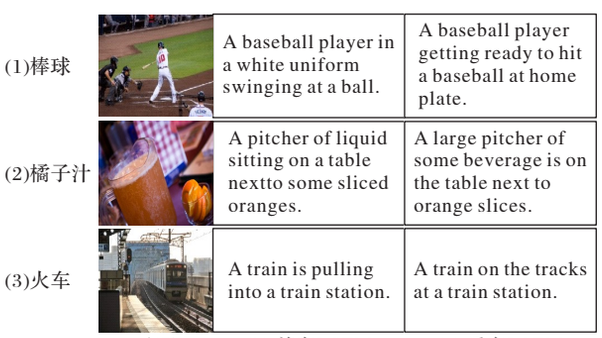

图6　不同LSTMs的图像语义描述效果对比

Fig. 6　Image captioning effect comparison of different LSTMs

图7　不同模型的图像语义描述效果对比

Fig. 7　Image captioning effect comparison of different models

根据以上实验分析可知,本文所构建的模型能够准确地识别图像中物体,并且能细致地描述所检测到的物体之间的关系;对于图像中较小的物体,本文的模型识别更精确,可以根据图像中人物和场景推测出图像中的人物动作;相较于其他图像语义描述方法,本文的模型在衡量指标上的描述效果更好。

3　结语

本文提出了一种多特征提取的图像语义描述算法,在Flickr8k、Flickr30k和MSCOCO数据集上的实验结果表明了本文算法在图像语义描述任务中的有效性。与现有的算法相比,本文算法通过提取图像高级语义信息和利用生成语句的双向信息,不仅能准确地描述目标细节,而且能更准确地表现行为和场景。简单的图像语义描述算法有很多局限性,更倾向于捕获全局特征信息,难以在现实场景中应用,未来考虑细粒度图像语义描述生成,如图像段落描述、图像和语言的双向检索等。

参考文献(References)

[1] KULKARNI G, PREMRAJ V, DHAR S, et al. BabyTalk: understanding and generating simple image descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[2] MITCHELL M, DODGE J, GOYAL A, et al. Midge: generating image descriptions from computer vision detections [C]// Proceedings of the 2012 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2012: 747-756.

[3] ELLIOTT D, KELLER F. Image description using visual dependency representations [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2013: 1292-1302.

[4] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every picture tells a story: generating sentences from images [C]// Proceedings of the 2010 European Conference on Computer Vision, LNCS 6314. Berlin: Springer, 2010: 15-29.

[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: 207-218.

[6] 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.

[7] KUZNETSOVA P, ORDONEZ V, BERG A, et al. Generalizing image captions for image-text parallel corpus [C]// Proceedings of the 2013 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2013:790-796.

[8] MASON R, CHARNIAK E. Nonparametric method for data-driven image captioning [C]// Proceedings of the 2014 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014: 592-598.

[9] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal Recurrent Neural Networks (m-RNN) [EB/OL]. [2020-11-17]. https://arxiv.org/pdf/1412.6632.pdf.

[10] KIROS R, SALAKHUTDINOV R, ZEMEL R S. Unifying visual-semantic embeddings with multimodal neural language models [EB/OL]. [2020-11-17]. https://arxiv.org/pdf/1411.2539.pdf.

[11] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with

convolutions [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1-9.

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

[13] XU K, BA JI, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention [C]// Proceedings of the 2015 32nd International Conference on Machine Learning. New York: JMLR. org , 2015: 2048-2057.

[14] LI L, TANG S, DENG L, et al. Image caption with global-local attention [C]// Proceedings of the 2017 31st AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4133-4139.

[15] LUO M, CHANG X, LI Z, et al. Simple to complex cross-modal learning to rank [J]. Computer Vision and Image Understanding, 2017, 163: 67-77.

[16] HE X, SHI B, BAI X, et al. Image caption generation with part of speech guidance [J]. Pattern Recognition Letters, 2019, 119: 229-237.

[17] YANG J, SUN Y, LIANG J, et al. Image captioning by incorporating affective concepts learned from both visual and textual components [J]. Neurocomputing, 2019, 328: 56-68.

[18] ZHAO D, CHANG Z, GUO S. A multimodal fusion approach for image captioning [J]. Neurocomputing, 2019, 329: 476-485.

[19] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector [C]// Proceedings of the 2016 European Conference on Computer Vision, LNCS 9905. Cham: Springer, 2016: 21-37.

[20] 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. Piscataway: IEEE, 2016: 770-778.

[21] WANG C, YANG H, MEINEL C. Image captioning with deep bidirectional LSTMs and multi-task learning [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(2s): Article No. 40.

[22] 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. Piscataway: IEEE, 2015: 3156-3164.

[23] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677-691.

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

[25] CHENG Y, HUANG F, ZHOU L, et al. A hierarchical multimodal attention-based neural network for image captioning [C]// Proceedings of the 2017 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017: 889-892.

[26] QU S, XI Y, DING S. Visual attention based on long-short term memory model for image caption generation [C]// Proceedings of the 2017 29th Chinese Control and Decision Conference. Piscataway: IEEE, 2017: 4789-4794.

[27] 王媛华. 基于多融合模型的图像语义描述研究[J]. 河南科技, 2019(14): 34-36. (WANG Y H. Image caption based on multi-fusion model [J]. Henan Science and Technology, 2019(14): 34-36.)

This work is partially supported by the National Key Research and Development Program of China (2017YFC0804400), the Key Research and Development Science and Technology Project of Xuzhou (KC19112).

ZHAO Xiaohu, born in 1976, Ph. D., professor. His research interests include mine internet of things, mine communication, monitoring and control, computer network, intelligent computing.

LI Xiao, born in 1994, M. S. candidate. Her research interests include computer vision processing, image caption.