计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330523
61(11): 2897−2908, 2024

一种基于双重语义协作网络的图像描述方法

江泽涛　朱文才　金　鑫　廖培期　黄景帆
（广西图像图形智能处理重点实验室（桂林电子科技大学）　广西桂林　541004）
（zetaojiang@126.com）

An Image Captioning Method Based on DSC-Net

Jiang Zetao, Zhu Wencai, Jin Xin, Liao Peiqi, and Huang Jingfan
（*Guangxi Key Laboratory of Image and Graphic Intelligent Processing* (*Guilin University of Electronic Technology*), *Guilin*, *Guangxi* 541004）

Abstract　As visual features closer to the text domain, the grid features extracted by the CLIP (contrastive language-image pre-training) image encoder are easy to convert into the corresponding semantic natural language, which can alleviate the semantic gap problem, so it may become an important source of visual features in the image captioning in the future. However, this method does not consider that the division of image content may cause a complete object to be divided into several grids. The segmentation of the objects will inevitably lead to the lack of a complete expression of the object information in the feature extraction results, and further lead to the lack of an accurate expression of the object and the relationship between the objects in the generated sentence. Aiming at the phenomenon of grid features extracted by CLIP image encoder, we propose dual semantic collaborative network (DSC-Net) for image captioning. Specifically, dual semantic collaborative self-attention (DSCS) module is first proposed to enhance the expression of object information by CLIP grid features. Then dual semantic collaborative cross-attention (DSCC) module is proposed to integrate semantic information between grid and object to generate visual features, and to be used to predict sentences. Finally, dual semantic fusion (DSF) module is proposed to provide region-oriented fusion features for the above two semantic cooperation modules, and to solve the problem of correlation conflicts that may arise in the process of semantic cooperation. After a large number of experiments on the COCO dataset, the proposed model achieves a CIDEr score of 138.5% on the offline test set divided by Karpathy et al., and a CIDEr score of 137.6% in the official online test. Compared with the current mainstream image captioning methods, this result has obvious advantages.

Key words　image captioning; grid features; attention mechanism; dual semantic collaborative attention; dual semantic collaborative feature fusion

摘　要　CLIP（contrastive language-image pre-training）视觉编码器提取的网格特征作为一种更加靠近文本域的视觉特征，具有易转化为对应语义自然语言的特点，可以缓解语义鸿沟问题，因而未来可能成为图像描述任务中视觉特征的重要来源. 但该方法中未考虑图像内容的划分，可能使一个完整的目标被划分到

收稿日期：2023-06-19；**修回日期**：2023-12-19
基金项目：国家自然科学基金项目（62172118）；广西自然科学基金重点项目（2021GXNSFDA196002）；广西图像图形智能处理重点实验项目（GIIP2302, GIIP2303, GIIP2304）；广西研究生教育创新计划项目（YCSW2022269）；桂林电子科技大学研究生教育创新计划项目（2023YCXS046）
This work was supported by the National Natural Science Foundation of China (62172118), the Guangxi Natural Science Key Foundation (2021GXNSFDA196002), the Project of Guangxi Key Laboratory of Image and Graphic Intelligent Processing (GIIP2302, GIIP2303, GIIP2304), the Innovation Project of Guangxi Graduate Education (YCSW2022269), and the Innovation Project of GUET Graduate Education (2023YCXS046)
通信作者：朱文才（zhuwencai00@qq.com）

若干个网格中,目标被切割势必会导致特征提取结果中缺少对目标信息的完整表达,进而导致生成的描述语句中缺少对目标及目标间关系的准确表述.针对CLIP视觉编码器提取网格特征这一现象,提出一种基于双重语义协作网络(dual semantic collaborative network, DSC-Net)的图像描述方法.具体来说:首先提出双重语义协作自注意力(dual semantic collaborative self-attention, DSCS)模块增强CLIP网格特征对目标信息的表达能力;接着提出双重语义协作交叉注意力(dual semantic collaborative cross-attention, DSCC)模块,综合网格和目标2个层面的语义构造与文本相关的视觉特征,进行描述语句预测;最后提出双重语义融合(dual semantic fusion, DSF)模块,为上述的2个语义协作模块提供以区域为主导的融合特征,解决在语义协作过程中可能出现的相关性冲突问题.经过在COCO数据集上的大量实验,提出的模型在Karpathy等人划分的离线测试集上取得了138.5%的CIDEr分数,在官方在线测试中取得了137.6%的CIDEr分数,与目前主流的图像描述方法相比具有显著优势.

关键词 图像描述;网格特征;注意力机制;双重语义协作注意力;双重语义协作特征融合

中图法分类号 TP391.41

图像描述(image captioning)是一项计算机视觉领域与自然语言处理领域的交叉任务,是图像与视频转化为文字语音的重要桥梁,在图像与视频智能化服务中扮演重要角色,在机器人视觉、智能交通、智慧农业等诸多领域具有重要的潜在应用前景.由于图像描述研究属于起步阶段,具有非常大的挑战性.随着深度学习技术的发展,基于深度学习的方法在图像描述中成为主流方向[1].目前典型的方法是利用目标检测器提取图像的区域特征,然后在特征间进行目标语义关系建模,利用建模结果提高特征的语义表示能力,最后将特征送入文本生成模块,产生描述图像内容的自然语言.但是,该方法中使用的目标检测器未与文本进行过联合训练,使得提取到的区域特征与文本域间存在较大差距,继而导致将区域特征转化为自然语言时往往不够准确.

近年来,受到自然语言处理领域中下游任务的启发,研究者们提出使用编码-解码结构进行图像描述任务.早期的研究中使用卷积神经网络(convolutional neural network, CNN)提取视觉特征,使用循环神经网络(recurrent neural network, RNN)预测描述文本,并将其分别称为编码阶段和解码阶段.随着Transformer[2]结构的兴起,一些人开始尝试利用Transformer代替RNN进行文本生成,为了避免混淆,将CNN对图像的处理过程称为视觉特征提取,而将编码和解码特指为Transformer中的编码器和解码器,这种新的结构被称为基于Transformer的图像描述方法.基于RNN的图像描述方法参数量少,但描述精度较低、并行性差、训练周期长;基于Transformer的图像描述方法描述精度高、并行性好、训练周期短,但参数量和运算量都较高.随着硬件的发展,Transformer的参数和运算规模能够被越来越多的人接受,基于Transformer的图像描述方法逐渐成为了图像描述领域中的主流,被越来越多的研究者们使用.

在视觉特征提取方面,早期的研究中受到分类任务的启发,使用单个向量表示图像的视觉特征[3],但是这种方法会导致描述语句中的全部单词都由1个向量预测,忽略了不同单词在语义上的差异.后来,使用一组向量作为图像的视觉特征[4],每个向量对应原图中的1个网格,同时引入注意力机制关注单词间的语义差异,以便根据原图中不同位置的特征向量生成不同的单词,但是这种方法的缺点是很难感知到目标层面的语义信息,难以对目标及目标间的关系形成有效建模.再后来,使用一组通过目标检测器取得的特征向量作为图像的视觉特征[5],这种特征向量基于目标检测器对目标的精准定位能力,从目标附近的区域进行提取,是目前主流的视觉特征提取方式,提取到的特征被称为区域特征,利用区域特征可以对目标及目标间的关系进行有效感知,但是由于图像与文本间存在语义鸿沟,区域特征中丰富的语义信息很难转化为与之对应的自然语言.最近一些研究表明,利用经过视觉文本联合训练的视觉特征提取模型,即CLIP(contrastive language-image pre-training)[6]的视觉编码器,可以提取到更加靠近文本域的视觉特征,这种视觉特征更容易转化为对应语义的自然语言,缓解语义鸿沟问题.

由于CLIP视觉编码器无法像目标检测器一般对目标进行定位,故只能提取到网格形式的视觉特征,记为CLIP网格特征,特征提取过程中未考虑图像内容的网格划分行为可能破坏目标的完整性,导致CLIP网格特征中缺少对目标信息的完整表达,进而导致生成的描述语句中缺少对目标及目标间关系

的准确表述，这是该特征提取方法的不足之处.

针对区域特征生成的描述语句不够准确的问题，Guo 等人[7]提出将区域之间的相对位置关系添加到目标语义关系建模中，利用相对位置信息增强对目标间语义关联的表达能力；Pan 等人[8]提出在将 Transformer 中原本的单一维度建模改为对二阶特征交互建模，从多个维度提高对语义的表达能力. 这些方法基于区域特征的特点进行了设计，但目标检测器自身的缺陷限制了模型性能的进一步提高. 为此一些研究者尝试使用经过图像文本联合训练的视觉特征提取模型代替目标检测器，Shen 等人[9]提出利用 CLIP 的视觉编码器提取网格特征，使用这种网格特征替换区域特征，在相同的网络结构下取得了显著的性能提升，但是该团队仅仅证明了 CLIP 提取网格特征的有效性，却没有提出新的模型结构. Li 等人[10]提出利用 CLIP 的视觉编码器和文本编码器提取网格特征和相关的文本特征，通过视觉和文本 2 种特征来生成描述图像内容的语句，该团队虽然提出了新的模型，但将研究重心放在了对文本特征的处理上，未基于网格特征的特点进行设计. 综上所述，当前的研究中虽然证明了 CLIP 网格特征与传统的区域特征相比所具有的优势，但也要注意到其仍然存在缺陷，无法完全代替区域特征.

本文针对提取网格特征时可能将一个完整的目标划分到多个网格中，继而丢失完整目标语义信息的问题，提出了一种基于双重语义协作网络（dual semantic collaborative network, DSC-Net）的图像描述方法. 该方法设计了 DSC-Net 网络结构，其主要贡献有 3 个方面：

1）设计了双重语义协作自注意力（dual semantic collaborative self-attention, DSCS）模块，借助区域特征对目标的准确表示能力进行目标间的语义关联感知；然后利用重叠矩阵将感知结果映射到对应位置的网格上，融入网格特征的语义建模过程，使各网格特征向量得以同时汇聚整合网格间、目标间、目标内的相关联信息，进行基于网格内容、目标内容、目标间关系的多层次建模，完成信息的更新迭代，增强对目标层面语义的表达能力.

2）设计了双重语义协作交叉注意力（dual semantic collaborative cross-attention, DSCC）模块，借助区域特征对目标位置的捕捉能力，计算文本与目标的语义相关性；然后利用重叠矩阵将计算结果映射到对应位置的网格上，从而在视觉与文本的跨模态建模中综合目标和网格 2 个层面的信息构建上下文感知的视觉特征表示，进而利用特征表示中蕴含的视觉语义信息对下一个单词进行预测.

3）设计了双重语义融合（dual semantic fusion, DSF）模块，内置了 3 种确保区域特征主导地位的融合机制，为上述 2 个语义协作模块提供以区域为主导的融合特征，融合特征在保留区域特色的同时融入了网格信息，利用这种融合特征可以在不与网格关联冲突的情况下进行目标关联感知，从而解决语义协作过程中可能出现的相关性冲突问题.

本文方法在广泛使用的 COCO 数据集上进行大量实验，遵循 Karpathy 等人[11]的划分方式进行离线测试，取得了 138.5% 的 CIDEr 分数，在 COCO 图像描述在线测试平台上进行在线测试，取得了 137.6% 的 CIDEr 分数，与目前典型的图像描述方法相比具有明显优势.

1 相关工作

1.1 CLIP

CLIP[6]是一个用于视觉语言任务的预训练模型，由一个视觉编码器和一个文本编码器组成，2 个编码器均直接使用目前主流的特征提取网络，随意组合图像和文本作为训练样本，其中图像和文本语义一致的作为正样本，否则作为负样本，通过对比学习的方式展开视觉文本联合训练. 具体来说，首先利用 2 个编码器将样本中的图像和文本编码为向量的形式，然后计算 2 个向量的相似度得分，并在训练过程中通过对比损失函数控制正样本的相似度得分尽可能高和负样本的相似度得分尽可能低.

显然，利用 CLIP 取得的视觉特征和文本特征在表达相近的语义时将具有很高的相似度，所以与传统的区域特征相比，这种新的视觉特征更加接近文本域，可以更准确地转化为对应语义的自然语言，非常适合用于图像描述任务.

1.2 基于 Transformer 的图像描述

得益于 Transformer[2]的强大性能，近年来研究者们大多遵循 Transformer 结构对图像描述任务展开研究[7-10,12-15]. 具体来说，首先使用经过预训练的特征提取模型提取图像的视觉特征；接着使用 Transformer 的编码器通过多头注意力模块在视觉特征内部进行建模，建模过程中基于各视觉特征向量的语义进行关联感知，再利用感知结果汇聚整合与自身相关联的信息，实现更加精细化的视觉语义表达；最后将经过编码器处理的视觉特征送入 Transformer 解码器，在解码器中逐字生成描述图像内容的自然语言，每

个时间步下通过掩码注意力运算提取已生成单词的文本特征，与视觉特征进行跨模态建模，构建上下文感知的视觉特征表示，并根据该特征表示预测下一个单词. 同时，为了提高视觉语义的精细化程度和预测描述语句的准确程度，对编码器和解码器中的建模过程进行若干次堆叠，达到逐步求精的效果，并在建模后使用前馈神经网络进行处理，增强模型的泛化能力.

1.3 多头注意力

在基于 Transformer 的图像描述模型中进行精细化设计是目前进行图像描述相关研究的主流方法之一，而 Transformer 结构的核心是多头注意力(multi-head attention, MHA)模块，该模块可以有效地对输入之间的关系进行建模，并用建模结果对输入进行更新迭代，得到更加丰富和准确的表示. 多头注意力的运算过程可以用公式表示为

$$F_{\mathrm{MHA}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = concat(head_1, head_2, \cdots, head_h)\boldsymbol{W}^{\mathrm{o}}, \quad (1)$$

$$head_i = Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{q}}, \boldsymbol{K}\boldsymbol{W}_i^{\mathrm{k}}, \boldsymbol{V}\boldsymbol{W}_i^{\mathrm{v}}), \quad (2)$$

其中 $\boldsymbol{W}^{\mathrm{o}}, \boldsymbol{W}_i^{\mathrm{q}}, \boldsymbol{W}_i^{\mathrm{k}}, \boldsymbol{W}_i^{\mathrm{v}}$ 是可学习的参数矩阵，*Attention* 是缩放点积运算函数，用公式可以表示为

$$Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{q}}, \boldsymbol{K}\boldsymbol{W}_i^{\mathrm{k}}, \boldsymbol{V}\boldsymbol{W}_i^{\mathrm{v}}) = softmax\left(\frac{(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{q}})(\boldsymbol{K}\boldsymbol{W}_i^{\mathrm{k}})^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}\boldsymbol{W}_i^{\mathrm{v}}. \quad (3)$$

2 基于 DSC-Net 的图像描述方法

2.1 设计思想

CLIP 视觉编码器提取的网格特征在图像描述任务中展现出潜在的优势，但特征提取过程中未考虑图像内容的网格划分行为可能破坏目标的完整性，将一个完整的目标划分到多个网格中，这将导致特征提取结果中目标层面语义信息的缺失，继而增加后续处理中感知目标及目标间关系的难度. 为此，本文提出 DSC-Net 网络借助区域特征进行目标语义关联感知，再将感知结果注入对应位置的网格上，实现网格语义和目标语义协作的双重语义关联感知与建模. 该网络遵循目前主流的基于 Transformer 的图像描述框架进行设计，提出了 DSCS, DSCC, DSF 这 3 个全新的模块，如图 1 所示.

给定一张图片，首先使用 CLIP 视觉编码器和目标检测器提取网格特征和区域特征，并根据二者在原图中的位置关系构建重叠矩阵；接着在 DSCS 模块中利用重叠矩阵将区域特征中目标内和目标间的关系映射到对应位置的网格上，为网格特征进行基于网格内容、目标内容、目标间关系的多层次建模，增强对目标层面语义信息的表达能力；然后利用 DSF

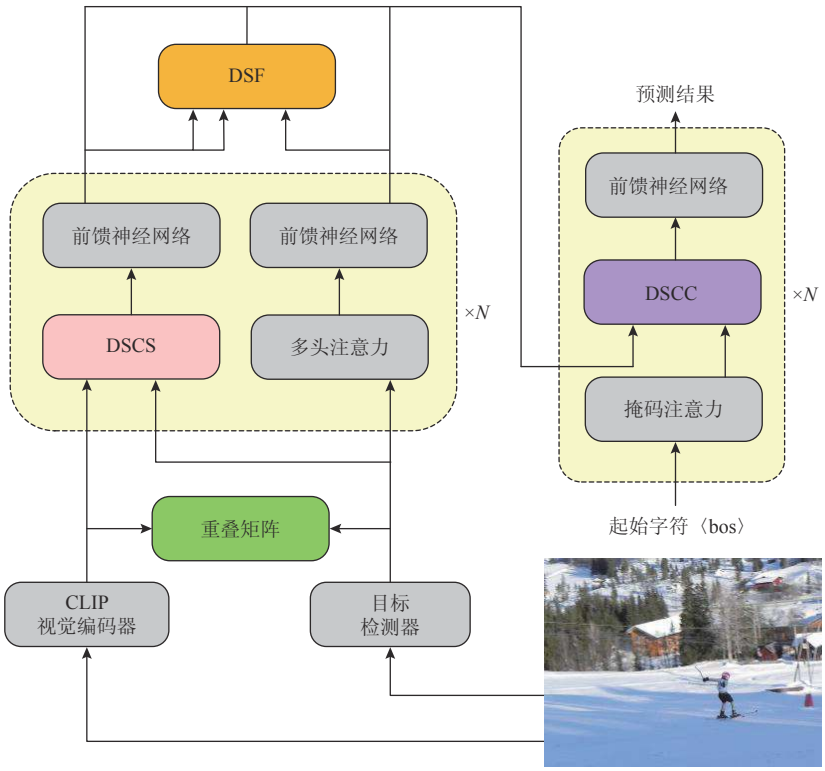


Fig. 1 DSC-Net structure diagram

图 1 DSC-Net 结构图

对视觉特征进行处理，生成以区域为主导的融合特征，避免语义协作过程中可能出现的相关性冲突问题；再将经过处理的视觉特征送入 DSCC 模块进行视觉与文本的跨模态建模，从网格和目标 2 个层面感知视觉特征与文本特征的语义关联，构建上下文感知的视觉特征表示，最后利用该特征表示预测下一个单词，即将视觉特征中的视觉语义信息转化为对应的自然语言.

2.2 双重语义协作自注意力模块

在一系列的视觉特征中，利用 CLIP 视觉编码器提取的网格特征具有靠近文本域，易转化为对应语义自然语言的优点，也有缺少目标层面语义信息的缺点，而这恰好是由目标检测器提取的区域特征所擅长的，为了增强网格特征对目标层面语义的表达能力，本文设计了 DSCS 模块. 目前，已有的自注意力模块都是为区域特征设计的，未考虑过目标语义不足的情况，所以无法解决网格特征的问题. 本文设计的 DSCS 模块如图 2 所示，首先通过相似度运算在 2 种特征内部分别进行网格和目标层面的语义关联感知；接着利用重叠矩阵将目标层面的语义关联感知结果映射到对应位置的网格上，在网格间形成对目标内容和目标间关系的语义关联感知结果，与原本基于网格内容的语义关联感知结果融合；最后利用融合后得到的双重语义关联感知矩阵进行基于网格内容、目标内容、目标间关系的多层次建模，达到增强网格特征目标语义表达能力的目的.

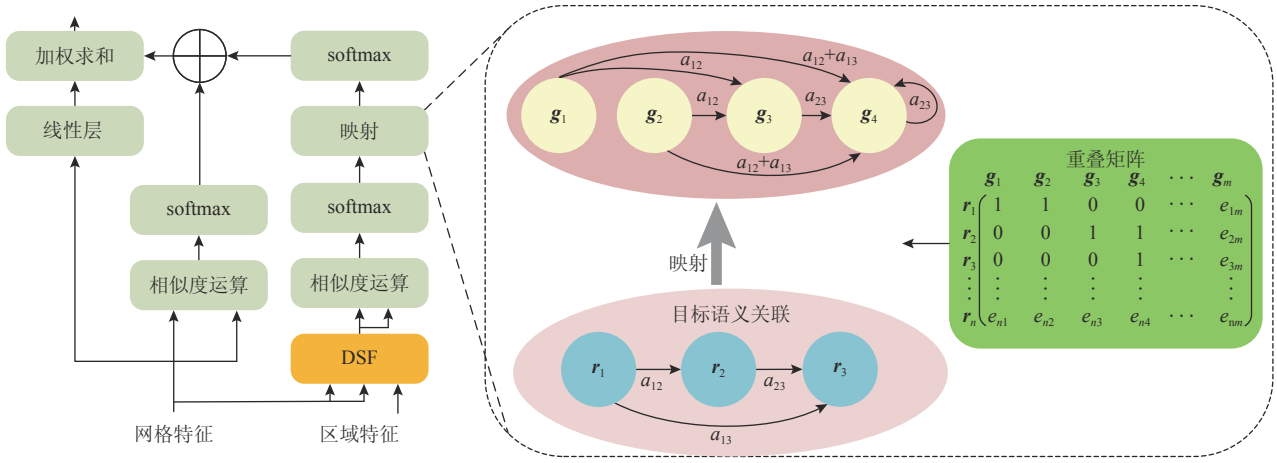


Fig. 2 DSCS structure diagram

图 2 DSCS 结构图

在上述语义关联感知过程中，网格层面的语义关联感知可以直接通过对网格特征的计算得到，而目标层面的语义关联感知则需要在 DSF 模块对区域特征处理后才能进行计算，这是为了避免在关联融合中出现“相关性冲突”，DSF 模块的结构和“相关性冲突”的细节将会在 2.3 节中进行详细介绍.

进行语义关联感知时，首先利用线性层将输入映射到 2 个不同的维度上作为查询和键，接着采用点积的方式计算相似度，最后通过多头的方式关注更多的语义相关信息. 语义关联感知的结果一般表现为矩阵形式，这个计算过程用公式可以表示为

$$Sim(\boldsymbol{Q},\boldsymbol{K})=\frac{(\boldsymbol{QW}^{\mathrm{q}})(\boldsymbol{KW}^{\mathrm{k}})^{\mathrm{T}}}{\sqrt{d}}, \tag{4}$$

$$\begin{aligned}\boldsymbol{M}_{\mathrm{GSS}}(\boldsymbol{G},\boldsymbol{G})&=concat(head_1^{g},head_2^{g},\cdots,head_h^{g}),\\ head_i^{g}&=softmax\left(Sim_i^{g}(\boldsymbol{G},\boldsymbol{G})\right),\end{aligned} \tag{5}$$

$$\begin{aligned}\boldsymbol{M}_{\mathrm{OSS}}(\boldsymbol{R}^{\mathrm{f}},\boldsymbol{R}^{\mathrm{f}})&=concat(head_1^{r},head_2^{r},\cdots,head_h^{r}),\\ head_i^{r}&=softmax\left(Sim_i^{r}(\boldsymbol{R}^{\mathrm{f}},\boldsymbol{R}^{\mathrm{f}})\right),\boldsymbol{R}^{\mathrm{f}}=F_{\mathrm{DSF}}(\boldsymbol{R},\boldsymbol{G},\boldsymbol{G}),\end{aligned} \tag{6}$$

其中 Sim 表示点积相似度运算函数，$\boldsymbol{M}_{\mathrm{GSS}}$ 表示网格语义关联矩阵，$\boldsymbol{M}_{\mathrm{OSS}}$ 表示目标语义关联矩阵，DSF 表示双重语义融合模块，$\boldsymbol{G}$ 表示网格特征，$\boldsymbol{R}$ 表示区域特征，$\boldsymbol{W}^{\mathrm{q}}$ 和 $\boldsymbol{W}^{\mathrm{k}}$ 表示可学习的参数矩阵. 为了将 $\boldsymbol{M}_{\mathrm{OSS}}$ 映射到网格间，需要使用重叠矩阵中记录的区域和网格在原图中的位置关系. 在构建重叠矩阵时，将网格与目标的数量作为矩阵的行和列，使用数字 1 和 0 表示对应编号的特征在原图中是否存在重叠，用公式可以将重叠矩阵的构建过程表示为

$$E_{ij}=\begin{cases}1, \text{目标和网格存在重叠},\\ 0, \text{目标和网格不存在重叠}.\end{cases} \tag{7}$$

利用重叠矩阵进行映射的过程如图 2 所示，其中 $\boldsymbol{r}_i$ 和 $\boldsymbol{g}_i$ 分别表示区域特征向量和网格特征向量，a_{ij} 表示目标间的语义关联情况，通过重叠矩阵确认区域与网格的对应关系，然后将区域间的语义关联映射到对应位置的网格上，将目标间的关联感知结果转化为网格间基于目标内容和目标间关系的语义关联

感知结果. 上述过程用公式可以表示为重叠矩阵与目标语义关联矩阵的 2 次矩阵乘法运算，这个映射过程用公式可以表示为

$$\boldsymbol{M}'_{\mathrm{OSS}} = softmax\left(\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{M}_{\mathrm{OSS}})\boldsymbol{E}\right). \tag{8}$$

最后，将上述结果与网格间原本存在的基于网格内容的语义关联感知结果以求和的方式进行融合，再利用融合后的关联矩阵进行基于网格内容、目标内容、目标间关系的多层次建模，对网格间、目标间、目标内的相关信息进行汇聚与整合，得到基于双重语义协作的更准确、更丰富的视觉表示，用公式可以表示为

$$\boldsymbol{G}' = softmax(\boldsymbol{M}_{\mathrm{GSS}} + \boldsymbol{M}'_{\mathrm{OSS}})\boldsymbol{G}\boldsymbol{W}^{\mathrm{v}}. \tag{9}$$

综上所述，在该模块中将双重语义协作的思想融入网格特征内部的建模过程，利用区域特征对目标的敏锐感知力增强了网格特征对目标信息的表达能力，缓解了网格特征中目标层面语义信息缺失的问题.

2.3 双重语义融合模块

使用区域特征辅助网格进行语义关联感知，虽然可以增加对目标间关系的感知能力，但也可能对原本基于网格内容的语义关联感知结果造成破坏. 试想一下，对于原图中的 2 个网格，当基于网格内容计算出的语义相关性与基于所属目标计算出的语义相关性产生冲突时，使用其中任何一方粗暴地改变另一方都是不合理的. 同时，本文的消融实验结果也证明，在双重语义协作过程中直接使用区域特征进行目标语义关联感知，并不能显著提高模型对图像内容的表述能力，本文将这种现象称为“相关性冲突”.

为了解决这一问题，本文提出在双重语义关联感知过程中使用以区域为主导的融合特征进行目标层面的语义关联计算，以产生不与网格关联冲突的目标语义关联感知结果. 这就要求融合过程中一方面要保留区域特征对目标敏感的特点，另一方面要融入网格信息. 基于双重语义协作的需要，其中保留区域对网格敏感的特点又显得尤其重要，因此本文在生成融合特征的 DSF 模块内精心设计了 3 种机制来保证区域特征的主导地位. DSF 模块结构如图 3 所示.

首先，以区域特征为查询，网格特征为键和值进行多头注意力运算，运算过程中计算查询和键的相似度，再以相似度为权重对值加权求和，通过这种方式计算出的结果由网格特征构成，却以接近区域特征为目标，这是确保区域特征主导地位的第 1 种机制. 接着，将区域特征和多头注意力的运算结果送入

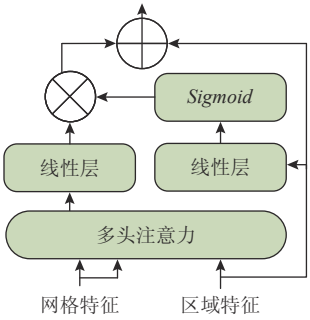


Fig. 3 DSF structure diagram

图 3 DSF 结构图

线性层引入可学习参数，再通过函数 *Sigmoid* 将区域特征中的元素转化为 0 和 1 之间的小数，与多头注意力的运算结果计算哈达玛积，在这个过程中，与区域特征相近的部分将被鼓励，反之被抑制，这是确保区域特征主导地位的第 2 种机制. 最后，将上述运算结果与区域特征求和，使原始的区域特征在最终的结果中固定占有一半的比重，这是确保区域特征主导地位的第 3 种机制. 上述过程用公式可以表示为

$$F_{\mathrm{DSF}}(\boldsymbol{R},\boldsymbol{G},\boldsymbol{G}) = F_{\mathrm{MHA}}(\boldsymbol{R},\boldsymbol{G},\boldsymbol{G})\boldsymbol{W}^{\mathrm{f}} \odot Sigmoid(\boldsymbol{R}\boldsymbol{W}^{r}) + \boldsymbol{R}, \tag{10}$$

其中 MHA 表示多头注意力模块，$\boldsymbol{W}^{\mathrm{f}}$和$\boldsymbol{W}^{r}$表示可学习的参数矩阵，⊙表示哈达玛积运算. 综上所述，该模块中通过精心设计的结构进行特征融合，使得双重语义协作过程中可以利用其中占据主导地位的区域特征进行目标语义关联感知，以及利用其中基于网格内容的语义信息避免冲突的产生，从而解决双重语义协作过程中可能出现的相关性冲突问题.

2.4 双重语义协作交叉注意力模块

在解码器中，需要在视觉特征和已生成文本间进行跨模态建模，构建上下文感知的视觉特征表示，然后利用该特征表示预测下一个单词，而网格特征的缺陷可能导致建模过程中丢失目标语义. 为此，本文提出 DSCC 模块，通过双重语义协作的方式进行视觉文本跨模态建模，综合网格和目标 2 个层面的语义信息构建上下文感知的视觉特征表示. 模块结构如图 4 所示.

在 Transformer 的解码器结构中，每个解码层接收相同的视觉特征，在堆叠过程中仅对描述语句逐步精细化地预测，因此，将 DSF 对区域特征的处理过程安排在解码层外，以保证各层中接收到的视觉信息相同. 解码层内部的 DSCC 模块中，首先计算文本特征与网格特征、融合特征间的语义关联情况；然后借助重叠矩阵将文本与目标的语义关联感知结果映射到文本与网格间，与原本基于网格内容的感知结

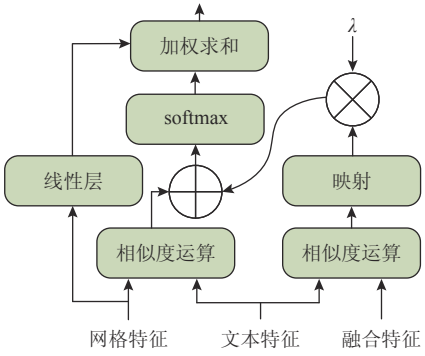


Fig. 4 DSCC structure diagram

图 4 DSCC 结构图

果融合，并引入超参数对映射过程进行控制，以避免注意力机制失效；最后利用融合后的关联矩阵整合视觉特征，综合目标和网格 2 个层面的语义信息构建上下文感知的视觉特征表示，用来进行下一个单词的预测. 公式为

$$\boldsymbol{M}_{\mathrm{TGS}}(\boldsymbol{T},\boldsymbol{G}') = Sim(\boldsymbol{T},\boldsymbol{G}'),\ \boldsymbol{M}_{\mathrm{TOS}}(\boldsymbol{T},\boldsymbol{R}_{\mathrm{f}}) = Sim(\boldsymbol{T},\boldsymbol{R}_{\mathrm{f}}),$$
$$\boldsymbol{R}_{\mathrm{f}} = F_{\mathrm{DSF}}(\boldsymbol{R},\boldsymbol{G},\boldsymbol{G}), \tag{11}$$

$$\boldsymbol{M}'_{\mathrm{TGS}} = \lambda\left(\boldsymbol{E}^{\mathrm{T}}(\boldsymbol{M}_{\mathrm{TOS}})\boldsymbol{E}\right) + \boldsymbol{M}_{\mathrm{TGS}}, \tag{12}$$

$$\boldsymbol{T}' = softmax(\boldsymbol{M}'_{\mathrm{TGS}})\boldsymbol{G}\boldsymbol{W}^{g}, \tag{13}$$

其中 $\boldsymbol{T}$ 表示文本特征，$\boldsymbol{M}_{\mathrm{TGS}}$ 表示文本特征与网格特征的语义关联矩阵，$\boldsymbol{M}_{\mathrm{TOS}}$ 表示文本特征与融合特征的语义关联矩阵，$\boldsymbol{M}'_{\mathrm{TGS}}$ 表示基于网格和目标的双重语义关联矩阵，$\boldsymbol{W}^{g}$ 表示可学习的参数矩阵，λ 表示超参数. 该过程的运算中引入了超参数，这是因为文本特征与视觉特征间不存在几何层面的相对位置关系，若此处使用与编码器中相同的处理方式，则会导致 $\boldsymbol{M}'_{\mathrm{TGS}}$ 中的元素相近，无法发挥出注意力机制应有的效果，继而导致描述精度下降.

综上所述，该模块中将双重语义协作的思想融入视觉文本的跨模态建模过程，综合目标和网格 2 个层面的语义信息构建上下文感知的视觉特征表示，利用该特征表示进行描述语句预测，避免结果中缺少目标相关的词汇.

2.5 损失函数

使用与主流方法相同的两阶段策略训练本文提出的模型：1）通过交叉熵损失进行预训练；2）通过强化学习[16]进行训练，以 CIDEr 分数作为奖励. 在第 1 阶段中，给定各时间步的真实标签 $Y_{1:T} = \{y_1, y_2, \cdots, y_T\}$，将各时间步的交叉熵损失之和作为描述语句的交叉熵损失，公式为

$$L_{\mathrm{XE}} = -\sum_{t=1}^{T}\ln(p_\theta(y_t|y_{1:t-1})), \tag{14}$$

其中 θ 表示模型的参数，$y_{1:t-1}$ 表示前 t–1 个时间步的预测结果，式（14）从分类任务的角度进行设计，但是描述任务中各时间步均依赖已有的结果进行预测，容易造成误差积累，同时交叉熵损失与评估指标不相关，所以该式仅能用于预训练而无法进一步提高描述的准确度. 在第 2 阶段中，将 CIDEr 分数作为奖励进行优化，目标是最大限度地减少负奖励期望，公式为

$$\nabla_\theta L_{\mathrm{RL}}(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\left(r(y^i_{1:T}) - b\right)\nabla_\theta \ln\left(p_\theta(y^i_{1:T})\right), \tag{15}$$

$$b = \frac{1}{k}\sum_{i=1}^{k} r(y^i_{1:T}),$$

其中 k 是束大小，r 是计算 CIDEr 分数的函数，b 是基线，通常使用均值作为基线. 式（15）从强化学习的角度进行设计，直接利用评估指标对模型进行优化，能够进一步提高描述准确度，同时避免了只用 1 个交叉熵损失函数带来的可能陷入局部最优的问题.

3 实验结果与分析

3.1 数据集及评价标准

实验在 COCO 数据集上进行，数据集共有 164 262 张图片，每张图片对应 5 条人工标注的描述语句，其中 123 287 张图片公开标注信息，供研究者下载使用，40 775 张图片用作线上测试，不公开标注信息. 本文与主流方法一样使用 Karpathy 等人[11]提供的划分方式，将 COCO 数据集可供下载的图片划分为训练集、验证集和测试集，其中训练集中包含 113 287 张图片，验证集和测试集各为 5 000 张图片. 此外，通过 COCO 提供的在线测试平台对未公开标注的图片进行在线测试，进一步衡量模型的性能. 采用 5 种在图像描述任务中广泛使用的评价指标来分析生成描述的质量，衡量模型性能，包括 BLEU（bilingual evaluation understudy），METEOR（metric for evaluation of translation with explicit ordering），ROUGE（recall-oriented understudy for gisting evaluation），CIDEr（consensus-based image description evaluation），SPICE（semantic propositional image caption evaluation）. 基于图像描述任务的特点，一般将 CIDEr 得分作为衡量描述语句质量的主要依据.

3.2 实验细节

为了公平对比，本文在设置实验时遵循主流方法中的超参数设置，将编码器中的编码层数量和解码器中的解码层数量均设置为 3，多头注意力的头数

设置为 8, Transformer 内部特征的维度设置为 512.在交叉熵损失训练阶段, 通过 4 轮迭代完成对模型的预热, 在预热时令学习率线性增长到 1×10^{-4}, 在第 5~10 轮迭代学习率固定为 1×10^{-4}, 在第 11~12 轮迭代学习率减小到 1×10^{-5}, 从第 13 轮迭代开始学习率设置为 1×10^{-6}, 并不再改变. 在完成 17 轮迭代后进入强化学习阶段, 强化学习阶段的学习率固定为 5×10^{-6}. 批次大小在交叉熵训练阶段设置为 50, 在强化学习阶段设置为 20, 这 2 个阶段都使用 Adam 优化器进行参数调整, 强化学习和测试阶段通过束搜索的方式生成描述, 束大小设置为 5.

3.3 对比实验

3.3.1 离线测试

表 1 展示了本文方法与当前主流方法在离线测试中的性能对比, 其中第 1 部分是主流方法使用传统视觉特征时的描述性能; 第 2 部分是将传统视觉特征替换为由 CLIP 视觉编码器(主干网络为 ResNet101)提取网格特征后的性能. 由于本文方法主要针对 CLIP 网格特征的不足进行设计, 所以为了保证公平, 主要与第 2 部分进行比较. 可以看到, 本文方法在离线测试中展现出优异的性能, 尤其在图像描述领域中最看重的 CIDEr 方面, 取得了 138.5% 的得分, 与传统视觉特征下表现最出色的 RSTNet 相比提高了 5.2 个百分点, 与利用 CLIP 网格特征复现后的 RSTNet 相比提高了 1.5 个百分点. 此外, 在 BLEU-1, BLEU-4, METEOR, ROUGE 方面取得了超越 RSTNet 的性能, 在 SPICE 方面与之相当.

Table 1 Performance Comparison of Our Method and Various Typical Methods in Offline Test

表 1 本文方法在离线测试中与各种典型方法的性能比较 %

方法	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
up-down[5](CVPR'18)	79.8	36.3	27.7	56.9	120.1	21.4
Transformer[17](ACL'18)	80.5	39.2	29.1	58.7	130.0	23.0
SGAE[12](CVPR'19)	80.8	38.4	27.4	58.6	127.8	22.1
M^2Transformer[13](CVPR'20)	80.8	39.1	29.2	58.6	131.2	22.6
X-Transformer[8](CVPR'20)	80.9	39.7	29.5	59.1	132.8	23.4
GET[14](AAAI'21)	81.5	39.5	29.3	58.9	131.6	22.8
APN[15](ICCV'21)		39.6	29.2	59.1	131.8	23.0
RSTNet[18](CVPR'21)	81.1	39.3	29.4	58.8	133.3	23.0
ReFormer[19](MM'22)	82.3	39.8	29.7	59.8	131.9	23.0
GAT[20](Expert Syst. Appl.'22)	80.8	39.7	29.1	59.0	130.5	22.9
CLIP-ViL[9](ICLR'22)		40.2	29.7		134.2	23.8
up-down*	81.3	39.4	29.2	59.3	131.9	22.8
Transformer*	81.6	40.6	29.9	59.8	136.2	**23.9**
RSTNet*	82.0	40.4	30.0	59.7	137.0	23.7
DSC-Net(本文)	**82.9**	**41.1**	**30.1**	**59.9**	**138.5**	23.7

注: 粗体数值表示最优值.带"*"的方法为本文复现的结果, 在复现中采用了与本文方法相同的环境, 因此对比更加公平.

本文方法在 CLIP 网格特征下超越了 RSTNet, 同时 RSTNet 在传统视觉特征下表现最佳, 这表明本文方法在结构设计方面超越了以往所有方法, 具有显著优势.

3.3.2 在线测试

通过将模型对官方测试集的预测结果提交到在线服务器, 进一步对本文方法进行客观公正的评估. 表 2 展示了在线测试中本文方法与当前主流方法的表现, 其中每个指标都展示了在 5 个参考字幕(c5)和 40 个参考字幕(c40)下的得分情况. 由于排行榜上展示的大多是集成模型的预测结果, 因此本文也参照之前的方法使用 4 个模型的集成进行在线测试. 可以看到, 本文提出的 DSC-Net 在图像描述领域中最看重的是在 CIDEr 上对于 c5 和 c40 分别取得了 135.8% 和 137.6% 的成绩, 与表现优秀的 RSTNet 相比在 CIDEr 的 c5 和 c40 上分别增长了 3.9 个百分点和 3.6 个百分点.

由于在线测试榜单上的其他方法没有使用 CLIP 网格特征, 为了保证公平, 本文对使用 CLIP 网格特征复现的 RSTNet 模型同样进行了在线测试, 令其与本文方法的单一模型进行在线测试结果比较, 可以看到本文提出的 DSC-Net 在单一模型的大部分指标上都取得了最优的性能, 特别是在 CIDEr 的 c5 上取得了 132.7 个百分点的成绩, 与 RSTNet 相比提高了 1.3 个百分点. 在线测试的结果充分证明了本文提出模块的有效性和本文方法的先进性.

3.4 消融实验

3.4.1 模块有效性分析

为了进一步评估本文方法的性能, 验证本文提出各模块的有效性, 对模型进行消融实验, 表 3 展示了针对本文提出模块进行消融实验的结果. 基于图像任务的需要, 遵循 Guo 等人[7]的方法对视觉特征进行相对位置编码, 并以此作为本次实验的基线模型, 以图像描述中最为看重的 CIDEr 得分作为评价指标. 可以看到, 仅使用基线模型可以取得 136.3% 的 CIDEr 得分, 增加 DSCS 模块后取得 137.4% 的 CIDEr 得分, 增加 DSCC 模块后可以取得 136.6% 的 CIDEr 得分, 同时增加 DSCS 和 DSCC 时可以取得

Table 2 Performance Comparison of Our Method and Various Typical Methods in Online Test

表 2 本文方法在在线测试中与各种典型方法的性能比较 %

方法	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
up-down[5](CVPR'18)	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE[12](CVPR'19)	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet[21](ICCV'19)	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
X-Transformer[8](CVPR'20)	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
M^2Transformer[13](CVPR'20)	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
GET[14](AAAI'21)	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
RSTNet[18](CVPR'21)	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
ReFormer[19](MM'22)	82.0	96.7					40.1	73.2	29.8	39.5	59.9	75.2	129.9	132.8
GAT[20](Expert Syst.Appl'22)	81.1	95.1	66.1	89.7	51.8	81.5	39.9	71.4	29.1	38.4	59.1	74.4	127.8	129.8
RSTNet[18](单一模型)	81.8	96.2	66.6	91.0	51.9	82.7	39.7	72.5	29.7	39.0	**59.4**	**74.5**	131.4	134.4
DSC-Net(本文单一模型)	**81.9**	**96.3**	**66.6**	**91.1**	**51.9**	**82.8**	**39.7**	**72.8**	**29.7**	**39.2**	59.3	74.4	**132.7**	**135.3**
DSC-Net(本文集成模型)	**82.8**	**96.9**	**67.8**	**92.3**	**53.2**	**84.5**	**40.9**	**74.7**	**30.1**	**39.6**	**60.1**	**75.3**	**135.8**	**137.6**

注：粗体数值表示最优值.

Table 3 Ablation Results on Modules

表 3 关于模块的消融结果 %

网络结构				CIDEr
Baseline	DSCS	DSCC	DSF	
✔				136.3
✔	✔			137.4
✔		✔		136.6
✔	✔	✔		137.1
✔	✔		✔	137.7
✔		✔	✔	137.5
✔	✔	✔	✔	**138.5**

注：粗体数值表示最优值；“✔”表示应用此模块.

137.1% 的 CIDEr 得分，同时增加 DSCS 和 DSF 时可以取得 137.7% 的 CIDEr 得分，同时增加 DSCC 和 DSF 时可以取得 137.5% 的 CIDEr 得分，同时使用所有模块时可以取得 138.5% 的 CIDEr 得分.

3.4.2 敏感性分析

在本文提出的 DSCC 模块中，引入了超参数 λ 来对矩阵中元素的值进行限制，为了得到合适的超参数取值，通过为 λ 赋予不同的值进行实验，同时进行敏感性分析，同样以 CIDEr 作为评价指标，实验结果如图 5 所示. 可以看到，当 λ=0.04 时取得了最高的 CIDEr 分数，所以采用 0.04 作为超参数的取值，且 λ 的变化对 CIDEr 分数的影响不大，所以认为模型对超参数的变化不敏感.

3.4.3 双重语义协作机制有效性分析

为了评估双重语义协作机制对提升描述性能的影响，设置实验进行验证，实验结果如表 4 所示. 其中第 1 行展示了仅使用网格特征进行描述的结果，可以取得 136.3% 的 CIDEr 分数，以此作为基线；第 2 行展示了同时使用区域特征和网格特征，但不使用双重语义协作机制的结果，可以取得 136.8% 的 CIDEr 分数，与基线相比提升 0.5 个百分点；第 3 行展示了同时使用区域特征和网格特征，但仅使用双重语义协作自注意力模块的结果，可以取得 137.7% 的 CIDEr 分数；第 4 行展示了同时使用区域特征和网格特征，

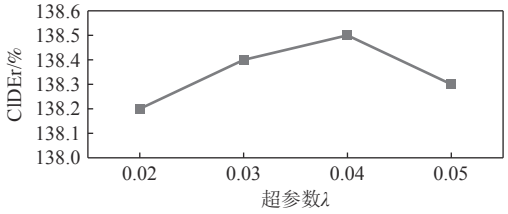


Fig. 5 Impact of hyperparameter on performance of model

图 5 超参数对模型性能的影响

Table 4 Ablation Experiment Results of DSCC and DSCS Modules

表 4 关于双重语义协作机制的消融实验结果 %

方法	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
w/o Region	82.0	40.3	29.8	59.8	136.3	23.5
w/o DSC	82.4	40.6	29.9	59.7	136.8	23.4
w/o DSCC	82.2	40.5	29.9	59.6	137.7	23.5
w/o DSCS	82.5	40.7	30.0	59.9	137.5	23.6
DSC-Net	**82.9**	**41.1**	**30.1**	**59.9**	**138.5**	**23.7**

注：加粗数值表示最优值.

但仅使用双重语义协作交叉注意力模块的结果, 可以取得 137.5% 的 CIDEr 分数; 第 5 行展示了使用区域特征和网格特征, 且同时使用 2 个模块进行处理的结果, 可以取得 138.5% 的 CIDEr 分数, 与基线相比提升 2.2 个百分点. 这表明本文提出 DSC-Net 对描述性能的提升主要得益于对区域和网格特点的准确分析和对双重语义协作结构的巧妙设计.

3.4.4 以区域为主导的特征融合有效性分析

为了评估以区域为主导的特征融合机制对提升描述性能的影响, 设置实验进行验证, 实验结果如表 5 所示. 其中第 1 行展示禁用 DSF 模块后 DSC-Net 的性能, 可以取得 137.1% 的 CIDEr 分数, 与表 4 中 w/o DSC 一栏展示的结果相比在性能方面提升不大; 第 2 行展示了在特征融合时将以区域特征为主导改为以网格特征为主导的结果, 记为 w/o RFD, 可以取得 137.3% 的 CIDEr 分数, 与禁用 DSF 模块相比提升 0.2 个百分点; 第 3 行展示了特征融合时以区域为主导的结果, 可以取得 138.5% 的 CIDEr 分数, 与禁用 DSF 模块相比提升 1.4 个百分点. 这证明了以区域为主导的特征融合有助于解决相关性冲突问题, 更证明了在双重语义融合模块中设置 3 种机制保证区域特征主导地位的合理性.

表 5 表明相关性冲突确实存在于双重语义协作过程中, 且会影响描述性能的提高. 进行特征融合可以缓解相关性冲突, 且融合过程中以区域为主导能够取得更大的性能提升.

3.5 定性分析

为了更直观地展示本文提出的模型在描述图像内容方面的出色性能, 特别是展示经过本文提出的双重语义协作机制处理后模型对待描述图像中重要目标的准确表述能力, 图 6 展示了本文提出的模型和基线模型生成的描述语句, 并与人工标注的语句进行对比. 如图 6(a)所示, 基线模型认为包含 2 头大象(two elephants), 本文提出的模型则认为是 1 头大象和 1 头牛(a elephant and a cow); 如图 6(b)所示, 基线模型认为包含 1 个年轻的男孩(a young boy), 本文提出的模型则判断出是 1 个棒球运动员(a baseball player); 如图 6(c)所示, 基线模型认为女人和孩子在行李旁边(next to bunch of luggage), 本文提出的模型则关注到孩子在拿着行李(child with her luggage); 如图 6(d)所示, 基线模型认为一群大象在动物园, 本文提出的模型则关注到有人出现(with people). 综上所述, 与基线模型相比, 本文提出的模型更容易关注到待描述图片中的目标和目标之间的关系, 所以对图像的描述结果往往更加符合其内容, 同时, 这也表明了本文提出的双重语义协作机制可以有效地解决网格特征对目标缺乏关注的问题.

Table 5 Ablation Experiment Results of Feature Fusion

表 5 特征融合的消融实验结果 %

方法	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
w/o DSF	82.2	40.6	30.0	59.7	137.1	23.6
w/o RFD	82.1	40.5	29.9	59.7	137.3	23.6
DSC-Net	**82.9**	**41.1**	**30.1**	**59.9**	**138.5**	**23.7**

注: 加粗数值表示最优值.

4 结 论

针对网格特征中缺少目标层面语义信息的问题, 本文提出一种基于 DSC-Net 的图像描述方法. 通过 DSCS 进行目标间的语义关联感知, 然后将其映射到

GT: An elephant and a rhino are grazing in an open wooded area.

基线: Two elephants walking in a field of grass.

本文: A elephant and a cow grazing in a field.

(a) 示例1

GT: A baseball player holding a bat while standing in a field.

基线: A young boy swinging a bat at a baseball game.

本文: A baseball player bolding a bat on a field.

(b) 示例2

GT: Women hugging a small girl with green and blue luggage.

基线: A woman holding a child next to bunch of luggage.

本文: A woman is hugging a child with her luggage.

(c) 示例3

GT: People are watching four elephants in a zoo.

基线: A group of elephants in a zoo.

本文: A herd of elephants in a zoo with people.

(d) 示例4

Fig. 6 Comparison of prediction results between DSC-Net and baseline

图 6 DSC-Net 与基线的预测结果对比

对应位置的网格上，融入网格特征的语义建模过程，增强网格特征对目标层面语义的表达能力；通过DSCC计算文本特征与区域特征之间的语义相关性，然后映射到对应位置的网格上，使视觉文本的跨模态建模过程可以综合网格和目标2个层面的语义信息构建上下文感知的视觉特征表示，最后将特征表示中蕴含的视觉语义信息精准地转化为自然语言；通过DSF进行以区域为主导的特征融合，在保留目标感知能力的同时融入网格信息，避免双重语义协作过程中可能出现的相关性冲突问题. 实验结果表明，本文提出的方法有效提高了图像描述模型预测描述语句的准确度，与目前的主流方法相比具有明显优势. 在下一步工作中，将探索更多为网格特征补充目标语义信息的策略，减少对区域特征的依赖，提高描述速度以便应用到轻量化的移动智能终端设备上.

作者贡献声明：江泽涛提出了论文整体思路并负责撰写和修改论文；朱文才完成算法设计与实验并撰写与修改论文；金鑫提出指导意见并修改论文；廖培期负责图表绘制；黄景帆参与了论文审阅与格式校正.

参 考 文 献

[1] Li Zhixin, Wei Haiyang, Zhang Canlong, et al. Research progress on image captioning[J]. Journal of Computer Research and Development, 2021, 58(9): 1951−1974 (in Chinese)
(李志欣, 魏海洋, 张灿龙, 等. 图像描述生成研究进展[J]. 计算机研究与发展, 2021, 58(9): 1951−1974)

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2017 : 5998−6008

[3] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156−3164

[4] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//Proc of the 32nd Int Conf on Machine Learning. New York: ACM, 2015: 2048−2057

[5] Anderson P, He Xiaodong, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6077−6086

[6] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[7] Guo Longteng, Liu Jing, Zhu Xinxin, et al. Normalized and geometry-aware self-attention network for image captioning[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10327−10336

[8] Pan Yingwei, Yao Ting, Li Yehao, et al. X-linear attention networks for image captioning[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10971−10980

[9] Shen Sheng, Li L H, Tan Hao, et al. How much can CLIP benefit vision-and-language tasks[C/OL]//Proc of the 10th Int Conf on Learning Representations. 2022[2022-01-29].https://openreview.net/forum?id=zf_Ll3HZWgy

[10] Li Yehao, Pan Yingwei, Yao Ting, et al. Comprehending and ordering semantics for image captioning[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 17990−17999

[11] Karpathy A, Li Feifei. Deep visual-semantic alignments for generating image descriptions[C]//Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3128−3137

[12] Yang Xu, Tang Kaihua, Zhang Hanwang, et al. Auto-encoding scene graphs for image captioning[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 10685−10694

[13] Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory transformer for image captioning[C]//Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10578−10587

[14] Ji Jiayi, Luo Yunpeng, Sun Xiaoshuai, et al. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network[C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 1655−1663

[15] Yang Xu, Gao Chongyang, Zhang Hanwang, et al. Auto-parsing network for image captioning and visual question answering[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 2197−2207

[16] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7008−7024

[17] Sharma P, Ding Nan, Goodman S, et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2556−2565

[18] Zhang Xuying, Sun Xiaoshuai, Luo Yunpeng, et al. RSTNet: Captioning with adaptive attention on visual and non-visual words[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 15465−15474

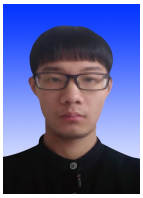
[19] Yang Xuewen, Liu Yingru, Wang Xin. ReFormer: The relational transformer for image captioning[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 5398−5406

[20] Wang Chi, Shen Yulin, Ji Luping. Geometry attention Transformer with position-aware LSTMs for image captioning[J]. Expert Systems with Applications, 2022, 201: 117174

[21] Huang Lun, Wang Wenmin, Chen Jie, et al. Attention on attention for image captioning[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4634−4643

Jiang Zetao, born in 1961. PhD, professor. His main research interests include image processing, computer vision, and artificial intelligence.

江泽涛, 1961 年生. 博士, 教授. 主要研究方向为图像处理、计算机视觉、人工智能.



Zhu Wencai, born in 1999. Master. His main research interests include image processing and computer vision.

朱文才, 1999 年生. 硕士. 主要研究方向为图像处理、计算机视觉.

Jin Xin, born in 1998. Master. His main research interests include image processing and computer vision.

金　鑫, 1998 年生. 硕士. 主要研究方向为图像处理、计算机视觉.

Liao Peiqi, born in 1995. Master. His main research interests include image processing and computer vision.

廖培期, 1995 年生. 硕士. 主要研究方向为图像处理、计算机视觉.

Huang Jingfan, born in 1999. Master. His main research interests include image processing and computer vision.

黄景帆, 1999 年生. 硕士. 主要研究方向为图像处理、计算机视觉.