浙江理工大学学报,第 49 卷,第 6 期,2023 年 11 月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2023.06.012

基于空间关系聚合与全局特征注入的视觉问答模型

陈巧红,漏杨波,方　贤
(浙江理工大学计算机科学与技术学院,杭州 310018)

摘　要:现有视觉问答模型缺乏视觉对象间关系的理解能力,导致复杂问题的答案预测准确率较差;针对该问题,提出了一种基于空间关系聚合与全局特征注入的视觉问答模型。该模型首先利用空间关系聚合视觉区域特征,将其转换为视觉全局特征,并将这些特征注入网络;然后引入双边门控机制进行特征融合,使模型能够根据不同的问题输入,自适应地调整视觉全局特征和视觉区域特征对答案预测的贡献度;最后将融合特征输入分类网络,得到预测结果。在 VQA 2.0 和 GQA 公开数据集上进行实验,结果表明:该模型在 VQA2.0 的测试—开发集、测试—标准集和 GQA 的数据集上的总准确率分别达到 71.12%、71.54%和 57.71%,优于 MCAN 和 SCAVQAN 等主流模型。该模型由于引入了具有空间关系的视觉全局特征,能够更好地提升视觉对象间关系的理解能力,有效提高了视觉问答模型的准确率。

关键词:视觉问答;空间关系聚合;全局特征注入;视觉区域特征;视觉全局特征;双边门控机制
中图分类号:TP181　　**文献标志码**:A　　**文章编号**:1673-3851 (2023) 11-0764-11
引文格式:陈巧红,漏杨波,方贤.基于空间关系聚合与全局特征注入的视觉问答模型[J]. 浙江理工大学学报(自然科学),2023,49(6):764-774.
Reference Format: CHEN Qiaohong, LOU Yangbo, FANG Xian. A visual question answering model based on spatial relationship aggregation and global feature injection[J]. Journal of Zhejiang Sci-Tech University, 2023, 49(6): 764-774.

A visual question answering model based on spatial relationship aggregation and global feature injection

CHEN Qiaohong, LOU Yangbo, FANG Xian
(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: A visual question answering model based on spatial relationship aggregation and global feature injection was proposed aiming at the problem that the existing visual question answering models lack understanding of the relationship between visual objects and have low forecast accuracy. First, spatial relations were used for the model to aggregate visual regional features, which were subsequently transformed into visual global features, and injected into the network; then, by introducing a bilateral gating mechanism for feature fusion, the model could control the contribution of visual global features and visual regional features to answer prediction in an adaptive manner according to different question inputs; finally, the fusion features were input into the classification network to obtain the prediction results. Experiments were conducted on VQA 2.0 and GQA public datasets, and the results showed that the model achieved overall accuracy of 71.12%, 71.54%, and 57.71% on VQA 2.0 test subsets Test-dev, Test-std, and GQA, superior to mainstream models MCAN and SCAVQAN. The model introduces visual global features with spatial relationships, which can better enhance the understanding ability of relationships between visual objects and effectively improve the accuracy of the visual question answering model.

收稿日期:2023-06-15　　网络出版日期:2023-09-08
基金项目:浙江省自然科学基金项目(LQ23F020021);浙江理工大学科研启动项目(22232262-Y)
作者简介:陈巧红(1978—　),女,浙江临海人,教授,博士,主要从事计算机辅助设计及机器学习方面的研究。
通信作者:方　贤,E-mail:xianfang@zstu.edu.cn

Key words: visual question answering; spatial relationship aggregation; global feature injection; visual regional feature; visual global feature; bilateral gating mechanism

0 引　言

近年来,深度学习的快速发展极大地推动了计算机视觉和自然语言处理领域的进步。视觉—语言理解研究,例如视觉问答[1-2]、图像字幕[3]、多模态情感分析[4]等,引起了研究人员的广泛关注。视觉问答任务一般要求所建模型根据给定图像和一个与图像内容相关的自然语言问题,给出准确的自然语言答案。这项任务在回答盲人的询问[5]、辅助医生进行临床分析和诊断[6]等场景具有广阔的应用前景。

目前,为了给出复杂问题的准确答案,视觉问答任务需要对模态信息进行有效地特征融合。在特征融合的相关研究中,最初的方法是采用逐元素求和或乘积来生成融合特征。Fukui 等[7]认为这些方法不能充分捕捉两种模态之间的复杂关系,因此提出了多模态紧凑双线性池化(Multimodal compact bilinear pooling, MCB)模型。该模型通过问题特征和视觉特征的向量外积进行特征融合,以捕获模态间的复杂关系;然而,随着输入特征维度的增加,该模型的参数量呈指数级增长,较大的参数量导致模型效率低下。为了解决这一问题,Ben-Younes 等[8]提出了一种基于块—超对角(Block-superdiagonal)张量分解的特征融合框架,平衡融合模型的表现力和复杂性,从而在减少模型参数量的同时提高了模型效率。不同于上述的浅层融合方法,Lao 等[9]提出了一种深层融合方法,即多级混合嵌入融合(Multi-stage hybrid embedding fusion, MHEF)方法,将二重嵌入融合(Dual embedding fusion, DEF)和潜在嵌入融合(Latent embedding fusion, LEF)相结合,并为这种融合方法设计了多阶段融合结构,以获取多样化的融合特征。然而,在特征融合过程中普遍存在一个挑战,即存在较多噪声信息。这些噪声信息来源于视觉和语言模态之间的固有差异,以及特征提取过程中的不确定性等多种因素,对视觉问答系统的性能产生不利影响。

为了消除特征融合过程中存在的噪声信息,Chen 等[10]在视觉问答任务中引入了注意力机制,在输入问题中学习图像区域的视觉注意力,以筛选出回答问题所需的关键视觉区域。此后,注意力机制被广泛应用于在多模态输入中提取有效信息。Yu 等[11]提出了深度模块化协同注意力网络(Modular co-attention network, MCAN),通过引入自注意力单元和引导注意力单元,以“编码器—解码器”(Encoder-decoder)结构来构建网络,以提升对视觉和问题的细粒度理解;但这种方法没有利用视觉特来构建问题特征注意力,忽略了问题特征注意力的重要性。为了克服这一缺陷,鲜荣等[12]提出了一种多模态双导向注意力网络,通过引入视觉特征来构建问题特征注意力,从而进一步加强模态间的交互。然而,上述模型[11-12]虽然能够利用多个注意力层捕获更深层次的视觉语言相关性,但参数量较大,易导致模型过拟合。受胶囊神经网络的启发,Zhou 等[13]提出了动态胶囊注意力(Dynamic capsule attention),采用动态单层胶囊注意力网络代替静态多层注意力网络。该研究将特征矩阵中的向量视为底层胶囊,通过这些底层胶囊的动态交互获得的上层胶囊,并将该上层胶囊作为注意力的输出;同时在一个注意力层上进行多步骤的注意力学习,从而大大降低了模型的参数量,有效避免了模型过拟合等问题的发生。

上述对视觉问答任务的相关研究[11-13]均采用了 Anderson 等[14]提出的自底向上-自顶向下(Bottom-up and top-down, BUTD[14])的注意力机制;受益于自底向上策略,采用这种方式可以使用预先训练的对象检测器来提取仅依赖视觉输入本身的显著区域特征。这些特征结合注意力机制,能够有效地捕捉视觉区域和单词之间的相关性,从而提升了视觉问答模型的性能。然而,这些视觉区域特征在获取过程中处于相互独立的状态,使得模型缺乏视觉对象间关系的理解能力,从而导致模型对复杂问题的答案预测准确率较差。

本文针对现有视觉问答模型缺乏视觉对象间关系的理解能力的不足,提出了一种基于空间关系聚合与全局特征注入的视觉问答模型。该模型首先利用空间关系聚合视觉区域特征,将其转换为视觉全局特征,以加强视觉区域特征中对象间的关系;其次,将视觉全局特征输入注意力模块进行学习,以提升模型对视觉对象间关系的理解能力;再次,通过采用双边门控机制进行特征融合,使模型能够根据不同的问题输入,自适应地调整视觉全局特征和视觉区域特征对答案预测的贡献度;最后,将融合特征输

入多层感知器和 Softmax 层,以获得答案预测。该模型提升了对视觉对象间关系的理解能力,其答案预测的准确度有望得到提升。

1 模型设计

1.1 整体结构

本文提出的基于空间关系聚合与全局特征注入的视觉问答模型的整体结构如图 1 所示。该模型首先使用视觉特征提取器从输入图像中获取视觉特征;其次利用词嵌入的方法从输入问题中获取问题特征;再次利用空间关系聚合视觉区域特征来形成视觉全局特征,并输入至注意力模块中进行注意力学习;从次通过层间聚合的方式,获取多层次的视觉全局特征,并过滤噪声信息;最后使用双边门控机制融合视觉区域特征、视觉全局特征和问题特征,并将融合特征输入分类器中,输出预测答案。

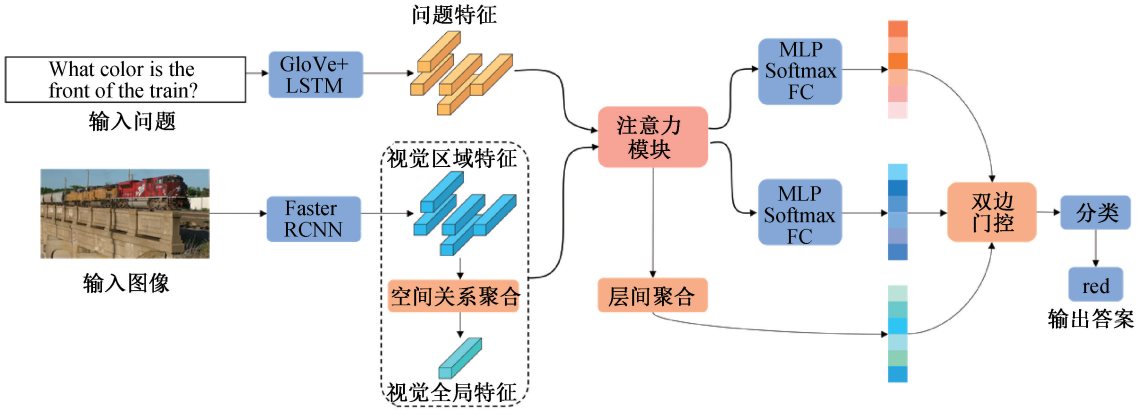


图 1 基于空间关系聚合与全局特征注入的视觉问答模型整体结构

1.2 视觉区域特征提取

对于模型中图像的输入,使用 Faster RCNN[15] 作为目标检测器,提取图像中显著区域。通过对每个对象区域进行非极大抑制,选取最相关的 M 个候选区域作为视觉区域特征。对于每张图像的输入,提取的视觉区域特征可以表示为 $\boldsymbol{V}=[v_1, v_2, \cdots, v_M]\in\mathbf{R}^{M\times t}$,每个视觉区域还拥有对应的边界框特征 $\boldsymbol{b}=[x, y, w, h]$,其中:$t$ 为视觉区域特征的维度,$\boldsymbol{v}_i\in\mathbf{R}^t$ 为图像第 i 个视觉区域特征,x、y 为边界框的中心坐标,w、h 分别为边界框的宽度和高度。

1.3 问题特征提取

对于模型的输入问题,首先根据空格和标点符号将问题分割为单词。然后,为了提高计算效率,将各问题中包含的单词数填充或截断至相同数目 N,并利用维度为 300 的 GloVe[16] 进行词嵌入。对于字典中没有的单词,选择随机向量进行初始化。最后,将这些单词向量输入 LSTM 网络,将 LSTM 网络隐藏层的输出作为问题特征。对于每个问题的输入,提取的问题特征可以表示为 $\boldsymbol{Q}=[\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_N]\in\mathbf{R}^{N\times l}$,其中:$l$ 为问题特征的维度,即 LSTM 隐藏层的维度;$\boldsymbol{q}_i\in\mathbf{R}^l$ 为问题第 i 个单词的问题特征。

1.4 视觉全局特征提取

视觉区域特征由 Faster RCNN 提取,每个视觉区域特征获取过程相互隔离,使得模型缺乏视觉对象间关系的理解能力,因此模型需要注入包含对象间关系的视觉全局特征。本文通过以下两个步骤来获取视觉全局特征:a)区域相关性学习;b)空间关系聚合。

1.4.1 区域相关性学习

在计算视觉区域间的相关性后,仅对高相关性的区域利用空间关系进行聚合,可以降低计算的复杂度,提高模型的运行效率,有效降低噪声信息的注入。具体步骤如下:对于模型视觉区域特征的输入 $\boldsymbol{V}$,首先将每个区域特征 $\boldsymbol{v}_i$ 与问题特征最后一个向量 $\boldsymbol{q}_N$ 进行连接,然后通过变换矩阵 $\boldsymbol{W}_e\in\mathbf{R}^{(t+l)\times r}$ 转换为嵌入特征 $\boldsymbol{e}_i\in\mathbf{R}^r$,最后将所有嵌入特征合并成嵌入矩阵 $\boldsymbol{E}\in\mathbf{R}^{M\times r}$。嵌入特征 $\boldsymbol{e}_i$ 的计算过程可用式(1)表示:

$$\boldsymbol{e}_i=[\boldsymbol{v}_i,\boldsymbol{q}_N]\boldsymbol{W}_e,\quad i=1,2,\cdots,M \tag{1}$$

最后通过式(2)计算得到相关性矩阵 A:

$$\boldsymbol{A}=\boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} \tag{2}$$

其中:$\boldsymbol{A}_{ij}=\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{e}_j$ 表示在特定问题下视觉区域 i 与视觉区域 j 之间的相关性。

1.4.2 空间关系聚合

使用简单池化 $\boldsymbol{g}=\frac{1}{M}\sum_{i=0}^{M}\boldsymbol{v}_i$ 对视觉区域特征进行聚合,所形成的视觉全局特征缺少对象间空间关系的信息。因此本文使用空间关系来聚合视觉区域特征,具体过程如图 2 所示。

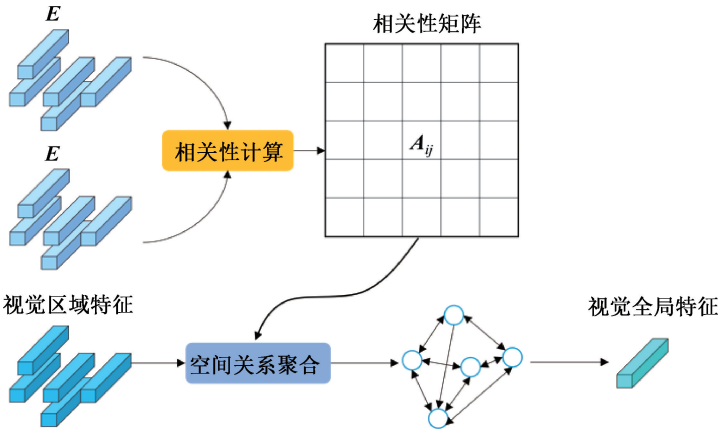


图 2　视觉全局特征获取流程

为了避免噪声信息的注入,对于每个视觉区域特征 $\boldsymbol{v}_i$,根据相关性矩阵 $\boldsymbol{A}$ 获取相关性最高的 m 个视觉区域特征。以视觉区域特征 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$ 为例,利用边界框特征计算空间关系特征 $\boldsymbol{s}_{ij}$,$\boldsymbol{s}_{ij}$ 可用式(3)表示:

$$\boldsymbol{s}_{ij}=\left[\frac{x_i-x_j}{w_i},\frac{y_i-y_j}{h_i},\frac{w_i}{w_j},\frac{h_i}{h_j}\right]\in\mathbf{R}^4 \tag{3}$$

获得的空间特征 $\boldsymbol{s}_{ij}$ 将用于聚合视觉区域特征 $\boldsymbol{v}_i$ 和 $\boldsymbol{v}_j$,得到 $\boldsymbol{p}_{ij}$,为每个隔离处理的视觉区域特征添加对象间关系信息。具体聚合过程可用式(4)表示:

$$\boldsymbol{p}_{ij}=f_{\mathrm{MLP}}([\boldsymbol{v}_i,\boldsymbol{v}_j,\boldsymbol{s}_{ij}\boldsymbol{W}_s]) \tag{4}$$

其中:$\boldsymbol{W}_s\in\mathbf{R}^{4\times t}$ 为变换矩阵,$\boldsymbol{p}_{ij}\in\mathbf{R}^t$,$f_{\mathrm{MLP}}$ 为多层感知器。依次将 $\boldsymbol{v}_i$ 与相关性最高的 m 个视觉区域特征进行聚合,将相关性作为权重,对这些特征进行加权求和,得到视觉区域特征 $\boldsymbol{v}_i$ 的聚合特征 $\boldsymbol{p}_i\in\mathbf{R}^t$。然后合并所有视觉区域特征的聚合特征,获得聚合矩阵 $\boldsymbol{P}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_M]\in\mathbf{R}^{M\times t}$。为了便于将视觉全局特征注入注意力网络,并减少后续计算的参数量,通过前馈网络将聚合矩阵 $\boldsymbol{P}$ 转换为一维向量,转换后获得视觉全局特征 $\boldsymbol{g}\in\mathbf{R}^t$。通过这种方式生成的视觉全局特征 $\boldsymbol{g}$,包含了对象间关系信息,与视觉区域特征相结合后,生成视觉特征 $\boldsymbol{C}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_M,\boldsymbol{g}]\in\mathbf{R}^{(M+1)\times t}$,并将其输入模型的注意力网络中。

1.5　注意力模块

在视觉问答模型中,由于答案的预测仅需要视觉和问题输入中的关键信息,因此在特征融合之前需要对问题特征和视觉特征进行注意力学习来减少噪声信息的干扰。Vaswani 等[17]提出了缩放点积注意力机制,作为 Transformer 的核心组件,使用点积来衡量查询和键之间的相关性。缩放点积注意力机制可用式(5)表示:

$$f_{\mathrm{Att}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

其中:f_{Att} 表示缩放点积注意力机制,$\boldsymbol{Q}$ 为查询,$\boldsymbol{K}$ 为键,$\boldsymbol{V}$ 为值,d_k 为查询和键的共同维度。缩放点积注意力首先计算查询和键之间的相似度,为了防止其结果过大,统一除以缩放因子 $\sqrt{d_k}$。随后通过 Softmax 函数计算分配给每个值的权重。为了拓展注意力机制探索子空间的能力,在缩放点积注意力机制的基础上,提出了多头注意力机制[17],通过同时使用 h 个并行的缩放点积注意,再将 h 个并行的缩放点积注意力输出进行拼接,生成最终的结果。该模块可用式(6)—(7)表示:

$$f_{\mathrm{Mul}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\mathrm{Concat}}(\boldsymbol{o}_1,\cdots,\boldsymbol{o}_h)\boldsymbol{W}_{\mathrm{o}} \tag{6}$$

$$\boldsymbol{o}_i=f_{\mathrm{Att}}(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\mathrm{K}},\boldsymbol{V}\boldsymbol{W}_i^{\mathrm{V}}) \tag{7}$$

其中:$\boldsymbol{W}_{\mathrm{o}}$ 为变换矩阵,$\boldsymbol{W}_i^{\mathrm{Q}}$、$\boldsymbol{W}_i^{\mathrm{K}}$ 和 $\boldsymbol{W}_i^{\mathrm{V}}$ 为头部映射矩阵,f_{Concat} 表示特征拼接,f_{Mul} 表示多头注意力机制。

本文模型遵循深度模块化共同注意网络[11]的设计思路,对于问题特征的输入 $\boldsymbol{Q}$,通过单层自意力模块的学习输出 $\boldsymbol{Q}_1\in\mathbf{R}^{N\times d}$,其过程可用式(8)—(9)表示:

$$\boldsymbol{Q}'=f_{\mathrm{LayerNorm}}(f_{\mathrm{Mul}}(\boldsymbol{Q},\boldsymbol{Q},\boldsymbol{Q})+\boldsymbol{Q}) \tag{8}$$

$$\boldsymbol{Q}_1=f_{\mathrm{LayerNorm}}(f_{\mathrm{FFN}}(\boldsymbol{Q}')+\boldsymbol{Q}') \tag{9}$$

其中:$f_{\mathrm{LayerNorm}}$ 为层归一化,f_{FNN} 为多层感知机。由于该自注意力模块不会改变 $\boldsymbol{Q}$ 的维度,因此可以将该模块堆叠 L 次来捕获单词间更深层次的相关性。经过 L 层自注意力模块的学习,能够捕获单词之间的语义特征和长距离依赖特征,增加重要单词的权重。例如当提出问题为“Is the girl sitting on the horse?”时,模型将会重点关注“girl”、“sitting”和“horse”,从而推断出更准确的答案。

对于视觉特征 $\boldsymbol{C}$,注意力模块在式(8)—(9)的

基础上,添加跨模态注意力学习,通过单层注意力模块的学习输出 $\boldsymbol{C}_1 \in \mathbf{R}^{(M+1)\times d}$,其过程可以用式(10)—(12)表示:

$$\boldsymbol{C}' = f_{\text{LayerNorm}}(f_{\text{Mul}}(\boldsymbol{C},\boldsymbol{C},\boldsymbol{C}) + \boldsymbol{C}) \tag{10}$$

$$\boldsymbol{C}'' = f_{\text{LayerNorm}}(f_{\text{Mul}}(\boldsymbol{C}',\boldsymbol{Q}_L,\boldsymbol{Q}_L) + \boldsymbol{C}') \tag{11}$$

$$\boldsymbol{C}_1 = f_{\text{LayerNorm}}(f_{\text{FNN}}(\boldsymbol{C}'') + \boldsymbol{C}'') \tag{12}$$

其中:$\boldsymbol{Q}_L$ 为问题特征经过 L 层自注意力学习后的输出。不同于上述自注意力模块,通过额外引入问题特征对视觉特征的跨模态注意力学习,使得模型能够聚焦与问题最相关的视觉内容上,关注与问题最相关的视觉特征。由于该注意力模块不会改变 $\boldsymbol{C}$ 的维度,因此同样可以将该模块堆叠 L 次形成深度注意力网络。

经过 L 层深度模块化共同注意网络的学习,问题特征的输出为 $\boldsymbol{Q}_L = [\boldsymbol{q}_1^L, \boldsymbol{q}_2^L, \cdots, \boldsymbol{q}_N^L]$,视觉特征的输出为 $\boldsymbol{C}_L = [\boldsymbol{v}_1^L, \boldsymbol{v}_2^L, \cdots, \boldsymbol{v}_M^L, \boldsymbol{g}^L]$,从中截取出视觉区域特征 $\boldsymbol{V}_L = [\boldsymbol{v}_1^L, \boldsymbol{v}_2^L, \cdots, \boldsymbol{v}_M^L]$,然后分别计算视觉和问题中各特征的权重 $\boldsymbol{A}_{\text{Visual}}$ 和 $\boldsymbol{A}_{\text{Question}}$,计算过程可用式(13)—(14)表示:

$$\boldsymbol{A}_{\text{Visual}} = \text{Softmax}(f_{\text{MLP}}(\boldsymbol{V}_L)) \tag{13}$$

$$\boldsymbol{A}_{\text{Question}} = \text{Softmax}(f_{\text{MLP}}(\boldsymbol{Q}_L)) \tag{14}$$

其中:$A_{\text{Visual}} \in \mathbf{R}^M$, $A_{\text{Question}} \in \mathbf{R}^N$。视觉区域的注意力特征 $\widetilde{\boldsymbol{v}} \in \mathbf{R}^d$ 和问题的注意力特征 $\widetilde{\boldsymbol{q}} \in \mathbf{R}^d$ 表示为对所有特征根据权重进行加权求和,其计算过程可用式(15)—(16)表示:

$$\widetilde{\boldsymbol{v}} = \boldsymbol{A}_{\text{Visual}}\boldsymbol{V}_L \tag{15}$$

$$\widetilde{\boldsymbol{q}} = \boldsymbol{A}_{\text{Question}}\boldsymbol{Q}_L \tag{16}$$

1.6 层间聚合

Wang 等[18]认为在深层网络中仅利用较高层网络信息而忽略底层网络信息,会丢失特征信息。为了获取更加全面的信息,本文通过聚合注意力模块各层中的视觉全局特征,以融合所有低级和高级信息。最简单的方式就是使用 LSTM 网络进行层间聚合,通过遗忘门、输入门和输出门在深层特征中保留浅层特征中重要的信息,加强各层特征间的联系,从而生成最终的视觉全局表示。首先将注意力模块各层视觉全局特征 $\boldsymbol{G} = [\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_L] \in \mathbf{R}^{L\times d}$ 输入 LSTM 网络中,具体流程可用式(17)表示:

$$\boldsymbol{h}_i = f_{\text{LSTM}}(\boldsymbol{g}_i, \boldsymbol{h}_{i-1}) \ (i = 1,2,\cdots,L) \tag{17}$$

其中:f_{LSTM} 为长短期记忆网络,$\boldsymbol{h}_{i-1} \in \mathbf{R}^d$ 为上一个时序单元隐藏层输出,$\boldsymbol{h}_i \in \mathbf{R}^d$ 为当前时序单元隐藏层输出。然后根据不同问题,为每个 LSTM 隐藏层输出分配不同的权重,其中第 i 个隐藏层输出特征 $\boldsymbol{h}_i$ 的权重 a_i^h 计算过程可用式(18)—(19)表示:

$$w_i = \widetilde{\boldsymbol{q}}^{\mathrm{T}}\boldsymbol{h}_i \tag{18}$$

$$a_i^h = \frac{\exp w_i}{\sum_{k=1}^{L}\exp w_k} \tag{19}$$

其中:w_i 为问题注意力特征和第 i 个隐藏层输出特征 $\boldsymbol{h}_i$ 的相似度结果。最后通过对所有 LSTM 隐藏层输出特征进行加权求和获得最终的视觉全局特征 $\widetilde{\boldsymbol{g}} \in \mathbf{R}^d$,具体过程可用式(20)所示:

$$\widetilde{\boldsymbol{g}} = \sum_{i=1}^{L} a_i^h \boldsymbol{h}_i \tag{20}$$

1.7 双边门控机制

对于上述注意力模块的输出,引入双边门控机制用于特征融合,使模型能够根据不同的问题输入,以自适应的形式控制视觉全局特征和视觉区域特征对答案预测的贡献度。其中用于控制贡献度的门控值 G 计算过程可用式(21)表示:

$$G = \text{Sigmoid}(\boldsymbol{W}_2^G(\boldsymbol{W}_1^G([\widetilde{\boldsymbol{q}}, \widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{g}}]))) \tag{21}$$

其中:$\boldsymbol{W}_1^G \in \mathbf{R}^{(3\times d)\times d}$ 和 $\boldsymbol{W}_2^G \in \mathbf{R}^d$ 为变换矩阵。在特征融合过程中,通过使用门控值 G 控制视觉区域特征的权重,使用门控值$(1-G)$控制视觉全局特征的权重。特征融合过程可用式(22)表示:

$$\boldsymbol{f} = f_{\text{LayerNorm}}(\boldsymbol{W}_{\widetilde{\boldsymbol{v}}}^f G\widetilde{\boldsymbol{v}} + \boldsymbol{W}_{\widetilde{\boldsymbol{q}}}^f\widetilde{\boldsymbol{q}} + \boldsymbol{W}_{\widetilde{\boldsymbol{g}}}^f(1-G)\widetilde{\boldsymbol{g}}) \tag{22}$$

其中:$W_{\widetilde{\boldsymbol{v}}}^f$、$W_{\widetilde{\boldsymbol{q}}}^f$、$W_{\widetilde{\boldsymbol{g}}}^f$ 为变换矩阵,$\boldsymbol{f} \in \mathbf{R}^u$ 为融合特征。

1.8 损失函数

与以往工作相同,本文将视觉问答视为多标签分类任务,其中候选答案数量为 N_{ans}。将融合特征 $\boldsymbol{f}$ 送入由线性层所组成的分类器中,使用 Sigmoid 函数将结果控制在 0~1 之间,作为模型预测每个候选答案的概率。具体分类过程可用式(23)表示:

$$\boldsymbol{A}' = \text{Sigmoid}(f_{\text{MLP}}(\boldsymbol{f})) \tag{23}$$

其中:$A' \in \mathbf{R}^{N_{\text{ans}}}$ 为模型预测候选答案的概率。训练过程中,采用二元交叉熵作为损失函数,损失值 L 计算过可用式(24)表示:

$$L = -\sum_{i}^{N_{\text{ans}}} a_i \log(a_i') - (1-a_i)\log(1-a_i') \tag{24}$$

其中:a_i 为第 i 个候选答案的标签值,a_i' 为第 i 个候选答案的预测值。

2　实验与结果分析

2.1　数据集

在 2 个公共基准数据集 VQA 2.0[19]、GQA[20]上验证本文模型中各个模块的性能,并与其他主流模型进行比较。

VQA 2.0 数据集由 VQA 1.0[1]数据集更新而来,包含 2.04×10^5 张图像和 1.10×10^6 个问题,是目前视觉问答领域最常用的大型公共数据集。与 VQA 1.0 数据集不同,VQA 2.0 数据集包含更大的问题样本,解决了 VQA 1.0 数据集中答案分布不平衡的问题,并使数据集在语言偏见方面更平滑。数据集被分为三个子集:训练集、验证集和测试集。VQA 2.0 数据集中样例如图 3 所示,问题分为三种类型:是/否(Yes/No)、计数(Num)和其他(Other),并且每个问题包含 10 个相应的答案。

(a) 是/否问题　　(b) 计数问题　　(c) 其他问题

图 3　VQA 2.0 数据集样例

GQA 是一个由真实世界的图像与合成问题所组成的数据集,其问题相较于其他视觉问答数据集更具复杂性和多样性。这些问题需要推理、常识推断以及对图像场景的深入理解,要求模型不仅能够理解问题的表面含义,还要具备更高层次的推理能力。除此之外,GQA 数据集通过一种平滑技术来减少问题的偏差,从而平衡二元问题和开放问题的答案分布。数据集包含大约 2.20×10^7 个问题和 1.13×10^5 张图像,其中训练集、验证集、测试集和挑战集各占 70%、10%、10%和 10%。

2.2　评价指标

对于 VQA 2.0 数据集,本文依据 Agrawal 等[1]的工作,当预测答案占 10 个人工标注答案中 3 个以上时,才会被认为是完全正确。评估公式可用式(25)表示:

$$\mathrm{acc}(a)=\min\left\{\frac{\mathrm{count}(a)}{3},1\right\} \tag{25}$$

其中:a 表示模型预测答案;$\mathrm{count}(a)$为预测答案在 10 个人工标注答案中所占的数量。

对于 GQA 数据集,除了总准确率、二元问题准确率以及开放问题准确率的标准精度指标之外,引入 4 个额外的指标来进一步评价模型,即一致性、有效性、合理性和分布性。一致性用于度量不同问题的回答一致性,对于新问题的答案不应该与之前的答案相矛盾。有效性用于检查给定的答案是否在问题回答范围内。合理性用于度量问题的答案是否合理或有意义。分布性用于度量预测答案分布和真实答案分布之间的总体匹配,判断模型是否不仅预测了最常见的答案,而且预测了不太常见的答案。

2.3　实验设置

本文模型由 PyTorch 框架进行构建,使用 Nvidia GeForce RTX 3090 显卡作为硬件平台进行模型训练。实验主要参数设置如表 1 所示:对于问题特征的提取,VQA 2.0 数据集上句子长度被设置为 $N=14$,GQA 数据集上句子长度被设置为 $N=29$,问题特征维度 $l=512$。对于视觉特征的提取,通过 Faster RCNN 提取图像中概率最高的 $M=100$ 个视觉特征,视觉特征维度 $t=2048$;注意力模块中视觉特征和问题特征将被转换为统一维度 $d=1024$,注意力模块堆叠层数为 $L=6$,多头注意力包含 $h=8$ 的缩放点积注意力,缩放点积注意力的维度 $c=128$。特征融合过程中融合特征维度 $u=1024$;对于 VQA 2.0 数据集,选取 $N_{\mathrm{ans}}=3129$ 个在训练集中最常见的答案作为多分类问题的预测向量,而 GQA 数据集则选取 $N_{\mathrm{ans}}=1843$。上述的实验参数设置与 MCAN 模型[11]相同,可以清晰比较模型的性能差异。

训练过程中,使用 Adam 优化器($\beta_1=0.9$,$\beta_2=0.98$)对模型训练 13 个周期,其中前 10 个周期的学习率为 0.0001,之后每个周期的学习率下降

1/10。为了防止模型过拟合,在每个全连接层之后采用值为0.5的Dropout。

表1 实验主要参数

模型参数	参数值
问题特征数 N	14(VQA 2.0),29(GQA)
视觉特征数 M	100
视觉特征维度 t	2048
问题特征维度 l	512
融合特征维度 u	1024
注意力模块特征维度 d	1024
注意力模块堆叠层数 L	6
多头注意力平行头数 h	8
多头注意力平行头维度 c	128
答案特征维度 N_{ans}	3129(VQA 2.0),1843(GQA)

2.4 消融实验

由于基于空间关系聚合与全局特征注入的视觉问答模型由多个模块组成,为了分析各个模块在模型中的作用,在VQA 2.0数据集上进行消融实验,通过在验证集上展示该模型不同变体下的结果来评估本文模型中不同模块的贡献。

在评估各个模块有效性之前,对区域相关性学习模块中所提出计算空间特征 $\boldsymbol{s}_{ij}$ 的不同参数 m 进行消融实验,实验结果如表2所示。实验发现,通过计算不同数量的空间特征来聚合区域特征,对性能的影响较大。其中 $m=4$ 时模型总准确率最高,过少的空间关系导致重要信息的缺失,过多的空间关系则导致模型中噪声信息的引入。因此,后续的消融实验将仅对每个视觉区域特征与其最高相关性的 m 个区域利用空间关系进行聚合。

表2 最相关区域数目的参数选择实验结果

最相关区域数目 m	不同类型问题的准确率/%			总准确率/%
	是/否问题	计数问题	其他问题	
1	84.90	49.52	58.70	67.34
2	85.07	49.43	58.74	67.41
4	85.21	49.49	58.75	67.48
8	85.04	49.35	58.68	67.36

基于空间关系聚合与全局特征注入的视觉问答模型消融实验的模型变体为:

a)基线模型:将MCAN[11]模型作为基线模型。

b)基线模型+池化聚合:通过在基线模型中引入简单池化的方式聚合视觉区域特征形成视觉全局特征,并注入注意力模块中进行注意力学习,后续特征融合采用线性多模态融合。

c)基线模型+空间关系聚合:利用空间关系聚合视觉区域特征形成视觉全局特征,注入注意力模块中进行注意力学习,后续特征融合采用线性多模态融合。

d)基线模型+空间关系聚合+层间聚合:在空间关系聚合的基础上,聚合注意力网络层间视觉全局特征。

e)基线模型+空间关系聚合+双边门控机制:在空间关系聚合的基础上采用双边门控机制替换线性多模态融合。

f)Full model:本文所提出的完整模型,在空间关系聚合的基础上,同时使用层间聚合和双边门控机制。

消融实验的结果如表3所示。

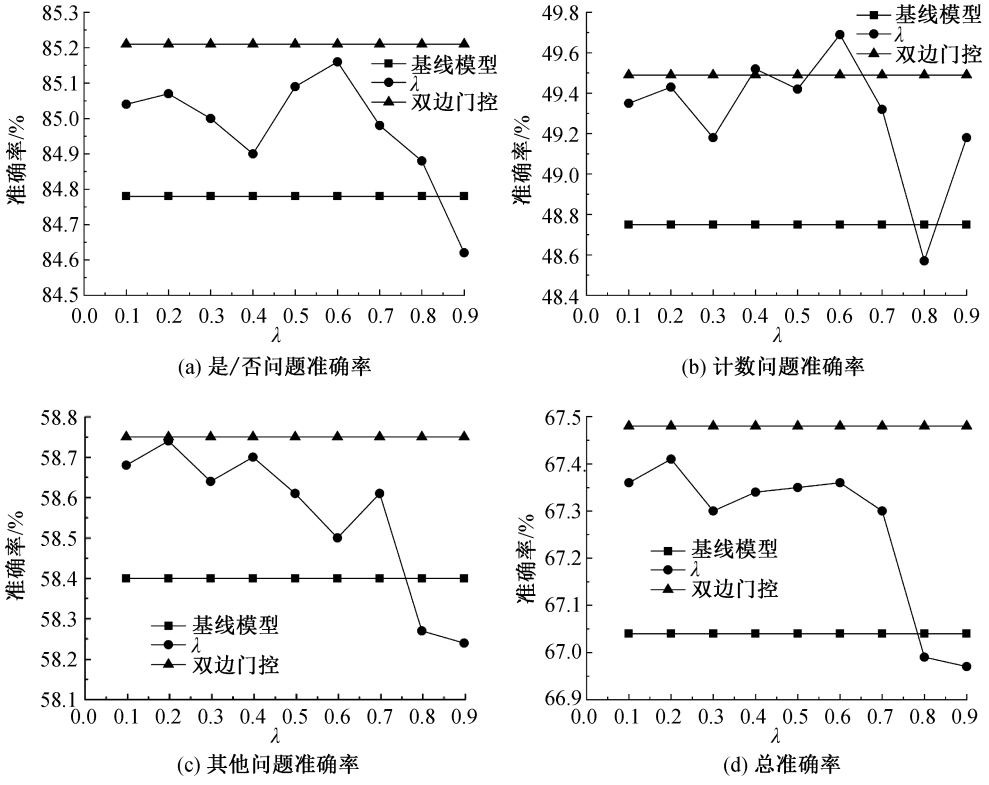
表3 基于空间关系聚合与全局特征注入的视觉问答模型消融实验结果

模型	总准确率/%
基线模型	67.04
基线模型+池化聚合	67.08
基线模型+空间关系聚合	67.34
基线模型+空间关系聚合+层间聚合	67.39
基线模型+空间关系聚合+双边门控机制	67.42
完整模型	67.48

实验结果显示:利用不同方式形成视觉全局特征中,简单池化聚合和利用空间关系聚合相较于基线模型均有所提升。由于简单池化方式仍无法解决的模型缺乏对象关系信息的缺陷,仅仅提升了0.02%。而空间关系聚合所生成的视觉全局特征,通过将区域与其相关性最高的其他区域利用空间关系进行聚合,从而加强了视觉对象间的关系,相较于基线模型,总体准确度有0.30%的提升。对于添加注意力网络层间聚合的模型,模型总体准确率提升了0.05%,这是因为通过时序模型融合所有低级和高级信息,加强了各层特征间的联系,使得最终输出的视觉全局特征更为全面。采用双边门控机制替换线性模态融合来进行特征融合,模型的总体准确率从67.34%提升至67.42%,由此可以看出,利用自适应的方式控制视觉区域特征和视觉全局特征的权重,模型可以根据具体的问题来决定是需要更多的区域信息还是全局信息,从而提升预测精度。最终比较完整模型和基线模型总准确率提升了0.44%。

本文模型通过双边门控机制以自适应的形式控制视觉全局特征和视觉区域特征的贡献。在VQA 2.0数据集上评估该模块的有效性,将包含双边门控的模型与固定权重系数($\lambda=0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$)的对比模型进行比较。

实验结果如图 4 所示,具有双边门控机制的模型在“是/否问题准确率”、“其他问题准确率”和“总准确率”上具有最佳效果,表明双边门控机制有效。而在回答计数问题时,固定权重系数为 $\lambda=0.4$ 和 $\lambda=0.6$ 时,对比模型的性能要优于本文模型,但其总准确率仍然不如本文模型。这是因为双边门控机制使模型能够根据问题,自适应地调节视觉全局特征和视觉区域特征对于答案预测的贡献度,从而提高模型预测精度。



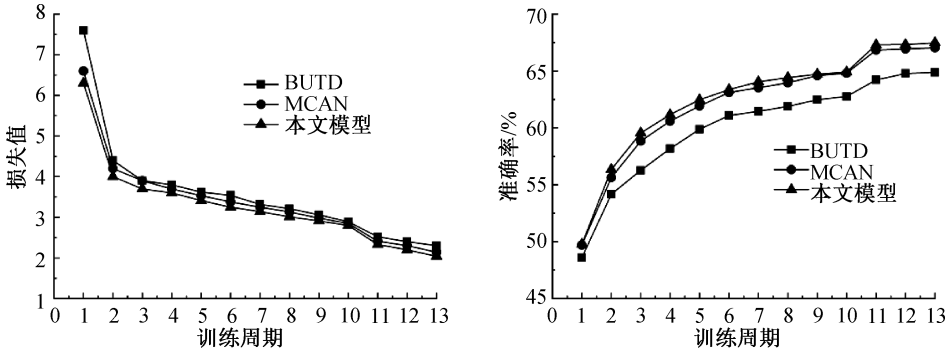
(a) 是/否问题准确率　(b) 计数问题准确率

(c) 其他问题准确率　(d) 总准确率

图 4　模型采用不同方式进行特征融合时回答各类问题的准确率曲线

2.5　收敛性分析

图 5 表示模型在训练过程中损失值和准确度随着训练周期增加的变化曲线,其中 BUTD[14] 和 MCAN[11] 为基线模型。由图 5(a)可以看出,模型在训练过程中损失值随着训练周期的增加稳定下降,在第 1 个训练周期处波动幅度最大,在第 11 个训练周期处损失逐渐平稳。相较于 BUTD 和 MCAN 模型,本文模型在训练过程中损失值收敛更快,需要的训练周期数更少。另一方面,从图 5(b)可以看出,在训练过程中模型准确率随着损失的减少而增加,在第 1 个训练周期处显著增加,随后平稳增加,并且在第 13 个训练周期处准确率到达最大值。根据损失值和准确率变化曲线可以看出,本文模型的拟合能力和表现能力均都要优于 BUTD 和 MCAN 模型。



(a) 训练过程中模型损失值变化曲线　(b) 训练过程中模型准确率变化曲线

图 5　训练过程中模型损失值和准确率变化曲线

2.6 可视化实验

本文通过分析视觉的区域贡献度以验明模型的有效性和可解释性。视觉区域贡献度实验结果如图 6 所示,其中:图 6(a)—(c)为本文模型针对特定样例的视觉区域贡献度,图 6(d)—(f)为基线模型对特定样例的视觉区域贡献度;区域贡献度的值显示在每个边界框的左上角,值越大该区域用于回答问题的贡献越大。对于问题“Is there a man dressed in blue?”和“How many lights on the front of the train are lit?”,两者虽然均能回答出正确的答案,但是通过贡献度可视化可以看出,本文模型对于回答该问题时所需要的图像关键区域的贡献度更高。而对于问题“What hand is the person holding out?”,本文模型能够正确回答出答案“left”,基线模型则回答错误。这是因为本文模型利用空间关系来聚合视觉区域特征所形成的视觉全局特征,并将该特征注入到模型中,从而提升了模型对视觉对象间关系的理解能力,能够理解用于回答问题的图像信息,从而提升了模型预测的准确度。

图 6 样例的视觉区域贡献度

2.7 模型总体性能

本文模型与 BUTD、MCAN 和 MESAN 等近年来的视觉问答主流模型,在 VQA 2.0 数据集的测试-开发集与测试-标准集上的性能对比实验,结果如表 4 所示。

表 4 模型在 VQA 2.0 数据集上的实验结果

模型	测试-开发集实验准确率/%				测试-标准集实验总准确率/%
	是/否问题	计数问题	其他问题	总准确率	
BUTD[14]	81.82	44.21	56.05	64.32	65.67
MCAN[11]	86.82	53.26	60.72	70.63	70.90
MESAN[21]	87.05	53.21	60.72	70.71	71.08
Re-Att[22]	87.00	53.06	60.19	70.43	70.72
MGSA[23]	86.96	53.21	60.84	70.76	71.12
SCAVQAN[24]	86.96	53.49	60.95	70.82	71.14
本文模型	87.73	54.07	61.12	71.12	71.54

由表 4 的实验结果可以得出:利用提出的空间关系聚合模块与全局特征注入能够提高答案预测的准确率。BUTD 模型[14]通过采用自底向上的注意力机制提取视觉区域特征。可以看出本文模型的性能要明显优于 BUTD 模型,这是因为本文模型在视觉区域特征的基础上构建视觉全局特征,通过自适应门控动态选择与输入问题相关的视觉特征,从而提升了模型对视觉对象间关系的理解能力。相比于基线模型 MCAN,本文模型在 VQA 2.0 数据集的测试-开发集上达到 71.12%的总准确率,在测试-标准集上达到了 71.54%的总准确率,均高于采用深层联合注意网络的 MCAN 模型。与 MCAN 不同,

MESAN[21]采用基于 top-k 的显式选择,仅关注指定数量的问题词,从而减少无关信息所造成的干扰。Re-Att[22]用基于问题重建初始注意力图的方法,使模型对问题的理解更加准确。MGSA[23]在自注意力过程中利用其他模态信息,动态调节模态内的注意力权重和流量,有效过滤了自注意力过程中的噪声信息。SCAVQAN[24]采用一种基于阈值的稀疏共同注意视觉网络,通过设置阈值来滤出图像和问题中对于回答问题最有用的信息,从而提升了模型的整体性能。上述视觉问答模型(MESAN、Re-Att、MGSA 和 SCAVQAN)均通过改进注意力机制来有效过滤特征中存在的噪声信息,然而,它们在建模时忽略了视觉对象间的关系。为弥补这一不足,本文模型不仅注重注意力学习,还引入了空间关系聚合模块,以加强对视觉对象间关系的建模。这使得模型可以更好地理解图像内容,从而在 VQA 2.0 数据集的各类答案预测指标上均具有一定的优势。

本文还在 GQA 数据集上进行实验,实验结果如表 5 所示。相较于 BUTD 模型,本文模型性能得到大幅度提升,总准确率达到 57.71%。对比作为基线模型的 MCAN,本文模型在保持相似的有效性的同时,其他所有指标上均表现出更好的效果。与近期主流的视觉问答模型 SCAVQAN 相比,除了分布性指标外,其他指标均存在一定竞争力,这表明本文模型具有优秀的性能。

表 5　模型在 GQA 数据集上的实验结果

模型	精度指标			非精度指标			
	二元问题准确率/%	开放问题准确率/%	总准确率/%	一致性/%	有效性/%	合理性/%	分布性
CNN+LSTM	63.26	31.80	46.55	74.57	96.02	84.25	7.46
BUTD[14]	66.64	34.83	49.74	78.71	96.18	84.57	5.98
MCAN[11]	75.56	40.38	56.84	87.19	96.85	85.32	1.31
SCAVQAN[24]	76.00	40.54	57.13	87.99	96.82	84.96	1.09
本文模型	76.43	41.21	57.71	88.27	96.97	85.51	1.12

3　结　论

本文提出了基于空间关系聚合与全局特征注入的视觉问答模型,通过空间关系聚合来生成视觉全局特征,有效增强了视觉对象之间的关联,提升对视觉对象间关系的理解能力,有效地提高了答案预测的准确率。该模型利用相似度矩阵来实现视觉区域特征的有效聚合,降低了聚合过程中的计算量,提高了模型收敛速度;将包含空间关系的视觉全局特征输入到注意力网络中,显著提升了模型对视觉对象间关系的理解能力;在此基础上,引入了双边门控机制,有助于模型筛选出用于回答问题的关键视觉信息。在 VQA 2.0 和 GQA 数据集上的实验结果表明:本文模型在各个指标上均优于其他主流模型;与去除空间关系聚合模块和双边门控机制的本文模型进行对比,本文设计的各个模块均起到了重要作用。后续研究可考虑视觉区域对象间显式和隐式的关系信息,进一步提升模型对视觉对象间关系的理解能力。

参考文献:

[1] Agrawal A, Lu J S, Antol S, et al. VQA: visual question answering [J]. International Journal of Computer Vision, 2017, 123(1): 4-31.

[2] 闫悦, 郭晓然, 王铁君, 等. 问答系统研究综述[J/OL]. 计算机系统应用. (2023-06-12)[2023-06-15]. https://doi.org/10.15888/j.cnki.csa.009208.

[3] 王源顺, 段迅, 吴云. 一种新的 seq2seq 的可控图像字幕的生成方法[J]. 计算机应用研究, 2021, 38(11): 3510-3516.

[4] 陈巧红, 孙佳锦, 孙麒, 等. 基于多层跨模态注意力融合的图文情感分析[J]. 浙江理工大学学报(自然科学版), 2022, 47(1): 85-94.

[5] Le T, Nguyen H T, Le Nguyen M. Multi visual and textual embedding on visual question answering for blind people[J]. Neurocomputing, 2021, 465: 451-464.

[6] Liu B, Zhan L M, Xu L, et al. Medical visual question answering via conditional reasoning and contrastive learning[J]. IEEE Transactions on Medical Imaging, 2023, 42(5): 1532-1545.

[7] Fukui A, Park D H, Yang D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 457-468.

[8] Ben-Younes H, Cadene R, Thome N, et al. BLOCK: bilinear superdiagonal fusion for visual question answering and visual relationship detection [J].

Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8102-8109.

[9] Lao M R, Guo Y M, Pu N, et al. Multi-stage hybrid embedding fusion network for visual question answering [J]. Neurocomputing, 2021,423: 541-550.

[10] Chen K, Wang J, Chen L C, et al. ABC-CNN: An attention based convolutional neural network for visual question answering[EB/OL]. (2016-04-03)[2023-06-15]. https://arxiv.org/abs/1511.05960.

[11] Yu Z, Yu J, Cui Y H, 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, CA, USA. IEEE, 2020: 6274-6283.

[12] 鲜荣, 何小海, 吴晓红, 等. 基于多模态双向导向注意的视觉问答[J]. 太赫兹科学与电子信息学报, 2021, 19(1): 156-161.

[13] Zhou Y Y, Ji R R, Su J S, et al. Dynamic capsule attention for visual question answering[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 9324-9331.

[14] Anderson P, He X D, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 6077-6086.

[15] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[16] 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). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1532-1543.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Advances in Neural Information Processing Systems. Long Beach: Curran Associates, 2017: 5998-6008.

[18] Wang Q, Li F X, Xiao T, et al. Multi-layer representation fusion for neural machine translation [EB/OL]. (2020-02-16)[2023-06-15]. https://arxiv.org/abs/2002.06714.

[19] 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, HI, USA. IEEE, 2017: 6325-6334.

[20] 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, CA, USA. IEEE, 2020: 6693-6702.

[21] Guo Z H, Han D Z. Multi-modal explicit sparse attention networks for visual question answering[J]. Sensors, 2020, 20(23): 6758.

[22] Guo W Y, Zhang Y, Yang J F, et al. Re-attention for visual question answering[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2021, 30: 6730-6743.

[23] 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):413-423.

[24] Guo Z H, Han D Z. Sparse co-attention visual question answering networks based on thresholds[J]. Applied Intelligence, 2023, 53(1): 586-600.

(责任编辑:康 锋)