计算机科学
COMPUTER SCIENCE

基于空间和多层级联合编码的图像描述算法

方仲俊，张静，李冬冬

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900159

基于空间和多层级联合编码的图像描述算法

方仲俊[1,2]　张　静[1]　李冬冬[1,2]
1 华东理工大学信息科学与工程学院　上海 200237
2 苏州大学江苏省计算机信息处理技术重点实验室　江苏 苏州 215031
(y30190781@mail.ecust.edu.cn)

摘　要　图像描述是图像理解领域的热点研究课题之一,它是结合计算机视觉和自然语言处理的跨媒体数据分析任务,通过理解图像内容并生成语义和语法都正确的句子来描述图像。现有的图像描述方法多采用编码器-解码器模型,该类方法在提取图像中的视觉对象特征时大多忽略了视觉对象之间的相对位置关系,但它对于正确描述图像的内容是非常重要的。基于此,提出了基于Transformer的空间和多层级联合编码的图像描述方法。为了更好地利用图像中所包含的对象的位置信息,提出了视觉对象的空间编码机制,将各个视觉对象独立的空间关系转换为视觉对象间的相对空间关系,以此来帮助模型识别各个视觉对象间的相对位置关系。同时,在视觉对象的编码阶段,顶部的编码特征保留了更多的贴合图像语义信息,但丢失了图像部分视觉信息,考虑到这一点,文中提出了多层级联合编码机制,通过整合各个浅层的编码层所包含的图像特征信息来完善顶部编码层所蕴含的语义的信息,从而获取到更丰富的贴合图像的语义信息的编码特征。文中在MSCOCO数据集上使用多种评估指标(BLEU,METEOR,ROUGE-L和CIDEr等)对提出的图像描述方法进行评估,并通过消融实验证明了提出的基于空间的编码机制以及多层级联合编码机制能够辅助产生更为准确有效的图像描述语句。对比实验结果表明,所提方法能够产生准确、有效的图像描述并优于大多数最新的算法。

关键词:图像描述;Transformer;空间编码机制;多层级联合编码机制;注意力机制
中图法分类号　TP183

Spatial Encoding and Multi-layer Joint Encoding Enhanced Transformer for Image Captioning

FANG Zhong-jun[1,2],ZHANG Jing[1] and LI Dong-dong[1,2]
1 School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China
2 Provincial Key Laboratory for Computer Information Processing Technology,Soochow University,Suzhou,Jiangsu 215031,China

Abstract　Image captioning is one of the hot research topics in the field of computer vision. It is a cross-media data analysis task that combines computer vision and natural language processing. It describes the image by understanding the content of the image and generating captions that are both semantically and grammatically correct. Existing image captioning methods mostly use the encoder-decoder model. This kind of methods mostly ignore the relative position relationship between visual objects when extracting the visual object features in image,and the relative position relationship between objects is very important for generating accurate captioning. Based on this,this paper proposes a spatial encoding and multi-layer joint encoding enhanced transformer for image captioning. In order to make better use of the position information contained in the image,this paper proposes a spatial encoding mechanism for visual objects,which converts the independent spatial relationship of each visual object into the relative spatial relationship between visual objects to help the model to recognize the relative spatial relationship between each visual object. At the same time,in the encoder part of visual objects,the top encoding feature retains more semantic information that fits the image but loses part of the visual information of the image. Taking this into account,this paper proposes a multi-level joint encoding mechanism to improve the semantic information contained in the top encoding layer by integrating the image feature information contained in each shallow encoding layer,so as to obtain richer semantic features that fit the image. This paper evaluates the proposed image captioning method by multiple evaluation indicators(BLEU,METEOR,ROUGE-L,CIDEr,etc.) on the MSCOCO dataset. The ablation experiment proves that the spatial encoding mechanism and the multi-level joint encoding mechanism proposed in this paper can be helpful in generating more accurate and effective image captions. Comparative experimental results show that the proposed method in can produce accurate and effective image caption and is superior to most of the latest methods.

到稿日期:2021-09-22　返修日期:2022-03-09
基金项目:国家自然科学基金(61806078)
This work was supported by the National Natural Science Foundation of China(61806078).
通信作者:张静(jingzhang@ecust.edu.cn)

Keywords Image captioning, Transformer, Spatial encoding mechanism, Multi-level joint encoding mechanism, Attention mechanism

1 引言

图像描述任务是图像理解领域的重要研究内容，其结合了计算机视觉和自然语言处理两大领域的技术。它利用计算机视觉相关的技术提炼出了图像中关键内容的特征信息，然后利用自然语言处理的方法生成符合人们认知的而且语义和句法都正确的描述语句，因此这个任务非常具备挑战性。图像描述的发展经历了几个阶段，包括基于模板的方法[1-2]、基于检索的方法[3-4]以及编码器-解码器范式[5-6]等。

随着自然语言处理中机器翻译的发展，编码器-解码器的模型结构被广泛用于图像描述的研究，并逐渐成为了主流的方法。该模型结构通过编码器来对相关的图像进行编码，以提取出图像中关键的特征信息，然后通过解码器来利用这些提取好的特征生成符合人们认知的图像描述句子，而且该解码器结合了自然语言处理的技术，因此生成的语句往往在语义和语法上都非常符合人们的认知。

对于图像视觉特征的获取，在编码器的阶段多采用了VGG[7]，ResNet[8]等卷积神经网络对整个图像进行编码，从而得到全图的图像特征。而这种全图的图像特征往往包含了很多的无用信息，Faster-RCNN[9]可以用来获取图像细粒度的各个对象的特征信息，过滤掉与生成自然语句无关的特征信息，从而提炼出更加精准的图像特征信息。随着计算机视觉的发展，这些视觉的特征提取器的准确率也越来越高，从而能帮助模型更好地识别出图像的视觉对象。然而，图像中的各个视觉对象之间在生成描述性的语句时必然存在联系，而单纯采用更好的特征提取器来提取各个视觉对象的特征是远远不够的。图1给出了当前在图像描述任务中视觉特征提取的主要方法。

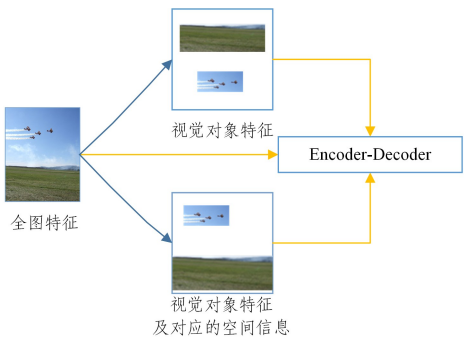


图1 图像描述领域的不同的特征处理方式

Fig. 1 Different feature processing methods in image captioning

在图像描述的方法中，一开始提取的是图像的全图特征，然而全图的特征往往携带了很多对生成最终的描述无效的特征，因此有人引入了Faster-RCNN来提取图像中的视觉特征，从而生成相应的图像描述。但提取出的视觉对象特征丢失了视觉对象特征在图像中的空间信息，而实际上视觉对象的位置关系信息对于准确有效的图像描述非常重要。因此，本文提出了基于视觉对象的空间位置信息的编码机制，用于解决Transformer在编码阶段提取出的视觉对象特征缺失空间信息的问题。

本文提出了一种基于Transformer的空间和多层级联合编码机制的模型SEMJET(Spatial Encoding and Multi-layer Joint Encoding Enhanced Transformer)。该模型通过获取到图像对象所对应的空间坐标这一额外信息，提出空间编码机制来显式地对图像的各个特征进行编码，从而更好地帮助模型理解每个视觉对象特征之间的位置关系。同时，为了获取到更加丰富的编码特征信息，本文通过Transformer模型的编码结构块来得到不同层次的特征信息，并提出多层级联合编码机制对这些不同层次的特征信息进行融合，建立起各个层级的特征之间隐含的复杂联系，从而完善高层特征所蕴含的语义信息，并生成更加准确的图像描述。本文的主要贡献如下：

(1)提出了一种空间编码机制，通过对各个视觉对象的空间信息进行编码，从而获取到视觉对象之间的相对位置关系，帮助模型更好地理解和描述对象之间的相对位置。

(2)提出了多层级联合编码机制，通过融合各个浅层的编码层所包含的图像信息和语义信息来完善顶部编码层所蕴含的语义信息，以得到更丰富的贴合图像的语义描述。

(3)实现了基于Transformer的空间和多层级联合编码的图像描述方法，通过将各个视觉对象的独立位置信息转换为视觉对象间的相对位置关系，来获取更加贴合图像的语义信息，以实现更为准确的图像描述。

本文第2节介绍了与本文研究内容相关的图像描述领域的相关工作；第3节详细介绍了本文模型的原理；第4节通过实验验证了本文模型的有效性；最后总结全文并展望未来。

2 相关工作

近年来，图像描述领域绝大多数的模型都是基于编码器-解码器结构的。本文将从经典的编码器解码器结构、语义信息增强的图像描述方法、基于Transformer的编码器解码器结构这3个方面来介绍图像描述的相关工作。

经典的基于编码器-解码器的图像描述方法大多基于CNN-LSTM架构，Vinyals等[10]提出了一种基于深度循环架构的生成模型，该生成模型结合了计算机视觉中的CNN(Convolutional Neural Networks)和机器翻译中的LSTM(Long Short-Term Memory)等技术，能够以一种端到端的方法生成多样的描述图像的句子，解决了基于模板的方法生成固定模式语句的问题。考虑到在生成句子时上下文注意到的图像区域不同，Xu等[11]将注意力机制引入到图像描述领域，使得每个时间步的LSTM关注到的视觉区域是不一样的，从而产生更准确的句子。Lu等[12]把视觉哨兵引入基于编码器-解码器的结构，从而可以自适应地决定是否关注图像的视觉特征或自然语言的语义描述。Chen等[13]提出了一种基于空间和通道的注意力机制，充分地利用了CNN具备空间和

通道的双重特征,从而生成更准确的语句。大多数方法采用顶层 LSTM 的输出作为最终的解码信息来生成相关的描述,而 Guo 等[14]提出了双重预测网络来充分利用 LSTM 中的层级信息,从而获取到更加丰富的解码特征信息。相比利用全图特征来解码生成相关的图像描述的方法,Anderson 等[5]在 LSTM 中提出了一种自下而上的机制,从视觉对象级别来计算注意力,从而避免了全图特征中的无效特征对描述语句生成的影响。

考虑到额外信息对图像描述的重要性,一些方法将额外信息加入编码解码的过程中,提出了语义信息增强的图像描述方法。Gan 等[15]提出了语义组成网络,在生成自然语句时,他们利用从图像中提取到的额外信息来帮助 LSTM 获取到更好的参数结果。由于引入了更加丰富的额外语义信息,因此可以利用更多更全面的特征来帮助模型获取到更好的图像描述结果。不同于 Gan 等[15]把图像中的额外的信息输入到每个 LSTM 的时间步中,Yao 等[16]提出了基于图像属性的 LSTM 架构,对额外的信息在各个不同时间步中的输入展开了研究,使得图像和语义可以共同作用,从而生成更准确的图像描述。Feng 等[17]主要利用基于注意力的 Reset 模型来改进编码器,从而获取更准确对象之间的关系的语义信息,如“inside”“cover”和“overlap”。鉴于在每个时间步的各个层级的 LSTM 的作用是不一样的,Li 等[18]提出了基于视觉和语义的 LSTM 的框架,他们在低层的 LSTM 中更加关注低级的视觉特征,由于高层的 LTSM 与生成的高级语义描述更接近,因此他们使高层的 LSTM 更关注于提取到的额外的语义信息。相比使用层级 LSTM 来处理额外信息,Zhang 等[19]提出的 pLSTM-A 以并行的 LSTM 来利用额外信息和编码信息。为了更好地利用视觉对象的区域关系,Zhang 等[20]提出了 VRAtt-Soft,该模型利用各个对象的绝对位置来隐含地探索多个区域之间的视觉关系。除了利用各个对象的绝对位置以外,Pei 等[21]提出 VREA 来学习每对视觉对象的相对位置关系。

经典的 CNN-LSTM 范式存在着无法在训练示例中并行化的问题[22],从而导致算法效率低下,而 Transformer 的编码器-解码器结构由于其高度的并行性被成功应用于图像描述。Li 等[6]提出了交互的注意力机制,通过在 Transformer 的解码阶段将视觉信息和语义信息进行交互,来获取更好的模型性能。相比 Li 等[6]利用了对象的语义信息,He 等[23]提出了对象关系的 Transformer 模型,通过利用对象的空间信息来改进注意力。为了解决非关键对象产生的噪声问题,Wang 等[24]提出了基于门控的注意力机制,减少了非关键对象的干扰,从而更有效地捕获视觉对象的全局信息。

考虑到图像中对象的空间位置关系对于描述图像的内容十分重要,本文提出了一种基于 Tansformer 的空间和多层级联合编码机制的模型。本文通过空间编码机制来帮助模型更好地构建对象特征之间的空间关系,将各个对象的绝对位置转换为所有视觉对象之间的相对位置关系,帮助模型更好地识别对象之间的关系;同时,本文提出的空间编码机制不仅限于每对视觉对象的相对位置关系,还可以更好地帮助模型识别所有对象之间的空间关系。深度学习网络的问题之一是梯度消失,在 Transformer 的编码层中,随着层数的加深,顶部编码层携带着更多的语义信息,却丢失了部分的图像信息,因此本文对编码器的各个层级的编码特征采用多层级联合编码机制,以底部编码层保留的图像所固有的图像信息反哺顶部编码层所丢失的部分图像特征,以获取到顶部编码层的贴合图像的语义信息,从而得到更为准确的图像描述信息。

3 模型框架

本节首先介绍了提出的模型的整体结构,然后分别介绍了该方法的两个主要贡献点:空间编码机制和多层级联合编码机制。

3.1 整体模型结构

图 2 给出了本文模型 SEMJET 的框架。本文提出了空间和多层级联合编码机制来改善 Transformer 的模型。视觉对象的空间信息在真实的世界中常被人们用来推理视觉对象的关系,本文通过空间编码机制引入视觉对象的空间信息,从而帮助模型更好地识别视觉对象在图中的各个位置。并且,本文提出多层级联合编码机制来建立 Transformer 的各个层级的特征之间隐含的复杂联系,这将有助于完善高层特征所蕴含的语义信息,以生成更加准确的图像描述。

本文通过 Faster-RCNN 获取到图片中的各个视觉对象的特征以及对应的空间坐标信息。本文提出了一种空间编码机制,将各个视觉对象对应的绝对位置关系转换为相对位置关系,并对各个视觉对象的特征进行编码,从而帮助模型更好地理解视觉对象之间的空间关系。然后将这些附带空间信息的视觉对象特征送入 Transformer 的编码层进行特征之间的编码。

编码层的第一个模块为多头注意力机制,如式(1)—式(3)所示。

$$\boldsymbol{M}(\boldsymbol{X})=\boldsymbol{W}_O\boldsymbol{C}(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h) \tag{1}$$

$$\boldsymbol{head}(\boldsymbol{X}_t)=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{2}$$

$$\boldsymbol{Q}=\boldsymbol{X}_t\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{X}_t\boldsymbol{W}_K,\boldsymbol{V}=\boldsymbol{X}_t\boldsymbol{W}_V \tag{3}$$

其中,$\boldsymbol{X}$ 为附带空间信息的视觉对象特征,$\boldsymbol{W}$ 是参数,C 表示向量连接操作,$\boldsymbol{M}$ 表示由多个单头注意力输出的特征组合而成的特征,即多头注意力特征,$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 分别是 $\boldsymbol{X}$ 的独立线性映射的不同特征,d 表示向量的维数,$\boldsymbol{head}$ 表示单头注意力输出的特征。随着网络层级的加深,会出现梯度消失的现象,因此 Transformer 的编码层通过残差连接和归一化来减轻这种现象,残差连接和归一化的计算式如式(4)和式(5)所示。

$$\boldsymbol{T}=\mathrm{LayerNorm}(\boldsymbol{X}+\boldsymbol{M}(\boldsymbol{X})) \tag{4}$$

$$\boldsymbol{X}'=\mathrm{LayerNorm}(\boldsymbol{T}+FFN(\boldsymbol{T})) \tag{5}$$

其中,$\boldsymbol{X}$ 为附带空间信息的视觉对象特征,$\boldsymbol{M}$ 为多头注意力输出的特征,LayerNorm 为归一化操作,FFN(Feed-Forward Networks)为前向传播网络,$\boldsymbol{X}'$为第一层 Transformer 的编码层的输出向量。前向传播网络的计算式如式(6)所示。

$$FFN(\boldsymbol{T})=\rho(\boldsymbol{T}\boldsymbol{W}_{f1}+b_{f1})\boldsymbol{W}_{f2}+b_{f2} \tag{6}$$

其中,ρ 为激活函数,$\boldsymbol{T}$ 为经过残差连接和归一化操作以后

输出的特征向量，W 为参数，b 为偏置项。

在经过 L 层的 Transformer 的编码层后，在底层的特征更加注重图像固有的信息，在高层的编码层会更加注重图像所携带的语义信息，本文通过多层级联合编码机制的方法来对各个编码层的输出特征进行编码，从而获取到更加丰富的编码层的输出特征信息。然后将这种多层级联合编码的特征输入到 Transformer 的解码层进行解码。

本文模型生成的单词序列分布的计算式如下：

$$\log p(\boldsymbol{C}|\boldsymbol{I})=\sum_{t=1}^{T}\log p(c_t|c_1,\cdots,c_{t-1},\boldsymbol{I}) \tag{7}$$

其中，$\boldsymbol{I}$ 表示输入的图像，c_t 表示在每个时间步生成的单词，$\boldsymbol{C}$ 表示一句描述图像的语句。

任务的目标是最小化损失函数，如式(8)所示。

$$loss=-\log p(\boldsymbol{C}|\boldsymbol{I}) \tag{8}$$

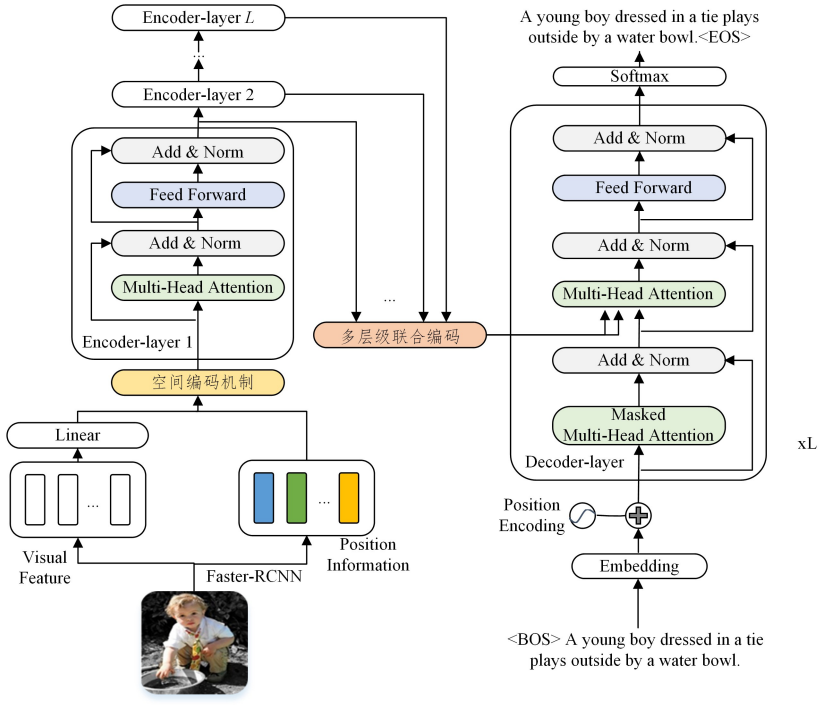


图 2 SEMJET 模型的框架

Fig. 2 Framework of SEMJET model

3.2 空间编码机制

由于 Faster-RCNN 所提取的视觉对象直接送入到了 Transformer 的编码层，不能很好地帮助模型识别各个对象之间的空间关系。基于这一点，本文引入了额外的空间坐标信息，并且提出了空间编码机制的方法来对各个视觉特征进行空间编码。

本文通过 Faster-RCNN 获取到的额外的空间绝对位置信息如式(9)所示。

$$\boldsymbol{\lambda}=(\boldsymbol{x},\boldsymbol{y},\boldsymbol{w},\boldsymbol{h}) \tag{9}$$

其中，$\boldsymbol{x}$ 表示各个视觉对象的 x 轴的值，$\boldsymbol{y}$ 表示各个视觉对象的 y 轴的值，$\boldsymbol{w}$ 和 $\boldsymbol{h}$ 分别表示各个视觉对象的宽度和长度。在获取到额外的空间坐标信息以后，本文引入了能够学习各个对象空间坐标信息之间的相对位置的正弦和余弦函数，来对空间坐标信息进行位置编码，正弦函数和余弦函数的计算式如式(10)、式(11)所示。

$$P_{(pos,2i)}=\sin(pos/10000^{2i/d}) \tag{10}$$

$$P_{(pos,2i+1)}=\cos(pos/10000^{(2i+1)/d}) \tag{11}$$

其中，pos 是空间坐标信息的位置，d 表示向量的维数的常量。在维度为偶数 $2i$ 时，使用正弦函数来编码；在维度为奇数 $2i+1$ 时，使用余弦函数来编码。通过正弦和余弦函数可以获取到各个对象之间的相对位置关系，然后通过线性变换，将编码以后的空间坐标信息映射到与视觉特征相同的特征空间中，对各个视觉对象特征进行相应的空间编码，如式(12)和式(13)所示。

$$\boldsymbol{R}=(\boldsymbol{x}+\boldsymbol{P}_{(1)},\boldsymbol{y}+\boldsymbol{P}_{(2)},\boldsymbol{w}+\boldsymbol{P}_{(3)},\boldsymbol{h}+\boldsymbol{P}_{(4)}) \tag{12}$$

$$\boldsymbol{X}=\boldsymbol{W}\boldsymbol{R}+\boldsymbol{V} \tag{13}$$

其中，$\boldsymbol{x}$ 表示各个视觉对象的 x 轴的值；$\boldsymbol{y}$ 表示各个视觉对象的 y 轴的值；$\boldsymbol{w}$ 和 $\boldsymbol{h}$ 分别表示各个视觉对象的宽度和长度；$\boldsymbol{P}$ 表示由正弦函数和余弦函数组成的位置编码机制，其下标表示空间坐标信息的位置；$\boldsymbol{W}$ 是参数；$\boldsymbol{V}$ 表示从 Faster-RCNN 提取的各个视觉对象特征；$\boldsymbol{R}$ 是由位置编码机制获取的各个对象之间的相对位置关系。通过这样的空间编码机制，模型可以从各个视觉对象的绝对位置信息中学习到各个视觉对象的相对位置的关系，并利用其来对各个视觉对象进行空间编码，从而学习到各个视觉对象之间的相对位置关系。

3.3 多层级联合编码机制

Transformer 的各个编码层的输出都有其特定的含义，底部的编码层的输出保留了更多图像所固有的信息，而越顶部的编码层则保留了更多的贴合图像的语义信息。为了对 Transformer 的各个编码层的输出特征的有效信息进行融合，本文提出了多层级联合编码机制的方法，通过融合 Transformer 的各个编码层所提取出的有效的语义信息和图像信息，来获取到更丰富的贴合图像的语义信息，从而帮助解码部分生成更准确的描述。

首先，本文获取到不包含顶层的 Transformer 的各个编码层的输出，如式(14)所示。

$$\boldsymbol{X}_{\text{ulti}}=\boldsymbol{W}_m C(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_{l-1}) \tag{14}$$

其中，$\boldsymbol{W}$ 是参数，C 表示向量连接操作，$\boldsymbol{X}_{l-1}$ 是第$(l-1)$层的编码特征。由于在解码阶段，模型除了识别图像所包含的视觉信息，更重要的是识别编码层所携带的语义信息，因此本文将编码器输出的层级信息用来增强编码器最顶层的输出，使得在编码阶段的输出不仅尽可能多地包含了图像的语义信息，还涵盖了随着网络层级的加深而丢失掉的图像信息，从而建立起各个层级的特征之间隐含的复杂联系，这将有助于完善高层特征所蕴含的语义信息，以生成更加准确的图像描述。多层级联合编码特征的计算式如式(15)、式(16)所示。

$$\boldsymbol{X}_{\text{enhanced}}=\boldsymbol{X}_l\otimes\sigma(\boldsymbol{X}_{\text{multi}}) \tag{15}$$

$$\boldsymbol{X}\text{X}_{\text{mje}}=LayerNorm(\boldsymbol{X}_{\text{enhanced}}+\boldsymbol{X}_l) \tag{16}$$

其中，$\otimes$表示点乘操作，这个操作通过使用编码器底部的输出特征来对编码器顶部的特征进行增强，使得多层级联合编码以后的特征具备更强的特征表现能力，从而达到更好的解码效果。

4 实验

4.1 实验数据集及配置

本文的实验使用的是 2014 年的 MSCOCO[25] 的数据集，这种数据集包含了 12 387 张图像，每张图像包含 5 句对应的图像描述。使用 Karpathy 的划分方法[26]，它将 MSCOCO 数据集划分为训练集、验证集和测试集。，然后使用 4 个常见的评估指标 BLEU[27]，METEOR[28]，ROUGE-L[29] 和 CIDEr[30] 来评估本文模型的有效性。

本文使用的服务器是 Linux 操作系统 Ubuntu16.04，使用 Pytorch 的深度学习框架来展开本文的实验，GPU 为显存为 8GB 的 Geforce RTX 2080，内存为 32 GB，CUDA 版本为 10.0，服务器 CPU 为英特尔 E5-2630。

本文通过 Faster-RCNN[9] 来提取最大上限为 50 的视觉对象特征以及该视觉对象特征所对应的对象的坐标轴、宽和高，对于数据集中的图像描述语句的预处理，本文将其全部转化成了小写，并且统计了在语句中至少出现过 5 次的单词，将它们组成单词词典，大小为 10201。本文使用“UNKNOW”作为单词不在单词词典中的标识，并且采用“BOS”和“END”分别作为生成单词序列的起始标志和结束标志。本文模型的编码层和解码层的层数默认为 6。

在训练阶段，本文使用了在图像描述领域广泛使用的 Adam[31] 优化器，然后在起始阶段使用较小的学习率来进行热身训练，学习率设为 0.0003，并且在第 3 个轮次之后以 0.5 的速率进行衰减，一共训练 15 轮。在测试阶段，本文采取了 BEAM SEARCH 大小为 3 的测试策略来获取候选描述语句。

4.2 对比实验结果分析

表 1 列出了本文提出的基于 Transformer 的空间和多层级联合编码的图像描述模型 SEMJET 与当前主流模型的对比结果。对比的模型包括：SCN[15]，LSTM-A5[16]，ARN[17]，VS-LSTM[18]，NIC[10]，SA[11]，Adaptive[12]，SCA-CNN[13]，ETA[6]，ORT[23]，TSSM[24]，pLSTM-A[19]，VRAtt-Soft[20] 和 VREA[21] 模型。

由表 1 可知，SEMJET 在评估指标 BLEU-1，BLEU-2，BLEU-3，BLEU-4，METEOR，ROUGE-L 和 CIDEr 上的分数分别为 76.9%，60.9%，47.4%，36.9%，28.4%，57.2%，117.3%，在多数指标上都优于其他模型。相比 SA，Adaptive 这些传统的编码器-解码器的模型而言，本文模型 SEMJET 在 BLEU-4 指标上平均提升了约 8.3%。相比 ETA，ORT 和 TSSM 这些基于 Transformer 的模型，本文模型 SEMJET 在 CIDEr 指标上分别提升了约 18%，5%和 4%。以 DPN 为例，该模型利用了两个 LSTM 的层级信息的平均输出来生成最终的描述。相比 DPN，本文模型利用 Transformer 的多个编码块来提取各个层次的图像特征，并对特征做了进一步融合，将融合的解码特征作用于 Transformer 的各个解码块，这一过程建立起了各个层级的特征之间隐含的复杂联系，将有助于完善高层特征所蕴含的语义信息，从而获取到更加丰富的编码特征，因此本文提出的 SEMJET 模型的结果优于 DPN。

表 1 不同图像描述模型指标对比

Table 1 Comparison of indicators of different image captioning models

(单位：%)

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
NIC	66.6	46.1	32.9	24.6	—	—	—
SA	71.8	50.4	35.7	25.0	23.0	—	—
Adaptive	74.2	58.0	43.9	32.2	26.6	—	108.5
SCA-CNN	71.9	54.8	41.1	31.1	25.0	—	—
DPN	72.6	—	—	32.0	24.7	53.4	—
SCN	72.8	56.6	43.3	33.0	25.7	—	101.2
LSTM-A5	73.4	56.7	43.0	32.6	25.4	54.0	100.2
ARN	72.8	56.2	42.3	32.3	26.0	54.2	102.0
VS-LSTM	76.3	59.4	44.8	34.3	26.9	—	110.2
ETA	72.2	—	—	31.9	25.7	53.4	99.2
ORT	75.6	—	—	33.5	27.6	56.0	112.6
TSSM	77.0	—	—	36.3	27.5	56.6	113.7
pLSTM-A	76.3	60.0	46.2	35.2	26.8	56.9	108.3
VRAtt-Soft	**79.2**	—	—	**36.9**	28.3	**60.9**	114.3
VREA	77.6	—	—	**36.9**	27.9	56.8	115.5
SEMJET(ours)	76.9	**60.9**	**47.4**	**36.9**	**28.4**	57.2	**117.3**

相比 SCN, LSTM-A, ARN, pLSTM-A 以及 VS-LSTM 这些利用额外信息的模型，本文提出的 SEMJET 模型利用了额外的基于对象坐标的空间信息，并且对对象的空间信息进行了进一步的编码，从而获取到视觉对象之间的相对空间位置关系，以帮助模型更好地理解视觉对象之间的空间关系。VRAtt-Soft 引入视觉对象的绝对坐标，通过注意力机制让模型隐式地学习各个视觉对象之间的关系。本文提出的 SEMJET 模型通过位置编码机制将视觉对象的绝对位置关系转换为视觉对象之间相对的位置关系，可以更好地帮助模型直接学习到各个视觉对象之间的位置关系，SEMJET 在 CIDEr 指标上提升了约 3%。VREA 使用视觉关系推理模型来利用每对视觉对象之间的相对位置关系，而 SEMJET 模型通过一种新颖的空间编码机制的方法来对各个视觉特征进行空间编码，从而利用所有视觉对象之间的相对位置关系，从对比结果来看，本文在 CIDEr 指标上和 METEOR 上分别提升了约 1.8%和 0.5%。

4.3 消融实验结果分析

表 2 列出了消融实验的结果，复现的实验结果包括基于图中对象特征的经典的编码器解码器模型 BUTDA[5]、基于图中对象特征的 Transformer 模型、结合空间编码机制的 Transformer 模型 SET、结合多层级联合编码机制的 Transformer 模型 MJET 和基于 Transformer 的空间和多层级联合编码机制的 Transformer 模型 SEMJET。

表 2 各个模块的消融实验结果

Table 2 Ablation experiment results of each module

Method	BLEU-1/%	BLEU-2/%	BLEU-3/%	BLEU-4/%	METEOR/%	ROUGE-L/%	CIDEr/%	Times/min
BUTDA	75.6	59.5	46.0	35.6	27.4	56.2	110.5	763
Transformer	76.0	60.0	46.7	36.2	28.2	56.8	115.3	557
SET	76.5	60.6	47.1	36.6	28.2	56.8	116.4	636
MJET	76.6	60.7	47.3	36.8	28.3	57.0	116.7	504
SEMJET	**76.9**	**60.9**	**47.4**	**36.9**	**28.4**	**57.2**	**117.3**	**501**

表 2 中，将 BUTDA 和 Transformer 进行对比可以看出，本文所采用的 Transformer 的模型结构的性能明显优于经典的编码器解码器模型。同时，从 BUTDA 和基于 Transformer 的模型（Transformer, SET, MJET 和 SEMJET）运行时间的对比可以看到，BUTDA 花费了 763 min，在运行时间方面平均多花了 191 min，这是因为基于 CNN-LSTM 的模型在解码部分每个时间步的输入都需要依赖上一个时间步，而基于 Transformer 的模型在训练阶段通过解决 RNN 的递归问题来提供并行序列建模能力以及出色的序列建模性能，并且实验结果表明基于 Transformer 的模型的性能优于基于 CNN-LSTM 的模型。通过将 SET 和 Transformer 进行对比，利用本文提出的空间编码机制来改进 Transformer，结果表明 SET 的 CIDEr 指标优于实验中的 Transformer 的模型，证明了本文提出的空间编码机制可以帮助模型有效地理解各个视觉对象之间的空间位置关系。同时，通过将 MJET 和 Transformer 进行比较发现，本文提出的结合多层级联合编码机制的 Transformer 在各项指标上也优于实验中的 Transformer 的模型，证明了本文提出的多层级联合编码机制可以有效增强 Transformer 的顶层编码层的输出，以获取到更丰富的对象特征。将 SET, MJET 和 SEMJET 进行比较可以发现，结合了空间编码机制和多层级联合编码机制的 Transformer 模型在各项指标上都明显优于结合了单个机制的 Transformer 模型，这证明本文提出的两个机制是可以互相促进的，可以帮助模型更好地理解视觉对象之间的关系，以及结合多层级的特征输出以达到特征增强的效果。

表 3 列出了对模型的关键参数进行消融实验的结果。由于底部的编码层的输出保留了更多图像所固有的信息，而越是顶部的编码层则保留了更多的贴合图像的语义信息，本文提出了多层级联合编码机制来利用底部和顶部的特征信息，因此对 SEMJET 模型的层数这一关键参数也进行了消融实验。本文提出的 SEMJET 模型默认的层数为 6 层，记为 SEMJET。我们将变化了层数以后的模型记为 SEMJET-N，N 表述变化以后的模型层数，比如 SEMJET 模型的编码层和解码层为 8 层时，记为 SEMJET-8。由表 3 可知，本文提出的 SEMJET 在编码层和解码层的层数都为 4 时，SEMJET 模型在绝大多数指标上的效果最好，SEMJET-4 的 CIDEr 指标超过 SEMJET 的 CIDEr 指标 1.2%，达到了 118.5%，同时 SEMJET-4 的运行时间也相对较短，只需要接近 7 h 就可以完成训练。这说明融合 Transformer 的各个编码层的特征需要合适的编码层的层数来提取出有效的贴合图像的语义信息，并且以恰当的解码层的层数来进行解码，从而更好地匹配图像所生成描述语句的语义复杂度。

表 3 本文提出的 SEMJET 模型关键参数的消融实验结果

Table 3 Ablation experiment results of key parameters of the proposed SEMJET

Method	BLEU-1/%	BLEU-2/%	BLEU-3/%	BLEU-4/%	METEOR/%	ROUGE-L/%	CIDEr/%	Times/min
SEMJET-8	76.4	60.2	46.6	36.1	28.2	56.8	116.2	547
SEMJET-7	76.0	60.0	46.5	36.0	28.2	56.8	115.2	469
SEMJET	**76.9**	60.9	47.4	36.9	**28.4**	**57.2**	117.3	501
SEMJET-5	76.3	60.5	47.1	36.7	**28.4**	57.1	116.6	419
SEMJET-4	76.7	**61.0**	**47.7**	**37.2**	28.3	**57.2**	**118.5**	409
SEMJET-3	76.5	60.7	47.3	36.8	28.2	56.9	116.9	408
SEMJET-2	76.4	60.4	47.0	36.5	28.1	56.8	116.4	**336**

4.4 定性实验结果分析

为了对实验结果有一个直观的理解，本文选取了一些实验结果展示在图 3 中，其中包含了图片、5 个人工标注的句子以及 SEMJET 模型生成的描述。可以看出，SEMJET 可以很好地捕捉图像中的视觉对象之间的空间关系以及图像的细节信息。例如图 3 中第一个样例的“large”和“next to”、图 3 中第二个样例的“in front of”以及第三个样例的“standing in”，这些生成的语句能够涵盖图像对象的空间信息和细节，并且模型生成的描述基本符合语法和语义，有效证明了本文提出的 SEMJET 的方法能够帮助模型理解每个视觉对象特征之间的位置关系，并且建立起各个层级的特征之间隐含的复杂联系，从而完善高层特征所蕴含的语义信息以生成更加准确的图像描述。

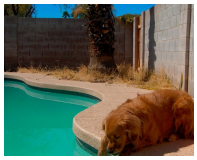
图像	描述
	人工标注的句子 1：A golden retriever laying down on the side of a pool. 人工标注的句子 2：A large brown dog laying next to a blue pool. 人工标注的句子 3：A golden retriever sleeps at the edge of the pool. 人工标注的句子 4：A swimming pool in a yard that has a cinderblock wall all the way around it and a dog sitting at the edge of the pool. 人工标注的句子 5：A dog laying down next to a pool in a backyard. 生成的描述：A large brown dog laying next to a pool of water.
	人工标注的句子 1：A bunch of people sit in an open court yard. 人工标注的句子 2：A small group of people standing around a ball patio. 人工标注的句子 3：A group of people walking around a parking lot. 人工标注的句子 4：A group of people in front of a white building. 人工标注的句子 5：Many people on a courtyard under a clock. 生成的描述：A group of people standing in front of a white building.
	人工标注的句子 1：A ram is looking at the camera and standing on some grass. 人工标注的句子 2：A large sheep bends his head towards some grass. 人工标注的句子 3：A very large sheep is standing in the grass. 人工标注的句子 4：A sheep looks at the camera，by the side of the road. 人工标注的句子 5：A woolly sheep stands in the grass looking at the camera. 生成的描述：A large sheep standing in a field of grass.

图 3 SEMJET 模型生成的图像描述

Fig. 3 Captions generated by SEMJET model

结束语 本文提出了基于 Transformer 的空间和多层级联合编码机制的模型，首先使用 Faster-RCNN 来获取图像中的各个对象特征以及额外的空间信息，提出空间编码机制显式地对各个对象特征进行编码，从而帮助模型更好地注意到图像中各个对象之间的相对空间关系；然后通过 Transformer 的多个编码结构块来获取各个层级的图像特征信息；最后通过多层级联合编码机制来获取到更加丰富的图像特征信息，用于在解码器阶段生成更加准确的描述语句。消融实验和对比实验验证了本文提出的基于 Transformer 的空间和多层级联合编码机制的图像描述算法的有效性。

接下来，将针对解码器，结合图像中的全局图像、对象特征信息、分割的对象实例信息和对象对应的语义信息，以及由粗到细的颗粒层次进行图像特征的解码，提出了更为准确有效的多层次图像描述方法。

参 考 文 献

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

[2] LU J，YANG J，BATRA D，et al. Neural baby talk[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7219-7228.

[3] DEVLIN J，CHENG H，FANG H，et al. Language models for image captioning：The quirks and what works[C]// Association for Computational Linguistics(ACL). 2015：100-105.

[4] WANG C，YANG H，BARTZ C，et al. Image captioning with deep bidirectional LSTMs[C]// Proceedings of the 24th ACM international conference on Multimedia. 2016：988-997.

[5] 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.

[6] LI G，ZHU L，LIU P，et al. Entangled transformer for image captioning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：8928-8937.

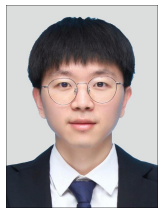
[7] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv：1409. 1556，2014.

[8] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：770-778.

[9] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，39(6)：1137-1149.

[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. 2015: 3156-3164.
[11] 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. PMLR, 2015: 2048-2057.
[12] 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. 2017: 375-383.
[13] 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. 2017: 5659-5667.
[14] GUO Y, LIU Y, DE BOER M H T, et al. A dual prediction network for image captioning[C]// 2018 IEEE International Conference on Multimedia and Expo. IEEE, 2018: 1-6.
[15] GAN Z, GAN C, HE X, et al. Semantic compositional networks for visual captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5630-5639.
[16] YAO T, PAN Y, LI Y, et al. Boosting image captioning with attributes[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4894-4902.
[17] FENG Y, LAN L, ZHANG X, et al. AttResNet: Attention-based ResNet for Image Captioning[C]// Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. 2018: 1-6.
[18] LI N, CHEN Z. Image Cationing with Visual-Semantic LSTM [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 793-799.
[19] ZHANG J, LI K, WANG Z. Parallel-fusion LSTM with synchronous semantic and visual information for image captioning[J]. Journal of Visual Communication and Image Representation, 2021, 75: 103044.
[20] ZHANG Z, WU Q, WANG Y, et al. Exploring region relationships implicitly: Image captioning with visual relationship attention[J]. Image and Vision Computing, 2021, 109: 104146.
[21] PEI H, CHEN Q, WANG J, et al. Visual Relational Reasoning for Image Caption[C]// 2020 International Joint Conference on Neural Networks. IEEE, 2020: 1-8.
[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances inNeural Information Processing Systems. 2017: 5998-6008.
[23] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: transforming objects into words[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 11137-11147.
[24] WANG D, HU H, CHEN D. Transformer with sparse self-attention mechanism for image captioning[J]. Electronics Letters, 2020, 56(15): 764-766.
[25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// European Conference on Computer Vision. Cham: Springer, 2014: 740-755.
[26] KARPATHY 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.
[27] 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.
[28] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]// Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005: 65-72.
[29] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// TextSummarization Branches Out. 2004: 74-81.
[30] VEDANTAM R, LAWRENC E 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.
[31] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]// Proceedings of the 3rd International Conference for Learning Representations. 2015.



FANG Zhong-jun, born in 1997, postgraduate, is a member of China Computer Federation. His main research interests include computer vision, neural networks and image captioning.

ZHANG Jing, born in 1978, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include computer vision, image/video information retrieval, image annotation and so on.

(责任编辑:何杨)