基于改进Transformer的多尺度图像描述生成

崔 衡[1],张海涛[2],杨 剑[3],杜宝昌[1]

(1. 辽宁工程技术大学 软件学院,辽宁 葫芦岛 125105;2. 汕头职业技术学院 计算机系,广东 汕头 515071;3. 信息工程大学 地理空间信息学院,河南 郑州 450052)

摘 要: Transformer模型被广泛应用于图像描述生成任务中,但存在以下问题:①依赖复杂神经网络对图像进行预处理;②自注意力具有二次计算复杂度;③Masked Self-Attention缺少图像引导信息。为此,提出改进Transformer的多尺度图像描述生成模型。首先,将图像划分为多尺度图像块以获取多层次图像特征,并将其通过线性映射作为Transformer的输入,避免了复杂神经网络预处理的步骤,从而提升了模型训练与推理速度;其次,在编码器中使用线性复杂度的记忆注意力,通过可学习的共享记忆单元学习整个数据集的先验知识,挖掘样本间潜在的相关性;最后,在解码器中引入视觉引导注意力,将视觉特征作为辅助信息指导解码器生成与图像内容更为匹配的语义描述。在COCO 2014数据集上的测试结果表明,与基础模型相比,改进模型在CIDEr、METEOR、ROUGE和SPICE指标分数方面分别提高了2.6、0.7、0.4、0.7。基于改进Transformer的多尺度图像描述生成模型能生成更加准确的语言描述。

关键词: 图像描述;Transformer模型;记忆注意力;多尺度图像;自注意力

DOI: 10.11907/rjdk.231488 开放科学(资源服务)标识码(OSID):

中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-7800(2024)007-0160-07

Multi-scale Image Captioning Generation Based on Improved Transformer

CUI Heng[1], ZHANG Haitao[2], YANG Jian[3], DU Baochang[1]

(1. *Software College, Liaoning Technical University, Huludao* 125105, *China*; 2. *Computer Department, Shantou Polytechnic, Shantou* 515071, *China*; 3. *School of Geospatial Information, Information Engineering University, Zhengzhou* 450052, *China*)

Abstract: The Transformer model is widely used in image description generation tasks, but it has the following problems: ① relying on complex neural networks for image preprocessing; ② Self attention has a quadratic computational complexity; ③ Masked Self Attention lacks image guidance information. To this end, an improved Transformer based multi-scale image description generation model is proposed. Firstly, the image is divided into multi-scale image blocks to obtain multi-level image features, which are then linearly mapped as input to the Transformer, avoiding the steps of complex neural network preprocessing and improving model training and inference speed; Then, linear complexity memory attention is used in the encoder to learn the prior knowledge of the entire dataset through learnable shared memory units and explore potential correlations between samples; Finally, visual guided attention is introduced into the decoder, using visual features as auxiliary information to guide the decoder in generating semantic descriptions that better match the image content. The test results on the COCO 2014 dataset show that compared to the base model, the improved model has improved scores on CIDEr, METEOR, ROUGE, and SPICE indicators by 2.6, 0.7, 0.4, and 0.7, respectively. The multi-scale image description generation model based on improved Transformer can generate more accurate language descriptions.

Key Words: image captioning; Transformer model; memory attention; multi-scale image; self-attention

收稿日期: 2023-05-11

基金项目: 国家自然科学基金项目(42130112);国家重点研发计划项目(2017YFB0503500);KartoBit Research Network开放课题基金项目(KRN2201CA)

作者简介: 崔衡(1996-),男,辽宁工程技术大学软件学院硕士研究生,研究方向为机器视觉与多模态学习;张海涛(1974-),男,博士,汕头职业技术学院计算机系教授、硕士生导师,研究方向为多模态学习、图像图像处理;杨剑(1985-),男,博士,信息工程大学地理空间信息学院讲师,研究方向为时空数据挖掘与知识发现;杜宝昌(1998-),男,硕士,辽宁工程技术大学软件学院硕士研究生,研究方向为机器视觉与多模态学习。本文通讯作者:张海涛。

0　引言

图像描述生成是一项利用自然语言描述视觉图像内容的任务,需要准确识别图像中的对象特征,理解并建立视觉与文本之间的关系,因此具有一定的挑战性。该领域具有巨大的人机交互潜力,可用于帮助视力障碍人群理解图像内容、帮助驾驶员观察驾驶时可能出现的紧急情况以及自动生成医学影像诊断报告等。这些潜在的益处鼓舞了学者们对该技术的进一步研究与探索。

1　相关研究

早期图像描述任务通常采用基于检索的方法[1]。该方法对输入图像与已知图像库进行相似度匹配,如果找到与输入图像相似度较高的样本图片,则根据样本图片的描述直接生成输入图像的描述。该方法存在诸多限制,例如指定图像库中的描述常常不精确或无法描述某些情况,同时该方法需要大量图像库样本,增加了资源使用成本。

近年来,基于神经网络的编码器—解码器模型成为图像描述领域的主流方法[2-3]。编码器将输入图像编码为一组向量,通过循环神经网络将该组向量解码为一句语义描述。在该类模型中,编码器通常采用卷积神经网络作为高效的特征提取器,如 VGGNet[4]、ResNet[5]和 EfficientNet[6]等经典模型可以将输入图像转换为高层次的语义信息,以应对多种复杂场景和多样化视觉特征。解码器通常采用循环神经网络,如长短期记忆网络(Long Short-Term Memory, LSTM)[7]和门控循环单元(Gate Recurrent Unit, GRU)[8]等经典模型利用词向量、注意力机制[9]等技术自动学习并合理组织文本信息,从而生成与图像内容相关的自然语言描述。目前,许多学者对基于神经网络的编码器—解码器模型在图像描述领域中的应用进行了研究。例如,Xu 等[10]将注意力机制引入到图像描述生成任务中,通过对编码器提取的深层次特征向量进行加权求和来指导解码过程,使模型能够聚焦图像的重要区域,更符合人类视觉习惯;Anderson 等[11]提出一种融合自下而上与自上而下注意力机制的方法,其运用 Faster R-CNN 自下而上的机制编码图像特征[12],以便更有效地关注物体级别而非同等大小的图像区域;通过自上而下的机制计算特征权重并生成最终描述,从而实现更加准确和丰富的图像描述;Yang 等[13]提出一种融合视觉常识与多层全局特征的图像描述生成方法,该方法利用线性注意力机制挖掘对象间的视觉语义关系,并将视觉特征与区域特征融合起来;同时采用 AoA 机制增强全局特征,以达到更为准确的图像描述生成效果。

目前,基于自注意力机制的 Transformer 模型成为最流行的解码器模型,其具有高效的并行化处理能力,能够实现输入与输出之间的全局依赖关系,已在部分图像描述任务中取得了突出效果[14]。常见基于 Transformer 的图像描述生成模型结构如图 1 所示。以往基于 Transformer 的图像描述生成模型存在 3 个主要问题:一是依赖预训练的复杂神经网络(如 Faster-RCNN 或 VGGNet 等)提取图像深层次特征,使网络模型更加复杂;二是编码器中的自注意力机制具有二次计算复杂度,增加了训练和推理的计算开销;三是简单利用原始 Transformer 解码器中的 Masked Self-Attention 进行词嵌入信息交互,缺少图像信息的引导,不利于生成高质量的图像描述语句。

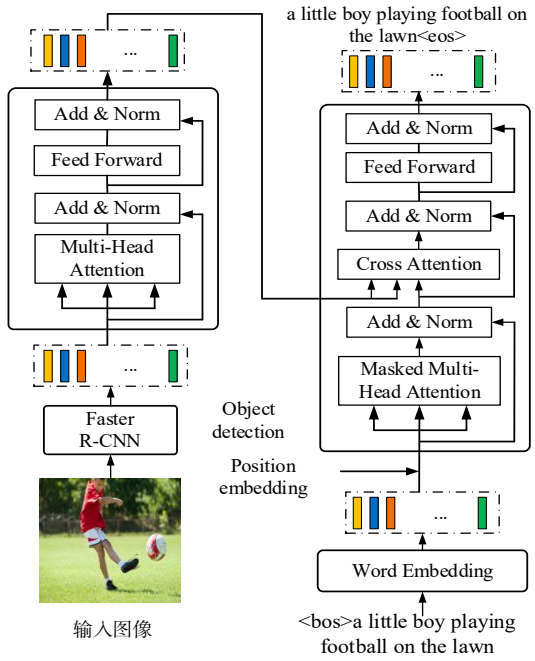

Fig. 1　Image captioning generation model based on Transformer

图 1　基于 Transformer 的图像描述生成模型

为解决上述问题,本文提出改进 Transformer 的多尺度图像描述生成模型。贡献之处在于:①将图像划分为多尺度的图像块以获取多层次的图像特征,将其通过简单线性映射作为 Transformer 的输入,而不需要预训练复杂神经网络提取图像特征;②在 Transformer 编码器中使用线性复杂度的记忆注意力,通过外部记忆单元在数据集中挖掘特征之间的潜在关系,学习更全面、丰富的特征表示;③改进 Masked Self-Attention,采用视觉特征指导描述词汇的信息交互。在 COCO 2014 数据集上的验证实验结果表明,以上改进提升了图像描述生成质量。

2　改进模型

针对以往模型存在的 3 个主要问题进行针对性改进:①使用多尺度 Vision Transformer(Multi-scale Vision Transformer, MViT)将图像划分为多尺度的图像块以获取多层次的图像特征,将其通过简单线性映射作为 Transformer 的输入,利用自注意力机制的全局交互能力获取全局信息的

图像特征；②将具有线性复杂度的记忆注意力引入 Transformer 编码器中，通过外部记忆单元在数据集中挖掘特征之间的潜在关系，学习更丰富的特征表示；③使用视觉引导注意力（Vision Leaded Attention，VLA）将视觉特征作为辅助信息引入 Transformer 解码器中，采用视觉特征指导自然语言描述生成。改进模型结构如图 2 所示。

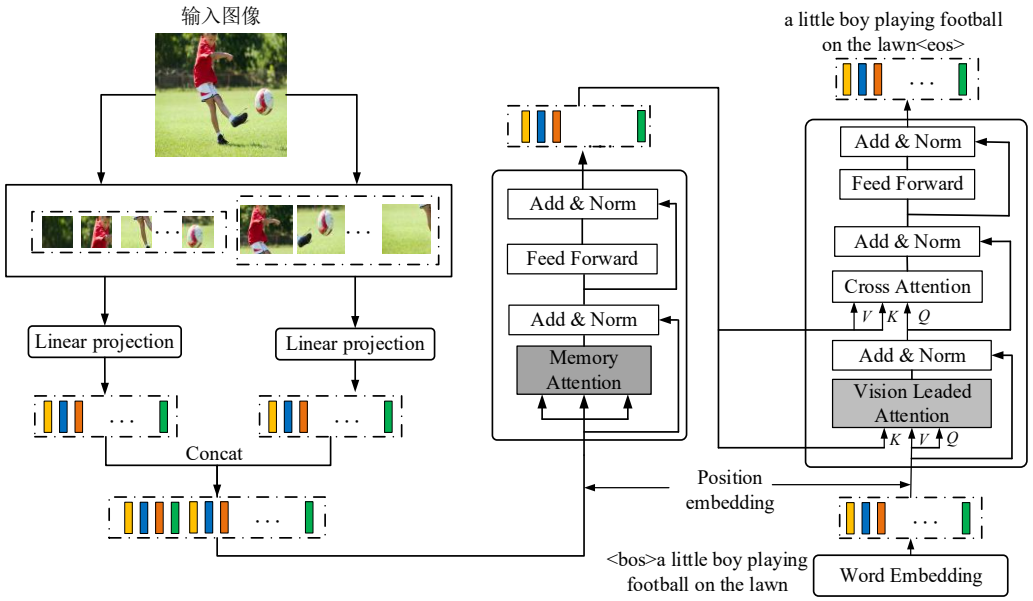

Fig. 2 Multi-scale image captioning generation model based on improved Transformer

图 2 改进 Transformer 的多尺度图像描述生成模型

2.1 MViT

以往的图像描述生成任务依赖于预训练的卷积神经网络模型（如 ResNet）提取空间特征或 Faster R-CNN 提取自下而上的特征，大大提升了模型复杂度。为此，本文在 Vision Transformer 的基础上进行改进。对于输入图像 $P \in \mathrm{R}^{H \times W \times C}$，首先重塑成小（small，s）图像块和大（large，l）图像块 $P^{\eta} \in \mathrm{R}^{N^{\eta} \times (P^{\eta^2} \cdot C)}$，其中 (H, W) 为原始图像的分辨率，C 为图像通道数，$(P^{\eta} \times P^{\eta})$ 为每个图像块的分辨率，$N^{\eta} = HW/P^{\eta^2}$ 为图像块数量，$\eta \in \{s, l\}$。由于 Transformer 编码器使用维度一致的特征向量，大小不同的图像块通过线性映射转换为维度一致的特征向量。对不同大小的特征向量进行拼接，得到一组特征向量 P，拼接可描述为：$P = \mathrm{Concat}(P^s, P^l)$。然后对其进行线性映射得到图像块特征序列 I，将 I 输入自注意力机制中进行特征交互，以获取全局交互后的图像特征向量序列。

2.2 记忆注意力

自注意力机制能够交互全局信息，并且能够从输入序列中提取重要信息，不仅被广泛用于自然语言处理、图像处理等领域，而且被用于图像描述生成等任务中。自注意力模型结构如图 3（a）所示。首先将图像块特征向量序列 I 分别通过线性层 W_q、W_k 和 W_v 映射到查询向量 $Q \in \mathrm{R}^{N \times d}$、键向量 $K \in \mathrm{R}^{N \times d}$ 和值向量 $V \in \mathrm{R}^{N \times d}$，并通过点乘计算得到查询向量与键向量之间的相似性，即权重矩阵 A；然后根据注意矩阵对值向量进行加权求和操作。自注意力公式表示为：

$$Q = IW_q, K = IW_k, V = IW_v \tag{1}$$

$$A = (\alpha)_{i,j} = \mathrm{SoftMax}(\mathrm{QK}^{\mathrm{T}} / \sqrt{d/h}) \tag{2}$$

$$I_{out} = AV \tag{3}$$

式中：W_q、W_k 和 W_v 为可学习的权重矩阵；d 为特征维度，h 为多头注意力头数，缩放因子 $\sqrt{d/h}$ 的作用为使梯度稳定。式（2）通过计算特征向量 Q 与 K 之间的相似性得出注意权重矩阵 A，$\alpha_{i,j}$ 表示第 i 与第 j 个元素之间的相似性；式（3）对值向量 V 进行加权求和操作得到自注意力的输出。

尽管自注意力机制已广泛应用于图像描述任务中，但其仍存在显著缺点。首先，自注意力机制具有二次计算复杂度 $O(dN^2)$，其中 N 为特征序列中元素的个数，这会使得计算成本增加，限制其可扩展性和效率；其次，自注意力机制仅考虑每个样本不同区域之间的关系，忽略了整个数据集中不同样本之间的潜在关系，这可能会限制模型的性能和灵活性。对于图像描述任务而言，特征表示的质量与生成句子的质量密切相关，因此需要探索更优秀的特征学习方法和模型架构。为此，本文在模型编码器中使用记忆注意力编码图像特征。如图 3（b）所示，不同于自注意力从特征向量映射到键向量和值向量，记忆注意力使用独立于输入特征的记忆单元作为键、值向量，能够捕获整个训练集具有信息性的部分。记忆注意力计算输入特征与外部记忆单元 $M \in \mathrm{R}^{S \times d}$ 之间的注意力，其中 S 表示记忆单元维度。计算方式为：

$$Q = W_q I \tag{4}$$

$$A = (\alpha)_{i,j} = \mathrm{DoubleNorm}(QM_k^T) \tag{5}$$

$$F_{out} = AM_v \tag{6}$$

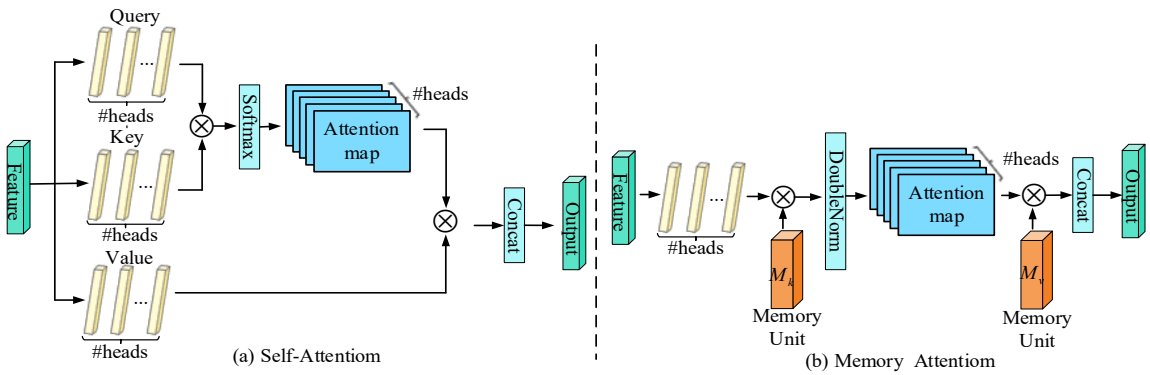

Fig. 3　Self-attention versus memory attention

图3　自注意力与记忆注意力

式(4)通过可学习的权重矩阵 W_q 将输入特征 I 映射为查询向量 Q。式(5)中 DoubleNorm 表示双重归一化;M_k 和 M_v 为独立于输入向量的可学习参数,通过两个线性层实现,可以采用端到端的方式通过反向传播进行优化,并作为整个训练数据集的共享记忆单元;$(\alpha)_{i,j}$ 表示查询向量 Q 中第 i 个元素与 M_k 中第 j 个元素之间的相似性;A 为根据训练集中学习到的先验知识所推断出的注意权重矩阵。式(6)通过注意权重矩阵 A 更新输入图像特征,由此得到记忆注意力的一次计算复杂度为 $O(dSN)$。

此外,在求解注意权重时,自注意力机制通过 Softmax 函数获得值向量的权重,使注意权重矩阵归一化为 $\sum_j \alpha_{i,j}=1$。然而,自注意力的注意权重矩阵是通过矩阵乘法计算得到,对输入特征的规模大小非常敏感,规模过大或过小均会破坏注意力机制计算任意两点键相似性的特性。为此,在记忆注意力中使用双重归一化对注意权重矩阵的行和列进行归一化处理。具体公式见式(7)—式(9)。

$$(\tilde{\alpha})_{i,j}=QM_k^T \tag{7}$$

$$\hat{\alpha}=\exp(\tilde{\alpha}_{i,j})/\sum_k \exp(\tilde{\alpha}_{k,j}) \tag{8}$$

$$\alpha_{i,j}=\hat{\alpha}_{i,j}/\sum_k \hat{\alpha}_{i,k} \tag{9}$$

式(7)表示外部注意力中计算得到的注意矩阵;式(8)表示对注意矩阵第一维度进行 Softmax 运算,用于消除输入特征规模大小对注意力的影响;式(9)表示对注意矩阵第二维度进行 L1-Norm 运算,最终求得 Q 与 M 之间的注意力分数。

2.3　VLA 机制

以往研究中的 Masked Self-Attention 模块仅用于单词间的信息交互。为使图像特征信息更有效地指导图像描述语句生成,本文设计了 VLA 模型,结构如图 4 所示。将 VLA 作为解码过程的辅助框架,综合视觉关系优化词序列内各个词汇之间的注意力分布,有助于充分发挥图像信息对解码过程的影响并生成更具图像语义相关性的描述。

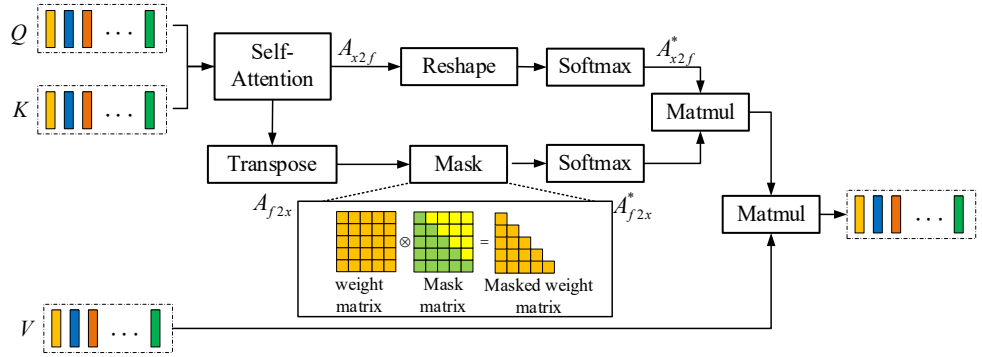

Fig. 4　Vision leaded attention mechanism

图4　VLA 机制

VLA 机制是一种双重注意力机制,其对具有图像特征引导的点乘运算得出的权重 A_{x2f} 进行分解,得到词序列—图像注意权重矩阵 A^*_{x2f} 和图像—词序列注意权重矩阵 A^*_{f2x}。给定一组词嵌入的输入序列 $X=\{x_1,x_2,...,x_k\}$,将其作为 VLA 中的查询向量 Q 和值向量 V;图像特征表示 $F=\{f_1,f_2,...,f_n\}$ 作为 VLA 中的值向量 K。如式(10)所示,首先通过 Q 与 K 点乘运算得到 $A_{x2f}\in \mathrm{R}^{k\times n}$,表示词嵌入序列 X 与图像特征 Z 之间的相关性(权重系数);然后对 A_{x2f} 进行转置得到图像—词序列注意力 $A_{f2x}\in \mathrm{R}^{n\times k}$。在真实任务中,词都是按照时序输入的,意味着当前时刻无法观察到后续输入,因此将 Mask 覆盖到 A_{f2x} 上得到 A^*_{f2x},使其只依赖于 t 时间之前生成的词序列信息,最终图像—词序列注意力表示可由式(11)得到。对词序列—图像注意力表示 A_{x2f} 进行 reshape 操作得到 A^*_{x2f},具体如式(12)所示。将 A^*_{x2f} 与

A^*_{f2x} 的乘积作为具有视觉影响的相似性度量值，并作用于词序列 X，则 VLA 的输出可表示为式(13)。综上所述，图像中不同区域的视觉特征均可参与到词嵌入序列中各个词之间关系程度的衡量中。

$$A_{x2f} = QK^T / \sqrt{d/h} \tag{10}$$

$$A^*_{f2x} = \mathrm{Softmax}(\mathrm{Mask}(A^T_{x2f})) \tag{11}$$

$$A^*_{x2f} = \mathrm{Softmax}(\mathrm{reshape}(A_{x2f})) \tag{12}$$

$$VLA(X, F, X) = A^*_{x2f} A^*_{f2x} X \tag{13}$$

2.4 目标函数

首先将交叉熵损失函数(Cross Entropy，XE)作为目标函数优化模型。表示为：

$$L_{XE}(\Theta) = -\sum_{t=1}^{T} \log(p_\Theta(y^*_t | y^*_{1:t-1})) \tag{14}$$

式中：$y^*_{1:T}$ 为真实标注的句子；Θ 为模型参数。

为使生成的描述语句更为准确，采用自临界序列训练策略优化 CIDEr 度量指标。表示为：

$$L_R(\Theta) = -\mathrm{E}_{y_{1:T} \sim p_\Theta}[r(y_{1:T})] \tag{15}$$

式中：$r(\cdot)$ 表示 CIDEr 的得分；L_R 的梯度可近似表示为：

$$\nabla_\Theta L_R(\Theta) \approx -(r(y^s_{1:T}) - r(\hat{y}_{1:T}))\nabla_\Theta \log p_\Theta(y^s_{1:T}) \tag{16}$$

式中：$y^s_{1:T}$ 为一个采样的句子；$r(\hat{y}^s_{1:T})$ 为从当前模型中获得的解码分数。

3 实验方法与结果分析

3.1 实验环境

实验基于 Linux 服务器展开，使用 PyTorch1.7.1 深度学习框架对模型进行训练、验证与测试。硬件环境配置为：Intel(R) Core (TM) 7-12700k CPU@3.6 GHz 处理器和显存 12 G 的 NIADIA Geforce RTX3060Ti GPU。

3.2 实验数据集

在目前图像描述生成任务中应用最广泛的 MS COCO 数据集[15]上进行实验。该数据集涵盖了人、动物和日常复杂场景等多种主题，且每张图片都被标注了 5 条人工描述语句。为确保对模型性能进行客观准确的评估，引用 Karpathy 等[16]提出的数据拆分方法对数据集进行划分，将原始数据集分为 3 个部分：113 287 张图片模型训练，5 000 张图片用于模型验证，5 000 张图片用于模型测试。同时，将所有句子转化为小写，删除标点符号，并剔除出现次数小于 5 次的单词，最终得到由 10 201 个单词组成的单词表，如此以来可确保生成的描述语句简洁明了、符合语言规范，并且有利于提高模型性能。

3.3 评价指标

为了更加全面地评估模型性能，使用目前常用的图像描述生成评价指标，包括 BLEU(Bilingual Evaluation Understudy)、METEOR(Metric for Evaluation of Translation with Explicit Ordering)、ROUGE_L(Recall-Oriented Understudy for Gisting Evaluation)、SPICE(Semantic Propositional Image Caption Evaluation)和 CIDEr(Consensus-based Image Description Evaluation)，分别简写为 B-1/2/3/4、M、R、S 和 C。其中，BLEU 用于评估机器翻译句子的保真度和流畅度，通过计算生成句子与真实句子的 n 元组匹配分数来衡量，划分为 B-1、B-2、B-3、B-4 几个等级，B-1 衡量的是单词级别的准确性，B-4 衡量的是句子的流畅性，本文选择 B-1 和 B-4 作为指标；METEOR 同时考虑精确率、召回率和对齐率，弥补了 BLEU 的不足；ROUGE_L 用于评估生成句子的灵活性和充分性；SPICE 对场景图的对象、属性和关系进行量化，其考虑了候选描述的语义内容，而不是语法及句子流畅程度；CIDEr 用于判断图像描述质量，从句子准确性、语法准确度等方面对描述句子进行评估。

3.4 实验结果

3.4.1 模型整体性能

为验证本文模型的有效性，在 MS COCO 数据集上使用一些主流算法模型进行比较，对照模型包括 GCN-LSTM[2]、Up-Down[10]、SCST[17]、RFNet[18]、ORT[19]、GET[20]、SATIC[21]、DLCT[22]、DIFNet[23]、TSG[24]。比较结果如表 1 所示。可以看出，本文模型的大多数指标优于对照模型。其中，与经典的 DLCT 算法相比，本文模型仅有 B-1 指标次之，在 B-4、R、C 和 S 指标上分别高出 0.7、0.8、1.3 和 0.2；与基线网络 SATIC 相比，本文模型所有指标均有较大领先；与最新的 TSG 模型相比，本文模型只有 S 得分次之，其余指标均更优。综上所述，本文改进措施有效。

Table 1 Comparison of experimental results of different models

表 1 不同模型性能指标比较

模型	B-1	B-4	M	R	C	S
SCST	-	34.2	26.7	55.7	114.0	-
RFNet	79.1	36.5	27.7	57.3	121.9	21.2
Up-Down	79.8	36.3	27.7	56.9	120.4	21.6
GCN-LSTM	80.7	38.2	28.6	58.6	128.0	22.3
ORT	80.5	38.5	28.7	58.3	128.4	22.7
GET	81.5	39.5	29.3	58.9	131.6	22.8
SATIC	80.8	39.4	28.6	59.3	132.2	22.2
DLCT	**81.7**	39.8	29.3	58.9	133.8	22.7
DIFNet	**81.7**	40.0	**29.7**	59.4	134.5	22.8
TSG	-	39.8	29.6	59.5	133.1	**23.4**
本文模型	81.6	**40.5**	29.3	**59.7**	**134.8**	22.9

注：最优结果以**粗体**表示。

3.4.2 时间性能

为验证 MViT 的有效性，选择基于 Faster RCNN 与 Transformer 的图像描述生成模型(E2N)进行时间性能比较实验。结果见表 2。可以看出，除 METEOR 得分与 E2N 模型持平外，本文模型其余指标均更优。因为 E2N 模型主体由 Faster R-CNN 和 Transformer 组成，其使用复杂的神经网络对图像进行预处理且原始自注意力机制具有二次计算复杂度，模型训练与推理时间均较长。得益于结构较为简单的 MViT 结构，本文模型的训练和推理时间大幅度降低，分别仅为 31 min/轮和 10 min/轮，由此说明 MViT 能有效改善以往图像描述生成任务依赖于复杂神经网络的现象，提

升模型训练与推理速度。

Table 2　Experimental results of time performance comparison

表 2　时间性能比较实验结果

算法	训练时间（min/轮）	推理时间（min/轮）	B-1	B-4	M	R	C	S
E2N	52	21	80.8	39.6	**29.3**	58.5	131.7	22.6
本文模型	31	10	**81.6**	**40.5**	**29.3**	**59.7**	**134.8**	**22.9**

3.4.3　消融实验

通过消融实验验证 3 个改进模块对模型性能的影响，结果见表 3。其中，Transformer 表示未作修改的模型，+MViT 表示使用多尺度 Vision Transformer，+MA 表示使用记忆注意力，+VLA 表示使用视觉引导注意力。可以看出，单独使用任何一个改进模块均能提高模型性能，且对模型性能的提高程度依次为 VLA＞MA＞MViT；同时使用两个改进模块可进一步提高模型性能，同时使用 3 个改进模块时模型性能达到最佳，说明每个改进都很有必要。

3.4.4　定性分析

从 MS COCO 2014 数据集中随机选取 3 张图片进行定性实验，每张图片对应 1 条人工标注语句（GT）、1 条 Transformer 基线模型生成的描述语句以及 1 条本文模型生成的描述语句。具体比较情况见表 4。本文模型对图片 1 的描述为“地板上放着一个红色花瓶，上面装饰着长长的粉红色雏菊”，Transformer 基线模型生成的描述为“地板上的红色容器中的鲜花”，很明显基线模型缺少“长长的、粉红色”等细节描述，且没有识别出花的名称。本文模型对图片 2 中的描述为“一群骑自行车的人在自行车道上骑行”，Transformer 基线模型生成的描述为“一群人正在骑车”。基线模型生成的句子虽然正确，但本文模型识别到的信息更加完整，场景描述更准确。本文模型能准确识别出图 3 中的吃饭行为，而 Transformer 不能。综上所述，本文模型不仅可以更加准确地识别出图像中的细节与背景信息，而且语义表达更加清晰，同时表述方式也更贴近人类对图像的描述方式。

Table 3　Ablation experiment result

表 3　消融实验结果

模型	B-1	B-4	M	R	C	S
Transformer	80.4	37.9	28.7	58.4	131.1	22.1
+MViT	80.5	38.0	28.7	58.3	132.2	22.2
+MA	80.6	38.1	28.8	58.5	132.4	22.3
+VLA	81.0	38.9	29.0	58.8	133.0	22.6
+MViT+MA	81.2	39.0	29.1	59.1	132.5	22.7
+MViT+VLA	81.4	39.3	29.2	59.3	133.8	22.7
+MA+VLA	81.5	40.3	**29.3**	59.4	134.0	22.8
+MViT+MA+VLA	**81.6**	**40.5**	**29.3**	**59.7**	**134.8**	**22.9**

Table 4　Qualitative analysis results

表 4　定性分析结果

序号	图片	模型	描述文本
1		GT	a red vase with long pink daisies set on a floor
		Transformer	flowers in a red vase on the floor
		本文模型	on the floor sits a red vase adorned with long pink daisies
2		GT	a group of bicyclists are riding in the bike lane
		Transformer	a group of people are cycling
		本文模型	a group of bicyclists are riding in the bike lane
3		GT	a group of people are sitting around a dinner table
		Transformer	a group of people are sitting around a table
		本文模型	a group of people sitting at a table，having dinner

4　结语

本文提出一种基于改进 Transformer 的多尺度图像描述生成模型，摆脱了对复杂神经网络的依赖，提高了模型训练与推理速度，同时提升了描述语句生成质量，在 MS COCO 数据集上的比较实验和消融实验验证了本文模型及各改进模块的有效性。该模型可应用于自动生成医学影像诊断报告、自动驾驶等领域，具有较大的人机交互潜力。本文主要在编码器端降低模型复杂度，但解码器端的模型依旧存在二次复杂度，会影响模型的训练与推理速度。后续研究拟改进解码端注意力机制，降低模型复杂度，进一

步提升模型的训练与推理速度。

参考文献：

[1] SUN C, GAN C, NEVATIA R. Automatic concept discovery from parallel text and visual corpora[C]//IEEE International Conference on Computer Vision, 2015: 2596-2604.

[2] VYAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning[C]//15th European Conference on Computer Vision, 2018: 711-727.

[3] YU Y W, SHI S C, WANG H J. Image caption based on Bert word vectors and ordered memory network[J]. Software Guide, 2023, 22(3): 125-133.
俞艺文, 施水才, 王洪俊. 基于Bert词向量与有序记忆网络的图像描述[J]. 软件导刊, 2023, 22(3): 125-133.

[4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [DB/OL]. https://arxiv. org/pdf/1409. 1556. pdf.

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

[6] TAN M X, LE O V. EfficientNet: rethinking model scaling for convolutional neural networks[DB/OL]. https://arxiv. org/pdf/1905. 11946. pdf.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[8] CHO K, AN B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. Computer Science, 2014, 21(4): 62-72.

[9] 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.

[10] XU K, BA J L, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015: 2048-2057.

[11] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 6077-6086.

[12] 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 and Machine Intelligence, 2017, 39(6): 1137-1149.

[13] YANG Y, FANG X L, SHANG J, et al. Image caption generation method fused with visual commonsense and enhanced multi-layer global features [P]. China, 202110642157. 4, 2021-09-10.
杨有, 方小龙, 尚晋, 等. 融合视觉常识和增强多层全局特征的图像描述生成方法[P]. 中国, 202110642157. 4, 2021-09-10.

[14] XIAN T, LI Z, ZHANG C, et al. Dual global enhanced Transformer for image captioning[J]. Neural Networks, 2022, 148: 129-141.

[15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Zurich: 13th European Conference on Computer Vision, 2014: 740-755.

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

[17] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1179-1195.

[18] JIANG W H, MA L, JIANG Y G, et al. Recurrent fusion network for image captioning[C]//Munich: 15th European Conference on Computer Vision, 2018: 510-526.

[19] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: transforming objects into words[C]//Annual Conference on Neural Information Processing Systems, 2019: 11135-11145.

[20] JI J, LUO Y, SUN X, et al. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network[C]//Proceedings of the AAAI, 2021: 1655-1663.

[21] ZHOU Y, ZHANG Y, HU Z, et al. Semi-autoregressive transformer for image captioning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 3139-3143.

[22] LUO Y, JI J, SUN X, et al. Dual-level collaborative transformer for image captioning[C]//Thirty-Fifth AAAI Conference on Artificial Intelligence, 2021: 2286-2293.

[23] WU M R, ZHANG X Y, SUN X S, et al. DIFNet: boosting visual information flow for image captioning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 17999-18008.

[24] YANG X, PENG J, WANG Z, et al. Transforming visual scene graphs to image captions[DB/OL]. https://arxiv. org/abs/2305. 02177.

（责任编辑：尹晨茹）