计算机科学
COMPUTER SCIENCE

基于跨模态信息过滤的视觉问答网络

何世阳, 王朝晖, 龚声蓉, 钟珊

引用本文

何世阳, 王朝晖, 龚声蓉, 钟珊. 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): 85-91.

HE Shiyang, WANG Zhaohui, GONG Shengrong, ZHONG Shan. Cross-modal Information Filtering-based Networks for Visual Question Answering [J]. Computer Science, 2024, 51(5): 85-91.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300202

基于跨模态信息过滤的视觉问答网络

何世阳[1]　王朝晖[2]　龚声蓉[1,3]　钟　珊[3]
1 苏州大学计算机科学与技术学院　江苏 苏州 215008
2 苏州大学东吴学院　江苏 苏州 215006
3 常熟理工学院计算机科学与工程学院　江苏 苏州 215500
(bujiayana@163.com)

摘　要　视觉问答作为多模态任务,瓶颈在于需要解决不同模态间的融合问题,这不仅需要充分理解图像中的视觉和文本,还需具备对齐跨模态表示的能力。注意力机制的引入为多模态融合提供了有效的路径,然而先前的方法通常将提取的图像特征直接进行注意力计算,忽略了图像特征中含有噪声和不正确的信息这一问题,且多数方法局限于模态间的浅层交互,未曾考虑模态间的深层语义信息。为解决这一问题,提出了一个跨模态信息过滤网络,即首先以问题特征为监督信号,通过设计的信息过滤模块来过滤图像特征信息,使之更好地契合问题表征;随后将图像特征和问题特征送入跨模态交互层,在自注意力和引导注意力的作用下分别建模模态内和模态间的关系,以获取更细粒度的多模态特征。在 VQA2.0 数据集上进行了广泛的实验,实验结果表明,信息过滤模块的引入有效提升了模型准确率,在 test-std 上的整体精度达到了 71.51%,相比大多数先进的方法具有良好的性能。

关键词:视觉问答;深度学习;注意力机制;多模态融合;信息过滤
中图分类号　TP391

Cross-modal Information Filtering-based Networks for Visual Question Answering

HE Shiyang[1],WANG Zhaohui[2],GONG Shengrong[1,3] and ZHONG Shan[3]
1 School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215008,China
2 Soochow College,Soochow University,Suzhou,Jiangsu 215006,China
3 School of Computer Science and Engineering,Changshu Institute of Technology,Suzhou,Jiangsu 215500,China

Abstract　As a multi-modal task,the bottleneck of visual question answering(VQA) is to solve the problem of fusion between different modes. It requires not only a full understanding of vision and text in the image,but also the ability to align cross-modal representation. The introduction of the attention mechanism provides an effective path for multi-mode fusion. However,the previous methods usually calculate the extracted image features directly,ignoring the noise and incorrect information contained in the image features,and most of the methods are limited to the shallow interaction between modes,without considering the deep semantic information between modes. To solve this problem,a cross-modal information filtering network(CIFN) is proposed. Firstly,the feature of the problem is taken as the supervision signal,and the information filtering module is designed to filter the feature information of the image,so that it can better fit the representation of the problem. Then the image features and problem features are sent to the cross-modal interaction layer,and the intra-modal and inter-modal relationships are modeled respectively under the action of self-attention and guided attention,so as to obtain more fine-grained multi-modal features. Extensive experiments have been conducted on VQA2.0 data sets,and the experimental results show that the introduction of information filtering module effectively improves the model accuracy,and the overall accuracy of test-std reaches 71.51%,which has good performance compared with the most advanced methods.

到稿日期:2023-03-26　返修日期:2023-08-09
基金项目:国家自然科学基金(61972059,42071438);江苏省自然科学基金(BK20191474,BK20191475);吉林大学符号计算与知识工程教育部重点实验室(93K172021K01)
This work was supported by the National Natural Science Foundation of China(61972059,42071438),Natural Science Foundation of Jiangsu Province,China(BK20191474,BK20191475) and Key Laboratory of Symbolic Computing and Knowledge Engineering, Ministry of Education, Jilin University(93K172021K01).
通信作者:龚声蓉(shrgong@suda.edu.cn)

Keywords Visual question answering, Deep learning, Attention mechanism, Multi-modal fusion, Information filtering

1 引言

视觉问答即回答给定图像的问题，需要综合分析视觉特征和语言特征，是多模态融合领域的研究热点。随着深度学习的发展，诸多视觉语言任务被提出并取得了重大突破，如图像文本匹配[1-2]、图像字幕[3-5]和视觉问答[6-8]。相比其他视觉语言任务，VQA 更具挑战性，因为它需要在更精细语义层面理解图像和问题，并将分散在两个空间的特征映射到同一特征空间，在融合特征的基础上推测答案。

为了对齐视觉和文本两个不同模态的语义信息，早期的视觉问答框架采用直接融合的方式来实现答案分类[9]。Shih 等[10]将注意力机制引入 VQA，旨在根据问题学习图像的视觉注意。Ren 等[11]表明，仅学习视觉注意不能充分理解图像内容，为此提出同时学习视觉和文本注意以增强特征表达能力。

上述方法多是基于图像全局特征构建的模型，注意力多集中于模态间的粗交互，阻碍了对多模态特征细粒度的理解。为此，Anderson 等[7]提出自下而上地获取图像显著区域，并为每个区域生成相关特征向量来代替全局图像特征，从而使注意力得以在区域水平上计算。而后 Kim 等[12]提出了一种密集共同注意模型，以加强文本和视觉之间的密集交互。然而，此类方法未能充分考虑模态内的关系，针对这一问题，Yu 等[8]提出了深度模块化共同注意力网络，旨在通过自注意力单元建模密集的模态内交互，同时通过引导注意力单元实现跨模态交互。然而多数基于 Transformer[13]结构的模型，在建模模态内关系时，通常直接进行自注意力建模，未曾考虑特征中包含的噪声和不正确信息所带来的影响。

针对上述问题，本文提出了一种基于跨模态信息过滤的视觉问答网络(Cross-modal Information Filtering-based Networks for Visual Question Answering，CIFN)，即针对提取的两种模态特征，首先通过一个问题监督下的信息过滤模块来过滤图像特征，以去除不相关的信息；然后通过自注意力和引导注意力来建模模态内和模态间的关系，以提取图像和问题更为丰富的语义特征，最后将两种模态进行融合以实现答案预测。

综上所述，本文的主要贡献如下：

1)提出了一种新的基于跨模态信息过滤的视觉问答网络 CIFN，用于对齐不同模态的语义信息，以实现跨模态深度交互。

2)设计了一个问题监督下的信息过滤模块，该模块通过集中于与问题相关的区域，以过滤图像中的噪声，为后续建模多模态关系提供了有效的引导作用。

3)在 VQA2.0 数据集上的大量实验结果表明，CIFN 相比目前的先进模型有较大提升。

2 相关工作

2.1 视觉问答

近年来，多模态融合问题受到学术界广泛关注，其中以融合视觉和文本特征的 VQA 最具代表性。随着数据集和评价平台[6]的建立，各种 VQA 方法相继涌现。相比早期的直接特征融合，基于注意力机制的 VQA 方法已成为主流方式，通过聚焦于局部关键特征，使得模型得以有效地消除冗余信息。

Yang 等[14]提出了堆叠注意力网络模型(Stacked Attention Network，SAN)，该模型通过多次迭代，不断利用问题表征获取与答案相关的图像区域。Anderson 等[7]提出自下而上与自上而下相结合的注意力机制，通过对象属性检测器获取图像显著区域以代替全局特征，并通过自上而下的机制为目标区域生成注意力分布。

Lu 等[15]设计了一个共同注意框架，交替学习文本注意和视觉注意。Yu 等[16]提出了两阶段共同注意框架，即通过问题自注意和问题引导的图像注意来剔除不相关信息，以便同时关注图像、问题和答案的多模态注意[17-18]，可以定位更多对象的双线性方法[19]。相关研究[11,20-24]探索了共同注意机制，包括图像注意和文本注意，分别突出问题中关键词和图像区域。这些基于共同注意力的方法缺乏密集的模态交互，为改善这一问题，Yu 等[8]提出了基于密集的共同注意力方法。

2.2 注意力机制

近年来，自注意力机制已被广泛应用于各种视觉和语言任务。相比其他注意力机制，自注意力被训练为获取内部依赖关系。Vaswani 等[13]详细描述了一个基于自注意力构建的模型，并在神经机器翻译任务上显著优于基于 RNN 的模型。

门机制由来已久，Hochreiter 等[25]首次提出使用门机制来控制信息的输入和输出量，通过这种精巧的设计，门机制在机器翻译任务中取得了优异的成绩，而后这种思维被广泛应用于各个领域。在口语理解任务[26]中，门机制被用于将意图语义再现标记为槽标记。在视觉问答任务上，Rahman 等[27]通过添加注意力门，使得模型性能得以提升。

3 模型结构

VQA 任务的问题定义如下：给定一个基于图像 I 的问题 Q，目标是预测答案 a。在实践中通常将 VQA 任务定义为分类问题：

$$p(a|I,Q)=\emptyset(f(I,Q)),a\in A \tag{1}$$

其中，$f(\cdot)$为融合函数，$\emptyset(\cdot)$为分类器，a 为从候选答案集 A 中预测的答案。

本文提出的 CIFN 模型由 3 个部分组成，分别是特征

提取模块、跨模态信息过滤模块以及多模态融合与答案预测模块。模型的整体结构如图 1 所示。

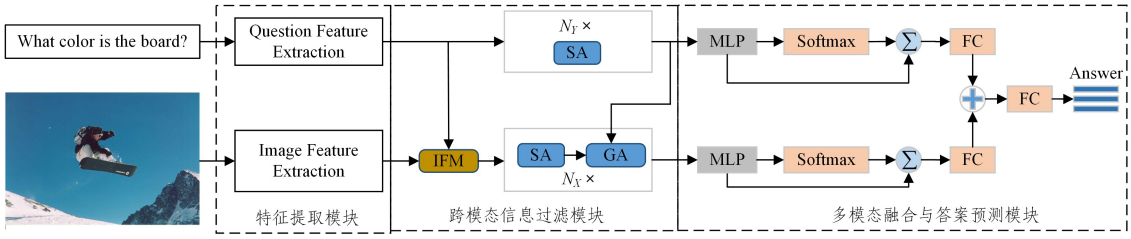

图 1　跨模态信息过滤网络的总体架构

Fig. 1　Overall architecture of the proposed cross-modal information filtering network

3.1　特征提取模块

视觉特征提取采用了基于区域的图像特征，在实践中本文采用在 Visual Genome[28] 预训练的 Faster R-Cnn 提取对象特征，其主干是在 ImageNet[29] 上预训练的 ResNet-101[30]。对于每幅图像，依据概率提取前 K 个对象区域，对不足的在模型中使用零向量进行末尾填充。并通过平均卷积池化生成图像特征输入 $X=\{x_1,x_2,\cdots,x_K\}\in R^{K*d_x}$，其中每个对象区域表示为 $x_i\in R^{d_x}$。

问题编码由词嵌入单元和 LSTM 单元组成。对于给定的问题 T，首先对其修剪，将单词数量固定为最长 M 个，摒弃多余单词，对于不足的在模型中则使用零向量进行末尾填充。使用 Glove[31] 进行词嵌入，将问题中单词编码为 300 维向量 $\boldsymbol{w}_m\in R^{300}$；随后将词嵌入序列 $W=\{w_m\in R^{300}\}_{m=1}^{M}$ 输入 LSTM 网络以获取问题文本表征 $Y\in R^{M\times d_y}$。

3.2　跨模态信息过滤模块

跨模态信息过滤模块由信息过滤模块(IFM)和跨模态交互层(CIL)组成。本节首先介绍 IFM 的定义及作用，随后介绍跨模态交互层的具体结构。

3.2.1　信息过滤模块

Nguyen 等[32] 指出，对于预训练检测器提取的图像特征可能含有噪声或不正确的信息这一问题，引入注意力可以有效解决。受此启发，本文设计了一个信息过滤模块(见图 2)，用于过滤图像特征信息使其更加契合问题表征。在实践中，该模块有利于帮助理解每个 ROI 和问题的关系。为了过滤图像特征，本文以问题为监督信息，因此信息过滤模块以图像和问题表征为输入，以优化后的图像特征作为输出。

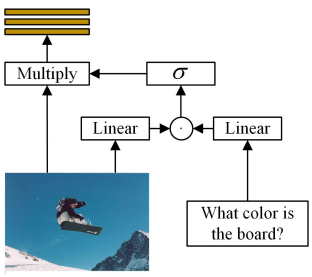

图 2　信息过滤模块示意图

Fig. 2　Illustration of information filtering module

图 2 给出了信息过滤模块的结构，对于给定的视觉特征 $X=\{x_1,x_2,\cdots,x_K\}$，本文选取 LSTM 最后隐层状态 $h_q\in R^{d_q}$ 来引导优化图像特征，为后续建模模态内关系起先导作用。其具体表达式如下：

$$X=P\cdot X \tag{2}$$

其中，$P=\{p_1,p_1,\cdots,p_K\}$ 为注意力权重系数，p_i 权重表示了图像中第 i 目标区域与其对应问题的相关性，权重系数的计算式如下：

$$p_i=\sigma(S^{\mathrm{T}}(W_y^{\mathrm{T}}h_y\circ W_x^{\mathrm{T}}x_i)) \tag{3}$$

其中，$S\in R^d$，$W_y\in R^{d_y\times d}$，$W_x\in R^{d_x\times d}$ 为可学习参数。

3.2.2　跨模态交互层

跨模态交互层包括编码器和解码器两个部分，其核心组件是自注意力(Self-Attention，SA)和引导(Guide-Attention，GA)。

在介绍 SA 和 GA 前，本文简要回顾自注意力机制。自注意力机制即将问题或图像表征通过线性变化映射为一组查询、键和值矩阵，三者共同构成了缩放点积注意的输入，为便于计算通常将维度统一为 d。通过计算查询与键的点积，然后除以 $\sqrt{d}$，并用 softmax 函数来生成注意力权重。因此，对于给定的 Q,K,V，注意力值是通过从 Q 和 K 学习到的权重对 V 的加权和。其表达式如下：

$$Att(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{4}$$

为了从不同角度来增强特征表达能力，文献[13]引入了多头注意力，多头注意力由 h 个平行头组成，其中每个头包括一个缩放点积函数。多头注意力输出如下：

$$f=MH(Q,K,V)=Contact(Att_1,Att_2,\cdots,Att_h)W \tag{5}$$

$$Att_i=Att(Q\,W_i^Q,K\,W_i^K,V\,W_i^V) \tag{6}$$

其中，$W_i^Q,W_i^K,W_i^V\in R^{d\times d_h}$，下标 i 表示第 i 个头，在合并 h 个头并通过一个线性映射函数 $W\in R^{h\times d_h\times d}$ 后，得到最终的注意力输出。

针对多模态问题，本文以多头注意力为基准构建了 SA 和 GA，如图 3 所示，SA 用于捕获模态内信息，GA 用于获取模态间的信息。其中编码器由若干堆叠的 SA 组成，SA 包含一个多头自注意力子层和一个前馈层，同时在两个子层上分别应用了残差连接和层归一化以防止模型过拟合。对于输入的问题，特征首先通过多头自注意力学习每对词 (y_i,y_j) 之间的相关性；然后通过前馈层来对多头注意力的输出进行 ReLu 和 Dropout 等。上级 SA 的输出将作为下级 SA 的输入，表达式如下：

$$f=MH(Y^l,Y^l,Y^l) \tag{7}$$

$$Y^{l+1}=FFN(f)=FC_2(\max(0,FC_1(f))) \tag{8}$$

其中，FC_1,FC_2 为带偏置的线性函数，$l\in L_Y$。

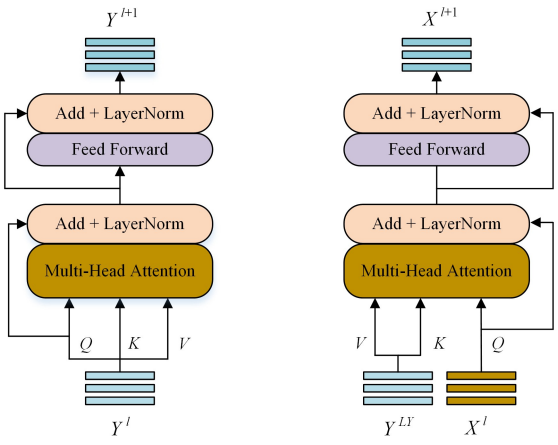

图 3 自注意(SA)和引导注意(GA)

Fig. 3 Illustration of SA and GA

解码器则采取了 SA+GA 的方式。首先解码器的 SA 对输入图像进行自注意力学习,以建模图像目标区域间的关系。

$$X^l = MH(X^l, X^l, X^l) \tag{9}$$

GA 内部结构与 SA 相同,不同的是输入来自视觉和语言两个模态。即以经过 SA 建模后的图像表征为查询矩阵 $\boldsymbol{Q}$,来自编码器 SA 最后一层输出 Y^{L_Y} 作为 K 和 V,以实现对图像中关键区域的注意。最后通过一个前馈层来处理 GA 输出。

$$f = MH(X^l, Y^{L_Y}, Y^{L_Y}) \tag{10}$$

$$X^{l+1} = FFN(f) = FC_2(\max(0, FC_1(f))) \tag{11}$$

其中,FC_1,FC_2 为带偏置的线性函数,$l \in L_X$。

最终通过多层迭代跨模态信息过滤模块输出图像表征 X^{L_X} 和问题表征 Y^{L_Y}。

3.3 多模态融合与答案预测

经过跨模态信息过滤模块后,模型得到了语义信息更加丰富的视觉特征 X^{L_X} 和文本特征 Y^{L_Y},需要采用一种策略将两者进行融合后来进行答案预测,如图 4 所示。

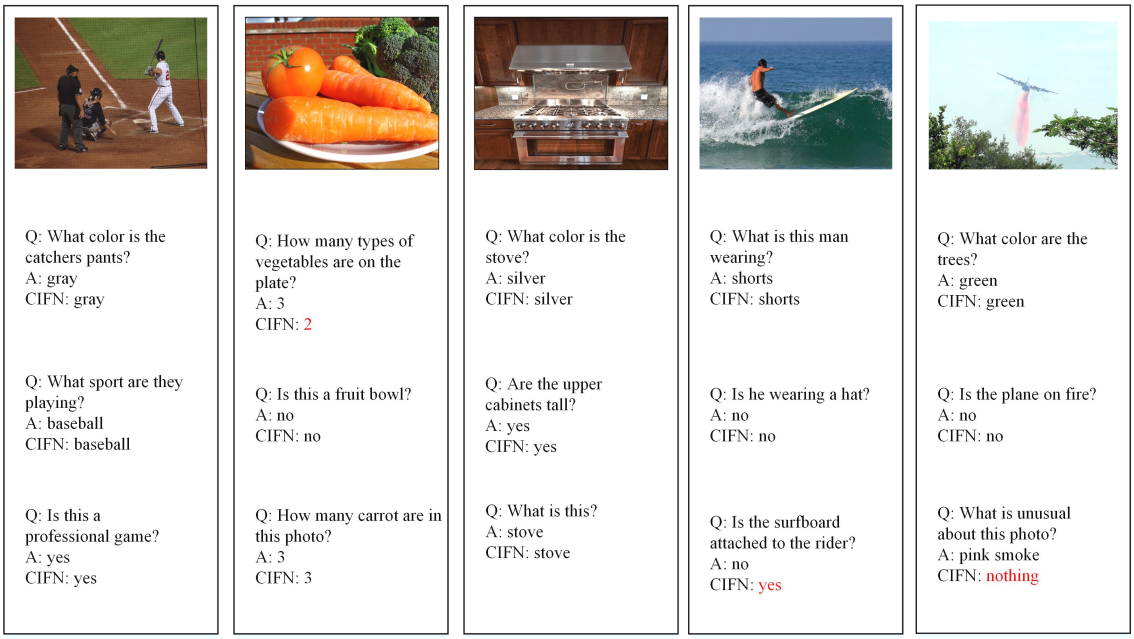

图 4 CIFN 在 VQA2.0 数据集上的预测结果(电子版为彩图)

Fig. 4 Prediction results on VQA2.0 dataset

首先将跨模态信息过滤模块输出的视觉特征 X^{L_X} 和文本特征 Y^{L_Y} 分别通过 MLP 网络,以生成具有单个头部分布的注意力分布,然后对权值进行累加求和,得到最终的视觉特征 $\hat{X}^{L_X}$ 和文本特征 $\hat{Y}^{L_Y}$。

$$\alpha = \mathrm{Softmax}(MLP(X^{L_X})), \hat{X}^{L_X} = \sum_{i=1}^{K} \alpha_i X_i^{L_X} \tag{12}$$

$$\beta = \mathrm{Softmax}(MLP(Y^{L_Y})), \hat{Y}^{L_Y} = \sum_{i=1}^{M} \beta_i Y_i^{L_Y} \tag{13}$$

其中,$\alpha \in R^K$,$\beta \in R^M$。随后通过两个线性函数,并进行元素相加来获得最终融合特征 S。

$$S = FC_X(\hat{X}^{L_X}) + FC_Y(\hat{Y}^{L_Y}) \tag{14}$$

其中,$FC_X \in R^{d \times d}$,$FC_Y \in R^{d \times d}$。

将得到的融合特征输入 LayerNorm 中进行稳定训练,最后通过一个全连接层进行答案预测。

$$Z = LayerNorm(S) \tag{15}$$

$$a = \mathrm{Sigmoid}(Linear(Z)) \in R^N \tag{16}$$

其中,$a \in R^N$,N 为候选答案集大小。

本文使用二元交叉熵损失(BCE)作为损失函数来进行训练,表达式如下:

$$loss = -[y \log \tilde{y} + (1-y)\log(1-\tilde{y})] \tag{17}$$

其中,y 是真实答案,$\tilde{y}$ 为模型预测值。

4 实验

4.1 数据集

本文中的训练阶段和测试阶段均在 VQA2.0 数据集[33]上进行,该数据集是最常用的 VQA 基准数据集,它包含来自 MS-COCO 数据集[34]中的图像相关的人类注释问答对,分为训练集(83×10^3 张图像和 444×10^3 个问答对)、验证集(41×10^3 张图像和 214×10^3 个问答对)、测试集(81×10^3 张图像和 448×10^3 个问答对),每张图像有 3 个问题,每个问题是由 10 位不同答案注释者分别给出 1 个答案。此外,对于每个

问题，有两个相似的图像，因此存在相异的答案，这可以减少数据集的偏差和不平衡。官方评测指标为 $Acc(ans)=\min\left\{\frac{num(humans\ give\ ans)}{3},1\right\}$，该指标在回答问题时对注释者间的差异具有鲁棒性。此外还有两个用于在线评估的测试集 test-dev 和 test-std，结果包括 yes/no、nums、other 和总体精度(all)。

4.2 实现细节

为了评估所提方法，CIFN 遵循了 MCAN[8]。其中对象检测器提取的目标区域数量 K、问题词数量 M 分别设置为 100 和 14；输入图像特征 d_v、问题特征 d_q、融合特征 d_S 分别为 2048，1024，1024。多头注意力的隐藏维度 d_h 设置为 1024，头数 h 设置为 8，每个头隐藏维度为 $d_h=d/h=128$。

对于 VQA2.0 数据集，本实验遵循文献[7]中的策略，选择训练集中出现次数超过 8 次的正确答案作为候选答案，因此候选答案集为 $N=3\,129$。训练中使用 Adam Solver[35]($\beta_1=0.9$，$\beta_2=0.98$)和二元交叉熵(BCE)来训练 CIFN，总共训练 13 轮，批处理大小为 64，基础学习率为 1×10^{-4}，dropout 设为 0.1，学习率衰减为 1/5。实验环境如表 1 所列。

表 1 实验环境

Table 1 Experimental environment

系统	Linux Ubuntu 16.04
硬件环境	CPU Intel Xeon Gold 6131 GPU Tesla P100 16GB
软件环境	pytorch 3.6 CUDA 8.0

4.3 消融研究

消融实验在 VQA2.0 上进行，实验结果如表 2 所列，主要就 IFM 和 CIL 层数进行探究。首先研究了 CIL 层数对模型性能的影响，实验中分别将层数设置为 2～8。从表中可以看出，不同的级联层数对模型性能有不同的影响。随着级联层数的加深，整体准确率持续提高，但超过 6 层后准确率开始下降，由此可知当 $L=6$ 时模型最优，因此，本文后续工作均建立在 $L=6$ 的基础上。如表 2 所列，当未引入信息过滤模块时模型性能下降，由此表明信息过滤模块具有较强的信息过滤能力。

表 2 在 VQA2.0 test-dev 上的消融实验

Table 2 Ablation studies on VQA 2.0 test-dev

Method		Test-dev			
		Y/N	Num	Other	All
CIL(层数)	2	86.95	53.22	61.19	70.91
	4	87.19	53.61	61.21	71.05
	6	87.43	53.82	61.42	71.27
	8	87.33	53.48	61.22	71.10
IFM	w/o	87.09	53.24	61.09	70.94

4.4 实验结果

为了证明本文方法的有效性，本文将其与之前的方法在 VQA2.0 上进行了比较。选取的是近年来有代表性的模型，BUTD[7]是 2017 VQA 挑战赛的冠军，首次提出使用自下而上的方式提取基于区域的图像特征，以代替此前基于网格的图像特征；BAN[12]提出了 8 头的双线性注意力，有效解决了计算复杂性的问题，超越了同期其他方法；MCAN[8]设计了自注意单元以建模模态内(词到词、区域到区域)关系、引导注意力单元以实现跨模态交互(词到区域)，优于此前最先进的技术；ViLT[36]是一个以 Transformer 编码器为核心的视觉语言预训练模型，通过对图像进行 Patch 映射，将图像处理简化为与文本处理相同的无卷积方式，因此图像和问题得以作为一个整体被送入 Transformer 编码器进行预训练；QD-GFN[37]通过问题监督下的 3 个图注意网络分别对图像的语义、空间和隐式关系进行建模，同时采用对象过滤机制以去除图像中包含的与问题无关的对象；MCAoAN[27]受 AoA 模型的影响，对多头注意力输出再次应用注意力，以此来增强原始注意力值与输入查询的关系。实验结果表明该方法相比基线方法具有更好的性能。

对比实验如表 3 所列，本文方法在 test-dev 和 test-std 上的总体准确率分别达到了 71.27%和 71.51%。通过分析表 2 可知，本文模型在 Y/N 及 other 指标上明显优于其他模型。

表 3 与 SOTA 模型在 VQA2.0 上的比较

Table 3 Comparison with SOTA on VQA2.0

Method	Test-dev				Test-std
	Y/N	Num	Other	All	All
BUTD	81.82	44.21	56.05	65.32	65.67
BAN	85.03	49.80	59.97	69.14	—
MCAN	86.82	53.26	60.72	70.63	70.90
QD-GFN	86.45	54.41	60.52	70.51	70.71
ViLT	—	—	—	70.85	—
MCAoAN	87.05	53.81	60.97	70.9	71.14
CIFN(ours)	**87.43**	53.82	**61.42**	**71.27**	**71.51**

4.5 定性分析

本文选取了 VQA2.0 数据集上的 5 幅图像问题对，使用本文方法分别对其进行预测，如图 4 所示，红色表示模型预测与实际不符。图 4 中第 2 例，模型在回答“How many types of vegetables are on the plate”时，给出了与注释不同的答案，导致这一错误的原因可能是，“tomato”身份界定不清晰，模型将其归类为“fruit”。第 5 例中，模型在回答“What is unusal about this photo”时，不能判断出此图有何不同，而给出了“nothing”的回答，表明模型对复杂问题的处理能力欠缺。模型在某些领域的错误并不能掩盖其整体性能，如在回答一些常见计数、颜色和是否问题上，模型表现出了较强的鲁棒性。

图 5 给出了 CIFN 的可视化结果，实验样例来自于 VQA2.0。鉴于篇幅原因，仅选取了 1 个样例，针对其展示了 5 幅子图，第一列为原图，第二列从上至下依次表示第 0 层和第 5 层问题自注意力；第三列从上至下依次表示为第 0 层和第 5 层问题引导的图像注意力，其中颜色深浅表示权重大小。

从图 5 中可以看出，SA(ques)-0 的注意力成垂直分布，“How”“this”等词具有显著的注意力权重，但随着多轮迭代，在 SA(ques)-5 中“bikinis”一词获得了较高的权重，由此表明，问题中的关键词“bikinis”被正确识别。GA(ques，img)-0 的关注多集中于“bikinis”，GA(ques，img)-5 则对“people”和“bikinis”都赋予了较高的权重，这表明模型不只是关注“bikinis”，同时关注与之相关联的“people”，这是回答问题的关键。综上，经过多层迭代交互，问题自注意力将权重集中于关键词，而问题引导的图像注意则聚焦于与问题相关的区域。

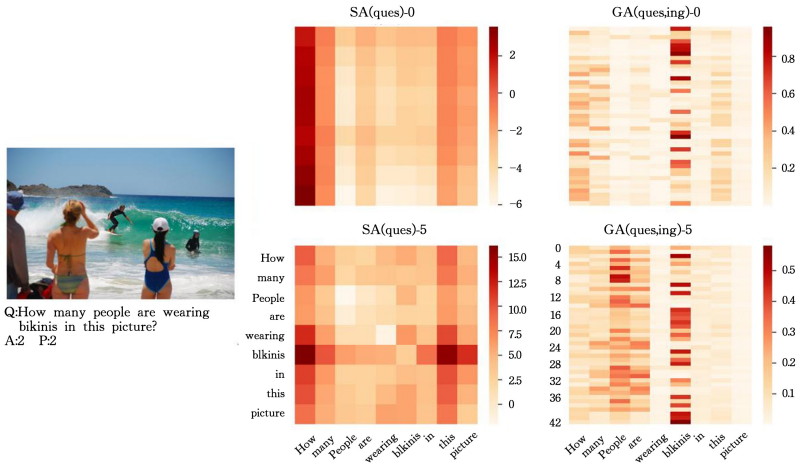

图 5 注意力单元学习到的注意力图的可视化

Fig. 5 Visualization of attention diagram learned by attention unit

结束语 本文设计了一个跨模态信息过滤网络(CIFN),其核心是一个信息过滤模块(IFM)和一个深度级联的跨模态共同注意力层(CIL),通过问题监督下的信息过滤模块,来实现对图像特征的优化,随后通过深度迭代的跨模态交互模块,建模密集的模态内和模态间交互。在 VQA2.0,通过与现有模型进行比较,大量实验结果表明,IFM 的引入能够有效过滤图像特征噪声,同时跨模态深度交互促进了图像和问题更细粒度的对齐,实现了两种模态的密集交互。

由于实验条件限制,未能就跨模态交互的不同变体和不同的多模态融合策略进行研究。后续将就这两个方面开展深入探究,逐渐改进模型结构;同时考虑采用更先进的文本编码器和其他数据集,以提高其泛化能力。

参 考 文 献

[1] YAN F, MIKOLAJCZYK K. Deep correlation for matching images and text[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 3441-3450.

[2] WANG Y, YANG H, QIAN X, et al. Position focused attention network for image-text matching[C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2019: 3792-3798.

[3] YOU Q, JIN H, WANG Z, et al. Image captioning with semantic attention[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 4651-4659.

[4] LI G, ZHU L, LIU P, et al. Entangled transf-ormer for image captioning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 8928-8937.

[5] NGUYEN K, TRIPATHI S, DU B, et al. In defense of scene graphs for image captioning[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 1407-1416.

[6] ANTOL S, AGRAWAL A, LU J, et al. Vqa: Visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2015: 2425-2433.

[7] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 6077-6086.

[8] YU Z, YU J, CUI Y, et al. Deep modular co-attention networks for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 6281-6290.

[9] MALINOWSKI M, ROHRBACH M, FRITZ M. Ask your neurons: A neural-based approach to answering questions about images[C]// Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE Press, 2015: 1-9.

[10] SHIH K J, SINGH S, HOIEM D. Where to look: Focus regions for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 4613-4621.

[11] REN S, HE K, GIRSHICK R, et al. Faster rcnn: Towards real-time object detection with region proposal networks[C]// Advances in NeuralInformation Processing Systems 28. Cambridge: MIT Press, 2015: 91-99.

[12] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[C]// Advances in Neural Information Processing Systems 31. Cambridge: MIT Press, 2018: 1571-1581.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems 30. Cambridge: MIT Press, 2017: 5998-6008.

[14] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 21-29.

[15] LU P, LI H, ZHANG W, et al. Co-attending freeform regions and detections with multi-modal multiplicative feature embedding for visual question answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018: 7218-7225.

[16] YU Z，YU J，FAN J，et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering [C]// Proceedings of the IEEE International Conference on Computer Vision. New York：IEEE Press，2017：1839-1848.

[17] ZHOU B，TIAN Y，SUKHBAATAR S，et al. Simple baseline for visual question answering[J]. arXiv，2015，1512. 02167.

[18] SCHWARTZ I，SCHWING A，HAZAN T. High-orderattention models for visual question answering[C]// Advances in Neural Information Processing Systems 30. Cambridge：MIT Press，2017：3664-3674.

[19] BENYOUNES H，CADENE R，CORD M，et al. Mutan：Multimodal tucker fusion for visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2612-2620.

[20] NAM H，HA J W，KIM J. Dual attention networks for multimodal reasoning and matching[C]// Proceedings of the IEEE Conferenceon Computer Vision and Pattern recognition. New York：IEEE Press，2017：299-307.

[21] NGUYEN D K，OKATANI T. Improved fusion ofvisual and language representations by densesymmetric coattention for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2018：6087-6096.

[22] FU P C，YANG G，LIU X M，et al. Visual Question Answering Model Based on Spatial Relation and Frequency Feature[J]. Computer Engineering，2022，48(9)：96-104.

[23] PENG L，YANG Y，BIN Y，et al. Word-to-region attention network for visual question answering[J]. Multimedia Tools and Applications，2019，78：3843-3858.

[24] GUAN W，WU Z，PING W. Question-oriented cross-modal coattention networks for visual question answering[C]// 2022 2nd International Conference on Consumer Electronics and Computer Engineering. New York：IEEE Press，2022：401-407.

[25] HOCHREITER S，SCHMIDHUBER J. Long short term memory[J]. Neural Computation，1997，9(8)：1735-1780.

[26] LI C，LI L，QI J. A self-attentive model with gate mechanism for spoken language understanding[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2018：3824-3833.

[27] RAHMAN T，CHOU S H，SIGAL L，et al. An improved attention for visual question answering [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2021：1653-1662.

[28] KRISHNA R，ZHU Y，GROTH O，et al. Visual g-enome：Connecting language and vision using crowd sourced dense image annotations[J]. International Journal of Computer Vision，2017，123：32-73.

[29] RUSSAKOVSKY O，DENG J，SU H，et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision，2015，115：211-252.

[30] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2016：770-778.

[31] PENNINGTON J，SOCHER R，MANNING C D. Glove：Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg：ACL，2014：1532-1543.

[32] NGUYEN B X，DO T，TRAN H，et al. Coarse-to-fine reasoning for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2022：4558-4566.

[33] 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]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Press，2017：6904-6913.

[34] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft coco：Common objects in context[C]// Computer Vision ECCV 2014：13th European Conference. Berlin：Springer，2014：740-755.

[35] KINGMA D P，BA J. Adam：A method for stochastic optimization[C]// 3rd International Conference on Learning Representations. Ithaca，2015.

[36] KIM W，SON B，KIM I. Vilt：Vision and language transformer without convolution or region supervision[C] // International Conference on Machine Learning. New York：ACM，2021：5583-5594.

[37] QIAN Y，HU Y，WANG R，et al. Question Driven Graph Fusion Network For Visual Question Answering[C]// 2022 IEEE International Conference on Multimedia and Expo. New York：IEEE Press，2022：1-6.

HE Shiyang，born in 1995，postgraduate. His main research interests include machine learning and computer vision.

GONG Shengrong，born in 1966，Ph.D，professor，Ph.D supervisor. His main research interests include image and video processing，pattern recognition and computer vision.

（责任编辑：喻藜）