ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 1,Jan. 2023,pp. 209－219
DOI:10. 16337/j. 1004-9037. 2023. 01. 018

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

基于图神经网络和引导向量的图像字幕生成模型

佟国香，李乐阳

（上海理工大学光电信息与计算机工程学院，上海 200093）

摘　要：近年来，深度学习已在图像字幕技术研究中展现其优势。在深度学习模型中，图像中对象之间的关系在图像表示中起着重要作用。为了更好地检测图像中的视觉关系，本文基于图神经网络和引导向量构建了图像字幕生成模型（YOLOv4-GCN-GRU，YGG）。该模型利用图像中被检测到的对象的空间和语义信息建立成图，利用图卷积神经网络（Graph convolutional network，GCN）作为编码器对图的每个区域进行表示。在字幕生成阶段，额外训练一个引导神经网络来产生引导向量，从而辅助生成模型自动生成语句。基于MSCOCO图像数据集的对比实验表明，YGG模型具有更好的性能，将CIDEr-D的性能从138.9％提高到了142.1％。

关键词：图像字幕；空间语义图；图卷积神经网络；引导向量；生成模型

中图分类号：TP3　　**文献标志码**：A

Image Caption Generation Model Based on Graph Neural Network and Guidance Vector

TONG Guoxiang，LI Yueyang

（College of Optical-Electrical and Computer Engineering，University of Shanghai for Science and Technology，Shanghai 200093，China）

Abstract：In recent years，deep learning has shown its advantages in the research of image caption technology. In deep learning model，the relationship between objects in image plays an important role in image representation. In order to better detect the visual relationship in the image，an image caption generation model（YOLOv4-GCN-GRU，YGG）is constructed based on graph neural network and guidance vector. The model uses the spatial and semantic information of the detected objects in the image to build a graph，and uses graph convolutional network（GCN）as an encoder to represent each region of the graph. In the process of decoding，an additional guidance neural network is trained to generate guidance vector，so as to assist the decoder to automatically generate sentences. Comparative experiments based on MSCOCO image dataset show that YGG model has better performance，and the performance of CIDEr-D is improved from 138.9％ to 142.1％.

Key words: image caption; spatial semantic map; graph convolution neural network; guidance vector; generation model

---

基金项目：国家重点研发计划项目（2018YFB1700902）。
收稿日期：2022-01-03；**修订日期**：2022-04-18

引　言

图像字幕生成[1]是用自然语言描述图像视觉内容的任务，该任务基本可以用一个视觉检测系统和一个能够生成有意义且语法正确的语句的语言模型来解决。目前图像字幕生成技术的研究方法可以分为3类：基于模板的方法、基于检索的方法以及基于神经网络的方法。

本文主要基于深度学习展开图像字幕生成研究。基于神经网络的方法是从机器翻译中获取灵感，通常使用两个循环神经网络（Recurrent neural network, RNN）[2]分别作为编码器和解码器[3]。早期，Vinyals等[4]采用RNN作为编码器，将长短期记忆网络（Long short-term memory, LSTM）或门循环单元（Gate recurrent unit, GRU）作为解码器将图像向量输出成语句，取得了较好的效果；Mao等[5]使用注意力机制解决图像字幕生成问题，解码器可以像人眼一样将注意力集中在图像的不同区域，从而提高视觉提取能力；为进一步改进注意力机制，Lu等[6]通过引入视觉哨兵，使模型可以自适应地关注图像中更重要的区域；You等[7]提出了一种基于语义的注意力模型，融合了全局图像特征和语义信息选择性地关注语义概念上的区域。由于在传统方法中，字幕的对数似然分数与人类对质量的评估不具有较好的相关性，Liu等[8]使用策略梯度方法来直接优化SPICE和CIDEr的线性组合（SPICE分数），从而使得生成字幕确在语法上更流畅，更符合人类习惯；Zhu等[9]于2021年采用一种网络架构搜索策略来更好地设计基于RNN的图像字幕自动解码器模块（AutoCaption），探索了基于共享参数的强化学习方法，以便为解码器生成更多的结构；Pan等[10]提出了X-线性注意网络（X-linear attention network, X-LAN），该网络将X-线性注意块集成到图像字幕模型的图像编码器和句子解码器中，以利用更高阶的模内和模间交互。此外，图像字幕生成可以视作文本生成图像任务的反向工作，与之相关的最新研究也具有一定的借鉴意义。如Xu等[11]提出了一种注意力生成对抗网络，由注意力驱动来优化细粒度文本到图像的生成；Johnson等[12]提出了一种从场景图生成图像的方法，能够显式地推理文本中对象及其关系。图像和文本之间的转换涉及计算机视觉和自然语言领域，这就要求图像字幕生成任务既要保证文本的空间感知质量，又要保证图像与文本之间的结构相似性和语义一致性。

图像字幕生成任务关注于视觉关系。视觉关系检测涉及对对象的定位和识别，以及对象之间的交互及分类[13]。将每种可能的语义关系进行组合，并将视觉关系检测看做一个分类任务研究图像中对象之间的语义关系。Dai等[14]利用深度学习解决视觉关系检测问题，并将预测输出表述为（主语、谓语、宾语）三元组，利用三元组之间的空间配置和统计依赖关系来推断它们的类标签；Li等[15]提出了一个多级场景描述网络，以端到端方式解决了3个视觉任务；Lu等[16]利用语义词嵌入来微调关系的可能性，最后将预测关系中的对象定位为图像中的边界框；Xu等[17]使用场景图来明确地建模对象及其关系，利用端到端模型解决了场景图推理问题，并通过消息传递迭代学习来改进预测；Yao等[18]提出使用图卷积神经网络（Graph convolutional network, GCN）来对物体的空间和语义信息进行表示，从而学习区域级特征。

神经网络在图像字幕生成中的应用越来越受到关注，因此，本文基于深度学习方法，设计了基于图神经网络和引导向量的图像字幕生成模型，用于重构图像中的位置和语义信息，完成字幕语句生成任务。主要创新点包括：（1）通过改进损失函数的YOLOv4算法对图像中的对象进行检测，实现了对图像信息的采集。（2）采用空间节点权重分配算法和语义关系分类器，将图像中的信息重构成空间位置图和语义关系图，以表达图像中两两对象的相对位置和语义关系。（3）在字幕生成阶段，提出了一种引导网络，用于生成引导向量，为图像中的不同对象分配权重，辅助模型完成字幕语句生成。

1　字幕生成模型

字幕生成模型的构建过程总体可分为3部分：图像信息采集、图像信息重构以及基于引导向量的字

幕生成。

1.1 图像信息采集

为了适应更广泛的应用场景,本文通过改进YOLOv4(You only look once)算法的损失函数,并将其作为目标检测模块,从而实现对原始图像信息的采集,如图1所示。改进的损失函数为

$$\mathrm{Loss}=L_{\mathrm{IoU}}+L_{\mathrm{confidence}}+L_{\mathrm{class}} \tag{1}$$

式中:IoU(Intersection over union)表示预测边界框与真实边界框相交的比例,L_{IoU}为边界框的位置损失;$L_{\mathrm{confidence}}$为置信损失;L_{class}为分类损失。3种损失计算分别表示为

$$\begin{cases} L_{\mathrm{IoU}}=1-\mathrm{IoU}+d^2/c^2+\alpha v \\ L_{\mathrm{confidence}}=\sum_{i=0}^{S^2}K\left[-\lg p+\mathrm{BCE}(\hat{n},n)\right] \\ L_{\mathrm{class}}=\sum_{j=0}^{B}1_{i,j}^{\mathrm{noobj}}\left[-\lg(1-p_{\mathrm{c}})\right] \end{cases} \tag{2}$$

式中:$\mathrm{BCE}(\hat{n},n)=-\hat{n}\lg n-(1-\hat{n})\lg(1-n)$;$c$和$d$分别为两个边界框中心之间的距离和其对角线距离;$S$为网格数;$B$为每个网格对应的编号;$\hat{n}$和$n$分别代表第$i$个网格的第$j$个先验框的真实类别和预测类别;$p$代表当前类别的可能性;$\alpha$、$v$为计算参数,$K$为权重,可分别表示为

$$\alpha=\frac{\nu}{(1-\mathrm{IoU})+\nu},\quad v=\frac{4}{\pi}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right),\quad K=1_{i,j}^{\mathrm{obj}} \tag{3}$$

式中:w^{gt}和h^{gt}为图像真实边界框的宽度和高度;w和h为预测边界框的宽度和高度;K表示如果第i个网格的第j个先验框中存在对象,则值为1,否则为0。

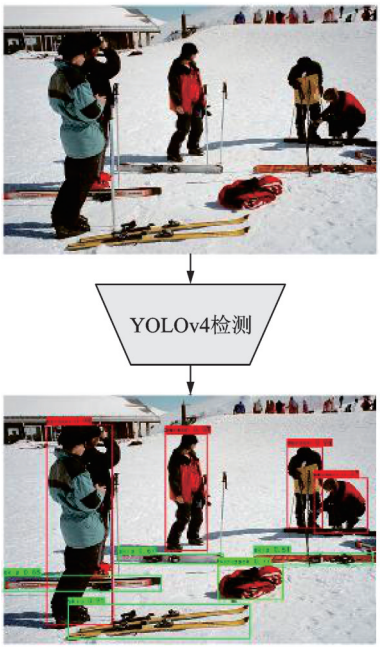


图1 图像信息采集

Fig.1 Image information acquisition

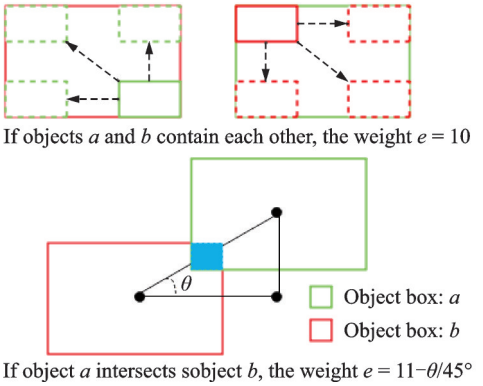
1.2 图像信息重构

空间节点权重分配原理如图2所示。这里将对象之间的空间关系表示为(v_1,v_2),即对象1相对于对象2的几何位置。空间位置图可以定义为

$$G_{\mathrm{space}}=(V,E_{\mathrm{space}}) \tag{4}$$

式中:V代表图像中的对象节点集合;E_{space}代表对象之间的空间位置关系集合。

由于空间位置图是一种无向图,所以利用IoU、对象之间的相对距离和角度来对对象节点之间的边进行权重分配,节点权重分配如算法1所示。对象a和对象b之间进行节点权重分配,当对象框处于相互包含状态,认定权重为10。当对象框处于相交状态,认定权重在0~9范围内取值。权重之间没有大小之分,只用于分类。该算法挖掘了图像中对象之间的空间关系,构造了一种空间位置图,用于表达一幅图像中每两个对象之间的相对位置关系。



(a) If objects a and b contain each other, the weight $e=10$

(b) If object a intersects sobject b, the weight $e=11-\theta/45°$

图2 空间节点分配

Fig.2 Spatial node assignment

算法 1 节点权重分配算法

Input：节点两两搭配个数 N；每个节点对象的坐标 (x,y)；节点对象的中心坐标 $(\hat{x},\hat{y})$

Output：节点之间边的权重 e_i

While $i<N$ do

$$\mathrm{IoU}_i \leftarrow \frac{\mathrm{Intersection}(v_1,v_2)}{\mathrm{Union}(v_1,v_2)};$$

$$d_i \leftarrow \sqrt{(x_{v_i}-x_{v_j})^2+(y_{v_i}-y_{v_j})^2};$$

$$\theta_i \leftarrow \arctan\left(\frac{\hat{x}_{v_1}-\hat{x}_{v_2}}{\hat{y}_{v_1}-\hat{y}_{v_2}}\right);$$

if Inside (v_1,v_2) and $\mathrm{IoU}_i \leftarrow 1$ do

$e_i \leftarrow 10$;

end

if Cover (v_1,v_2) and $\mathrm{IoU}_i \leftarrow 1$ do

$e_i \leftarrow 10$;

end

if $\mathrm{IoU}_i<1$ and $\mathrm{IoU}_i \geqslant 0.8$ do

$e_i \leftarrow 9$;

end

if $\mathrm{IoU}_i>0$ and $\mathrm{IoU}_i<0.8$ do

$$e_i \leftarrow 11-\left\lceil\frac{\theta_i}{45^\circ}\right\rceil;$$

else $e_i \leftarrow 0$;

end

end

上述空间位置图仅仅重构了图像中对象的几何位置信息，而缺少了两两对象之间的语义表达。本文受到文献[16]的启发，将图像中对象的语义关系检测视作分类问题，并在 Visual Genome 数据集上预训练语义关系分类器，能将两两对象间的语义关系分类出来，如图 3 所示。

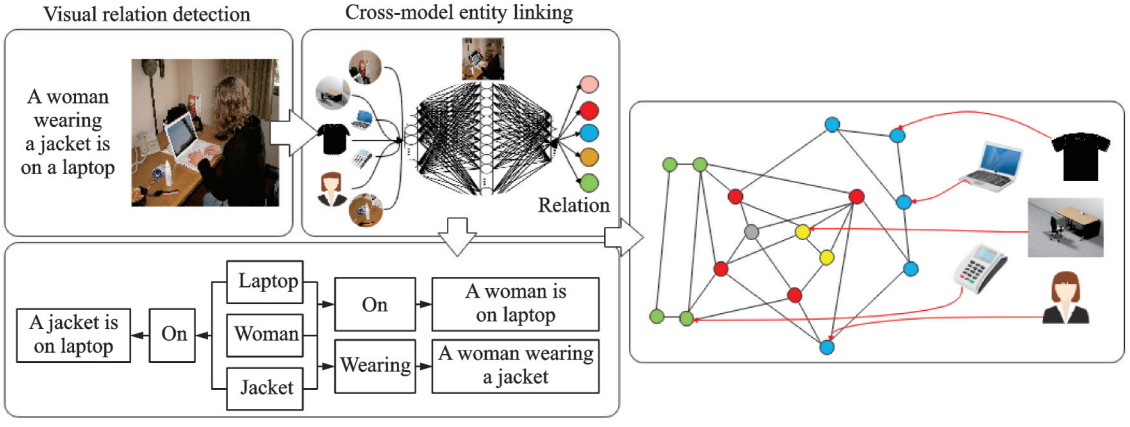


图 3 语义关系分类器

Fig.3 Semantic relation classifier

在预训练语义关系分类器时,使用的损失函数定义为

$$\text{Loss}=-[\,y\lg\hat{y}+(1-y)\lg(1-\hat{y})\,] \tag{5}$$

式中:y表示真实类别;$\hat{y}$表示预测当前类别的概率。

语义信息图的构建步骤如下:

步骤1 将合并框从原图像中截取下来,输入到YOLOv4网络计算,由于YOLOv4的特征图具有原始图像的强语义特征,是不同检测层的特征聚合体。故将最后一层特征图(38×38×255)作为合并框的嵌入表示向量,该向量包含合并框中两个对象的特征。

步骤2 将合并框向量进行全局池化,用于缩小特征图的宽高且不损失信息。将数据(维度=512)输入到一层隐藏层为1 024个神经元分类器中,该分类器会输入每种类别和非关系类别的SoftMax类概率。

步骤3 如果非关系类别的概率小于0.5,则当前用于分类的两个对象之间没有语义关系,节点之间不具有边;否则,取类概率最大的类别作为边标签。语义关系图可以定义为

$$G_{\text{sem}}=(V,E_{\text{sem}}) \tag{6}$$

式中:V代表图像中的对象节点集合;E_{sem}代表对象之间的语义关系集合。

空间位置图和语义关系图的构建实现了图像信息的重构,使得字幕生成模型提取丰富的信息。本文引入GCN网络对其进行上下文编码,编码过程如算法2所示。

算法2 基于空间位置图和语义关系图的GCN前向传播算法

Input: 空间位置图$G_{\text{space}}(V,E_{\text{space}})$;语义关系图$G_{\text{sem}}(V,E_{\text{sem}})$;邻接矩阵$\boldsymbol{A}$,$\forall\boldsymbol{A}\in\mathbf{R}^{K\times K}$;图中的节点个数$K$;$\boldsymbol{I}$是单位矩阵,$\forall\boldsymbol{I}\in\mathbf{R}^{K\times K}$;滤波器参数的矩阵$\boldsymbol{\theta}^{(l)}$;dense层的权重参数$\theta_{\text{dense}}$;节点的特征向量$\boldsymbol{H}^{(0)}$,$\boldsymbol{H}^{(0)}\in\mathbf{R}^{K\times 512}$

Output: 节点向量Out

$\tilde{\boldsymbol{A}}\leftarrow\boldsymbol{A}+\boldsymbol{I}$;

$D_{ij}\leftarrow\sum_j\tilde{A}_{ij}$;

for k=1,2,…,K do

$\tilde{D}_{kk}\leftarrow D_{kk}$//对角矩阵;

end

for l=0,1 do

$\boldsymbol{H}^{(l+1)}\leftarrow\text{ReLU}\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{A}\tilde{\boldsymbol{D}}^{\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{\theta}^{(l)}\right)$, $\forall l\in\{0,1\}$/*在编码器的图卷积层,每个节点使用相邻节点信息*/;

end

Out$\leftarrow$SoftMax($\boldsymbol{H}^{(2)}\theta_{\text{dense}}$)/*将图卷积层的输出馈送到dense层,利用SoftMax函数给出最终概率输出*/。

1.3 基于引导向量的字幕生成

在字幕生成阶段,在将图像中对象的空间和语义信息转换为文本信息时,不仅需要准确描述图像的内容,还需要适应当前训练的语言模型。因此,本文基于BP神经网络设计了一种引导网络(Guide network, GN),对原始图像的特征建模,将其输出作为生成模型的引导信息,记作引导向量$\boldsymbol{l}$。为了减少训练时间且最大程度地提高模型的优化能力,引导网络的隐藏层设定为2层,神经元个数分别设置为512和1 024,输出维度和上述GCN网络输出的特征向量$\boldsymbol{a}_t$维度相同。加入引导向量之后的特征向量表达为

$$\boldsymbol{x}_t=\boldsymbol{a}_t+W_t\boldsymbol{l}_t \tag{7}$$

式中：$\boldsymbol{a}_t$为GCN输出的特征向量；$\boldsymbol{W}_t$为需要更新的学习参数；$\boldsymbol{l}_t$为引导网络生成的引导向量。

在训练阶段，引导网络可以作为一个组件被插入到当前的GRU生成模型中，并以端到端的方式进行训练。利用基于引导网络分配注意力信息的GRU模型，对GCN输出的区域特征向量[19]进行解码，即字幕生成的主要步骤如下：

步骤1 对空间位置图和语义关系图编码后，使用式(8)对两种特征向量进行平均池化，得到$\bar{\boldsymbol{a}}_t$。

$$\bar{\boldsymbol{a}}_t=\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{a}_t^{(i)} \tag{8}$$

步骤2 由式(8)获得的特征向量随时间展开输入到第一层GRU单元，表达式为

$$h_t^1=f_1[h_{t-1}^2,\boldsymbol{W}_s\boldsymbol{w}_t,x_t] \tag{9}$$

式中：$\boldsymbol{W}_s\in\mathbf{R}^{D_s^1\times D_s}$代表输入向量$\boldsymbol{W}_t$的变换矩阵表示；$h_t^1\in\mathbf{R}^{D_h}$为第一层GRU单元的输出；$f_1$为第一层GRU单元内的更新函数。

步骤3 根据第一层GRU单元的输出，所有区域特征的归一化后的概率分布为

$$a_{t,i}=\boldsymbol{W}[\tanh(\boldsymbol{W}_f\boldsymbol{l}_t^{(1)}+\boldsymbol{W}_h h_t^1)] \tag{10}$$

$$\lambda_t=\mathrm{softmax}(\boldsymbol{a}_t) \tag{11}$$

式中：$a_{t,i}$为$\boldsymbol{a}_t$的第i个元素；$\boldsymbol{W}\in\mathbf{R}^{1\times D_a}$，$\boldsymbol{W}_f\in\mathbf{R}^{D_a\times D_a}$且$\boldsymbol{W}_h\in\mathbf{R}^{D_a\times D_h}$，三者均为变换矩阵；$\lambda\in\mathbf{R}^K$代表归一化后的概率分布，其第$i$个元素$\lambda_{t,i}$是$l_i^{(1)}$的概率。

步骤4 将λ_t其作为第二层GRU单元的输入，从而预测出当前时间步的单词w_{t+1}。

通过这种机制，每个时间步长的误差信息都可以传播到引导网络，从而实现了引导向量的自适应学习。

2 实 验

2.1 数据集及参数设置

实验基于Visual Genome数据集对YOLOv4网络和语义分类器进行预训练，且基于MSCOCO测试集对YGG模型进行性能评估。随机选择20%的图像样本用于评估，其余图像用于训练。为提高模型的泛化能力，防止模型过拟合，在训练之前对数据集进行的处理包括：对图像进行随机水平翻转和随机裁剪等方法进行数据增强，将训练集中所有描述转换为小写，丢弃出现次数小于5次的罕见词等。经处理后的数据分布如表1所示。本文基于Pytorch实现图像字幕生成模型。在训练过程中，使用了Adam优化器。设置初始学习率为0.001，mini-batch size为32，最大迭代次数为5 000。在GRU的语句生成部分使用Beam search策略，并设置Beam size为3。

表1 不同数据集参数对比

Table 1 Comparison of different dataset parameters

数据集	训练数据	验证数据	测试数据	描述	单词表大小	图像主题
MSCOCO	82 783	40 504	40 775	5/40	9 957	综合
Visual Genome	11 788	5 122	5 347	50	3 147	综合

2.2 参数分析

选用BLEU-1、BLEU-2、BLEU-3、BLEU-4、METEOR、ROUGE-L和CIDEr-D等评价指标来评估模型性能。YGG模型在字幕生成阶段使用了Beam search这种启发式图搜索策略，该策略具有超参数Beam size，模型在最佳参数下会收敛到最优。因此，本文分析了Beam size在测试阶段对模型的影响。基于CIDEr-D和METEOR指标，将不同条件下的YGG模型进行了对比，并把Beam size控制在{1,3,5,7,9,10}范围内，MSCOCO数据集下模型性能随参数的变化如图4所示。

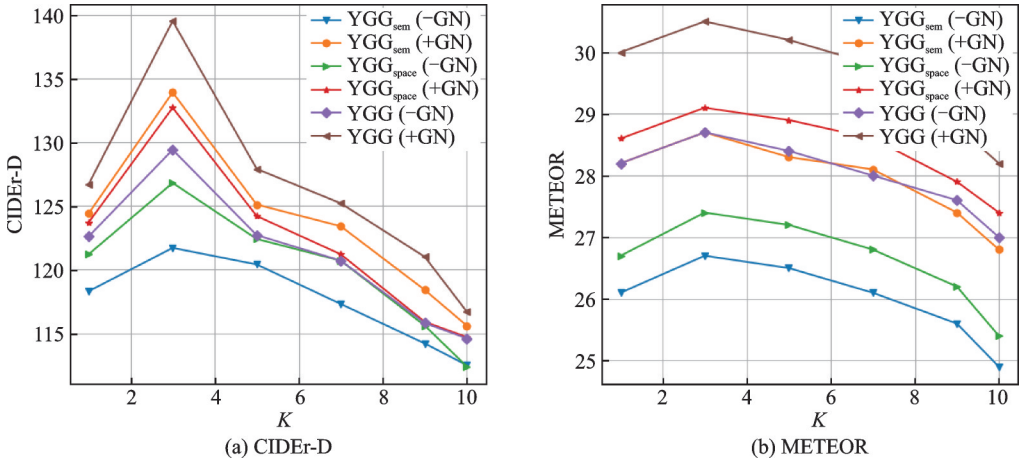


图4 Beam size对CIDEr-D和METEOR的影响

Fig.4 Effect of beam size on CIDEr-D and METEOR

从图4可以看出,Beam size K的设置对模型的影响较大:当K大于3时,CIDEr-D和METEOR逐渐下降,这表明真实语句和生成语句的相似度和语义一致性降低;直到K为10时,指标分数达到最低点。因此,在YGG模型中,设置K为3较为合适。

2.3 算法对比实验

为了更全面地评估YGG模型的效果,本文将文献[9-10,20-24]的方法划分为3类,即基于Global的方法、基于Grid的方法以及基于Self-attention的方法。将提出的YGG模型与最新方法在MSCOCO数据集上进行了对比试验,其中c5和c40分别表示数据集中每个图像有5个或40个字幕描述,结果如表2所示。可以看出,第1类和第2类方法得分普遍比较低,这是由于前者全局处理原始图像,没有考虑局部区域;后者对原始图像进行网格划分,相较于前者有效地利用局部区域信息,但仍然缺少细粒度的细节提取能

表2 不同方法在MSCOCO数据集上的性能对比

Table 2 Performance comparison of different methods on MSCOCO dataset

	方法	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
		c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
基于Global	PG-SPIDEr[20]	74.3	—	59.1	—	44.5	—	33.2	—	25.7	—	55.0	—	101.3	—
	PG-BCMR[20]	75.4	—	57.8	—	43.3	—	32.2	—	25.1	—	54.4	—	100.0	—
基于Grid	MaBi-LSTM[21]	79.3	—	61.2	—	47.3	—	36.8	—	28.1	—	56.9	—	116.6	—
	Stack-Cap(XE)[22]	76.2	—	60.4	—	46.4	—	35.2	—	26.5	—	—	—	109.1	—
	Stack-Cap(C2F)[22]	78.6	—	42.5	—	47.9	—	36.1	—	27.4	—	—	—	120.4	—
基于Self-attention	VinVL[23]	81.9	96.9	66.9	92.4	52.6	84.7	40.4	74.9	**30.6**	**40.8**	**60.4**	**76.8**	134.7	138.7
	GET[24]	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
	DLCT[9]	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.3	135.4
	AutoCaption[10]	82.5	96.6	67.8	91.9	53.3	84.2	41.1	74.3	30.3	40.1	**60.4**	76.0	135.9	138.9
	X-Transformer[24]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
	NG-SAN[24]	80.8	95.0	65.4	89.3	50.8	80.6	38.8	70.2	29.2	38.4	58.7	74.0	126.3	128.6
	YGG$_{sem}$	81.7	95.5	67.1	90.7	51.4	84.1	39.8	71.4	28.7	39.1	57.2	71.9	133.9	134.4
	YGG$_{space}$	82.0	96.3	66.9	91.2	50.6	82.7	40.3	73.6	29.1	40.5	58.6	73.3	132.7	135.2
	YGG	**83.1**	**97.2**	**68.3**	**92.8**	**53.7**	**84.9**	**41.2**	**75.1**	30.5	40.5	59.7	75.3	**139.5**	**142.1**

力。基于Self-attention的方法在各个指标上的得分普遍比较高，这是由于该类方法引入注意力机制，不仅利用局部区域信息，而且有效地关注上下文信息和细粒度细节。本文使用GCN对图像的空间位置和语义信息进行上下文表示，YGG_{sem}代表只利用语义信息的YGG模型，YGG_{space}代表只利用空间位置信息的YGG模型。以c5数据为例，两者在BLEU-1、BLEU-2、CIDEr-D指标上的得分都高于平均水平，即81.85%、66.81%以及109.23%。这表明GCN可以从语义和空间层面提取有效信息，提高模型对图像的信息提取能力。从整体考虑，YGG在CIDEr-D指标上分别达到了139.5%和142.1%，相较于该指标上的最优方法AutoCaption分别提高了3.6%和3.2%。此外，YGG在其他4种指标上的得分相较于当前最优模型都有了不同程度的提升。研究结果表明，将空间位置和语义信息相结合，能够最大程度上发挥YGG模型字幕生成效果。

2.4 消融实验

为了检验引导网络的作用，对不同情况下的YGG模型进行消融实验，结果如表3所示。从表3可以看出，添加引导网络的模型在不同图像信息重构方式下性能都有所提升，其中，在c5数据上，YGG_{sem}、YGG_{space}以及完整YGG模型的CIDEr-D指标分别提升了12.2%、5.9%以及10.1%；在c40数据上，YGG_{sem}、YGG_{space}以及完整YGG模型的CIDEr-D指标分别提升了7.2%、7.6%以及7.2%。实验表明，采用的引导网络能够有效地提高模型的精度，使得模型更加稳健。

表3 消融实验结果

Table 3 Ablation results

方法	BLEU-1		METEOR		CIDEr-D	
	c5	c40	c5	c40	c5	c40
YGG_{sem}(−GN)	80.2	92.3	26.7	37.1	121.7	127.2
YGG_{sem}(+GN)	81.7	95.5	28.7	39.1	133.9	134.4
YGG_{space}(−GN)	80.9	94.7	27.4	38.6	126.8	127.6
YGG_{space}(+GN)	82.0	96.3	29.1	40.5	132.7	135.2
YGG(−GN)	82.3	96.1	28.7	38.4	129.4	134.9
YGG(+GN)	**83.1**	**97.2**	**30.5**	**40.5**	**139.5**	**142.1**

引导网络在一定程度上实现了图像注意力的分配，使得模型在字幕生成任务上对图像中的不同对象进行选择性表示，图像中不同对象的权重分布情况如图5所示，图中第1行为YGG模型分配权重，第2行为真实字幕权重。可以看出，引导网络对每张图像的选择性表示接近于真实比例，并且能够准确地给予蕴含重要信息的对象以更大的权重，这表明引导网络具有较好的权重分配能力。

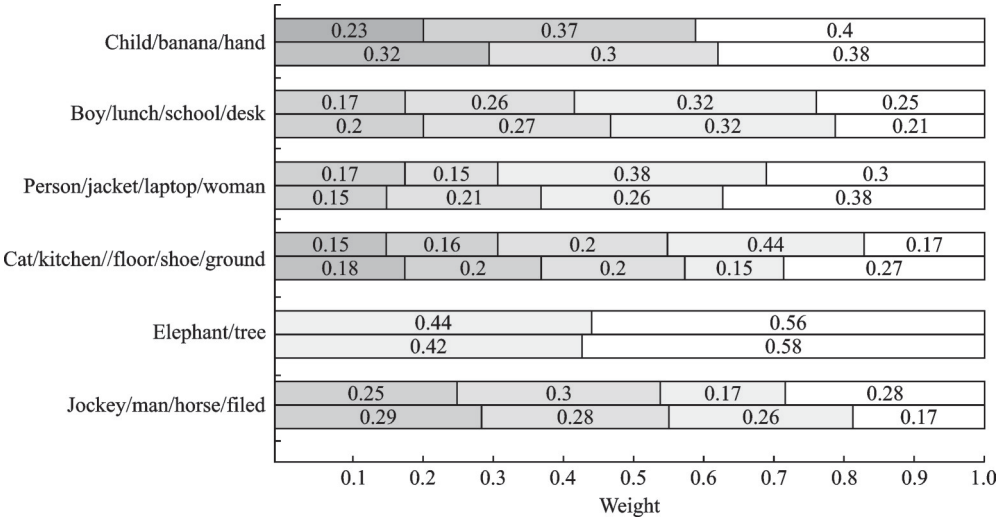


图5 YGG分配权重和真实权重的对比

Fig.5 Comparison between assigned weight and real weight of YGG model

2.5 定性分析

图6展示了几个具体的示例,每个示例包括目标检测后的图像、检测到的语义关系图以及生成字幕。从图中可以看出,YGG模型能够产生“packed in”“written on”“looking at”和“on top of”等复杂语义关系,将对象和语义关系进行组合,产生了更准确的字幕语句。基于MSCOCO数据集的图像字幕对比示例如表4所示。YGG模型可以很好地指出图像中的对象,并准确描述对象的特征,如颜色、性别和年

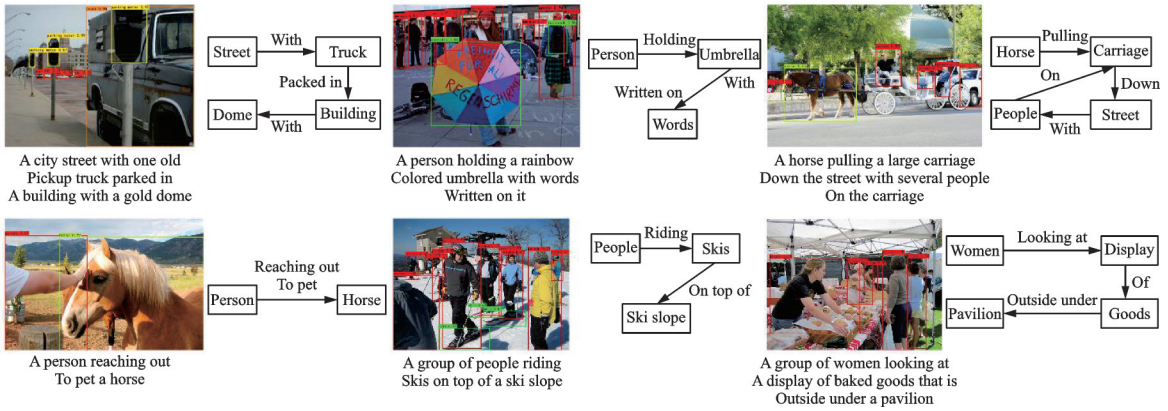


图6 YGG模型生成的一些字幕示例

Fig.6 Several image captioning examples generated by YGG model

表4 生成字幕与真实字幕的对比

Table 4 Comparison of generated caption with real caption

图片	真实字幕	生成字幕	准确度
	A little boy holding a banana in his hands	A little child holding a banana on her hand	0.866
	People were skiing in the snow	A group of people were skiing in the snow with skis	0.743
	White and black goats standing and sitting in A grassy field	Two goats standing in a grassy field	0.786
	A person in a black jacket is on a white laptop	A woman wearing a jacket is on a laptop	0.795
	A pretty young lady holding up a pizza	Two young people holding up a pizza	0.823
	The man is playing a game of tennis On the court	A man is playing tennis on the playground	0.851
	A black cat laying on a kitchen floor next to A pair of black shoes	A black cat lying on the ground next to black shoes	0.824
	An elephant pushing up against a tree trunk	An elephant is hitting a tree	0.782
	A jockey riding a horse on a horse track	A man riding a horse on the grass field	0.829

龄等。生成的语句并没有明显的语法错误,符合自然语言的使用规则。以第6张图像为例,YGG能够产生新的描述信息,并且更符合图像场景。

3 结束语

本文提出了一种基于图神经网络和引导网络的图像字幕生成模型,以图像中的空间位置信息和对象之间的语义信息构造空间位置和语义图,将其作为图卷积神经网络的输入,并提出将引导网络嵌入到GRU生成模型中辅助语句生成。在MSCOCO图像数据集上的对比实验和消融实验结果表明,本文提出的YGG模型能准确地生成图像字幕。未来可以在生成模型中添加注意力机制来提取更加全面的图像信息,从而获取丰富的视觉表示,同时添加绘画、视频等模型数据,设计新的多模态的网络结构,从而进一步提升字幕生成能力。

参考文献:

[1] 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. [S.l.]: IEEE, 2017: 7008-7024.

[2] 郑昌艳,张雄伟,曹铁勇,等.一种基于LSTM-RNN的喉振传声器语音盲增强算法[J].数据采集与处理,2019,34(4): 615-624.
ZHENG Changyan, ZHANG Xiongwei, CAO Tieyong, et al. A blind enhancement algorithm for throat microphone speech based on LSTM recurrent neural networks[J]. Journal of Data Acquisition and Processing, 2019, 34(4): 615-624.

[3] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 4878-4886.

[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. [S.l.]: IEEE, 2015: 3156-3164.

[5] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN)[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2014: 2984-2998.

[6] 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. [S.l.]: IEEE, 2017: 375-383.

[7] 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. [S.l.]: IEEE, 2016: 4651-4659.

[8] 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. [S.l.]: IEEE, 2017: 873-881.

[9] ZHU X, WANG W, GUO L, et al. AutoCaption: Image captioning with neural architecture search[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2020: 1235-1248.

[10] PAN Y, YAO T, LI Y, et al. X-linear attention networks for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 10971-10980.

[11] XU T, ZHANG P, HUANG Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 1316-1324.

[12] JOHNSON J, GUPTA A, LI F F. Image generation from scene graphs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 1219-1228.

[13] 李家宁,田永鸿.神经形态视觉传感器的研究进展及应用综述[J].计算机学报,2021,44(6): 1258-1286.
LI Jianing, TIAN Yonghong. Recent advances in neuromorphic vision sensors: A survey[J]. Chinese Journal of Computers, 2021, 44(6): 1258-1286.

[14] DAI B, ZHANG Y, LIN D. Detecting visual relationships with deep relational networks[C]//Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 3076-3086.
[15] LI Y, OUYANG W, ZHOU B, et al. Scene graph generation from objects, phrases and region captions[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 1261-1270.
[16] LU C, KRISHNA R, BERNSTEIN M, et al. Visual relationship detection with language priors[C]//Proceedings of European Conference on Computer Vision. [S.l.]: Springer, 2016: 852-869.
[17] XU D, ZHU Y, CHOY C B, et al. Scene graph generation by iterative message passing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 5410-5419.
[18] YAO T, PAN Y, LI Y, et al. Exploring visual relationship for image captioning[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2018: 684-699.
[19] 叶继华,万叶晶,刘长红,等.基于多子空间直和特征融合的人脸识别算法[J].数据采集与处理,2016,31(1): 102-107.
YE Jihua, WAN Yejing, LIU Changhong, et al. Face recognition algorithm of feature fusion based on multi-subspaces direct sum[J]. Journal of Data Acquisition and Processing, 2016, 31(1): 102-107.
[20] GE H, YAN Z, ZHANG K, et al. Exploring overall contextual information for image captioning in human-like cognitive style [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 1754-1763.
[21] GUO L, LIU J, ZHU X, et al. Normalized and geometry-aware self-attention network for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 10327-10336.
[22] ZHANG P, LI X, HU X, et al. Vinvl: Revisiting visual representations in vision-language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 5579-5588.
[23] LUO Y, JI J, SUN X, et al. Dual-level collaborative transformer for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021, 35(3): 2286-2293.
[24] GU J, CAI J, WANG G, et al. Stack-captioning: Coarse-to-fine learning for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2018.

作者简介:

佟国香(1968-),女,博士,副教授,研究方向:嵌入式系统设计与开发、人工智能及数据科学,E-mail: tonggx@usst.edu.cn。

李乐阳(1997-),**通信作者**,男,硕士研究生,研究方向:深度学习、机器学习,E-mail: muzili1212@163.com。

(编辑:王静)