



# 基于 Transformer 的多方面特征编码图像描述生成算法

衡红军, 范昱辰, 王家亮

(中国民航大学, 计算机科学与技术学院, 天津 300300)

**摘 要:** 由目标检测算法提取的目标特征在图像描述生成任务中发挥着重要作用, 但仅使用对图像进行目标检测的特征作为图像描述任务的输入会导致除关键目标信息以外的其余信息获取缺失, 且生成的文本描述对图像内目标之间的关系缺乏准确表达。为了解决上述不足, 提出了用于编码图像内目标特征的目标 Transformer 编码器, 以及用于编码图像内关系特征的转换窗口 Transformer 编码器, 从不同角度对图像内不同方面的信息进行联合编码。通过拼接方法将目标 Transformer 编码的目标特征与转换窗口 Transformer 编码的关系特征相融合, 以达到图像内部关系特征和局部目标特征融合的目的, 最终使用 Transformer 解码器将融合后的编码特征解码生成对应的图像描述。通过在 MS-COCO 数据集上的广泛实验, 以及与当前经典模型算法的比较, 结果表明, 所提出的算法模型性能明显优于基线模型, 实验结果表明在 BLEU-4、METEOR、ROUGE-L、CIDEr 得分可达到 38.6%, 28.7%, 58.2% 和 127.4%, 优于传统图像描述网络模型, 能够生成更详细更准确的图像描述。

**关键词:** 图像描述; 转换窗口; 多头注意力机制; 多模态; Transformer

开放科学(资源服务)标志码(OSID):



## Multifaceted feature coding image caption algorithm based on Transformer

Heng Hongjun, Fan Yuchen, Wang Jialiang

(School of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

**【Abstract】** Object features extracted by object detection algorithm play an ever-growing critical role in the generation of image caption. However, only using the features of object detection as the input of image caption task will lead to the loss of other information except the key object information and generate a caption that lacks of accurate expression of the relationship between the image object. To solve the disadvantages mentioned above, an Object Transformer encoder for encoding object features in image and a Shift Window Transformer for encoding relational features in image are proposed to make joint efforts to encode different aspects of information in image from different angles. The object features of the Object Transformer encoder are fused with the relational features of the Shift Window Transformer by splicing method, so as to achieve the purpose of fusion of the internal relational features and local object features. Finally, a Transformer decoder will be utilized to decode the fused coding features and generate the corresponding image caption. Extensive experiments on MS-COCO data set and comparison with the current classical model algorithm show that the performance of the proposed algorithm is significantly better than the baseline model. Experimental results indicate that the scores of BLEU-4, METEOR, Rouge-L and CIDEr can reach 38.6%, 28.7%, 58.2% and 127.4% respectively, better than the traditional image caption algorithm. Besides, it can also generate more detailed and accurate caption.

**【Key words】** image caption; shift window; multi-headed attention mechanism; multimodal; Transformer

DOI: 10.19678/j.issn.1000-3428.0064450

### 0 概述

图像描述是将图像的视觉内容转换为符合人类描述习惯的自然语言语句的任务, 是一项结合计算机视觉和自然语言处理的多模态任务。图像描述的

挑战不仅存在于识别图像中的目标与目标之间的关系, 而且还存在于不同模态下实现相同语义的转换以及生成人类描述习惯的句子。

现有的图像描述生成方法有基于模板的方法

**基金项目:** 国家自然科学基金资助项目 (U1333109)

**作者简介:** 衡红军, 1968-, 男, 博士学位, 副教授; 范昱辰 (通信作者), 1996-, 男, 硕士研究生, 图像描述主研方向; 王家亮, 1983-, 男, 博士学位, 讲师。E-mail: 2019052041@cauc.edu.cn

<sup>[1][2]</sup>、基于检索的方法<sup>[3]</sup>和基于编码-解码的方法。目前主流图像描述方法倾向于采用基于神经网络的编码器-解码器结构<sup>[4-7]</sup>。早期的图像描述的编码器-解码器结构使用卷积神经网络(Convolutional Neural Networks, CNN)作为编码器对输入图像进行编码,使用循环神经网络(Recurrent Neural Networks, RNN)作为解码器对编码器产生的结果进行解码。这些方法模型都由一个图像  $I$  作为模型的输入,每个时间戳产生的单词的概率  $P(S|I)$  作为模型的输出,最终生成的句子  $S = \{W_1, \dots, W_n\}$  为图像描述语句。

现有的图像描述模型多采用原始图像或对原始图像进行目标检测得到的目标特征向量作为模型输入,这两种方案均致力于更加准确的描述图像内的关键目标,但却造成了对图像内部其余信息(图像背景信息、目标之间的关系信息等)的获取缺失,导致生成的图像描述存在误差和局限性。

为了兼顾准确描述图像内部目标的同时对图像内部目标之间的关系进行合理表达,本文提出一种目标 Transformer 和转换窗口 Transformer 的联合编码模型。对于给定图像,采用本文提出的目标 Transformer 编码器编码目标视觉特征,同时使用转换窗口 Transformer 编码器编码图像内部关系特征。本文采用拼接方法将编码后的视觉特征与编码后的图像内部关系特征进行融合。将融合后的编码向量使用 Transformer 解码器解码,最终生成对应图像内容的描述。通过实验表明,相比基线模型,本文所提出的模型在各项评价指标上得到了更高的评分,在准确表述图像内目标的同时能够更加合理表达图像内目标之间的关系。

## 1 相关工作

2014 年,谷歌提出了 Neural Image Caption Generator<sup>[5]</sup>,这是一个首先使用 CNN 作为编码器,RNN 作为解码器的神经网络模型,展现出了良好的性能。随着研究的深入,研究人员发现人类观察图像中的内容时,会从复杂的图像内容中找出关键点,并将注意力集中于此,因此研究人员基于人类注意力机制启发,设计了加入视觉注意力机制的神经网络模型<sup>[8]</sup>用于图像描述。注意力的加入使模型可以选择性地关注图像的特定区域,而不是无偏好地关注整个图像。Jiasen 等<sup>[9]</sup>注意到在生成描述过程中并非每个单词均来源于图像,也有可能来源于已生成的描述本身(如一些介词、连词的生成),因此设计了自适应注意力(adaptive attention),让模型自行选择应关注于图像还是描述语句。随着目标检测技术精度的提升,Anderson 等<sup>[10]</sup>提出了一种目标检测引导的注意力机制,它被证明可以提高图像描述的准确率。

综上所述,图像描述任务的研究由刚开始对图像的无偏关注,到加入注意力机制的辅助,再到目标检测方法的加入,研究者一直致力于对图像内目标内容的精确识别。但对于图像描述任务,不仅仅需要准确描述目标,更需要对目标之间的互动关系进行准确表达,如果目标之间的互动关系表达错误,则会造成描述与图像内容严重不符。

2017 年,谷歌提出了 Transformer 模型<sup>[11]</sup>,用于解决 Seq2Seq(Sequence to Sequence)问题。Transformer 模型也遵循编码器-解码器架构,但编码器和解码器没有使用卷积、池化等网络架构,而是

完全依靠自注意机制的并行化架构来捕捉序列依赖。此后, Transformer 在自然语言处理(Neural Language Processing, NLP) 任务中取得了优异的成绩。但 Transformer 在计算机视觉领域的表现却不尽如人意。研究者一度认为 Transformer 模型并不适用于计算机视觉任务, 直至 ViT(Vision Transformer)<sup>[12]</sup> 模型的出现, 才使研究人员将眼光重新聚焦于 Transformer 相关模型, 经过长期实践证明 Transformer 在计算机视觉领域也能取得比传统 CNN 模型更强的性能。2021 年微软亚洲研究院提出了 Swin Transformer<sup>[13]</sup>, 其结果比 ViT 更好, 并明显优于 CNN 模型, 进一步提升了 Transformer 在计算机视觉领域的应用。通过实验研究发现, Swin Transformer 不仅在图像分类任务中表现出色, 而且在计算图像内部的关系方面也有良好的效果。

得益于 Transformer 近几年在自然语言处理领域和计算机视觉领域的突出表现, 本文借鉴 Swin Transformer 和基于 Encoder-Decoder 框架的 ViT 的原理, 放弃了传统的 CNN 和 RNN 模型, 使用与 Transformer 相关的多头注意力机制来处理图像特征并生成与图像对应的描述。架构的原理如图 1 所示, 并有如下特点:

- 1)使用目标 Transformer 对目标检测得到的局部目标特征进行编码;
- 2)使用转换窗口 Transformer 对整张图像内容进行编码, 用于编码图像内部潜在的关系信息;
- 3)在解码过程中, 使用 Transformer 解码器代替传统的 RNN 解码器。

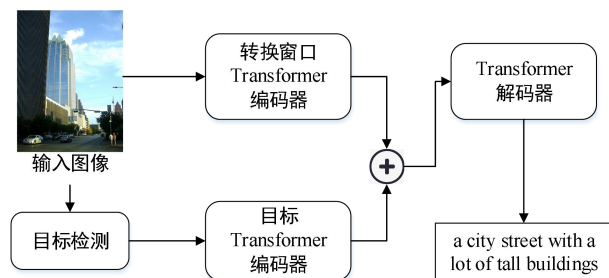


图1 算法模型简要结构图

Fig.1 Brief structure diagram of algorithm model

## 2 多方面特征编码

为了提高图像描述的准确性, 本文从融合不同方面的特征表示的角度出发, 重新设计了网络结构, 如图 2 所示。1.1 节介绍了目标 Transformer 编码器对目标特征进行编码的方法; 1.2 节介绍了转换窗口 Transformer 对图像内部关系特征进行编码的方法, 1.3 节介绍了特征融合以及 Transformer 解码器的解码方法。

### 2.1 目标 Transformer 编码器

首先使用 Faster R-CNN<sup>[14]</sup>对图像  $I$  进行检测得到图像的  $k$  个区域特征  $\{r_1, \dots, r_k\}$ , 每个图像特征向量首先通过一个嵌入层进行处理, 该层通过一个全连接层将特征向量的尺寸从 2048 维降至 512 维, 然后通过一个 ReLu 激活函数和 Dropout 层处理后生成的向量作为目标 Transformer 编码器的输入向量。

目标 Transformer 编码器总共有 6 层, 每一层由一个多头注意力层和一个前馈神经网络组成。集合  $\{x_1, \dots, x_n\}$  为经过目标检测并嵌入后的  $N$  个目标特征向量的集合, 则  $x_n$  表示为经过目标检测并嵌入得到的第  $n$  个目标所对应的特征向量。所有经过目标检测并嵌入得到的特征向量所拼接成的矩阵作为第一个编码层的输入, 第 2-6 个编码层均使用前一层编码层的输出作为输入。对编码层中的每个多头注意

力层，每一层中“头”的数量设为 8，为  $N$  个特征向量分别计算查询向量  $Q_o$ 、键向量  $K_o$  和值向量  $V_o$  的

计算方法如公式 1 所示：

$$Q_o = XW_{Q_o}, K_o = XW_{K_o}, V_o = XW_{V_o} \quad (1)$$

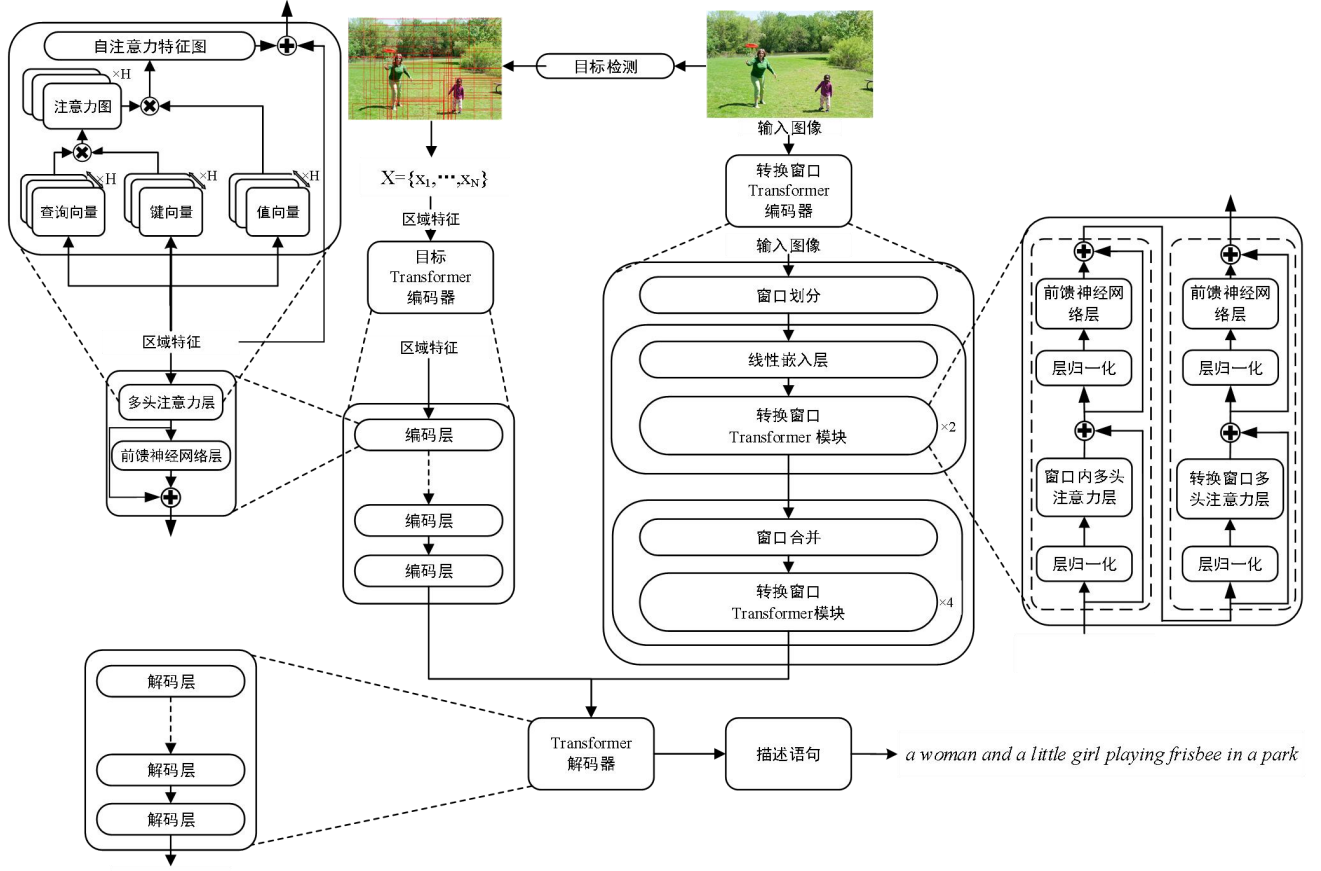


图 2 算法模型详细结构图

Fig.2 Detailed structure diagram of algorithm model

其中  $X$  为包含所有的输入向量  $\{x_1, \dots, x_N\}$  所拼接成的矩阵， $W_{Q_o}, W_{K_o}, W_{V_o}$  为可学习的权重矩阵。

不同的两个目标区域之间的相关性分数计算方法如公式 2 所示：

$$\Omega_o = \frac{Q_o K_o^T}{\sqrt{d_k}} \quad (2)$$

其中  $\Omega_o$  为形状为  $N \times N$  的权重矩阵，其中的元素  $\omega_{mn}$  表示为第  $m$  个特征区域和第  $n$  个特征区域之间的相关性得分。本文对  $d_k$  的设定与文献[15]中相同，设为 64，代表查询向量，键向量和值向量的维度。

多头注意力的计算方法如公式 3 所示：

$$\begin{aligned} H(X) &= \text{Attention}(Q_o, K_o, V_o) \\ &= \text{soft max}(\Omega_o) V_o \end{aligned} \quad (3)$$

由于本节将多头注意力中“头”的数量设置为 8，因此，需要通过公式 1、公式 2 和公式 3 重复计算 8 次来分别表示 8 个“头”。计算完成后，将各个“头”矩阵拼接后与可学习的参数矩阵  $W_o$  相乘，多头注意力计算方法如公式 4 所示：

$$\text{MHAttention}(Q_o, K_o, V_o) = \text{Concat}(H_1, \dots, H_8) W_o \quad (4)$$

残差结构和层归一化均被应用在多头注意力层和前馈神经网络层中，如公式 5、公式 6、公式 7 所示：

$$X = \text{LayerNorm}(X + \text{MHAttention}(Q_o, K_o, V_o)) \quad (5)$$

$$\text{FFN}(X) = \max(0, XW_1 + b_1)W_2 + b_2 \quad (6)$$

$$X = \text{LayerNorm}(X + \text{FFN}(X)) \quad (7)$$

其中公式 5 中的参数  $X$  为当前层的输入数据  $X$ ，最终得到的  $X$  作为当前编码层的输出。公式 6 和公式 7 表示为将多头注意力层的输出  $X$  输入至前馈神经网络进行计算的计算方法，其中  $W_1, W_2$  和  $b_1, b_2$  分别为可学习的权重和偏置量。

## 2.2 转换窗口 Transformer 编码器

由于 Swin-Transformer<sup>[13]</sup>在目标检测及语义分割任务中均有出色的表现，其中 Shift Window 操作可以实现不同窗口内信息的交互。因此，基于 Swin Transformer 的 Shift Window 思想，设计了转换窗口 Transformer 编码器。

如图 2 所示，在转换窗口 Transformer 编码器中，每个转换窗口 Transformer 模块中含有两个子模块分别为窗口多头注意力模块和转换窗口多头注意力模块，与目标 Transformer 的设置相同，为减小训练误差并消除奇异样本数据，残差结构和层归一化均被应用与多头注意力模块和转换窗口多头注意力模块。

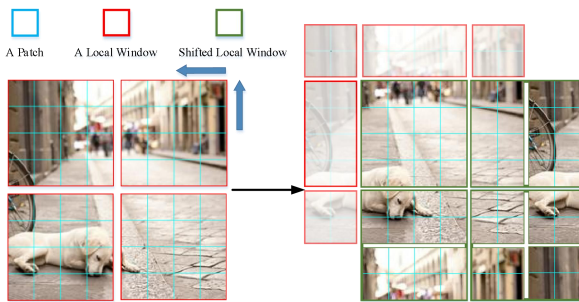


图 3 转换窗口方法示意图

Fig.3 Shift window method schematic

为了提高计算效率，本文以不重叠的方式将输入图像均匀地分割为多个窗口，只在局部窗口内计算自注意力。这样就导致了窗口之间缺乏信息交互，

因此本文将 Swin Transformer 模型中的 Shift Window 的思想引入转换窗口 Transformer 编码器。

如图 3 所示，基于窗口的多头注意力模块采用正常的窗口划分策略，将一个大小为  $8 \times 8$  的图像均匀的分为了  $2 \times 2$  个大小为  $4 \times 4$  ( $M=4$ ) 的窗口。为了实现窗口之间信息的交互，本文使用了转换窗口的方法，将  $\left\lfloor \frac{M}{2} \right\rfloor, \left\lfloor \frac{M}{2} \right\rfloor$  像素从规则划分的窗口中循环替换，实现窗口间内容的交互。在这种转换之后，一个局部窗口内可能有图像中的图连续像素块组成，因此采用了遮盖机制，将自注意力的计算限制在每个子窗口内。

在转换窗口 Transformer 中，以图像矩阵作为输入，首先通过图像分割层进行处理，窗口集合  $\{y_1, \dots, y_M\}$  为输入图像中均匀划分的  $M$  个子区域而构成的集合， $y_m$  则代表第  $m$  个划分的子区域对应的特征向量。为  $M$  个窗口子区域分别计算查询向量  $Q_{sw}$ ，键向量  $K_{sw}$  和值向量  $V_{sw}$  的计算方法如公式 8 所示：

$$Q_{sw} = YW_{Q_{sw}}, K_{sw} = YW_{K_{sw}}, V_{sw} = YW_{V_{sw}} \quad (8)$$

其中  $Y$  为包含所有的输入窗口子区域特征向量  $\{y_1, \dots, y_M\}$  所拼接成的矩阵， $W_{Q_{sw}}, W_{K_{sw}}, W_{V_{sw}}$  为可学习的权重矩阵。

两个窗口子区域之间的相关性分数计算方法如公式 9 所示：

$$\Omega_{sw} = \frac{Q_{sw} K_{sw}^T}{\sqrt{d}} \quad (9)$$

其中  $\Omega_{sw}$  是一个形状为  $M \times M$  的权重矩阵，其中的元素  $\omega_{mn}$  表示第  $m$  个窗口子区域和第  $n$  个窗口子区域之间的关系得分。 $d$  的值为查询向量与键向量之间的维度比，表示为  $\dim(Q_{sw}) / \dim(K_{sw})$ 。



计算自注意力的方法与目标 Transformer 不同，其方法公式 10 所示：

$$\begin{aligned} H(Y) &= SWAttention(Q_{sw}, K_{sw}, V_{sw}) \\ &= \text{soft max}(\Omega_{sw} + B)V_{sw} \end{aligned} \quad (10)$$

其中参数  $B$  的含义为窗口子区域之间的相对位置偏置量，本文中对  $B$  的设定与文献[13]中相同，存在一个偏差矩阵  $\hat{B} \in \mathbb{R}^{(2M-1) \times (2M-1)}$ ， $B$  的值取自  $\hat{B}$ 。

如图 2 所示，转换窗口 Transformer 存在两个子转换窗口 Transformer 模块，本文将两个子转换窗口 Transformer 模块的“头”的数量分别设定为 6 和 12 并进行计算，多头注意力计算方法如公式 11 所示：

$$\begin{aligned} SWMHAttention(Q_{sw}, K_{sw}, V_{sw}) \\ = \text{concat}(H_1, \dots, H_N)W_{sw} \end{aligned} \quad (11)$$

其中  $N$  为“头”的数量， $W_{sw}$  为可学习的权重矩阵。

转换窗口 Transformer 也使用了残差结构和层归一化，其方法与 2.1 节目标 Transformer 所介绍的方法相同，因此不再赘述。

### 2.3 Transformer 解码器

对于目标 Transformer 编码器编码的目标特征向量  $X$  和转换窗口 Transformer 编码器编码的关系特征向量  $Y$ ，本文采用向量拼接的方式将两个特征向量进行融合，如公式 12 所示：

$$F = \text{Concat}(X, Y) \quad (12)$$

如图 4 所示解码器结构，编码结果  $F$  作为解码器的一部分输入用于计算解码器中的键向量  $K_D$  和值向量  $V_D$ ，计算方法如公式 13 所示：

$$K_D = FW_{K_D}, V_D = FW_{V_D} \quad (13)$$

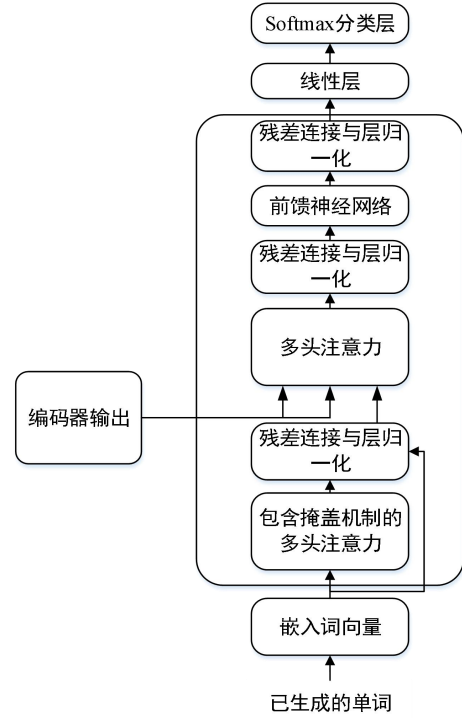


图 4 Transformer 解码器结构图

Fig.4 Transformer decoder structure diagram

其中  $W_{K_D}, W_{V_D}$  为可学习的权重矩阵，而查询向量  $Q_D$  需要将之前时间戳生成的单词经过嵌入后计算多头注意力得到，而后将得到的解码器查询向量  $Q_D$ ，键向量  $K_D$  和值向量  $V_D$  计算多头注意力后送入前馈神经网络产生输出，其计算多头注意力的方法与 2.1 节中目标 Transformer 的多头注意力计算方法完全相同，因此不再赘述。值得注意的是，解码器在训练过程中对输入单词采用遮盖方法计算多头注意力，这是因为使用 Ground Truth 中存在未来信息，而在实际生成文本描述语句过程中是无法预知的，因此使用遮盖机制保证训练与测试过程的一致性。

对于解码器的输出，经过一个线性层扩展至词汇总表长度后输入 Softmax 分类层进行分类得到当前时间戳的输出单词，计算方法如 14 所示：

$$W_{new} = \text{Soft max}\left(\text{Linear}\left(W_{output}\right)\right) \quad (14)$$

其中 $W_{output}$ 表示为解码器解码结果, $W_{new}$ 为当前时间戳生成的单词。接下来会一直重复解码过程,直至解码结果与单词表中结束符一致,代表该模型对当前图像的文本描述语句生成完毕。

### 3 实验结果与分析

#### 3.1 数据集与实验环境

为了评估本文提出方法的有效性,本文采用的数据集为 MSCOCO 2014(Common Objects in COntext 2014)<sup>[16]</sup>数据集。MSCOCO 数据集可以用于图像分类、目标检测、语义分割、图像描述等任务。数据集中包含了图像包括 91 类目标,328,000 余张图像和 2,500,000 余个标签。本文采用 Karpathy 等在文献[6]中对数据集的划分方法将数据集分为训练集、验证集和测试集,其中包含 11300 余张训练图像,5000 张验证图像和 5000 张测试图像,其中每张图像对应 5 句英文描述性语句。

实验环境使用 Ubuntu 18.04 64 位系统,采用 Pytorch 深度学习框架进行训练和测试,硬件配置为 Intel i9-9900k CPU, Nvidia RTX 2080TI 显卡(11G 显存)。

#### 3.2 评价指标

为了对本文提出算法模型的有效性和先进性做出合理评估,实验采用广泛被应用于图像描述的几个客观量化评分方法,其中包括 BLEU(Bilingual Evaluation Understudy)<sup>[17]</sup>、CIDEr(Consensus-based Image Description Evaluation)<sup>[18]</sup>、METEOR(Metric for Evaluation of Translation with Explicit ORdering)<sup>[19]</sup>、ROUGE-L(Longes common subsequence based Recall-Oriented Understudy for

Gisting Evaluation)<sup>[20]</sup>。

#### 3.3 模型主要参数设置

本实验中,首先对图像数据进行预处理,按照 RGB 格式读取图片,将图片调整大小为  $224 \times 224$  像素,使用 Imagenet<sup>[21]</sup>上预训练的 ResNet-101<sup>[22]</sup>作为基础的 CNN 进行图像的特征提取,使用 Faster R-CNN<sup>[23]</sup>进行目标检测。使用 ResNet-101 的中间特征作为 Faster R-CNN 的输入,RPN(Region Proposal Network)为识别的目标生成边界框,使用非最大抑制法丢弃 IoU(intersection-over-union)超过阈值 0.7 的重叠边界框,然后使用 Rol(region-of interest)池化层将所有的边界框特征向量转换为相同维度,剩余的 CNN 层被用于预测标签和细化每个边界框,最终将所有预测概率值低于阈值 0.2 的边界框丢弃,使用平均池化的方法为剩余的每一个边界框生成一个 2048 维的向量作为目标 Transformer 编码器的输入。将调整大小后的图像作为转换窗口 Transformer 编码器的输入,并将转换窗口 Transformer 中划分窗口的长宽值大小设定为 4 个像素。

本实验将语料库规模设为出现频次超过 5 次的单词并将语料库中的单词进行独热(one-hot)编码。分批处理图像时,单次输入图像 batch size 数量设为 10。使用 Dropout 舍弃单元来提高模型在数据集上的泛化能力,并将 Dropout 值设为 0.1。在训练中使用了集束搜索的方法,将 beam 的值设为 3。在模型训练过程中使用了交叉熵损失和文献[24]中提出的 CIDEr-D 优化强化学习方法。定义训练轮次数为 50 轮,前 30 轮使用交叉熵损失进行训练,后 20 轮使用 CIDEr-D 优化强化学习方法进行训练。本文使用

Pytorch 自带的 Adma(Adaptive Moment Estimation) 网络优化算法, 其中将  $\beta_1$  和  $\beta_2$  的值分别设置为 0.9 和 0.999。

### 3.4 消融实验

#### 3.4.1 Transformer 结构有效性分析

表 1 不同编码器解码器的消融实验结果对比

Table 1 Comparison of ablation results of different encoders and decoders

算法	BLEU-4	METEOR	ROUGE-L	CIDEr
Up-Down+LSTM	32.9	26.5	55.6	105.6
ObjTrans+Trans	32.6	27.3	55.4	111.2
ObjTrans+ViT+Trans	32.4	26.5	54.8	105.6
ObjTrans+SWTrans+Trans	34.4	27.6	56.2	122.2
ObjTrans+SWTrans+Trans(beam=3)	35.7	27.8	56.3	124.2

为了验证本文采用的 Transformer 结构相较于 CNN、RNN 相关结构的先进性, 将本文方法与经典的 Up-Down 算法<sup>[10]</sup>进行比较。使用控制变量的思想设计以下消融实验: (1)将编码器替换为目标 Transformer; (2)目标 Transformer 与 ViT 的组合和目标 Transformer 和转换窗口 Transformer 的组合, 将 LSTM 解码器替换为 Transformer 解码器; (3)本文方法, 即使用目标 Transformer 以及转换窗口 Transformer 联合编码结构; (4)在本文方法基础上使用 beam size 为 3 的波束搜索。在相同数据集, 相同的训练条件下, 使用交叉熵损失对模型训练了 30 轮。结果如表 1 所示, 可以得出, 将编码器和解码器分别替换为 Transformer 结构之后其性能在各项指标上均有提升。

#### 3.4.2 转换窗口 Transformer 有效性分析

为了验证转换窗口 Transformer 提取关系信息的有效性, 本文在实验中使用无位置编码的方法和按

照目标边界框由大到小进行位置编码的方法与转换 Transformer 编码器进行比较。在相同数据集, 相同的训练条件下, 使用交叉熵损失对模型训练了 30 轮, 结果如表 2 所示。

表 2 不同位置嵌入方式与转换窗口 Transformer 编码方式消融实验结果对比

Table 2 Comparison of ablation results between different embedding methods and shift window Transformer encoding method

方法	CIDEr
无位置嵌入	109.6
按目标边界框大小嵌入	108.9
转换窗口 Transformer	112.4

结果表明, 通过转换窗口 Transformer 获取全局特征的方法, 最终的到评价指标 CIDEr 的值明显高于无位置编码和按边界框由大到小进行编码方法所得到的 CIDEr 值。

### 3.5 实验结果对比与分析

#### 3.5.1 定量分析

本文算法与 Google NIC(Google Neural Image Caption)<sup>[5]</sup>、Soft-Atten<sup>[8]</sup>、Hard-Atten<sup>[8]</sup>、Deep VS(Deep Visual-Semantic alignments)<sup>[6]</sup>、AFAR(attention feature adaptive recalibration)<sup>[26]</sup>、MSM(Multimodal Similarity Model)<sup>[25]</sup>、MMFFN(Multi-attention and Multi-scale Feature Fusion)<sup>[27]</sup>、ASIA (Attention-guided image captioning)<sup>[28]</sup>、GO-AMN (Gated Object-Attribute Matching Network)<sup>[29]</sup>算法对比。

结果如表 3 所示, 本文算法的 CIDEr 值可以达到 127.4%, BLUE-4 的得分可以达到 38.6%, METEOR 的值可以达到 28.7%, ROUGE-L 的值可以达到 58.2%。在相同的数据集、相同的训练条件下, 本文算法的性能指标得分最高。



表3 不同图像描述算法的实验结果

Table 3 Experimental results of different image description algorithms

算法	BLEU-4	METEOR	ROUGE-L	CIDEr
GoogleNIC[5]	22.7	23.7	—	85.5
Soft-Atten[8]	24.3	23.9	—	—
Hard-Atten[8]	25.0	23.0	—	—
Adaptive[9]	33.2	26.6	—	108.5
DeepVS[6]	23.0	19.5	—	66.0
AFAR[26]	28.5	23.3	—	83.6
MSM[25]	32.5	25.1	—	98.6
GO-AMN[29]	35.2	27.0	—	109.8
Up-Down[10]	36.3	27.7	56.9	120.1
MMFFN[27]	36.4	27.2	56.9	116.7
ASIA[28]	37.8	27.7	—	116.7
本文算法	<b>38.6</b>	<b>28.7</b>	<b>58.2</b>	<b>127.4</b>

## 3.5.2 定性分析


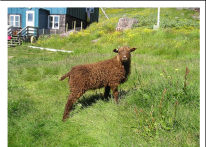


	Ours: a cup of coffee sitting next to a computer keyboard Up-Down: a computer keyboard and a mouse sitting on a desk GT1: a cup of coffee sitting next to a computer keyboard GT2: a desk that has a keyboard and a cup on it GT3: a desk with a cup and a keyboard
	Ours: a brown sheep standing on a lush green hillside Up-Down: a brown sheep standing in a field of grass GT1: a brown sheep standing on top of a lush green field GT2: a sheep standing in a field in front of a house GT3: a chocolate colored sheep standing in the grass
	Ours: a young boy swinging a tennis racket at a tennis ball Up-Down: a young man holding a tennis ball on a court GT1: a person hitting a tennis ball with a tennis racket GT2: a man on a tennis court that has a racquet GT3: a boy hitting a tennis ball on the tennis court
	Ours: a plate of food with eggs and tomatoes on a table Up-Down: a group of food on a plate GT1: a white plate topped with breakfast food and baked beans GT2: a plate of breakfast food including eggs and sausage GT3: a plate of breakfast food with a silver tea pot

图5 算法结果定性对比图

Fig.5 Qualitative comparison of algorithm results

在模型训练完成后,选取测试集中的图像结果与基线模型 Up-Down 模型的实验结果以及数据集中给出的标准描述语句作比较如图 5 所示,可以看出 Up-Down 模型生成的描述和图像内容具有一定的关联性,在逻辑上是正确的,而本文提出的模型得到的结果对于图像细节和图像内目标之间的关系描述更加准确生动。例如,在第三幅图中 Up-Down 模型生成的“holding a tennis ball”内容与图像内的

视觉信息并不一致,而本文模型生成的“swinging a tennis racket at a tennis ball”对图像内的视觉信息的描述更加准确,把图像内目标之间的关系描述的更加生动,则再次证明了本文提出算法在捕捉图像内目标之间关系的有效性。

## 4 结束语

本文设计了基于转换窗口 Transformer 的图像描述生成算法。该算法在使用设计了目标 Transformer 和转换窗口 Transformer 两个编码器,分别对 Faster R-CNN 目标检测提取的图像和整张图像编码后进行特征融合,使用 Transformer 解码器代替传统 RNN 模型。证明了 Transformer 结构在图像描述任务上的有效性。本文算法的图像描述效果以及 BLEU、CIDER、METEOR、ROUGE-L 等评价指标上,相较于基线模型,都取得了较高的得分,其中 BLEU-4 和 CIDEr 得分达到了 38.6%和 127.4%。实验结果表明,本文提出的转换窗口 Transformer 方法提高了模型的图像内部关系识别能力,提升了描述的准确性,提高了模型的泛化能力。下一步工作是利用转换窗口方法的优良性能显式的提取图像的内部关系,明确图像内所含关系的具体信息,进一步提高图像描述模型的内部关系表达能力。

## 参考文献

- [1] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(12): 2891-2903.
- [2] Li S, Kulkarni G, Berg T, et al. Composing simple image descriptions using web-scale n-grams[C]//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. 2011: 220-228.
- [3] Hodosh M, Young P, Hockenmaier J. Framing image

- description as a ranking task: Data, models and evaluation metrics[J]. *Journal of Artificial Intelligence Research*, 2013, 47: 853-899.
- [4] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Explain images with multimodal recurrent neural networks. In *NIPS Workshop on Deep Learning*, 2014.
- [5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.
- [6] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [7] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen. Encode, review, and decode: Reviewer module for caption generation. In *NIPS*, 2016.
- [8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.
- [9] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.
- [10] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. “Bottom-up and top-down attention for image captioning and visual question answering.” In *CVPR*, 2018.
- [11] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. *arXiv*, 2017.
- [12] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. *arXiv preprint arXiv:2010.11929*, 2020.
- [13] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[J]. *arXiv preprint arXiv:2103.14030*, 2021.
- [14] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. *Advances in neural information processing systems*, 2015, 28: 91-99.
- [15] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. *arXiv*, 2017.
- [16] T.-Y. Lin, M. Maire, S. Belongie et al., “Microsoft coco: common objects in context,” in *European Conference on Computer Vision*, pp. 740–755, Springer, Berlin, Germany, 2014.
- [17] 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.
- [18] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015: 4566-4575.
- [19] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language[C]//*Proceedings of the ninth workshop on statistical machine translation*. 2014: 376-380.
- [20] Lin C Y. Rouge: A package for automatic evaluation of summaries[C]//*Text summarization branches out*. 2004: 74-81.
- [21] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//*2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009: 248-255.
- [22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [23] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. *Advances in neural information processing systems*, 2015, 28: 91-99.
- [24] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical Sequence Training for Image Captioning[C]//*IEEE. IEEE*, 2016.
- [25] Yao T, Pan Y, Li Y, et al. Boosting Image Captioning with Attributes[C]//*IEEE International Conference on Computer Vision*. IEEE Computer Society, 2017:4904-4912.
- [26] 韦人予, 蒙祖强. 基于注意力特征自适应校正的图像描述模型[J]. *计算机应用*, 2020(S01):45-50.
- [27] 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,

039(002):354-359.

- [28] 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.

- [29] Yun J, Xu Z W, Gao G L. Gated Object-Attribute Matching Network for Detailed Image Caption[J]. Mathematical Problems in Engineering, 2020.