DOI:10.11918/202404002

跨模态自适应特征融合的视觉问答方法

陈巧红,项深祥,方 贤,孙 麒

(浙江理工大学 计算机科学与技术学院,杭州 310018)

摘 要:为提高视觉问答(VQA)中跨模态融合与交互的精确度,减少多模态特征信息的丢失,提出了一种新颖的基于跨模态自适应特征融合的视觉问答方法。首先,该方法设计了卷积自注意力单元,包含自注意力层和空洞卷积层,前者用于捕捉全局特征信息,后者用于捕捉视觉对象间的空间关系。其次,通过自适应特征融合层,将全局关系与空间关系进行有效结合,使模型在处理图像特征时能够同时考虑全局关系和视觉对象之间的关联性,从而克服了传统注意力机制忽视空间关系的问题。最后,基于不同模态特征在答案预测中贡献程度的差异,该方法还构建了多模态门控融合模块,根据多模态特征间的重要程度自适应地融合特征,减少多模态信息的丢失,同时不会带来额外的计算资源开销。研究结果表明,该方法在未使用额外数据集预训练的情况下,在VQA2.0的测试-开发集、测试-标准集和GQA数据集上的整体准确率分别达到71.58%、72.00%、58.14%,显著优于传统自注意力方法,该研究成果可为跨模态特征融合领域提供了重要的参考和借鉴。

关键词:视觉问答(VQA);特征融合;多模态;注意力机制;门控机制

中图分类号:TP391.41;TP391.1 **文献标志码**:A **文章编号**:0367-6234(2025)04-0094-11

Visual question answering method based on cross-modal adaptive feature fusion

CHEN Qiaohong, XIANG Shenxiang, FANG Xian, SUN Qi

(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: To enhance the accuracy of cross-modal fusion and interaction in visual question answering (VQA) while mitigating the loss of multimodal feature information, we propose a novel cross-modal adaptive feature fusion approach for VQA. First, the method designs a convolutional self-attention unit consisting of self-attention layers and dilated convolution layers-the former captures global feature information while the latter extracts spatial relationships between visual objects. Subsequently, an adaptive feature fusion layer effectively integrates global relationships with spatial correlations, enabling the model to simultaneously consider both global contextual information and inter-object spatial relationships during image feature processing, thereby addressing the limitation of traditional attention mechanisms in overlooking spatial relationships. Furthermore, based on the varying contributions of different modal features to answer prediction, we construct a multimodal gated fusion module that adaptively combines features according to their relative importance, effectively reducing information loss across modalities without introducing additional computational overhead. Experimental results demonstrate that our method achieves overall accuracies of 71.58%, 72.00%, and 58.14% on the VQA2.0 test-dev, test-std, and GQA datasets respectively, significantly outperforming traditional self-attention approaches without requiring additional pre-training datasets. This research provides valuable insights and serves as an important reference for cross-modal feature fusion studies.

Keywords: visual question answering (VQA); feature fusion; multimodal; attentional mechanisms; gating mechanisms

视觉问答(visual question answering,VQA)[1-2]是计算机视觉和自然语言处理交叉领域任务,近年来已成为相关领域的研究焦点。它需要利用问题和图片中的特征构建模型,从而预测出问题所对应的答案。与其他图像任务不同,视觉问答任务要求完全理解图像中的信息来获得准确的答案。视觉问答任务的问题类型多样,涵盖了图像识别、目标检测、图像分割等多个图像任务。例如“图片中这个物体是什么?”是一个图像识别问题,而“图中有几艘船?”是一个目标检测[3]和图像分割[4]的问题。此外,视觉问答还包含复杂的问题,需要考虑空间关系和常识推理,例如“男人和电视之间是什么?”和“女人为什么哭?”,因此,视觉问答是一个复杂的任务,包含了多个人工智能问题。在视觉问答领域,若模

收稿日期:2024-04-01;**录用日期**:2024-06-03;**网络首发日期**:2025-03-14
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20250314.1036.004
基金项目:浙江省自然科学基金(LQ23F020021)
作者简介:陈巧红(1978—),女,博士,教授
通信作者:方 贤,xianfang@zstu.edu.cn

型不能实现文本和视觉特征之间细粒度的跨模态交互,将会限制模型建立模态信息之间的关联,进而导致模型输出错误的答案, 因此,进行有效的跨模态交互与融合是视觉问答研究的关键所在。

视觉问答技术中往往采用注意力机制来进行跨模态特征间的交互。注意力机制[5]是深度学习领域近年来的一项新发展,已经在视觉[6]、文本[7]和语音[8]等单模态任务中取得了显著成功,随后被研究者用于多模态领域。在相关研究中,Chen 等[9]提出了使用视觉注意力从问题中学习图像区域的初步方法,但是该方法没有使用文本注意力提前提取问题中的关键词。随后,Anderson 等[10]提出自底向上 - 自顶向下(bottom-up and top-down,BUTD)的注意力机制。具体而言,自底向上利用预训练的目标检测模型来提取显著区域特征,慢慢替代了利用卷积神经网络生成的网格特征。然而,BUTD 忽略了不同模态特征间的交互。为解决该问题,Yang 等[11]提出了堆叠注意力网络(stacked attention networks,SAN)机制。SAN 利用堆叠注意力进行多模态特征间的交互,但交互过程相对粗糙。针对多模态交互粗糙的问题,Kim 等[12]提出了双线性注意力网络(bilinear attention networks,BAN),BAN 能够实现任意图像区域特征和任意问题文本特征的密集交互。然而 BAN 随着参数量的增加,性能却得不到明显地提高。为解决该问题,鉴于自然语言处理中 Transformer 的成功,Yu 等[13]提出了深度模块化共同注意网络(modular co-attention networks,MCAN)来学习模态内和模态间的交互关系。随着注意力层数的增加,MCAN 的性能提升显著,但该模型在设计上忽视了文本特征对视觉特征的潜在影响。除此之外,MCAN 中的多模态特征融合只是简单地采用向量逐元素求和,缺乏更有效的跨模态特征融合,这会影响模型输出答案的准确率, 因此需要探索更有效的多模态特征融合方式来提高模型性能。为解决 MCAN 忽视视觉特征对文本特征影响的问题,陈巧红等[14]提出了一种基于多模态门控自注意力机制的视觉问答模型(multimodal gate self attention,MGSA)。MGSA 利用其他模态的特征作为通道调节门,以过滤目标模态特征的自注意力学习输出结果。MGSA 改进了跨模态注意力交互,但是模型仍缺乏有效的跨模态融合。为弥补跨模态融合粗糙的问题,Zhang 等[15]提出了一种多头注意力融合网络(multi head attention fusion network,MHAFN),它能够通过多个分支实现多层次的多模态融合,捕捉单词、区域以及它们之间的细粒度和复杂关系。同时,MHAFN 还能够捕捉到不同的注意力分布,以关注推断答案所需的多个不同视觉和文本组件。此外,Zhang 等[16]提出了一种名为 TGAM(transformer gate attention model)的视觉问答模型,用于融合不同模态之间的信息并减少干扰,提高了跨模态融合的精度。

鉴于 bert[17]模型的成功,预训练模型引起了研究者的广泛关注。研究者们开始思考在多模态领域是否可以应用预训练大模型的可能性。在相关研究中,Lu 等[18]提出了首个视觉语言多模态预训练大模型(vision-and-language BERT, ViLBERT)。ViLBERT 模型在 bert 模型的基础上增加了对图像信息的处理,并通过预训练使得模型能够同时理解文本和视觉之间的关系,从而在多模态任务上表现出色。随后,Tan 等[19]基于 bert 模型提出了语言跨模态预训练模型(learning cross-modality encoder representations from transformers,LXMERT)。与 ViLBERT 模型不同的是,LXMERT 模型采用了交叉模态注意力机制,以便模型能够更好地理解图像和文本之间的关系。虽然预训练大模型在提升准确度方面取得了显著的进展,但面临着计算资源需求高、微调困难以及额外数据集需求高等问题。

受到应用于图像字幕中的双重注意力(attention on attention,AoA)模块[20]的启发,Rahman 等[21]提出了深度模块化共同双重注意网络(modular co-attention on attention network,MCAoAN),用双重注意力替换了 MCAN 中的传统的注意力机制,并用注意力网络来进行跨模态融合。在 MCAoAN 中存在两个问题需要改进。首先,在多模态特征交互阶段,受限于注意力机制没法捕捉位置信息,模型往往易忽略视觉对象的空间关系。其次,MCAoAN 在多模态特征融合阶段采用了传统的注意力机制,导致模型的计算复杂度高且对模型精度提升不高。

上述模型存在一个共同的问题,即它们在特征交互阶段忽视了视觉对象之间的空间关系。此外,在跨模态特征融合阶段,这些模型所采用的融合方法都相对较为粗糙,无法准确地捕捉不同模态之间的关联和重要性。针对上述问题,本文提出了基于跨模态自适应特征融合的视觉问答方法。该方法包含卷积注意力交互模块和多模态门控融合模态。其中,卷积注意力交互模块包含了自注意力层和空洞卷积层,分别捕捉全局特征信息和视觉对象间的空间关系,并设计了自适应特征融合层来实现全局关系与空间关系的融合,使得模型在处理图像特征时

能够更全面地考虑全局关系,并且同等重视视觉对象之间的关联性,解决了忽视视觉对象空间关系的问题。此外,多模态门控融合模块以多模态特征间的重要程度来自适应地融合多模态特征,避免特征融合中多模态特征的信息丢失,解决了跨模态融合粗糙的问题。这种模型可以有效地提高视觉问答模型的性能,为视觉问答领域的研究提供了一种新的思路和方法。

1 视觉问答模型的架构设计

如图 1 所示,给定图像及相应的问题,视觉问答的任务是根据图像的内容及问题的语义,从一组候选答案中返回概率最大的答案。基于跨模态自适应特征融合的视觉问答方法主要包含文本特征提取、视觉特征提取、卷积注意力交互模块和多模态门控融合模块 4 个部分。

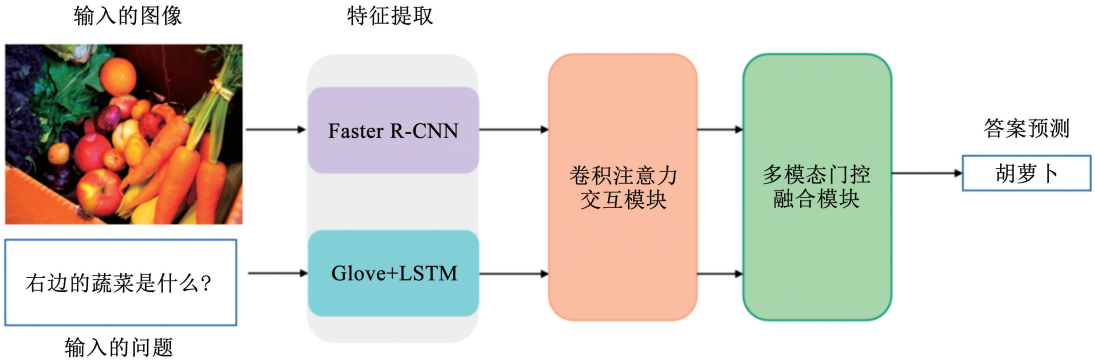

图 1 基于跨模态自适应特征融合的视觉问答模型

Fig. 1 Visual question answering model based on cross-modal adaptive feature fusion

1.1 文本特征提取

在处理 VQA-v2[22] 数据集时,首先对每个问题进行标记。由于 VQA-v2 数据集中大部分的问题的长度不超过 14 个单词,并且只有 0.25% 的问题超过了这个长度限制,因此对于长度不足 14 个单词的问题,使用零填充技术以达到 14 个单词的长度要求,而对于长度超过 14 个单词的问题,将超出部分丢弃。接下来使用一个大规模预训练的 Glove 单词嵌入模型[23]将每个单词转换为单词向量。然后,使用长短期记忆网络(long short-term memory, LSTM)[24]将单词向量转换为文本特征,对于每个单词 i,$\boldsymbol{x}_i$ 为 dx 维的文本特征向量,则输入的文本表示为 $\boldsymbol{X}=[x_1, x_2, \cdots, x_n]$,其中:$n$ 为问题的长度,dx 为 LSTM 输出的维度。

1.2 视觉特征提取

为确保与其他模型的公平比较,本文遵循 Anderson 等[10]在 BUTD 中的视觉特征提取方法。该方法采用了基于 ResNet-101[25]的预训练视觉对象检测器(Faster R-CNN[26]),并在 VG 数据集[27]上进行了训练,用于检测输入图像。对象检测器会选择最相关的 m 个候选区域,对于每个候选区域 j,$\boldsymbol{y}_j$ 是dy 维的视觉特征向量,则输入的图像表示为 $\boldsymbol{Y}=[y_1, y_2, \cdots, y_m]$,其中 $m \in [10, 100]$ 为候选区域的数量。

1.3 卷积注意力交互模块

如图 2 所示,卷积注意力交互模块包含问题编码器和图像解码器,其中包括多个单元块。如图 3 所示,卷积自注意力单元(convolutional self-attention unit, ConvSA)由自注意力层、空洞卷积层和自适应特征融合层组成。本文将重点介绍自注意力层、空洞卷积层、特征融合层以及问题编码器和图像解码器结构。

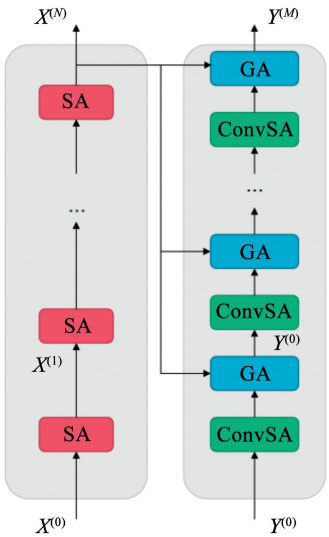

图 2 卷积注意力交互模块

Fig. 2 Convolutional attention interaction module

1.3.1 自注意力层

目前流行的自注意力机制通常是指 Vaswani 等[5]提出的多头标度点积注意力。如图 3 所示的自注意力层的输入包括查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$。为简化计算,通常它们的维度都设置为 d。

先计算查询向量 $\boldsymbol{Q}$ 和键向量 $\boldsymbol{K}$ 之间的相似度分数,再用 Softmax 激活函数来将得分归一化,最后使相似度分数与值向量 $\boldsymbol{V}$ 相乘得到输出向量 $\boldsymbol{F}$。自注意力的计算公式为

$$\boldsymbol{F} = f(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

多头自注意力是指具有 N 个头的并行计算,每个头执行单独的自注意力操作。将这些头部连接在一起以形成最终的全局特征 I。多头自注意力的计算公式为:

$$\mathbf{head}_j = f(\boldsymbol{Q}\boldsymbol{W}_j^{Q},\boldsymbol{K}\boldsymbol{W}_j^{K},\boldsymbol{V}\boldsymbol{W}_j^{V}) \tag{2}$$

$$\boldsymbol{I} = \mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = [\mathbf{head}_1,\mathbf{head}_2,\cdots,\mathbf{head}_N]\boldsymbol{W} \tag{3}$$

式中, $\boldsymbol{W}_j^{Q}$、$\boldsymbol{W}_j^{K}$、$\boldsymbol{W}_j^{V}$、$\boldsymbol{W}$ 分别为权重矩阵。

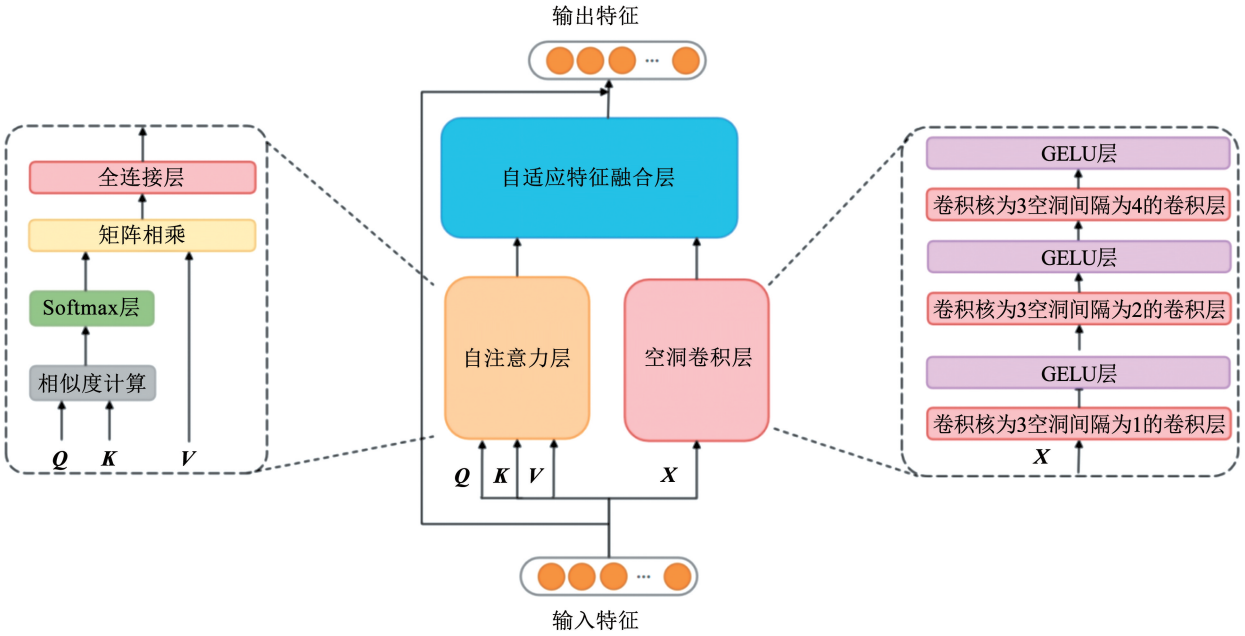

图 3　卷积自注意力单元结构

Fig. 3　Convolutional self-attention unit structure

1.3.2　空洞卷积层

如图 3 所示的卷积层,传统的自注意力机制先计算查询向量 $\boldsymbol{Q}$ 和键向量 $\boldsymbol{K}$ 之间的相似度分数,这导致任意特征间的距离都是相等的,使注意力能够有效地获取全局关系,这也是自注意力机制在自然语言处理中大放异彩的原因。但自注意力机制忽略了视觉对象中的空间关系,依靠卷积神经网络的局部感知性,本文提出在传统自注意力,加入卷积神经网络来获取视觉对象中的空间关系。这样的设计充分利用了自注意力的全局关系和卷积神经网络的局部感知性,从而综合考虑了特征之间的全局和空间关系。如果设计连续的 $n \times n$ 卷积层会导致网络的权重系数变多,所以本文在卷积层中采用空洞卷积的方法来获得指数增长的感受野。

在卷积层设计 3 个卷积核为 3 的空洞卷积层,空洞间隔分别为 1、2、4。为了使空洞卷积层的输入和输出尺寸相同,采用边缘填充,所有的卷积步长为 1。在每一个空洞卷积层之后都加上一个 GELU 激活函数层。通过最后一个空洞卷积层和 GELU 激活函数层的输出为包含丰富视觉对象空间的空间特征 $\tilde{\boldsymbol{X}}$。空洞卷积层的计算过程如下:

$$\boldsymbol{X}' = \mathrm{GELU}(\mathrm{Conv}(\boldsymbol{X})) \tag{4}$$

$$\boldsymbol{X}'' = \mathrm{GELU}(\mathrm{Conv}(\boldsymbol{X}')) \tag{5}$$

$$\tilde{\boldsymbol{X}} = \mathrm{GELU}(\mathrm{Conv}(\boldsymbol{X}'')) \tag{6}$$

式中: Conv 为空洞卷积块,GELU 为激活函数。

1.3.3　自适应特征融合层

受门控机制在图像字幕应用[28]上的启发,本文利用 Sigmoid 激活函数对空间特征 $\tilde{\boldsymbol{X}}$ 进行归一化,得到自适应门向量 $\boldsymbol{G}$。随后,将自注意力层的输出全局特征 $\boldsymbol{I}$ 和自适应门向量 $\boldsymbol{G}$ 根据逐元素乘法相乘,得到融合特征 $\tilde{\boldsymbol{I}}$。由于 Sigmoid 激活函数的输出值为一个 0 ~ 1 的门向量 $\boldsymbol{G}$。当门控向量接近 1 时,逐元素乘法能够保留全局特征 $\boldsymbol{I}$ 中的大部分信息;相反,当门控信号接近 0 时,逐元素乘法的结果会趋向于 0,从而抑制全局特征 $\boldsymbol{I}$。因此,门控结构可以实现对两个输入的融合,其中门控向量决定了两个输入在融合过程中的权重分配,从而实现有效地自适应融合,上述过程分别如下式所示:

$$\boldsymbol{G} = \mathrm{Sigmoid}(\tilde{\boldsymbol{X}}) \tag{7}$$

$$\tilde{\boldsymbol{I}} = \boldsymbol{I} \odot \boldsymbol{G} \tag{8}$$

式中⊙为元素乘法。整个卷积注意力 ConvSA 如下:

$$\mathrm{ConvSA}(\boldsymbol{X}) = \mathrm{MultiHead}(\boldsymbol{X},\boldsymbol{X},\boldsymbol{X}) \odot \mathrm{Sigmoid}(\tilde{\boldsymbol{X}}) \tag{9}$$

1.3.4 问题编码器和图像解码器结构

受 Transformer[5] 中编码器 – 解码器结构的启发,本文设计了问题编码器和图像解码器。如图 2 所示,问题编码器的输入是文本特征 $\boldsymbol{X}$,文本特征 $\boldsymbol{X}$ 被递归地传递到 N 个编码器块 $\{b_{\mathrm{enc}}^1, b_{\mathrm{enc}}^2, \cdots, b_{\mathrm{enc}}^N\}$,用于在文本模态内进行交互。以相同的方式,图像解码器的输入是视觉特征 $\boldsymbol{Y}$ 和问题编码器的最后一层输出 $\boldsymbol{X}^N$,它们被递归地传递到 M 个解码器块 $\{b_{\mathrm{dec}}^1, b_{\mathrm{dec}}^2, \cdots, b_{\mathrm{dec}}^M\}$,用于文本和视觉模态之间的交互以及视觉模态内的交互。问题编码器和图像解码器的计算过程分别如下:

$$\boldsymbol{X}^i = b_{\mathrm{enc}}^i(\boldsymbol{X}^{i-1}) \tag{10}$$

$$\boldsymbol{Y}^j = b_{\mathrm{dec}}^j(\boldsymbol{Y}^{j-1}, \boldsymbol{X}^N) \tag{11}$$

式中:$i \in \{1,2,\cdots,N\}$,$j \in \{1,2,\cdots,M\}$,$\boldsymbol{X}^0 = \boldsymbol{X}$ 和 $\boldsymbol{Y}^0 = \boldsymbol{Y}$,每个 b_{enc}^i 由一个自注意力(self-attention,SA)单元组成,每个 b_{dec}^j 由一个卷积自注意力(ConvSA)单元和导向注意力(guided attention,GA)单元组成。

1.4 多模态门控融合模块

VQA 的数据通常包含视觉模态和文本模态。这些模态中每一种的特征都不能被忽视,忽视一种模态的特征会导致不正确的结果。给定图像和问题,视觉问答的答案通常取决于提问的方式,因此文本特征是主要模态,视觉特征是辅助模态。如图 4 所示,在向量表示空间中,文本向量将基于其特征产生相应的位置为 $\boldsymbol{P}_{\mathrm{a}}$,向量空间位置中的真实答案表示为 $\boldsymbol{P}_{\mathrm{c}}$。这两个位置通常会有相对较大的偏差,如果加上视觉模态偏移向量,那么文本向量将在向量空间中发生较大的偏移,将到达一个新的位置作为 $\boldsymbol{P}_{\mathrm{b}}$,那么 $\boldsymbol{P}_{\mathrm{b}}$ 将更接近 $\boldsymbol{P}_{\mathrm{c}}$。然而,有时候辅助模态的偏移向量并不一定比主要模态的向量更为重要。因此,需要自适应地调整辅助模态的权重,以确保在特征融合过程中,不同模态的贡献能够合理平衡。

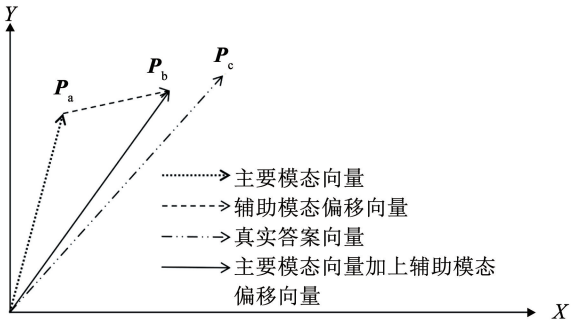

图 4 向量表示空间

Fig. 4 Vector representation space

多模态门控融合模块的结构见图 5。在卷积注意力交互模块之后,输出的文本特征 $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$ 和视觉特征 $\boldsymbol{Y} = [y_1, y_2, \cdots, y_m]$ 已经包含了丰富的注意力信息。给定文本特征 $\boldsymbol{X}$ 和视觉特征 $\boldsymbol{Y}$,采用两个全连接层(fully connected laye,FC)组成的多层感知机(multi-layer perceptron,MLP)和一个 Softmax 激活函数,计算出文本特征权重和视觉特征权重;最后将特征权重与特征 $\boldsymbol{X}$ 和特征 $\boldsymbol{Y}$ 做矩阵相乘。上述过程分别如下式所示:

$$\mathrm{MLP}(\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{W}_x^a)\boldsymbol{W}_x^b \tag{12}$$

$$\mathrm{MLP}(\boldsymbol{Y}) = (\boldsymbol{Y}\boldsymbol{W}_y^a)\boldsymbol{W}_y^b \tag{13}$$

$$\tilde{\boldsymbol{X}} = \sum_{i=1}^{n} \mathrm{Softmax}(\mathrm{MLP}(\boldsymbol{X})\boldsymbol{x}_i) \tag{14}$$

$$\tilde{\boldsymbol{Y}} = \sum_{i=1}^{m} \mathrm{Softmax}(\mathrm{MLP}(\boldsymbol{Y})\boldsymbol{y}_i) \tag{15}$$

式中,$\boldsymbol{W}_x^a$、$\boldsymbol{W}_x^b$、$\boldsymbol{W}_y^a$、$\boldsymbol{W}_y^b$ 分别为权重矩阵。

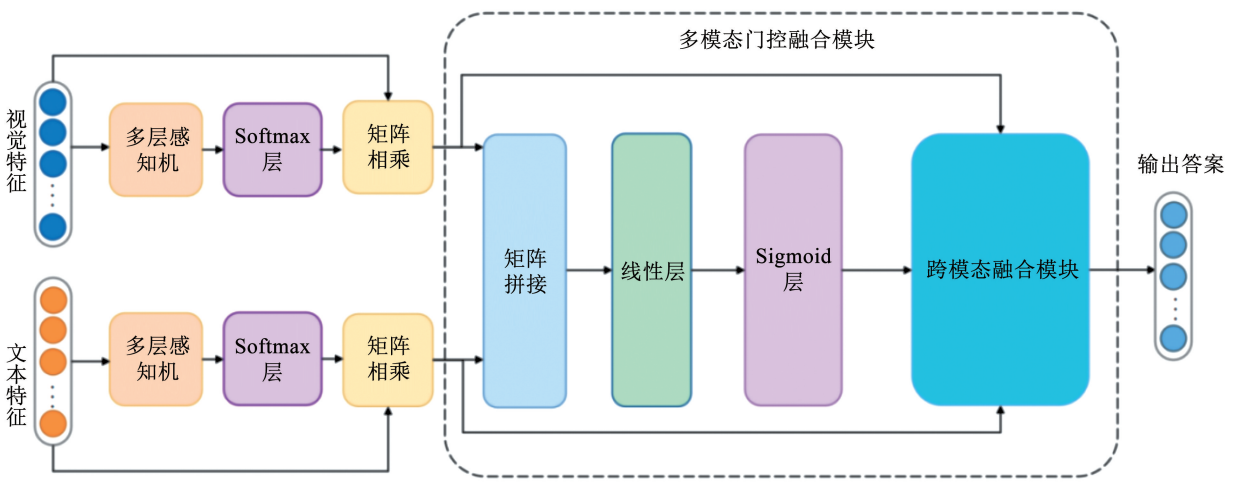

图 5 多模态门控融合模块结构

Fig. 5 Multimodal gated fusion module structure

之后,两种模态的向量捕捉它们各自模态内的相关性,并获得丰富的注意力特征。然后使这两种模态的向量通过多模态门控融合模块。在多模态门控融合模块中,本文首先连接两种模态的向量。再通过一层线性层和 Sigmoid 激活函数生成视觉模态门向量 $\boldsymbol{G}$,视觉模态门向量 $\boldsymbol{G}$ 的计算公式如下:

$$\boldsymbol{G} = \mathrm{Sigmoid}(\boldsymbol{W}[\tilde{\boldsymbol{X}};\tilde{\boldsymbol{Y}}] + \boldsymbol{b}) \tag{16}$$

式中 $\boldsymbol{W}$ 为权重矩阵。

在跨模态融合模块中,将刚才所提取的视觉模态门向量 $\boldsymbol{G}$ 乘以视觉向量 $\tilde{\boldsymbol{Y}}$ 来计算视觉模态偏移向量 $\bar{\boldsymbol{Y}}$,使得 $\bar{\boldsymbol{Y}}$ 的大小随文本特征 $\tilde{\boldsymbol{X}}$ 和视觉特征 $\tilde{\boldsymbol{Y}}$ 之间的相对重要程度而自适应调整。视觉位移向量 $\bar{\boldsymbol{Y}}$ 的计算公式如下:

$$\bar{\boldsymbol{Y}} = \boldsymbol{G}\tilde{\boldsymbol{Y}} \tag{17}$$

通过上述公式,可以得到视觉位移向量 $\bar{\boldsymbol{Y}}$。接着引入了一个大小控制为 0 ~ 1 的缩放因子 γ 来控制视觉偏移向量 $\bar{\boldsymbol{Y}}$ 的大小,这是为了防止视觉偏移向量 $\bar{\boldsymbol{Y}}$ 的大小与文本特征 $\tilde{\boldsymbol{X}}$ 相比过大。随后通过将视觉偏移向量 $\bar{\boldsymbol{Y}}$ 与文本特征 $\tilde{\boldsymbol{X}}$ 相加来生成多模态融合特征向量 $\boldsymbol{F}$。多模态融合特征向量 $\boldsymbol{F}$ 的计算公式为:

$$\gamma = \min\left(\frac{\|\tilde{\boldsymbol{X}}\|_2}{\|\bar{\boldsymbol{Y}}\|_2}, 1\right) \tag{18}$$

$$\boldsymbol{F} = \tilde{\boldsymbol{X}} + \gamma\bar{\boldsymbol{Y}} \tag{19}$$

最后,将多模态融合特征向量 $\boldsymbol{F}$ 通过归一化层和线性层来预测最终的答案。

2　实验及结果分析

为验证本文提出的模型的有效性,将在主流的视觉问答数据集上进行实验。本文首先介绍了两个常用的视觉问答数据集,以及实验细节和参数设置。随后,对比了本文所提出的模型与基准模型,并进行了消融实验以验证各个模块的作用。最后,进行了对提出模型和基准模型的样例分析。

2.1　数据集介绍

为验证本文提出的模型,使用 VQA-v2 和 GQA 数据集进行实验研究。

VQA-v2 数据集[1]来自 COCO 数据集,其中包含图像和问答对。每张图片包含 3 个或 3 个以上的问题,每个问题都有 10 个可能的答案,由不同的人标注,最频繁的答案被视为基本答案。数据集分为训练集、验证集和测试集。其中测试集分为 test-dev 和 test-std。评估指标包括 3 种不同类型问题的准确率(是/否、数字和其他)以及总体准确率。VQA-v2 数据集的具体细节见表 1。

表 1　VQA-v2 数据集中样本的统计

Tab. 1　Statistics of samples in VQA-v2 dataset

类别	图片数	问题数
训练集	82 783	443 757
验证集	40 504	214 354
测试集	81 434	447 793
总数	204 721	1 105 904

GQA 数据集[29]是一个新提出的 VQA 数据集,以真实世界图像的组成问题为特征。它旨在提供视觉理解能力的准确指示,并减轻先前 VQA 数据集中广泛存在的语言先验。与 VQA-v2 数据集相比,GQA 数据集是通过利用视觉基因组场景图结构来生成的,以较少的语言偏见来创建不同的推理问题,因此回答问题需要更复杂的推理技巧。GQA 数据集由两部分组成(即平衡拆分和全分法)。平衡拆分由具有重新采样问答分布的 QA 对组成。按照通常的做法,本文使用平衡拆分进行训练和评估。GQA 数据集被分为 70% 的训练集、10% 的验证集、10% 的测试集和 10% 的挑战集,具体细节见表 2。VQA-v2 数据集的评价指标主要侧重于答案的准确率,而 GQA 数据集的评估指标则更全面,不仅需要考虑答案的准确率,还需评估答案结果的有效性、一致性和合理性,以全面评估模型在视觉问答任务中的性能。

表 2　GQA 数据集中样本的统计

Tab. 2　Statistics of samples in GQA dataset

类别	图片数	问题数
训练集	72 140	943 000
验证集	10 234	132 062
测试集	398	12 578
挑战集	2 987	95 336
总数	85 759	1 182 976

2.2　评价指标

对于 VQA-v2 数据集,评估指标包括 3 种类型问题的准确率(是/否、数字和其他)以及总体准确率。针对“是/否”类型的问题,答案的标注只有两种情况:“是”和“否”。此类型的问题不存在含义相同但表述不同的情况。然而,在某些情况下,尽管预测的答案与标注不同,但从人工评判的角度来看,它们可以被认为具有相同的含义,因此被视为正确答案。因此,本文遵循 Agrawal 等[1]的工作,使用投票机制对预测答案的准确率进行打分,具体如下:

$$A_{cc}(a) = \min\left(\frac{\text{count}(a)}{3},1\right) \tag{20}$$

式中：$A_{cc}(a)$为答案 a 是否为正确答案，count(a)为 10 个不同标注人员对答案 a 投票的数量。仅当前所预测的答案在 10 个人工标注答案中占 3 个及以上时，才认为是正确答案。其中与准确率相关的评价指标都依赖于上述投票机制得到。

对于 GQA 数据集，除了评估二元问题（是/否问题）、开放问题（数字和其他）和总体的准确率等评价指标外，还引入了 3 个额外的指标来进一步评估模型，即有效性、合理性和一致性。一致性用于检验模型在面对不同问题时的回答是否一致，即对于一个新问题的答案不应该与之前的答案相矛盾；有效性用于检验模型的回答是否在问题涉及的范围内；合理性用于检验模型的回答是否符合常识。

为了能够客观地反映模型的复杂度和资源消耗情况，本文增加了计算复杂度和资源消耗相关的指标，例如模型参数量和计算量（FLOPs）。模型参数量对应了模型的空间复杂度，而模型计算量对应了模型的时间复杂度。参数量和计算量为评估模型性能和确定实际部署可行性提供了重要依据。

2.3 实验细节与参数设置

本文模型在 2 块 NVIDIA GeForce RTX 3090 显卡上进行训练和测试，操作系统为 Ubuntu 16.04，使用 Python 3.9 作为编程语言，深度学习框架为 PyTorch 1.13.0。

本文的实验参数遵循 MCAN 提出的实验方案。模型超参数如下：所有自注意力单元中隐藏层的维度为 512，答案的词汇量大小为 3 129，模型基础学习率设置为 $\min(1.25\times10^{-5}T,5.00\times10^{-5})$，其中 T 为训练周期数。模型一共训练 12 个周期，在 10 个训练周期之后，每个训练周期的学习率下降为原来的 1/10。此外，本文使用交叉熵（binary cross-entropy，BCE）作为提出模型的损失函数。

2.4 基准模型

为验证本文所提出的模型，将模型与之前的 SOTA 模型进行了比较，其中包括经典的基线模型。具体模型如下。

1）BUTD[10]。将描述生成问题转化为图像生成问题，并使用 bottom-up 和 top-down 注意力机制来生成图像描述和回答视觉问答问题。这种方法在图像描述生成和视觉问答问题上都取得了很好效果。

2）QG-SRGR[30]。利用门控图推理网络的信息传递机制和控制特征信息的聚合，让模型具有关系推理能力。

3）BAN[12]。提出了双线性注意力网络，用于学习问题词与图像区域之间的稠密互动关系。

4）MCAN[13]。基于 Transformer 架构，提出深度模块化共同注意力网络，使用自注意力和导向注意力来进行模态内和模态间的交互。MCAN 获得 2019 年 VQA 挑战赛的冠军。

5）MCAoAN[21]。在 MCAN 的基础上改进了注意力，使用 attention on attention 模块来替换传统注意力模块，过滤了图像中的干扰信息。

6）TCAN[31]。在 MCAN 的基础上改进了注意力，针对自注意力层增加了位置编码，使得自注意力可以捕获位置信息。

7）CIFN[32]。基于 MCAN 优化，在注意力交互模块之前增加跨模态信息过滤模块，解决了多数方法忽略模态间深层语义信息的问题。

8）Vilbert[18]。将自然语言处理中流行的 bert 架构扩展到多模态双流模型中，用大量的额外数据集进行预训练。

9）LXMERT[19]。大型 Transformer 预训练模型，用大量的图像和文本数据集对模型进行预训练，在 VQA-v2 和 GQA 数据集上获得了最先进的结果。

2.5 结果对比分析

表 3 展示了本文提出的模型与基准模型在 VQA-v2 数据集上的实验结果。结果显示，本文提出的模型优于以往的 SOTA 模型。在测试 - 标准集上，相比于基准模型 BUTD，总体精度提升了 6.26%。表 3 的第 1、3、5、7、8 行表明，目前较为流行的基于传统自注意力的视觉问答模型虽然在模态内和模态间的交互上表现出一定的有效性，但这些模型忽略了视觉对象的空间关系，因此在性能上仍有明显的不足。而本文中引入的卷积注意力交互模块则能够同时捕获全局关系与空间关系特征，并通过自适应融合来提升模型输出答案的准确率。表 3 的第 6 行表明，尽管目前主流的基于注意力机制的模态融合方法在一定程度上提升了模型性能，但注意力机制的计算成本较高。相对而言，本文采用的多模态门控融合模块在计算资源需求低，且跨模态融合效果并不弱于注意力机制。表 3 的第 9 行表明，尽管本文提出的模型的参数量仅为 LXMERT 的 1/3，且未使用额外数据集对模型进行预训练，但仍具有相当的竞争力。本文提出的模型在更少的参数量下能够取得相当的效果。

表 3　VQA-v2 数据集上所提的模型与基准模型的对比结果

Tab. 3　Comparison of proposed model and baseline model on the VQA-v2 dataset

对比模型	参数量/M	是否使用额外数据集预训练	测试 - 开发集/%			测试 - 标准集/%	
			总体	是/否	计数	其他	总体
BUTD[10]	25	否	65.32	81.82	44.21	56.05	65.67
QG-SRGR[30]		否	66.98	82.82	47.68	56.62	67.34
BAN + Counter[12]	91	否	70.04	85.42	54.04	60.52	70.53
Vilbert[18]		是	70.55				70.92
MCAN[13]	58	否	70.63	86.82	53.26	60.72	70.90
MCAoAN[21]	72	否	70.90	87.05	53.81	60.97	71.14
TCAN[31]		否	70.92	87.21	53.94	60.93	
CIFN[32]		否	71.27	87.43	53.82	61.42	71.51
LXMERT[19]	228	是	**72.42**				**72.54**
本文模型	71	否	71.58	**87.47**	**55.51**	**61.62**	72.00

为了评估基于跨模态自适应特征融合的视觉问答方法的泛化能力,本文在 GQA 数据集上进行了实验。表 4 展示了本文提出的模型和基准模型在 GQA 数据集上的实验结果。结果表明,在 GQA 数据集上,本文提出的模型仍然能够取得良好的效果。

表 4　GQA 数据集上所提的模型与基准模型的对比结果

Tab. 4　Comparisonof proposed model and baseline model on the GQA dataset　%

对比模型	总体	二元问题	开放问题	有效性	合理性	一致性
BUTD[10]	53.38	67.78	40.72	96.62	84.81	77.62
BAN[12]	56.19	73.31	41.13	96.77	**85.58**	81.85
MCAN[13]	56.64					
MCAoAN[21]	56.74	74.45	41.18	96.72	85.49	85.74
LXMERT[19]	**60.00**					
本文模型	58.14	**76.91**	**41.63**	**96.85**	85.47	**88.13**

2.6　消融实验

本文提出模型由多个模块组成,为了分析各个模块在模型中的作用,本文在 VQA-v2 数据集上对完整模型进行消融实验分析,探讨每个模块的作用。视觉问答模型的变体如下:

1)基线模型。本文采用 MCAN 作为基线模型。MCAN 是视觉问答领域中经典的基线模型,其通过传统的注意力机制来实现模态内和模态间的特征交互,同时采用向量逐元素相加来实现跨模态特征融合。

2)基线模型 + 卷积注意力交互模块。该模型在视觉编码器一侧采用卷积自注意力来替换传统的自注意力,但在其他方面与基线模型保持一致。

3)基线模型 + 多模态门控融合模块。该模型采用多模态门控融合来替换向量逐元素相加,但在其他方面与基线模式保持一致。

4)基线模型 + 卷积注意力交互模块 + 多模态门控融合模块。该模型为提出的完整模型。

消融实验在 VQA-v2 数据集和 GQA 数据集上进行,实验结果见表 5、6。

表 5　所提模型在 VQA-v2 数据集上的消融实验结果

Tab. 5　Ablation experimental results of the proposed model on the VQA-v2 dataset

模型	参数量/M	计算量/M	总体/%	(是/否)/%	计数/%	其他/%
基线模型	58	1 034	70.63	86.82	53.26	60.72
基线模型 + 卷积注意力交互模块	71	1 664	71.27	87.31	55.33	61.15
基线模型 + 多模态门控融合模块	58	1 034	70.95	87.14	53.00	61.19
基线模型 + 卷积注意力交互模块 + 多模态门控融合模块	71	1 664	**71.58**	**87.47**	**55.51**	**61.62**

表 6　所提模型在 GQA 数据集上的消融实验结果

Tab. 6　Ablation experimental results of the proposed model on the GQA dataset　%

模型	总体	二元问题	开放问题	有效性	合理性	一致性
基线模型	56.72	75.17	40.49	96.73	85.41	88.03
基线模型 + 卷积注意力交互模块	57.46	76.03	41.13	96.83	85.34	87.56
基线模型 + 多模态门控融合模块	57.27	75.28	41.43	**96.96**	**85.53**	88.04
基线模型 + 卷积注意力交互模块 + 多模态门控融合模块	**58.14**	**76.91**	**41.63**	96.85	85.47	**88.13**

表 5 的实验结果表明,在基线模型中采用卷积自注意力来替换传统自注意力有助于提高总体准确率,特别是对于“计数”和“其他”类型的问题,这是由于卷积自注意力可以同时提取视觉对象的全局关系和空间关系,并且在自适应特征融合层里自适应地调整对象的全局关系和空间关系之间的比重,从而提升视觉特征质量。但卷积自注意力增加了模型的参数量和计算量,会带来额外的计算资源消耗。若仅通过简单的向量逐元素相加来实现跨模态特征融合,将极大地限制模型的潜力。为了更有效地融合文本模态和视觉模态的信息,本文引入了多模态门控融合模块,使得模型能够基于不同模态的重要性程度,自适应地调整并分配各自的权重,从而优化模型的性能。且多模态门控融合模块的参数量和计算量相比注意力模块可以忽略不计,能够有效地提高模型性能。

表 6 的实验结果表明,卷积注意力交互模块对模型的准确度提升显著,而多模态门控融合模块提升了输出答案的有效性和合理性,在两个模块的协作下,所有指标都得到了显著的提高。在 VQA-v2 数据集和 GQA 数据集上整体准确率比基线模型分别提高了 0.95% 和 1.42%,证明了卷积注意力交互模块和多模态门控融合模块的有效性。

为了探索最佳的卷积自注意力层的架构,本文构建了卷积自注意力层的不同变体,见表 7。实验结果显示,将卷积自注意力放入文本编码器中会降低模型的性能,这是因为文本数据缺乏空间关系。相反,将卷积自注意力用于视觉解码器一侧,可以提高模型的性能,因为卷积操作可以有效地捕捉视觉对象的空间关系,使提取到的特征信息更加丰富。因此,本文的模型在文本编码器中使用传统的自注意力,在视觉解码器一侧使用卷积自注意力,这种设计有助于提高模型的表现。

为了更好地研究文本模态和视觉模态哪个更适合作为基于跨模态自适应特征融合的视觉问答方法的主要模态,本文构建了多模态门控融合模块的不同变体(见表 8),其他设置保持不变。结果表明,将文本模态设置为主要模态可以提供更好的性能。这是因为问题中的词汇数量较少,而图片中目标对象的数量很多,因此相对于视觉模式,文本模式更加重要,将文本模态设置为主要模态。

表 7 卷积注意力交互模块的不同变体在 VQA-v2 数据集上的实验结果

Tab. 7 Experimental results of different variants of convolutional attention interaction module on the VQA-v2 dataset

文本编码器	视觉解码器	总体/%	(是/否)/%	计数/%	其他/%
传统自注意力	传统自注意力	70.86	87.08	53.11	61.03
卷积自注意力	传统自注意力	70.13	86.08	52.59	60.49
传统自注意力	卷积自注意力	**71.58**	**87.47**	**55.51**	**61.62**
卷积自注意力	卷积自注意力	70.58	86.54	52.76	60.99

表 8 多模态门控融合模块的不同变体在 VQA-v2 数据集上的实验结果

Tab. 8 Experimental results of different variants of multi-modal gated fusion module on the VQA-v2 dataset

文本模态	视觉模态	总体/%	(是/否)/%	计数/%	其他/%
主要模态	辅助模态	**71.58**	**87.47**	**55.51**	**61.62**
辅助模态	主要模态	71.37	87.36	55.35	61.33

2.7 样例分析

为了体现本文所提出模型的泛化性能与其对多模态预测的准确率,从 VQA-v2 数据集中选取了一些样例进行测试。图 6 展示了本文提出模型和 MCAN 获得的一些定性结果。结果得出本文提出的模型比 MCAN 在“是/否”和“计数”的准确率更高。MCAN 的特征交互方法为传统自注意力,多模态特征融合使用的向量相加的方式来进行模态融合。而本文所提出的模型在跨模态特征交互上采用空洞卷积结合自注意力的方式,弥补了 MCAN 中忽视视觉对象空间关系的缺点。此外,本文在跨模态特征融合阶段采用多模态门控融合网络来对多模特特征进行跨模态自适应融合。

图 6　用 MCAN 和本文提出模型获得的一些定性结果的样本

Fig. 6　Qualitative results samples obtained with MCAN and the proposed model

3　结　论

1)通过改进传统自注意力单元,设计了卷积自注意力单元。该单元块中的自注意力层和空洞卷积层分别用于捕捉全局特征信息和视觉对象间的空间关系,然后通过自适应特征融合层来自适应地融合视觉特征中的全局关系与空间关系,有效地解决大多数视觉问答任务中传统自注意力忽视视觉对象空间关系的问题。

2)基于不同模态特征在答案预测中贡献程度的差异,设计了多模态门控融合模块。该模块引入门控机制,按照多模态特征的重要程度自适应地融合文本和视觉模态特征,且不会带来额外的计算资源开销。

3)通过实验验证,本文提出的方法有效提升了视觉问答模型的性能。本文提出的模型在不使用额外数据集预训练的情况下,在 VQA-v2 数据集上达到 72% 的准确率,具有一定竞争力。

参考文献

[1]AGRAWAL A, LU Jiasen, ANTOL S, et al. VQA: visual question answering[J]. International Journal of Computer Vision, 2017, 123(1): 4. DOI: 10.1007/s11263-016-0966-6

[2]张丰硕, 李豫, 李向前, 等. 一种消减多模态偏见的鲁棒视觉问答方法[J]. 北京大学学报(自然科学版), 2024, 60(1): 23
ZHANG Fengshuo, LI Yu, LI Xiangqian, et al. Reducing multi-model biases for robust visual question answering [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2024, 60(1): 23. DOI: 10.13209/j.0479-8023.2023.072

[3]冯俊健, 李彬, 田联房, 等. 多视图交叉一致性学习的半监督水面目标检测[J]. 哈尔滨工业大学学报, 2023, 55(4): 107
FENG Junjian, LI Bin, TIAN Lianfang, et al. Semi-supervised surface object detection based on multi-view cross-consistency learning[J]. Journal of Harbin Institute of Technology, 2023, 55(4): 107. DOI: 10.11918/202201067

[4]房超, 王小鹏, 李宝民, 等. 基于自适应结构元素的改进分水岭图像分割方法[J]. 哈尔滨工业大学学报, 2023, 55(5): 59
FANG Chao, WANG Xiaopeng, LI Baomin, et al. Improved watershed image segmentation method based on adaptive structural elements[J]. Journal of Harbin Institute of Technology, 2023, 55(5): 59. DOI:10.11918/202204057

[5]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: Curran Associates Inc., 2017: 6000. DOI: 10.48550/arXiv.1706.03762

[6]郭玲, 于海雁, 周志权. 基于 SimAM 注意力机制的近岸船舶检测方法[J]. 哈尔滨工业大学学报, 2023, 55(5): 14
GUO Ling, YU Haiyan, ZHOU Zhiquan. Nearshore ship detection method based on SimAM attention mechanism[J]. Journal of Harbin Institute of Technology, 2023, 55(5): 14. DOI: 10.11918/202201069

[7]肖琳, 陈博理, 黄鑫, 等. 基于标签语义注意力的多标签文本分类[J]. 软件学报, 2020, 31(4): 1079
XIAO Lin, CHEN Boli, HUANG Xin, et al. Multi-label text classification method based on label semantic information [J]. Journal of Software, 2020, 31(4): 1079. DOI: 10.13328/j.cnki.jos.005923

[8]张宇, 张鹏远, 颜永红. 基于注意力 LSTM 和多任务学习的远场语音识别[J]. 清华大学学报(自然科学版), 2018, 58(3): 249
ZHANG Yu, ZHANG Pengyuan, YAN Yonghong. Long short-term memory with attention and multitask learning for distant speech recognition [J]. Journal of Tsinghua University (Science and Technology), 2018, 58(3): 249. DOI: 10.16511/j.cnki.qhdxxb.2018.25.016

[9]CHEN Kan, WANG Jiang, CHEN L C, et al. ABC-CNN: an attention based convolutional neural network for visual question answering[J]. ArXiv e-Prints, 2015: arXiv: 1511.05960. DOI: 10.48550/arXiv.1511.05960

[10]ANDERSON P, HE Xiaodong, 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. Salt Lake City: IEEE, 2018: 6077. DOI: 10.1109/CVPR.2018.00636

[11] YANG Zichao, HE Xiaodong, GAO Jianfeng, et al. Stacked attention networks for image question answering[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 21. DOI: 10.1109/CVPR.2016.10

[12] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[C]//32nd Conference on Neural Information Processing Systems (NIPS 2018). Montréal: arXiv, 2018: 1. DOI: 10.48550/arXiv.1805.07932

[13] YU Zhou, YU Jun, CUI Yuhao, et al. Deep modular co-attention networks for visual question answering[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 6274. DOI: 10.1109/cvpr.2019.00644

[14] 陈巧红, 漏杨波, 孙麒, 等. 基于多模态门控自注意力机制的视觉问答模型[J]. 浙江理工大学学报(自然科学版), 2022, 47(3): 413
CHEN Qiaohong, LOU Yangbo, SUN Qi, et al. Visual question answering model based on multimodal gate self-attention mechanism [J]. Journal of Zhejiang Sci-Tech University (Natural Sciences Edition), 2022, 47(3): 413. DOI: 10.3969/j.issn.1673-3851(n).2022.03.017

[15] ZHANG Haiyang, LI Ruoyu, LIU Liang. Multi-head attention fusion network for visual question answering[C]//2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei: IEEE, 2022: 1. DOI: 10.1109/ICME52920.2022.9859639

[16] ZHANG Haotian, WU Wei. Transformer gate attention model: an improved attention model for visual question answering[C]//2022 International Joint Conference on Neural Networks (IJCNN). Padua: IEEE, 2022: 1. DOI: 10.1109/IJCNN55064.2022.9891887

[17] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 4171. DOI: 10.18653/V1/N19-1423

[18] LU Jiasen, BATRA D, PARIKH D, et al. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [EB/OL]. 2019: 1908.02265. https://arxiv.org/abs/1908.02265v1

[19] TAN Hao, BANSAL M. LXMERT: learning cross-modality encoder representations from transformers[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: ACL, 2019: 5099. DOI: 10.18653/v1/d19-1514

[20] HUANG Lun, WANG Wenmin, CHEN Jie, et al. Attention on attention for image captioning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 4633. DOI: 10.1109/ICCV.2019.00473

[21] RAHMAN T, CHOU S H, SIGAL L, et al. An improved attention for visual question answering[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville: IEEE, 2021: 1653. DOI: 10.1109/CVPRW53098.2021.00181

[22] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the V in VQA matter: elevating the role of image understanding in visual question answering[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6325. DOI: 10.1109/CVPR.2017.670

[23] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2014: 1532. DOI: 10.3115/v1/d14-1162

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735. DOI: 10.1162/neco.1997.9.8.1735

[25] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770. DOI: 10.1109/CVPR.2016.90

[26] REN Shaoqing, HE Kaiming, 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. DOI: 10.1109/TPAMI.2016.2577031

[27] KRISHNA R, ZHU Yuke, GROTH O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32. DOI: 10.1007/s11263-016-0981-7

[28] 陈巧红, 裴皓磊, 孙麒. 基于视觉关系推理与上下文门控机制的图像描述[J]. 浙江大学学报(工学版), 2022, 56(3): 542
CHEN Qiaohong, PEI Haolei, SUN Qi. Image caption based on relational reasoning and context gate mechanism[J]. Journal of Zhejiang University (Engineering Science), 2022, 56(3): 542. DOI:10.3785/j.issn.1008-973X.2022.03.013

[29] HUDSON D A, MANNING C D. GQA: a new dataset for real-world visual reasoning and compositional question answering[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 6693. DOI: 10.1109/CVPR.2019.00686

[30] 兰红, 张蒲芬. 问题引导的空间关系图推理视觉问答模型[J]. 中国图象图形学报, 2022, 27(7): 2274
LAN Hong, ZHANG Pufen. Question-guided spatial relation graph reasoning model for visual question answering[J]. Journal of Image and Graphics, 2022, 27(7): 2274. DOI: 10.11834/jig.200611

[31] 杨旭华, 庞宇超, 叶蕾. 利用可交谈多头共注意力机制的视觉问答[J]. 小型微型计算机系统, 2024, 45(8): 1901
YANG Xuhua, PANG Yuchao, YE Lei. Talking co-attention networks for visual question answering[J]. Journal of Chinese Computer Systems, 2024, 45(8): 1901. DOI: 10.20009/j.cnki.21-1106/TP.2023-0061

[32] 何世阳, 王朝晖, 龚声蓉, 等. 基于跨模态信息过滤的视觉问答网络[J]. 计算机科学, 2024, 51(5): 85
HE Shiyang, WANG Zhaohui, GONG Shengrong, et al. Cross-modal information filtering-based networks for visual question answering[J]. Computer Science, 2024, 51(5): 85. DOI: 10.11896/jsjkx.230300202

(编辑 张 红)