计算机科学
COMPUTER SCIENCE

基于外部先验和自先验注意力的图像描述生成方法

李永杰, 钱艺, 文益民

引用本文

李永杰, 钱艺, 文益民. 基于外部先验和自先验注意力的图像描述生成方法[J]. 计算机科学, 2024, 51(7): 214-220.

LI Yongjie, QIAN Yi, WEN Yimin. Image Captioning Generation Method Based on External Prior and Self-prior Attention [J]. Computer Science, 2024, 51(7): 214-220.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600167

基于外部先验和自先验注意力的图像描述生成方法

李永杰　钱　艺　文益民
广西图像图形与智能处理重点实验室(桂林电子科技大学)　广西 桂林 541004
(jayli1998@foxmail.com)

摘　要　图像描述是一种结合计算机视觉和自然语言处理的跨模态任务,旨在理解图像内容并生成恰当的句子。现有的图像描述方法通常使用自注意力机制来捕获样本内的长距离依赖关系,但这种方式不仅忽略了样本间的潜在相关性,而且缺乏对先验知识的利用,导致生成内容与参考描述存在一定差异。针对上述问题,文中提出了一种基于外部先验和自先验注意力(External Prior and Self-prior Attention,EPSPA)的图像描述方法。其中,外部先验模块能够隐式地考虑到样本间的潜在相关性进而减少来自其他样本的干扰信息。同时,自先验注意力能够充分利用上一层的注意力权重来模拟先验知识,使其指导模型进行特征提取。在公开数据集上使用多种指标对EPSPA进行评估,实验结果表明该方法能够在保持低参数量的前提下表现出优于现有方法的性能。

关键词:图像描述;自注意力机制;潜在相关性;外部先验模块;自先验注意力
中图分类号　TP391

Image Captioning Generation Method Based on External Prior and Self-prior Attention

LI Yongjie,QIAN Yi and WEN Yimin
Guangxi Key Laboratory of Image and Graphic Intelligent Processing(Guilin University of Electronic Technology),Guilin,Guangxi 541004,China

Abstract　Image captioning,a multimodal task that combines computer vision and natural language processing,aims to comprehend the content of images and generate appropriate textual captions. Existing image captioning methods often employ self-attention mechanisms to capture long-range dependencies within samples. However,this approach overlooks the potential correlations among different samples and fails to utilize prior knowledge,resulting in discrepancies between the generated content and reference captions. To address these issues,this paper proposes an image description approach based on external prior and self-prior attention(EPSPA). The external prior module implicitly considers the potential correlations among samples and removes interference from other samples. Meanwhile,the self-prior attention effectively utilizes attention weights from previous layers to simulate prior knowledge and guide the model in feature extraction. Evaluation results of EPSPA on publicly available datasets using multiple metrics demonstrates its superior performance compared to existing methods while maintaining a low parameter count.

Keywords　Image captioning,Self-attentive mechanisms,Potential associations,External prior module,Self-prior attention

1　引言

图像描述是同时涉及计算机视觉(CV)与自然语言处理(NLP)两个不同领域的任务,它是多模态的研究热点之一。该任务需要运用适当的词汇和语法来对图像进行描述,在帮助人们更好地理解图像、为视觉障碍者提供辅助服务和提高搜索引擎效率及环境交互等方面有着广泛应用。

尽管图像描述领域发展迅速,但仍面临着以下挑战:1)需要克服图像与文本的异构性,实现跨模态信息的提取和对齐;2)图像中存在大量显式和隐式的视觉语义信息,需要合理使用这些信息,并有针对性地生成描述;3)需要根据上下文或先验知识进行推理,进而生成流畅的描述。

近年来,受机器翻译模型[1]的启发,大多数图像描述生成方法通常采用编码-解码架构。文献[2-4]先利用卷积神经网络(CNN)对输入图像进行编码,再利用循环神经网络(RNN)来建模并生成一系列单词输出。然而,这种方法使得模型难以捕捉到长距离的上下文信息,并且难以生成长文本描述。随后,文献[5-7]证明了基于自注意力机制的Transformer在CV和NLP领域拥有巨大潜力。得益于自注意力机制强大的长距离关系建模能力,基于Transformer框架的方法逐渐成为图像描述领域的主流方法。目前,基于自注意力机制的图像描述生成方法已经取得了很多研究成果。例如,Huang等[8]提出AoA(Attention on Attention)模块,该模块通过扩展传统的注意力机制来确定注意力结果和查询结果之间的

到稿日期:2023-06-21　返修日期:2023-10-25
基金项目:广西重点研发计划项目(桂科AB21220023);国家自然科学基金(62366011);广西图像图形与智能处理重点实验室项目(GIIP2306);桂林电子科技大学研究生教育创新计划项目(2023YCXB11)
This work was supported by the Key R&D Program of Guangxi(AB21220023),National Natural Science Foundation of China(62366011),Guangxi Key Laboratory of Image and Graphic Intelligent Processing(GIIP2306) and Innovation Project of GUET Graduate Education(2023YCXB11).
通信作者:文益民(ymwen@guet.edu.cn)

相关性。此外,Cornia 等[9]利用额外的记忆向量来学习和编码先验知识,并在解码阶段使用多层连接来融合低级和高级特征。上述方法虽然取得了不错的效果,但仍存在一些问题尚未解决。一方面,自注意力机制在计算时只考虑到样本内部的特征信息,却忽略了输入样本与其他样本之间的潜在相关信息,而这类不同样本之间的联系对特征提取有着至关重要的影响。例如,当前样本存在一些噪声或冗余信息时,模型可以利用其他样本中包含的相关信息来减少当前样本中的干扰,达到增强模型泛化能力的效果。另一方面,在缺少先验知识引导时,自注意力机制可能会导致模型难以有效地聚焦于重要的特征。但现有方法在利用先验知识时往往需要大量的数据集和额外的内存开销。

针对上述问题,本文提出了一种基于外部先验和自先验注意力的图像描述生成方法(EPSPA)。该方法在编码端提出了外部先验模块(External Prior Module,EPM)和自先验注意力模块(Self-prior Attention,SPA),在解码端使用经典的 Transformer 解码器。具体来说,为捕捉到不同样本之间的潜在相关性,本文借鉴简洁高效的多层感知机(Multilayer Perceptron,MLP)设计了 EPM。与自注意力机制需要计算多个权重矩阵相比,EPM 由投影和静态参数化的循环矩阵组成,参数量较少且计算复杂度较低。此外,本文在原有自注意力机制的基础上,在计算时将上一层的注意力权重向后传递,进而提出了一种新的自注意力机制,本文称之为自先验注意力模块。

本文的主要贡献包括两个方面:

1)在编码端,提出了一种基于外部先验和自先验注意力的方法 EPSPA,用于解决 Transformer 编码器多头自注意力机制在生成描述过程中忽略样本间联系和缺少先验知识指导的问题。其中,在外部先验模块中,所有的输入共同更新一个权重矩阵,使得这个矩阵可以隐式地代表全部样本中最具信息量的特征。另一方面,自先验注意力模块在不增加额外内存开销的前提下,将上一层的注意力权重作为模拟的先验知识向后传递,进而指导模型提取到更显著的图像特征。

2)在公开数据集上对本文提出的 EPSPA 方法进行了实验,和其他方法相比,EPSPA 在参数量更少的同时在各项指标上取得了更优越的效果。

2 相关工作

很多研究者针对图像描述领域进行了诸多探索,并取得了显著进展。目前,主流的图像描述方法都是基于编码器-解码器架构的。本文将从基于 CNN-RNN 的图像描述、基于 Transformer 的图像描述,以及给本文提供灵感的 MLP 在视觉特征提取中的应用 3 个方面来进行介绍。

2.1 基于 CNN-RNN 的图像描述

近年来,CNN 和 RNN 的方法在图像描述领域取得了显著进展。这类方法通常采用 CNN 模型来提取图像特征,然后将其输入 RNN 模型以生成描述语句。例如,Vinyals 等[2]提出一个端到端的神经网络模型,使用 Inception-V3[10]作为 CNN 模块,使用长短期记忆网络(LSTM)作为 RNN 模块,并引入注意力机制来动态选择图像区域以进行描述。Xu 等[11]则使用 ResNet-101[12]作为 CNN 模块,并在 RNN 模块使用自适应注意力机制[13],以便在生成每个单词的同时考虑图像和文本信息。此外,还有一些工作尝试使用更复杂的 RNN 结构或引入其他信息来提高图像描述的质量和多样性。例如,Anderson 等[14]使用自下而上的注意力机制来预先检测出图像中的物体,并将其编码为特征向量,然后选择相关的物体生成描述。Li 等[15]则利用场景图[16]来表示图像中的物体、属性和关系,以便生成更丰富和逻辑性强的描述语句。然而,上述方法也存在一些问题,如缺乏对图像中细节和背景知识的考虑,以及生成语句过于简单或重复。

2.2 基于 Transformer 的图像描述

Transformer[5]是一种基于自注意力机制的神经网络架构,最初被用于自然语言处理任务,如机器翻译和文本摘要。近年来,Transformer 也被广泛应用于图像描述任务,即根据给定的图像生成一段描述性文本。它由编码器和解码器组成,编码器将图像分割成若干区域,并提取每个区域的特征向量;解码器则根据编码器的输出和已生成的单词来预测下一个单词。Transformer 可以利用多头自注意力机制来捕捉图像和文本之间的关系,从而生成更准确和流畅的描述。目前,基于 Transformer 框架的图像描述已有很多研究。例如,Herdade 等[17]在编码器中加入对象之间的空间关系以进一步提取图像特征;Fang 等[18]提出空间编码机制和多层级联合编码机制,帮助模型更好地理解对象之间的相对位置;Huang 等[8]提出 AOA 模块,该模块通过扩展传统的注意力机制来确定注意力结果和查询结果之间的相关性;Cornia 等[9]提出多层的网格状连接来融合低级和高级的图像特征;Guo 等[19]加入对象的相对几何关系,以弥补 Transformer 无法对输入对象的几何结构进行建模的缺陷。尽管基于 Transformer 的图像描述方法取得了巨大成功,但也存在不足之处。具体表现在 Transformer 的训练需要大量计算资源和数据资源。例如,在处理长序列时,需要计算所有位置之间的注意力权重,这会使得计算复杂度呈二次增长,从而带来巨大的计算开销,可能导致模型过拟合,同时还存在模型在不同领域和场景下的适应性问题。

2.3 MLP 在视觉特征提取中的应用

近期,多层感知机(MLP)在视觉任务上取得了令人瞩目的成果,引发了学术界的广泛关注。最近的一些工作表明,只要合理地设计输入输出和层间交互方式,MLP 可以在图像分类任务上达到与 CNN 或 Transformer 相当甚至超越它们的性能。例如,Tolstikhin 等[20]提出了 MLP-Mixer,它将图像划分为多个块,并将每个块映射为一个向量,然后使用两种类型的 MLP 层来处理这些向量,一种是沿着空间维度进行交互的 channel-mixing MLPs,另一种是沿着通道维度进行交互的 token-mixing MLPs,这样就可以实现跨空间位置和跨通道特征的信息融合。其他工作也探索了如何利用 MLP 来实现高效且强大的视觉分类模型。例如,Liu 等[21]提出了 gMLP,它使用门控线性单元(Gated Linear Unit)来替换传统的激活函数,并引入了空间感知门控模块(Spatially Aware Gating Module),可以根据不同位置调节信息流动。Guo 等[22]提出了一个具有线性复杂度的外部注意机制,简单地使用两个 MLP 和两个标准化层就可以很容易取代现有架构中的自注意机制。

基于上述研究,本文采用 MLP 来设计外部先验模块。通过初始化一个共有的权重矩阵,可以隐式地学习样本间的潜在相关性,并且还提出了自先验注意力机制,用于加强对图片特征的提取。

3 本文方法

本章对 EPSPA 的结构进行介绍。方法框架图如图 1 所示。首先介绍生成模型的总体结构，其次介绍用于捕获样本间潜在相关性的外部先验模块和模拟先验知识的自先验注意力模块，最后对整个编码和解码的流程以及学习策略进行详细阐述。

3.1 模型概述

本文提出的图像描述生成模型整体也采用编码-解码架构。如图 1 所示，首先采用预训练 Faster R-CNN[23] 提取输入图像的视觉特征，并将其表示为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$，$\boldsymbol{X}\in\boldsymbol{R}^{N\times d}$，其中 N 为图像区域数量，d 为特征维数。$\boldsymbol{x}_n\in\boldsymbol{R}^d$ 代表第 n 个区域的视觉特征。将 $\boldsymbol{X}$ 输入包含外部先验模块(EPM)和自先验注意力模块(SPA)的编码层中得到输出 $\widetilde{\boldsymbol{x}}^i$，$L$ 层编码层依次堆叠，更高的编码层可以利用上一层的输出，最终得到整个编码器的输出 $\widetilde{\boldsymbol{X}}^L$。随后将人工标注参考描述 $\widetilde{\boldsymbol{Y}}$ 和 $\widetilde{\boldsymbol{X}}^L$ 一同作为解码层的输入，同样经过 L 层堆叠进而生成最终的图像描述 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_T\}$，其中 T 为生成描述的长度。上述内容可表示为：

$$\begin{aligned}\widetilde{\boldsymbol{X}}^L&=Encoder\ Layer\ with\ EPSPA(\boldsymbol{X})\\ \boldsymbol{Y}&=Decoder\ Layer(\widetilde{\boldsymbol{X}}^L,\widetilde{\boldsymbol{Y}})\end{aligned}\tag{1}$$

3.2 外部先验模块

经典的自注意力机制仅考虑到样本内部之间的联系，却忽略了样本间的潜在相关性。例如，两张图像可能具有相同的主题、颜色、纹理、形状等特征，这些特征可以构成图像之间的潜在相关性。自注意力机制可能会过度关注某些特定样本中的噪声信息而缺乏对其他样本中全局特征的关注。并且在计算过程中，它需要存储所有位置的嵌入表示和对应的注意力权重，这会占用大量内存。

为此，本文提出具有线性复杂度的外部先验模块(EPM)。EPM 与经典自注意力机制的区别在于它不使用相似度作为权重矩阵，而是采用一个循环矩阵让模型从原始数据中学习权重。由于 EPM 只需要存储一个固定大小的矩阵和序列的嵌入表示，因此能显著减少内存的占用。循环矩阵[24]由一组固定权重循环构成，矩阵中的每一行都是上一行循环右移得到的。因为权重矩阵是根据不同的输入图像协同更新的，在训练过程中，每当新的图像数据经过网络时，都会根据该图像计算梯度并更新权重。这些更新不仅基于当前图像的特征，还受到之前所有图像数据的影响，因为之前的图像数据已经对权重进行了初步的调整。因此，随着训练的进行，权重矩阵 $\boldsymbol{W}$ 会逐渐保留来自不同图像的特征信息。这种机制使模型能编码丰富的上下文信息，从而有助于模型理解不同图像之间的联系。为了增强模型的泛化能力，本文在 EPM 中引入多头机制，将输入 $\boldsymbol{X}$ 沿着特征维度分成 k 个子张量，每个子张量 $\boldsymbol{X}_i$ 对应一个矩阵 $\boldsymbol{W}_i$ 和偏置 $\boldsymbol{b}_i$。

具体来说，给定从输入图像中提取的一组图像区域 $\boldsymbol{X}$，其操作可以表示如下：

$$\begin{aligned}EPM(\boldsymbol{X})&=Concat(head_1,head_2,\cdots,head_H)\\ head_i&=\boldsymbol{W}_i Norm(\boldsymbol{X}_i)+\boldsymbol{b}_i\end{aligned}\tag{2}$$

其中，EPM 代表外部先验模块，H 代表多头的数量，$head_i$ 代表第 i 个头。将权重矩阵 $\boldsymbol{W}$ 初始化为一个 $N\times N$ 的循环矩阵，$Norm$ 表示归一化操作。

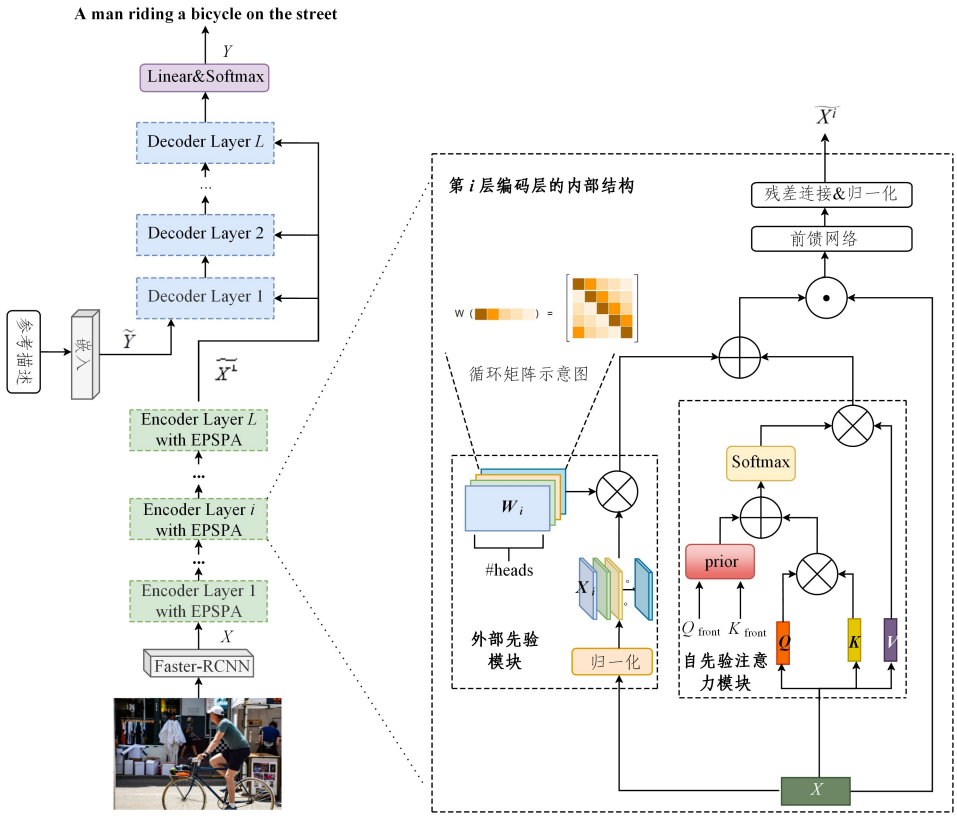

图 1 EPSPA 方法框架图

Fig. 1 Framework of EPSPA

3.3 自先验注意力模块

经典自注意力机制在每个时刻更新加权后的注意力向量,让模型在不同时刻关注到图像的不同区域,从而充分利用图像中的区域特征信息。它将输入投影到 3 个矩阵 $\boldsymbol{Q}$,$\boldsymbol{K}$ 和 $\boldsymbol{V}$,即先计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的注意力权重,然后将这些注意力权重与 $\boldsymbol{V}$ 相乘,表示对图像特征进行加权求和。其操作可以表示为:

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_K,\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V$$
$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{3}$$

其中,$\boldsymbol{W}_Q$,$\boldsymbol{W}_K$,$\boldsymbol{W}_V$ 为可学习的映射矩阵。

但自注意力机制的每次计算都缺少先验知识的引导,这可能会导致模型对一些无关的信息关注过多。为了对模型进行引导,使注意力相关性更强,本文提出了自先验注意力模块。具体来说,当前编码层在进行计算时会将上一层的注意力权重作为参数补充,并传递到后一层。这使得注意力能关注到更相关的信息,同时对模型起到一定的指导作用。值得注意的是,为了减少冗余计算,本文没有像传统方法那样在自先验注意力模块中使用多头,而是仅仅使用单头将特征映射到一个空间中,让模型更专注于当前特征。其操作表示为:

$$SPA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V},prior)=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}+prior\right)\boldsymbol{V}$$
$$prior=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{front}}\boldsymbol{K}_{\mathrm{front}}^{\mathrm{T}}}{\sqrt{d}}\right) \tag{4}$$

其中,SPA 代表自先验注意力,prior 中的 $\boldsymbol{Q}_{\mathrm{front}}$ 和 $\boldsymbol{K}_{\mathrm{front}}$ 来自于前一层,第一层的 prior 初始化为空。自先验注意力在当前注意力权重的基础上加入之前的权重,然后像经典注意力一样进行加权求和。

3.4 生成模型的整体流程

为了在生成描述时更好地利用图像之间的潜在相关信息以及注意力计算中的层间信息,本文在编码端把 EPM 和 SPA 提取到的特征信息通过一种自适应门控机制融合起来。它能够在融合视觉特征的过程中调节两者所占的权重,具体可表示为:

$$\boldsymbol{X}_{\mathrm{Enhance}}=\boldsymbol{X}\odot(EPM(\boldsymbol{X})+SPA(\boldsymbol{X})) \tag{5}$$

其中,$\boldsymbol{X}_{\mathrm{Enhance}}$ 表示融合了外部先验模块和自先验注意力模块的视觉特征,$\odot$ 表示元素级乘法。

为了使模型具有非线性,增加一组前馈网络(Feed Forward Network,FFN),FFN 使用两个全连接层进行变换。其表示如下:

$$\boldsymbol{F}(\boldsymbol{Z})_k=\boldsymbol{U}\sigma(\boldsymbol{V}\boldsymbol{Z}_k+b)+c \tag{6}$$

其中,$\boldsymbol{Z}_k$ 表示输入集的第 k 个向量,$\boldsymbol{F}(\boldsymbol{Z})_k$ 表示第 k 个向量的输出,$\sigma(\cdot)$是 ReLU 激活函数,$\boldsymbol{V}$ 和 $\boldsymbol{U}$ 是可学习的权重矩阵,b 和 c 是偏差项。

最后经过残差连接和归一化操作,得到编码层的完整定义如下:

$$\widetilde{\boldsymbol{X}}=AddNorm(F(\boldsymbol{X}_{\mathrm{Enhance}})) \tag{7}$$

其中,$AddNorm$ 表示残差连接和归一化操作的组合,$\widetilde{\boldsymbol{X}}$ 表示编码层的输出。

对于模型的解码层,如式(1)所示。本文参照标准的 Transformer,将编码器最后一层的输出 $\widetilde{\boldsymbol{X}}^L$ 以及参考描述 $\widetilde{\boldsymbol{Y}}$ 输入解码层,生成最终的描述。

3.5 学习策略

本文参考标准的图像描述训练方法,把训练分成两个阶段。首先对每个时刻生成的单词采用交叉熵损失函数进行训练,然后使用强化学习来微调生成的描述。

对于交叉熵损失,首先给定真实标签 $\boldsymbol{y}^*_{1:t-1}$ 来预测下一个单词,将模型参数记为 θ,目标是最小化模型的损失函数 L_{XE},即在每个时刻,最大化真实单词 $\boldsymbol{y}_t^*$ 的概率。具体表示为:

$$L_{\mathrm{XE}}(\theta)=-\sum_{t=1}^{T}\log(P_\theta(\boldsymbol{y}_t^*\mid\boldsymbol{y}^*_{1:t-1})) \tag{8}$$

其中,T 为句子的长度。使用交叉熵损失函数在训练过程中对每个单词进行独立优化,导致生成的描述会存在一定偏差。为了更好地生成相应的描述,本文使用 CIDEr-D[25] 分数作为奖励,以提高模型的最终表现。因此,通过自我批判的序列训练(SCST)[26] 不断优化 CIDER-D 的最终梯度表达式为:

$$\nabla_\theta L_{RL}(\theta)=-\frac{1}{k}\sum_{i=1}^{k}(r(y^i_{1:T})-b)\nabla_\theta\log P_\theta(y^i_{1:T}) \tag{9}$$

其中,b 表示使用贪婪算法来作为模型生成描述对应的奖励分数,$b=(\sum_i r(y^i_{1:T}))/k$,$k$ 是 beam size,$r(y^i_{1:T})$是 CIDEr-D 评分函数。在序列的预测过程中,本文采用集束搜索策略(Beam Search)[27],即每个时刻从概率分布中采样概率最大的前 k 个单词,并在解码过程中始终保留置信度最高的前 k 个文本序列。最后,将置信度最高的序列作为预测的文本描述。基于强化学习的训练方法直接在 CIDEr-D 评价指标上优化了描述的生成,使得模型生成的描述更加完整、流畅。

4 实验

本章首先介绍实验所用的数据集、评价指标和网络参数设置。然后为了验证 EPSPA 方法的有效性,进行了消融实验,并将所提方法与其他主流方法进行对比。最后从定性的角度对实验结果进行分析。

4.1 数据集与评价指标

本文使用的 MSCOCO[28] 数据集是当前图像描述领域经常使用的大型公开数据集,该数据集包含超过 12 万张图像,其中每幅图像至少有 5 条人工标注的参考描述。实验遵循 Karpathy 等[3] 提供的分割方法,其中 5 000 张图片用于验证,5 000 张图片用于测试,其余图片用于训练。

在测试阶段,本文使用了一套完整的评价指标,分别为 BLEU[29],METEOR[30],ROUGE[31],CIDEr[25] 和 SPICE[32]。其中,BLEU 是用于评估机器生成文本质量的指标;METEOR 是用于评估自动机器翻译的指标;ROUGE 基于最长公共子串来计算准确率;CIDEr 用于评测生成描述和参考描述的相似度;SPICE 是一种基于场景图和语义概念的评估指标,用于衡量生成语句是否描述了图像中各个对象之间的关系。

4.2 实验设置

为了编码图像区域特征,本文使用了 Faster R-CNN 和在 Visual Genome dataset[33] 数据集上微调的 ResNet-101[12],每一个区域用一个特征向量表示。为了编码参考描述,本文使用独热编码,并将它们投影到 512 维后再输入到模型。

在模型进行交叉熵损失训练时,本文将每一层的维度 d

设置为 512，head 设置为 8，warm-up 设置为 20 000，在每个前馈网络后采用概率为 0.9 的 dropout。如在训练过程中，验证集的 CIDEr 连续下降 5 个训练周期，则进入强化学习阶段。随后在对 CIDEr 的优化过程中，使用固定的学习速率 5×10^{-6}。本文采用 Adam 优化器[34]进行训练，批处理大小为 50。在测试阶段，模型使用集束搜索方法[27]，其中 beam size 设置为 5。

4.3 消融实验与分析

为了验证本文所提方法 EPSPA 的有效性，实验采用标准的 Transformer 作为基础网络。根据文献[9]的研究，基于 Transformer 的模型编解码层数设置为 3 时，达到最好的效果。因此，消融实验中的所有方法均使用 3 层编码层和解码层，并且表中所有实验均使用了强化学习。表 1 列出了在 MSCOCO 数据集上外部先验模块与基线方法 Transformer 的对比结果，其中 B-1，B-4，M，R，C 和 S 分别表示 BLEU-1，BLEU-4，METEOR，ROUGE-L，CIDEr 和 SPICE。EPM 表示使用外部先验模块替换 Transformer 中的多头自注意力机制。

表 1 EPM 消融实验结果对比

Table 1 Comparison of EPM ablation experiment results

Models	B-1	B-4	M	R	C	S
Transformer	79.6	36.5	27.8	57.0	123.6	21.1
EPM	**79.8**	**37.2**	**28.4**	**57.5**	**125.9**	**21.3**

由表 1 可以看出，本文仅使用外部先验模块替换原有 Transformer 编码器中的多头自注意力机制，模型在各个指标上都有所提高，其中 CIDEr 值提高了 2.3。说明在编码器端不使用任何注意力计算，仅仅依靠外部先验模块获取样本间的潜在相关性就可以给模型带来提升，证明了外部先验模块的有效性。

表 2 列出了 EPM 结合自先验注意力与普通自注意力在数据集上的对比结果。其中，EPM＋SA 表示同时使用外部先验模块和一个经典单头自注意力，EPM＋SPA 表示使用自先验注意力。

表 2 自先验注意力与经典注意力的对比结果

Table 2 Comparison results of SPA and classical attention

Models	B-1	B-4	M	R	C	S
EPM＋SA	80.6	38.4	28.8	58.0	129.5	22.5
EPM＋SPA	**80.8**	**38.6**	**29.2**	**58.5**	**131.5**	**22.7**

如表 2 所列，同时使用 EPM 和注意力机制，模型的性能得到了提升。这说明 EPM 可以很好地与注意力机制相结合，进而提高模型生成描述的质量。相比表 1 中仅仅使用 EPM，添加自注意力后 B-1 提高了 0.8，CIDEr 提高了 3.6。而将普通注意力更换为本文的自先验注意力后，指标再一次获得提升，其中表示生成描述和参考描述相似度的 CIDEr 指标达到了 131.5。说明自先验注意力模拟先验知识进行传递可以辅助模型生成质量更高的描述，这证明了自先验注意力优于经典自注意力。

4.4 对比实验与分析

将 EPSPA 与当前主流的图像描述生成方法在 MSCOCO 数据集上进行实验对比，包括使用特征网格注意的 SCST[26]方法、Anderson 等[14]基于自上而下注意力的 Up-Down 方法、Jiang 等[35]基于循环融合网络的 RFNet 方法、Yao 等[36]基于图神经网络的 GCN-LSTM 方法、Yang 等[37]基于自动编码场景图的 SGAE 方法、Huang 等[8]基于 AOA 模块的 AoA-Net 方法、Liu 等[38]基于视觉关联与上下文双注意力的 VR-CRA 方法、Herdade 等[17]基于加入对象间空间关系的 ORT 方法，以及 Cornia 等[9]基于记忆向量来学习和编码先验知识的 M2 方法。表 3 列出了 EPSPA 和上述方法在 MSCOCO 上的实验结果。

表 3 EPSPA 与其他对比方法在 MSCOCO 数据集上的实验结果

Table 3 Results of EPSPA and comparision methods on MSCOCO dataset

	B-1	B-4	M	R	C	S
SCST[26]	—	34.2	26.7	55.7	114.0	—
Up-Down[14]	79.8	36.3	27.7	56.9	120.1	21.4
RFNet[35]	79.1	36.5	27.7	57.3	121.9	21.2
VRCRA[38]	80.6	37.9	28.4	58.2	123.7	21.8
GCN-LSTM[36]	80.5	38.2	28.5	58.3	127.6	22.0
SGAE[37]	80.8	38.4	28.4	58.6	127.8	22.1
ORT[17]	80.2	38.6	28.7	58.4	128.3	22.6
AoANet[8]	80.2	38.9	29.2	**58.8**	129.8	22.4
M2[9]	80.8	**39.1**	29.2	58.6	131.2	22.6
EPSPA	**80.8**	38.6	**29.3**	58.5	**131.5**	**22.7**

由表 3 可知，EPSPA 与其他方法相比表现出了良好的性能。EPSPA 在 CIDEr 指标和 SPICE 指标上均超过了所有对比方法，同时在 BLEU-1 和 METEOR 指标上具有最强的竞争力，并且其 CIDEr 的指标提高了 0.3。这充分说明本文提出的 EPSPA 方法可以一定程度地利用样本间的潜在相关性，使模型获得良好的性能表现与泛化能力，从而提高模型生成描述的准确性。

4.5 性能参数比实验

为了证明 EPSPA 方法的轻量级性能，将其与当前主流的图像描述方法在参数量和描述的生成质量上进行了综合实验对比。由于外部先验模块不依赖复杂的注意力计算，仅仅引入一个固定大小的循环矩阵，因此外部先验模块的计算复杂度比自注意力机制要低。同时，相较于使用多头注意力的主流方法，本文只使用了单头自先验注意力以减少参数量。图 2 中，横轴代表参数量，纵轴代表 CIDEr 指标。

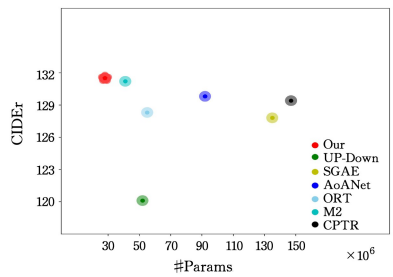

图 2 性能-参数图

Fig. 2 Performance-Parameters chart

实验结果表明 EPSPA 方法不仅参数量在所有方法中是最少的，而且 CIDEr 指标是最高的。

4.6 实验结果定性分析

为了对 EPSPA 的性能有一个直观的理解，本文选取

部分实验结果，如图 3 所示。其中 GT 表示参考描述。经过对比可以很直观地发现，EPSPA 可以更好地捕捉图像中对象的颜色、对象之间的空间关系以及图像细节。

如图 3(a)和图 3(b)所示，Transformer 只能描述摩托车和消防栓的大致方位，但是 EPSPA 可以捕获样本间的潜在相关性，借助从其他图片中学习到的颜色特征把摩托车和消防栓的颜色描述出来。同时，EPSPA 使用自先验注意力可以关注到图像中更相关的区域，从而可以生成“truck”和“brick sidewalk”。

如图 3(c)所示，Transformer 由于无法全面理解图像，导致其生成了“holding a bat”这一不太准确的描述。而 EPSPA 通过图片的潜在相关性对“bat”和“ball”进行关联，从而生成了“swinging a bat at a ball”这一更准确的描述。

如图 3(d)所示，EPSPA 可以检测出更多的细节，能把句子表达得更完整。EPSPA 生成的句子不仅能把“wearing”这个动作表达出来，更能够判断出男子的表情“smiling”，这是 Transformer 无法实现的。

正如上述例子所展现的，EPSPA 可以捕获图片中更详细的上下文信息，指导模型关注更相关的区域，进而生成更准确的图像描述。

GT: a motorcycle is parked in front of a truck.

Transformer: a motorcycle parked in front of a building.

EPSPA: a **blue** motorcycle parked in front of a **truck**.

(a)

GT: a fire hydrant on a busy city sidewalk.

Transformer: a fire hydrant on the side of a street.

EPSPA: a **yellow** fire hydrant on a brick **sidewalk**.

(b)

GT: a baseball player swinging a bat at home plate.

Transformer: a baseball player holding a bat on a field.

EPSPA: a baseball player **swinging** a bat at a ball.

(c)

GT: a man in a black shirt and orange bow tie.

Transformer: a man in a blue shirt and a red tie.

EPSPA: a man **wearing** an orange bow tie and **smiling**.

(d)

图 3 EPSPA 模型生成的图像描述

Fig. 3 Image captions generated by EPSPA model

结束语 本文提出了一种基于外部先验和自先验注意力的图像描述生成方法 EPSPA。该方法考虑到了不同样本之间的潜在相关性，以及利用经过学习的注意力权重模拟先验知识，使得生成的描述更准确。首先，为了联系样本间的视觉语义，本文提出了外部先验模块，它利用共享权重矩阵的优势来获取不同样本之间的潜在相关性。其次，在不增加额外内存的前提下，自先验注意力能够通过传递层间信息模拟先验知识，进而指导模型进行特征提取。最后，将经过融合形成的视觉特征进行解码，生成更准确的图像描述。实验结果表明，该方法在保持最低参数量的同时能够展现出超越当前主流方法的性能。

在未来的工作中，可以在编码端尝试融入更多的图像特征信息，在解码端尝试对图像特征进行更深层次的交互，提出描述更为准确的轻量化图像描述方法。

参 考 文 献

[1] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. arXiv: 1409.3215, 2014.

[2] 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.

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

[4] CORNIA M, BARALDI L, CUCCHIARA R. Show, control and tell: A framework for generating controllable and grounded captions[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8307-8316.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv: 1706.03762, 2017.

[6] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[7] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv: 2010.11929, 2020.

[8] HUANG L, WANG W, CHEN J, et al. Attention on attention for image captioning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4634-4643.

[9] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed-memory transformer for image captioning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10578-10587.

[10] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[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] 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.

[13] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv: 1508.04025, 2015.

[14] 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.
[15] LI Y, PAN Y, YAO T, et al. Comprehending and ordering semantics for image captioning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:17990-17999.
[16] JOHNSON J, KRISHNA R, STARK M, et al. Image retrieval using scene graphs[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3668-3678.
[17] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: Transforming objects into words[J]. arXiv: 2106. 10887, 2019.
[18] FANG Z J, ZHANG J, LI D D. Image description algorithm based on spatial and multi-level joint coding[J]. Computer Science, 2022, 49(10): 151-158.
[19] 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. 2020:10327-10336.
[20] TOLSTIKHIN I O, HOULSBY N, KOLESNIKOV A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in Neural Information Processing Systems, 2021, 34: 24261-24272.
[21] LIU H, DAI Z, SO D, et al. Pay attention to mlps[J]. Advances in Neural Information Processing Systems, 2021, 34: 9204-9215.
[22] GUO M H, LIU Z N, MU T J, et al. Beyond self-attention: External attention using two linear layers for visual tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 5436-5447.
[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] NEUMANN P M, PRAEGER C E. Cyclic matrices over finite fields[J]. Journal of the London Mathematical Society, 1995, 52(2): 263-284.
[25] VEDANTAM R, ZITNICK C L, PARIKH D. Cider: Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566-4575.
[26] 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. 2017:7008-7024.
[27] WANG P, NG H T. A beam-search decoder for normalization of social media text with application to machine translation[C]// Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013:471-481.
[28] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014:740-755.
[29] 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.
[30] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]// Proceedings of the ACL workshop on Intrinsicand Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005:65-72.
[31] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Proceedings of the Workshop on Text Summarization Branches Out. 2004:74-81.
[32] ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: Semantic propositional image caption evaluation[C]// Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing, 2016:382-398.
[33] 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.
[34] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv: 1412. 6980, 2014.
[35] JIANG W, MA L, JIANG Y G, et al. Recurrent fusion network for image captioning[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:499-515.
[36] YAO T, PAN Y, LI Y, et al. Exploring visual relationship for image captioning[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:684-699.
[37] YANG X, TANG K, ZHANG H, et al. Auto-encoding scene graphs for image captioning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10685-10694.
[38] LIU M F, SHI Q, NIE L Q. Image Captioning Based on Visual Relevance and Context Dual Attention[J]. Journal of Software, 2022, 33(9): 3210-3222.

LI Yongjie, born in 1998, postgraduate. His main research interests include computer vision, neural networks and image captioning.

WEN Yimin, born in 1969, Ph. D, professor, Ph. D supervisor, is a distinguished member of CCF(No. 06757D). His main research interests include machine learning, computer vision and big data analytics.

(责任编辑:杨雪敏)