计算机科学
COMPUTER SCIENCE

一种基于多模态深度特征融合的视觉问答模型

邹芸竹, 杜圣东, 滕飞, 李天瑞

引用本文

邹芸竹, 杜圣东, 滕飞, 李天瑞.一种基于多模态深度特征融合的视觉问答模型[J]. 计算机科学, 2023, 50(2): 123-129.

ZOU Yunzhu, DU Shengdong, TENG Fei, LI Tianrui. Visual Question Answering Model Based on Multi-modal Deep Feature Fusion [J]. Computer Science, 2023, 50(2): 123-129.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200303

一种基于多模态深度特征融合的视觉问答模型

邹芸竹[1]　杜圣东[1,2]　滕　飞[1]　李天瑞[1,2]
1 西南交通大学计算机与人工智能学院　成都 611756
2 综合交通大数据应用技术国家工程实验室　成都 611756
(zyz590@my.swjtu.edu.cn)

摘　要　大数据时代,随着多源异构数据的爆炸式增长,多模态数据融合问题备受研究者的关注,其中视觉问答因需要图文协同处理而成为当前多模态数据融合研究的热点。视觉问答任务主要是对图像和文本两类模态数据进行特征关联与融合表示,最后进行推理学习给出结论。传统的视觉问答模型在特征融合时容易缺失模态关键信息,且大多数方法停留在数据之间浅层的特征关联表示学习,较少考虑深层的语义特征融合。针对上述问题,提出了一种基于图文特征跨模态深度交互的视觉问答模型。该模型利用卷积神经网络和长短时记忆网络分别获取图像和文本两种模态数据特征,然后利用元注意力单元组合建立的新型深度注意力学习网络,实现图文模态内部与模态之间的注意力特征交互式学习,最后对学习特征进行多模态融合表示并进行推理预测输出。在 VQA-v2.0 数据集上进行了模型实验和测试,结果表明,与基线模型相比,所提模型的性能有明显提升。

关键词:视觉问答;多模态特征融合;注意力机制;深度学习;数据融合
中图法分类号　TP391.41

Visual Question Answering Model Based on Multi-modal Deep Feature Fusion

ZOU Yunzhu[1],DU Shengdong[1,2],TENG Fei[1] and LI Tianrui[1,2]
1 Institute of Computer and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
2 National Engineering Laboratory of Integrated Transportation Big Data Application Technology,Chengdu 611756,China

Abstract　In the era of big data,with the explosive growth of multi-source heterogeneous data,multi-modal data fusion has attracted much attention of researchers,and visual question answering(VQA) has become a hot topic in multi-modal data fusion due to its image and text fusion processing characteristics. Visual Q&A task is mainly based on the deep feature fusion association and representation of image and text multi-modal data,and inference learning of the fusion feature results,so as to get the conclusion. Traditional visual question answering models tend to miss key information and mostly focus on the superficial modal feature association representation learning between data,but less on the deep semantic feature fusion. To solve the above problems,this paper proposes a visual question answering model based on cross-modal deep interaction of of graphic features. The proposed method uses convolutional neural network and LSTM network to obtain the data features of image and text modes respectively,and builds a novel deep attention learning network based on combination of meta-attention units,to realize interactive learning of attention features within or between modes of image and text. At last,we represent the learning features so as to output the results. The model is tested and evaluated on VQA-v2.0 dataset. Compared with the traditional baseline model,the experimental results show that the performance of the proposed model is significantly improved.

Keywords　Visual question answering,Multi-modal feature fusion,Attention mechanism,Deep learning,Data fusion

1　引言

随着大数据时代的到来,对多源异构数据的分析和处理越来越重要,特别是对于多模态数据的融合学习已成为当下深度学习研究的热点[1]。多模态学习可以利用不同模态的重复信息和互补信息进行映射与融合表示,从不同模态的数据中提取融合特征,最终实现跨模态学习的预测、回归和分类。目前,在多模态特征融合研究方面,主要以视觉图像和语言文本两类模态数据作为研究对象,也获得了计算机视觉、自然语言处理等领域研究者的广泛关注,并在视觉问答[2]、图像

到稿日期:2021-12-28　返修日期:2022-06-26
基金项目:国家科技重大专项(2020AAA0105101)
This work was supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (2020AAA0105101).
通信作者:杜圣东(sddu@swjtu.edu.cn)

字幕[3]、视频描述[4]、图文匹配[5]等方向取得了显著的进展。

视觉问答(Visual Question Answering, VQA)是其中一项具有挑战性的任务。给定一张图片和与图片相关的问题,视觉问答的目标是结合图片的视觉信息推理出问题的正确答案,因此视觉问答任务需要对图像和文本进行有效的深度特征融合处理。

视觉问答的示例模型如图 1 所示,包括两个核心步骤:1)从图像和问题中提取各自的特征;2)将提取的图像和文本多模态特征利用设计的视觉问答模型进行学习和理解并得出结论。步骤 1 提取特征的方法很多,例如:使用卷积神经网络[6-7],从图像模态中提取特征;使用 LSTM[8] 自然语言处理模型从问题文本中提取特征。而视觉问答模型的性能差异主要体现在步骤 2 中的多模态学习,即如何将两种模态的特征进行有效的融合表示和学习理解。

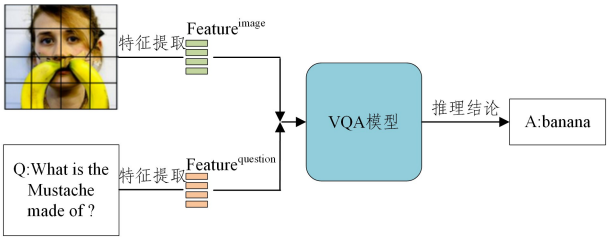

图 1 VQA 示例模型[9]

Fig. 1 VQA sample model[9]

2 相关工作

视觉问答早期研究中采用的跨模态学习方法多是基于简单的特征组合。如 Zhou 等[10]将词袋模型表示的问题特征与图像的卷积特征直接拼接整合;Antol 等[9]使用哈达玛积的方式对图文特征进行组合,模型在实验数据集上表现较好。也有研究者通过对问题和图像特征建模,使用贝叶斯方法统计特征的共现概率以推断问题答案,如 Malinowski 等[11]通过语义分割识别图像目标区域,然后使用贝叶斯方法对特征的空间关系进行建模,计算答案概率;Kafle 等[12]将 VQA 任务由开放式问答转变为包含选项的选择题任务,使用二次判别分析建模了图像特征在给定问题特征、答案类型下的条件概率;Ren 等[13]选择将单词的嵌入特征和图像的卷积特征输入设计的"VIS+LSTM"模型中产生预测结果;Lin 等[14]在不同的维度下计算特征之间的互信息量以得到分类器结果,实现了基于用户行为足迹的多模态融合身份认证。

研究人员进一步将图像和文本两种模态联合嵌入(Joint Embedding)到公共特征空间得到全局特征,再使用融合模型进行分类预测。如 Fukui 等[15]使用池化方法将高维空间内联合表示的两种特征向量做傅里叶变换,从而实现多模态特征的组合;Kim 等[16]提出了一种多模态残差网络,用于学习两类模态的特征联合表示;Meng 等[17]提出了一种新的空间离散余弦动态参数网络,实现了问题与图像特征融合的 Hash 过程;Gu 等[18]使用 logSoftmax 函数对点云和图像特征矩阵得到分类置信度以实现不同模态特征的线性融合操作。然而,图像和文本特征在公共空间内联合表示时容易损失关键信息,影响模型后续的学习和分类预测。针对上述问题,Lu 等[19]通过计算图像的注意力权重和问题的注意力权重,获得了更准确的预测结果;Nguyen 等[20]建立了基于共同注意力层的在视觉表示和语言表示之间的对称架构,用于图像问题对之间的交互;Yu 等[21]提出了一种线性池化方法与注意力机制相结合的深度学习模型;Yan 等[22]结合自底向上的注意力机制和记忆网络建立视觉问答模型。还有一些研究者[23]提出的模型也是计算两种模态的注意力权重信息,用于多模态特征的融合。然而上述方法主要是进行多模态特征的浅层交互,难以对图像、问题两种模态的特征进行深度的特征交互学习,也很难进一步实现多模态特征的融合表示和学习推断。

针对上述问题,研究者将改良模型的方案转向了如何实现图像与文本模态特征的深度融合。如 Chen 等[24]提出了一种新的多模态编码-解码注意力网络,该网络将问题与图像各自的关键区域进行关联,从而捕获丰富且合理的模态特征;Fu 等[25]在模块化注意力网络模型的基础上增加了关系网络,建立起候选对象之间的关联关系,并提出用离散余弦变换增加频率特征,以提高模型的细粒度识别能力;Zou 等[26]利用图注意力机制建模学习问题的视觉关系表示,结合问题特征并协同注意编码,加强了问题词与对应图像区域之间的依赖性。

本文提出了一种多模态深度特征融合的视觉问答模型(Visual Question Answering model of Cross-modal Deep Interaction, CDI-VQA)。模型使用不同的深度神经网络,对两种模态特征进行自注意力学习建模与协同注意力建模;通过提出的注意力学习网络实现深度的跨模态交互式学习,提取出带有丰富注意力权重信息的问题和图像多模态特征;最后通过多模态融合注意力加权后的图像、问题特征,并将融合特征传入分类器中,结合答案文本建立映射,完成分类预测。

本文工作的主要贡献如下:

(1)提出了一种新的基于多模态深度特征融合的视觉问答模型 CDI-VQA,实现了跨模态特征的深度融合。

(2)设计了一种新的协同注意力构建方式,利用自注意力和交互注意力两种元注意力单元,通过单层内两次递进的跨模态特征交互实现图像特征与文本特征各自指导对方的注意力权重学习。

(3)在 VQA v2.0 数据集上进行对比实验,在预测输出答案的准确率指标上,验证了本文提出的模型相比基线模型性能有了较大提升。

3 模型结构设计

CDI-VQA 模型分为 3 个部分:第一部分如图 2(a)所示,主要是分别完成对图像(视觉场景)与文本(问题)模态数据的深度特征抽取;第二部分如图 2(b)所示,主要是将模型第一部分获取的图像特征表示和文本特征表示作为输入,基于构建的新型深度协同注意力网络完成对多模态特征的深度学习交互;第三部分如图 2(c)所示,将经协同注意力网络处理后

的图文特征进行多模态融合，结合训练集中的答案标签数据训练分类器，进行分类预测并输出结果。

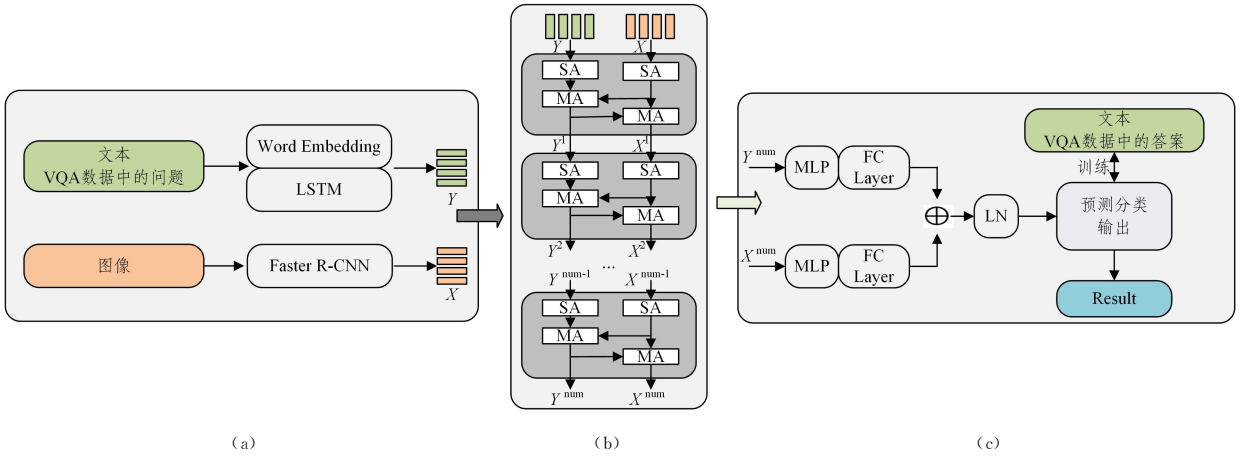

图 2 模型结构设计图

Fig. 2 Design drawing of model structure

3.1 图像文本特征抽取

首先对模型的输入数据进行特征抽取表示，如图 2(a)所示。对输入的图像数据 X_{in}，使用开源 Faster-RCNN[27] 模型进行学习训练，得到图像特征 $X\in\mathbb{R}^{m\times dx}$，它是 VQA 图像数据的特征集合。对输入的问题文本 Y_{in}，先对文本进行预处理划分单词，然后使用开源 GloVe 模型实现单词 Embedding，向量化表示输入的问题文本数据，然后传入 LSTM 网络，抽取得到问题文本特征 $Y\in\mathbb{R}^{n\times dy}$。

3.2 协同注意力网络层构建

基于“多头”注意力[28](Multi-head Attention)机制，构建自注意力(Self-Attention，SA)和交互注意力(Mutual-Attention，MA)两类元注意力单元用于模块化组合成深度协同注意力网络，以进行图文多模态特征的交互学习。“多头”注意力计算式如式(1)所示：

$$\begin{aligned}&MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_h)\boldsymbol{W}^O\\&head_i=Attention(\boldsymbol{QW}_i^Q,\boldsymbol{KW}_i^K,\boldsymbol{VW}_i^V)\end{aligned}\tag{1}$$

其原理是在点积注意力机制的基础上，按“头”的个数将输入等分为 h 份，对等分后的 h 份数据分别通过不同的权重 $\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$ 映射得到新的 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$，以计算相应的 Attention 值，计算式为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\sum\frac{1}{z}\exp\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\tag{2}$$

然后将分割计算的结果重新连接，映射到原始的向量维度，得到注意力特征结果。

式(2)中，z 是归一化因子，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 是注意力宏观理解下的 *Key-Value* 对，两者均是接收到的一种模态特征数据；网络输入的另一种模态特征数据 $\boldsymbol{Q}$ 作为主体，与 $\boldsymbol{K}$ 内积后进行 Softmax 得到相似度概率，最后计算对 $\boldsymbol{V}$ 加权求和的结果。

在“多头”注意力机制基础上建立 SA 单元，如图 3(a)所示，该单元仅输入单种模态 X 作为式(2)的 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$。SA 单元通过“多头”注意力层获取 *self-attention*，学习特征集 X 内成对样本$\langle x_i,x_j\rangle$之间的关系，对所有成对实例的相似度加权求和得到注意力处理后的特征。特征信息随后进入使用 ReLU 激活函数的全连接层和另一个用于归一化的全连接层，实现特征的空间变换，提升模型的表达能力，最终输出得到 *Attention*(*SA*)，它是输入模态 X 的所有特征 x_i 之间的归一化相似度重构集合。

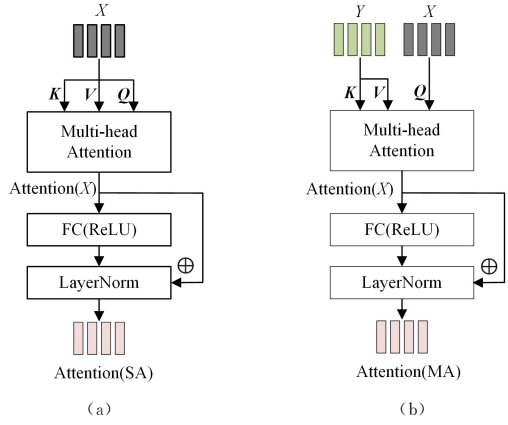

图 3 元注意力单元设计

Fig. 3 Design of meta attention unit

同理建立 MA 元注意力单元，如图 3(b)所示。MA 单元与 SA 的不同点在于，单元的输入使用了 X 和 Y 两种模态特征，模态 Y 的特征数据用于指导模态 X 的注意力学习，即模态 Y 的特征是式(2)的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$，模态 X 的特征作为计算主体 $\boldsymbol{Q}$。MA 单元学习特征集 X 各元素与特征集 Y 所有元素的成对样本$\langle x,y_j\rangle$之间的关系，利用 Y 指导 X 学习。最终输出 *Attention*(*MA*)，即输入模态 X 的各特征 x_i 与输入模态 Y 的所有特征交叉学习后的归一化相似度重构集合。

将两种元注意力单元进行模块化组合，得到新型的协同注意力层结构，如图 4 所示。

协同注意力层结构包含两个 SA 单元、两个 MA 单元，其实现共分 3 个步骤：

(1)SA(Text)单元和 SA(Image)单元并行化处理，分别实现文本和图像模态内的自注意力特征建模，有利于单模态内全局信息的捕捉和关键特征的获取。

(2)实现协同注意力层结构的第一次跨模态特征交互。新型层结构模拟人类“先看图像，然后带着图像信息浏览问题”的自然行为，使用 MA(Text)单元，将经自注意力处理后

的图像特征作为“指导”提供 MA 单元所需的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$；将自注意力处理后的文本特征作为 MA 单元所需的 $\boldsymbol{Q}$，实现协同注意力建模与第一次跨模态特征交互。

(3)实现协同注意力层结构的第二次跨模态特征交互。使用 MA(Image)单元，利用文本特征帮助获取图像特征中的关键信息，此时将步骤(2)交互学习后的文本特征作为 $\boldsymbol{K}$ 和 $\boldsymbol{V}$，自注意力处理后的图像特征作为协同注意单元的主体 $\boldsymbol{Q}$，完成最终的跨模态特征交互。

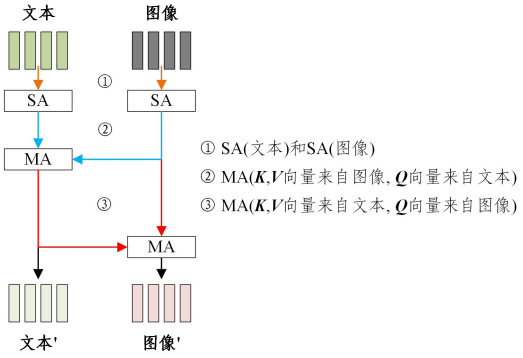

图 4 协同注意力层结构

Fig. 4 Co-attention layer structure

上述协同注意力层的设计与已有工作(如文献[20,29])不同：首先本文提出的协同注意力层结构是非对称的，且能更好地模拟人类面对图像问题的思考与行为方式，自然地学习两种模态各自最相关联的特征信息；其次本文提出的模型更加简洁，仅使用两个 SA 单元和两个 MA 单元的模块化组合就能实现图像和文本的深度跨模态交互，如图 4 所示，单层协同注意力结构只需要最初输入的图像和文本特征就能实现特征交互的功能，不再需要额外的特征向量或参数。

单个协同注意力层结构输出的结果可以作为新的协同注意力层的输入，将多个注意力层串联堆叠，得到最终如图2(b)所示的深度串联注意力学习网络，协同注意力层(Co-Attention Layer)简写为 CAL。假设模型共堆叠 Num 层 CAL，第 num 层可表示为 CAL_{num}，输入的图像特征和问题特征分别表示为 $X^{\text{num}-1}$ 和 $Y^{\text{num}-1}$，作为下一个串联 CAL 层的输入，如式(3)所示：

$$(X^{\text{num}}, Y^{\text{num}}) = CAL_{\text{num}}(X^{\text{num}-1}, Y^{\text{num}-1}) \tag{3}$$

特别地，对于 CAL_1，其输入图像特征和文本特征分别为 $X^0 = X$ 以及 $Y^0 = Y$。深度串联注意力学习网络的输出为 $X^{\text{Num}} \in \mathbb{R}^{m \times d}$ 和 $Y^{\text{Num}} \in \mathbb{R}^{n \times d}$。

3.3 特征融合与分类预测输出

对图像特征 X 和问题特征 Y 进行协同注意力学习，输出的图像特征 X^{Num} 和文本特征 Y^{Num} 各自携带有丰富的图像区域和问题单词的注意力权重信息。使用 MLP 学习特征信息，得到归一化的权重概率，计算式如式(4)所示：

$$att^{y} = \frac{e^{\text{MLP}(Y^{\text{Num}})}}{\sum e^{\text{MLP}(Y^{\text{Num}})}} \tag{4}$$

利用新权重概率对特征加权求和得到最终的图像特征 $\boldsymbol{x}^*$ 和问题特征 $\boldsymbol{y}^*$，如式(5)所示，$\boldsymbol{x}^*$ 的计算方法类同。

$$\boldsymbol{y}^* = \sum_{i=1}^{n} att_i^{y} \boldsymbol{y}_i^{\text{Num}} \tag{5}$$

然后基于双线性池化(Bilinear Pooling)的思想，将计算得到的图像特征 $\boldsymbol{x}^*$ 和问题特征 $\boldsymbol{y}^*$ 使用融合函数进行融合，如式(6)所示：

$$\boldsymbol{res} = LayerNorm(\boldsymbol{A}_x^{\mathrm{T}} \boldsymbol{x}^* + \boldsymbol{A}_y^{\mathrm{T}} \boldsymbol{y}^*) \tag{6}$$

其中，$\boldsymbol{A}_x^{\mathrm{T}}, \boldsymbol{A}_y^{\mathrm{T}} \in \mathbb{R}^{d \times d_{\text{res}}}$ 是两个线性投影矩阵，由融合前设置的全连接层参数决定，d_{res} 是融合特征的共同维度，LayerNorm 层在输入序列张量的最后一个维度求均值和方差，然后对输入特征标准化，便于模型后续的预测分类。

多模态模态特征融合后得到融合特征 $\boldsymbol{res}$，随后进入 N-分类器(Classifier)，建立输入融合特征 $\boldsymbol{res}$ 与输出预测答案 $result$ 之间的映射关系。其中 N 是训练集使用的答案(Answer)标签中出现频率较高的标签数量。模型使用交叉熵损失函数，计算式如式(7)所示：

$$Loss = -\sum_{v}^{N} y_v \log(p_v) \tag{7}$$

其中，N 为标签数量，y_v 是对预测结果的标记值，p_v 是预测分类结果为第 v 类的概率。

综上，本文建立了图文特征跨模态深度交互的视觉问答模型 CDI-VQA。

4 模型实验及结果对比分析

基于 VQA-v2.0 数据集进行模型实验和分析评估。实验运行环境为 Ubuntu 18.04，采用 Pytorch1.8-cuda10-gpu-vnc 作为深度学习模型的框架。实验硬件环境为：CPU：4Core，运行内存 8GB，GPU：1Core，类型为 TITAN_V，存储内存大小 256GB。

实验使用的深度学习开源库包括 Faster R-CNN 模型[27]和 OpenVQA 平台[21]。

4.1 数据集介绍

VQA v2.0 数据集[30]是视觉问答研究中使用最广泛的数据集，该数据集的图像来源于 MS-COCO 数据集，基于真实场景且内容丰富，包含了与图像相关的由人工注释、收集的问答对。VQA v2.0 数据集中，每张图片包含若干个问题，每个问题对应 10 个来自不同收集者的“基本真实答案(Ground Truth Answer)”。实验中使用的数据分为训练集和测试集两部分，包含 443 757 个训练问题对应 4 437 570 个训练答案，214 354 个测试问题对应 2 143 540 个测试答案。

4.2 模型性能评估方法及评估指标

视觉问答属于开放式任务，实际验证需要尽可能减小句法和语义正确性对评估的影响，因此数据提供者对问题答案的长度进行了限制。数据集中每个问题包含 10 个参考答案，出现次数最多的答案被确认为标准答案。只需将 CDI-VQA 模型预测得到的问题答案与标准答案进行对比，并将模型所有问题的预测结果进行总结，就能计算得到模型的准确率(Accuracy，Acc)，即模型的评价指标。

由于 VQA v2.0 数据集的问题种类超过了 20 种，模型对不同类型问题预测正确答案的难易程度是不同的，因此实验中针对性地选出若干种问题类型，分别计算通过 CDI-VQA 模型预测输出的结果对应某类型问题的准确率。

我们将答案类型中的“是/否”(Yes/No)和“数字”(Num-

ber)分别作为一类，将其他类型的答案归为“其他”(Other)一类，共同参与 CDI-VQA 模型预测结果的评估。模型性能评估所使用的评价指标为不同问题类型的准确率，如表 1 所列。

表 1 不同类别问题的准确率定义

Table 1 Definition of accuracy of different types of problems

参数名称	参数含义
Acc_{All}	模型对全部问题进行预测的答案的平均准确率
$Acc_{Yes/No}$	模型对“是/否”类型问题所预测答案的平均准确率
Acc_{Number}	模型对“数字”类型问题所预测答案的平均准确率
Acc_{Other}	模型对“其他”类型问题所预测答案的平均准确率

4.3 实验参数设置

实验参数设置如下：输入的问题特征维度 d_y、输入的图像特征维度 d_x，以及融合的特征数据维度 d_{res} 分别为 512，2048 和 1024，N 设置为 3129，参数由深度学习开源库 OpenVQA 提供。

“多头”注意力单元的潜在维度 d 设置为 512，同时将“头数”h 设置为 8，由于需要将输入的向量按照 $head$ 个数等分为 h 份，因此“头”的潜在维度 $d_h=64$；新型注意力学习网络的批处理大小设置为 64。设置基础学习率为 1×10^{-4}；$dropout$ 参数设置为 10%。

实验选用的基线模型为 MFB[21] 模型和 MFH[31] 模型，二者都是使用注意力方法的多模态数据融合模型。MFB 模型在基本的多模态双线性池化方法的基础上增加了协同注意力机制，分别学习文本注意力和图像注意力；MFH 模型是 MFB 模型的改进，它将原操作细化分为扩张(Expand)和压缩(Squeeze)两个阶段，通过基础注意力层的堆叠计算得到更高阶的信息。

4.4 实验结果对比分析

实验首先基于深度注意力网络中的协同注意力层数 Num 对 CDI-VQA 模型的性能影响进行了对比实验分析与性能优化，$Num\in[2,8]$；然后将优化后的 CDI-VQA 模型与解决视觉问答任务的两种参考模型进行性能比较。

本文设计实验探讨深度注意力网络的协同注意力层数 Num 对模型处理不同类型问题的准确率的影响。通过 Num 的不同取值，得到的 CDI-VQA 模型处理各类型问题的准确率结果和批处理速度，如表 2 所列，并依据表 2 的准确率数据绘制出了图 5 所示的折线图。下文将详细探讨协同注意力层数对模型处理不同类型问题的性能影响。

表 2 不同协同注意力层数的 CDI-VQA 模型评估结果

Table 2 Evaluation results of CDI-VQA model with different collaborative attention layers

Num	Acc_{All}/%	Acc_{Other}/%	$Acc_{Yes/No}$/%	Acc_{Number}/%	$Speed$/(s/batch)
2	49.82	39.83	68.74	33.15	0.6673
3	49.64	40.95	66.28	34.21	0.7245
4	**50.60**	**41.44**	68.55	33.65	0.7598
5	50.44	40.54	**69.14**	34.13	0.8156
6	50.48	40.68	68.88	**34.59**	0.8679
7	49.38	39.04	68.37	33.88	0.9375
8	49.63	39.17	68.55	33.76	**0.9784**

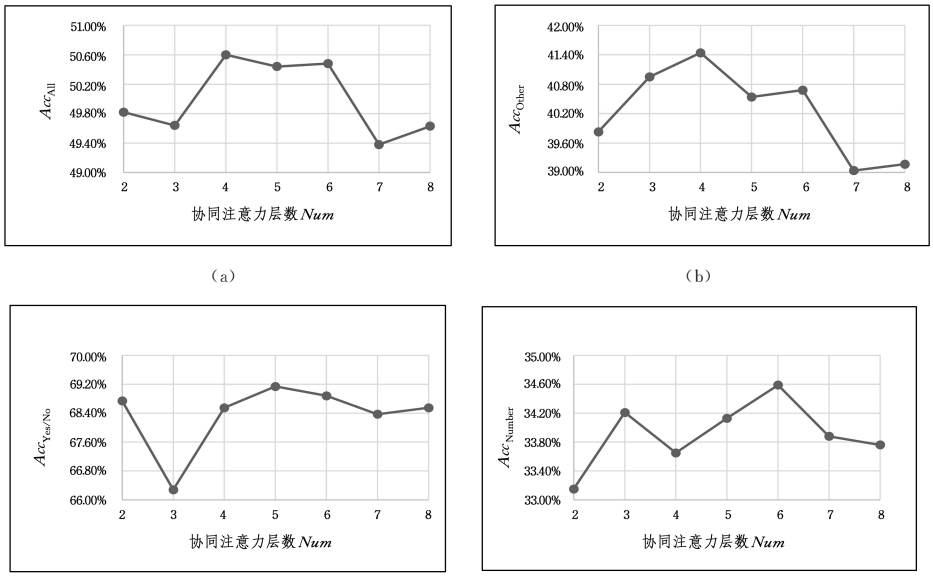

图 5 不同协同注意力层数对模型的性能影响结果示例图

Fig. 5 Sample graph of effect of different CAL on model performance

CDI-VQA 模型预测所有问题结果的平均准确率综合反映了模型的性能。从 $Num=2$ 开始，Num 值逐渐上升，模型预测结果的准确率也随之提升，当 $Num=4$ 时 Acc_{All} 取得最大值 50.60%，且平均准确率 Acc_{All} 在 $Num=\{4,5,6\}$ 时均超过 50%，达到饱和；当 $Num>6$ 后，CAL 层数过大，导致参与权重数量增加，训练梯度不稳定，模型泛化能力下降，性能降低。层数 Num 对模型处理 Other 类型问题的平均准确率的影响与处理所有问题的影响类似：实验初期随着 Num 值的上升，Acc_{Other} 的值不断增大，$Num=4$ 时取得最大值 41.44%；随着 Num 的上升，Acc_{Other} 开始下降。

模型处理 Other 类型问题与所有问题的结果变化趋势近似，表明了选取问题的合理性与泛用性。层数 Num 对模型处理 Yes/No 类型问题的平均准确率的影响较小，除了 $Num=3$ 时结果反映的准确率较低，其余 Num 值得到的准确率结果都超过了 68.3%且不存在较大波动。这是由于 Yes/No 类型问题的答案只能是{Yes，No}二者选其一，问题性质

简单，预测输出 *result* 的表示范围窄，参数的调整对模型预测该类型问题的性能影响不大。接着，实验研究了层数 *Num* 对模型处理 Number 类型问题的平均准确率的影响，在初期，准确率同样随 *Num* 值的上升而提高，在 $Num=6$ 时 Acc_{Number} 得到最高准确率 34.59%；$Num>6$ 后，模型对“数字”类型问题所预测答案的平均准确率开始下降。我们认为，Number 类型问题的答案主要取决于选取图像特征的有效区域数量，*Num* 值的提高会使模型结合文本特征，更专注学习图像的关键区域特征，但过高的 *Num* 值会弱化模型预测 Number 类型问题的准确率，导致模型性能降低。

最后，*Num* 的取值同样会对模型的批处理速度造成影响。通过数据我们可以明显地观察到，随着 *Num* 值的增大，处理单个 batch 所花的时间不断增加（*Speed*(s/batch)），在 $Num=8$ 时取得最大值 0.9784/(s/batch)。模型越复杂，机器计算处理的速度越慢，需花费的时间越多。

部分实验预测结果示例如图 6 所示，也验证了表 2 的数据结果，模型对 Yes/No 类型的问题的预测准确率更高，对 Number 类型的问题预测准确率偏低。另外，如图 6 的第三个问题所示，模型因受到知识系统的限制，输出的预测结果仅对有限的答案参数选项数据进行建模，这种情况下就难以获得最佳答案。

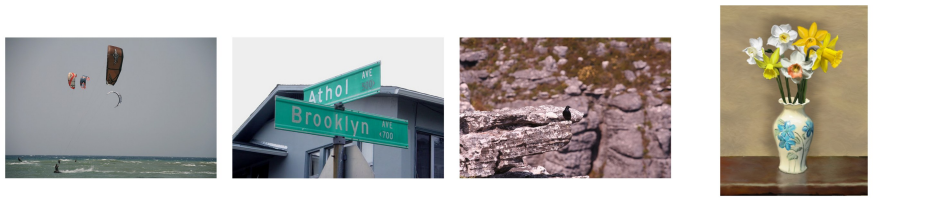

图 6 视觉问答实验结果展示样例

Fig. 6 Example of VQA experiment results

通过对比分析深度注意力网络中的协同注意力层数 *Num* 对 CDI-VQA 模型性能的影响，我们发现 *Num* 取值在 {4,5,6} 时得到了准确率较高的输出结果。

将 $Num=\{4,5,6\}$ 的 3 种 CDI-VQA 模型与基线模型在相同的部署环境和设备条件下基于 VQA v2.0 数据集进行了比较，对比结果如表 3 所列。

表 3 CDI-VQA 模型与基线模型的性能比较

Table 3 Performance comparison between CDI-VQA model and baseline model

（单位：%）

Model type	Acc_{All}	Acc_{Other}	$Acc_{Yes/No}$	Acc_{Number}
MFB[21]	41.66	26.15	66.14	29.87
MFH[31]	48.92	37.66	68.01	33.71
CDI-VQA ($Num=4$)	**50.60**	**41.44**	68.55	33.65
CDI-VQA ($Num=5$)	50.44	40.54	**69.14**	34.13
CDI-VQA ($Num=6$)	50.48	40.68	68.88	**34.59**

本文提出的 CDI-VQA 模型相比两种基线模型的进步在于：CDI-VQA 对单模态的自注意力建模相比基线模型更好地把握了单个模态的全局特征信息，降低了关键特征的损失；CDI-VQA 模型模块化地组建了协同注意力层 CAL，模拟人类对图像问题的处理行为，通过两次跨模态的特征交互，让文本特征和图像特征互为“指导”，较两种基线模型更系统、全面地进行跨模态特征交互，进一步提升了文本和图像多模态特征融合的质量；CDI-VQA 模型利用协同注意力层输出可作为另一协同注意力层输入的性质，将多个协同注意力层串联堆叠得到深度协同注意力学习网络，较基线模型的组成结构更加自然严谨，网络层数更深，因而提升了预测结果的准确率和模型的性能。实验结果表明：本文设计的 CDI-VQA 模型在性能上优于其他两类模型，如表 3 所列，评价指标中的 Acc_{All} 和 Acc_{Other} 均在 $Num=4$ 的 CDI-VQA 模型下取得了最大值；$Acc_{Yes/No}$ 在 $Num=5$ 的 CDI-VQA 模型下取得了最大值；Acc_{Number} 在 $Num=6$ 的 CDI-VQA 模型下取得了最大值。

结束语 本文提出了一种基于多模态深度特征融合的视觉问答模型 CDI-VQA。该模型使用不同的神经网络分别从图像和文本数据中抽取特征，然后利用抽取的特征进行模态内部和模态之间的注意力建模；在模型的注意力学习网络设计中，两种模态的特征相互作为注意力权重学习的参考，实现了图文模态特征的深度交互；最后通过多模态融合函数融合注意力加权后的图像信息和文本语义，并将融合特征传入分类器结合答案文本数据预测结果。CDI-VQA 模型对两种模态特征进行并行化的自注意力建模，保证了单模态关键特征的有效获取；同时模型构建了新型协同注意力单元层，模拟了人对图像问题的思考方式和自然行为，两次模态特征交互使得图像特征与文本特征各自指导对方的注意力权重学习，实现了跨模态交互式学习；CDI-VQA 模型将多个协同注意力层串联堆叠得到深度注意力学习网络，进一步提升了模型性能。实验使用 VQA-v2.0 数据集进行模型的评估分析与训练优化，并与基线模型在相同的实验条件下进行了比较，结果表明本文提出的 CDI-VQA 视觉问答模型具有更佳的性能。

当前的研究只设计出了一种协同注意力层构建方式，且仅探讨了注意力网络的层数对模型性能的影响，还需深入研究其他因素，例如“多头”注意力的头数、相似度运算方法的选择等对模型性能的影响，这将是未来研究的重点内容。

参 考 文 献

[1] WU A M, JIANG P, HAN Y H. Survey of Cross-media Question Answering and Reasoning Based on Vision and Language [J]. Computer Science, 2021, 48(3): 71-78.

[2] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based lo-

calization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:618-626.
[3] DU H J,LIU X L. Image description generation method based on inhibitor learning [J]. Journal of Image and Graphics,2020,25(2):333-342.
[4] XU S K,NI C H,JI C C,et al. Image Caption of Safety Helmets Wearing in Construction Scene Based on YOLOv3 [J]. Computer Science,2020,47(8):233-240.
[5] LEE K H,CHEN X,HUA G,et al. Stacked cross attention for image-text matching[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:201-216.
[6] ZHOU Y X,YU J. Design of Image Question and Answer System Based on Deep Learning [J]. Computer Application and Software,2018,35(12):199-208.
[7] ZHUANG M Q,TAN X H,FAN Y C,et al. 3D Animation Expression Generation and Emotional Supervision Based on Convolutional Neural Network [J]. Journal of Chongqing University of Technology(Natural Science),2022,36(01):151-158.
[8] XU S,ZHU Y X. Study on Question Processing Algorithms in Visual Question Answering [J]. Computer Science, 2020, 47(11):226-230.
[9] ANTOL S,AGRAWAL A,LU J,et al. Vqa:Visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:2425-2433.
[10] ZHOU B,TIAN Y,SUKHBAATAR S,et al. Simple baseline for visual question answering [J]. arXiv:1512. 02167,2015.
[11] MALINOWSKI M,FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input [J]. Advances in Neural Information Processing Systems,2014,27:1682-1690.
[12] KAFLE K,KANAN C. Answer-type prediction for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4976-4984.
[13] REN M,KIROS R,ZEMEL R. Exploring models and data for image question answering [J]. Advances in Neural Information Processing Systems,2015,28:2953-2961.
[14] LIN M Q,ZHANG X M. Identity Authentication of Multi-Modal Fusion Based on Behavioral Footprint[J]. Computer Engineering,2021,47(10):116-124.
[15] 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. 2016:457-468.
[16] KIM J H,LEE S W,KWAK D,et al. Multimodal residual learning for visual qa[C]// Advances in Neural Information Processing Systems. 2016:361-369.
[17] MENG X S,JIANG A W,LIU C H,et al. Visual Question Answering based on Spatial-DCTHash Dynamic Parameter Network [J]. SCIENTIA SINICA Informationis,2017,47(8):1008-1022.
[18] GU L,JI Y,LIU C P. Classification Method of Three-Dimensional Point Cloud Based on Multiple Modal Feature Fusion[J]. Computer Engineering,2021,47(2):279-284.
[19] LU J,YANG J,BATRA D,et al. Hierarchical question-image co-attention for visual question answering [J]. Advances in Neural Information Processing Systems,2016,29:289-297.
[20] NGUYEN D K,OKATANI T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:6087-6096.
[21] 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. 2017:1821-1830.
[22] YAN R Y,LIU X L. Visual Question Answering Model Based on Bottom-up Attention and Memory Network [J]. Journal of Image and Graphics,2020,25(5):993-1006.
[23] WANG Y L,ZHUO Y F,WU Y J,et al. Question Answering Algorithm on Image Fragmentation Information Based on Deep Neural Network [J]. Journal of Computer Research and Development,2018,55(12):2600-2610.
[24] CHEN C,HAN D,WANG J. Multimodal encoder-decoder attention networks for visual question answering [J]. IEEE Access,2020,8:35662-35671.
[25] 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.
[26] ZOU P R,XIAO F,ZHANG W J,et al. Multi-Modele Co-Attention Network for Visual Question Answering [J]. Computer Engineering,2022,48(2):250-260.
[27] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149.
[28] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need [J]. arXiv:1706. 03762,2017.
[29] LI L. Research on Collaborative Attention Model and Deep Correlated Networks for Visual Question Answer [D]. Xiamen:Huaqiao University,2020.
[30] NIU Y L,ZHANG H W. Survey on Visual Question Answering and Dialogue [J]. Computer Science,2021,48(3):87-96.
[31] YU Z,YU J,XIANG 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.

ZOU Yunzhu, born in 1999, postgraduate. His main research interests include data fusion and fake news detection.

DU Shengdong, born in 1981, Ph.D, associate professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning and knowledge graph.

(责任编辑:何杨)