文章编号:1672-3961(2025)03-0080-08 DOI:10.6040/j.issn.1672-3961.0.2024.018

融合多尺度视觉和文本语义特征的图像描述生成算法

李丰,文益民*
(桂林电子科技大学计算机与信息安全学院, 广西 桂林 541004)

摘要:为了解决目标检测器预训练数据集与图像描述生成任务数据集存在类别差异导致物体识别错误,以及不同场景样本规模存在差异导致模型对少见场景中对象间关系理解不足的问题,提出融合多尺度视觉和文本语义特征的图像描述生成算法(multi-scale visual and textual semantic feature fusion for image captioning, MVTFF-IC)。多尺度视觉特征融合(multi-scale visual feature fusion, MVFF)模块通过图注意力网络对全局、网格和区域特征进行建模,以获取更具代表性的视觉表征;深度语义融合模块(deep semantic fusion module, DSFM)通过交叉注意力机制整合包含对象关系的文本语义特征,以生成更准确的描述。在微软常见物体场景(Microsoft common objects in context, MSCOCO)数据集上的试验结果表明,MVTFF-IC 基于共识的图像描述评价指标 C_{DIEr} 达到 136.7,优于许多现有的流行算法,能够更准确地捕捉图像中的关键信息,生成高质量的描述。

关键词:图像描述;图注意力网络;视觉特征;文本特征;注意力机制
中图分类号:TP391 **文献标志码**:A
引用格式:李丰,文益民. 融合多尺度视觉和文本语义特征的图像描述生成算法[J]. 山东大学学报(工学版),2025,55(3):80-87.
LI Feng, WEN Yimin. Multi-scale visual and textual semantic feature fusion for image captioning[J]. Journal of Shandong University (Engineering Science), 2025, 55(3):80-87.

Multi-scale visual and textual semantic feature fusion for image captioning

LI Feng, WEN Yimin*
(School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China)

Abstract: To address issues caused by category differences between the pre-training dataset of the object detector and the dataset for the image captioning task, which could lead to object recognition errors, as well as variations in sample sizes across different scenes that could result in the model′s insufficient understanding of relationships between objects in rare scenes, the multi-scale visual and textual semantic feature fusion for image captioning (MVTFF-IC) was proposed. The multi-scale visual feature fusion (MVFF) module modeled global, grid, and regional features using a graph attention network to obtain more representative visual representations. The deep semantic fusion module (DSFM) integrated textual semantic features, including object relationships, through a cross-attention mechanism to generate more accurate descriptions. Experimental results on the Microsoft common objects in context (MSCOCO) dataset showed that MVTFF-IC achieved a consensus-based image description evaluation C_{DIEr} of 136.7, outperforming many popular existing algorithms, demonstrating its ability to capture key information more accurately in images and generate high-quality descriptions.

Keywords: image captioning; graph attention network; visual feature; textual feature; attention mechanism

0 引言

图像描述生成是计算机视觉和自然语言处理交叉领域的一项复杂挑战,旨在使计算机能够理解图像并生成准确的自然语言描述,在多个领域均具有重要价值,如图像检索系统、图像标注及视障人士辅助工具。要生成准确的图像描述,关键是所提模型要能尽可能挖掘图像中的对象和关系,尽可能全面地理解图像。

随着多年发展,图像描述生成的巨大成功得到一系列方法支持。在这些方法中,学者们考虑将许

收稿日期:2024-01-24
基金项目:国家自然科学基金资助项目(62366011);广西重点研发计划资助项目(桂科 AB21220023);广西图像图形与智能处理重点实验室资助项目(GIIP2306);桂林电子科技大学研究生教育创新计划资助项目(2023YCXB11)
第一作者简介:李丰(1997—),男,广东东莞人,硕士研究生,主要研究方向为图像描述生成。E-mail:21032303066@ mails.guet.edu.cn
*通信作者简介:文益民(1969—),男,湖南桃江人,教授,博士生导师,博士,主要研究方向为机器学习与数据挖掘。E-mail:ymwen@ guet.edu.cn

多种视觉特征作为图像描述生成模型的输入。全局特征由于粗粒度的特性,不足以满足现代模型的需求[1]。但有研究指出,全局特征在指导和选择有吸引力的对象和关系方面仍具有一定价值[2]。网格特征因其能够提供更为细致的上下文信息而受到广泛关注[3]。但由于单个网格区域往往无法涵盖完整的对象信息,网格特征也存在局限性。将目标检测器[4]提取的区域特征作为输入的图像描述生成方法[5]逐渐占据主导地位。输入特征主要受目标检测器预训练数据集的影响,该数据集通常与图像描述生成模型所用数据集存在分类类别差异,导致生成图像描述时可能发生物体的误判或漏判等问题。因此,为提高模型对图像视觉信息的理解能力,用于图像描述生成的图像视觉表征有待深入研究与探索。

由于图神经网络(graph neural network, GNN)擅长捕捉实体间复杂的相互作用,基于 GNN 的方法在图像描述领域逐渐受到关注[6-8]。这些方法都通过 GNN 对图像中的对象进行建模,生成更优质的视觉特征,在训练过程中通过真实标注的反馈学习对象间的关系词,深入理解对象间的具体关系。由于数据集分布差异,当某些特定场景的图像样本数量较少时,模型可能无法充分理解这些场景下的关系知识,导致关系推断的准确性下降。

为缓解目标检测器提取的区域特征作为输入时导致物体漏判或错判的问题,更好帮助模型学习对象之间的关系词,本研究建议引入网格特征以提供图像中非目标区域信息,缓解目标检测器预训练数据集的影响,弥补区域特征缺陷,通过全局特征引导网格和区域特征融合。同时,在模型前向传播过程中融入包含对象关系的文本语义信息,指导模型更好地理解图像中对象之间的关系。

本研究提出一种融合多尺度视觉和文本语义特征的图像描述生成算法(multi-scale visual and textual semantic feature fusion for image captioning, MVTFF-IC)。初始阶段涉及提取图像的全局、网格和区域视觉特征,依据文献[9]中跨模态检索技术,检索与图像及其子区域高度相关的文本语义特征。为提升视觉表征质量,提出多尺度视觉特征融合(multi-scale visual feature fusion, MVFF)模块,通过计算全局、网格与区域特征间相似度,构建视觉特征图,利用图注意力网络(graph attention network, GAT)进行建模,以增强视觉特征表达;提出深度语义融合模块(deep semantic fusion module, DSFM),整合描述图像中对象关系的文本语义信息,增强对象间关系描述。2 个模块协同工作以提高模型对图像内容的理解能力,实现更完整、准确的图像描述。

1　相关工作

图像描述生成方法大致分为 3 类[10]:基于模板的方法[11]、基于检索的方法[12]和基于编码器-解码器的方法。基于模板和基于检索的方法为早期的传统方法,前者预先定义一个语言模板,将图像中检测关键的视觉概念填入语言模板,获得完整的句子;后者将图像描述生成视为一项检索任务,将输入图像作为查询,用于从训练集现成的句子中检索出语义相似度最高的句子。这两种方法的共同缺点是不够灵活,生成的句子不自然,不适应图像内容。

目前,图像描述生成任务的主要方法依赖于编码器-解码器架构,将图像作为输入,利用深度学习方法生成自然语言描述。文献[6]提出场景图自动编码器(scene graph auto-encoder, SGAE),将语义场景图映入基于循环神经网络(recurrent neural network, RNN)的编码器-解码器中,利用语言的归纳偏差提升图像描述生成性能;文献[7]提出对象关系 Transformer(object relation Transformer, ORT),基于物体检测器获得区域特征,通过几何注意力机制纳入输入对象之间的空间关系信息;文献[8]提出双图卷积网络(dual graph convolutional networks, Dual-GCN),设计对象级和图像级图卷积网络,对象级图卷积网络挖掘对象间关系,图像级图卷积网络引入相似图像信息加深模型对场景的理解;文献[13]提出图卷积网络与长短时记忆网络结合(graph convolutional networks plus long short-term memory, GCN-LSTM),将语义和空间对象关系整合到图像编码器中;文献[14]提出注意力上的注意力网络(attention on attention network, AoANet),在多头注意力之上提出 Attention on Attention 模块,提炼注意力权重,以便更精确识别对生成当前词最相关的图像区域;文献[15]提出配备 X-线性注意力块的 Transformer(Transformer with X-linear attention blocks, X-Transformer),同时利用空间和通道双线性注意力捕捉输入的单模态或多模态特征之间的二阶交互;文献[16]提出网格记忆 Transformer(meshed-memory Transformer, M^2 Transformer),设计一种记忆增强注意力,通过整合先验知识学习图像区域之间关系的多级表示,在解码阶段使用网状连接利用低级和高级特征。上述方法仅利用区域特征进行建模,但区域特征的应用受目标检测器分类数量限

制及背景信息缺失等问题影响,限制其性能。

为突破区域特征的局限性,文献[1]提出全局增强型 Transformer(global enhanced Transformer, GET),设计全局增强编码器,用于嵌入全局特征,引导模型选择有吸引力的目标区域,对目标区域的关系进行建模,通过全局自适应解码器引导图像描述生成;文献[17]提出双层协作 Transformer(dual-level collaborative Transformer, DLCT),根据网格和区域特征的边界框设计几何对齐图,通过设计的局部抑制交叉注意力模块对边界框相交特征进行建模,减缓网格和区域特征融合带来的冗余和语义噪声问题。以上方法都基于视觉特征进行建模,但图像描述生成是一个视觉语言跨模态任务,视觉特征难以提供明确的对象属性和对象间关系信息。文献[18]提出增强理解与推理能力的图像描述(enhance understanding and reasoning ability for image captioning, EURAIC)算法,利用从图像中检测到的核心物体语义特征引导视觉特征,引入外部知识网络,获取图像固有内容以外的信息。

受上述方法的启发,本研究提出一种多模态综合感知网络,引入一种新颖的多尺度视觉信息和多模态信息融合策略,尽可能挖掘更多有用的视觉特征,同时引入文本语义特征,增强对象之间关系表示,生成更准确的描述。

2 方法

本研究提出一种融合多尺度视觉和文本语义特征的图像描述生成算法 MVTFF-IC,其中 MVFF 模块对齐全局、网格和区域视觉特征,实现特征互补;DSFM 模块引入文本语义特征,增强对象间的关系表示,生成高质量的描述。

2.1 特征提取

给定一张原始图像 $\boldsymbol{I}$,采用预先训练好的目标检测模型提取图像的目标区域特征 $\boldsymbol{F}_{\text{regions}}=[\boldsymbol{r}_1\ \boldsymbol{r}_2\cdots\ \boldsymbol{r}_M]\in\mathbf{R}^{M\times 2\,048}$,其中 $\boldsymbol{r}_{\text{m}}$ 为第 m 个区域特征,$m=1,2,\cdots,M$,M 为检测出的目标区域数量。为学习更丰富的图像场景级特征,本研究采用基于对比语言-图像对预训练(contrastive language-image pre-training, CLIP)模型的图像编码端 CLIP-I 对图像进行编码,得到图像的全局特征 $\boldsymbol{F}_{\text{global}}\in\mathbf{R}^{1\times 2\,048}$。按 5、9 等分的比例切分图像,得到图像子区域,使用 CLIP-I 对图像子区域进行编码,得到图像网格特征 $\boldsymbol{F}_{\text{grids}}=[\boldsymbol{g}_1\ \boldsymbol{g}_2\cdots\ \boldsymbol{g}_N]\in\mathbf{R}^{N\times 2\,048}$,其中 $\boldsymbol{g}_n$ 为第 n 个网格特征,$n=1,2,\cdots,N$,N 为网格特征数量。为实现全局、网格和区域视觉特征的优势互补,本研究设计一个多尺度视觉特征融合模块,通过 GAT 对 3 种视觉特征进行建模,得到提炼后的视觉特征 $\boldsymbol{V}_{\text{F}}$。此外,本研究采用文献[9]提出的跨模态图文检索方法获得图像及子区域图像最相关的文本特征 $\boldsymbol{T}_{\text{F}}=[\boldsymbol{t}_1\ \boldsymbol{t}_2\cdots\ \boldsymbol{t}_K]\in\mathbf{R}^{K\times 512}$,其中 $\boldsymbol{t}_k$ 为第 k 个文本特征,K 为文本特征数量。采用一个深度语义融合模块学习 $\boldsymbol{V}_{\text{F}}$ 和 $\boldsymbol{T}_{\text{F}}$ 之间的关系,生成语义信息更加丰富的特征,送入文献[16]提出的网状解码器中,生成最终的图像描述。本研究提出的 MVTFF-IC 整体框架如图 1 所示。

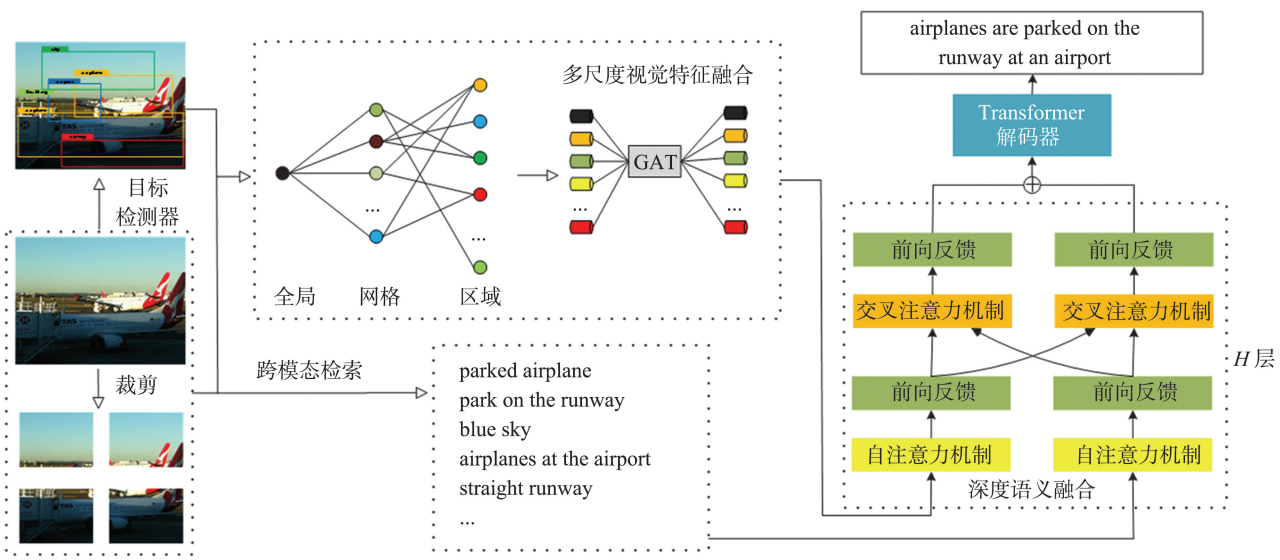

图 1 MVTFF-IC 概览

Fig.1 Overview of the MVTFF-IC

2.2　多尺度视觉特征融合

由于所有网格特征都覆盖在全局特征上,本研究将所有网格特征与全局特征直接相连,即通过全局特征指导视觉特征融合。鉴于网格特征 $\boldsymbol{f}_i$ 和区域特征 $\boldsymbol{f}_j$ 表示的图像区域之间可能没有关联,生硬地将两块不相关的区域进行相连可能会产生语义噪声,因此采用皮尔逊相关系数 $\rho(\boldsymbol{f}_i,\boldsymbol{f}_j)$ 表示二者之间的相关性,决定是否相连,$\rho(\boldsymbol{f}_i,\boldsymbol{f}_j)$ 的计算式为

$$\rho(\boldsymbol{f}_i,\boldsymbol{f}_j)=\frac{\mathrm{cov}(\boldsymbol{f}_i,\boldsymbol{f}_j)}{\sigma(\boldsymbol{f}_i)\times\sigma(\boldsymbol{f}_j)},\tag{1}$$

式中,$\mathrm{cov}(\boldsymbol{f}_i,\boldsymbol{f}_j)$ 为 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$ 的协方差,$\sigma(\boldsymbol{f}_i)$、$\sigma(\boldsymbol{f}_j)$ 分别为 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$ 的标准差。假设邻接矩阵为 $\boldsymbol{A}$,$\boldsymbol{A}\in\mathbf{R}^{M\times N}$。当 $\rho(\boldsymbol{f}_i,\boldsymbol{f}_j)>\Omega$ 时,$\boldsymbol{A}$ 中 (i,j) 位置的元素为 1,表示 $\boldsymbol{f}_i$ 和 $\boldsymbol{f}_j$ 直接相连,其中 Ω 为皮尔逊相关系数的阈值,$\Omega\in[-1,1]$。

得到邻接矩阵 $\boldsymbol{A}$ 和视觉特征 $\boldsymbol{V}_{\mathrm{F}}=[\boldsymbol{F}_{\mathrm{global}}\quad\boldsymbol{F}_{\mathrm{grids}}\quad\boldsymbol{F}_{\mathrm{regions}}]$,令所有视觉特征 $\boldsymbol{V}_{\mathrm{F}}=[\boldsymbol{f}_1\quad\boldsymbol{f}_2\quad\cdots\quad\boldsymbol{f}_L]\in\mathbf{R}^{L\times 2\,048}$,其中 L 为视觉特征数量,维度为 2 048。利用 GAT 进行视觉特征融合,$\boldsymbol{f}_i$ 与 $\boldsymbol{f}_j$ 之间的注意力系数

$$e_{ij}=\mathrm{LeakyReLU}(\boldsymbol{a}^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{f}_i\parallel\boldsymbol{W}\boldsymbol{f}_j]),\tag{2}$$

式中,LeakyReLU 为激活函数,$\boldsymbol{a}$、$\boldsymbol{W}$ 均为可训练参数矩阵,$\parallel$ 为连接操作。使用 softmax 函数对注意力系数进行标准化,得到权重系数

$$c_{ij}=\mathrm{softmax}_j(e_{ij})=\frac{\exp(e_{ij})}{\sum_{p\in N_i}\exp(e_{ip})},\tag{3}$$

式中,e_{ip} 为 $\boldsymbol{f}_i$ 与其邻居节点 $\boldsymbol{f}_p$ 的注意力系数,N_i 为邻接矩阵 $\boldsymbol{A}$ 中 $\boldsymbol{f}_i$ 的邻居节点数。对 $\boldsymbol{f}_i$ 所有相邻的特征 $\boldsymbol{f}_j$ 与相应的权重系数 c_{ij} 进行加权求和,得到更新后的特征表示

$$\boldsymbol{f}'_i=\tau\Big(\sum_{j\in N_i}c_{ij}\boldsymbol{W}\boldsymbol{f}_j\Big),\tag{4}$$

式中 τ 为非线性层。得到融合后的视觉特征

$$\boldsymbol{V}_{\mathrm{F}}=[\boldsymbol{f}'_1\quad\boldsymbol{f}'_2\quad\cdots\quad\boldsymbol{f}'_L]\in\mathbf{R}^{L\times 512}。\tag{5}$$

2.3　深度语义融合模块

为提高模型对于对象之间关系的推理能力,采用交叉注意力机制对视觉特征 $\boldsymbol{M}_{\mathrm{v}}^0=\boldsymbol{V}_{\mathrm{F}}$ 和文本特征 $\boldsymbol{M}_{\mathrm{t}}^0=\boldsymbol{T}_{\mathrm{F}}$ 进行特征融合。将第 $(l-1)$ 层输出的视觉特征 $\boldsymbol{M}_{\mathrm{v}}^{(l-1)}$ 和文本特征 $\boldsymbol{M}_{\mathrm{t}}^{(l-1)}$ 送入第一层的自注意力模块,分别学习两种特征内部关系表示,利用前馈网络进行非线性变换,得到更显著的视觉和文本特征表示,分别为

$$\boldsymbol{S}_{\mathrm{Av}}^l=\mathrm{FFN}(\mathrm{MHAtt}(\boldsymbol{M}_{\mathrm{v}}^{(l-1)},\boldsymbol{M}_{\mathrm{v}}^{(l-1)},\boldsymbol{M}_{\mathrm{v}}^{(l-1)})),\tag{6}$$

$$\boldsymbol{S}_{\mathrm{At}}^l=\mathrm{FFN}(\mathrm{MHAtt}(\boldsymbol{M}_{\mathrm{t}}^{(l-1)},\boldsymbol{M}_{\mathrm{t}}^{(l-1)},\boldsymbol{M}_{\mathrm{t}}^{(l-1)})),\tag{7}$$

式中,FFN 为前馈网络,MHAtt 为多头自注意力机制。随后,通过两个独立的交叉注意力机制,捕捉视觉和文本特征之间复杂的相互作用,得到第 l 层的文本与视觉相互融合后的特征表示,分别为

$$\boldsymbol{M}_{\mathrm{v}}^l=\mathrm{FFN}(\mathrm{MHCAtt}(\boldsymbol{S}_{\mathrm{Av}}^l,\boldsymbol{S}_{\mathrm{At}}^l,\boldsymbol{S}_{\mathrm{At}}^l)),\tag{8}$$

$$\boldsymbol{M}_{\mathrm{t}}^l=\mathrm{FFN}(\mathrm{MHCAtt}(\boldsymbol{S}_{\mathrm{At}}^l,\boldsymbol{S}_{\mathrm{Av}}^l,\boldsymbol{S}_{\mathrm{Av}}^l))。\tag{9}$$

利用这一模块可以通过视觉和文本特征相互作用促进视觉和文本特征的增强。将得到的特征 $\boldsymbol{F}=[\boldsymbol{M}_{\mathrm{v}}\quad\boldsymbol{M}_{\mathrm{t}}]\in\mathbf{R}^{(L+K)\times 512}$ 送入 M^2 Transformer 的网状解码器中[16],生成最终的描述 $\boldsymbol{Y}$。

3　试验

3.1　数据集和评价标准

3.1.1　数据集

本研究在微软常见物体场景(Microsoft common objects in context, MSCOCO)数据集上进行试验,评估所提模型的有效性。MSCOCO 数据集是图像描述生成任务中通用的数据集,包含123 287张有标注图片,每张图片都有 5 个不同的标题。为确保与其他方法进行公平比较,本研究采用文献[19]中的划分方法,将训练集、验证集、测试集分别划分为113 287、5 000、5 000 张。

3.1.2　评价指标

为评估本研究提出的 MVTFF-IC 的有效性,采用 5 种在图像描述生成任务中广泛使用的评价指标衡量模型的性能,包括双语互译质量评估指标 B_{leu}[20]、带有明确排序的翻译评估指标 M_{eteor}[21]、以召回率为导向的摘要评价指标 $R_{\mathrm{ouge\text{-}l}}$[22]、基于共识的图像描述评价指标 C_{IDEr}[23] 和语义命题图像描述评价指标 S_{pice}[24]。其中,B_{leu} 是比较参考描述与生成描述之间 n-gram 的重合度,重合度越高,代表生成的描述越准确,其中 n-gram 为连续 n 个单词组成的序列,后续试验测试了 1-gram 和 4-gram 的得分,表示为 B_{leu}-1 和 B_{leu}-4;M_{eteor}更注重句子中单词的召回率和准确率;$R_{\mathrm{ouge\text{-}l}}$通过计算句子之间的最长公共子串得到准确率和召回率;C_{IDEr}对句子之间的每个 n-gram 执行词频-逆文档频率(term frequency-inverse document frequency, TF-IDF)加权,计算 TF-IDF 权重向量之间的余弦相似度,衡量参考描述和生成描述之间的一致性,评价图像描述的一致性和丰富度;S_{pice}将参考描述和生成描述转化为句子场景图,分析句子之间的对象、属性及其之间的关系。

3.2　试验设置

本研究使用文献[25]提出的物体-语义对齐预

训练（object-semantics aligned pre-training，OSCAR）模型提取图像区域特征，区域特征的最大数量设定为50。将图像划分为5等分和9等分两种网格形式，利用CLIP-I提取相应的网格特征。为得到与图像及其子区域最相关的文本特征，本研究采用文献[9]提出的跨模态图文检索方法，即利用视觉基因组数据集的标注构建由主语-属性和主语-谓语-宾语构成的文本语料库，计算CLIP图像及其子区域与文本语料库中的文本特征的余弦相似度，最高的前4个余弦相似度作为文本特征。在整个训练、验证和测试阶段，目标检测器、图像编码器和文本编码都保持冻结状态，随机失活率设为0.2。试验平台为搭载Tesla V100 32 G的工作站，编程语言为Python3.7。

3.3 消融分析

3.3.1 皮尔逊系数的影响

本研究验证皮尔逊系数ρ对图像描述生成的影响。当ρ较小时，更多的区域特征和网格特征相连接；随着ρ增大，相似度更高的区域特征和网格特征才会连接。基于此，构建邻接矩阵进行视觉特征建模。皮尔逊系数的影响如表1所示。由表1可知，当ρ设置为0.5时，C_{IDEr}达到136.7，模型性能最好。后续试验将ρ设为0.5。此外，由于ρ接近0时，视觉特征的融合方法与普通的注意力机制相似，故可推断通过计算不同类型视觉特征之间的相关性构建特征交互矩阵，对多尺度视觉特征进行建模融合，能够有效指导不同类型视觉特征之间的交互。

表1 皮尔逊系数的影响

Table 1 Effect of Pearson's coefficient

ρ	B_{leu}-1/%	B_{leu}-4/%	M_{eteor}/%	$R_{\mathrm{ouge\text{-}l}}$/%	C_{IDEr}	S_{pice}/%
0.3	81.8	39.8	29.6	59.3	135.5	23.2
0.4	81.8	39.8	29.6	59.1	135.7	23.2
0.5	81.9	39.6	29.6	59.1	136.7	23.2
0.6	82.0	40.0	29.7	59.4	136.3	23.4
0.7	82.0	39.8	29.7	59.3	136.0	23.4

3.3.2 文本特征引入的影响

为评估加入文本特征的效果，本研究进行一系列试验，主要研究3种情况：不使用图像视觉特征，只使用文本特征输入标准的注意力机制生成图像描述语句；仅使用图像视觉特征，通过MVFF融合3种视觉特征，生成的结果送入标准注意力机制中生成图像描述；同时使用图像视觉特征和文本特征生成图像描述，即本研究提出的MVTFF-IC模型。试验结果如表2所示。由表2可知：当仅使用文本特征时，C_{IDEr}仅为124.7；仅使用图像视觉特征时，C_{IDEr}明显提高，达到134.5，是图像视觉特征在图像描述生成任务占据主导地位的原因；在图像视觉特征的基础上引入文本语义特征后，C_{IDEr}进一步提高到136.7，表明文本语义特征能够有效提高图像描述生成的性能。

表2 文本特征的影响

Table 2 Effects of text features

图像	文本	B_{leu}-1/%	B_{leu}-4/%	M_{eteor}/%	$R_{\mathrm{ouge\text{-}l}}$/%	C_{IDEr}	S_{pice}/%
×	√	78.9	35.9	28.2	56.9	124.7	22.1
√	×	81.8	39.7	29.7	59.4	134.5	23.8
√	√	81.9	39.6	29.6	59.1	136.7	23.2

注：“√”和“×”分别表示使用或不使用图像和文本输入。

3.3.3 MVFF和DSFM的效果

为评估本研究提出的MVFF和DSFM模块的有效性，先验证不使用这两个模块的效果，再对这两个模块分别进行消融试验，试验结果如表3所示。

表3 MVFF和DSFM的效果

Table 3 Effects of MVFF and DSFM

MVFF	DSFM	B_{leu}-1/%	B_{leu}-4/%	M_{eteor}/%	$R_{\mathrm{ouge\text{-}l}}$/%	C_{IDEr}	S_{pice}/%
×	×	81.5	39.2	29.3	58.8	135.6	23.0
√	×	81.8	39.5	29.6	59.1	136.3	23.1
×	√	81.7	39.4	29.5	59.0	136.1	23.0
√	√	81.9	39.6	29.6	59.1	136.7	23.2

注：“√”和“×”分别表示使用或不使用MVFF和DSFM模块。

为评估不使用这两个模块的效果，将视觉特征和文本特征直接串联融合，送入普通的注意力机制，生成图像的自然语言描述，试验结果如表3第一行所示。

为评估MVFF模块的有效性，本研究将视觉特征输入MVFF模块，将提炼后的视觉特征与文本特征共同输入普通的注意力模块中，生成图像描述结果，试验结果如表3第二行所示。与不使用MVFF模块的情况相比，性能指标得到全面改善，C_{IDEr}提高0.7。

本研究评估了只使用DSFM模块的情况，即将3种视觉特征直接连接后与文本特征一起送入DSFM模块中，试验结果如表3第三行所示，与不使用DSFM的情况相比，C_{IDEr}提高0.5。

为评估整合两个模块生成图像描述的性能，在表3第四行列出二者共同协作的结果，表明两个模块协作有助于提升图像描述生成的性能指标。

3.4 试验结果定量分析

本研究与先进算法进行比较，结果如表4所示，每个评估指标的最佳结果均以粗体标出。将本研究提出的MVTFF-IC与完全依赖于视觉特征的先进模型进行比较。Up-Down[4]、GCN-LSTM[13]、SGAE[6]、ORT[7]、Dual-GCN[8]、AoANet[14]、X-

Transformer[15]和 M² Transformer[16]等模型只采用单一类型的视觉特征,通过改进注意力机制提高图像描述生成的性能;GET 通过利用全局特征,指导模型选择重要区域,对区域之间的关系进行建模[1];DLCT 根据网格和区域特征的边界框设计几何对齐图,利用提出的局部抑制交叉注意力模块对两种视觉特征进行建模[17]。GET 和 DLCT 这两种方法结合多种视觉特征,设计特征对齐的方法提高性能。由表 2 试验结果可知,本研究提出的 MVFF 模块对视觉特征的利用优于以上算法。

表 4　与先进算法比较

Table 4　Comparison with advanced algorithms

方法	B_{leu}-1/%	B_{leu}-4/%	M_{eteor}/%	R_{ouge-l}/%	C_{IDEr}	S_{pice}/%
Up-Down	79.8	36.3	27.7	26.9	120.1	21.4
GCN-LSTM	80.5	38.2	28.5	58.3	127.6	22.0
SGAE	80.8	38.4	28.4	58.6	127.8	22.1
ORT	80.5	38.6	28.7	58.4	128.3	22.6
Dual-GCN	82.2	39.7	29.7	59.7	129.2	
AoANet	80.2	38.9	29.2	58.8	129.8	22.4
X-Transformer	80.9	39.7	29.5	59.1	132.8	23.4
M² Transformer	80.8	39.1	29.1	58.4	131.2	22.6
GET	81.5	39.5	29.3	58.9	131.6	22.8
DLCT	81.4	**39.8**	29.5	59.1	133.8	23.0
EURAIC	80.9	39.5	29.4	59.4	130.3	
Xmodal-ctx	81.5	39.7	**30.0**	**59.5**	135.9	**23.7**
MVTFF-IC	**81.9**	39.6	29.6	59.1	**136.7**	23.2

为确保试验的公平性,本研究将所提 MVTFF-IC 与同时利用视觉和语义信息作为输入的先进模型进行比较。其中,文献[18]提出 EURAIC,引入一个外部知识网络,由核心对象的语义信息构建,与区域特征融合生成图像描述;文献[9]提出超越预训练的目标检测器 Xmodal-ctx,利用 CLIP 预训练模型从 Visual Genome 数据集挖掘属性和关系,通过挖掘到的文本语义信息与图像视觉信息进行连接操作后送入 M² Transformer 生成图像描述结果。Xmodal-ctx 有多个特征提取器版本,本研究只与基于 Oscar 的版本比较。由表 4 的试验结果可知:本研究提出的 MVTFF-IC 在 B_{leu}-1 和 C_{IDEr} 方面优于上述所有算法,B_{leu}-1 和 C_{IDEr} 分别比最先进的算法提高 0.4 和 0.8 百分点,主要得益于本研究提出的 MVFF 有效实现了视觉特征之间的优势互补,同时通过 DSFM 结合包含对象关系的文本语义信息,使生成的描述更加准确;其他指标也取得与先进算法接近的性能。

3.5　准确性分析

S_{pice}用于评估生成描述的准确性。将真实标注和 MVTFF-IC 生成的预测语句分别通过斯坦福场景图解析器进行解析[26],转换为语义属性(包括对象、属性和关系);根据语义属性分别构建场景图,将真实标注与预测语句的场景图进行匹配,计算出正确匹配的对象、属性和关系所占比例,得到对象、属性和关系的准确率。语义分类准确率结果如表 5 所示。由表 5 可以看出,本研究提出的 MVTFF-IC 在所有类别中的准确率都取得了优异的结果。

表 5　准确性分析

Table 5　Accuracy analysis

方法	准确率/%		
	对象	属性	关系
Up-Down	4.5	8.9	62.0
Dual-GCN	17.1	31.0	75.0
MVTFF-IC	18.3	33.1	76.4

3.6　试验结果定性分析

由 MVTFF-IC 生成的图像描述结果如图 2 所示。为说明 MVTFF-IC 的优势,本研究选择 Dual-GCN 作为基线模型,将生成的图像描述结果与真实标注共同展示。由图 2 可以看出,基线模型对于图像中对象分类和对象之间的关系理解依然存在问题,如 tarmac 误判为 road,seagull 误判为 bird,snowboard 和 skateboard 误判为 board,next to a bicycle 误判为 and a bicycle,riding a snowboard 误判为 with a board 等。

图 2　MVTFF-IC 生成的图像描述结果

Fig.2　Image captioning results generated by MVTFF-IC

上述研究结果表明,本研究提出的 MVTFF-IC 具有更强的图像理解能力,能有效关注图像中关键对象、对象与对象之间的关系等关键要素。因此,MVTFF-IC 能够生成更加合理、准确和生动的图像描述语句。

4 结论

本研究提出一种融合多尺度视觉和文本语义特征的图像描述生成算法 MVTFF-IC,旨在实现对图像视觉信息的全面理解,为非名词的生成提供显性指导。在 MVTFF-IC 中,提出多尺度视觉特征融合模块 MVFF,通过特征相似度引导视觉特征间的交互融合,实现不同视觉特征的有效对齐和互补;提出深度语义融合模块 DSFM,通过引入表示对象属性和对象间关系的细粒度信息,指导图像描述文本生成。未来,本研究旨在探索如何更细粒度、更深层次地融合视觉与语义信息,以便更有效地挖掘图像信息,期望在图像描述生成领域实现更大突破。

参考文献:

[1] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE, 2009: 248-255.

[2] 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 Conference on Artificial Intelligence. [S. l.]: AAAI, 2021: 1655-1663.

[3] JIANG H, MISRA I, ROHRBACH M, et al. In defense of grid features for visual question answering[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10267-10276.

[4] 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 & Machine Intelligence, 2017, 39(6):1137-1149.

[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 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 6077-6086.

[6] YANG X, TANG K, ZHANG H, et al. Auto-encoding scene graphs for image captioning[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 10685-10694.

[7] HERDADE S, KAPPELER A, BOAKEY K, et al. Image captioning: transforming objects into words[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2019: 11137-11147.

[8] DONG X, LONG C, XU W, et al. Dual graph convolutional networks with Transformer and curriculum learning for image captioning[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM, 2021: 2615-2624.

[9] KUO C W, KIRA Z. Beyond a pre-trained object detector: cross-modal textual and visual context for image captioning[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 17969-17979.

[10] BERNARDI R, CAKICI R, ELLIOTT D, et al. Automatic description generation from images: a survey of models, datasets, and evaluation measures[J]. Journal of Artificial Intelligence Research, 2016, 55: 409-442.

[11] SOCHER R, KARPATHY A, LE Q V, et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 207-218.

[12] KULKARNI G, PREMRAJ V, ORDONEZ V, et al. Babytalk: understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[13] YAO T, PAN Y, LI Y, et al. Exploring visual relationship for image captioning[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 684-699.

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

[15] PAN Y, YAO T, LI Y, et al. X-linear attention networks for image captioning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10971-10980.

[16] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed-memory Transformer for image captioning [C]//Proceedings of the 2020 IEEE/CVF Conference

on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 10578-10587.

[17] LUO Y, JI J, SUN X, et al. Dual-level collaborative Transformer for image captioning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 2286-2293.

[18] WEI J, LI Z, ZHU J, et al. Enhance understanding and reasoning ability for image captioning[J]. Applied Intelligence, 2023, 53(3): 2706-2722.

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

[20] 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. Stroudsburg, USA: ACL, 2002: 311-318.

[21] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments [C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA: ACL, 2005: 65-72.

[22] LIN C Y. ROUGE: a package for automatic evaluation of summaries [C]//Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain: ACL, 2004: 74-81.

[23] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: consensus-based image description evaluation[C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015: 4566-4575.

[24] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE: semantic propositional image caption evaluation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Amsterdam, Netherlands: Springer, 2016: 382-398.

[25] LI X, YIN X, LI C, et al. Oscar: object-semantics aligned pre-training for vision-language tasks[C]// Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 2020: 121-137.

[26] SCHUSTER S, KRISHNA R, CHANG A, et al. Generating semantically precise scene graphs from textual descriptions for improved image retrieval[C]// Proceedings of the fourth Workshop on Vision and Language. Lisbon, Portugal: ACL, 2015: 70-80.

(编辑:孙亚彤)

(上接第 79 页)

ZHOU Liming, ZHANG Yang, FU Daiguang, et al. Wave field and time-frequency characteristics of ground penetrating radar for underground cavity of road[J]. Journal of Tongji University (Natural Science), 2024, 52(1): 77-85.

[24] LU H P, PLATANIOTIS K N, VENETSANOPOULOS A N. MPCA: multilinear principal component analysis of tensor objects[J]. IEEE Transactions on Neural Networks, 2008, 19(1): 18-39.

[25] ZHU L, WANG X J, KE Z H, et al. BiFormer: vision transformer with bi-level routing attention [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 10323-10333.

[26] 常致富, 周风余, 王玉刚, 等. 基于深度学习的图像自动标注方法综述[J]. 山东大学学报(工学版), 2019, 49(6): 25-35.
CHANG Zhifu, ZHOU Fengyu, WANG Yugang, et al. A survey of image captioning methods based on deep learning[J]. Journal of Shandong University (Engineering Science), 2019, 49(6): 25-35.

[27] KLAMBAUER G, UNTERTHINER T, MAYR A, et al. Self-normalizing neural networks[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: [s.n.], 2017: 971-980.

(编辑:郭少华)