基于 Transformer 视觉特征融合的图像描述方法

白雪冰[1,2]，车进[2,3*]，吴金蔓[2,3]，陈玉敏[2,3]

（1. 宁夏大学前沿交叉学院，宁夏 中卫 755000；2. 宁夏大学宁夏沙漠信息智能感知重点实验室，宁夏 银川 750021；
3. 宁夏大学电子与电气工程学院，宁夏 银川 750021）

摘 要：现有图像描述方法只利用区域型视觉特征生成描述语句，忽略了网格型视觉特征的重要性，并且均为两阶段方法，从而影响了图像描述的质量。针对该问题，提出一种基于 Transformer 视觉特征融合的端到端图像描述方法。首先，在特征提取阶段，利用视觉特征提取器提取出区域型视觉特征和网格型视觉特征；其次，在特征融合阶段，通过视觉特征融合模块对区域型视觉特征和网格型视觉特征进行拼接；最后，将所有的视觉特征送入语言生成器中以生成图像描述。该方法各部分均基于 Transformer 模型实现，实现了一阶段方法。在 MS-COCO 数据集上的实验结果表明，所提方法能够充分利用区域型视觉特征与网格型视觉特征的优势，BLEU-1、BLEU-4、METEOR、ROUGE-L、CIDEr、SPICE 指标分别达到 83.1%、41.5%、30.2%、60.1%、140.3%、23.9%，优于目前主流的图像描述方法，能够生成更加准确和丰富的描述语句。

关键词：图像描述；区域型视觉特征；网格型视觉特征；Transformer 模型；端到端训练

源代码链接：https://github.com/auh-12/TVFF

中图分类号：TP391.41 **文献标志码**：A **DOI**：10.19678/j.issn.1000-3428.0068402

Image Captioning Method Based on Transformer Visual Features Fusion

BAI Xuebing[1,2], CHE Jin[2,3*], WU Jinman[2,3], CHEN Yumin[2,3]

(1. School of Advanced Interdisciplinary, Ningxia University, Zhongwei 755000, Ningxia, China;
2. Ningxia Key Laboratory of Intelligent Sensing for Desert Information, Ningxia University, Yinchuan 750021, Ningxia, China;
3. School of Electronic and Electrical Engineering, Ningxia University, Yinchuan 750021, Ningxia, China)

【Abstract】 Existing image captioning methods only use regional visual features to generate description statements and ignore the importance of grid visual features. Moreover, as these methods are two-stage approaches, image captioning quality is affected. To address this issue, this study proposes an end-to-end image captioning method based on the visual feature fusion of Transformer. First, in the feature extraction stage, the visual feature extractor is used to extract regional and grid visual features. Second, in the feature fusion stage, the regional and grid visual features are concatenated using a visual feature fusion module. Finally, the visual features are sent to the language generator to realize image captioning. All components of the method are implemented based on the Transformer model, which is a one-stage method. The experimental results on the MS-COCO dataset show that the proposed method can fully utilize the respective advantages of regional and grid visual features, with the BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE metrics reaching 83.1%, 41.5%, 30.2%, 60.1%, 140.3%, and 23.9%, respectively, indicating that the proposed method is superior to mainstream image captioning methods and can generate more accurate and rich description statements.

【Key words】 image captioning; regional visual features; grid visual features; Transformer model; end-to-end training

0 引言

图像描述是一个跨领域、多模态的交叉研究方向，其利用计算机视觉生成图像特征，利用自然语言处理生成描述内容，在自动翻译、智能监控、智能搜索等领域具有极高的研究价值。

随着深度学习的飞速发展，目前的图像描述方法主要采用基于深度学习的“编码器-解码器”模型。编码器的功能可描述为视觉特征提取器，它利用更快区域卷积神经网络（Faster R-CNN）、卷积神经网络（CNN）等提取图像视觉特征。解码器的功能可描述为语言生成器，它利用长短期记忆（LSTM）网络、循环神经网络（RNN）等生成自然语言描述[1-5]。

收稿日期：2023-09-17 **修回日期**：2023-11-21

基金项目：国家自然科学基金（62366042）；宁夏自然科学基金（2023AAC03127）。

通信作者 E-mail：* koalache@126.com

由于视觉特征提取器获取的视觉特征直接影响着图像描述的质量[6-8]，因此在以前的研究中图像描述任务主要集中在网格型视觉特征(在均分的网格上提取的视觉特征，通常使用 CNN 或者 Vision Transformer 等实现)，该特征虽然能覆盖图像信息，但是缺少区域对象信息，导致图像描述细粒度欠缺。为解决这一问题，区域型视觉特征(物体检测模型检测到的有代表性的对象区域特征，通常使用 Faster R-CNN 等实现)成为目前图像描述任务的研究热点[9-10]，该特征能展现出图像中的物体信息，但是其只利用区域特征，会使物体之间的背景信息缺失。此外，目前主流的图像描述方法大多为两阶段的形式，即将视觉特征提取器和语言生成器分开训练，虽然该方法展现出了良好的性能，但是视觉特征提取器和语言生成器所采用的数据以及训练方式完全不同，导致模型的性能很难实现新突破。一阶段模型可以解决视觉特征提取器和语言生成器无法统一训练的问题，有着极大的研究价值。

本文提出一种完全采用 Transformer 模型[11]进行特征提取与融合的图像描述方法。视觉特征提取器将提取的网格型视觉特征、区域型视觉特征以及两者融合后的视觉特征一起送入语言生成器，生成最终的图像描述。所提方法在视觉特征提取器和语言生成器中均采用 Transformer 模型，从而实现一阶段模型。

1 相关工作

1.1 图像描述方法

RNN 在自然语言处理领域取得了优秀表现，因此，研究人员将其应用于图像描述任务。文献[12]首次将机器翻译框架应用于图像描述，提出了采用 CNN 提取图像视觉特征、利用 RNN 生成图像描述的 m-RNN 模型。为了获得更好的解码能力，文献[13]使用 LSTM 代替 RNN 作为语言生成器，提出了经典的 NIC 模型，这种结构在图像描述任务中取得了良好的性能，此后大量的研究人员一直基于此框架进行研究。

然而，CNN 只能提取出网格型视觉特征，导致语言生成器无法对图像的部分区域进行精准解析。随着目标检测技术的发展，研究人员开始在图像描述任务中使用区域型视觉特征。文献[14]利用目标检测器 Faster R-CNN 提取图像区域型视觉特征，经过自上而下的注意力机制动态分配区域权重以生成图像描述。为了提升图像描述的细粒度表达，文献[15]将 Faster R-CNN 提取的图像视觉信息和场景语义信息结合起来共同指导图像描述生成。文献[16]利用视觉区域聚合模块形成紧凑的视觉区域表征，再利用交叉注意力机制学习更有代表性的语义信息，进一步提高了图像描述性能。

目前的图像描述方法多采用区域型视觉特征与辅助信息相结合的形式，忽略了网格型视觉特征在上下文信息关联方面的作用，使得图像描述在完整性、连贯性方面仍有不足。

1.2 Transformer 模型

Transformer 模型在计算机视觉和自然语言处理领域取得了巨大成功，由于其内部的自注意力机制和交叉注意力机制可以充分实现特征交互，因此越来越多的研究人员开始在图像描述任务中使用 Transformer 模型。文献[17]通过高阶的模态内和模态间交互，增强输出特征的代表性。文献[18]在多头注意力机制中融入相对位置信息，改善了视觉特征之间的方向感知。文献[19]发现网格型视觉特征的准确率与区域型视觉特征相近，但前者运行速度却快后者一个量级。文献[20]在提取出网格型视觉特征的基础上，将分割特征作为附加的视觉信息源，增强了视觉内容的预测。

然而，上述方法均为两阶段形式，即先利用预训练好的视觉特征提取器提取出图像的视觉特征，再单独训练语言生成器。此类方法无法在图像描述中实现视觉特征提取器和语言生成器网络架构的完全统一，影响了图像描述的性能。

2 模型设计

为充分发挥区域型视觉特征和网格型视觉特征的优点，使整体模型成为一阶段模型，本文提出了基于 Transformer 视觉特征融合的端到端图像描述方法，模型整体框架如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。从图 1 可看出，本文模型分为视觉特征提取器和语言生成器 2 个模块。首先，将一张原始图像送入骨干网络，通过融合不同阶段的特征图，形成多尺度特征图；然后，把多尺度特征图分别送入 Deformable DETR (Deformable Detection Transformer)的解码器和 Transformer 编码器中，得到区域型视觉特征和网格型视觉特征；随后，通过视觉特征融合模块得到融合后的视觉特征；最后，将区域型视觉特征、网格型视觉特征以及融合后的视觉特征一起送入 Transformer 解码器中，生成图像描述。

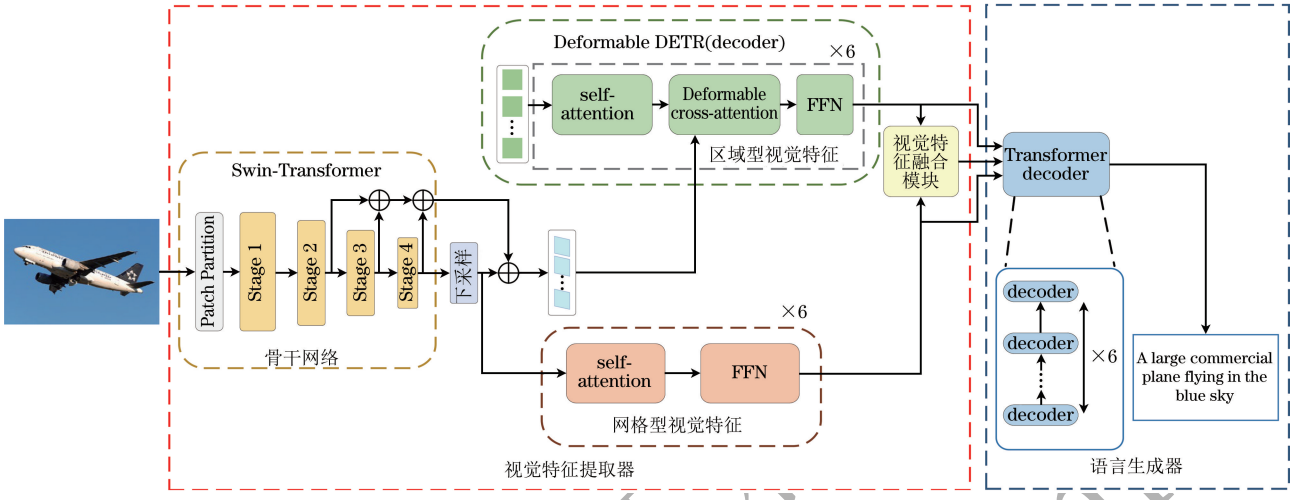

图 1 模型的整体框架

Fig. 1 Overall framework of the model

2.1 视觉特征提取器

2.1.1 骨干网络

图像描述是“图像→图像视觉特征→图像描述”的过程，因此，图像视觉特征提取在图像描述中尤为重要。与 CNN 架构的模型相比，Transformer 架构的模型更善于捕捉特征之间的交互信息，具有更优的性能表现。因此，本文采用 Transformer 模型的变体 Swin Transformer 作为骨干网络，骨干网络结构如图 2 所示。

在图 2 中，Patch Partition 模块将输入的图像在通道方向上进行展平，图像的形状从 $H \times W \times 3$ 变为 $(H/4) \times (W/4) \times 48$，$H$ 和 W 分别为图像的高度和宽度。随后，通过 4 个阶段得到不同的特征信息，其中，阶段 1 先通过 Linear Embedding 层，而剩下 3 个阶段先通过 Patch Merging 层，使特征图的尺寸变为输入的一半，通道数变为输入的 2 倍。除此之外，每一个阶段均包含 Swin Transformer Block，它的结构与 Transformer Block 类似，不同之处是多头自注意力(MSA)换成窗口多头自注意力(W-MSA)和移动窗口多头自注意力(SW-MSA)，使得非局域窗口间可以实现交互和联系，大幅提升了模型的特征信息表达能力。

为了充分利用不同尺度的图像特征信息，本文模型分别提取骨干网络中的第 2、第 3、第 4 阶段的输出特征图，并得到第 4 阶段经过下采样的特征图，再把提取出的 4 个部分特征图进行合并，合并后的特征图用于生成区域型视觉特征和网格型视觉特征。

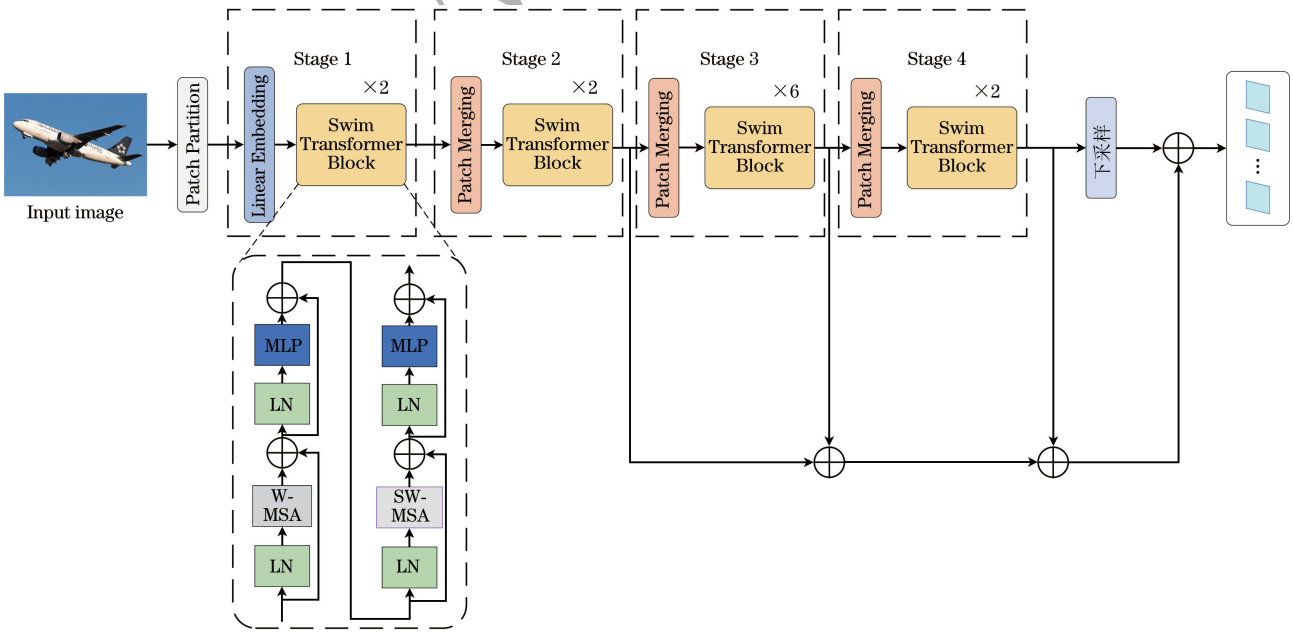

图 2 骨干网络结构

Fig. 2 Backbone network structure

2.1.2 区域型视觉特征

本文使用 Deformable DETR 的变体 Deformable Transformer 提取区域型视觉特征。因为在骨干网络部分已经提取出图像的多尺度特征图，所以此部分仅使用 Deformable Transformer 的解码器。提取区域型视觉特征的网络架构如图 3 所示。

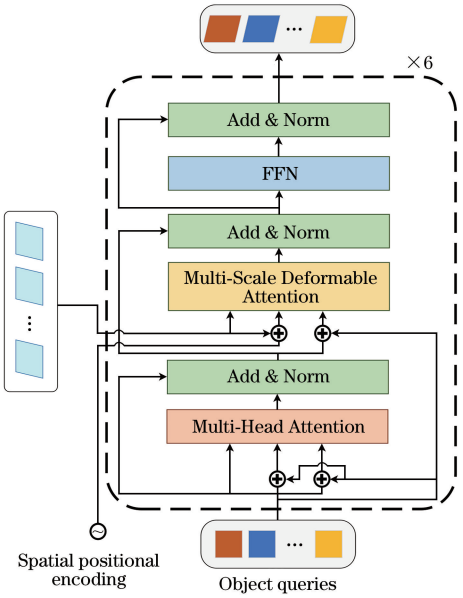

图 3 提取区域型视觉特征的网络架构

Fig.3 Network architecture for extracting regional visual features

由图 3 可知，Deformable Transformer 的解码器共有 6 层，第 2～第 6 个解码器都使用前一层解码器的输出作为输入。解码器首先随机初始化 Z 个可学习的 Object queries，经过多头自注意力层和残差连接 & 归一化层处理后，进入多尺度可变形注意力层。在多尺度可变形注意力层中，输入端包括前一层的输出、骨干网络生成的多尺度特征图、空间位置编码以及 Object queries。为了保持维度一致，骨干网络生成的多尺度特征图需要进行线性变换，通过可学习的投影矩阵将其映射为 d 维。此处的空间位置编码 $\boldsymbol{P}_{2i}$、$\boldsymbol{P}_{2i+1}$ 分别使用正弦和余弦三角函数，如式(1)、式(2)所示：

$$\boldsymbol{P}_{2i}=\sin\left(\frac{\boldsymbol{p}_{\mathrm{os}}}{10\ 000^{2i/d}}\right) \tag{1}$$

$$\boldsymbol{P}_{2i+1}=\cos\left(\frac{\boldsymbol{p}_{\mathrm{os}}}{10\ 000^{2i/d}}\right) \tag{2}$$

式中：$\boldsymbol{p}_{\mathrm{os}}$ 表示词向量的序列位置；i 是维度的索引。

多尺度可变形注意力层的计算公式如式(3)所示：

$$\boldsymbol{M}_A=\sum_{m=1}^{M}\boldsymbol{w}'_m\left[\sum_{l=1}^{L}\sum_{k=1}^{K}A_{mlqk}\cdot\boldsymbol{w}_m\boldsymbol{x}_l(\varphi_l(\boldsymbol{p}_q)+\Delta\boldsymbol{p}_{mlqk})\right] \tag{3}$$

式中：$\boldsymbol{x}_l$ 表示输入的多尺度特征图；$\boldsymbol{p}_q$ 表示查询元素的归一化二维坐标；$\varphi_l(\boldsymbol{p}_q)$ 表示将归一化的坐标重新映射到对应层的坐标；$\Delta\boldsymbol{p}_{mlqk}$ 表示采样的位置偏移量，用二维坐标表示；A_{mlqk} 表示注意力权重；$\boldsymbol{w}'_m$ 和 $\boldsymbol{w}_m$ 表示 2 组全连接层的权重参数；M 表示多头注意力机制中头的数目；L 表示多尺度金字塔维度；K 表示每一个头中考虑的附近像素点的个数；$\boldsymbol{M}_A$ 为多尺度可变形注意力层的计算结果。

多尺度可变形注意力层的输出再经过 2 个残差连接 & 归一化层和 1 个前馈神经网络层后，便可输出图像的区域型视觉特征 $\boldsymbol{R}_g$。

2.1.3 网格型视觉特征

本文使用 Transformer 编码器提取网格型视觉特征，网络架构如图 4 所示。

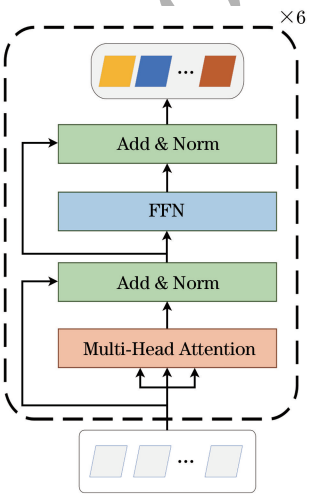

图 4 提取网格型视觉特征的网络架构

Fig.4 Network architecture for extracting grid based visual features

从图 4 可以看出，网络架构共有 6 层，第 2～第 6 层都使用前一层的输出作为输入。网络架构的输入端为骨干网络最后一层的多尺度特征图。首先通过可学习的投影矩阵将最后一层的多尺度特征图映射为 d 维，再经过多头自注意力层、残差连接 & 归一化层以及前馈神经网络层，得到图像的网格型视觉特征 $\boldsymbol{G}_r$。

2.1.4 视觉特征融合模块

为了在语言生成器中充分发挥区域型视觉特征和网格型视觉特征的作用，本文通过视觉特征融合模块，将区域型视觉特征 $\boldsymbol{R}_g$ 和网格型视觉特征 $\boldsymbol{G}_r$ 直接进行拼接，得到融合后的视觉特征 $\boldsymbol{F}_u$。

2.2 语言生成器

本文在语言生成器部分使用 Transformer 解码器输出图像描述内容，Transformer 解码器共有 6 层，语言生成器结构如图 5 所示。

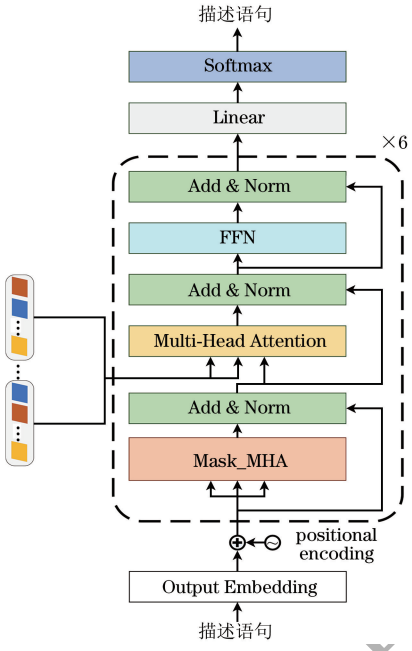

图 5 语言生成器架构

Fig. 5 Language generator architecture

在进入语言生成器之前，首先将每张图像的描述语句通过添加开始标记<s>和结束标记<e>的方式进行预处理，然后用分词工具对整个文档进行分词，每个词对应一个编号，并以字典的形式存储编号和词。在语言生成器中，描述语句通过查找字典确定对应的数字序列，再通过嵌入层转换成词嵌入向量，最后添加位置编码后输入到 Transformer 解码器中。

在 Transformer 解码器中，词嵌入矩阵 $\boldsymbol{X}$ 首先通过掩码多头注意力模块，其中，掩码通过上三角矩阵实现，它起到了遮挡当前词之后信息的作用。接着将掩码多头注意力模块的输出送入残差连接 & 归一化层中得到注意力特征矩阵 $\boldsymbol{D}_m$，如式(4)所示：

$$\boldsymbol{D}_m=\text{Norm}(\boldsymbol{X}+\text{Mask\_MHA}(\boldsymbol{X},\boldsymbol{X},\boldsymbol{X})) \tag{4}$$

在多头注意力模块部分，由于输入端既包括前一层的输出，又包括区域型视觉特征、网格型视觉特征以及融合后的视觉特征。因此，原始 Transformer 解码器中的多头注意力模块需要进行调整，调整后的多头注意力模块如图 6 所示。

从图 6 可以看出，3 个并行的多头注意力模块输入前后操作相同。以区域型视觉特征为例，在多头注意力模块的输入端，$\boldsymbol{D}_m$ 为查询向量，$\boldsymbol{R}_g$ 为键向量和值向量，经过多头注意力模块后，得到 $\boldsymbol{c}_r$，然后将其与 $\boldsymbol{D}_m$ 拼接，再通过可学习的投影矩阵将其映射为 d 维，并使用 Sigmoid 函数实现归一化，得到 $\boldsymbol{s}_r$。以同样的方式可以得到网格型视觉特征归一化输出 $\boldsymbol{s}_g$、融合视觉特征归一化输出 $\boldsymbol{s}_m$。将 3 个归一化输出结果分别与多头注意力模块的输出结果相乘，并将相乘结果相加，得到多头注意力模块的输出 $\boldsymbol{c}_l$，如式(5)所示：

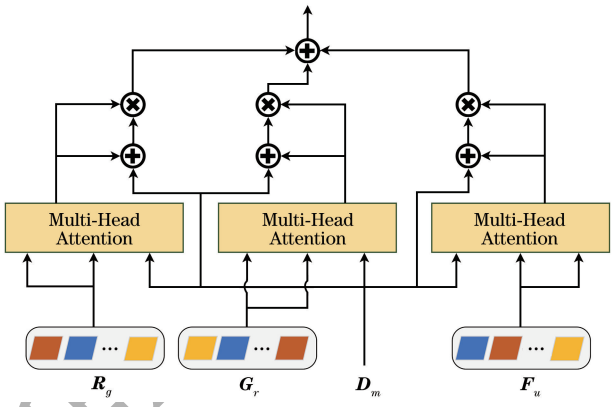

图 6 调整后的多头注意力模块

Fig. 6 Adjusted multi-head attention module

$$\boldsymbol{c}_l=\boldsymbol{s}_r\otimes\boldsymbol{c}_r+\boldsymbol{s}_g\otimes\boldsymbol{c}_g+\boldsymbol{s}_m\otimes\boldsymbol{c}_m \tag{5}$$

式中：$\boldsymbol{c}_g$ 和 $\boldsymbol{c}_m$ 分别为网格型视觉特征与融合视觉特征经过多头注意力层的输出结果。

在上述结果的基础上，经过残差连接 & 归一化层与前馈神经网络层后，得到解码器的输出。最后，将解码器的输出送入线性层和 Softmax 分类层，得到当前时间点的输出单词。

2.3 模型训练

与主流算法相同，本文的训练过程由 2 个阶段组成：第 1 阶段采用交叉熵损失函数对模型进行训练；第 2 阶段采用强化学习的策略对模型进行进一步优化。

在第 1 阶段，图像的真实描述表示为 $\{y_1,y_2,\cdots,y_T\}$，用 θ 表示模型中的参数，最小化模型的交叉熵损失值 $N_{\text{XE}}(\theta)$，如式(6)所示：

$$N_{\text{XE}}(\theta)=-\sum_{t=1}^{T}\log_a(p_\theta(y_t\mid y_{1:t-1})) \tag{6}$$

式中：T 为描述语句的长度。

交叉熵损失函数容易实现，方便训练，但存在训练目标和评估指标不一致等问题。为了得到更好的生成结果，本文遵循 SCST[21] 的方法，将 CIDEr 评价指标作为奖励，在训练过程中引入强化学习，通过减少负奖励期望进行优化，如式(7)所示：

$$N_{\text{RL}}(\theta)=-\mathbb{E}_{y_{1,T}\sim p_\theta}[r(y_{1:T})] \tag{7}$$

式中：$N_{\text{RL}}(\theta)$ 为损失值；$\mathbb{E}$ 为期望；$r(\cdot)$ 表示 CIDEr 的得分；$y_{1:T}$ 是使用蒙特卡洛采样的方法后模型依据 p_θ 生成的描述语句。

结合式(7)，根据 SCST 方法可以求得梯度损失，如式(8)所示：

$$\nabla_\theta N_{RL}(\theta)=$$
$$-\frac{1}{n}\sum_{t=1}^{T}(r(y_{1:T})-b)\nabla_\theta\log_a(p(y_t)) \quad (8)$$

式中：b 表示奖励分数；n 为序列数量。

基于强化学习的训练方法以 CIDEr 评价指标作为优化目标，解决了交叉熵损失在训练中存在的问题，能够大幅提升模型性能。

3 实验结果与分析

3.1 数据集与评价指标

本文实验采用 MS-COCO 2014[22]数据集，该数据集共 123 287 张图像，每张图像均包含 5 个图像描述语句。本文按照文献[23]所提供的数据集划分方式，将数据集划分为包含 113 287 张图像的训练集、包含 5 000 张图像的验证集和包含 5 000 张图像的测试集。

在测试阶段，本文采用常见的 BLEU[24]、METEOR[25]、ROUGE-L[26]、CIDEr[27]、SPICE[28]评价指标对图像描述质量进行评估。BLEU 包含 BLEU-1(简写为 B1)、BLEU-4(简写为 B4)，用来评价模型输出语句与参考语句之间一元组和四元组词汇重叠程度；METEOR(简写为 M)评价语句的流畅度；ROUGE-L(简写为 R)评价文本摘要质量；CIDEr(简写为 C)评价生成句子与参考语句之间的语义相似度；SPICE(简写为 S)评价生成的描述是否有效地描述了图像中的对象、属性及它们之间的关系。

3.2 实验设置

实验在 Windows 10 操作系统上进行，实验代码基于深度学习框架 PyTorch 1.13 版本和 Python 3.9 编写。GPU 为 GeForce RTX 3090(24 GB)，CPU 为 Intel® Core™ i5-12400F(32 GB)。特征维度 $d=512$，可学习的 Object queries 数量 $Z=100$，多尺度金字塔维度 $L=4$，多头注意力机制中头的数目 $M=8$，每一个头中考虑的附近像素点的个数 $K=4$。

Deformable Transformer 解码器采用文献[29]的预训练模型，总的训练周期为 30：前 10 轮使用交叉熵损失函数进行训练，批量大小为 8，第 1 轮训练学习率为 1×10^{-4}，此后学习率为 1×10^{-5}；后 20 轮使用强化学习方法进行训练，批量大小为 4，学习率为 5×10^{-6}。2 个阶段的参数优化均使用 Adam 优化器，描述语句生成采用集束搜索方式，束大小为 5。

3.3 定量分析

为了验证本文模型的效果，将本文模型与近几年出现的主流两阶段模型进行对比，包括 Adaptive[30]、SCST[21]、Up-Down[14]、VRCDA[31]、GCN-LSTM[32]、AoANet[9]、M^2 Transformer[33]、GET[34]、X-Transformer[17]、DRT[18]、RSTNet[35]、DLCT[36]以及 DIFNet[20]。Adaptive 和 SCST 使用的视觉特征通过 ResNet101 提取，RSTNet、DLCT 和 DIFNet 的视觉特征通过 ResNeXt101 提取，这 4 个模型均使用网格型视觉特征。剩余模型均使用通过 Faster R-CNN 提取出的区域型视觉特征。表 1 列出了本文模型与主流模型的对比结果，所有数值均以百分数呈现，最优结果加粗标注，表中的“—”代表原论文没有该评价指标得分。

表 1 本文模型与其他主流模型的性能比较

Table 1 Comparison of performance between the model in this paper and other mainstream models %

模型	B1	B4	M	R	C	S
Adaptive[30]	74.2	33.2	26.6	—	108.5	—
SCST[21]	—	34.2	26.7	55.7	114.0	—
Up-Down[14]	79.8	36.3	27.7	56.9	120.1	21.4
VRCDA[31]	80.6	37.9	28.4	58.2	123.7	21.8
GCN-LSTM[32]	80.9	38.3	28.6	58.5	128.7	22.1
AoANet[9]	80.2	38.9	29.2	58.8	129.8	22.4
M^2 Transformer[33]	80.8	39.1	29.2	58.6	131.2	22.6
GET[34]	81.5	39.5	29.3	58.9	131.6	22.8
X-Transformer[17]	80.9	39.7	29.5	59.1	132.8	23.4
DRT[18]	81.7	40.4	29.5	59.3	133.2	23.3
RSTNet[35]	81.1	39.3	29.4	58.8	133.3	23.0
DLCT[36]	81.4	39.8	29.4	59.1	133.8	23.0
DIFNet[20]	81.7	40.0	29.7	59.4	136.2	23.2
本文模型	**83.1**	**41.5**	**30.2**	**60.1**	**140.3**	**23.9**

由表 1 可知，本文模型在所有指标上都获得了显著地提升，具有很大的性能优势。与性能次优的 DIFNet 相比，本文模型在 CIDEr 指标上提升了 4.1 个百分点，说明本文模型生成的描述语句与参考语句具有极高的语义相似度。

表 2 列出了本文模型与主流模型在 MS-COCO 官网上的在线测试结果对比，表中的 c5 和 c40 表示一张图像包含的描述语句个数分别为 5 和 40。由表 2 可以看出，本文模型在所有评价指标上均展现出了良好的性能。其中，本文模型在 CIDEr(c5/c40)评价指标上的得分为 135.4%/137.9%，优于其他主流模型，进一步证明了本文模型的有效性。

表 2 本文模型与其他主流模型在在线测试集上的性能比较

Table 2 Comparison of performance between the model in this paper and other mainstream models on online test sets %

模型	B1		B4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Adaptive[30]	74.8	92.0	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
SCST[21]	78.1	93.7	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down[14]	80.2	95.2	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
VRCDA[31]	80.4	94.8	37.8	69.2	28.3	37.3	58.2	73.3	123.2	125.6
GCN-LSTM[32]	80.8	95.2	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
AoANet[9]	81.0	95.0	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M^2 Transformer[33]	81.6	96.0	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
GET[34]	81.6	96.1	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
X-Transformer[17]	81.3	95.4	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
DRT[18]	82.7	96.5	**40.9**	73.6	29.6	39.0	**59.8**	**75.0**	132.2	133.9
RSTNet[35]	81.7	96.2	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
DLCT[36]	82.0	96.2	40.2	73.2	29.5	39.1	59.4	74.8	131.0	133.4
DIFNet[20]	82.1	96.4	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
本文模型	**82.8**	**96.9**	40.8	**74.3**	**30.0**	**39.6**	**59.8**	**75.0**	**135.4**	**137.9**

实验除了对图像描述生成的句子进行质量评估外，还对本文模型与 RSTNet、DLCT、DIFNet 进行模型参数量、计算量以及推理时间的比较，实验结果如图 7 所示。

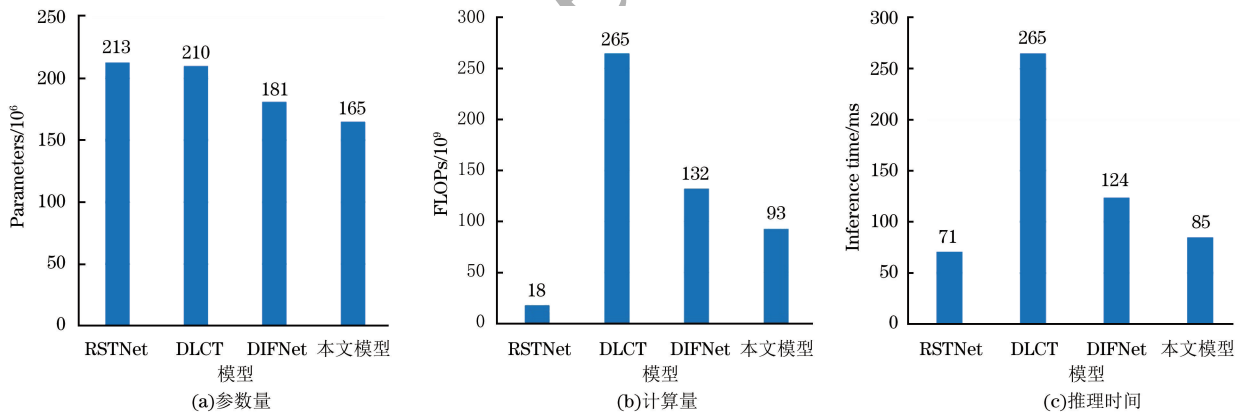

图 7 参数量、计算量、推理时间的比较结果

Fig.7 Comparison results of parameter quantity, computational complexity, and inference time

由图 7 可知，本文模型参数量最少，且在计算量上仅高于 RSTNet，远低于 DLCT，这说明所提模型复杂度较低。本文模型在推理时间上与 RSTNet 相差 14 ms，远低于另外 2 种模型，但 RSTNet 属于两阶段模型，在实际使用中需要进行数据存取、转换等操作，这些操作将产生额外的耗时，而本文模型不存在上述问题。

3.4 消融实验结果分析

本文通过控制变量的思想，设计不同模型来验证视觉特征融合的有效性。消融实验模型分为：1)只使用区域型视觉特征的模型；2)只使用网格型视觉特征的模型；3)本文所提视觉特征融合的模型。消融实验结果如表 3 所示。从表 3 可以看出，本文模型单独使用区域型视觉特征的性能优于单独使用网格型视觉特征的性能，而融合的视觉特征性能最优。这证明视觉特征融合有助于模型在检测出图像中物体特征的前提下更好地理解视觉对象之间的关系，从而达到特征增强的

效果。

由于本文模型在后 20 轮使用了强化学习方法进行训练，因此进行未使用强化学习方法与使用强化学习方法的对比实验，结果如表 4 所示。

表 3 消融实验结果对比

Table 3 Comparison of ablation experiment results %

模型	区域型视觉特征	网格型视觉特征	B1	B4	M	R	C	S
本文模型	√	×	82.5	41.3	30.1	59.8	139.7	23.6
	×	√	81.3	40.6	29.3	58.9	138.3	23.2
	√	√	**83.1**	**41.5**	**30.2**	**60.1**	**140.3**	**23.9**

表 4 使用强化学习前后的效果对比

Table 4 Comparison of effects before and after using reinforcement learning %

模型	Cross-Entropy Loss	CIDEr Optimization	B1	B4	M	R	C	S
本文模型	√	×	79.2	38.5	28.7	58.3	124.2	21.9
	√	√	**83.1**	**41.5**	**30.2**	**60.1**	**140.3**	**23.9**

从表 4 可看出，使用了强化学习方法后，模型的各项指标均有不同程度的提升，其中，CIDEr 指标上升幅度较大，表明强化学习是直接从 CIDEr 指标出发来优化模型的生成效果，避免了只用交叉熵函数可能引起的陷入局部最优等问题。

3.5 定性分析

本文选取了部分实验结果进行图像描述的直观展示，包含示例图片、本文模型的图像描述以及数据集中人工标注的 5 个句子，实验结果如图 8 所示。

	本文模型：A teddy bear sitting on **a couch** reading a book. 数据集中人工标注的句子： 1. A teddy bear is positioned to read a text book. 2. A teddy bear sits in a chair with an open book. 3. Teddy bear comically reading text book on kitchen table. 4. Teddy bear in a chair with a book opened up in front of it. 5. A brown teddy bear sitting in front of an open book.
	本文模型：An old red truck parked **in front of** a house. 数据集中人工标注的句子： 1. An old red truck sits in a driveway. 2. Old worn red truck parked in a driveway near a cactus. 3. The old truck is parked in the driveway of a house. 4. An old red pickup truck parked in a driveway next to a cactus. 5. A red truck in the drive way of a home.
	本文模型：A man riding a wave on a surfboard **in the ocean**. 数据集中人工标注的句子： 1. A man on a surfboard riding the wave. 2. A man on a surfboard riding a large wave. 3. The surfer is riding a wave in the ocean. 4. A surfer wearing red shirt surfing on a wave. 5. A young man surfing an eight foot wave.
	本文模型：**Two young boys** playing baseball **on a field**. 数据集中人工标注的句子： 1. A picture of a kid picking up a ball. 2. A boy reaches for the ball during a baseball game. 3. Two boys playing baseball run to catch the ball. 4. Two young boys run towards a baseball during a baseball game. 5. The boys are playing baseball in the field.

图 8 本文模型的图像描述与人工标注结果对比

Fig. 8 Comparison between the image description of the model in this paper and the results of manual annotation

在图 8 中，第 1 个示例中的“a couch”、第 2 个示例中的“in front of”以及第 3 个示例中的“in the ocean”在人工标注的句子中未体现，但是在本文模型中均有体现。第 4 个样例的“two young boys”、“on a field”在人工标注的句子中出现，但未同时出现在一个句子中。可以看出，本文模型不仅能识别出图像中的视觉对象，还能够捕捉视觉对象之间的位置关系和图像的详细信息。

4 结束语

本文提出一种基于 Transformer 视觉特征融合的端到端图像描述方法。该方法充分利用区域型视觉特征和网格型视觉特征的优点，在同时提取 2 个视觉特征的基础上，通过视觉特征融合模块进行融合，得到语义更加丰富的视觉信息，引导模型生成更加符合图像内容且质量更高的描述。此外，模型整体基于 Transformer 实现，解决了视觉特征提取器和语言生成器分开训练的问题，实现了一阶段模型。在 MS-COCO 数据集上的实验结果表明，本文方法能够识别到更全面、更准确的物体，并学习它们之间的视觉语义关系，生成更生动、更详细的描述，从而提升图像描述的性能。下一步将尝试在语言生成器中引入语言句法和语义信息，以增强模型的可解释性。

参考文献

[1] LI Z X, LIN L, ZHANG C L, et al. A semi-supervised learning approach based on adaptive weighted fusion for automatic image annotation[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(1): 1-23.

[2] 衡红军，范昱辰，王家亮. 基于 Transformer 的多方面特征编码图像描述生成算法[J]. 计算机工程，2023，49(2)：199-205.
HENG H J, FAN Y C, WANG J L. Multifaceted feature coding image caption generation algorithm based on Transformer[J]. Computer Engineering, 2023, 49(2): 199-205. (in Chinese)

[3] 卓亚琦，魏家辉，李志欣. 基于双注意模型的图像描述生成方法研究[J]. 电子学报，2022，50(5)：1123-1130.
ZHUO Y Q, WEI J H, LI Z X. Research on image captioning based on double attention model[J]. Acta Electronica Sinica, 2022, 50(5): 1123-1130. (in Chinese)

[4] CHEN F L, ZHANG D Z, HAN M L, et al. VLP: a survey on vision-language pre-training[J]. Machine Intelligence Research, 2023, 20(1): 38-56.

[5] YANG X, ZHANG H W, GAO C Y, et al. Learning to collocate visual-linguistic neural modules for image captioning[J]. International Journal of Computer Vision, 2023, 131(1): 82-100.

[6] FENG Y M, LAN L, ZHANG X, et al. AttResNet: attention-based ResNet for image captioning[C]// Proceedings of 2018 International Conference on Algorithms, Computing and Artificial Intelligence. New York, USA: ACM Press, 2018: 1-6.

[7] HOSSAIN M Z, SOHEL F, SHIRATUDDIN M F, et al. A comprehensive survey of deep learning for image captioning[J]. ACM Computing Surveys, 2019, 51(6): 1-36.

[8] WANG D F, HU H F, CHEN D H. Transformer with sparse self-attention mechanism for image captioning[J]. Electronics Letters, 2020, 56(15): 764-766.

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

[10] GUO L T, LIU J, ZHU X X, et al. Normalized and geometry-aware self-attention network for image captioning[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 10327-10336.

[11] 赵敬伟，林珊玲，梅婷，等. 基于 YOLACT 与 Transformer 相结合的实例分割算法研究[J]. 半导体光电，2023，44(1)：134-140.
ZHAO J W, LIN S L, MEI T, et al. Research on instance segmentation algorithm based on YOLACT and Transformer[J]. Semiconductor Optoelectronics, 2023, 44(1): 134-140. (in Chinese)

[12] MAO J H, XU W, YANG Y, et al. Deep captioning with multimodal Recurrent Neural Networks (m-RNN)[EB/OL]. [2023-08-05]. https://arxiv.org/abs/1412.6632.

[13] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 3156-3164.

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

[15] 李志欣，魏海洋，黄飞成，等. 结合视觉特征和场景语义的图像描述生成[J]. 计算机学报，2020，43(9)：1624-1640.
LI Z X, WEI H Y, HUANG F C, et al. Combine visual features and scene semantics for image captioning[J]. Chinese Journal of Computers, 2020, 43(9): 1624-1640. (in Chinese)

[16] 宋井宽，曾鹏鹏，顾嘉扬，等. 基于视觉区域聚合与双向协作的端到端图像描述生成[J]. 软件学报，2023，34(5)：2152-2169.
SONG J K, ZENG P P, GU J Y, et al. End-to-end image captioning via visual region aggregation and dual-level collaboration[J]. Journal of Software, 2023, 34(5): 2152-2169. (in Chinese)

[17] PAN Y W, YAO T, LI Y H, et al. X-linear attention networks for image captioning[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 10968-10977.

[18] SONG Z L, ZHOU X F, DONG L H, et al. Direction relation Transformer for image captioning[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM Press, 2021: 5056-5064.

[19] JIANG H Z, MISRA I, ROHRBACH M, et al. In defense of grid features for visual question answering[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 10264-10273.

[20] WU M R, ZHANG X Y, SUN X S, et al. DIFNet: boosting visual information flow for image captioning[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022:

18020-18029.
[21] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1179-1195.
[22] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[EB/OL]. [2023-08-05]. https://arxiv.org/abs/1405.0312.
[23] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 3128-3137.
[24] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. [S. l.]: ACL, 2001: 311-318.
[25] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[EB/OL]. [2023-08-05]. https://aclanthology.org/W05-0909/.
[26] LIN C Y. ROUGE: a package for automatic evaluation of summaries[EB/OL]. [2023-08-05]. https://aclanthology.org/W04-1013/.
[27] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: consensus-based image description evaluation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 4566-4575.
[28] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE: semantic propositional image caption evaluation[EB/OL]. [2023-08-05]. https://arxiv.org/abs/1607.08822.
[29] NGUYEN V Q, SUGANUMA M, OKATANI T. GRIT: faster and better image captioning Transformer using dual visual features[EB/OL]. [2023-08-05]. https://arxiv.org/abs/2207.09666.
[30] LU J S, XIONG C M, PARIKH D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 3242-3250.
[31] 刘茂福,施琦,聂礼强. 基于视觉关联与上下文双注意力的图像描述生成方法[J]. 软件学报, 2022, 33(9): 3210-3222.
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. (in Chinese)
[32] YAO T, PAN Y W, LI Y H, et al. Exploring visual relationship for image captioning[EB/OL]. [2023-08-05]. https://arxiv.org/abs/1809.07041.
[33] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed-memory Transformer for image captioning[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 10575-10584.
[34] JI J Y, LUO Y P, SUN X S, et al. Improving image captioning by leveraging intra- and inter-layer global representation in Transformer network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1655-1663.
[35] ZHANG X Y, SUN X S, LUO Y P, et al. RSTNet: captioning with adaptive attention on visual and non-visual words[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2021: 15460-15469.
[36] LUO Y P, JI J Y, SUN X S, et al. Dual-level collaborative Transformer for image captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2286-2293.

编辑 吴云芳