山西大学学报(自然科学版)49(2):263—271,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024104

基于图文理解增强的教科书视觉问答方法

胡景畅[1],强鹏鹏[1],谭红叶[1,2*],王宏宇[1],慕永利[3]
(1.山西大学 计算机与信息技术学院,山西 太原 030006;
2.计算智能与中文信息处理教育部重点实验室,山西 太原 030006;
3.智林信息技术股份有限公司,山西 太原 030000)

摘 要:教科书视觉问答是智慧教育领域的一项多模态任务,需要深度理解教科书图像、文本和问题以推理正确答案。然而,现有通用领域视觉问答方法在这一任务中表现不佳。主要原因是:首先,这些方法仅能简单地识别物体属性,缺乏学科信息,且容易受到与问题无关的冗余信息干扰;其次,这些方法难以捕捉文本关键信息。针对上述问题,提出基于图文理解增强的教科书视觉问答方法,主要包括3个模块:(1) 文本编码与理解:利用大语言模型提取问题关键词,并在文本中检索与问题关键词相关语句,来增强文本理解,消除冗余信息干扰。(2) 图像编码与描述:根据问题关键词,在图像描述中采用问题-图像注意力机制生成问题约束的细粒度图像描述语句,从而增强图像理解能力。(3) 答案预测:使用预训练视觉语言模型,将文本信息与视觉信息融合,以提高模型推理能力。实验结果表明:该文提出的方法能够较好地理解教科书图文信息,有效提升答案预测准确率。在测试集和验证集上准确率分别提高了1.82%和1.72%。

关键词:视觉问答;智慧教育;图像描述;图文理解增强
中图分类号:TP391.41 **文献标志码**:A **文章编号**:0253-2395(2026)02-0263-09

Enhancing Image and Text Comprehension for Textbook Visual Question Answering

HU Jingchang[1], QIANG Pengpeng[1], TAN Hongye[1,2*], WANG Hongyu[1], MU Yongli[3]
(1. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China;*
2. *Key Laboratory of Ministry of Education Intelligence and Chinese Information, Shanxi University, Taiyuan* 030006*, China;*
3. *Zhilin Information Technology Co, Ltd, Taiyuan* 030000*, China*)

Abstract: Textbook Visual Question Answering is a multi-modal task in the field of smart education that requires a deep understanding of textbook images, text, and questions to infer the correct answers. However, existing generic Visual Question Answering methods perform poorly in this task. The main reasons are as follows: Firstly, these methods can only simply recognize object attributes, lack disciplinary information, and are susceptible to interference from redundant information unrelated to the questions. Secondly, they struggle to capture key information in the texts. To solve these problems, a textbook visual question-answering method based on image description enhancement is proposed, which mainly includes three modules: (1) Text encoding and understanding: Utilizing large language models to extract keywords from questions and retrieve relevant statements in the text related to the question keywords to enhance text understanding and eliminate interference from redundant informations. (2) Image encoding and description:

收稿日期:2024-01-10;**修回日期**:2024-06-14
基金项目:国家自然科学基金(62076155);太原市小店区-山西大学产学研合作项目“短答案自动评分技术在综合评价系统中的推广与应用”(202301S06)
作者简介:胡景畅(1997—),男,山西省晋中人,硕士研究生,研究方向为自然语言处理、视觉问答。E-mail:hjc_kk@163.com
* **通信作者**:谭红叶(TAN Hongye),E-mail:hytan_2006@126.com
引文格式:胡景畅,强鹏鹏,谭红叶,等.基于图文理解增强的教科书视觉问答方法[J].山西大学学报(自然科学版),2026,49(2):263-271. DOI:10.13451/j.sxu.ns.2024104.

Employing a question-image attention mechanism in image descriptions to generate fine-grained image description statements constrained by questions based on question keywords, thereby enhancing image understanding ability. (3) Answer prediction: using a pre-trained visual-language model to fuse text information with visual information to improve the model's reasoning ability. Experimental results on relevant datasets demonstrate that the proposed method effectively improves the understanding of textbook information, thereby enhancing answer prediction accuracy. The accuracy of the test set and the verification set was improved by 1.82% and 1.72%, respectively.

Key words: visual question answering; intelligent education; image caption; image-text comprehension enhancement

0 引言

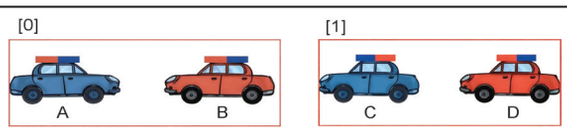
随着全球信息化浪潮的兴起,智慧教育对传统的教育思维、教育理念、教学模式、教学内容与方法等产生深远的影响,推动了教育形式和学习方式的巨大变革。

教科书视觉问答任务(Textbook Question Answering, TQA)[1]是智慧教育领域的一项多模态任务,需要模型充分理解图像、上下文、问题的前提下,预测问题答案,属于视觉问答(Visual Question Answering, VQA)[2-4]研究范畴。目前,VQA主要基于深度学习方法,如VISUABERT[5]和VILBERT(Vision-and-Language BERT)[6]等。虽然这些方法在通用领域简单VQA任务上(对图像中物体形状、种类、数量等简单属性的问题)表现优异,但在TQA上性能较差,主要原因是:(1)TQA任务更为复杂,文本与图片均蕴含丰富的学科信息,要求模型具有更强的图文交互理解能力;(2)VQA任务的文本信息一般只涉及问题,但TQA不仅包含问题,还包含比较长的描述相关知识的段落信息,需要模型具备较好的长文本关键信息捕捉能力。如表1所示示例,图片[0]与图片[1]均包含两辆顶部放置磁铁的不同颜色的小汽车(蓝色和红色),文本Passage部分描述了与磁铁相关的物理知识。模型需要准确识别图中小汽车中磁铁极性信息,并结合Passage描述的磁极作用原理知识,进行图文交互理解与推理预测出正确答案。

目前已有研究者将图像描述应用在通用领域的VQA任务中以实现较好的图文交互,但针对该任务的图像描述只能对图中物体的数量、种类、色彩等简单属性进行描述。因此如何生成蕴含学科知识、适应TQA任务的图像描述仍是一个挑战。此外,模型如何捕捉TQA任务中长文本关键信息并进行图文信息融合是另外一个挑战。

表1 教科书视觉问答示例

Table 1 Example of textbook visual question answer

Image	
Question	Choose the correct statement about [0] and [1] based on the above information.
Answer Choices	A. [0] will repel, [1] will attract; B. [0] will attract, [1] will repel; C. both [0]and [1] will attract; D. both [0] and [1] will repel
Answer	B
Passage	Consider two cars carrying large magnets on top of them as per [0] and [1]. Red indicates north pole whereas blue indicates south pole. When you place the north pole of one magnet near the south pole of another magnet, they are attracted to one another. When you place like poles of two magnets near each other, they repel.

为了解决上述挑战,本文提出了基于图文理解增强的教科书视觉问答模型,提高TQA任务的性能。主要包含三个模块:(1)文本理解与编码模块:利用具备丰富知识和强大推理能力的大语言模型提取问题关键词,并在文本中检索相关信息,然后再基于这些问题关键词提取上下文相关语句。(2)图像理解与编码模块:在图像描述模型中,采用问题-图像注意力机制生成与问题高度相关且蕴含一定学科知识的图像描述来增强图像理解能力;(3)答案预测模块:将得到的文本信息(问题、上下文相关语句、图像描述)与视觉信息融合,输入预训练视

觉语言模型,进行视觉问答推理。

综上所述,本文的主要贡献可以总结如下:

(1)提出了一种基于图文理解增强的TQA方法,通过问题约束的细粒度图像描述,使模型能够理解图像中物体蕴含的学科知识,提高了模型在TQA任务中的图文理解能力与推理能力。

(2)利用具有丰富知识的大语言模型(ChatGPT)理解问题与上下文,并抽取问题关键词以及上下文相关语句,以消除冗余信息的干扰。在图像描述中,使用注意力机制生成问题相关的图像描述语句,增强了图文交互能力。

(3)在Visuo-linguistic question answering(VLQA)[7]数据集上进行实验,证明该方法可以有效提高文本与视觉模态的信息理解能力。

1 相关工作

1.1 视觉问答

视觉问答是涉及自然语言处理和计算机视觉的一个多模态任务。自2015年第一个VQA数据集被提出后,各种VQA方法相继涌现。早期研究者为解决多模态信息融合问题,普遍采用拼接嵌入方法,将图像与文本两种信息通过简单机制(如串联、逐元素乘或逐元素加)组合,送入线性分类器或神经网络进行答案预测。例如,Antol等[2]利用卷积神经网络(Convolutional Neural Networks,CNNs)进行图像表示,采用长短期记忆网络(Long Short-term Memory,LSTM)进行文本表示,并利用元素乘法将二者联合嵌入,对答案进行分类。

近年来,基于注意力的VQA方法[6]成为主流。该方法利用局部特征并对其赋予不同权重,解决了全局特征带来的信息冗余问题。例如,Lu等[7]提出了一种多层协同注意力模型(Hierarchical Question-image Co-attention for Visual Question Answering,HIECOATT),在词汇、短语和问题三个层次上共同关注图像和问题。

尽管上述方法在VQA任务中表现良好,但在问题较为复杂且包含长文本信息的TQA任务上表现较差。

1.2 教科书视觉问答

教科书视觉问答[1]自提出以来就受到了广泛的关注。一些研究人员利用图方法关注推理过程。如:Li等[8]通过构建实体关系图(Contradiction Entity-relationship Graph,CERG)来理解空间分析规则。Kim等[9]提出了基于GCN(Graph Convolution Network)[10]从上下文图中提取知识特征作为相关问题背景知识,并在模型训练中引入一种新的自监督学习过程解决特定领域的问题。Ma等[11]提出一种细粒度的关系抽取方法,对所构造图的结点进行推理。但是上述方法忽略了多模态输入表征的巨大潜力。

Gomez-Perez和Ortega[12]利用预训练Transformer[13]进行文本编码,并利用自下而上和自上而下的注意力进行多模态融合,显著提高了性能。

但是直接从图像与文本中学习得到的特征表示不足以回答TQA任务中的问题,该任务还需要学科知识来增强特征表示,同时增强图文交互能力,以提高模型在理解图像和文本的能力,才能正确回答问题。

1.3 图像描述

图像描述[14]是根据一张或多张图片,生成一句或一段描述图片的文本。近年来,基于深度学习的方法由于可直接从大量数据中学习图像到语句的映射,可以生成更加准确的描述,因此得到广泛应用。例如:Mao等[15]采用基于CNN-RNN(Convolutional Neural Network-Recurrent Neural Network)的编码器-解码器框架,其中CNN[16]负责图像识别,并输入给RNN[17]成文本描述。但是编码器-解码器方法在处理固定长度向量时通过增加一个上下文向量来对每个输入进行解码,以增强图像区域和单词的相关性,获取更多图像语义细节。Pan等[18]提出一种统一的X-Linear注意力网络,充分利用双线性池对视觉信息进行集中利用。Guo等[19]提出几何感知自注意力,有效地考虑了图像中对象之间的相关几何关系。

目前图像描述的方法[20]已被应用到VQA中,生成的语句基本为对图像中物体属性的描述,忽略了图像与文本的联系,使得图像描述直接应用在TQA任务中效果较差。

2 方法

2.1 模型总体框架

教科书视觉问答可以形式化定义为四元组(T,I,Q,A),给定上下文T、图片I、问题Q以及与问题Q相关的候选答案集A,目标是根据公

式(1)预测问题答案$\hat{a}\in A$:

$$\hat{a}=\arg\max_{a\in A} P(a|T,I,Q)。\tag{1}$$

图1展示了本文模型的总体框架,模型主要由三个模块组成:(1)文本编码与理解模块,提取问题关键词,并根据问题关键词提取文本中相关语句。(2)图像编码与描述模块,根据问题关键词,引导模型生成与问题相关的图像描述,为模型推理问题提供了丰富的图像信息。(3)答案预测模块,将图片、图像描述、问题、选项和文本信息输入到VL-BERT模型中,预测正确答案。

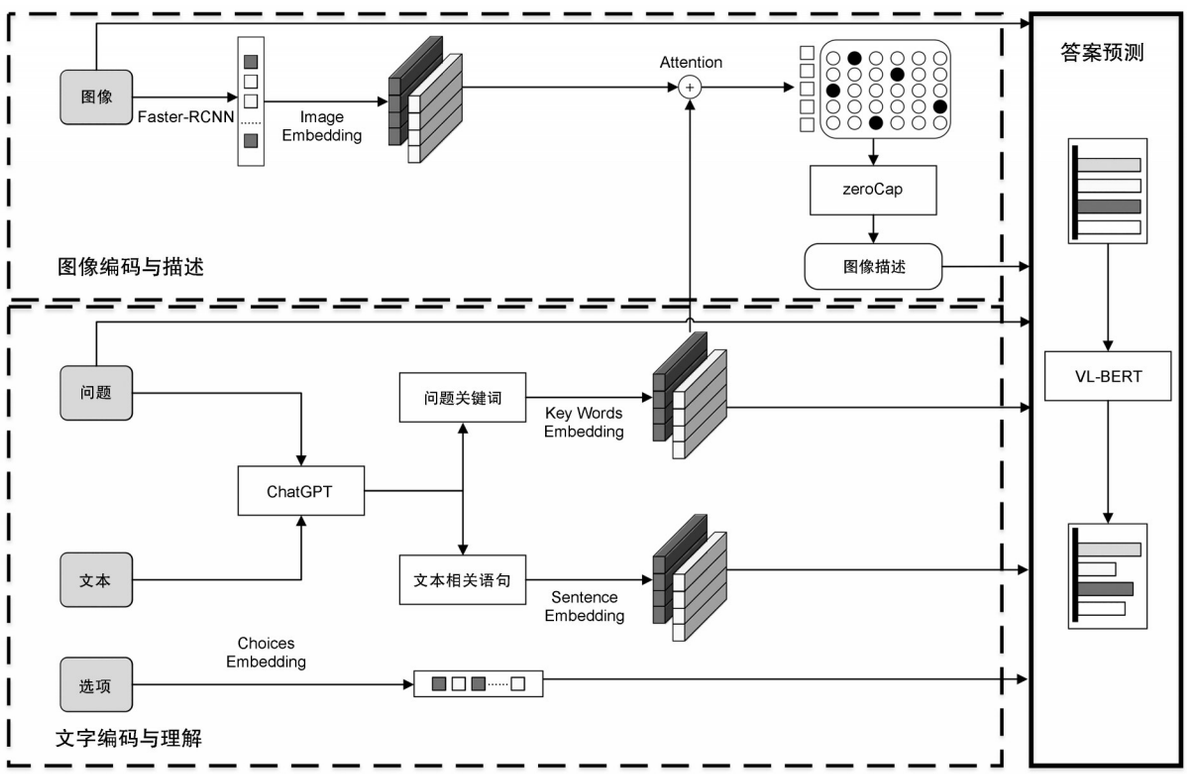


图1 模型框架图

Fig. 1 Framework of model

2.2 文本编码与理解

提取问题关键词和文本相关语句。给定问题$Q=[q_1,q_2,\cdots,q_i,\cdots,q_n]$、文本$T=[t_1,t_2,\cdots,t_j,\cdots,t_m]$,其中$q_i$表示问题中的第$i$个词,其中$t_j$表示文本中的第$j$句话。将问题$Q$和文本$T$输入大语言模型ChatGPT[21]中,通过提示,让其提取问题Q中的关键词$K=[k_1,k_2,\cdots,k_i,\cdots,k_k]\in Q$,以及上下文$T$的相关语句$L=[l_1,l_2,\cdots,l_i,\cdots,l_l]$,其公式可以定义为:

$$K,L=\text{ChatGPT}(Q,T)。\tag{2}$$

采用预训练语言模型BERT[22]对上下文T、问题Q、选项A问题、关键词K以及相关语句L进行编码,以获得特征表示。计算公式如公式(3)—(7)所示:

$$W_Q=\text{BERT}(Q),\tag{3}$$

$$W_T=\text{BERT}(T),\tag{4}$$

$$W_A=\text{BERT}(A),\tag{5}$$

$$W_K=\text{BERT}(K),\tag{6}$$

$$W_L=\text{BERT}(L),\tag{7}$$

其中W_Q、W_T、W_A是利用Bert编码将问题Q、上下文T和选项A转为词嵌入的结果。$W_Q=\{w_1,w_2,\cdots,w_n\}\in\mathbf{R}^{n*D}$,$W_T=\{w_1,w_2,\cdots,w_m\}\in\mathbf{R}^{m*D}$,$W_A=\{w_1,w_2,\cdots,w_l\}\in\mathbf{R}^{l*D}$,$W_K=\{w_1,w_2,\cdots,w_s\}\in\mathbf{R}^{s*D}$,$W_L=\{w_1,w_2,\cdots,w_j\}\in\mathbf{R}^{j*D}\mathbf{R}^{j*D}$其中$n$、$m$、$l$、$s$、$j$表示$Q$、$T$、$A$、$K$、$L$中token的最大个数,$D$是Bert隐层的固定维数,*表示卷积。

2.3 图像编码与理解

此模块的主要目的是生成图像描述语句,丰富文本内容,增强对图像的理解。本文使用Faster-RCNN[23]提取图像特征,使用改进的ZeroCaption[24]模型进行图像描述。

图像特征提取,使用Faster-RCNN提取图

像I的区域特征,用被检测物体的特征作为图片的表征,其可以表示为:

$$B=\text{FasterRCNN}(I)。\tag{8}$$

每张图片提取到的RoI组成一个特征矩阵$B\in\mathbf{R}^{N_j*D_j}$。其中N_j表示RoI区域个数,D_j表示特征维度。

通过基于问题引导的注意力机制将问题特征与图片特征融合,使模型的注意力集中在问题关注的物体上,在保证图片中物体完整性的前提下摒弃图片背景等冗余信息的干扰,关注问题最相关的物体。

具体过程如下:将问题Q关键词的编码结果$W_K=[w_{k1},w_{k2},\cdots,w_{ki},\cdots,w_{ks}]$与图像特征$B=[b_1,b_2,\cdots,b_i,\cdots,b_t]$融合,得到融合特征$F_w=[f_{w1},f_{w2},\cdots,f_{wi},\cdots,f_{wn}]$,其表示如公式(9)所示:

$$f_{wi}=\text{Attention}(w_{ki},b_i)。\tag{9}$$

在模型中使用基于Transformer的Language Model(LM)从初始提示(问题关键词)开始,不断推断下一个单词,最后输出一句图像描述。LM利用Transformer中的注意力机制生成单词序列。序列生成过程可以形式化为公式(10):

$$X_{i+1}=\text{LM}(x_i,[(U_j^i,V_j^i)]_{j<i,1<l<r}),\tag{10}$$

其中x_i是生成句子中第i个单词,U_j^i,V_j^i是上下文中的键值和token值,r表示Transformer中的索引层数。

综上所述,使用ZeroCaption进行推行图像描述,具体实现过程如图2所示,其可以表示为公式(11):

$$C=\text{ZeroCaption}(f_{wi},k_i)。\tag{11}$$

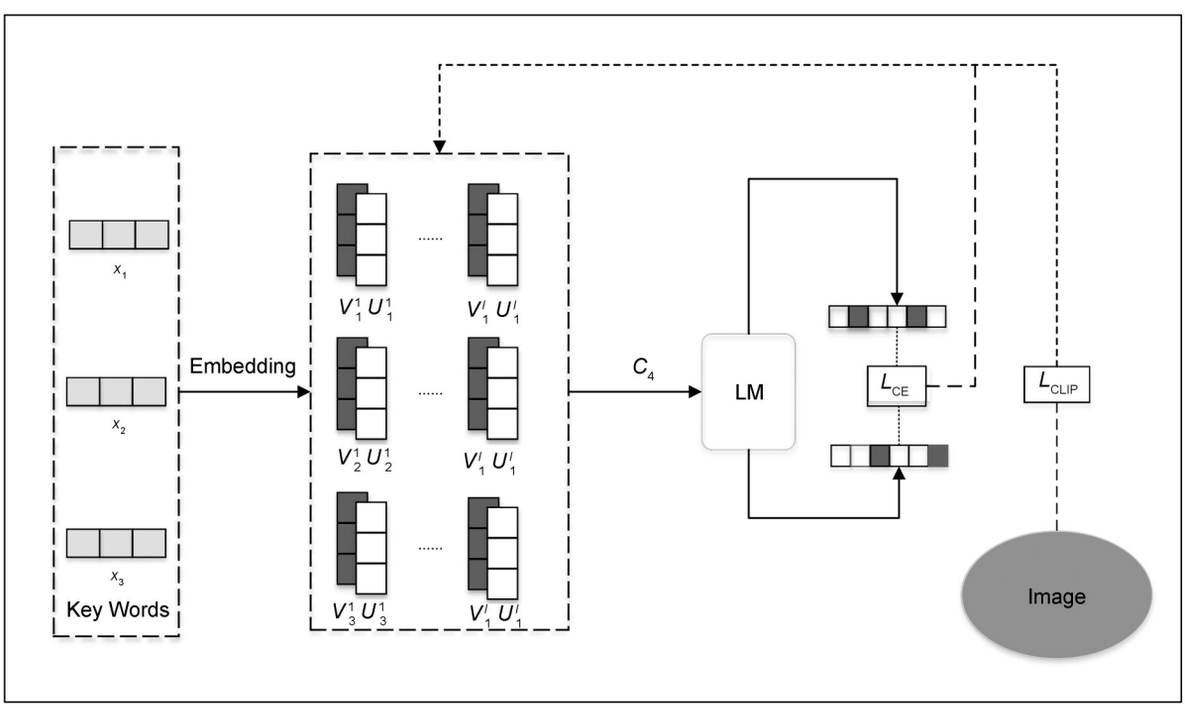


图2 图像描述模块

Fig. 2 Image description module

使用CLIP(Contrastive Language-image Pre-training)损失约束模型生成与关键词意思相关的语句。使用CLIP损失选取最高CLIP匹配分数的单词。CLIP损失即模型得到的下一个token x_{i+1}的CLIP潜在分布与目标分布之间的交叉熵损失:

$$L_{\text{CLIP}}=\text{CE}(p_i,x_{i+1})。\tag{12}$$

通过优化生成图像描述$C=\{c_i|i=1,2,3,\cdots,n\}$,$n$表示图像描述的长度。

2.4 答案预测

模型通过对问题和图像的编码、多模态融合、上下文感知等步骤,将视觉和语言信息结合起来,输入到预训练视觉语言模型VL-BERT中,得到预测答案:

$$Y=\text{VLBERT}'(B_I,W_Q,W_T,W_C,W_A),\tag{13}$$

其中B_I,W_Q,W_T,W_C,W_A分别表示图像,问题、

上下文、图像描述以及选项的编码后的结果。

3 实验

3.1 数据集

我们在VLQA数据集上进行了实验。该数据集由多种在线资源(书籍、百科全书、网络爬虫、现有数据集、标准化测试等)策划而成,有9 267条图像-文本对数据。形式为真/假、文本选择题、图像选择题和序列预测。数据集情况如表2—表3所示。

表2 数据集划分信息

Table 2 Information of dataset partitioning

类型	数量
训练集	7 413
验证集	927
测试集	927

表3 数据集题目信息

Table 3 Information of dataset title

类型	每题候选答案数	题目总数
真/假	2	1 272
文本选择	4	4 676
序列预测	4	1 088
图像选择	4	1 172
图像选择	2	1 088

3.2 基线

Random Baseline VLQA数据集包含4项和2项多项选择题,其中每个答案选择可能以25%和50%的概率被选中。随机基线的性能为31.36%。

VL-BERT[25]单流统一模型。主干是(多模态)Transformer注意力模块,将视觉和语言嵌入特征作为输入,其中,每个输入的token要么是输入句子中的一个单词,要么是输入图像中的感兴趣区域(RoI),再加上某些特殊token(CLS\SEP\IMG\END)来消除不同输入格式的歧义。通过堆叠多层多模态Transformer注意力模块,生成的表征在聚集和对齐视觉语言信息方面具有丰富的能力。

VisualBERT[8]包含了一组Transformer层,借助自注意力把输入一段文本中的元素和一张相关的输入图像中的区域隐式地对齐起来。可以在没有任何显式监督的情况下建立语言元素和图像中区域之间的联系,而且也对句法关系和追踪(根据描述建立动词和图像区域之间的关系)有一定的敏感性。

VilBERT视觉语言联合表征模型,扩展了BERT语言模型,以联合表征文本和图像。引入独立的视觉和语言处理流,再通过共同注意力Transformers层进行交互。这种结构可以适应每个模态的不同处理需求,并在不同的表征深度(Transformer层的数量)下提供模态之间的交互。

LXMERT(Language X Multimodal ERiT)[26]是一个大规模Transformer模型,它有三个编码器:对象关系编码器、语言编码器和跨模态编码器。利用掩码语言建模、掩码目标预测、跨模态匹配、图像问答等预训练任务,在大量"图像句子对"数据集上对模型进行预训练。

BLIP(Bootstrapped Language-image Pretraining)[27]是一种encoder-decoder混合多模态结构。可以灵活地在视觉理解任务上和生成任务上面迁移,BLIP通过自展标注(bootstrapping the captions),可以有效地利用带有噪声的web数据,其中标注器(captioner)生成标注,过滤器(filter)去除有噪声的标注。

BLIP-2[28]利用预训练的视觉模型和语言模型来提升多模态效果和降低训练成本。预训练的视觉模型能够提供高质量的视觉表征,预训练的语言模型则提供了强大的语言生成能力依赖。模型由预训练的Image Encoder,预训练的Large Language Model,和一个可学习的Q-Former组成。

3.3 实验设置

实验在NVIDIA A100下进行模型训练和调试,采用VL-Bert预训练模型为基线。由于VL-Bert通过向Bert模型添加视觉信息输入来完成多模态任务。因此,本文将参数初始化为与原始Bert相同,Batch size为16,Epochs为10,Learning Rate为1×10^{-5}。

通过调用接口的形式,将问题与上下文输入到ChatGPT提取问题的关键词,以及抽取上下文语句。

图像描述在ZeroCaption模型中使用问题引导的注意力机制完成。模型采用FasterRCNN来提取RoIs。视觉内容嵌入由FasterRCNN+Resnet-101产生,在视觉基因组上预先训练的参数初始化,用于对象检测。具体来说,每幅图像最多选择100个检测分数高于0.5的RoI。最少从一幅图像中选择10个RoI。

为了将VLQA数据适应VQA风格模型,将所有图像组合成一个(在多个图像的情况下)作为单一的视觉输入,并将段落和问题连接为单一的语言输入。

3.4 实验结果与分析

本文提出的图文理解增强的研究方法,结合了一个额外的文本信息(图片描述),增加了图文交互。图像描述使模型能够理解图像的内容。通过生成文字描述,可以结合文本内容深层次理解图像中语义信息,使图像与学科信息相关联。有利于提升模型对图像的感知和理解能力。这对教科书视觉问答任务提供了很好的图像理解能力。

表4给出了本文方法在VLQA验证集和测试集上的实验结果。从结果观察可以得到,本文的方法在测试集上准确率比基线VL-BERT高出约1.82%。在验证集上高出1.72%。

表4 不同模型的实验结果

Table 4 Experimental results of different models

模型	测试集准确率/%	验证集准确率/%
Random	31.36	31.36
VL-BERT	35.92	34.60
VisualBERT	33.17	34.17
ViLBERT	34.70	35.25
LXMERT	36.41	37.82
BLIP	35.75	35.73
BLIP-2	36.83	35.38
Ours	**37.74**	**36.32**

3.5 消融实验

表5为消融实验的结果,其中"-Image caption"表示模型去掉图像描述,"-Attention"表示去掉问题引导的注意力机制,"-ChatGPT"表示去掉大语言模型。从数据可以看出,ChatGPT对模型的影响最大,得到的准确率明显降低。ChatGPT在本实验中提取问题关键词以及理解上下文,影响图像描述生成语句的质量,是本实验的关键环节。同时,可以看到在去掉图像描述、注意力机制之后模型的性能有所下降,这也证明了融入图像描述以及注意力机制的有效性。

表5 消融实验结果

Table 5 Results of ablation test

模型	测试集准确率/%	验证集准确率/%
-Image caption -Attention	35.92	34.60
-Image caption	36.21	35.93
-Attention	36.71	36.26
-ChatGPT	36.05	35.14
Ours	**37.74**	**36.32**

4 样例分析

本文对错误样例进行了分析。如表6所示,其中Q表示给定的问题,C是利用ZeroCaption生成的图像描述,选项A中红色标注的代表真实答案,P代表预测答案。一方面,在预测正确的案例中可以观察到,类似于表6d中这种不能由图像和问题直接推导出的答案得到了正确的预测,图像描述中出现的"oven","restaurant"关键词信息帮助模型预测答案。另一方面,样例b和c的问题模型均回答错误。主要原因是模型欠缺学科知识,导致图像描述不够准确,未能对答案推理起到作用。具体的,c示例的图像描述中,描述为"容器中装有不同液体",而问题实际上关注的是"密度"。同样,在b的情况中,题目要求"确定呼吸时肋骨的运动方向",而生成的图像描述为"人体结构的示意图",显然这样描述并未对解答问题提供实质性帮助。模型在应对此类问题时,需要依赖相关的知识背景进行正确的推理。

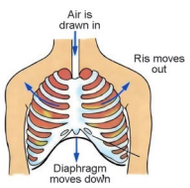
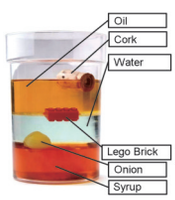
表6 样例分析示例

Table 6 Examples of sample analysis

图像(I)	问题(Q)	图像描述(C)	选项(A)	预测答案(P)
a 	Which of the following is a correct pair of balanced forces for a rescue helicopter?	A picture of a helicopter with four directions,lift、drag 、gravity、friction	A. lift and thrust B. drag and gravity C. lift and gravity D. friction and gravity	C

续表6 样例分析示例

Continued Table 6 Examples of sample analysis

图像(*I*)	问题(*Q*)	图像描述(*C*)	选项(*A*)	预测答案(*P*)
b 	Which of the following is not correct about Exhalation?	A diagram of the structure of the human body	A. Ribs move inside. B. Diaphragm moves up C. Air is drawn out D. Ribs move outside.	A
c 	Based on [0]. which of the following is true for density of water d(W) and density of Lego brick d(B)?	Image of a container containing liquids of different densities	A. d(W)>d(B) B. d(W)<(B) C. d(W)=(B) D. Cannot be answered from [0]	B
d [0]  [1] 	Choose the correct order of the events I to IV above to prepare Lasagna.	[0]:the image describes about Lasagna ersatz potatoes that are cooked in the oven at the restaurant. [1]:The key word is lasagna, an image of a dish in an oven.	A. III—I—II—IV B. III—II—I—IV V. II—III—IV—I D. II—IV—III—I	D

5 总结

在本文中,我们提出了一种增强图文理解的教科书视觉问答方法,提升了模型的推理能力。该方法首先利用大语言模型精准地提取问题关键词,并根据不同的问题在文本中检索相关信息。接着,通过问题引导的注意力推理机制,引导模型生成与关键词紧密相关的图像描述。最终,将提取到的信息与图像进行深度融合,实现了对教科书问题的精准推理。实验结果表明,本文提出的方法能够较好地理解教科书图文信息,并有效提升了教科书视觉问答的准确率。然而,我们也发现模型在捕捉图像中细粒度信息方面尚存在不足,且缺乏相关领域的专业知识。为了进一步优化模型性能,下一步我们将重点关注如何让模型更准确地捕捉图像的细粒度区域。同时,我们还将探索如何通过引入额外知识为不同的细粒度类别提供垂直领域的专属知识信息。通过这些改进措施,我们提高模型在教科书视觉问答领任务上的准确率。

参考文献:

[1] KEMBHAVI A, SEO M, SCHWENK D, *et al*. Are you Smarter than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2017: 5376-5384. DOI: 10.1109/CVPR.2017.571.

[2] ANTOL S, AGRAWAL A, LU J S, *et al*. VQA: Visual Question Answering[C]//2015 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2015: 2425-2433. DOI: 10.1109/ICCV.2015.279.

[3] CAO Q X, LIANG X D, LI B L, *et al*. Interpretable Visual Question Answering by Reasoning on Dependency Trees[J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, **43** (3): 887-901. DOI: 10.1109/TPAMI.2019.2943456.

[4] YU J, ZHU Z H, WANG Y J, *et al*. Cross-modal Knowledge Reasoning for Knowledge-based Visual Question Answering[J]. *Pattern Recognit*, 2020, **108**: 107563. DOI: 10.1016/j.patcog.2020.107563.

[5] Li L H, YATSKAR M, YIN D, *et al*. VisualBERT: A Simple and Performant Baseline for Vision and Language [EB/OL]. (2019-08-09) [2024-01-01]. https://arxiv.org/abs/1908.03557.

[6] LU J, BATRA D, PARIKH D, *et al*. ViLBERT: Pretraining Task-agnostic Visiolinguistic Representations for Vision-and-language Tasks[EB/OL]. (2019-08-06) [2024-01-01]. https://arxiv.org/abs/1908.02265.

[7] SAMPAT S K, YANG Y, BARAL C. Visuo-linguistic Question Answering (VLQA) Challenge[EB/OL]. (2020-11-18) [2024-01-01]. https://arxiv.org/abs/2005.00330.

[8] LI J, SU H, ZHU J, *et al*. Textbook Question Answering

Under Instructor Guidance with Memory Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 3655−3663.

[9] KIM D, KIM S, KWAK N. Textbook Question Answering with Multi-modal Context Graph Understanding and Self-supervised Open-set Comprehension[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3568−3584. DOI: 10.18653/v1/p19-1347.

[10] KIPF T N, WELLING M. Semi−supervised classification with graph convolutional networks[C]//International Conference on Learning Representations. Toulon: ICLR, 2017.

[11] MA J, LIU J, WANG Y X, *et al*. Relation-aware Fine-grained Reasoning Network for Textbook Question Answering[J]. *IEEE Trans Neural Netw Learn Syst*, 2023, **34**(1): 15−27. DOI: 10.1109/TNNLS.2021.3089140.

[12] GOMEZ-PEREZ J M, ORTEGA R. ISAAQ - Mastering Textbook Questions with Pre-trained Transformers and Bottom-up and Top-down Attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 5469−547. DOI: 10.18653/v1/2020.emnlp-main.441.

[13] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is All You Need[J]. *NeurIPS*, 2017, 30. DOI: 10.48550/arXiv.1706.03762.

[14] VINYALS O, TOSHEV A, BENGIO S, *et al*. Show and Tell: a Neural Image Caption Generator[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 3156−3164. DOI: 10.1109/CVPR.2015.7298935.

[15] MAO J, XU W, YANG Y, *et al*. Deep Captioning with Multimodal Recurrent Neural Networks[C]//International Conference on Learning Representations. San Diego: ICLR, 2015.

[16] SILVER D, HUANG A, MADDISON C J, *et al*. Mastering the Game of Go with Deep Neural Networks and Tree Search[J]. *Nature*, 2016, **529**(7587): 484−489. DOI: 10.1038/nature16961.

[17] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent Neural Network Regularization[C]//International Conference on Learning Representations. San Diego: ICLR, 2015.

[18] PAN Y W, YAO T, LI Y H, *et al*. X-linear Attention Networks for Image Captioning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 10968−10977. DOI: 10.1109/CVPR42600.2020.01098.

[19] GUO L T, LIU J, ZHU X X, *et al*. Normalized and Geometry-aware Self-attention Network for Image Captioning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 10324−10333. DOI: 10.1109/CVPR42600.2020.01034.

[20] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[21] WU T Y, HE S Z, LIU J P, *et al*. A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development[J]. *IEEE/CAA J Autom Sin*, 2023, **10**(5): 1122−1136. DOI: 10.1109/JAS.2023.123618.

[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2019: 4171-4186.

[23] GIRSHICK R. Fast R-CNN[J]. *Computer Science*, 2015. DOI:10.1109/ICCV.2015.169.

[24] TEWEL Y, SHALEV Y, SCHWARTZ I, *et al*. ZeroCap: Zero-shot Image-to-text Generation for Visual-semantic Arithmetic[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 17897−17907. DOI: 10.1109/CVPR52688.2022.01739.

[25] SU W, ZHU X, CAO Y, et al. VL-BERT: Pre-training of Generic Visual-linguistic Representations[C]//International Conference on Learning Representations. Addis Ababa: ICLR, 2020.

[26] TAN H, BANSAL M. LXMERT: Learning Cross-modality Encoder Representations from Transformers [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2019: 5099-5110.

[27] LI J, LI D, XIONG C, et al. BLIP: Bootstrapping Language-image Pre-training for Unified Vision-language Understanding and Generation[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 18887-18900.

[28] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping Language-image pre-training with Frozen Image Encoders and Large Language Models[C]//International Conference on Machine Learning. Honolulu: PMLR, 2023: 23018-23040.