浙江理工大学学报，2022，47(3)：413-423
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2022.03.017

基于多模态门控自注意力机制的视觉问答模型

陈巧红，漏杨波，孙　麒，贾宇波
（浙江理工大学信息学院，杭州　310018）

摘　要：针对现有视觉问答模型中自注意力机制过滤噪声信息能力较差的问题，提出了一种基于多模态门控自注意力(Multimodal gate self-attention，MGSA)机制的视觉问答模型。该模型在自注意力模块中利用其他模态特征作为通道调节门，以过滤目标模态特征自注意力学习的输出结果；同时结合跨模态双导向注意力机制与堆叠注意力模块，共同学习协同注意力和深层注意力；最后将包含丰富注意结果的视觉和语言特征进行特征融合，通过分类网络获得预测结果。在视觉问答公开数据集 VQA-v2 上进行实验，结果表明：该模型在 Test-dev 和 Test-std 两个测试子集的总准确率分别为 70.76%和 71.12%，优于当前主要模型；变体模型间的性能对比验证了模型中各模块的有效性。该模型具有较强的过滤噪声信息能力，有效提升了视觉问答模型的性能。

关键词：视觉问答；多模态；门控自注意力；双导向注意力；特征融合
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2022) 05-0413-11

Visual question answering model based on multimodal gate self-attention mechanism

CHEN Qiaohong，*LOU Yangbo*，*SUN Qi*，*JIA Yubo*
(School of Information Science and Technology，Zhejiang Sci-Tech University，Hangzhou 310018，China)

Abstract：To address the problem of poor noise information filtering ability in the self-attention mechanism of the existing visual question answering model，a visual question answering model based on multimodal gate self-attention (MGSA) mechanism was proposed in this paper. In the self-attention module，other modal features were used as the channel adjustment gate to filter the output results of self-attention learning with target modal features. Combined with the cross modal dual guided-attention mechanism and the stacked attention，it jointly learned co-attention and deep attention. Finally，through the feature fusion of visual features and language features containing rich attention results，the prediction results were obtained by means of classification network. The experiment was carried out on the visual question answering public data set VQA-v2. The results show that the total accuracy of this model in Test-dev and Test-std subsets is 70.76% and 71.12%，respectively，which is superior to the current main models. Through the performance comparison between variant models，the effectiveness of each module in the model has been verified，indicating that the proposed model has strong ability to filter noise information and it can effectively improve the performance of the visual question answering model.

Key words：visual question answering；multimodal；gate self-attention；dual guided-attention；feature fusion

---

收稿日期：2021－11－11　　网络出版日期：2022－03－18
基金项目：浙江理工大学中青年骨干人才培养经费项目
作者简介：陈巧红(1978－　)，女，浙江临海人，副教授，博士，主要从事计算机辅助设计及机器学习方面的研究。

0 引 言

近年来,计算机视觉和自然语言处理相结合的任务,如图像字幕[1-2]、视觉问答[3-5]、视觉蕴含[6]等,受到了广泛关注。其中,视觉问答作为一项完整的人工智能任务,旨在准确回答给定图像或视频的自然语言问题,目前主要应用于回答盲人的询问[7]、辅助医生进行临床分析和诊断[8]等。相较于其他多模态任务,视觉问答任务在预测答案过程中要求同时对语言特征和视觉特征进行更好的细粒度语义理解和视觉推理。如果视觉和语言特征之间的联系不紧密,那么模型的预测精度则会下降,因此如何更好地建立各模态之间的关系,是视觉问答任务最具挑战的难题之一。

注意力机制是建立模态间关系的常用方法,其原理是通过模仿人类的视觉工作,获取更多与关注目标相关的细节信息,从而提高模型预测精度。在有关视觉问答任务的注意力机制研究中,Chen 等[9]最早从输入问题中学习图像区域的视觉注意力,从而筛选图像关键区域,但该方法忽视了模型通过语言注意力提取关键词的重要性。Lu 等[10]提出了协同注意力(Co-attention)的概念,与之前只关注视觉注意力的研究不同,该研究不仅将问题用于引导图像注意力,而且将图像用于引导问题注意力,然而协同注意力虽然能够有效选择问题关键词和图像关键区域,但两者交互所产生的庞大计算量,将导致模型训练时间过长。Kim 等[11]提出了双线性注意力网络(Bilinear attention networks, BAN[11]),用于学习问题词与图像区域之间的稠密互动关系,与 Yu 等[12]提出的多模态分解双线性池(Multi-modal factorized bilinear pooling, MFB[12])一样,均可以通过级联的方式,形成更为复杂的深度视觉推理共同注意网络,然而深层模型相比浅层模型,其参数量的增加对模型性能的提升微乎其微。Anderson 等[1]提出了一种自底向上一自顶向下的注意力机制,采用 Faster RCNN[13]提取图像的区域对象特征,以代替空间网格特征进行注意力的学习,然而该模型仍存在缺陷,其并未对每个图像区域和问题词之间的密集交互关系进行注意力学习。Yu 等[14]在多头注意力机制(Multi-head attention[15])的基础上,提出了适用于多模态任务的自注意力(Self-attention)和导向注意力(Guided-attention),用来学习模态内部和模态之间的密集交互关系。

上述对视觉问答任务的相关研究,均通过对注意力机制的设计,有效地建立了图像与问题之间的关系,但忽视了自注意力机制过滤噪声信息能力较差的缺陷。以往的模型在目标模态进行自注意力学习过程中,会保留较多噪声信息,从而阻碍模型后续的学习,最终降低模型的性能。例如,当提出问题为“Is the girl sitting on the horse?”时,在进行自注意力学习过程中,图像信息应当侧重建立“girl”和“horse”之间的关系。尽管信息只在相同模态内传播,但其他模态的信息也应该被考虑,并用于过滤目标模态中与预测真实答案无关的信息。

图像字幕任务中 AoA[16](Attention on attention)模块通过拓展传统注意力的方法确保查询值和注意力结果的相关性。受此启发,本文提出了一种基于门控自注意力机制的视觉问答模型,通过使用其他模态作为通道调节门,过滤目标模态在自注意力学习过程的输出结果,从而提高视觉问答模型的性能。该模型在自注意力模块中利用其他模态特征作为通道调节门,过滤目标特征自注意力学习的输出结果,从而减少噪声信息;同时,利用跨模态双导向注意力机制与堆叠注意力模块,共同学习协同注意力和深层注意力,从而获取用于预测真实答案的关键信息;最后将包含丰富注意结果的视觉和语言特征进行特征融合[17],通过多层感知机与 Sigmoid 层获得预测答案。由于该模型在注意力模块中有效地过滤噪声信息,模型的性能将得到提升。

1 方 法

1.1 总体流程

给定图像 $\boldsymbol{I}$ 以及相应问题 $\boldsymbol{T}$,视觉问答任务需要同时理解图像内容和问题语义,从一组候选答案中返回概率最高的答案。本文提出的基于多模态门控自注意力机制的视觉问答模型,主要包括以下四个部分:a)视觉和语言特征提取;b)模态内和模态间注意力学习;c)特征融合;d)多标签分类答案预测。首先,该模型将输入的问题进行分词,通过 Glove[18]词嵌入获取对应的词向量,并将词向量输入门控循环(Gated Recurrent Unit, GRU)[19]网络中,获取问题的语言特征;同时,对于图像输入,通过 Faster RCNN 目标检测器提取概率最高的数个候选区域作为视觉特征;其次,视觉和语言特征通过多个注意力模块的注意力学习,筛选出用于预测真实答案的关键信息;再次,富含丰富注意力信息的视觉和语言特征经过注意力衰减模块后,输入到线性多模态融合函数中获得融合特征;最后将融合特征输入分类网络,以获取预测结果。模型的整体流程如图 1 所示。

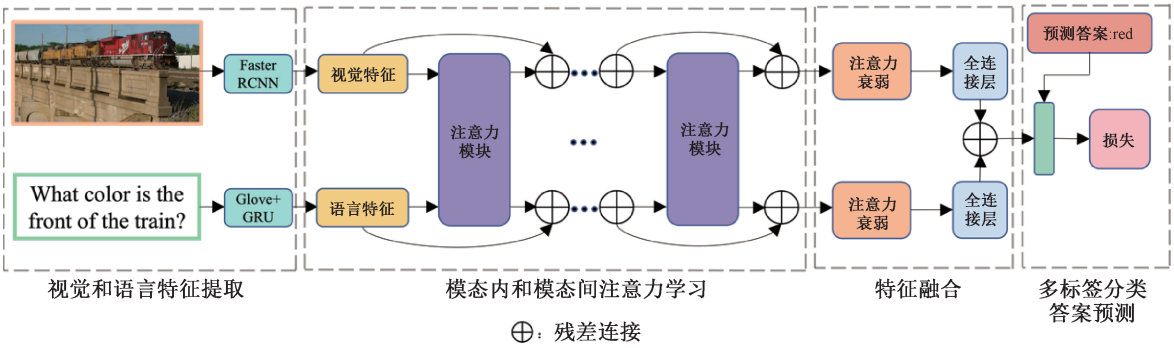


图 1　基于多模态门控自注意机制的视觉问答模型总体流程

1.2　视觉和语言特征提取

为获得视觉特征和语言特征，本文采用自底向上一自顶向下模型[1]的方法。将 Faster RCNN 作为目标检测器，通过提取每张图像中概率最高的 M 个检测目标，将其作为视觉特征。对于每张图像输入 $\boldsymbol{I}$，提取的视觉特征可以表示为 $\boldsymbol{X}\in\mathbf{R}^{M\times dim_X}$。其中：$M$ 为视觉特征的数量；dim_X 为视觉特征的维度；$\boldsymbol{x}_i\in\mathbf{R}^{dim_X}$ 为图像第 i 个视觉区域特征。Faster RCNN 目标检测器为预训练模型，其参数通过 Visual Genome[20] 数据集预训练获得，因此不参与视觉问答模型损失的反向传播。

对于模型的问题输入，首先进行预处理，根据空格和标点符号将问题分割为单词形式；其次将各问题包含的单词数填充或截断到相同长度 N；再次使用 GloVe 模型进行词嵌入；最后将词嵌入输入 GRU 网络获得语言特征。对于每个问题输入 $\boldsymbol{T}$，语言特征可以表示为 $\boldsymbol{Y}\in\mathbf{R}^{N\times dim_Y}$。其中：$N$ 为统一单词数目；dim_Y 为语言特征的维度（即 GRU 隐藏层的维度）；$\boldsymbol{y}_i\in\mathbf{R}^{dim_Y}$ 为问题第 i 个单词的语言特征。视觉和语言特征提取模块可以用方程(1)—(2)表示：

$$\boldsymbol{X}=\mathrm{RCNN}(\boldsymbol{I};\boldsymbol{\theta}_{\mathrm{FasterRCNN}}) \tag{1}$$

$$\boldsymbol{Y}=\mathrm{GRU}(\boldsymbol{T};\boldsymbol{\theta}_{\mathrm{GRU}}) \tag{2}$$

其中：$\boldsymbol{\theta}_{\mathrm{Faster\ RCNN}}$ 和 $\boldsymbol{\theta}_{\mathrm{GRU}}$ 分别表示用于获取视觉特征和语言特征的网络参数。$\boldsymbol{\theta}_{\mathrm{Faster\ RCNN}}$ 为预训练模型参数，因此在训练过程中参数固定；而 $\boldsymbol{\theta}_{\mathrm{GRU}}$ 随整体模型训练进行学习。

1.3　模态内和模态间注意力学习

本文所提出的模型通过堆叠注意力模块进行模态内和模态间注意力学习，注意力均基于多头注意力机制实现。对于每层堆叠注意力模块，视觉和语言特征首先通过门控自注意力机制学习模态内信息；然后利用跨模态双导向注意力机制，实现视觉引导语言注意和语言引导视觉注意，即模态间注意力学习；最后将该层的输出结果作为下一层堆叠注意力模块的输入，学习深层注意力。由于经过多层堆叠注意力模块的注意力学习，视觉和语言特征均包含丰富的注意力结果，故模型能够有效地联系图像和问题之间的信息。

1.3.1　缩放点积注意力和多头注意力

Vaswani 等[15]在 Transformer 结构中提出一种新颖的注意力机制，即缩放点积注意力机制。该注意力机制最早运用于自然语言处理中，由于其表现出优秀的性能，随后也应用于计算机视觉和多模态任务中。缩放点积注意力机制可以简单描述为：将查询(Query)和一组键值对(Key-Value)映射到一个输出上。其整体流程如图 2(a)所示：首先通过函数计算查询和键之间的相似度，为了防止相似度结果数值过大，统一除以缩放因子；然后通过 softmax 函数将相似度转换为值的权重；最后对所有值进行加权求和，从而获得查询和键之间的注意力。计算公式如方程(3)所示：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

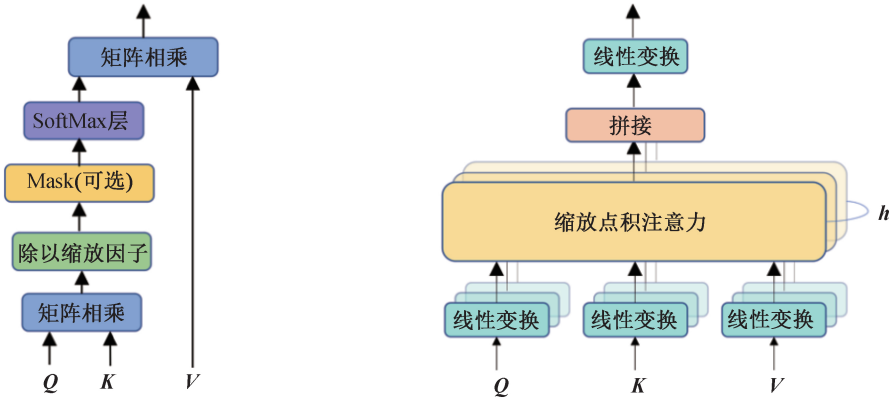
其中：$\boldsymbol{Q}$ 为查询，$\boldsymbol{K}$ 为键，$\boldsymbol{V}$ 为值，d_k 为查询和键的共同维度。多头注意力机制如图 2(b)所示，是缩放点积注意力机制的改进。通过拼接 h 个并行的缩放点积注意力，将拼接结果作为多头注意力的输出，从而允许模型联合关注来自不同位置不同子空间表示的信息。其计算公式可以用方程(4)—(5)表示：

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h)\boldsymbol{W}^{\mathrm{o}} \tag{4}$$

$$\boldsymbol{head}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{Query}},\boldsymbol{K}\boldsymbol{W}_i^{\mathrm{Key}},\boldsymbol{V}\boldsymbol{W}_i^{\mathrm{Value}}) \tag{5}$$

1.3.2　视觉问答任务中的自注意力和导向注意力

Yu 等[14]将上述的多头注意力机制引入到视觉问答任务中，分别设计了自注意力(Self-Attention)单元和导向注意力(Guided-Attention)单元。自注意力单元由多头注意力层和逐点前馈层组成，用来

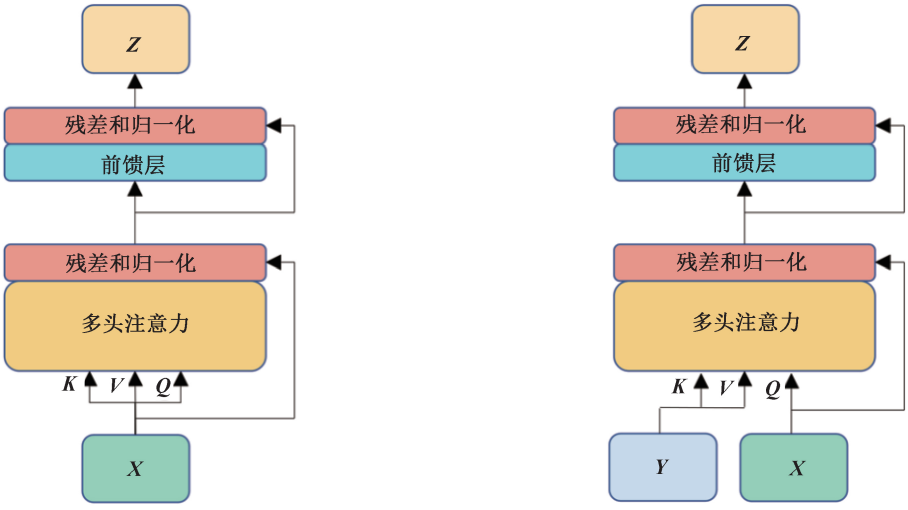


(a) 缩放点积注意力 (b) 多头注意力

图 2 缩放点积注意力和多头注意力流程图

学习模态内信息的关系,其结构如图 3(a)所示:自注意力单元首先将同一模态特征 $\boldsymbol{X}=[\boldsymbol{x}_1;\cdots;\boldsymbol{x}_m]\in\mathbf{R}^{m\times dx}$ 作为输入;然后利用多头注意力学习特征 $\boldsymbol{X}$ 中$<\boldsymbol{x}_i,\boldsymbol{x}_j>$两两之间关系,并输出包含丰富自注意力信息的特征 $\boldsymbol{Z}\in\mathbf{R}^{m\times d}$;最后逐点前馈层将 Z 作为输入,通过包含 ReLU 和 Dropout 的全连接网络,获得最终结果。除此之外,多头注意力层和前馈层的输出都应用归一化操作和残差连接,从而解决多层神经网络训练困难的问题,并加速了模型的收敛。导向注意力单元如图 3(b)所示,表示为 $\boldsymbol{Y}$ 引导 $\boldsymbol{X}$ 进行注意力学习。其输入 $\boldsymbol{X}=[\boldsymbol{x}_1;\cdots;\boldsymbol{x}_m]\in\mathbf{R}^{m\times dx}$ 和 $\boldsymbol{Y}=[\boldsymbol{y}_1;\cdots;\boldsymbol{y}_m]\in\mathbf{R}^{n\times dy}$ 为两种不同模态的特征,用于学习模态间信息的关系。

虽然视觉问答任务中使用自注意力单元,能够帮助模型建立图像和问题之间的联系,但在完全不依靠其他模态信息的情况下进行自注意力学习,仍然会保留较多的噪声信息。这些与预测真实答案无关的信息,将会阻碍模型后续的学习。因此,本文提出了门控自注意力机制,通过将其他模态信息作为通道调节门,用于过滤目标模态自注意力学习过程中的输出结果,从而减少噪声信息,以提高视觉问答模型的性能。



(a) 自注意力单元 (b) 导向注意力单元

图 3 自注意单元和导向注意力单元结构图

1.3.3 门控自注意力机制

本文以自注意力单元为原型,构建了如图 4 所示的门控自注意力单元。

以语言特征 $\boldsymbol{Y}=[\boldsymbol{y}_1,\cdots,\boldsymbol{y}_N]\in\mathbf{R}^{N\times dim}$ 为例,门控自注意力的整体流程为:首先每个语言特征被转化为查询(Query)、键(Key)和值(Value),其计算过程如方程(6)所示:

$$\begin{cases}\boldsymbol{Y}_{\text{Query}}=\boldsymbol{Y}\cdot\boldsymbol{W}_{\text{Query}},\\ \boldsymbol{Y}_{\text{Key}}=\boldsymbol{Y}\cdot\boldsymbol{W}_{\text{Key}},\\ \boldsymbol{Y}_{\text{Value}}=\boldsymbol{Y}\cdot\boldsymbol{W}_{\text{Value}}\end{cases}\tag{6}$$

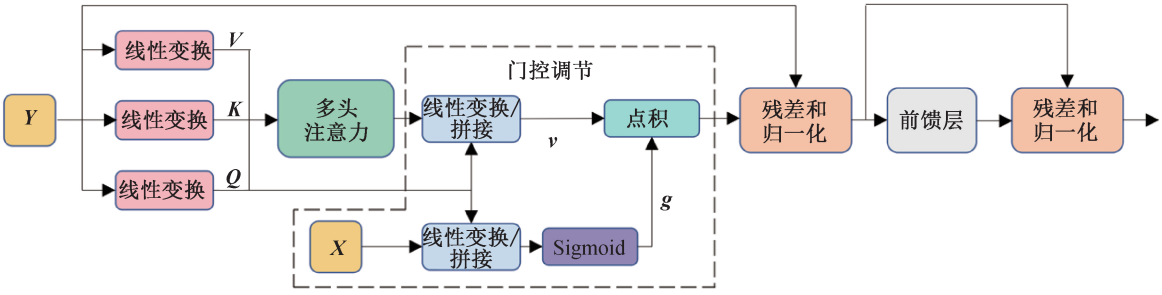


图 4　门控自注意力单元结构示意图

其中：$\boldsymbol{Y}_{\text{Query}}, \boldsymbol{Y}_{\text{Key}}, \boldsymbol{Y}_{\text{Value}} \in \mathbf{R}^{N\times dim}$，分别表示语言特征的查询、键和值，$dim$ 为模态特征的统一维度，$\boldsymbol{W}_{\text{Query}}, \boldsymbol{W}_{\text{Key}}, \boldsymbol{W}_{\text{Value}} \in \mathbf{R}^{dim\times dim}$。其次使用多头注意力对语言特征模态内信息进行重新构建。相比于缩放点积注意力，多头注意力能够进一步提升特征的表示能力，从而获得更好的细粒度语义理解，提升模型的性能，计算过程如方程(7)—(9)所示：

$$Z = \text{MultiHead}(\boldsymbol{Y}_{\text{Query}}, \boldsymbol{Y}_{\text{Key}}, \boldsymbol{Y}_{\text{Value}}) = \text{Concat}(\boldsymbol{head}_1, \cdots, \boldsymbol{head}_h)\boldsymbol{W}^{\circ} \tag{7}$$

$$\boldsymbol{head}_i = \text{Attention}(\boldsymbol{Y}_{\text{Query}}\boldsymbol{W}_i^{\text{Query}}, \boldsymbol{Y}_{\text{Key}}\boldsymbol{W}_i^{\text{Key}}, \boldsymbol{Y}_{\text{Value}}\boldsymbol{W}_i^{\text{Value}}) \tag{8}$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_h}}\right)\boldsymbol{V} \tag{9}$$

其中：$\boldsymbol{W}_i^{\text{Query}}, \boldsymbol{W}_i^{\text{Key}}, \boldsymbol{W}_i^{\text{Value}} \in \mathbf{R}^{dim\times dh}$，是第 i 个缩放点积平行头 $\boldsymbol{head}_i$ 的投影矩阵；$dh = dim/h$ 为平行头输入的特征维度；$\boldsymbol{W}^{\circ} \in \mathbf{R}^{h*dh\times dim}$；Concat()表示矩阵的拼接，经过多头注意力重建后的语言特征表示为 $\boldsymbol{Z} \in \mathbf{R}^{N\times dim}$。再次将多头注意力的输出结果进行门控调节，门控调节单元包含两条信息流的输入：a)通过多头注意力重建后的特征 $\boldsymbol{Z}$ 和查询信息 $\boldsymbol{Y}_{\text{Query}}$ 所组成的模态内信息，b)视觉特征 $\boldsymbol{X}$ 和查询信息 $\boldsymbol{Y}_{\text{Query}}$ 所组成的通道调节门信息。门控调节单元中两条信息输入流 $\boldsymbol{v}$ 和 $\boldsymbol{g}$ 可以通过如方程(10)—(12)计算得到：

$$\hat{\boldsymbol{X}} = \boldsymbol{W}_{\text{transform}}\boldsymbol{X} \tag{10}$$

$$\boldsymbol{g} = \sigma(\boldsymbol{W}_1^{\text{gate}}\boldsymbol{Y}_{\text{Query}} + \boldsymbol{W}_2^{\text{gate}}\hat{\boldsymbol{X}}) \tag{11}$$

$$\boldsymbol{v} = \boldsymbol{W}_1^{\text{intra}}\boldsymbol{Y}_{\text{Query}} + \boldsymbol{W}_2^{\text{intra}}\boldsymbol{Z} \tag{12}$$

其中：$\boldsymbol{W}_1^{\text{gate}}, \boldsymbol{W}_2^{\text{gate}}, \boldsymbol{W}_1^{\text{intra}}, \boldsymbol{W}_2^{\text{intra}} \in \mathbf{R}^{dim\times dim}$；$\boldsymbol{v}, \boldsymbol{g} \in \mathbf{R}^{N\times dim}$；$\sigma$ 为 Sigmoid 函数。为保证特征数量相同，视觉特征 X 进行线性变换，权重矩阵为 $\boldsymbol{W}_{\text{transform}} \in \mathbf{R}^{N\times M}$。为了过滤噪声信息，将模态内信息 $\boldsymbol{v}$ 和通道调节门 $\boldsymbol{g}$ 进行点积运算。从而防止自注意力过程中保留过多与预测真实答案无关的信息，影响模型后续的训练。门控过滤过程如方程(13)—(14)所示，其中⊙表示矩阵的点积运算，LayerNorm()表示层归一化：

$$\hat{\boldsymbol{v}} = \boldsymbol{g} \odot \boldsymbol{v} \tag{13}$$

$$\boldsymbol{Y}' = \text{LayerNorm}(\boldsymbol{Y} + \hat{\boldsymbol{v}}) \tag{14}$$

将门控调节的输出 $\boldsymbol{Y}'$ 输入前馈层，并进行残差连接和归一化操作，获得整个门控自注意力单元的最终输出结果。

1.3.4　堆叠注意力模块

模态内和模态间注意力学习是通过 L 个相同的注意力模块堆叠而成，包含门控自注意力和跨模态导向注意力，其结构如图 5 所示。与 MCAN[14] 网络中单导向注意力不同，在视觉问答任务中，图像和问题之间具有天然的对称性，即语言特征可以用于引导视觉注意，而视觉特征也应该被用于引导语言注意，因此在模型中使用跨模态双导向注意力。由于注意力模块的输入和输出特征维度相同，可以将上一个注意力模块的输出作为下一个注意力模块的输入，从而形成深层注意力模型。每个注意力模块结束位置都进行残差连接，防止网络过深带来的梯度消失问题。

1.4　特征融合

经过 L 个注意力模块的学习后，此时视觉特征的输出为 $\boldsymbol{X}^{(L)} = [\boldsymbol{x}_1^{(L)}; \cdots; \boldsymbol{x}_M^{(L)}] \in \mathbf{R}^{M\times dim}$；语言特征则为 $\boldsymbol{Y}^{(L)} = [\boldsymbol{y}_1^{(L)}; \cdots; \boldsymbol{y}_N^{(L)}] \in \mathbf{R}^{N\times dim}$；均包含了关于问题词和图像区域注意力权重的丰富信息。与 MCAN 中做法一致：首先将 $\boldsymbol{X}^{(L)}$ 和 $\boldsymbol{Y}^{(L)}$ 输入到注意力衰弱模型(全连接层-ReLU-Dropout－全连接层)中；然后通过 softmax 函数计算出特征权重 $\boldsymbol{A}_{\text{visual}} = [a_1^{\text{visual}}; \cdots; a_M^{\text{visual}}] \in \mathbf{R}^M$，$\boldsymbol{A}_{\text{lang}} = [a_1^{\text{lang}}; \cdots; a_N^{\text{lang}}] \in \mathbf{R}^N$；最后特征权重与 $\boldsymbol{X}^{(L)}$ 和 $\boldsymbol{Y}^{(L)}$ 进行加权求和，获得视觉和语言特征向量用于多模态特征融合：

$$\boldsymbol{A}_{\text{visual}} = \text{softmax}(\text{MLP}(\boldsymbol{X}^{(L)})) \tag{15}$$

$$\boldsymbol{A}_{\text{lang}} = \text{softmax}(\text{MLP}(\boldsymbol{Y}^{(L)})) \tag{16}$$

$$\bar{\boldsymbol{x}} = \sum_{i=1}^{M} a_i^{\text{visual}}\boldsymbol{x}_i^{(L)} \tag{17}$$

$$\bar{\boldsymbol{y}} = \sum_{i=1}^{N} a_i^{\text{lang}}\boldsymbol{y}_i^{(L)} \tag{18}$$

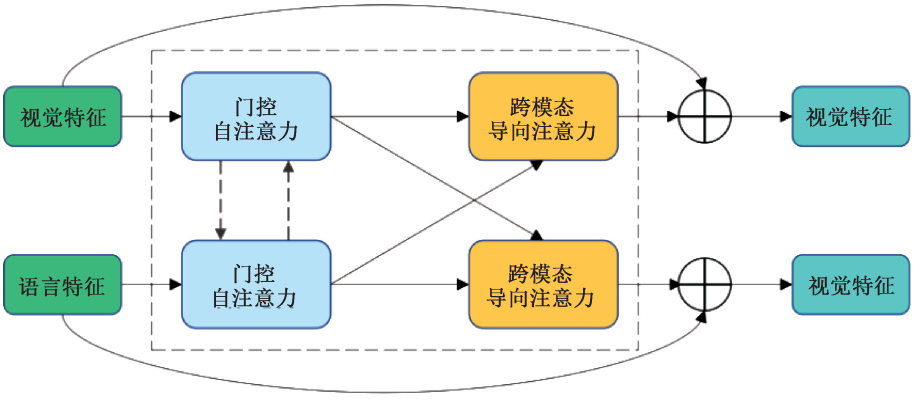


图 5 注意力模块

将计算出的 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 输入方程(19)所示的线性多模态融合函数中,得到融合特征 $\boldsymbol{f}$:

$$\boldsymbol{f}=\mathrm{LayerNorm}(\boldsymbol{W}_{\mathrm{visual}}^{\mathrm{fuse}}\boldsymbol{x}+\boldsymbol{W}_{\mathrm{lang}}^{\mathrm{fuse}}\boldsymbol{y}) \tag{19}$$

其中:$\boldsymbol{W}_{\mathrm{visual}}^{\mathrm{fuse}},\boldsymbol{W}_{\mathrm{lang}}^{\mathrm{fuse}}\in\mathbf{R}^{dim\times df}$,$df$ 为融合特征的维度。

1.5 多标签分类答案预测

融合特征经过线性变换后,将维度 d_z 转化为候选答案维度 N_{ans},通过 Sigmoid 函数得到答案预测结果。本文使用二元交叉熵(BCE)作为损失函数训练 N_{ans} 分类问题,损失函数如方程(20)所示:

$$C_{\mathrm{loss}}=-\sum_{i}^{N_{\mathrm{ans}}}a_i\log(a'_i)-(1-a_i)\log(1-a'_i) \tag{20}$$

其中:N_{ans} 为多分类中类型的数目,a'_i 为第 i 个类的预测值,a_i 为第 i 个类的标签值。

2 实验与结果分析

2.1 数据集

2.1.1 VQA-v2

VQA-v2 数据集源于 MS COCO[21] 数据集,并在其基础上添加了人工标注的问题和答案,是目前 VQA 任务中最常用的基准数据集。为解决上一代 VQA-v1 数据集的缺陷,VQA-v2 数据集中每个问题都对应两张不同的图片,并有着两种不同的回答,有效防止模型通过简单记忆来表现出优秀的性能。数据集包含大约 1.11×10^6 个问题和 2.04×10^5 张图片,其中训练集、验证集和测试集各占 40%、20% 和 40%。VQA v2 数据集样例如图 6 所示,问题被分为三种类型:是/否(Yes/No)、其他(Other)和计数(Num),并且每个问题包含 10 个人工标注答案。

2.1.2 VQA-CP v2

VQA-CP v2[22] 数据集是评估 VQA 模型泛化能力的诊断基准,通过对 VQA-v2 数据集进行重新划分,使得分割后训练集和测试集的答案分布显著不同。例如以“What sport…”作为开头的问题中,“tennis”是训练集中最常见的答案,而“skiing”则是测试集中最常见的答案。训练集包含约 1.21×10^5 张图片和 4.38×10^5 个问题,测试集包括约 9.80×10^4 张图片和 2.20×10^5 个问题。

(a) 是/否问题 (b) 其他问题 (c) 计数问题

图 6 VQA v2 数据集样例

2.2 评价指标

由于 VQA-v2 和 VQA-CP v2 数据集中每个问题都包含 10 个人工标注的答案,因此问题的答案不唯一。为了统一评估标准,本文采用 Antol 等[23] 提

出的方法，评估方程如(21)所示：

$$\mathrm{acc}(a)=\min\left\{\frac{\#\,\mathrm{humansprovidedthatanswer}a}{3},1\right\} \tag{21}$$

即当前预测答案占 10 个人工标注答案中三个以上时，才会被认为是完全正确。

2.3　实验参数设置

本文模型在 Nvidia GeForce RTX 3090 显卡，Ubuntu16.04 操作系统，CUDA 11.1 支持下的 PyTorch 1.9.0 框架上进行实验。对于图片特征提取，本文通过 Faster RCNN 提取每张图片中概率最高的 $M=100$ 个视觉特征。对于问题特征提取，本文将每个问题进行填充或截取至统一长度 $N=14$ 并输入到隐藏层维度为 512 的 GRU 网络中，GRU 网络的输出作为问题特征。其中视觉特征维度 dim_X、语言特征维度 dim_Y 和融合特征维度 df 分别为 2048、512 和 1024。为方便起见，视觉特征和语言特征进入注意力模块前将转化为统一维度 $dim=512$。模型中多头注意力包含 $h=8$ 个平行缩放点积注意力。选取 $N_{\mathrm{ans}}=3000$ 个在训练集中最常见的答案作为多分类问题的预测向量。并分别对注意模块堆叠数 $L\in\{1,2,4,6,8,12\}$进行了实验。

训练过程中，模型的 Dropout 层均使用 0.5 的概率，以 64 为训练批度在 Adam 优化器($\beta_1=0.9$，$\beta_2=0.98$)上训练总计 16 个周期，前 10 个周期的学习率为 0.0001，10 个周期后，每个周期的学习率下降 1/10。

2.4　消融实验

基于门控自注意力机制的视觉问答模型由多个模块组成，为了验证门控自注意力和双导向注意力模块的有效性，在 VQA-v2 数据集上对该模型进行消融实验。实验中，为了控制变量以下模型中注意力模块层数 $L=1$，多头注意力中平行头数 $h=8$，统一特征维度 $dim=512$，视觉问答模型的变体如下所示：

a)基线模型：该模型在注意力模块中仅使用自注意力和问题引导图像的单导向注意力。

b)基线模型＋门控自注意力：该模型在注意力模块中使用门控自注意力和问题引导图像的单导向注意力。

c)基线模型＋双导向注意力：该模型在注意力模块中使用自注意力和双导向注意力，即问题引导图像注意力和图像引导问题注意力。

d)基线模型＋门控自注意力＋双导向注意力：该模型为本文所提出的完整视觉问答模型，注意力模块如图 5 所示，使用门控自注意力和双导向注意力，即问题引导图像注意力和图像引导问题注意力。

本文将变体模型在 VQA-v2 训练集上进行训练，并通过 VQA-v2 验证集进行测试。实验结果如表 1 所示，可以得出如下结论：通过对比仅添加门控自注意力的模型和基线模型，前者在各类问题的预测精度上均有提升，并且总准确率提升了 0.19%，从而验证了门控自注意力模块的有效性。这是因为在目标模态自注意力过程中添加来自其他模态的信息作为通道调节门，能够有效过滤噪声信息，从而提升模型性能。为了验证双导向注意力模块的有效性，将仅添加双导向注意力的模型和基线模型进行对比，前者在 VQA-v2 数据集上总准确率提升了 0.30%。这是因为图像和问题具有天然的对称性，在包含问题引导图像的跨模态注意力学习中，通过添加图像引导问题的跨模态注意力，可以加强图像特征和问题特征的密集交互关系，丰富问题特征的表示，使得问题特征包含更细致的语义信息，从而提升模型性能。将本文所提出的两个模块均添加到基线模型上，总准确率达到 65.95%，相较于只添加单个模块的模型各类问题的预测精度均有所提升，可以看出本文所提出的两个模块不存在冲突。

表 1　基于多模态门控自注意力机制的视觉问答模型消融实验结果　%

模型	总准确率	是/否	计数	其他
基线模型	65.42	83.21	44.95	57.23
基线模型＋门控自注意力	65.61	83.55	45.12	57.37
基线模型＋双导向注意力	65.72	83.41	45.26	57.49
基线模型＋门控自注意力＋双导向注意力	**65.95**	**83.64**	**45.37**	**57.64**

为了研究注意力模块堆叠层数对模型性能的影响，本文分别对模型中注意力模块堆叠层数 $L\in\{1,2,4,6,8,12\}$进行实验，结果如表 2 所示。从表 2 中可以看出，L 在一定范围内时，注意力模块堆叠层数的增加可以提升模型的性能。这是因为随着深度的增加，模型能够学习到问题和图像之间更深层的关系，推进语言特征和视觉特征进行更好的细粒度语义理解和视觉推理，最终提升模型的性能。实验结果表明，当 $L=6$ 时模型达到最佳效果。由于模型深度过深时会导致浅层的重要信息遗忘，因此当 $L>8$时，低于模型最佳性能。

表 2　不同注意力模块堆叠层数 L 下的模型总准确率　%

堆叠层数	1	2	4	6	8	12
MGSA	65.95	66.43	67.18	**67.42**	**67.42**	67.38

深层模型往往存在模型参数量大,并且随着深度增加模型性能提升不明显的问题。为了展现 MGSA 模型的优势,本文将 MGSA 与同为深层模型的 BAN 进行比较。BAN 是一种通过残差求和的方式来增加网络深度的模型,其最高实现八个“一瞥”,在深层模型中表现出优秀的性能。其中单层 BAN 参数量为 4.92×10^7,而本文所提出的单层 MGSA 则为 2.21×10^7;当 BAN 深度为 8 时其性能达到最佳,此时模型参数量为 9.06×10^7,而本文所提出的模型 MGSA 在深度为 6 时性能达到最佳,其模型参数量为 6.41×10^7;即使 MGSA 的深度到达 8 时,模型的参数量仅仅为 8.09×10^7,均比同为深层模型的 BAN 低。而且从图 7 可以看出,在增加 BAN 的模型深度时,其模型表现出的性能并未得到明显地提升,而本文所提出的 MGSA 模型在提升模型深度的同时,模型性能提升显著。

由于模型在 VQA-v2 数据集上性能的提升,可能是模型利用单峰偏差(文本偏见)来提供正确的答案,而无需利用图像信息。这些利用单峰偏差的模型泛化能力较差,并在域外数据集上的性能将大幅度下降。为了研究本文所提出的 MGSA 模型的泛化能力,将其在 VQA-CP v2 数据集上进行实验,实验结果如表 3 所示。从表 3 中可以发现,同为基于

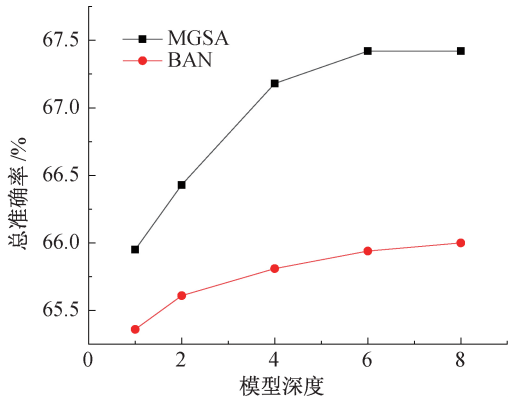


图 7 MGSA 和 BAN 模型深度与准确率关系曲线

自底向上机制的深层模型 BAN 和本文所提出的 MGSA 模型,BAN 模型在 VQA-CP v2 上各指标的表现均差于基线模型 BUTD,而 MGSA 各指标相较于 BUTD,分别有 0.25%、0.61%、1.04% 和 0.68%的提升。这是因为 MGSA 模型中视觉特征和语言特征之间是充分进行交互的,无论是门控自注意力还是双导向注意力都能够保证语言特征在注意力学习过程中均涉及视觉特征信息,有效防止了模型忽略视觉特征信息所导致的单峰偏差。配合消除文本偏见的集合模型 LPF[24],本文所提出的 MGSA 模型在 VQA-CP v2 上能够达到 55.81%的总准确率。

表 3 模型在 VQA CP-v2 数据集上的总准确率 %

模型	VQA-CP v2				VQA v2			
	是/否	计数	其他	总准确率	是/否	计数	其他	总准确率
BUTD	45.21	11.96	42.98	39.49	81.82	44.21	56.05	64.32
BUTD+LPF	88.61	23.78	46.57	55.34	—	—	—	—
BAN	41.55	12.43	41.40	37.03	85.31	50.93	60.26	69.52
BAN+LPF	88.13	18.59	40.03	50.76	—	—	—	—
MGSA(本文)	45.46	12.57	44.02	40.17	**86.96**	**53.21**	**60.84**	**70.76**
MGSA(本文)+LPF	**88.97**	**24.23**	**47.37**	**55.81**	—	—	—	—

为了寻找模型的最佳效果,对模型的部分参数进行了实验,实验结果如表 4 所示。在同时使用包含双导向注意力和门控注意力的 6 层注意力模块网络中,多头注意力使用不同数量的平行头对模型的预测精度有一定的影响。实验表明,当使用 8 个平行头时,模型达到最佳效。对视觉特征和语言特征在进入注意力模块时的统一维度 dim 最优取值展开研究,分别对 dim 取 512、1024 和 2048 时进行了实验,从表 4 中可以看出当统一维度为 1024 时模型的效果最佳。

表 4 不同参数下模型的总准确率

模型参数	参数值	总准确率/%
多头注意力的平行头数目 h	1×512	66.89
	4×128	67.17
	8×64	**67.42**
	16×32	67.23
统一维度 dim	512	67.42
	1024	**67.61**
	2048	67.29

2.5 定性分析

近期相关研究[25-26]改进了 Grad-CAM 来推导模型中每个图像特征对最终答案的贡献度,并将其运用于视觉问答任务中。为了更清晰地研究门控自注意力对模型预测所起到的作用,本文通过 Chen 等[27]提出的式(22)计算了第 i 个对象特征对真实答案 a 的贡献度:

$$s(a,\boldsymbol{v}_i)=S(P_{\mathrm{vqa}}(a),\boldsymbol{v}_i):=(\nabla_{\boldsymbol{v}_i}P_{\mathrm{vqa}}(a))^{\mathrm{T}}\mathbf{1} \quad (22)$$

其中：$P_{\text{vqa}}(a)$为模型预测答案 a 的概率，$\boldsymbol{v}_i$ 为输入图像的第 i 个对象特征。回答问题所需关键特征 v_i 的贡献度 $s(a,\boldsymbol{v}_i)$越高说明模型在模态内和模态间注意力学习过程中能够准确筛选出重要信息。

图像区域贡献度可视化的样例如图 8 所示，其中：图 8(a)—(c)均为使用门控自注意力的模型在回答问题时所提取的图像特征贡献度，而图 8(d)—(f)为普通自注意力模型所提取的图像特征贡献度。对比样例 1 和样例 4，可以看出：当询问“what is flying in the air?”时，虽然两个模型都能给出正确的答案“kites”，但是在添加了门控机制的自注意力模型中，关键特征的贡献度更高。而通过对比样例 2 和样例 5，可以看出：未添加门控机制的自注意力模型中各特征的贡献度较为平均，模型在预测过程中无法确认用于回答问题的关键图像特征，从而导致给出错误的答案。而添加了门控机制的自注意力模型中关键图像特征区域仍然能够保持较高的贡献度，这是因为普通自注意力机制结果仍然保留较多噪声信息，这些噪声信息会影响模型后续的训练，使得模型无法准确判断哪些对象特征是回答问题的关键，导致模型对各个对象特征无差别对待，最后模型预测出错误的答案。而添加门控机制的自注意力模型，在自注意力过程中就能过滤部分与真实答案不相关的信息，使得模型能够更加关注关键信息，从而提升模型性能。

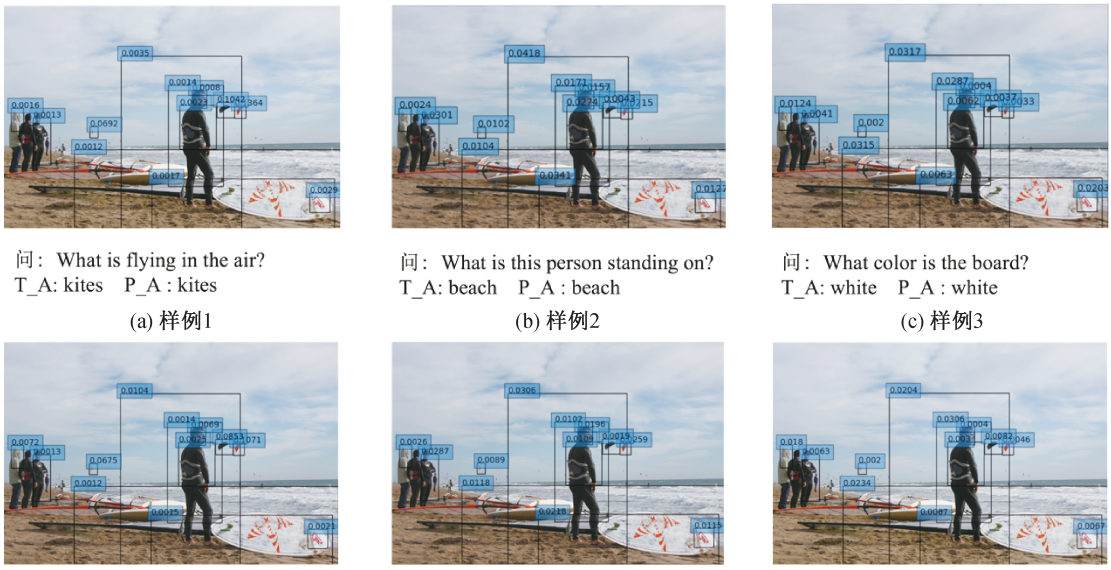


图 8　图像区域贡献度的样例

注：T_A 为真实答案；P_A 为预测答案。

2.6　模型总体性能

本文使用最佳单一模型与当前主要的视觉问答模型进行比较，实验结果如表 5 所示。BUTD 模型将 Faster RCNN 提取的图像对象特征，替代图像网格特征进行注意力学习。MFH 是最先进的双线性融合方法。BAN 提出了一种双线性注意力机制，利用每个通道的信息生成多模态注意力。DCN 提出了一种密集连接的共同注意力机制用于跨模态特征学习，在预测精度上也具有不错的表现。DFAF 是一个多层堆叠的网络，他动态的学习模态内部和模态之间的信息流来实现特征的融合。MLIN 也是一个多层的堆叠网络，通过少量多模态信息来建模多模态交互作用，有助于从全局角度对模态之间进行建模和编码，避免捕获太多视觉和语言之间的无关信息。MCAN 中提出了一种基于 Encoder-Decoder 的联合注意力学习方式，提高了模型的预测精度。而本文所提出的模型 MGSA，通过另一模态门控自注意力的学习，过滤掉与无关的信息，从而提高模型的预测精度，从表格中可以看出，总准确率达到 71.12%，并在 Yes/No，Other 和 All 指标上的准确度均具有竞争性。

3　结　语

本文提出了一种基于门控自注意力机制的视觉问答模型，使用其他模态信息作为通道调节门过滤目标模态自注意力学习的结果，从而减少与真实答案无关的信息；考虑到图像和问题具有天然的对称性，本文提出的模型通过跨模态双导向注意力学习

表 5 VQA v2 测试集上本文与当前主要视觉问答模型的准确度 %

模型	Test-dev				Test-std
	是/否	计数	其他	总准确率	总准确率
BUTD[1]	81.82	44.21	56.05	64.32	65.67
MFH[12]	—	—	—	66.12	—
BAN[11]	85.31	50.93	60.26	69.52	—
DCN[28]	83.51	46.61	57.26	66.87	66.97
DFAF[4]	86.09	**53.32**	60.49	70.22	70.34
MLIN[5]	85.96	52.93	60.40	70.18	70.28
MCAN[14]	86.82	53.26	60.72	70.63	70.90
MGSA(本文)	**86.96**	53.21	**60.84**	**70.76**	**71.12**

模态间的协同注意力;注意力模块通过堆叠的形式构成深层模型,加强注意力的学习。大量实验表明,采用门控自注意力机制的模型能够更好地过滤噪声信息,加强关键信息的权重,从而提升视觉问答模型的性能。与当前主要的视觉问答模型相比,本文所提出的模型具有一定的竞争力。

然而,在目前视觉问答任务中,均使用基于边界框(或区域)的视觉特征与问题特征相结合来预测最终的答案。这种设计标准使得模型仅包含图像的区域级信息,却忽略了图像整体的全局信息对答案预测的重要性。因此在后续研究中,可尝试提升模型对全局图像信息的理解,更好地预测答案。

参考文献:

[1] Anderson P, He X, 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: 6077-6086.

[2] Cornia M, Stefanini M, Baraldi L, et al. Meshed-memory transformer for image captioning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10575-10584.

[3] 闫茹玉, 刘学亮. 结合自底向上注意力机制和记忆网络的视觉问答模型[J]. 中国图象图形学报, 2020, 25(5): 993-1006.

[4] Gao P, Jiang Z, You H, et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 6632-6641.

[5] Peng G, You H X, Zhang Z P, et al. Multi-modality latent interaction network for visual question answering [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 5824-5834.

[6] Xie N, Lai F, Doran D, et al. Visual entailment: A novel task for fine-grained image understanding[EB/OL]. (2019-01-20)[2021-11-09]. https://arxiv.org/abs/1901.06706.

[7] Gurari D, Li Q, Stangl A J, et al. VizWiz grand challenge: Answering visual questions from blind people [C]//2018 IEEE/CVF Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3608-3617.

[8] Ren F J, Zhou Y Y. CGMVQA: A new classification and generative model for medical visual question answering[J]. IEEE Access, 2020, 8: 50626-50636.

[9] 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)[2021-11-09]. https://arxiv.org/abs/1511.05960.

[10] Lu J S, Yang J W, Batra D, et al. Hierarchical question-image co-attention for visual question answering[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: MIT, 2016, 29: 289-297.

[11] Kim J H, Jun J, Zhang B T. Bilinear attention networks [EB/OL]. (2018-10-19) [2021-11-09]. https://arxiv.org/abs/1805.07932.

[12] Yu Z, Yu J, Xiang C C, et al. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 5947-5959.

[13] 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.

[14] 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: IEEE, 2019: 6274-6283.

[15] 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.

[16] Huang L, Wang W M, Chen J, et al. Attention on attention for image captioning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 4633-4642.

[17] Fukui A，Park D H，Yang D，et al. Multimodal compact bilinear pooling for visual question answering and visual grounding[EB/OL]. (2016-09-24)[2021-11-09]. https://arxiv.org/abs/1606.01847.

[18] 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：Association for Computational Linguistics，2014：1532-1543.

[19] Chung J，Gulcehre C，Cho K H，et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2021-11-09]. https://arxiv.org/abs/1412.3555.

[20] Krishna R，Zhu Y，Groth O K，et al. Visual genome：Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision，2017，123(1)：32-73.

[21] Lin T Y，Maire M，Belongie S，et al. Microsoft COCO：Common objects in context[C]//European Conference on Computer Vision. Zürich，Switzerland：Springer，2014：740-755.

[22] Agrawal A，Batra D，Parikh D，et al. Don't just assume；look and answer：Overcoming priors for visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：4971-4980.

[23] Antol S，Agrawal A，Lu J S，et al. VQA：Visual question answering[C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago，Chile：IEEE，2015：2425-2433.

[24] Liang Z，Hu H，Zhu J. LPF：A language-prior feedback objective function for de-biased visual question answering[EB/OL]. (2021-06-23)[2021-11-09]. https://arxiv.org/abs/2105.14300.

[25] Jain S，Wallace B C. Attention is not explanation[EB/OL]. (2019-05-08)[2021-11-09]. https://arxiv.org/abs/1902.10186.

[26] Selvaraju R R，Lee S，Shen Y L，et al. Taking a HINT：Leveraging explanations to make vision and language models more grounded[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul，Korea (South)：IEEE，2019：2591-2600.

[27] Chen L，Yan X，Xiao J，et al. Counterfactual samples synthesizing for robust visual question answering[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle：IEEE，2020：10797-10806.

[28] Nguyen D K，Okatani T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：6087-6096.

(责任编辑：康　锋)