文章编号：2095-4980(2021)01-0156-06

基于多模态双向导向注意的视觉问答

鲜　荣，何小海*，吴晓红，卿粼波

(四川大学 电子信息学院，四川 成都 610065)

摘　要：针对视觉问答(VQA)任务中现存深度协同注意模型只考虑问题引导图像的单向注意方式，导致多模态学习交互性不足的问题，提出一种多模态双向导向注意力网络。该网络由多模态特征提取模块、双向导向注意力模块、特征融合模块以及分类器组成。将提取出的图像和问题特征分别经过层层注意后输出加权的注意特征；经过特征线性融合后送入 softmax 分类器，得到问题的预测答案；再结合计数模块提升模型的计数能力。结果表明，该模型在公共数据集 VQA v2.0 上表现良好，在 test_dev 和 test_std 测试子集上分别获得 70.77%、71.28%的总体分类准确率，与大多数先进模型相比，体现出一定优势。

关键词：视觉问答；深度协同注意；单向注意；双向导向注意；特征融合

中图分类号：TP391.41　　**文献标志码：**A　　**doi：**10.11805/TKYDA2020172

Visual Question Answering based on multimodal bidirectional guided attention

XIAN Rong，HE Xiaohai*，WU Xiaohong，QING Linbo

(School of Electronic Information，Sichuan University，Chengdu Sichuan 610065，China)

Abstract：Aiming at the problem that the existing deep collaborative attention models in the Visual Question Answering(VQA) task only consider the unidirectional attention of the question-guided image, which leads to the lack of interactivity of multimodal learning, a multimodal bidirectional guided attention network is proposed. The network consists of multimodal feature extraction module, bidirectional guided attention module, feature fusion module and classifier. The extracted image and question features are respectively output with weighted attention features after passing through layers of attention, and then the features are linearly merged into the softmax classifier to obtain the predicted answer to the question. Finally, the counting module is combined to improve the counting ability of the model. The results show that the model performs well on the public data set VQA v2.0, and obtains an overall classification accuracy of 70.77% and 71.28% on the test_dev and test_std, respectively, showing certain advantages compared with most advanced models.

Keywords：Visual Question Answering；deep collaborative attention；unidirectional attention；bidirectional guided attention；feature fusion

视觉问答(VQA)[1]旨在给出一幅图像，自动回答出与图像相关问题的答案。该任务涉及到视觉和文本两种模态的学习，架起了计算机视觉和自然语言处理领域间的桥梁。许多基于视觉—语言的多模态任务已取得显著进展，如：图像文本匹配[2]、视觉描述[3-4]以及视觉问答。然而，视觉问答较其他多模态任务更具挑战性，不仅需要细致理解图像和问题的语义，更需要结合视觉推理来得到一个正确的答案。最简单直接的方法是通过提取图像和问题的全局特征，再经过简单的特征融合，分类后产生一个预测答案。这种方法虽易操作，但丢失了重要的局部信息，不利于回答针对局部区域提出的问题。为获得更细节性的特征，VQA 引入了注意力机制，很大程度上提升了多模态任务的性能表现。

收稿日期：2020-04-24；**修回日期：**2020-06-04

基金项目：国家自然科学基金资助项目(61871278)；成都市产业集群协同创新资助项目(2016-XT00-00015-GX)；四川省科技计划资助项目(2018HH0143)

作者简介：鲜　荣(1995–)，女，在读硕士研究生，主要研究方向为计算机视觉。email:874293694@qq.com

*通信作者：何小海　email:hxh@scu.edu.cn

SHIH 等[5]提出从输入问题中学习图像区域的视觉注意力;PENG 等[6]提出一种根据问题中关键单词定位图像区域的注意力学习模型，消除问题中一些无关紧要的冗余信息；ANDERSON 等[3]提出一种自底向上—自顶向下的注意力机制，不再通过空间网格学习视觉注意，而是直接学习候选对象的注意力。除视觉注意外，学习文本注意问题中的关键词也很重要。为同时学习视觉和文本注意力，LU 等[7]提出协同注意力，学习图像注意和问题注意。但基于协同注意的大多数模型[8]只学习到多模态间的粗糙交互，并不能进一步推断出每个图像区域与每个问题单词间的关系。为克服基于协同注意的多模态交互不足性，密集协同注意力模型被提出，如双线性注意力网络(Bilinear Attention Networks，BAN)[9]，此类模型可以深入级联，支持更复杂的视觉推理，但较之于相应的浅层模型或粗糙交互的协同注意力模型，性能并没有明显地提升。YU 等[10]提出一种深度模块化协同注意网络(Modular Co-Attention Network，MCAN)，在以往深度协同注意模型基础上，同时在每个模态内构建密集自注意模型，学习图像中区域与区域之间、问题中单词与单词之间的关系，进一步丰富了图像和问题的特征表示。

上述研究表明，充分学习模态内的自相关性与多模态间的交互性对于 VQA 的发展至关重要。人们习惯根据问题中的关键词去聚焦图像中某具体区域，因此在构建多模态交互注意模型的工作中，大多数只考虑了通过问题特征构建图像特征的注意引导方式。但有时候人们视线也会首先关注图像内容并获取主要信息，再根据此信息过滤掉问题中一些不相关的单词信息，因此通过图像特征构建问题特征的注意力导向同样需要。为此，本文拟在深度模块化协同注意网络中设置的注意力单元基础上，提出图像引导问题注意模块，并联合协同注意提出双向导向注意力。最后提出网络结合 Counter 模块[11]，以期提高视觉问答模型对每类问题的回答正确率。

1　相关基础网络

1.1　注意力机制

注意力机制最初用于视觉图像领域，近几年，被广泛用于基于深度学习的自然语言处理领域。2017 年，在机器翻译任务中，VASWANI 等[12]提出一种 transformer 结构，采用放缩点积自注意力机制，将文本特征描述为 3 个矩阵，分别为查询(Query)和键值对(Key—Value)，经线性变换后得到对应的特征矩阵 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$，并作为注意力层的输入，计算文本之间的关系以及学习文本特征表示。同时，为保证模型在不同的表示子空间中学习到文本相关信息，将多个放缩点积注意力层进行拼接，进而提出多头自注意力机制。

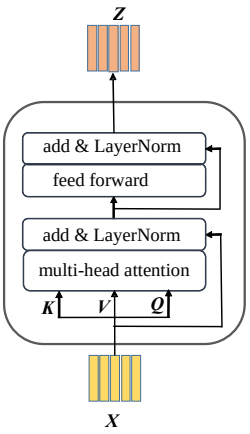

Fig.1 Self-Attention unit
图 1　自注意力单元

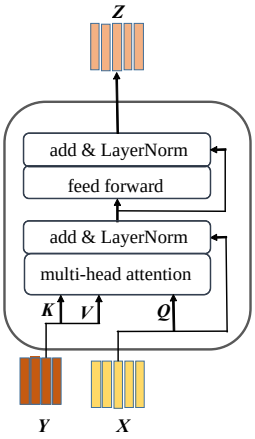

Fig.2 Guided-Attention unit
图 2　导向注意力单元

受此启发，YU 等[10]将多头注意力机制引入视觉问答领域。为满足多模态间的学习，通过改变多头注意力的输入，并设置自注意力(Self-Attention，SA)单元和导向注意力(Guided-Attention，GA)单元作为网络模型的基础。

SA 学习同一模态内各样本间的关系，结构如图 1 所示，由一个多头注意力层和逐点前馈层组成。多头注意力层的 3 个输入取同一模态特征矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1;\boldsymbol{x}_2;\cdots;\boldsymbol{x}_m]\in\mathbf{R}^{m\times d_x}$，对应关系为：$\boldsymbol{Q}\rightarrow\boldsymbol{x}_i,\boldsymbol{K}\rightarrow\boldsymbol{X},\boldsymbol{V}\rightarrow\boldsymbol{X}$。其中，$\boldsymbol{x}_i\in\boldsymbol{X}$。学习 $\boldsymbol{X}$ 矩阵中成对样本 $\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$ 之间的两两关系，并与 $\boldsymbol{X}$ 中所有向量进行加权求和，得到多头自注意力层的输出特征矩阵，再将得到的特征矩阵作为逐点前馈层的输入。其中，逐点前馈层通过 2 个全连接层(Fully Connected Layers，FC)、ReLU 激活层和 dropout 层实现。为方便优化模型，多头注意力层和逐点前馈层输出分别经过层归一化操作后通过残差连接。GA 以一种模态特征为导向，学习另一种模态更细致的特征表示，与 SA 结构相似，唯一不同在于多头注意力层的输入由同一种模态特征 $\boldsymbol{X}$ 变为 2 种不同模态特征 $\boldsymbol{X},\boldsymbol{Y}$，其中 $\boldsymbol{X}\in\mathbf{R}^{m\times d_x}$，$\boldsymbol{Y}=[\boldsymbol{y}_1;\boldsymbol{y}_2;\cdots;\boldsymbol{y}_n]\in\mathbf{R}^{n\times d_y}$，如图 2 所示。通过注意力学习 $\boldsymbol{X},\boldsymbol{Y}$ 中两两向量间的关系，以 $\boldsymbol{Y}$ 为导向学习构建 $\boldsymbol{X}$ 的特征语义，此时多头注意力输入对应关系为：$\boldsymbol{Q}\rightarrow\boldsymbol{x}_i,\boldsymbol{K}\rightarrow\boldsymbol{Y},\boldsymbol{V}\rightarrow\boldsymbol{Y}$，其中 $\boldsymbol{x}_i\in\boldsymbol{X}$。

1.2　协同注意模块

MCAN[10]中提出的协同注意模块由 L 个 MCA 层通过编解码方式架构组成，每个 MCA 层组成采用 SA—SGA 两条支路结构，如图 3 所示。协同注意模块的输入即第一个 MCA 层的输入。其中，SA 支路输入为问题中每个单词转换为词向量后再通过长短期记忆(Long Short-Term Memory，LSTM)网络得到的问题特征，SGA 支路由独立的 SA 和 GA 栈式连接，输入为经 Faster R-CNN 提取出的图像特征。对于 MCA$^{(1)}$，设置输入特征 $\boldsymbol{f}_Y^{(0)}=\boldsymbol{Y},\boldsymbol{f}_X^{(0)}=\boldsymbol{X}$。

如图 3 所示，协同注意模块的编码端由 L 个 SA 通过 stacking 方式级联而成，解码端由 L 个 SGA 层 stacking 级联组成，模块输出同样分为两路，分别为编码端经层层自注意输出的问题特征和解码端经自注意加导向注意输出的图像特征，记为 $\boldsymbol{f}_Y^{(L)},\boldsymbol{f}_X^{(L)}$。其中，编码端的输出作为解码端所有 GA 单元的输入之一参与图像特征的表示，从而实现以问题特征为导向，重新学习构建图像特征。

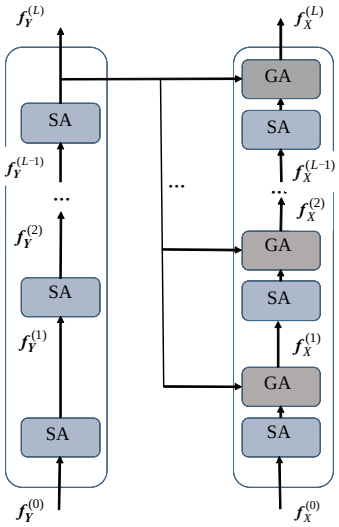

Fig.3 Collaborative attention module
图 3 协同注意模块

2 多模态双向导向注意力网络

2.1 双向导向注意力模块

MCAN 弥补了浅层协同注意力网络及粗糙交互的多模态学习网络其交互性不足的缺点，但同其他深度协同注意力网络一样，MCAN 仅考虑了以问题为导向，去关注图像的重要区域信息。本文提出以图像为导向，学习构建问题注意特征的导向注意力方式，在 MCAN 中提出的导向注意力单元的基础上，提出图像引导问题注意模块，其组成结构如图 4 中虚线框所示。以 GA 单元为基础，采用 stacking 级联方式，每一个 GA 单元代表独立的一层，经过 M 层的深度级联。如图 4 所示，$\boldsymbol{f}_Y^{(L)},\boldsymbol{F}_X$ 作为图像引导问题注意模块的输入，以 $\boldsymbol{F}_X$ 为注意力导向，重新学习构建问题特征，将模块的输出记为 $\boldsymbol{f}_{Y|X}^{(M)}$，用公式表示为：

$$\begin{cases}\boldsymbol{f}_{Y|X}^{(M)}=GA^{(M)}\left(\boldsymbol{y}_i,\boldsymbol{F}_X,\boldsymbol{F}_X\right)\\ \boldsymbol{f}_Y^{(L)}=SA^{(L)}\left(\boldsymbol{y}_i^{(0)},\boldsymbol{f}_Y^{(0)},\boldsymbol{f}_Y^{(0)}\right)\end{cases}\tag{1}$$

式中：$\boldsymbol{y}_i\in\boldsymbol{f}_Y^{(L)}$；$\boldsymbol{y}_i^{(0)}\in\boldsymbol{f}_Y^{(0)}$。$\boldsymbol{f}_Y^{(L)},\boldsymbol{F}_X$ 分别对应图 3 中协同注意模块的 2 个支路输出 $\boldsymbol{f}_Y^{(L)},\boldsymbol{f}_X^{(L)}$。

在 MCAN 中协同注意模块的基础上结合图像引导问题注意模块，构建双向导向注意力模块(Bidirectional Guided Attention Module，BGAM)。BGAM 包括问题引导图像注意和图像引导问题注意的双向导向注意，输入可分解成两条支路，分别为图像经目标检测网络提取出的视觉特征 $\boldsymbol{f}_X^{(0)}$，问题中每个单词处理成词向量经单层 LSTM 获得的问题特征 $\boldsymbol{f}_Y^{(0)}$。经过层层自注意和导向注意后，分别输出加权后的图像注意特征 $\boldsymbol{F}_X$ 和问题注意特征 $\boldsymbol{F}_Y$，可表示为：

$$\begin{cases}\boldsymbol{F}_X=SGA^{(L)}\left(SA^{(L)}\left(\boldsymbol{y}_i,\boldsymbol{f}_Y^{(0)},\boldsymbol{f}_Y^{(0)}\right),\boldsymbol{f}_X^{(0)}\right)\\ \boldsymbol{F}_Y=\boldsymbol{f}_Y^{(L)}\oplus\boldsymbol{f}_{Y|M}^{(M)}\end{cases}\tag{2}$$

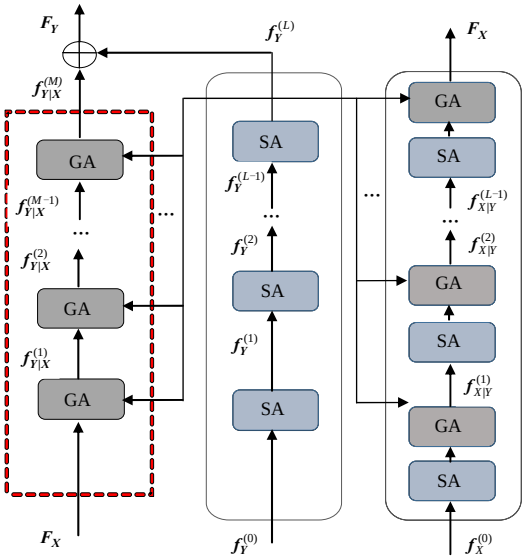

Fig.4 Bidirectional guided attention module
图 4 双向导向注意力模块

2.2 多模态双向导向注意力网络结构

基于 BGAM，本文提出一种多模态双向导向注意力网络(Multimodal BGMA，MBGAN)用于视觉问答任务。与其他深度级联的密集协同注意网络不同，本文扩展了导向注意机制的应用，丰富了问题特征的表示，最终输出的问题特征包含更多、更细致的语义信息，加强了问题与图像模态间的交互性，提高了预测答案的准确性。MBGAN 包括多模态特征提取模块、BGAM、多模态特征融合模块及分类器，其结构如图 5 所示。整个网络以图像及对应的问题—答案对作为输入。近年来视觉问答被视为多分类任务，与传统图像分类[13]不同，视觉问答给出问题类型中涉及目标计数问题，此时一般分类特征提取器不再适用。本文采用 ResNet-101 为基础骨架的 Faster R-CNN 模型，提取并输出图像中 m 个区域的中间特征。问题中先将每个单词转换为对应的词向量，再通过单层的 LSTM 得到问题中间特征。

多模态特征提取模块输出的图像和问题中间特征作为双向导向注意力模块的输入，经过层层自注意和导向注意输出加权的视觉注意特征和文本注意特征。

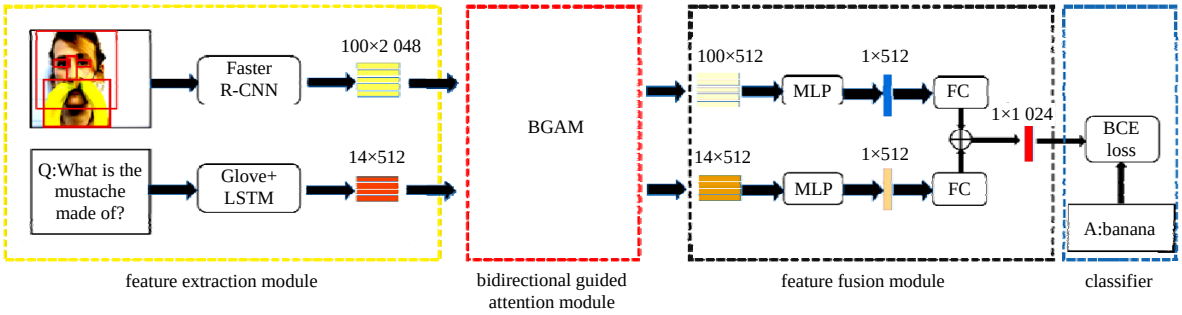

Fig.5 Structure of multimodal bidirectional guided attention network

图 5　多模态双向导向注意力网络结构

采用合理且高效的多模态特征融合方式是实现正确分类的关键。双向导向注意力模块输出的图像和问题特征分别记为 $\boldsymbol{F_X},\boldsymbol{F_Y}$，包含了丰富的图像区域以及问题单词的权重信息。为避免计算量过大，特征融合前利用文献[10]提出的 MLP 模型(FC-ReLU-Dropout-FC)对 $\boldsymbol{F_X},\boldsymbol{F_Y}$ 进行简化运算，记简化后的图像和问题特征分别为 $\boldsymbol{f}_x,\boldsymbol{f}_y$，两种模态特征简化运算过程一样。以 $\boldsymbol{f}_x$ 为例，可表示为：

$$\boldsymbol{f}_x=\sum_{i=1}^{m}\boldsymbol{\alpha}_i\boldsymbol{x}_i \tag{3}$$

$$\boldsymbol{\alpha}=soft\max\left[MLP(\boldsymbol{F}_X)\right] \tag{4}$$

式(4)中，$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_m]\in\mathbf{R}^m$，表示 m 个区域学习到的注意力权重。简化后的 $\boldsymbol{f}_x,\boldsymbol{f}_y$ 值通过一个线性函数进行特征融合，融合后的特征记为 $\boldsymbol{f}_z$，表示为：

$$\boldsymbol{f}_z=LayerNorm\left(\boldsymbol{W}_x^{\mathrm{T}}\boldsymbol{f}_x+\boldsymbol{W}_y^{\mathrm{T}}\boldsymbol{f}_y\right) \tag{5}$$

式中 $\boldsymbol{W}_x^{\mathrm{T}},\boldsymbol{W}_y^{\mathrm{T}}\in\mathbf{R}^{d\times d_z}$ 为 2 个线性映射矩阵，d_z 表示融合后的特征 $\boldsymbol{f}_z$ 的维度，LayerNorm 层起到稳定训练的作用。

取参与训练的所有答案中出现频率超过 8 次以上的前 N 个答案作为预测答案选项，用作分类中的类别标签。训练中输入的答案可视为真实标签。基于 $\boldsymbol{f}_z$，使用二进制交叉熵损失函数训练一个 N 路分类器，表示为：

$$L(o,t)=-\frac{1}{n}\sum_i\left[t_i\log o_i+(1-t_i)\log(1-o_i)\right] \tag{6}$$

式中：o 为输出的答案；t 为真实标签；n 为输入的所有问题数量。

3　Counter 模块

目前大多数 VQA 模型都是通过注意力机制获得加权后的特征向量，这些模型常用 softmax 函数对注意力权重的和归一化，但这对于回答计数问题存在不足性。如 2 个完全相同且重叠的物体通过目标检测网络后，检测出 2 个对象，并提取出 2 个相同的特征向量，注意力机制按等比例分配给每个对象 0.5 的权重，经权重求和归一化后，这 2 个对象便被平均为一个，此时已丢失了注意力图谱上所有有关计数可能的信息，任何采取权值归一化的模型都存在这个问题。并且有相关实验表明[14]，使用 sigmoid 函数进行权重归一化，不仅对计数问题没有帮助，反而降低非数字问题的准确率。因此，针对计数问题，ZHANG 等[11]提出一个可微分的神经网络组件，简称 Counter 模块，不仅可以有效避免注意力机制的基本限制，且表现出强大的计数功能。同时，该模块还可以解决同一对象重复计数的问题，其原理如图 6 所示。

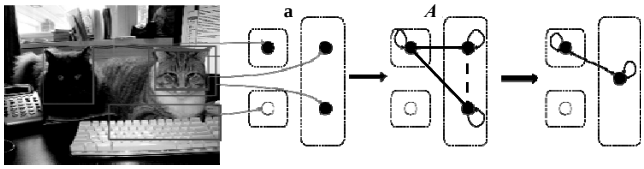

Fig.6 Schematic diagram of counter module

图 6　计数模块工作原理图

为解决同一对象重复计数问题，取经过注意力机制后得到的注意力图谱中的前 n 个最大权值向量 $\boldsymbol{a}=[a_1,a_1,\cdots,a_n]^{\mathrm{T}}$ 及所对应的目标标注框 $\boldsymbol{b}=[b_1,b_1,\cdots,b_n]^{\mathrm{T}}$ 作为该模块的输入，如图 6 所示，每个权值对应一个顶点，表示一个目标建议。橘色猫被检测 2 次，分配有 2 个权值，对应 2 个重复的目标建议。所有权值表示成相应的点图，黑色表示与计数问题相关的目标建议，白色代表不相关。对 $\boldsymbol{a}$ 进行外积操作得到一个邻接矩阵 $\boldsymbol{A}=\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$，表示成一个加权有向图。实线边连接相关目标建议，虚线边连接重复目标建议。使用微分操作删除和缩放重复目标建议内部的连接，及相关建议与重复目标建议之间多余的连接，最终估算出待统计目标的数量，并转换成对应的特征向量，作为补充信息融合到本文模型中。

4 实验结果及分析

4.1 数据集

VQA v2.0 是 VQA 任务最常用的基准数据集，包含 MS-COCO 数据集中的图像及相对应的人工标注的问答对。每幅图像对应 3 个问题，每个问题对应 10 个答案。数据集划分方式为：训练集，含 80 k 图像和 444 k 问答对；验证集，含 40 k 图像和 214 k 问答对；测试集，含 80 k 图像和 448 k 问答对。其中，测试集包含 2 个测试子集 test-dev 和 test-std。数据集中包括 3 种问题类型，分别为回答是或不是的问题(Yes/No)、目标计数问题(Num)和其他类问题(other)。

4.2 实验环境及参数设置

采用 Ubuntu16.04 操作系统，显卡为 NVIDIA Geforce GTX 1080 TI，显存 11 GB，CUDA10.0，并使用 cuDNN7.4 进行加速。整个实验基于 PyTorch 深度学习框架开展，采用 python 语言进行编程。

为与基准模型进行公平比较，本文同样进行了训练集增强，除 VQA v2.0 训练集外，还将其验证集以及来自 Visual Genome 中 VQA 样本子集一同进行训练。图像经过 Faster R-CNN 网络以 bottom-up 的方式提取一系列代表整个图像的目标区域特征，目标区域数量记为 $m\in[10,100]$。将问题先处理成单词，并最多截取 14 个单词，记每个问题句中单词数为 $n\in[1,14]$，使用 300-D GloVe 词嵌将每个单词转换为向量，通过单层 LSTM 网络获得问题特征。使用零填充方式固定 m,n 变量值，分别取值 100 和 14。实验中超参数设置为：输入图像特征维度 2 048，输入问题特征维度 512 以及融合后特征维度 1 024。多头注意力维度 d 设为 512，头数 h 取值 8，每个放缩点积维度 d_h 取值 64。图像引导问题注意模块级联层数 $M\in\{1,2,3,4,5,6\}$。双向导向注意模块中 L 取为 6。根据经验值，分类器中 N 取值 3 129，基础学习率设为 $\min(2.5te^{-5},e^{-4})$，所有训练样本共循环训练 17 次，t 从 1 开始。每次训练 batchsize 设为 64，验证和测试时 batchsize 设为 32。训练第 10 次以后，学习率分别在第 12,15,16,17 次时以 0.2 的速率衰减，在第 14 次时以 0.8 的速率衰减。采用 Adam 优化器，其中一阶矩衰减系数 $\beta_1=0.9$，二阶矩衰减系数 $\beta_2=0.98$。

4.3 实验结果及分析

为验证图像引导问题注意模块的有效性，本文选取不同的 M 值开展多次实验，不同取值在 test-dev 和 test-std 测试子集上总体准确率 all 的表现如表 1 所示。

表 1 不同 M 取值在测试子集上的总体准确率

Table1 Overall accuracy of different M values on the test subset

M	test-dev	test-std
	all/%	all/%
1	70.62	71.00
2	70.63	71.08
3	70.65	71.06
4	**70.76**	**71.09**
5	70.72	71.05
6	70.69	71.03

从表 1 可知，级联层数与准确率并不是呈线性相关，M 取值为 4 时，模型在 test-dev 和 test-std 两个测试子集上的总体准确率比其他取值情况都高，且高于大多数先进模型的准确率，验证了 M 取值为 4 的合理性。

为充分验证本文所提网络的有效性，选择多个近年来的先进网络模型，比较在 test-dev 和 test-std 两种测试子集上各类问题回答准确率，如表 2 所示。从表 2 可以看到，本文提出的 MBGAN 在 test-dev 测试子集上对 Y/N 指标较基准网络 MCAN 提高 0.4%，test-std 测试子集上总体准确率也有所提高，验证了该网络对于回答某些类问题有一定正向作用。同时可以看出，MBGAN 整体表现优于其他网络模型。

表 2 不同 VQA 模型在 test-dev 和 test-std 上各类问题回答准确率的对比

Table2 Comparison of different VQA models' answer accuracy on test-dev and test-std

network	test-dev				test-std			
	all/%	Y/N/%	Num/%	other/%	all/%	Y/N/%	Num/%	other/%
bottom-up[14]	65.32	81.82	44.21	56.05	65.67	82.20	43.90	56.26
counter[11]	68.09	83.14	51.62	58.97	68.41	83.56	51.39	59.11
BAN+Glove[9]	69.66	85.46	50.66	60.50	–	–	–	–
BAN+Glove+Counter[9]	70.04	85.42	**54.04**	60.52	70.35	–	–	–
MLIN[15]	70.18	85.96	52.93	60.40	70.28	–	–	–
DFAF[16]	70.22	86.09	53.32	60.49	70.34	–	–	–
MCAN[10]	70.63	86.82	53.26	60.72	70.90	–	–	–
MBGAN	70.76	**87.22**	53.17	60.68	71.09	87.32	53.1	61.09
MBGAN+Counter	**70.77**	87.03	53.63	**60.75**	**71.28**	**87.39**	**53.59**	**61.30**

最后验证结合 Counter 模块对问题回答准确率的影响，将本文所提出的 MBGAN 结合 Counter 模块后进行训

练，并与 MBGAN 及其他网络模型比较在两种测试子集上的表现。从表 2 可以看出，MBGAN 结合 Counter 模块后在 test-std 测试子集上每类问题回答准确率都有一定的提高，特别是针对 Num 类问题。对测试子集 test-dev 中 Num 类问题同样有所提高。从而证明了结合该模块对回答计数问题的有效性。同时，可以发现结合该模块后，两种测试子集总体回答准确率较单独的 MBGAN 有所提高，验证了 Counter 模块在提高计数问题回答准确率的同时，并不会限制注意力发挥其优势。最终，本文提出的方法在 test-dev 测试子集上获得 70.77%的准确率，test-std 测试子集上获得 71.28%的准确率，高于目前大多数 VQA 先进网络，验证了模型的有效性。

5　结论

本文针对 VQA 任务，提出一种多模态双向导向注意力网络。其中，提出图像引导问题注意模块，加强了视觉和文本两种模态间的充分交互，弥补了问题引导图像注意的单向交互的不足，更有利于理解图像和问题之间的语义相关性。最后结合 Counter 模块，提升了网络的计数功能。在数据集 VQA v2.0 的两种测试子集上获得不错的表现，且高于大多数模型，体现出其具有一定的优势。

参考文献：

[1] ANTOL S,AGRAWAL A,LU J,et al. VQA:Visual Question Answering[J]. International Journal of Computer Vision, 2017, 123(1):4–31.

[2] WANG L,LI Y,HUANG J,et al. Learning two-branch neural networks for image–text matching tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,41(2):394–407.

[3] ANDERSON P,HE X,BUEHLER C,et al. Bottom–up and top–down attention for image captioning and visual question answering[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,USA:IEEE, 2018: 6077–6086.

[4] WU Q,SHEN C,WANG P,et al. Image captioning and visual question answering based on attributes and external knowledge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018,40(6):1367–1381.

[5] SHIH K J,SINGH S,HOIEM D,et al. Where to look:focus regions for visual question answering[C]// Computer Vision and Pattern Recognition. Las Vegas,US:IEEE, 2016:4613–4621.

[6] PENG L,YANG Y,BIN Y,et al. Word–to–region attention network for visual question answering[J]. Multimedia Tools and Applications, 2019,78(3):3843–3858.

[7] LU J,YANG J,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,Spain:Curran Associates, 2016: 289–297.

[8] YANG C,JIANG M,JIANG B,et al. Co-attention network with question type for visual question answering[J]. IEEE Access, 2019(7):40771–40781.

[9] KIM J,JUN J,ZHANG B,et al. Bilinear attention networks[C]// Neural Information Processing Systems. Montreal,Canada: Curran Associates, 2018:1564–1574.

[10] YU Z,YU J,CUI Y,et al. Deep modular co-attention networks for visual question answering[C]// Computer Vision and Pattern Recognition. Long Beach,CA,US:IEEE, 2019:6281–6290.

[11] ZHANG Y,HARE J,PRUGEL–BENNETT A,et al. Learning to count objects in natural images for visual question answering[C]// International Conference on Learning Representations. Vancouver,Canada:[s.n.], 2018.

[12] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach,USA:Curran Associates, 2017:5998–6008.

[13] 杨晓敏,严斌宇,李康丽,等. 一种基于词袋模型的图像分类方法[J]. 太赫兹科学与电子信息学报, 2014,12(5):726–730. (YANG Xiaomin,YAN Binyu,LI Kangli,et al. An image classification method based on BoW model[J]. Journal of Terahertz Science and Electronic Information Technology, 2014,12(5):726–730.)

[14] TENEY D,ANDERSON P,HE X,et al. Tips and tricks for visual question answering:learnings from the 2017 challenge[C]// IEEE/CVF Computer Vision and Pattern Recognition. Salt Lake City,USA:IEEE, 2018:4223–4232.

[15] GAO P,YOU H,ZHANG Z,et al. Multi-modality latent interaction network for visual question answering[C]// International Conference on Computer Vision. Seoul,Korea(South):IEEE, 2019:5825–5835.

[16] GAO P,JIANG Z,YOU H,et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]// Computer Vision and Pattern Recognition. Long Beach,CA,USA:IEEE, 2019:6639–6648.