Computer Science and Application 计算机科学与应用, 2023, 13(6), 1188-1198
Published Online June 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.136116

Hans汉斯

基于交叉记忆注意力的视觉问答模型

文必成，常青玲，刘兴林，崔　岩*

五邑大学智能制造学部，广东　江门

收稿日期：2023年5月5日；录用日期：2023年6月1日；发布日期：2023年6月7日

摘　要

视觉问答是一项涉及图像和文本的多模态任务，给定一个图像和一个用自然语言表达的问题，视觉问答系统需要对视觉和文本信息同时进行复杂的理解，提供关于图像的这个问题的准确答案。现有的视觉问答模型在获取与问题相关的图像区域时，不能有效利用文本与图像信息的多层次特征信息，因此，我们使用自注意记忆层，使得所得特征的每一层包含之前的先验知识。同时利用交叉记忆模块，在解码端的所有引导注意力层中输入编码端的各级加权特征，通过引导注意力，融合低层次与高层次信息，使用多层次信息更好地关注图像特征中的关键区域。本文在VQA v2.0数据集上进行了对比实验，表明该模型能充分利用图像和文本的多层次特征信息，与当前主流模型相比更具优越性。

关键词

视觉问答，注意力机制，特征融合

Visual Question Answering Model Based on Cross Memory Attention

Bicheng Wen, Qingling Chang, Xinglin Liu, Yan Cui*

Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen Guangdong

Received: May 5th, 2023; accepted: Jun. 1st, 2023; published: Jun. 7th, 2023

Abstract

Visual question answering (VQA) is a multimodal task involving images and text. Given an image and a question expressed in natural language, a visual question answering system needs to understand both visual and textual information in a complex way to provide an accurate answer to this question about the image. The existing visual question answering model cannot effectively utilize

*通讯作者。

文章引用: 文必成, 常青玲, 刘兴林, 崔岩. 基于交叉记忆注意力的视觉问答模型[J]. 计算机科学与应用, 2023, 13(6): 1188-1198. DOI: 10.12677/csa.2023.136116

the multi-level feature information of text and image information when acquiring the image area related to the question. Therefore, we use the self-attention memory layer so that each layer of the obtained feature contains the previous prior knowledge. At the same time, the cross-memory module is used to input the weighted features of the encoder at all levels in all the guided attention layers of the decoder. By guided attention, the low-level and high-level information are fused, and the multi-level information is used to better focus on key areas in the image features. In this paper, we conduct comparative experiments on the VQA v2.0 dataset and show that the model can make full use of the multi-level feature information of images and text, and is superior to the current mainstream models.

Keywords

Visual Question Answering, Attention Mechanism, Feature Fusion

1. 引言

随着社交媒体的飞速发展，图片、文字以及视频等多种类型的数据量迅速增长，数据的模态逐渐多样化，视频、图片、文本等数据可以得到一个物体的各方面互补信息，促进了多模态学习[1]的发展。近来，多模态视觉问答任务[2]不断发展，逐渐应用在教育、医疗以及媒体等多个领域[3] [4]，其目的是开发一个系统来回答关于输入图像的特定问题。答案可以是以下任意形式：单词、短语、二进制答案、多选答案或填空答案。与其他任务相比，视觉问答任务需要对图像和文本进行细粒度的语义理解，图片和视频等视觉信息的表达能力和信息涵盖能力比文本更强，如何通过交互式的方法从视觉信息中提取信息、过滤信息以及推理信息，是视觉问答研究的热点方向。

视觉问答任务涉及到特征提取、知识推理、特征融合[5]等复杂技术，需要识别文本的语义信息以及图片中的物体属性以及空间关系等信息，还需要进行一定的推理。如图 1 所示，视觉问答任务作为一个典型的多模态问题，其主要步骤一共有三个：1) 对图像特征和文本特征的提取，利用计算机视觉以及自然语言处理的相关技术，获取图像和文本中丰富的语义信息；2) 在获取图像和文本的语义信息后，要求模型同时理解不同模态的特征，建立视觉与语言模态之间的关联；3) 如何利用融合后的特征推理得出问题的正确答案。为解决这些问题，众多视觉问答模型使用特征融合、注意力机制、模块化网络等多种方法。

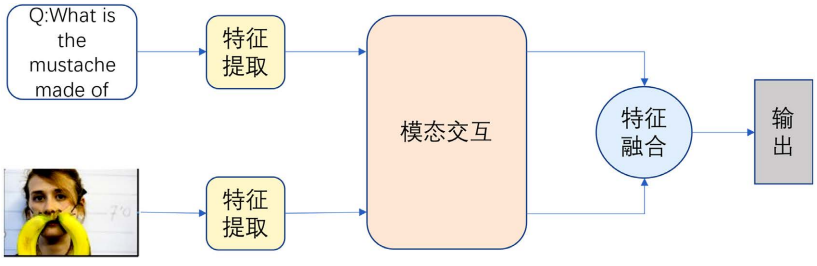

Figure 1. Composition of visual question answering system
图 1. 视觉问答系统组成

2. 相关工作

在过去几年中，有大量的深度神经网络模型被用来提高视觉问答任务的性能。早期的视觉问答模型使用全局特征的多模态融合，使用卷积神经网络[6]提取视觉特征，使用 LSTM [7]提取文本特征，将图像和问题共同映射到一个共享空间，然后使用特征融合模型进行融合，推断最终答案。基于注意力的方法也被广泛应用与视觉问答任务，注意力模块的目的是识别文本和图像区域最关键的信息，忽略无关信息，从而避免干扰。Anderson 等人[8]提出了一种自底向上和自顶向下相结合的注意力机制，利用目标检测网络 Faster R-CNN [9]来实现自底向上的注意力，使得提取出的每个图像区域都有一个特征向量，使用自顶向下的注意力机制决定每个特征区域的权重，将所得权重与注意力相结合，以获得更细粒度的图像理解。

除了理解图像的视觉内容，模型同时需要理解问题的文本内容，因此，协同注意力被用来同时学习文本注意力和视觉注意力。Nam [10]等人提出双重注意网络 DAN，利用文本注意力和视觉注意力使得两种模态相互关注，在推理时相互引导，在匹配时计算问题与图像的相似性，获得注意力特征。Yu 等人[11]提出了多模态分解双线性池模型，为解决特征表示的复杂性，从问题中提取文本特征，文本特征参与图像特征的提取，组合文本与图像特征，共同学习文本和图像注意。Cadene 等人[12]提出的用于关系推理的 MUREL 模型，通过端到端学习对真实图像进行推理，使用向量表示实现问题和图像区域之间的交互，并应用成对组合对区域关系进行建模。Yu 等人[13]提出的深度模块化协同注意网络 MCAN，使用自注意力和引导注意力对视觉信息和问题信息之间的模态内和模态间交互进行建模，通过自注意力与引导注意力的模块化组合，可以对文本和图像特征进行深度级联，实现共同关注。Chen 等人[14]提出 MEDAN 模型，通过将问题关键词和图像中的关键区域相关联来获取丰富合理的特征信息。Guo 等人[15]提出 re-attention 架构，首先计算每个单词－对象对的相似度来学习对象的初始注意力权重，然后通过基于答案的视觉对象重建图像注意力映射。Liu 等人[16]提出 DSACA 模型，通过使用新提出的自注意机制分别对空间和序列结构的内部依赖性进行建模。Sharma 和 Jalal [17]使用图像特征来回答与前景对象和背景区域相关的问题，并使用图神经网络对图像中图像区域对对象之间的关系进行编码，并生成图像说明，使用两个注意模块来利用彼此的知识。

上述研究人员所做的研究，均通过注意力机制建立了图像信息与文本信息的联系，但在模型交互的过程中没有充分利用输入信息的多层次特征和先验知识，不能对各区域之间的建立先验知识联系。例如，当给定编码一个人的区域和编码一个滑板的区域，在没有先验知识联系的情况下，很难推断出运动的概念。当建立充足的先验关系后，可以简单的从编码人和电脑的区域中推测出工作的概念。

本文以 MCAN [13]模型为基础架构，使用自注意记忆机制，进一步提升模型的性能。之前的模型在获取与问题相关的图像区域时，不能有效利用文本与图像信息的多层次特征，我们使用自注意记忆层，使得所得特征的每一层包含之前的先验知识，通过问题引导的注意力层获取与关键文本相关的图像区域。同时利用交叉记忆模块[18]，在解码端的所有层中输入编码端的各级特征，融合低层次与高层次信息，使用多层次信息更好的关注图像特征的关键区域。

3. 交叉记忆注意力模型

本文提出一种基于交叉记忆注意力的视觉问答模型，模型的整体架构如图 2 所示，图像特征的提取方法是采用 Faster R-CNN 网络作为特征提取网络对不同类别物体进行特征提取，文本采用 Glove [19] + LSTM 获得文本特征，利用自注意记忆层和问题引导的注意力层获取特征中的关键信息，不仅关注图像与文本中的关键信息，而且利用其中的多层次信息，关注与问题最相关的图像特征，采用编码器－解码器架构进行图片和文本信息的多级交互。最后，将所获得的图片和文本特征进行线性融合，在输出分类器上进行答案预测。

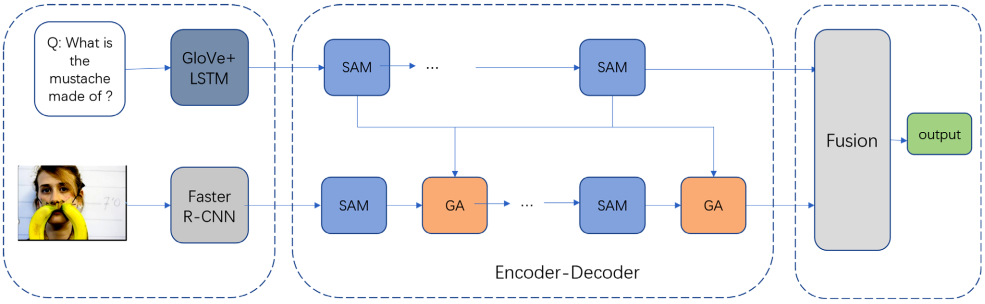

Figure 2. The overall architecture of the model
图 2. 模型整体网络架构

3.1. 交叉注意力网络层构建

如图 3 所示，给定图像 Y 的输入序列，以及所有编码层的输出 X_i，交叉注意力操作通过交叉注意力将 Y 连接到 X_i 中的所有元素。不只是关注最后一个编码层，而是对所有编码层进行交叉关注，对每个编码层的信息进行加权计算，这些多层次信息在调整后汇总在一起：

$$M\left(X_i,Y\right)=attention\left(W_qY,W_kX_i,W_vX_i\right) \tag{1}$$

如式(1)，$M\left(X_i,Y\right)$ 表示编码器－解码器交叉注意，使用解码器的查询以及编码器的键和值进行计算。

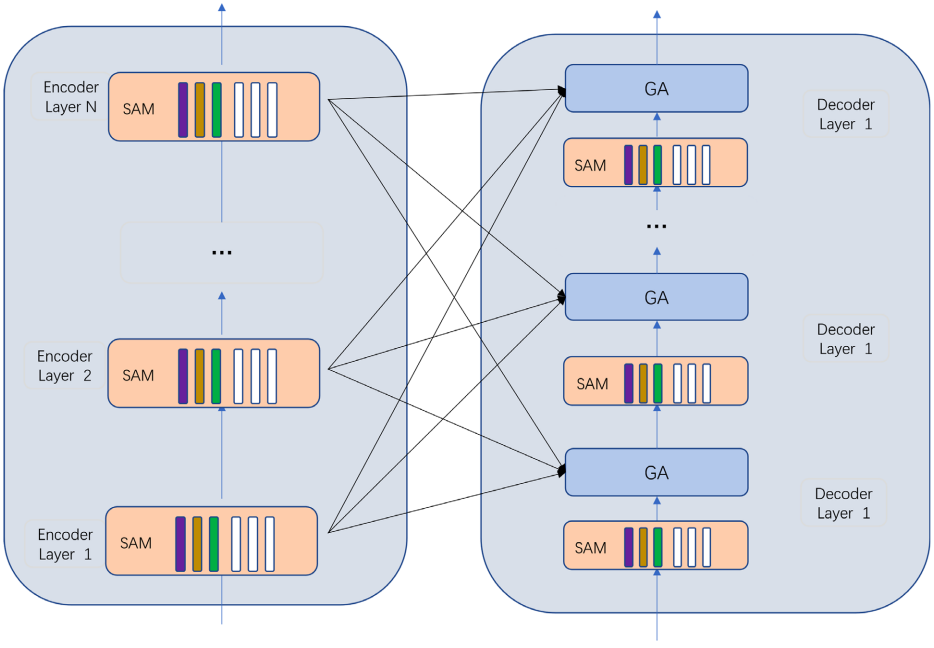

Figure 3. Encoder-decoder based on cross memory attention
图 3. 基于交叉记忆注意力的编码－解码器

$$F\left(X_i,Y\right)=\sum_{i=1}^{N}M\left(X_i,Y\right)\odot\alpha_i \tag{2}$$

如式(2)，α_i 是与交叉注意结果大小相同的权重矩阵。α_i 中的权重既调节了每个编码层的单个贡献，也调节了不同层之间的相对重要性。这些是通过测量使用每个编码层计算的交叉注意结果与输入查询之间的相关性来计算的，如下：

$$\alpha_i = \sigma\left(W_i\left[Y, M\left(X_i, Y\right)\right] + b_i\right) \tag{3}$$

式(3)中，σ 是 sigmoid 激活函数，W_i 是 $2d \times d$ 权重矩阵，b_i 是可学习的偏差矩阵。

3.2. 基于自注意记忆层的编码器

在编码器中，通过改进自注意层，使用自注意记忆层学习问题特征。传统的自注意力层使用多头注意力机制，由 N 个平行头组成，其中每个头可以表示为缩放的点积注意力函数，其中注意函数 $attn(Q,K,V)$ 作用于 Query，Key 和 Value 分别对应于查询、键和值。这个注意力函数的输出是加权平均向量。为此，我们首先计算 Q 和 K 之间的相似性分数；并使用 *Softmax* 将分数归一化。然后将归一化后的注意力权重与 V 结合，生成加权平均向量。d 是 Q 和 K 的维数，这两个维度是相同的。

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

然而，自注意力无法建立先验知识模型，不能有效利用多层次的特征关系，因此我们使用自注意力记忆层。如图 4，将自注意层中用于自我注意的键和值扩展为额外的“槽”M，可以对先验信息进行编码。为了表示先验信息不依赖于输入集 X，额外的键 M_k 和值 M_v 被实现为可直接更新的普通可学习向量。

$$att(Q,K,V) = attention\left(W_q X, K, V\right) \tag{5}$$

$$K = \left[W_k X, M_k\right] \tag{6}$$

$$V = \left[W_v X, M_v\right] \tag{7}$$

其中，M_k 和 M_v 是需要训练的可学习参数，通过添加可学习的键和值，保持查询操作不变，可以检索已学知识。同时，自注意记忆模块使用多头注意力，注意操作重复 h 次，使用不同的投影矩阵 W_q、W_k、W_v 和每个头部的不同可学习记忆 M_k、M_v。然后，将来自不同头部的结果连接起来，使用线性投影。

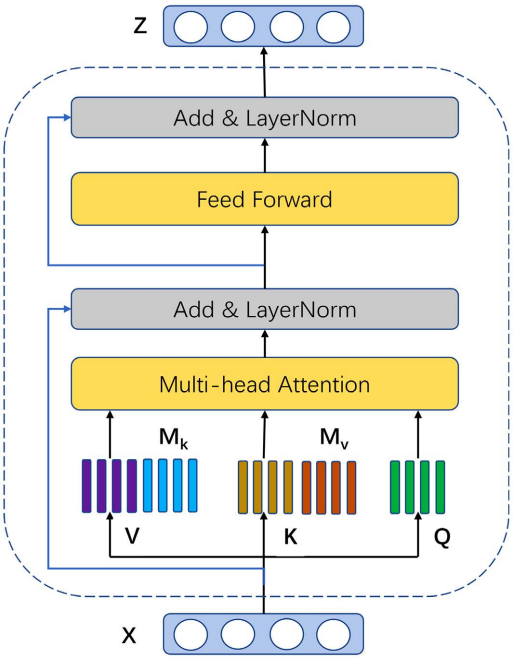

Figure 4. Self attention memory layer
图 4. 自注意记忆层

我们将多层自注意记忆层堆叠，每一层的输出经过前馈层，作为下一层的输入，这样前一层的输出可以直接馈送到下一层，这意味着输入特性的数量等于输出特性的数量，使得输入信息可以进行多层级联。

$$F(X)_i = U\sigma(VX_i + b) + c \tag{8}$$

其中，X_i 表示输入集的第 i 个向量，$F(X)_i$ 表示输出的第 i 个子向量，$\sigma(\ )$ 是 ReLU 激活函数，V 和 U 是可学习的权重矩阵，b 和 c 是偏差项

3.3. 基于问题引导注意力层的解码器

使用问题引导的注意力机制，如图 5，输入图像特征和文本特征，图像特征生成 query，文本特征生成 key 和 value，通过计算注意力权重自动忽略与给定问题无关的图像区域，而选择聚焦于重要的图像区域，利用问题表示寻找图片中与问题相关的区域。

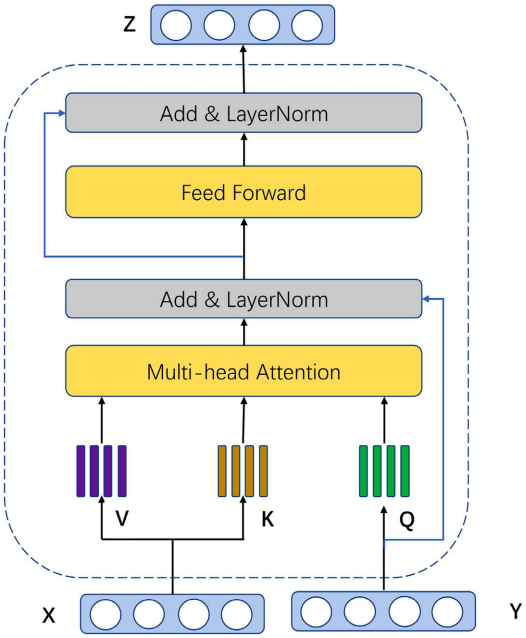

Figure 5. Guided attention layer
图 5. 引导注意力层

$$att(Q,K,V) = attention(W_q Y, W_k X, W_v X) \tag{9}$$

X 是输入的文本特征，Y 是输入的图像特征，W_q、W_k、W_v 需要训练的可学习矩阵。图像信息通过自注意力记忆层获得图像的关键区域信息，将文本信息和图像信息同时输入至问题引导的自注意力层，获得与关键词相关的图像区域。将自注意力记忆层与问题引导的注意力层进行堆叠构成解码端。

3.4. 答案预测

通过编码器与解码器的处理，输出的图像特征 Y_L 和文本特征 X_L 包含了问题信息和与问题相关的图像信息。因此，需要使用适当的融合机制结合这两种特征表示。为了获得对图像和问题共同理解的多模态融合特征，达到问题与图像信息进一步的语义对齐，将富含空间关系感知的图像特征和问题的特征向量进行多模态融合，多模态融合后的联合嵌入向量传递给多层感知器预测答案。在输出层，使用由一个全连接层(FC(d))-ReLU-Dropout(0.1)组成的线性层来得到融合特征，公式如下：

$$X' = \sum_{i=1}^{m} softmax\left(MLP\left(X_L\right)\right)x_i \tag{10}$$

$$Y' = \sum_{i=1}^{n} softmax\left(MLP\left(Y_L\right)\right)y_i \tag{11}$$

然后把 X' 和 Y' 放入如下的层中，归一化得到 Z：

$$Z = LayerNorm\left(W_x X' + W_y Y'\right) \tag{12}$$

其中，$W_x, W_y \in R^{d \times d_z}$，$d_z$ 是融合特征的共同维度。融合特征 Z 通过 Sigmod 函数投射到答案向量 $answer \in R^N$，这里 N 指的是训练集中出现频率最高的答案。损失函数使用 BCE 函数，在融合特征 Z 顶部训练一个 N 维分类器。

4. 实验

4.1. 数据集

VQA：在视觉问答任务中，VQA 是目前使用十分广泛的数据集，它的图像内容丰富多样，既有来源于基于微软 COCO 真实场景的数据集，又有来源于抽象场景的剪切画图片。数据集中与图像相关的问题/答案对是由人工注释生成的，并且鼓励人工注释者提供有趣且多样化的问题，与其他数据集相比，二值(即是/否)问题也采纳进来。这个数据集包括两个版本 VQA v1.0 和 VQA v2.0，与 VQA v1.0 版本相比，VQA v2.0 [20]版本的问题更加丰富，且减少了部分偏差，此外为了解决 VQA 数据集先验性问题，通过分别重新组织 VQA v1 和 VQA v2 数据集的训练和验证集来创建 VQA-CP v1 和 VQA-CP v2 数据集。问题的答案预测准确率计算如下式所示，依据可选答案在注释者给出答案中出现的次数，根据该评估指标，若模型预测的答案在该问题的 10 个可选答案中出现的次数大于等于 3，则 ACC 为 1，当预测答案在可选答案中出现的次数为 0、1、2 时，ACC 分别为 0、0.3 和 0.6。

$$Acc\left(ans\right) = \min\left\{\frac{\#\text{humans that said ans}}{3}, 1\right\} \tag{13}$$

数据集分为三部分：训练集(80,000 张图片和 444,000 个问答对)；验证集(40,000 张图片和 214,000 个问答对)；测试集(80,000 张图片和 448,000 个个问答对)。此外，还有两个测试子集，称为 test-dev 和 test-standard，用于在线评估模型性能。结果包括每种类型的三个精度：是/否(yes/no)，数字(number)，其他(other)和一个总体精度(overall)。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，即模型的评价指标。

4.2. 参数设置

问题文本输入 Glove 模型，得到的词嵌入序列通过 LSTM 网络生成问题特征。将文本序列划分为 14 个单词，不足的使用零填充补齐，得到的单词序列进行向量维度为 300 的词嵌入，最终输出的特征向量维度为 1024。图像特征通过预训练的 Faster R-CNN 模型提取，得到维度为 2048 的视觉特征。多头注意力头数设置为 8，训练过程中，训练批度设置为 64，使用 Adam 优化器，随机丢弃率为 0.1。

4.3. 实验结果及分析

4.3.1. 消融实验

在数据集 VQA v2.0 上对模型进行消融实验，以验证模型的有效性。为了比较各种设计对模型的影响，

本文设计了以下四种模型进行比较分析：

a) 基线模型：文本特征使用自注意力，图像特征使用自注意力和问题引导的注意力，将文本特征最后一层的输出作为问题引导的注意力层的输入。

b) 基线模型 + 自注意记忆层：文本和图像特征使用自注意记忆层，文本特征最后一层的输出作为问题引导的注意力层的输入。

c) 基线模型 + 交叉注意力结构：文本和图像特征使用自注意层，将文本特征的每一层都作为问题引导的注意力层的输入。

d) 基线模型 + 自注意记忆层 + 交叉注意力结构：文本特征使用自注意记忆层，将文本特征的每一层都作为问题引导的注意力层的输入。

实验结果如表 1 所示，可以得出结论，通过添加自注意记忆层，模型可以学到特征的多级信息，在预测精度上有所提升。在只使用交叉注意结构时，问题特征和图像特征进行多层次交互，可以更好地关注与问题最相关的图像区域。在添加自注意记忆层和交叉注意力结构后，模型准确率提高了 0.15%，文本特征包含多级信息，加强与图像特征的密集交互，使得特征中的多层次信息更好的关注图像特征的关键区域，增强了视觉特征表示。

Table 1. Comparison of different variant models
表 1. 不同变体模型比较

模型	验证集准确率%
基线模型	67.19
基线模型 + 自注意记忆层	67.26
基线模型 + 交叉注意力结构	67.21
基线模型 + 自注意记忆层 + 交叉注意力结构	67.34

4.3.2. 对比实验

模型在 VQA v2.0 的训练集上进行训练，并在验证集上评估模型的回答准确率，如图 6 所示为实验精确率变化。模型在 VQA v2.0 的训练集上进行训练，并在验证集上评估模型的回答准确率。可以看出在 13 轮时，模型在验证集上的精确率不再上升，最终采用训练轮次为 13 的模型，且在验证集上准确率为 67.34%。

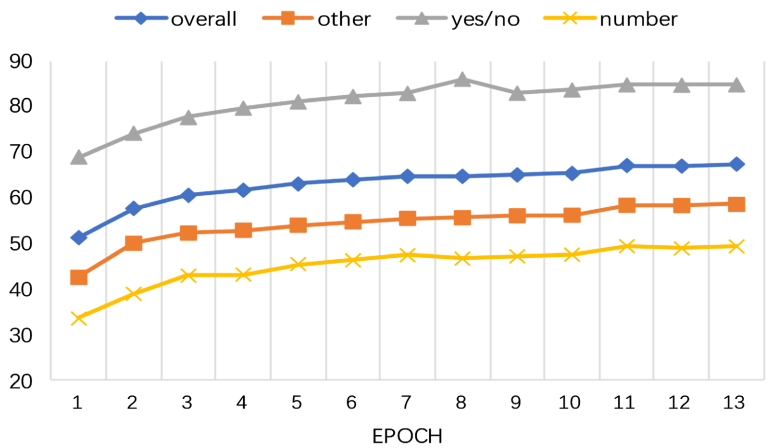

Figure 6. Changes in accuracy during model training
图 6. 模型训练过程中精确率变化

为了评价本文模型的性能，与其他模型进行了比较，表 2 显示了使用在线评估的 Test-dev 和 Test-std 的实验结果，从结果可以得出，本文提出的模型优于其他模型。采用本文模型在总体问题上的准确率为 70.85%，是/否问题的准确率为 86.91%，数值问题的准确率为 53.08%，其它类型的问题准确率为 61.03%。与 MCAN 模型相比，本文提出的模型可以有效利用问题特征的先验知识，更准确地关注图像中与问题最相关的区域，得到了高度相关的多模态特征，模型的精确率有了更大提升。

Table 2. Comparison of existing models in the VQA v2.0 dataset
表 2. 在 VQA v2.0 数据集上的模型比较

模型	Test-dev 准确率%				Test-std 准确率%
	yes/no	number	other	overall	overall
Language-only [20]	67.17	31.41	27.36	44.22	44.26
Bottom-up [8]	81.82	44.21	56.05	65.32	65.67
DA-NTN [21]	84.29	47.17	57.92	67.56	67.94
Mutan [22]	82.88	44.54	56.50	66.01	66.38
Scence GCN [23]	82.72	46.85	57.77	66.81	67.14
MuRel [12]	84.77	49.84	57.85	68.03	68.41
BAN + Counter [24]	85.42	54.04	60.52	70.04	70.35
MCAN [13]	86.82	53.26	60.72	70.63	70.90
本文	86.91	53.08	61.03	70.85	71.03

4.4. 可视化结果

为了与其他模型进行性能比较，我们在验证集上应用定性分析。如图 7 所示，给定一张图片以及针对这张图片的几个问题，GT (ground-truth)表示真实答案，然后是基线模型与本文模型的预测答案，红色表示模型预测的错误结果。从实验结果可以看出，本文提出的模型优于其他基线方法，能够有效地利用文本和图像特征的多层次信息，证明了模型的有效性。

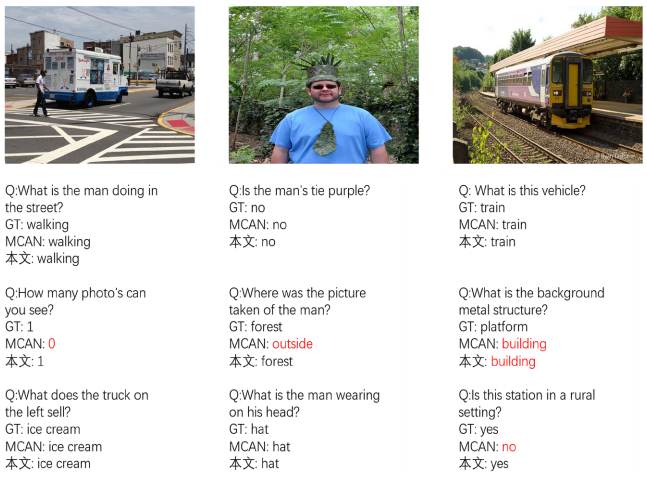

Figure 7. Comparison of model results
图 7. 模型结果比较

5. 结论

本文基于 MCAN 模型，结合交叉记忆注意模块，提高模型的细粒度识别能力，在编码端使用自注意记忆层，利用交叉结构使得解码端能充分利用多层次信息，获得与问题更相关的特征信息。在 VQA v2.0 数据集以及 test-dev 和 test-standard 验证集上的实验结果表明，该模型预测精确率有较大提高。后续将对输出分类器进行改进，将其中的特征融合方式进行改善，更大程度增强图像特征和文本特征的融合和推理能力。

基金项目

2021 年广东省重点建设学科科研能力提升项目(2021ZDJS095)；国家重点研发计划项目课题(2022YFA1602003)；五邑大学高层次人才科研启动项目(2019AL032)。

参考文献

[1] Stahlschmidt, S.R., Ulfenborg, B. and Synnergren, J. (2022) Multimodal Deep Learning for Biomedical Data fusion: A Review. *Briefings in Bioinformatics*, **23**, bbab569. https://doi.org/10.1093/bib/bbab569

[2] Antol, S., Agrawal, A., Lu, J., *et al*. (2015) VQA: Visual Question Answering. *Proceedings of the* 2015 *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 2425-2433. https://doi.org/10.1109/ICCV.2015.279

[3] Sharma, H. and Jalal, A.S. (2021) A Survey of Methods, Datasets and Evaluation Metrics for Visual Question Answering. *Image and Vision Computing*, **116**, Article ID: 104327. https://doi.org/10.1016/j.imavis.2021.104327

[4] Vu, M.H., Löfstedt, T., Nyholm, T., *et al*. (2020) A Question-Centric Model for Visual Question Answering in Medical Imaging. *IEEE Transactions on Medical Imaging*, **39**, 2856-2868. https://doi.org/10.1109/TMI.2020.2978284

[5] Zhang, D., Cao, R. and Wu, S. (2019) Information Fusion in Visual Question Answering: A Survey. *Information Fusion*, **52**, 268-280. https://doi.org/10.1016/j.inffus.2019.03.005

[6] Albawi, S., Mohammed, T.A. and Al-Zawi, S. (2017) Understanding of a Convolutional Neural Network. *Proceedings of* 2017 *International Conference on Engineering and Technology* (*ICET*), Antalya, 21-23 August 2017, 1-6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

[7] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Anderson, P., He, X., Buehler, C., *et al*. (2018) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *Proceedings of the* 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Lake City, 18-23 June 2018, 6077-6086. https://doi.org/10.1109/CVPR.2018.00636

[9] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[10] Nam, H., Ha, J.W. and Kim, J. (2017) Dual Attention Networks for Multimodal Reasoning and Matching. *Proceedings of the* 2017 30*th IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 2156-2164. https://doi.org/10.1109/CVPR.2017.232

[11] Yu, Z., Yu, J., Fan, J. and Tao, D. (2017) Multi-Modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. *Proceedings of the* 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 1839-1848. https://doi.org/10.1109/ICCV.2017.202

[12] Cadene, R., Ben-Younes, H., Cord, M. and Thome, N. (2019) MUREL: Multimodal Relational Reasoning for Visual Question Answering. *Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 1989-1998, https://doi.org/10.1109/CVPR.2019.00209

[13] Yu, Z., Yu, J., Cui, Y., *et al*. (2019) Deep Modular Co-Attention Networks for Visual Question Answering. *Proceedings of the* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 6281-6290. https://doi.org/10.1109/CVPR.2019.00644

[14] Chen, C., Han, D. and Wang, J. (2020) Multimodal Encoder-Decoder Attention Networks for Visual Question Answering. *IEEE Access*, **8**, 35662-35671. https://doi.org/10.1109/ACCESS.2020.2975093

[15] Guo, W., Zhang, Y. and Yang, J., *et al*. (2021) Re-Attention for Visual Question Answering. *IEEE Transactions on*

Image Processing, **30**, 6730-6743. https://doi.org/10.1109/TIP.2021.3097180

[16] Liu, Y., Zhang, X., Zhang, Q., *et al*. (2021) Dual Self-Attention with Co-Attention Networks for Visual Question Answering. *Pattern Recognition*, **117**, Article ID: 107956. https://doi.org/10.1016/j.patcog.2021.107956

[17] Sharma, H. and Jalal, A.S. (2022) An Improved Attention and Hybrid Optimization Technique for Visual Question Answering. *Neural Processing Letters*, **54**, 709-730. https://doi.org/10.1007/s11063-021-10655-y

[18] Cornia, M., Stefanini, M., Baraldi, L., *et al*. (2020) Meshed-Memory Transformer for Image Captioning. *Proceedings of the* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 10578-10587. https://doi.org/10.1109/CVPR42600.2020.01059

[19] Pennington, J., Socher, R. and Manning, C.D. (2014) Glove: Global Vectors for Word Representation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 25-29 October 2014, 1532-1543. https://doi.org/10.3115/v1/D14-1162

[20] Goyal, Y., *et al*. (2019) Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, **127**, 398-414. https://doi.org/10.1007/s11263-018-1116-0

[21] Bai, Y., Fu, J., Zhao, T. and Mei, T. (2018) Deep Attention Neural Tensor Network for Visual Question Answering. In: Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., Eds., *ECCV* 2018: *Computer Vision—ECCV* 2018, *Lecture Notes in Computer Science*, Vol. 11216, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-030-01258-8_2

[22] Ben-Younes, H., Cadene, R., Cord, M., *et al*. (2017) Mutan: Multimodal Tucker Fusion for Visual Question Answering. *Proceedings of the* 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2612-2620. https://doi.org/10.1109/ICCV.2017.285

[23] Yang, Z., Qin, Z., Yu, J. and Wan, T. (2020) Prior Visual Relationship Reasoning for Visual Question Answering. *Proceedings of* 2020 *IEEE International Conference on Image Processing* (*ICIP*), Abu Dhabi, 25-28 October 2020, 1411-1415. https://doi.org/10.1109/ICIP40778.2020.9190771

[24] Kim, J.H., Jun, J. and Zhang, B.T. (2018) Bilinear Attention Networks. *Proceedings of* 32*nd Conference on Neural Information Processing Systems* (*NIPS* 2018), Montréal, 3-8 December 2018, 31.