

# 多场景融合的细粒度图像描述生成算法

李欣晔, 张承强, 周雄图, 郭太良, 张永爱

(福州大学物理与信息工程学院, 福建 福州 350108)

**摘要:** 针对图像描述生成任务在不同场景下表现不佳的缺点, 提出一种融合卷积神经网络和先验知识的多场景注意力图像描述生成算法。该算法通过卷积神经网络生成视觉语义单元, 使用命名实体识别对图像场景进行识别和预测, 并使用该结果自动调整自注意力机制的关键参数并进行多场景注意力计算, 最后将得到的区域编码和语义先验知识插入 Transformer 文本生成器中指导句子的生成。结果表明, 该算法有效解决了生成的描述缺少关键场景信息的问题。在 MSCOCO 和 Flickr30k 数据集上对模型进行评估, 其中 MSCOCO 数据集的 CIDEr 得分达到 1.210, 优于同类图像描述生成模型。

**关键词:** 图像描述生成; 卷积神经网络; 命名实体识别; 多场景注意力; Transformer 结构

中图分类号: TP391.4 文献标志码: A DOI: 10.3969/j.issn.1006-2475.2021.09.001

## Multi-scene Fusion Algorithm for Fine-grained Image Caption

LI Xin-ye, ZHANG Cheng-qiang, ZHOU Xiong-tu, GUO Tai-liang, ZHANG Yong-ai

(College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China)

**Abstract:** In terms of the poor performance of image caption task in different scenes, a multi-scene image caption generation algorithm based on convolutional neural network and prior knowledge is proposed. The algorithm generates visual semantic units by convolutional neural network, then uses named entity recognition to identify and predict image scenes, uses the result of classifying to adjust the focusing parameter of self-attention mechanism automatically, and calculate the multi-scene attention score. Finally, the obtained region coding and semantic prior knowledge are inserted into Transformer text generator to guide sentence generation. The results show that the algorithm can effectively solve the problem that the caption lacks the key scene information. Evaluation indicators are used to evaluate the model on the MSCOCO dataset and Flickr30k dataset, and the CIDEr score of MSCOCO dataset reaches 1.210, which is better than similar image description generation models.

**Key words:** image caption; CNN; NER; multi-scene attention; Transformer structure

## 0 引 言

图像描述是对图像的语义层面进行分析和理解, 生成接近于人类语言的描述, 近年来成为计算机视觉领域备受关注的领域之一。在图像描述生成领域中, 早期模型主要基于模版和检索的方法, 当前主流的方法为基于深度学习的方法<sup>[1-2]</sup>。基于模版的方法类似于填词, 将图片中的视觉元素映射到语义空间后填入到句子模板中, 该方法简单易实现但生成的句式结构固定; 基于检索的方法, 类似于查找图像特征, 在数据

库中搜索相似的图像特征, 并从句子库中获取相似图像的描述, 缺点在于难以生成新颖的描述。近年来, 随着基于深度学习的物体检测模型在检测精度和速度上不断提升, 使得图像描述生成方法有了突破性的进展。Vinyals 等人<sup>[3]</sup>提出了一个端到端的卷积神经网络模型 (Neural Image Caption, NIC), 该模型联合了卷积神经网络 (Convolutional Neural Network, CNN)<sup>[4]</sup> 和循环神经网络 (Recurrent Neural Network, RNN), 使用长短期记忆网络 (Long Short Term Memory, LSTM)<sup>[5]</sup> 作为解码器来生成句子, 并将直接将图像

收稿日期: 2020-12-30; 修回日期: 2021-01-25

基金项目: 国家自然科学基金资助项目 (61775038); 国家自然科学基金青年基金资助项目 (61904031)

作者简介: 李欣晔 (1994—), 女, 福建福安人, 硕士研究生, 研究方向: 计算机视觉, E-mail: 598172095@qq.com; 张承强 (1993—), 男, 硕士研究生, 研究方向: 深度学习, 计算机视觉, E-mail: 2541594024@qq.com; 周雄图 (1982—), 男, 教授, 博士, 研究方向: 信息显示技术, E-mail: xtzhou@fzu.edu.cn; 郭太良 (1963—), 男, 研究员, 硕士, 研究方向: 物理电子学, E-mail: gtl\_fzu@hotmail.com; 张永爱 (1977—), 男, 研究员, 博士, 研究方向: 信息显示技术, E-mail: yongaizhang@fzu.edu.cn。

特征作为输入提供给 LSTM,在图像描述领域取得了巨大的突破。Mao 等人<sup>[6]</sup>提出了一种多模态递归神经网络(M-RNN),首次将图像描述任务分割成 2 个分支,使用卷积神经网络提取图像特征,循环神经网络进行文本嵌入,并将得到的特征输入到多模态层中预测句子。Xu 等人<sup>[7]</sup>进一步介绍了一种基于注意力的模型,使模型在生成相应单词时重点关注对应的区域。

目前,主流的基于深度学习的图像描述生成算法,相较于早期的方法已有了较大的改进,但仍存在一些局限性。仅仅依靠卷积神经网络提取出的视觉信息构建图像特征,无法有效地表达图像在语义层面的信息。虽然注意力机制可增强模型对重点区域信息的提取能力,但不同场景使用统一的注意力机制,会导致对图像信息的过度解读或信息缺失。LSTM 逐字预测单词特性,使用于修饰对象的属性特征词先于对象产生,这将导致生成的属性与对象无关。

本文针对图像描述生成任务在不同场景下表现不佳、逐字生成单词的缺陷,提出一种基于卷积神经网络和 Transformer<sup>[8]</sup>结构的图像描述生成算法,引入一种全新的多场景注意力机制。为了有效提取图像的深层信息,本文采用基于 ResNet101<sup>[9]</sup>的 Faster R-CNN<sup>[10]</sup>网络对数据集中的图像进行特征提取。模型设计一种基于图卷积网络(Graph Convolution Net-

work, GCN)<sup>[11-12]</sup>的多场景注意力机制,用于增强模型在不同场景下对区域特征的提取能力。同时,在文本生成部分改进了 Transformer 结构,加入了细粒度可控的区域编码,能够有效解决预测的属性特征词偏离目标对象的问题,使得生成的句子更贴合图像。

## 1 模型架构

图像描述生成任务通常遵循 ENCODER 和 DECODER 框架<sup>[13-14]</sup>。在 ENCODER 端对图像进行特征提取,获取图像的视觉信息并将其变为编码向量,本文中 ENCODER 端采用深度学习目标检测框架 Faster R-CNN 和 GCN;在 DECODER 端需要对已编码的图像信息进行解码操作,本文采用由自注意力机制和前馈层组成的 Transformer 语言模型来完成。本文算法在 ENCODER 端引入了多场景注意力机制,利用命名实体识别算法得到的场景信息对视觉特征进行自主可控的多场景注意力处理,并生成场景先验知识。融合多场景注意力机制处理得到的物体、属性、关系特征,作为输入传递给 DECODER 端。在语言解码模型中,将 ENCODER 端得到的先验知识作为持久存储向量,对图像区域之间的关系进行 Transformer 多级编码,生成更贴合图像场景的文本描述。本文算法使用 Faster R-CNN + Attention + GCN + Transformer 的基本框架来完成,其流程如图 1 所示。

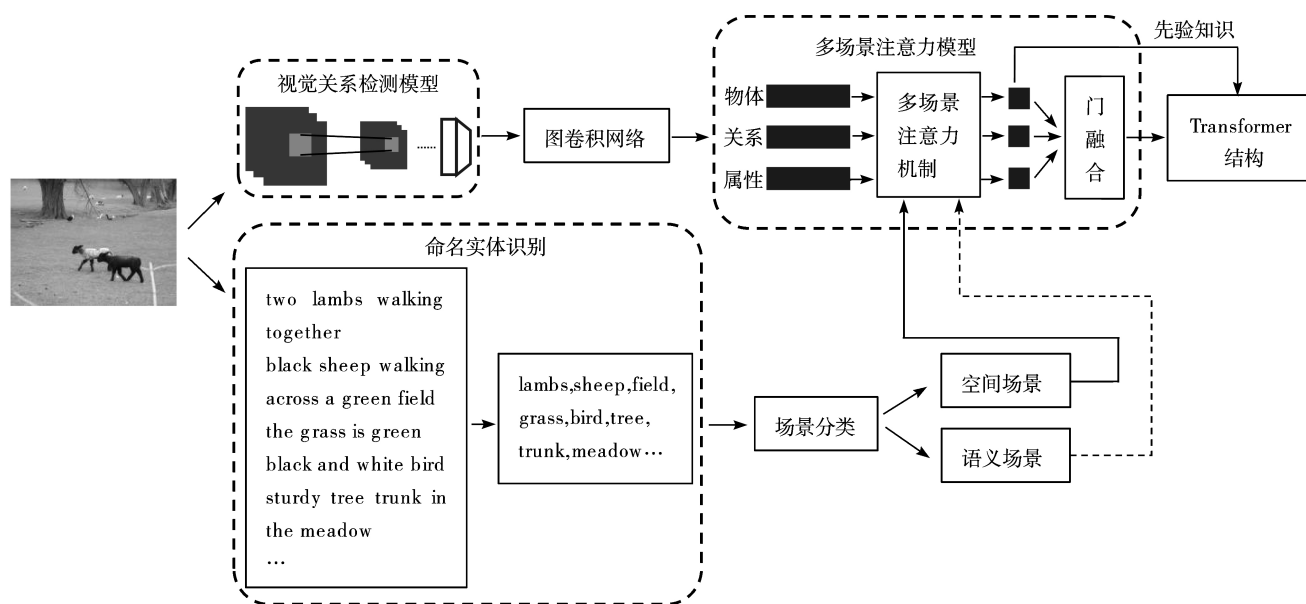


图 1 本文算法流程

### 1.1 多场景特征提取

本文使用预训练的 Faster R-CNN 作为物体检测器,使用多层感知器(Multi-Layer Perceptron, MLP)用于预测物体的属性,采用 MOTIFNET<sup>[15]</sup>作为物体关系检测器,将得到的视觉关系单元作为图卷积网络的

输入,生成物体、关系、属性 3 类节点。对 Visual Genome(VG)数据集<sup>[16]</sup>的文本标注内容,使用命名实体识别(Named Entity Recognition, NER)<sup>[17]</sup>对文本中的名词进行分类和预测,将数据集中的名词分为 7 个类别,分别是人物、动物、植物、风景、建筑、交通工具以

及其他,除“其他”类外,将其余 6 类名词设定为场景名词。经过多次实验,设定当输入图片 I 的标注文本 A 经过命名实体识别后得到的场景名词占有名词比例大于 75% 时,认为图片 I 为空间场景图,否则为语义场景图。为了提高模型对图像场景识别的准确度,在场景字典中插入场景标志位,空间场景图的标志位为 0,语义场景图的标志位为 1,并生成包含图片 id 和场景标志位的字典,将其编码后插入到多场景注意力机制中,对图像物体、关系、属性节点进行多场景注意力处理,得到的视觉特征和场景先验知识作为 Transformer 语言模型的输入,进行文本解码。视觉关系检测模型结构如图 2 所示。

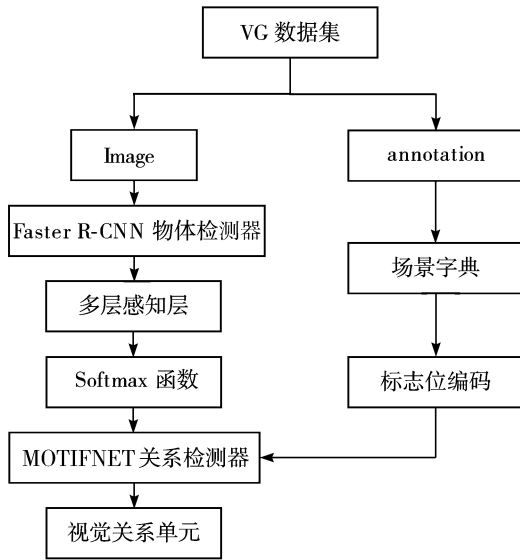


图2 关系检测模型

对具有标志位 0 的空间场景图,采用多场景注意力机制中的 hard-attention 机制<sup>[7]</sup>,重点关注局部区域的物体,设定一个  $t$  时刻的位置变量  $s_t$ ,表示在  $t$  时刻模型聚焦于某一个图像区域,为 one-hot 编码,权重  $\alpha_{t,i}$  为  $t$  时刻图像区域  $a_i$  被选中概率,并且每次只选取一个感兴趣区域。为此,引入变量  $s_{t,i}$ ,每个时刻  $t$  的序列为  $[s_{t,1}, \dots, s_{t,i}]$ ,当区域  $i$  被选中时概率值取 1,否则取 0,如公式(1)所示。 $s_t$  为生成第  $t$  个单词时注意力机制需要关注的位置信息,满足参数为  $\alpha_i$  的多元贝努利分布,如公式(2)所示。构建基于空间内的物体节点,在物体间添加一个关系节点和多个属性节点,建立空间物体节点指向对应属性节点的边和指向物体间关系节点的边,如图 3(a) 所示。

$$\hat{Z}_t = \sum_i s_{t,i} a_i \quad (1)$$

$$p(s_{t,i} = 1 | s_{j < t}, a) = \alpha_{t,i} \quad (2)$$

对具有标志位 1 的语义场景图,采用多场景注意力机制中的 soft-attention,不同于 hard-attention 机制仅关注一个感兴趣区域,soft-attention 会关注视觉关系单元中的全局信息,针对不同的位置计算不同的权

重,此时  $\alpha_{t,i}$  表示图像区域  $a_i$  在  $t$  时刻所有被选中的区域里的占比,通过对区域  $a_i$  与对应的权重  $\alpha_{t,i}$  进行加权可得到最终的注意力结果,如公式(3)所示。将物体在图像中的区域信息添加到图神经网络中作为物体节点,增加多个属性节点,如图 3(b) 所示。

$$E_{p(s_t | a)} [\hat{Z}_t] = \sum_{i=1}^L \alpha_{t,i} a_i \quad (3)$$

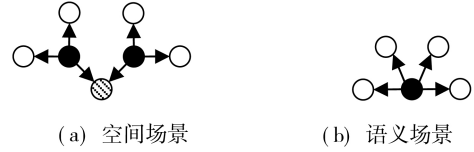


图3 不同场景下的节点图

图 3 中斜纹节点表示物体间的关系,黑色节点表示空间中的物体,白色节点为物体对应的属性。

## 1.2 Transformer 语言模型

经过上述处理后,输入数据已从图像转换为特征向量,由于语言模型逐字生成的特点,一旦上一时刻的单词预测不准确,将影响后续的预测准确结果,导致生成的描述偏离图像信息。为了克服逐字生成方式的缺点和自注意力机制的局限,本文提出细粒度可控的 Transformer 结构。将多场景注意力机制处理得到的场景信息作为先验知识 p-obj,在生成第一个单词时知道语言模型学习图像的主要场景信息,通过持久存储向量对图像区域之间的关系进行多级编码,避免出现文本偏离图像的情况。Transformer 结构的 encoder 部分由多头注意力机制和全连接前馈网络 (Feed Forward) 组成<sup>[18]</sup>,结构如图 4 所示。

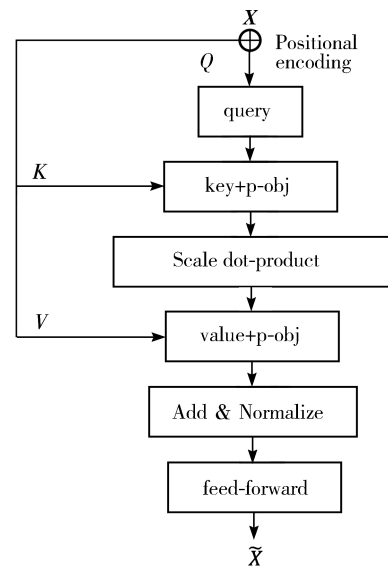


图4 插入 p-obj 的 Transformer encoder 结构

模型中  $X$  表示输入的图像特征向量,在输入注意力机制前,在 key 和 value 集合中扩展可以插入先验知识(p-obj)的向量槽,学习尚未从输入向量  $X$  处学到的知识,定义公式为:

$$T_{p\text{-obj}}(\mathbf{X}) = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \quad (4)$$

$$\mathbf{Q} = \mathbf{W}_q \mathbf{X} \quad (5)$$

$$\mathbf{K} = [\mathbf{W}_k \mathbf{X}, \mathbf{p}\text{-obj}_k] \quad (6)$$

$$\mathbf{V} = [\mathbf{W}_v \mathbf{X}, \mathbf{p}\text{-obj}_v] \quad (7)$$

其中,  $[\cdot, \cdot]$  表示拼接向量,  $\mathbf{W}_q$ 、 $\mathbf{W}_k$ 、 $\mathbf{W}_v$  表示注意力机制使用的投影的线性矩阵,  $\mathbf{p}\text{-obj}_k$  和  $\mathbf{p}\text{-obj}_v$  为多场景注意力机制中生成的先验知识矩阵。单个注意力机制结构将增强记忆的注意力操作重复 6 次, 然后将得到的单头注意力结果串联成多头注意力, 线性变换后输入 Add & Normalize 层进行归一化处理, 定义公式为:

$$\mathbf{F}(\mathbf{X})_i = \mathbf{U}_\sigma(\mathbf{V}\mathbf{X}_i + b) + c \quad (8)$$

$$\mathbf{Z} = \text{AddNorm}(T_{p\text{-obj}}(\mathbf{X})) \quad (9)$$

$$\tilde{\mathbf{X}} = \text{AddNorm}(\mathbf{F}(\mathbf{Z})) \quad (10)$$

其中,  $\mathbf{X}_i$  为输入集的第  $i$  个向量,  $\mathbf{F}(\mathbf{X})_i$  指示输出集的第  $i$  个向量,  $\sigma(\cdot)$  是 ReLU 激活函数,  $\mathbf{V}$  和  $\mathbf{U}$  是可学习的权重矩阵,  $b$  和  $c$  是偏差项。

### 1.3 损失函数

本文使用交叉熵损失函数来训练模型, 并使用强化学习在序列生成中进行识别。在使用强化学习训练时, 对 self-critical sequence 采用 beam search 的一种变体: 解码时, 在每个时刻从解码器的概率分布中采用前  $k$  个词, 并始终赋予前  $k$  个序列最高的概率值。由于序列的解码是迭代的, 在  $t$  时刻用于计算输出结果的 key 和 value 值在下一个迭代中被重用。将与人类的判断力接近的 CIDEr-D<sup>[19]</sup> 分数用作奖励, 因此样本的最终梯度表达式为:

$$\nabla_{\theta} L(\theta) = -\frac{1}{k} \sum_{i=1}^k ((r(\omega^i) - b) \nabla_{\theta} \log p(\omega^i)) \quad (11)$$

$$b = \frac{\sum_i r(\omega^i)}{k} \quad (12)$$

其中,  $\omega^i$  是 beam 中的第  $i$  个句子,  $r(\cdot)$  奖励函数,  $b$  为 baseline, 为得到采样序列获得的奖励的平均值。在预测时, 使用 beam search 再次进行解码, 并在最后一个 beam 中保持序列中最高预测概率。

## 2 实验过程

### 2.1 实验数据集以及训练环境

本文使用包含丰富物体种类、物体属性和区域性描述的 VG 数据集来训练物体检测器、属性分类器和关系检测器, 取每个物体得分最高的前 3 种属性, 预处理后得到 305 个对象、103 个属性特征和 64 种物体关系。为了评估本文算法在图像描述生成任务中的性能, 选择目前常用的 MSCOCO (Microsoft Common Object in Context)<sup>[13]</sup> 数据集和 Flickr30k<sup>[20]</sup> 数据集作为实验的验证集。采用文献 [21] 方法对数据集分割, 分

别选择 5000 张和 1000 张图片用于评估模型性能。

本实验使用深度学习框架 Pytorch 对模型进行训练和测试, 在 Ubuntu 16.04 64 bit 系统上完成, 硬件配置为: NVIDIA GeForce GTX 1080 显卡 (8 GB 显存)。

### 2.2 评价指标

为了评估模型生成的图像描述句子的质量, 采用基于规则的自动化评估指标, 主流的指标包括 BLEU<sup>[22]</sup>、METEOR<sup>[23]</sup>、CIDEr、ROUGE<sup>[24]</sup> 等。其中 BLEU 计算参考句子和模型生成的描述之间的 N-gram 相似度, 式 (13) 为 BLEU 的加权集合平均:

$$\text{BLEU}_N(c, s) = b(c, s) \exp\left(\sum_{n=1}^N \omega_n \log \text{cp}_n(c, s)\right) \quad (13)$$

METEOR 是将生成句子与参考句子进行广义单词组匹配, 进而计算得分, 与仅基于精度的 BLEU 指标相比, 该指标还考虑了召回率, 并且与人工评估具有更高的相关性。CIDEr 指标基于人类共识, 采用  $n$  元语法通过词频和倒排句子频率来衡量机器生成句子与参考句子的相似性, 相较于其他方法, 更符合人类评价标准。

### 2.3 实验主要参数设置

对于物体检测器, 使用预训练的 Fast R-CNN 模型以及 ResNet-101 框架, 获得 2048 维图像特征向量。为了构造视觉单元的几何图形, 本文认为如果对象框满足 2 个条件, 则 2 个对象具有相互作用, 即  $r_2 < 0.2$  且  $r_4 < 0.5$ , 其中  $r_2$  和  $r_4$  是 Eon 中的 IoU 和相对距离。本实验中, Transformer 输入时采用正弦位置编码来表示序列内的单词位置, 将每一层的维数设置为 512, 多头注意力机制的数量设置为 6, 存储向量的长度设为 40, 在每个注意力和前馈层之后, 令 dropout 为 0.1。超参数设置如表 1 所示。选用 Adam 优化器<sup>[25]</sup> 在交叉熵损失下训练模型, 首先对学习率进行初始化, 设置初始学习率为  $5 \times 10^{-4}$ , 终止学习率为 0, 训练 5 个 epoch 后通过优化 CIDEr-D 奖励的方式, 使用强化学习进行微调, 为防止模型陷入局部最优, 每迭代 3 个 epoch 使用指数衰减对学习率进行调整, 衰减系数为 0.8。设置网络可接收数据量 batch size 为 50, 对数据集的标注进行采样时设置波束大小 beam size 为 5。

表 1 超参数设置





名称	值
优化器	Adam
初始学习率	0.0005
终止学习率	0
批量大小	50
波束大小	5
Transformer 的隐藏单元	512
多头注意力机制数量	6

### 3 实验结果对比

#### 3.1 模型训练过程及产生结果

使用 ResNet-101 和 LSTM 对模型进行训练,将其作为基准模型(baseline),在 baseline 的基础上,加入场景标志位,在每个视觉单元中实现空间单元和语义单元的多场景注意力计算,使用 Transformer 替代 baseline 的 LSTM 模型,并将得到的场景区域编码作为先验知识与 Transformer 的 encoder 部分融合。在 MSCOCO 数据集上的部分实验结果如表 2 所示。GT、baseline、本文模型分别代表 MSCOCO 数据集中人工标注的真实句子、基准模型生成的句子和本文模型生成的句子。从表 2 可以看出,本文模型生成的句子具备丰富的空间和语义含义,较好地描述了图像内容。与基准模型生成的句子相比,本文方法能够融合更细致的细节和对象关系,从而生成在多场景下更准确和更具描述性的标题,减少了非重点区域的物体关系,使句子更贴合图像主题。

表 2 本文模型与 GT、baseline 模型对比

序号	图像	模型	图像描述
a		GT	a man on a sail board rides through the water
		baseline	a man is holding a surfboard in the water
		本文模型	a handsome man is surfing with a professional surfboard
b		GT	a baseball player that is standing up at a plate
		baseline	a baseball player is swinging at a baseball
		本文模型	lots of spectators are watching the game between baseball players
c		GT	a orange cat with green eyes and long whiskers
		baseline	a cat is sitting on a chair with a blue hat
		本文模型	a orange cat stares at something in the distance
d		GT	an animal that is in the snow by themselves
		baseline	a polar bear is standing in a zoo
		本文模型	a large white polar bear is walking slowly in its territory

例如,表 2(a)中,baseline 的句子翻译为“男人拿着冲浪板”,虽然描述出了图像中的主要物体,但是为一个病句,而本文方法生成了有向的场景信息:“男人正在冲浪”,既构建了场景图,又得到了正确的主宾关系,并且生成了男人“帅气”的属性特征,句子流畅更贴近 GT;在表 2(b)中,本文模型能够准确识

别出观众和棒球运动员这 2 个主要对象及其关系,baseline 生成的句子中,“游泳”不符合图像信息。表 2(c)的图中信息较少,本文模型聚焦于对象所在的场景,关注图中猫的属性,生成了“盯”这一符合场景语义的单词,体现了模型对场景的适当想象,而 baseline 模型捕捉图像信息的能力不佳,生成的描述中出现了图中没有的“椅子”“帽子”,物体识别错误,对图像过度想象。表 2(d)生成了融合多场景下语义关系的“领地”属性,增加了句子的生动性和趣味性。由此表明,本文模型在图像多场景描述上,对空间关系和语义关系的处理能力较好,使句子生动、准确,更贴近人类描述。

#### 3.2 实验结果客观指标对比

为了使实验结果有说服力,本文分别在 MSCOCO 数据集和 Flickr30k 数据集上与近些年带有注意力机制的图像描述算法进行评价指标比较。表 3 和表 4 中 B-1、B-4 分别表示 BLEU 的参数  $N$  取值为 1、4 时对应的值,表 3 为本文模型与其他模型在 MSCOCO 验证集上的性能对比,CIDEr 得分达到 1.210,超过 baseline 方法 5.68%,B-1 指标可达到 0.806。表 4 为本文模型与其他模型在 Flickr30k 数据集上的性能对比,CIDEr 得分达到 0.615,超过 baseline 方法 4.41%。通过对比发现,在相同数据集和相同训练条件下,本文算法在常用的评估指标上的得分高于基于 CNN + LSTM 算法的图像描述方法。

表 3 模型在 MSCOCO 数据集上在线测试的性能对比

方法	B-1	B-4	METEOR	ROUGE-L	CIDEr
文献[26]的 Up-Down(baseline)	0.798	0.361	0.266	0.566	1.145
文献[27]的 CAVP	0.801	<b>0.385</b>	0.274	0.570	1.173
文献[28]的 GCN-LSTM	0.804	0.380	0.285	0.567	1.180
本文算法	<b>0.806</b>	0.382	<b>0.286</b>	<b>0.572</b>	<b>1.210</b>

表 4 模型在 Flickr30k 数据集上在线测试的性能对比

方法	B-1	B-4	METEOR	ROUGE-L	CIDEr
文献[26]的 Up-Down(baseline)	0.663	0.219	0.205	0.451	0.589
文献[27]的 CAVP	0.646	0.221	0.209	0.473	0.539
文献[28]的 GCN-LSTM	0.667	0.233	0.213	0.488	0.604
本文算法	<b>0.692</b>	<b>0.240</b>	<b>0.220</b>	<b>0.504</b>	<b>0.615</b>

### 4 结束语

本文提出了一种全新的多场景注意力机制,结合命名实体识别和 Transformer 结构,优化模型在多场景下对空间信息和语义信息的获取能力。采用 ResNet-101 对图像进行编码得到深度特征,采用多层感知层、关系检测器和基于命名实体识别的多场景标志

位得到贴合图像场景的视觉单元和先验知识。在语言模型部分,改进了 Transformer 结构,融合多场景分类的结果,采用细粒度区域编码的方法,通过存储编码向量,更新多场景的先验知识。实验结果表明,本文方法提高了模型对场景的理解能力,更贴合人类语言,在生成的结果和评价指标上优于同类型的其他模型。

#### 参考文献:

- [1] 陈龙杰,张钰,张玉梅,等. 基于多注意力多尺度特征融合的图像描述生成算法[J]. 计算机应用, 2019, 39(2): 354-359.
- [2] 张姣,杨振宇. 图像描述生成方法研究文献综述[J]. 智能计算机与应用, 2019(5): 45-49.
- [3] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.
- [4] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.
- [5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [6] MAO J H, XU W. Explain images with multimodal recurrent neural networks[J]. Computer Science, 2014, arXiv: 1410. 1090.
- [7] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// International Conference on Machine Learning. 2015: 2048-2057.
- [8] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: Transforming objects into words[C]// Advances in Neural Information Processing Systems. 2019: 11137-11147.
- [9] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision & Pattern Recognition. 2016: 770-778.
- [10] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.
- [11] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[J]. Machine Learning, 2016, arXiv: 1609. 02907.
- [12] BASTINGS J, TITOV I, AZIZ W, et al. Graph convolutional encoders for syntax-aware neural machine translation[J]. Computation and Language, 2017, arXiv: 1704. 04675.
- [13] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 3104-3112.
- [14] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. Computation and Language, 2014, arXiv: 1409. 0473.
- [15] ZELLERS R, YATSKAR M, THOMSON S, et al. Neural motifs: Scene graph parsing with global context[J]. Computer Vision and Pattern Recognition, 2017, arXiv: 1711. 06640.
- [16] 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(1): 32-73.
- [17] SHANG J B, LIU L Y, GU X T, et al. Learning named entity tagger using domain-specific dictionary[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2054-2064.
- [18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 5998-6008.
- [19] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: Consensus-based image description evaluation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4566-4575.
- [20] PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2641-2649.
- [21] KARPATY A, FEI-FEI L. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.
- [22] 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. 2002: 311-318.
- [23] SATANJEEV B. 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. 2005: 228-231.
- [24] LIN C Y, HOVY E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]// Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003: 150-157.
- [25] KINGMA D, BA J. Adam: A method for stochastic optimization[J]. Machine Learning, 2014, arXiv: 1412. 6980.
- [26] 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. 2018: 6077-6086.
- [27] ZHA Z J, LIU D Q, ZHANG H W, et al. Context-aware visual policy network for sequence-level image captioning[C]// Proceedings of the 26th ACM International Conference on Multimedia. 2018: 1416-1424.
- [28] YAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning[C]// Proceedings of the European Conference on Computer Vision. 2018: 684-699.