软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(8):2522−2544 [doi: 10.13328/j.cnki.jos.006215]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

视觉问答研究综述*

包希港, 周春来, 肖克晶, 覃 飙

(中国人民大学 信息学院,北京 100872)

通讯作者: 覃飙, E-mail: qinbiao@ruc.edu.cn

摘 要: 视觉问答是计算机视觉领域和自然语言处理领域的交叉方向,近年来受到了广泛关注.在视觉问答任务中,算法需要回答基于特定图片(或视频)的问题.自 2014 年第一个视觉问答数据集发布以来,若干大规模数据集在近 5 年内被陆续发布,并有大量算法在此基础上被提出.已有的综述性研究重点针对视觉问答任务的发展进行了总结,但近年来,有研究发现,视觉问答模型强烈依赖语言偏见和数据集的分布,特别是自 VQA-CP 数据集发布以来,许多模型的效果大幅度下降.主要详细介绍近年来提出的算法以及发布的数据集,特别是讨论了算法在加强鲁棒性方面的研究.对视觉问答任务的算法进行分类总结,介绍了其动机、细节以及局限性.最后讨论了视觉问答任务的挑战及展望.

关键词: 视觉问答;交叉方向;语言偏见;数据集分布;鲁棒性

中图法分类号: TP18

中文引用格式: 包希港,周春来,肖克晶,覃飙.视觉问答研究综述.软件学报,2021,32(8):2522−2544. http://www.jos.org.cn/1000-9825/6215.htm

英文引用格式: Bao XG, Zhou CL, Xiao KJ, Qin B. Survey on visual question answering. Ruan Jian Xue Bao/Journal of Software, 2021,32(8):2522−2544 (in Chinese). http://www.jos.org.cn/1000-9825/6215.htm

Survey on Visual Question Answering

BAO Xi-Gang, ZHOU Chun-Lai, XIAO Ke-Jing, QIN Biao

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: Visual question answering (VQA) is an interdisciplinary direction in the field of computer vision and natural language processing. It has received extensive attention in recent years. In the visual question answering, the algorithm is required to answer questions based on specific pictures (or videos). Since the first visual question answering dataset was released in 2014, several large-scale datasets have been released in the past five years, and a large number of algorithms have been proposed based on them. Existing research has focused on the development of visual question answering, but in recent years, visual question answering has been found to rely heavily on language bias and the distribution of datasets, especially since the release of the VQA-CP dataset, the accuracy of many models has been greatly reduced. This paper mainly introduces the proposed algorithms and the released datasets in recent years, especially discusses the research of algorithms on strengthening the robustness. The algorithms of visual question answering are summarized and their motivation, details, and limitations are also introduced. Finally, the challenge and prospect of visual question answering are discussed.

Key words: visual question answering; interdisciplinary direction; language bias; distribution of datasets; robustness

视觉问答任务是人工智能领域一项具有挑战性的任务,其属于计算机视觉和自然语言处理的交叉方向.然而在此之前,计算机视觉和自然语言处理是分开发展的,在各自的领域取得了重大的进步.随着计算机视觉和深

* 基金项目: 国家自然科学基金(61772534, 61732006)
Foundation item: National Natural Science Foundation of China (61772534, 61732006)
收稿时间: 2020-07-09; 修改时间: 2020-10-02; 采用时间: 2020-11-23; jos 在线出版时间: 2021-01-15

度学习的不断发展,许多计算机视觉任务取得了巨大的进展,如图像分类[1,2]、物体检测[3,4]和动作识别[5,6].但是上述任务只需对图像进行感知,不需要对图像进行整体的理解和推理.图像字幕任务[7−9]首先将两个领域结合起来,利用图像和文本作为输入训练模型以描述图像中的内容.

文本问答系统[10,11]在自然语言处理领域已经有了广泛的研究,不论是科研界还是工业界都有众多成果涌现,如淘宝的智能客服.随着问答系统在自然语言处理领域的成功应用,有研究提出将问答系统应用至视觉领域.随着自媒体的不断发展,图片和视频的数据量爆炸性增长,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,成为了一个亟需解决的问题,视觉问答任务在这一背景下被提出.

视觉问答任务是以图像(或视频)和与图像(或视频)有关的文本问题的多模态信息作为计算机的输入,计算机根据图片得到问题的正确答案.本文的内容主要是对基于图片的视觉问答任务进行总结,如图 1 中所示.视觉问答任务如今分为开放式和多项选择形式两个子任务:开放式的视觉问答任务答案不确定,由计算机给出正确答案,答案通常是几个单词或者一个简单的短语;多项选择形式的视觉问答任务存在候选答案,计算机在已给定的候选答案中选择正确答案.视觉问答任务与其他计算机视觉任务相比更具有挑战性:视觉问答任务中要回答的问题是在运行时给出,需要处理视觉和文本的多模态信息,问题答案的形式和如何得出答案是未知的;相反,其他计算机视觉任务由算法回答的单个问题是预先确定的,只有输入图像发生变化[12].视觉问答任务的问题是任意类型的,问题的类型主要包含如下几类:

- 物体识别——图像中有什么?
- 物体检测——图像中存在狗吗?
- 二元问题——包含是否的问题
- 属性分类——图像中的狗是什么颜色?
- 场景分类——图像中的场景最可能是?
- 计数问题——图像中共有几只狗?
- 文本相关——图像中指示牌的内容是什么?

除此之外,问题可能更为复杂,可能涉及图像中对象间的空间关系或者需要一定的外部知识,比如回答“图中的动物属于哺乳动物吗?”时需要知道哺乳动物含有哪些动物.视觉问答任务包含了大部分其他经典的计算机视觉任务,并且需要对图像进行一定的推理.

Fig.1 Samples of visual question answering

图 1 视觉问答的样本

图像字幕任务与视觉问答任务的输入类似,但视觉问答任务比图像字幕任务更为复杂:视觉问答任务需要对图片内容进行推理,并且常常需要图片之外的知识,额外知识的范围从常识到专业知识;而图像字幕任务只需描述图像中的内容.与图像字幕任务相比,视觉问答任务更易于评价,其答案通常只有一个或几个单词;而图像字幕任务的答案通常是一个或多个句子,需要检查内容描述与图像是否一致,并且需要确认句子语法和句法的正确性,尽管当前研究了高级评价指标,但这仍是一个需要不断完善的研究.

视觉问答任务的研究有很多现实的应用,如:可以帮助盲人和视障人士能够在网络或者现实世界获得更多的信息,甚至可以进行实时的人机交互,这将极大改善盲人和视障人士的生活条件和便捷性;改善人机交互的方式,可以通过自然语言来查询视觉内容,拓展智能机器人的问答功能;视觉问答系统可以用于图像检索领域,比如可以针对数据集中的所有图像问“图像中存在汽车吗”.视觉问答任务包含大部分计算机视觉相关任务,视觉问答任务的不断发展,必定会带来诸多领域的进步.

视觉问答任务自 2014 年提出以来取得了巨大进步:最开始的方法主要集中在以视觉特征和文本特征联合嵌入的方式;之后,随着注意力机制的提出,视觉问答模型将注意力机制引入,为问题的解答提供了可解释性,效果也有了重要的进步.组合式模型注重问题解答的推理过程,但在自然图像集上表现不佳.针对部分需要外部知识问题,以知识库为基础的模型在这部分问题的解答方面有所进步.

2014 年~2017 年,已有多篇综述针对视觉问答任务进行了介绍[13−16].但近几年,视觉问答任务的研究得到了众多关注,数据集和模型有了重要的进步.有研究发现,视觉问答模型强烈依赖训练集中的表面相关性,存在语言偏见的问题,即:由于训练集中特定问题-答案对的数量占比过多,导致问题与答案存在强烈的关联,比如问题“是什么颜色”的答案一般为白色,问题“是什么运动”的答案一般为网球.当回答测试集中的问题时,模型会依赖训练数据中的语言先验得出答案,而缺乏对图像中内容的关注.由于训练集和测试集中针对相同问题的答案分布相近,早期模型利用数据集的漏洞取得了很好的效果;随着 VQA 2.0 数据集[17],特别是 VQA-CP 数据集的提出,模型的效果大幅下降.Agrawal 等人[18]的研究表明:VQA-CP 数据集相较于 VQA 数据集只对其数据分布进行改变,模型的效果平均下降 30%左右,如 SAN 模型[19]的准确率从 55.86%(VQA v1),52.02%(VQAv2)下降至 26.88%(VQA-CP v1),24.96%(VQA-CPv2).这说明数据集的分布对模型的影响十分严重,模型的鲁棒性存在一定问题.

本文主要介绍了与视觉问答任务相关的方法模型、数据集以及评价标准,许多研究针对模型的鲁棒性进行改进,本文进行了重点的介绍.本文第 1 节对视觉问答任务的方法进行了总结,重点介绍了近几年在模型鲁棒性方面的研究.第 2 节主要介绍了视觉问答任务相关的数据集,对于早期数据集进行比较和分析,重点介绍了近年来新提出的有关模型鲁棒性的数据集.第 3 节对于模型的评价标准进行了介绍.第 4 节讨论了视觉问答任务存在的挑战和展望.

1 模型介绍

在过去的 7 年内,国内外研究人员提出了大量的视觉问答模型,本文将问答模型的基本解决方案总结为以下 4 步:

- 提取视觉特征(图像特征化);
- 提取文本特征(问题特征化);
- 特征融合;
- 得出答案.

对于图像特征的提取,早期主要采用在 ImageNet[20]上预训练的卷积神经网络直接提取图像特征,常见的卷积神经网络模型为 VGGNet[21]、ResNet [22]和 GoogLeNet[23].之后,随着注意力机制的加入,大部分研究采用将图像分块提取特征.Anderson 等人[24]利用目标检测网络 Faster R-CNN[25]提取图像中对象的特征,采用图像中部分对象特征作为输入,这是目前视觉问答任务中主流的视觉特征.对于问题的文本特征提取,方法包括单词袋(BOW)、长短期记忆(LSTM)编码器[26]、门控递归单元(GRU)[27]和跳跃思想向量[28].对于特征融合方面,大部分模型使用简单的机制(例如串联、逐元素乘法或逐元素加法)将图像和问题特征进行组合.对于如何产生答案,

针对开放式的视觉问答任务,大部分研究将视觉问答任务视为分类任务,将视觉特征和文本特征作为分类系统的输入,从训练数据中得出若干个最常见的答案,每个答案视为一个单独的类别.对于多项选择形式的视觉问答任务,大部分研究将其视为排名问题,训练系统对每个可能的多项选择的候选答案给出分数,然后选择最高分数的答案.

本节的如下部分按照模型中采用的主要方法将模型分为联合嵌入方法模型、注意力方法模型、基于组合式的方法模型、基于外部知识库的方法模型以及鲁棒性研究模型这 5 个类别,分别从方法的动机、细节以及局限性这 3 个方面介绍使用这些方法的模型.最后,报告了近年来提出的模型在 3 个主要数据集的效果.

1.1 联合嵌入方法

视觉问答任务的输入为视觉特征和文本特征的多模态信息,需要将两种特征映射到共同的特征空间,联合嵌入的方法最先在图像描述任务[7−9]中应用.视觉问答任务与图像描述任务的输入类似,但需要进一步推理才能得出答案.将视觉特征和文本特征映射至同一空间更有利于信息之间的交互和进一步推理答案,因此,联合嵌入方法进一步在视觉问答任务中发展.联合嵌入方法大多是采用卷积神经网络提取视觉特征,循环神经网络提取文本特征,将两种特征通过简单的机制(例如串联、逐元素乘法或逐元素加法)组合,将组合后的特征送入线性分类器或神经网络,大致流程如图 2 所示.

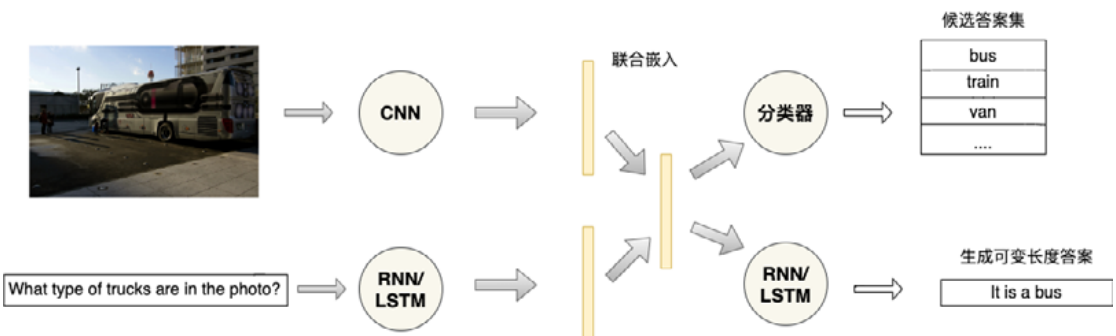

Fig.2 Framework of joint embedding methods

图 2 联合嵌入方法的框架

在视觉问答模型中最先利用联合嵌入方法是由 Malinowski 等人[29]提出的"Neural-Image-QA"模型,模型以卷积神经网络(CNN)和长短期记忆网络(LSTM)为基础,将视觉问答任务视为结合图像信息作为辅助的序列至序列(sequence to sequence)任务,最终生成的预测结果长度可变.首先由一个预训练的深度卷积神经模型提取图片特征,然后将图片特征和将问题词转化为词向量的文本特征作为长短期记忆网络的输入,每次输入将每个单词和图片特征输入至网络,直到将所有的问题特征信息输入.用同一个长短期记忆网络预测答案,直至产生结束符(⟨END⟩).模型的训练过程是结合视觉特征的长短期记忆网络的训练以及词向量生成器的训练.类似的工作[9]也采用长短时记忆网络生成可变长度的答案,但由于问题和答案的属性不同(例如两者的语法格式不同),应使用两个独立的长短时记忆网络处理更加合理.与上述两种生成式答案不同,Gao 等人[30]将视觉问答任务视为分类任务,将特征向量送入线性分类器,从预定义的词汇表中生成单字答案.在此基础上,Noh 等人[31]将 CNN 的全连接层中加入了动态参数预测层.利用递归神经网络将问题的文本特征产生候选权重,根据不同的问题对视觉输入产生的动态参数进行修改.

上述方法中,特征结合的方式有点乘、点加、连接等.由于图像和文本属于多模态信息,有大量的工作研究如何将两种特征进行融合.Fukui 等人[32]认为产生的联合向量表达能力不够,不足以捕捉多模态之间复杂的交互信息,因此提出多模态紧凑双线性池化模型(multimodal compact bilinear pooling,简称 MCB),在多模态特征融合时,使用双线性(外积、克罗内克积),但这会导致模型参数的数量急剧上涨.Fukui 等人通过 Tensor Sketch[33]算法降维和避免直接计算外积减少模型的参数,由于 MCB 模型需要输出高维度特征来保证鲁棒性,所以需要大

量的内存空间,限制了其适用范围.Kim 等人[34]提出了多模态低秩双线性池化模型(multimodal low-rank bilinear pooling,简称 MLB),MLB 模型是基于阿达玛积(Hadamard product)来融合两种特征.MLB 模型具有输出维度相对低、模型参数较少的优点,但是模型对超参数敏感,训练收敛速度慢.为了使得模型具有 MLB 模型输出低维度以及 MCB 模型具有鲁棒性的优点,Yu 等人[35]提出了多模态拆分双线性池化模型(multimodal factorized bilinear pooling,简称 MFB),将特征融合时用到的投影矩阵分解成两个低秩矩阵,大大减少了模型的参数和输出维度.Yu 等人[36]再次改进,提出了多模态因数化高阶池化模型(multi-modal factorized high-order pooling,简称 MFH). MFH 模型是将 MFB 模型中的操作分为扩张阶段和紧缩阶段,将 MFB 模型堆叠以得到高阶信息.在减少模型参数方面,Benyounes 等人[37]提出了一个多模态基于张量的塔克分解方法,用于参数化视觉和文本表示之间的双线性交互.此外,对于塔克分解,Benyounes 等人[37]设计了一种低秩矩阵分解来限制交互的秩,可以控制融合过程的复杂度,同时保持较好的、可解释的融合关系.Benyounes 等人[38]在此基础上提出了基于块超对角张量分解的双线性超对角融合.论文中借鉴了块项秩的概念,概括了已经用于多峰融合张量的秩和模态秩的概念.双线性超对角融合既能够表示模态间的精确交互,同时还保留单模态表示.一个双线性融合模型,其参数张量使用块项分解来构造.

在模型网络修改方面,Kim 等人[39]受深度残差结构的启发,提出了多模态残差网络(multimodal residual networks),在神经网络中加入多级残差连接,使得两个模态特征可以互相影响共同学习映射.Saito 等人[40]提出了“DualNet”整合两种操作,即两种模态特征元素级相加和相乘.Gao 等人[41]考虑了多种方式来进行模态融合,首先利用 *softmax* 操作和单层映射得到 k 组线性组合权重,权重与特征相乘后得到特征摘要向量,每个摘要向量都是单个特征的线性组合,与单个特征相比含有更高级的信息特征;将视觉摘要向量和文本摘要向量相乘得到 $k\times k$ 个视觉-文本摘要向量对,尝试建模每个单独的视觉-文本之间的关系以及在所有视觉-文本对之间传播更高阶的信息以对更复杂的关系进行建模,然后将结果相加,最后聚合信息以更新特征.

除将问题的文本特征和视觉特征作为模型的输入之外,Do 等人[42]发掘了三元组输入(图像,问题,答案)间的线性关联,输入的增加直接导致了模态融合时参数的增加.Do 等人使用 PARALIND 分解[43],有效地参数化 3 种输入间的交互.

单纯的联合嵌入方法不足以捕捉和建模所有特征信息,联合嵌入的过程不涉及对问题的理解以及对图片内容的推理,其属于视觉问答模型的基础部分,有很大的提升空间.由于视觉特征和文本特征中有很大部分信息对于解答问题没有帮助,直接将两种特征进行联合嵌入,会因为无关的信息影响最终的分类或答案生成.

1.2 注意力方法

上述大部分模型是将图片或问题提取的全部特征作为视觉问答模型的输入,但图片中含有大量与问题无关的信息,而问题中也存在需要重点关注的单词,将所有的特征全部输入最终会导致将大量噪声输入至分类器中,进而影响预测的准确率.注意力方法的目的是关注图片中与问题相关的区域或者关注问题中最关键的词,这一机制模拟了人脑的认知模式,即根据实际需求而将有限的注意力聚焦于事物的关键部分,从而大大加强了神经网络的理解能力.比如问题“图片中汽车的颜色是什么?”,问题中“汽车”和“颜色”是关键词,包含汽车的区域应该得到更多的关注.注意力方法在其他视觉领域和自然语言处理领域取得了很大的成功,比如对象识别[44]、图像字幕[45]以及机器翻译[46]等领域.Xu 等人[47]最先在与视觉问答任务相近的图像字幕任务中对注意力机制进行了探索,生成字幕时重点关注图片中的某一区域.

1.2.1 以问题为引导的注意力方法

早期的注意力方法是利用问题寻找图片中与问题相关的区域,Zhu 等人[48]将注意力方法与长短期记忆网络相结合,每一步将一个单词与视觉特征作为输入,输出该步的注意力图,将注意力图与视觉特征相乘就生成了新的视觉特征.Shih 等人[49]将视觉特征与文本特征简单相乘得到注意力权重,注意力权重的维度与视觉特征中区域的数量相等,权重的大小代表区域的重要程度,如图 3 中所示,注意力权重与视觉特征相乘后更新视觉特征. Yang 等人[19]认为视觉问答任务是一个推理的过程,单次获得的注意力权重不能准确地回答问题.因此,Yang 等人提出了堆叠关注网络模型(stacked attention network,简称 SAN).模型通过多次迭代,不断利用问题的文本特征

获得视觉特征区域的注意力,通过分析 SAN 模型不同层的输出,可以发现模型会更加关注图片中与问题有关的部分.实验结果表明:每次获得视觉特征区域注意力的过程都是一次推理的过程,能够关注更详细的内容.如图 4 所示,经过多次迭代,模型更加关注图中与问题相关的区域.

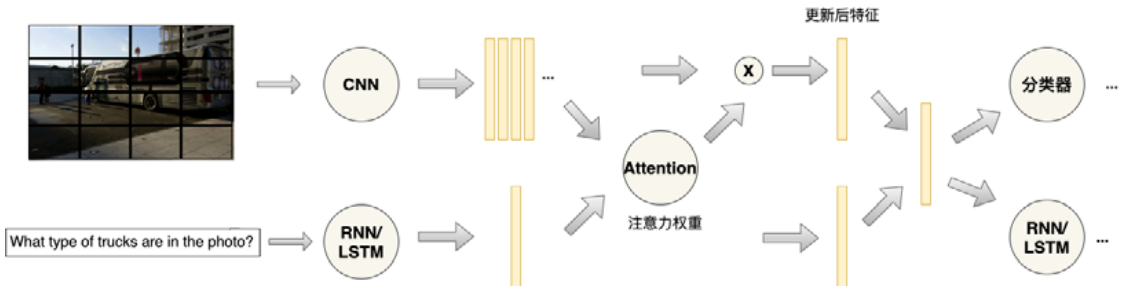

Fig.3 Framework of problem-guided attention methods

图 3 以问题为引导的注意力方法的框架

Fig.4 Visualization of the learned multiple attention layers[19]

图 4 学习的多个注意力层的可视化[19]

Patro 等人[50]认为,已有研究的注意力方法关注的区域与人类关注的图像区域并不相关.因此,Patro 等人提出通过一个或多个支持和反对范例来取得一个微分注意力区域,语义相近的范例和远语义范例之间存在差异,这样的差异能够引导注意力关注于一个特定的图像区域.实验证明了与基于图像的注意力方法相比,微分注意力更接近人类的注意力.

1.2.2 共同注意力方法

共同注意力方法不光考虑利用文本特征获得视觉特征的注意力,同样考虑得到问题的注意力,即问题中哪些单词更为重要.共同注意力模型是对称的,通过视觉特征可以引导产生问题的注意力,文本特征可以引导产生图片的注意力.Lu 等人[51]构建了一个层次结构,分别在单词层面、短语层面、句子层面构建共同注意力,提出了平行共同注意力和可选共同注意力两种构建方式:平行共同注意力是同时生成视觉注意力和文本注意力;而可选共同注意力是首先通过文本特征构建视觉注意力,利用得到的新视觉特征构建文本注意力.Nam 等人[52]认为:层次共同注意力模型[51]独立地执行了每一步的共同关注,而没有对之前的共同注意力输出进行推理.受内存网络启发,Nam 等人[52]提出通过视觉特征和文本特征共用的内存向量迭代更新视觉特征和文本特征,内存向量是通过将视觉特征和文本特征求和平均后分别得到视觉向量和文本向量,然后将两个向量相乘后得到的.利用内存向量与视觉或文本特征结合,分别生成视觉注意力和文本注意力.通过迭代的方式达到了推理的目的,进一步获得图片和问题的细节.不同于上述的共同注意力模型,Yu 等人[35]提出了多模态分解双线性池模型,文本注意力由问题单独推断,而视觉注意力的推断由文本注意力的参与.Yu 等人认为这与人类的反应一致,人们不需要借助图片也能抓住问题的重点.为了更好地获得图片中与问题有关的细节,Nguyen 等人[53]提出了层级递进的密集共同注意力的结构,其中使用了多头注意力,生成多个注意力图并将其平均.Yu 等人[54]提出的多层次注意力模型与之前不同的是并没有单独对问题求注意力,而是将注意力分成了语义注意力和上下文注意力,其中:上下

文注意力为以问题为引导的视觉注意力;语义注意力是通过卷积神经网络提取图片中的主要概念,将筛选出的概念与问题结合,形成语义概念注意力,即选出与问题相关的概念.不同于之前的方法,Wang 等人[55]提出了一种序列共同注意力方法,模型的输入为⟨问题,事实,图像⟩三元组,首先利用问题对事实进行加权,然后将加权的事实和初始问题表示相结合以指导图像加权.然后将加权的事实和图像区域一起用于指导问题进行加权,最后用问题和图像的注意力权重对事实再次进行加权构成整个循环.这意味着每一个注意力加权的过程都利用了其他过程的输出.Wu 等人[56]在视觉对话的研究中同样用到了序列共同注意力方法,其输入为⟨问题,历史对话,图像⟩的三元组,对 3 个输入进行互相加权,最后利用对抗生成算法使得生成的答案更像人类的回答.

共同注意方法学习了多模态实例的粗糙交互,而所学习的共同注意力不能推断出每个图像区域和每个问题词之间的相关性,这导致了共同注意模型的显著局限性.Yu 等人[57]认为,深度共同注意模型的瓶颈在于在每个模态中同时建模密集的自我注意(即问题的词对词关系和图像的区域对区域关系).他们提出了模块化共同关注网络(modular co-attention networks),如图 5 所示,通过共同注意力机制更新视觉特征和文本特征.网络框架的设计灵感来自于 Transformer 模型[58],模型设置了两个注意力单元,其中一个为自注意力单元进行模态内部交互和导向注意力单元进行模态之间交互.利用一个协同注意力模块层将两个单元串联起来,最后将多个模块层串联起来,组成 MCAN 网络.Gao 等人[59]认为:对视觉模态来说,每个图像区域不应该仅获得来自问题文本的信息,而且需要与其他图像区域产生关联.比如:对于“谁在滑板上?”这样的问题,模型应该把滑板对应的区域和滑板上方的区域关联起来;而对文本模态来说,使各个单词之间互相产生联系有助于提高模型对问题的理解.Gao 等人[59]同时考虑了模态内部关系和跨模态关系,分别构建了模态内部注意力单元和跨模态注意力单元更新视觉特征和文本特征.

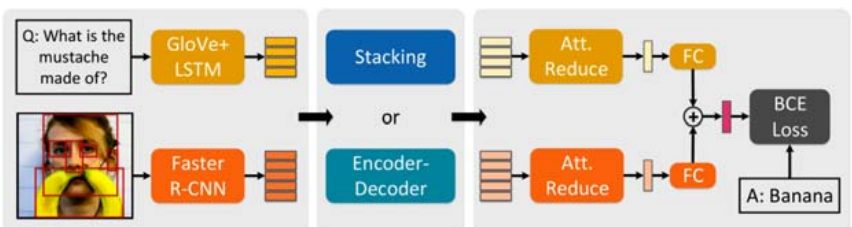

Fig.5 Overall flowchart of the deep modular co-attention networks[57]

图 5 深度模块共注意网络的总体流程图[57]

1.2.3 检测注意力方法

此前的图像注意力是基于卷积神经网络特征,这相当于把图片均等分割成若干区域,然后对其进行筛选.由于图片的分割,难免会破坏原有的对象,比如一个对象被分割为多个区域,如图 3 中左侧图像所示.Anderson 等人[24]利用目标检测网络 Faster R-CNN[25]来实现自底向上的注意力,将图片分割成一个个具体的对象来进行筛选,选择图片中前 K 个提议作为视觉特征,如图 6 中左侧图像所示,通过提取图中多个对象作为输入视觉特征.目前的主流模型均采用自底向上注意力生成的视觉特征.自上而下注意力即问题特征与各个提议的特征连接之后,通过非线性层和线性层得到视觉注意力,视觉注意力与视觉特征相乘得到更好的特征.Teney 等人[60]在此基础上对模型进行改进,采用多个技巧,如:分类器中使用 *sigmod* 输出,而不是传统的 *softmax* 输出,这样可以保证一个问题可能有多个正确答案;使用软分数作为地面真相目标,把任务作为候选答案分数的回归,而不是传统的分类;在所有非线性层中使用门控 *tanh* 激活函数;在随机梯度下降过程中使用大量小批次和对训练数据进行智能改组.

Lu 等人[61]并没有放弃原来那种基于卷积神经网络特征的开放式注意力(free-form attention),而是将开放式注意力与检测注意力结合,形成新的共同注意力.检测注意力作用受限于其检测类别的广度,如对于“今天天气怎么样?”这样的问题,如果目标检测网络不检测“天空”这个对象,则模型无法对这一问题做出准确回答.而开放式注意力就显示出了优势,因此,这两种注意力应是互补的.

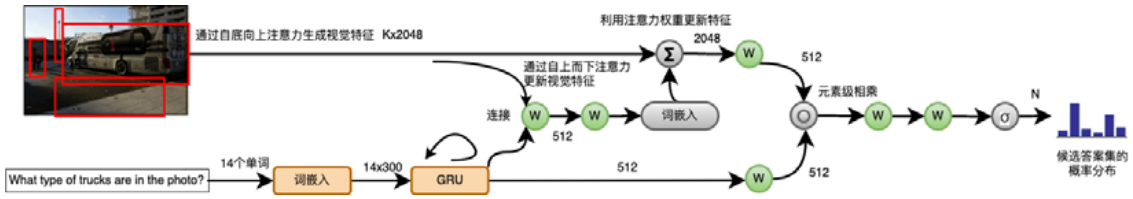

Fig.6 Overview of bottom-up and top-down attention model

图 6 自底向上和自上而下注意力模型的概述

1.2.4 关系注意力方法

Wu 等人[62]首次提出了关系注意力的概念,现有的大多数工作都集中在融合图像特征和文本特征来计算注意力分布,而不需要在不同图像对象之间进行比较.作为关注的主要属性,选择性取决于不同对象之间的比较.对象间的比较提供了更多信息,能够更好地分配注意力.对图中对象两两之间的关系进行建模,再用注意力机制对这些关系进行筛选.对于比较两个物体之间的关系,就是利用两个物体之间的特征进行差分操作.Cadene 等人[63]认为:目前的注意力机制相当于在给定问题的前提下,对每个图像区域打分后做信息加权.由于忽略了图像区域间空间和语义间的关联,所以不能做到有效地推理.Cadene 等人提出了 Murel 单元用于挖掘问题和图像区域间的细粒度关联,通过区域间关系的建模达到推理的目的,最后输出每个图像区域上下文感知的编码信息.如图 7 所示,通过对图像区域间的关系建模来获得上下文感知的嵌入特征.

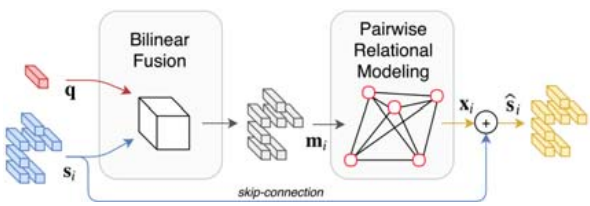

Fig.7 Overview of Murel cell[63]

图 7 Murel 模块概述[63]

图卷积网络(graph convolutional network,简称 GCN)是最近的研究热点,Li 等人[64]将图卷积网络应用至视觉问答任务.Li 等人认为,对象间视觉关系可以分为 3 大类:对象间的语义关系,主要体现为某个动作,比如孩子“吃”三明治;对象间的空间关系,主要体现两个对象间的相对位置,比如孩子和三明治“相交”(图像中的位置).以上两种关系被称为显式关系,因为它们是可以被明确命名的.但还有一些关系是无法语言表达,却对模型正确回答问题有重要帮助,称之为隐式关系.论文中用不同的图对 3 种关系建模,针对每一种关系训练一个关系编码器,最终将 3 个编码器进行综合,形成一个集成模型.

注意力方法与联合嵌入方法相比,显著地提高了模型在数据集上的准确率;同时,通过分析关于图像的注意力权重可以发现,模型会更关注于图像中与问题有关的区域,提供了回答问题的合理性.但是从问题类型中分析可以发现,注意力方法对于是/否问题的回答几乎没有帮助.注意力方法在回答问题的过程中没有进行推理的过程,仅仅是获得了更准确的视觉特征或文本特征.如何将视觉特征纳入推理的过程,仍需要进一步研究.

1.3 组合方法

上述方法中,主要是利用卷积神经网络和循环神经网络提取特征进行融合,训练过程缺乏具体推理的过程.而视觉问答任务本身是构成性的,比如问题“桌子上放的是什么?”,首先需要确定桌子的位置,然后需要确定桌子上方的位置,然后在桌子上方确定目标物体以及物体的类型.于是,有研究提出模块化网络解决视觉问答任务,针对不同的功能设计不同的模块,根据不同的问题将模块连接.模块化网络更易于监督,同样也提供了回答问题的可解释性,符合人类问答问题的逻辑思路.

Andreas 等人[65]首先将神经模块网络应用于视觉问答任务,其结构不同于传统的神经网络模型.神经模块

网络是一个整体,它是由多个模块化网络组合而成的.根据每个问题定制网络模型,神经模块网络是根据问题的语言结构动态生成的.首先,使用斯坦福大学提出的自然语言解析器[66]解析每个问题,获得通用的依赖关系表示[67];然后,以此分析出回答问题所需要的基础组成单元以及组成单元之间的联系,组成最终的布局网络.如图 8 中所示,神经模块网络回答“圆圈上面有红色的形状吗”时的推理过程.值得一提的是:网络中还使用长短期记忆网络(LSTM)作为问题编码器,目的是学习常识性知识和补充简化后丢失的信息.Andreas 等人[68]对神经模块网络的各个模块进行改进,在网络布局模块中加入了增强学习,从一组自动生成的布局候选中动态选择给定问题的最佳布局,动态地对每个实例的网络结构进行学习.Hu 等人[69]认为:目前的神经模块网络太过依赖语言解析器,并且仅限于解析器提供的模块配置,而不是从数据中学习.于是,Hu 等人提出了端到端模块网络,通过直接预测实例特定的网络布局来学习推理,而无需借助解析器.

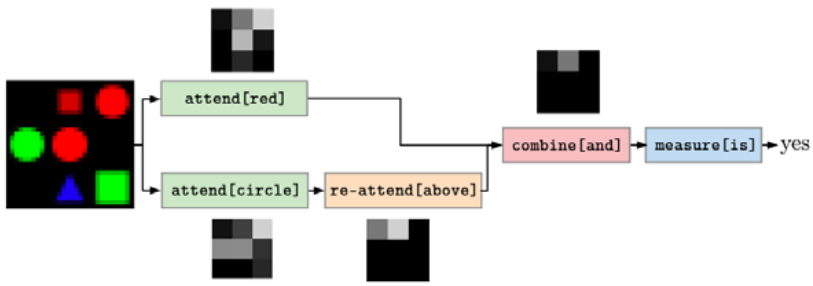

Fig.8 Overview of neural module networks[65]

图 8 神经模块网络概述[65]

动态内存网络最先由 Kumar 等人[70]提出,其是具有特定模块化结构的神经网络.Xiong 等人[71]将其应用至视觉问答领域,利用卷积神经网络提取视觉特征输入循环神经网络,将特征图使用激活函数的线性层映射到和问题的文本特征同一空间的向量,最后使用双向门循环单元获取特征.动态内存网络通过对数据多个部分之间的多次交互进行建模来解决需要复杂逻辑推理的任务.Noh 等人[72]提出的 RAU 模型也可以隐式执行合成推理,而无需依赖外部语言解析器.模型使用了多个可以解决视觉问答子任务的独立应答单元,这些应答单元以循环方式排列.

组合式模型目前主要应用于合成图像数据集中,在自然图像数据集中效果比较差,依赖于语言解析器的模型主要在进行语言逻辑的推理,并没将推理过程作用于图像中.但组合式模型潜力巨大,提供了解决视觉问答任务的可解释方式,这是符合人类回答问题的过程.目前的瓶颈可能在于提取的特征不足以开展推理过程,随着深度学习的不断进步,组合式方法可能会有着巨大的进步.

1.4 基于外部知识的方法

视觉问答任务是人工智能中一个非常具有挑战性的任务,回答问题需要理解图像的视觉内容,理解视觉内容的前提是知道一定的非视觉信息,如回答“图中有多少只哺乳动物?”,首先需要知道图中的动物是否属于哺乳动物,这种问题需要借助外部知识才能够回答.部分研究将视觉问答任务与知识库相结合,部分数据集的提出是专门针对这类方法的研究,如 KB-VQA 数据集[73]以及 FVQA 数据集[74].由于训练集中的知识是一定的,并不能完全覆盖回答问题的全部知识,所以若想回答有难度的问题,从外部获取知识是必要的.

Wang 等人[73]提出了名为“Ahab”的视觉问答框架:首先,通过卷积神经网络从图像中提取视觉概念;然后,在 DBpedia 知识库[75]内寻找相近的节点,总结查询的结果得出最终答案.但是“Ahab”框架需要通过设计的模板解析问题,这大大限制了能够回答问题的种类.为了解决需要模板解析问题的限制,Wang 等人[74]在此基础上通过长短期记忆网络和数据驱动的方法学习图像和问题到查询的映射.Wu 等人[76]通过卷积神经网络提取语义属性,从 DBpedia 知识库[75]中检索与之相关的外部知识,DBpedia 知识库中包含的简单描述通过 Doc2Vec 嵌入到固定大小的向量中.嵌入的向量被输入到长短期记忆网络模型中,然后与问题相结合,并最终生成答案.Wu 等

人[77]通过提取图像中的高级语义,将图像内容的内部表示与从通用知识库中提取的信息结合起来,特别允许询问关于图像内容的问题,即使图像本身不包含完整答案.如图 9 中所示,从知识库(在本例中是 DBpedia)和 Doc2Vec 编码的响应中挖掘知识,进一步编码问题的表示.

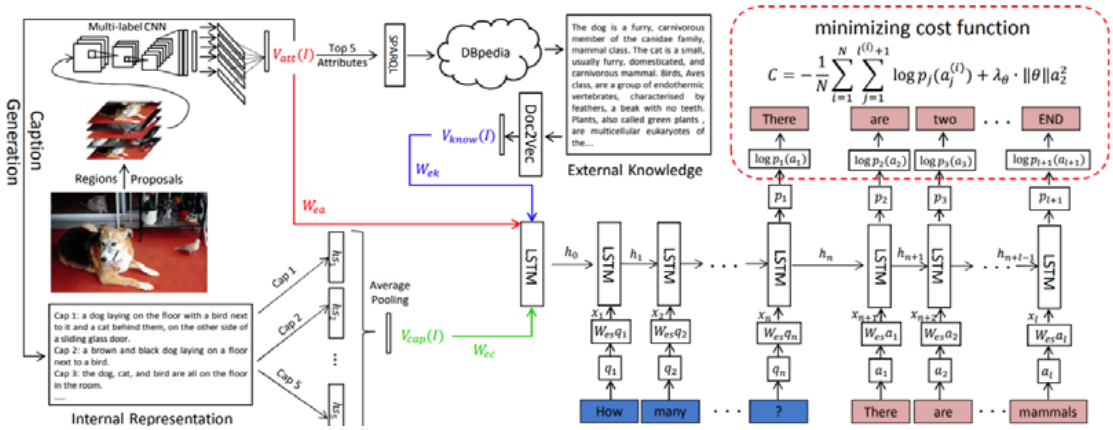

Fig.9 A VQA model with external knowledge[77]

图 9 具有外部知识的 VQA 模型[77]

由于大部分问题仅需要小量的先验知识,模型在通用数据集上的效果并不能在引入外部知识后得到显著的提升;并且,如何准确地查找所需的知识以及将获得的知识用于回答问题,如何得到一个合适的、可扩展的框架用于融合和自适应地选择相关的外部知识等问题,还需要进一步研究.

1.5 鲁棒性研究

近年来,视觉问答任务受到了广泛的关注,提出了很多深度学习模型,在不同数据集上展现了很大的进步,但是目前的视觉问答模型有着许多鲁棒性问题.从研究[16,30,78,79]中可以发现,目前的视觉问答模型受训练集表面相关性的影响很大.由于训练集计数问题的答案中“2”的比例很高,比如回答“图中有多少个…”的问题时,不论图中是什么物体,答案基本上都是“2”.模型可以利用训练集中的统计数据,问题类型与答案相关度很高,不需考虑图片的内容就可以得到正确答案.从 Shah 等人[80]的研究中可以发现:目前的视觉问答模型对于问题中语言变化十分敏感,在不改变问题含义的前提下,修改问题的句子结构或者增删某个单词,模型给出的答案随之改变.Zhang 等人[79]通过研究视觉问答模型对图像中有意义的语义变化的鲁棒性,分析了视觉问答模型中视觉的重要程度.Xu 等人[81]的研究表明:尽管使用了先进的注意力机制,但很容易用图像中很小的变化来欺骗视觉问答模型.Agrawal 等人[82]研究了视觉问答模型对训练和测试环境中答案分布变化的鲁棒性.

为了避免受数据集的表面相关性影响,有研究在改进数据集方面进行努力,创建更平衡的数据集.Zhang 等人[79]对所有二元问题收集了具有相反答案的互补抽象场景.Goyal 等人[17]把这个想法扩展到真实的图像和所有类型的问题.VQA v2 数据集[17]平衡答案分布,使每个问题至少存在两个答案不同的相似图像.VQA-CP v2 数据集[18]将 VQA v2 数据集进行诊断重构,其中,训练集中的问题答案分布与测试集中的明显不同,这可以避免视觉问答模型利用训练集中的偏见.

有的研究在改进模型方面进行努力,大部分的方法采用引入另一个只将问题作为输入的分支,如图 10 右侧中所示.Chen 等人[83]将模型的改进分为两类.

(1) 基于对抗的方式

Ramakrishnan 等人[18]将对抗性规则化(AdvReg)应用至视觉问答任务中,其引入了一个只考虑问题的模型,模型将视觉问答模型中的问题编码作为输入.将训练的过程视为视觉问答模型和只含有问题作为输入的模型进行对抗——阻止视觉问答模型在其问题编码中捕捉语言偏见.同时引入置信度量化,训练过程使得在考虑图像之后,模型置信度增加,通过显式地最大化两个模型之间的置信度差异,以鼓励模型重视视觉基础.Grand 等

人[84]研究了对抗性规则化的优缺点,其可能产生不稳定的梯度和在域内示例上的性能急剧下降.在训练过程中逐步引入正则化,有助于减轻这些问题.AdvReg 提高了对二元问题的泛化能力,但降低了对异质答案分布问题的性能.正则化模型往往过度依赖视觉特征,而忽略了问题中重要的语言线索.Belinkov 等人[85]在自然语言推理(natural language inference)任务上采用了相似的对抗策略,基准模型采用假设和前提来预测标签,而采用对抗策略的模型加入了只采用假设的分类器,或者针对一个假设随机采用一个前提进行训练.但是在 Grand 等人[84]的研究中显示:对抗性训练方法给梯度带来了很大的噪声,导致训练过程不稳定,可能导致性能的严重下降,引入正规化有助于缓解但不能完全解决这些问题.

(2) 基于融合的方式

基于融合的方式是将两个分支预测答案的分布在最后进行融合,并基于融合的答案分布导出训练梯度.基于融合方法的设计思想是,让目标视觉问答模型更多地关注于不能被只考虑问题模型正确回答的样本.Cadene 等人[86]提出了 RUBi 训练策略,通过将只考虑问题模型的预测答案分布经过 *sigmod* 操作之后视为掩码,然后将其与视觉问答模型的预测答案分布相乘.如图 10 所示,通过 RUBi 策略对现有模型进行语言去偏,若需要问答的问题存在严重的偏见,将掩码与视觉问答模型的预测答案分布相乘之后的损失会很小,则这个样本不会对模型的参数有很大改变;若需要回答问题的答案不常见,即通过只考虑问答模型得到的答案分布与视觉问答模型得到的答案分布差距很大,两者相乘之后的损失会很大,于是模型会更重视这个训练样本,对模型参数的影响也会很大.RUbi 策略使得模型更重视偏见更小的训练样本.Clark 等人[87]提出的方法分为两步:第 1 步训练一个带有偏见的模型,模型在训练集表现好,但是在这范围之外表现差;第 2 步再训练一个模型集成带偏见的模型,在测试集上只用第 2 个模型.具体实现中采用了答案分布中的偏见,给每个类型的问题出现的答案打分,每个类型的分数作为该候选答案的偏见,将偏见与模型的损失结合,达到减少训练过程中的损失,通过惩罚项使得模型对偏见高的答案关注更少.Mahabadi 等人[88]介绍了 3 种减少偏见的策略:第 1 种为直接将两个分支模型的预测答案分布相乘;第 2 种为 RUBi 策略并提供了修改 *sigmod* 操作的两种变体,即 RUBi+对数操作和 RUBi+标准化;第 3 种为修改模型的损失函数减少带有偏见样本的重要性,使模型更加关注回答难度高的样本.

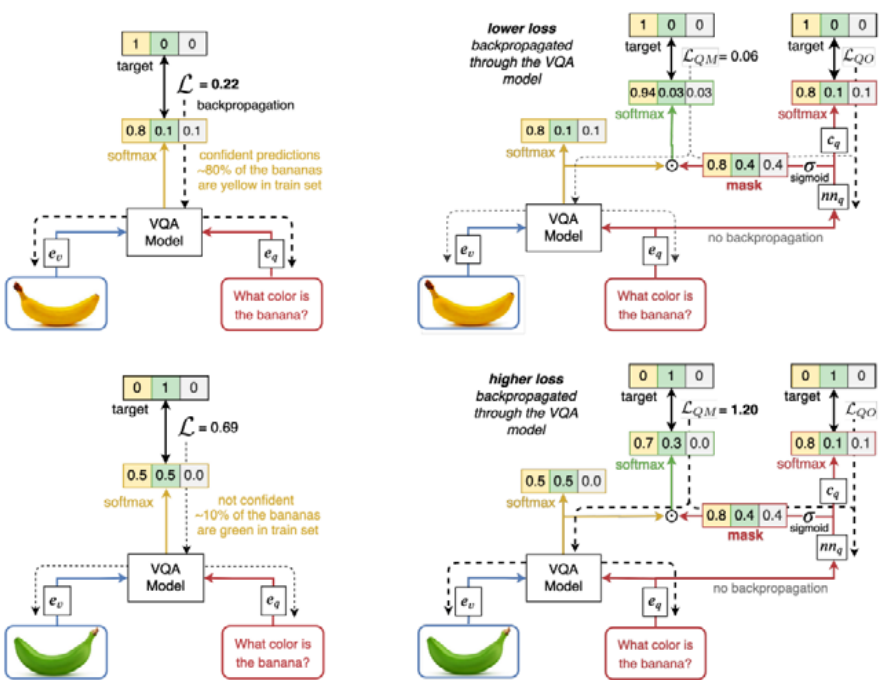

Fig.10 Detailed illustration of the RUBi impact on the learning[86]

图 10 RUBi 对学习影响的详细图示[86]

除增加分支对模型进行改进外,Wu 等人[89]在研究中发现:视觉问答模型被鼓励关注人类认为重要的图片区域,即使当视觉问答模型产生了错误的答案,也会关注重要的区域.当出现这种现象时,模型并不会纠正.论文中提出了一种“自我批评”的方法,直接批评不正确的答案对重要区域的敏感性.对于每个问答对,首先确定最影响模型预测正确答案的区域.当模型对这个问题的预测答案是错误的时候,惩罚它对这个区域的关注,保证了正确答案与其他答案相比更关注重要的区域.

但上述方法不能同时增加视觉问答模型的视觉可解释性和问题敏感度,模型应该更加注意与问题更相关的视觉区域,也就是针对正确的区域做出决定.模型应该对所讨论的语言变化敏感,也就是说应该注意问题的敏感词(重要的词),当敏感词变化的时候,得到的答案应该变化,模型的处理也应该有变化.Chen 等人[83]提出了与模型无关的反事实样本合成(CSS)训练策略.CSS 由两种不同的样本合成机制组成:V-CSS 和 Q-CSS.对于V-CSS,它通过掩盖原始图像中的关键对象来合成反事实图像.意味着这些对象对于回答某个问题很重要.然后,反事实图像和原始问题组成了一个新的图像问题对.对于 Q-CSS,它通过使用特殊标记“[MASK]”替换原始问题中的关键单词来合成反事实问题.同样,反事实问题和原始图像构成了新的视觉问题对.针对新生成的样本对采用动态答案分配机制构成完整的三元组样本.通过数据扩增,视觉问答模型被迫专注于所有关键对象和单词,从而显着提高了视觉可解释性和问题敏感性能力.

目前的视觉问答模型还有其他鲁棒性问题,如回答有关于图片中文本问题的准确率不高.Singh 等人[90]为了进一步研究回答有关图片中文本的问题,提出了 TextQA 数据集,TextQA 数据集中所有问题都需要对图片中的文本进行推理才能回答.同时提出了一个新的模型结构,在模型中加入了光学字符识别(optical character recognition)模块,它可以读取图像中的文本,模型可以在图像和问题的上下文中推理读取的文本,最终答案可以是通过文本和图像推理得到的答案或通过光学字符识别得到的文本.Biten 等人[91]同年提出了 ST-VQA 数据集,旨在强调在视觉问答过程中,利用图像中的高级语义信息作为回答关于文本问题的重要线索.论文中将传统视觉问答模型与场景文本检索(scene text retrieval)模型结合,将生成最可信的字符的金字塔状直方图(PHOC)特征与视觉特征连接.

视觉问答模型存在对问题敏感度高的鲁棒性问题,Shah 等人[80]针对这个问题提出了 VQA-Rephrasings 数据集,数据集中的每个问题有另外 3 个含义相同但句式等其他方面存在不同的改述问题.论文中提出了周期一致性的训练策略,该策略借鉴了 Cycle-GAN[92]的思想,首先通过视觉问答模型给出问题答案,通过答案生成原始问题的改述问题,视觉问答将改述问题作为输入得到新的答案.整个训练过程是缩小原始问题和改述问题之间、真实答案与两次生成的答案之间的损失,使得模型更加健壮,模型能针对相同含义的问题给出相同答案.

当前的视觉问答模型回答有关计数问题与其他类型问题相比准确率不高,Zhang 等人[93]提出造成计数类问题表现不佳的原因主要有:(1) 软注意力(soft-attention)的广泛运用;(2) 区别于标准的计数问题,对于视觉问答任务来说,没有明确的标签标定需要计数对象的位置;(3) 视觉问答模型的复杂性表现在不仅要处理计数类问题,同时还要兼顾其他复杂的问题;(4) 真实场景中,对某个对象区域可能存在多次重叠采样.论文中将相关的建议对象描述成点,对象间的内部与外部关系描述成边,最终形成图,通过设计策略取消重复采样对象内部和减半与其他对象之间的边,最终对象数量等于边数量的算术平方根.Acharya 等人[94]提出了世界上最大的开放式计数数据集 TallyQA 数据集[94],目前的数据集记数问题相对简单只需要对象检测,而 TallyQA 数据集中的问题属于复杂计数问题,只通过对象检测无法回答.论文中提出了新的计数方式——关系计数网络(RCN),其受到关系网络的启发,通过修改处理动态数量的图像区域并显式地合并背景信息,可以推断对象与背景图像区域之间的关系.

Shrestha 等人[95]提出:视觉问答模型并不能兼容自然图像的理解和合成数据集的推理,大部分模型在这两个领域不具有泛化能力.他们提出了通过将视觉特征和文本特征两次融合,在自然图像数据集和合成数据集上均得到了良好的效果.实验结果表明,第 1 次融合比较重要,若无第 1 次特征融合,模型的效果会下降约 4%.

视觉问答模型的鲁棒性研究是近几年的研究热点,由于问题类型的复杂性,模型不能兼顾所有类型的问题.数据集中答案的分布使得模型能够利用语言相关性正确地回答问题,但是其泛化能力差.大部分避免模型利用

语言相关性的方法均是引入一个仅考虑问题的分支,但是仍未从根本上解决问题,目前的模型在 VQA-CP 数据集上的准确率仍然很低.针对其他鲁棒性问题,比如有关于图片中文本的问题依赖于光学字符识别模块.计数问题是所有类型中最困难的一种问题,当前最有效的方法是将图中的对象和对象间的关系视为图,模型的准确率与目标检测的准确率有关.模型对于问题过于敏感表明模型对于问题并没有真正地理解,仍需要大量数据训练网络.模型对于自然图像和合成图像之间的泛化能力差的主要原因是自然图像中的信息过于复杂,目前的特征表示能力不足以对其进行推理.

1.6 模型效果介绍

表 1~表 3 介绍了近年来大部分最先进的模型在各个数据集上表现,并且介绍了模型使用的方法以及使用的视觉和文本特征.大部分模型使用了注意力方法,所有模型采用自底向上注意力得到的视觉特征,答案的生成方式均为分类.表 1 中,VQA 2.0 数据集含有验证测试集和标准测试集,模型在标准测试集上的效果略好于验证测试集.表 2 中,目前的模型在 VQA-CP 数据集上的效果欠佳,仍需要进一步的提高.表 3 中报告了各个模型在 TDIUC 数据集上的整体准确率(All)、每一类型准确率的算术均值(A-MPT)以及每一类型准确率的调和均值(H-MPT).

Table 1 State-of-the-art comparison on the VQA 2.0 dataset

表 1 数据集 VQA 2.0 的最新比较

模型方法	准确率		联合嵌入方法	注意力方法	组合方法	外部知识	鲁棒性研究	答案方式	视觉特征	文本特征
	test-dev	test-std								
BLOCK[38]	67.58	67.92	√	√	–	–	–	分类	UpDn	Skip-thought
MuRel[63]	68.03	68.41	√	√	–	–	–	分类	UpDn	GRU
RAMEN[95]	65.96	65.96	√	–	–	–	√	分类	UpDn	GRU
MCAN[57]	70.63	70.90	√	√	√	–	–	分类	UpDn	GloVe+LSTM
CTI[42]	66.00	67.40	√	√	–	–	–	分类	UpDn	GRU
MLIN-BERT[41]	71.09	71.27	√	√	–	–	–	分类	UpDn	Transformer
ReGAT[64]	70.27	70.58	√	√	–	–	–	分类	UpDn	GRU

Table 2 State-of-the-art comparison on the VQA-CP dataset

表 2 数据集 VQA-CP 的最新比较

模型方法	准确率	联合嵌入方法	注意力方法	组合方法	外部知识	鲁棒性研究	答案方式	视觉特征	文本特征
CSS[83]	58.95	√	√	–	–	–	分类	UpDn	LSTM
Learned-Mixin+H[87]	52.05	√	√	–	–	√	分类	UpDn	LSTM
RUBi[86]	47.11	√	√	–	–	√	分类	UpDn	GRU+Skip-thought
NSM[96]	45.80	√	–	√	–	–	分类	UpDn	GloVe
GVQA[82]	31.30	√	√	–	–	√	分类	UpDn	LSTM

Table 3 State-of-the-art comparison on the TDIUC dataset

表 3 数据集 TDIUC 的最新比较

模型方法	准确率			联合嵌入方法	注意力方法	组合方法	外部知识	鲁棒性研究	答案方式	视觉特征	文本特征
	All	A-MPT	H-MPT								
BLOCK[38]	85.96	71.84	65.52	√	√	–	–	–	分类	UpDn	Skip-thought
MuRel[63]	88.20	71.56	59.30	√	√	–	–	–	分类	UpDn	GRU
RAMEN[95]	86.86	72.52	–	√	–	–	–	√	分类	UpDn	GRU
DFAF[59]	85.55	–	–	√	√	–	–	–	分类	UpDn	GRU
QTA[97]	85.03	69.11	60.08	√	√	–	–	–	分类	UpDn	LSTM
MLI[41]	87.60	–	–	√	√	–	–	–	分类	UpDn	Transformer

2 数据集介绍

自从视觉问答任务被提出,大量数据集随之出现.视觉问答数据集的一般形式为⟨图像,问题,答案⟩的三元组,部分数据集还带有关于图像的注释.2014 年~2016 年,主要有 6 个包含自然图像的数据集:DAQUAR 数据集[98]、

COCO-QA 数据集[99]、FM-IQA 数据集[30]、VQA 数据集[100]、Visual7W 数据集[48]、Visual Genome 数据集[33].由于上述数据集已在综述[5,6]中详细介绍,在此便不多赘述,只对上述数据集存在的问题进行总结.下文主要详细介绍经过数据分布平衡的 VQA-CP 数据集[82]、研究图像文本的 TextVQA 数据集[90]、研究模型鲁棒性的 VQA-Rephrasings 数据集[80]、研究复杂计数问题的 TallyQA 数据集[94]以及研究模型可解释性的 VQA-X 数据集[101].

2.1 早期数据集分析

上述数据集都有其局限性,比如:DAQUAR 数据集和 COCO-QA 数据集在数据规模上比较小;DAQUAR 数据集中的图片比较杂乱,提出的问题难以回答,即使是人类回答的准确率也只有 50.2%;COCO-QA 数据集中的问题是由图片的注释自动生成的,存在高重复率的现象,难以支撑模型的训练和评价.相比较而言,Visual Genome 数据集、Visual7W 数据集和 COCO-VQA 数据集比较大,但是却存在一定的偏见,偏见既存在于针对图片的问题中,也存在于给出的答案中.在文献[102]中可以看到,仅将问题的特征输入模型进行训练就可以得到约 50%的准确率,这说明数据集中答案的分布不均衡.COCO-VQA 中以“是否存在一个”为开头的问题,79%的答案是“是”.Visual Genome 数据集中的问题一部分是关于图像整体内容的问题,这可能导致提问中的偏见.

为了减少数据分布对模型的影响,Goyal 等人[17]在 2017 年提出了 VQA 2.0 数据集.与 VQA 1.0 数据集相比,VQA 2.0 数据集规模更大,并且主要解决了答案不平衡的问题,针对两张不同的图像提问相同的问题,并且尽量使得到的答案相反.但是 VQA 2.0 数据集仍存在答案分布问题,训练集和测试集的答案分布相似,模型可以利用答案分布带来的偏见得到较高的准确率,降低了模型的泛化性.

由于评价指标存在的偏见,模型之间的性能比较不透明.Kafle 等人[14]提出了 TDIUC 数据集,将问题划分为 12 种类型,分别为“是否有对象”“对象种类识别”“计数”“颜色”“其他属性”“动作识别”“体育活动识别”“位置推理”“场景分类”“情绪理解”“用途”“错误”.TDIUC 数据集可以衡量视觉问答模型在每个类别中的性能,识别哪种问题是容易的还是困难的.为了进一步减少数据集中偏见的影响,分别计算了 12 种问题类型的准确性,同时计算最终的统一精度指标.总体指标是每个问题类型准确性的算术均值和调和均值,分别称为算术平均类型准确性和调和平均类型准确性.与算术平均类型准确性不同,调和平均类型准确性衡量系统在所有问题类型上均具有高分并偏向性能最低的类别的能力.

为了研究视觉问答模型的推理能力,有研究提出了 SHAPES 数据集[65]和 CLEVR 数据集[102],通过强调理解多个对象之间的空间和逻辑关系.这是对自然图像数据集的补充,在此之前的数据集中的图像均为自然图像,其中的问题不能衡量模型的推理能力.SHAPES 数据集由 244 个独特的问题组成,每个问题都与数据集中的 64 幅图像有关.所有问题都是二元的,答案为是或否.SHAPES 数据集中所有图像均为 2D 形状,不能代表真实世界的图像.CLEVR 数据集使用 3D 渲染的几何对象,数据集规模比 SHAPES 数据集规模大,包括 10 万张图像和 864 968 个问题.CLEVR 数据集中的问题测试了视觉推理的各个方面,包括属性标识、计数、比较、空间关系和逻辑运算.但 SHAPES 数据集和 CLEVR 数据集低估了视觉推理的重要性,相比较而言,模型在回答问题时更注重语言推理能力,比如回答“大球面左边的棕色金属物体左边的圆柱体的大小是多少?”需要严苛的语言推理能力,而对于视觉推理能力则有限.

上述讨论的数据集的大多是纯视觉问题和常识性问题,几乎没有需要“知识库级”的问题.为了更深入研究使用外部知识库的视觉问答的模型,有研究提出了 KB-VQA 数据集[73]和 FVQA 数据集[74].KB-VQA 数据集包含需要 DBpedia 中特定主题知识的问题,从 COCO 图像数据集[103]中收集了 700 幅图像,每幅图像收集 3 到 5 个问题-答案对,共 2 402 个问题.每个问题需要不同层次的知识,从常识到百科全书知识.FVQA 数据集仅包含涉及外部(非可视)信息的问题.数据集包含与 580 个视觉概念(234 个对象、205 个场景和 141 个属性)有关的 193 005 个候选支持事实,总共有 4 608 个问题.FVQA 数据集在每个问题/答案中都包含一个支持的事实(外部知识).

2.2 VQA-CP数据集

目前,数据集中存在训练集强语言相关性的问题,比如回答“香蕉是什么颜色的?”,回答通常是“黄色”,而这

种情况导致模型不需要查看图片的内容就可以回答这类问题.出现这种情况的一个原因是训练集和测试集有着相似的数据分布,模型会根据在训练集中产生的固有记忆偏差,忽略图像的内容,而在测试集中还能得到可观的性能.

针对训练集强语言优先级的问题,Aishwarya 等人[82]对数据集 VQA v1 和 VQA v2 重新划分,分别得到了 VQA-CP v1 和 VQA-CP v2 数据集,使得每个类型问题的答案分布在训练集和测试集之间是不同的.比如“什么运动?”这类问题,在训练集中最常见的答案是网球,而在测试集却是滑冰.通过对问题类型和答案类型的重新划分,能够减少在测试时依赖训练过程中产生的语言偏见.

在 VQA-CP 数据集中,测试集覆盖了绝大部分训练集中出现的概念,覆盖率在 VQA-CP v1 中是 98.04%,VQA-CP v2 是 99.01%.VQA-CP v1 的训练集前 1 000 个答案中,测试集答案的覆盖率为 95.07%(VQA-CP v2 为 95.72%),VQA- CP v1 训练集由 118K 张图像、245K 个问题和 2.5M 个答案组成(VQA-CP v2 训练集由 121K 幅图像、438K 个问题和 4.4M 个答案组成).VQA-CP v1 测验集由 87K 幅图像、125K 个问题和 13M 个答案组成(VQA-CP v2 测试集的 98K 幅图像、220K 个问题和 22M 个答案).Aishwarya 等人[82]报告了基线模型和现有视觉问答模型在 VQA-CP v1 和 VQA-CP v2 训练分割上的性能,几乎所有模型都出现了性能的大幅下降,这证明了之前的视觉问答模型利用了训练集的语言优先级.

2.3 TextVQA数据集

当前提出的视觉问答模型对于回答有关于图像文本问题的准确率很低,为了促进这类问题的研究,Singh 等人[90]提出了 TextVQA 数据集.TextVQA 要求模型阅读并推理图像中的文本,以回答关于它们的问题.具体来说,模型需要合并图像中出现的一种新的文本形式并对其进行推理,以回答 TextVQA 数据集中问题.其采用了 Open Images v3 数据集内的图像,选取的图像中包含文本(如广告牌、交通标志等),每个类别选取 100 幅图像.使用 OCR 模型 Rosetta[104]计算图像中的 OCR 盒的数量,将每个类别的 OCR 盒子的平均数量归一化,并用作每个类别的权重,以从类别中采样图像.从 Open Images v3 数据集的训练集中采样得到 TextVQA 数据集的训练集和验证集,从 Open Images v3 数据集的测试集采样得到 TextVQA 的测试集.每张图像有 1~2 个问题,每个问题由 10 名注释者给出答案.数据集共包含 45 336 个问题,其中,37 912 个问题是唯一的.TextVQA v0.51 中训练集包括 34 602 个问题、21 953 幅图像;验证集包括 5 000 个问题、3 166 幅图像;测试集包括 5 734 个问题、3 289 幅图像.

2.4 VQA-Rephrasings数据集

目前的视觉问答模型的鲁棒性不强,对于同一问题的不同表述,模型会给出不同的答案.为了进一步研究模型一致性和鲁棒性,提出了 VQA-Rephrasings 数据集[80].VQA-Rephrasings 数据集来自于 VQA v2 的验证数据集,其是对关于 4 万张图的 4 万个问题的改述生成的.这是首个能够进行一致性和鲁棒性视觉问答模型评估的数据集.数据集一共包含了 214 354 个问题和 40 504 张图片,随机采样了 40 504 个问题构成采样子集.作者用两阶段的方式对每个问题用人工标注的方式生成 3 个改写问题.

- 第 1 阶段,根据原始的问题-答案对改写问题,改写后的问题回答要与原始答案一致.
- 第 2 阶段,对第 1 阶段的问题进行语法和语义检查,不合规范的抛弃.

最后获得了 162 016 个问题(包括改写的 121 512 个和原始的 40 504 个)和 40 504 张图片,平均每张图片对应约 3 个改写问题.

2.5 TallyQA数据集

回答计数问题对于当前的视觉问答模型来说是一个严峻的挑战,但是当前存在的综合数据集的计数问题占比并不高,例如 COCO-QA 数据集[99]中约占 7%,VQA v1 数据集[100]中约占 10%,VQA v2 数据集[17]约占 10%以及 TDIUC 数据集[4]约占 20%.还有一些针对计数任务的 VQA 数据集如 CountQA 数据集[105]和 HowMany-QA 数据集[106]的规模并不大,并且上述数据集中很少有复杂的计数问题.简单的问题可以只用一个目标检测算法来解决,因此不能恰当地测试系统回答任意计数问题的能力,包括那些需要推理或属性识别的问题.

Acharya 等人[94]提出了新的数据集 TallyQA,旨在评估简单和复杂的计数问题,使计数问题和其他问题得到

准确的衡量.Acharya 等人使用 Amazon Mechanical Turk(AMT)收集新的复杂问题,并从其他数据集中导入简单和复杂问题.数据集的具体情况见表 4.

Table 4 Number of questions and images in the TallyQA dataset

表 4 TallyQA 数据集中问题和图像的数量

数据集分割	问题	图片
训练集	249 318	132 981
AMT收集	3 902	3 494
其他输入集引入	245 416	129 487
简单测试集	22 991	18 411
AMT收集	0	0
其他输入集引入	22 991	18 411
复杂训练集	15 598	14 051
AMT收集	15 598	14 051
其他输入集引入	0	0

2.6 VQA-X数据集

深度学习的可解释性是当前的研究热点和难点,视觉问答模型的可解释性同样是研究的难点.人类回答问题时是基于一定的事实,我们希望视觉问答模型得出答案同样是基于图像中事实或其他知识.为了研究视觉问答模型的可解释性,Huk 等人[101]提出了 VQA-X 数据集,其是在 VQA 数据集上得到.根据 Zitnick 等人[107]收集的注释,其中含有回答问题的年龄限制,Huk 等人选择 9 岁及 9 岁以上才能回答的问题.此外,Huk 等人还考虑了 VQA v2 数据集的互补对[17].互补对由一个问题和能够给出两个不同答案的两个相似图像组成.互补对能帮助理解解释模型是根据图像内容来给出解释,还是仅仅根据特定的问题类型记忆要考虑的内容.训练集中每一个问题答案对有一个文本解释,训练/测试集的每个问题答案对有 3 个文本解释.

3 评价标准

对于多项选择形式的视觉问答任务,算法得出的答案与正确答案容易比较;但开放式的视觉问答任务得出的答案通常为一个或多个单词,与图像字幕任务类似,难以对准确性进行评价.若将算法得出的答案与正确答案完全匹配则准确性过于严格,因为错误答案之间仍有严重程度之分,比如将得出的答案因为单复数的差别而判断为错误答案,与得出完全不相关的答案的惩罚程度相同则不太合适.而同一问题可能有多种合适的答案,比如问题“天空中正在飞的是什么?”,正确答案为“bird”,而回答“jay”或“fowl”与其意思相近.因此,有的研究提出了多种准确性评估的替代方法.

Malinowski 等人[98]提出两种方法进行模型准确性评价:一种是将预测答案与正确答案进行字符串匹配来确定最终的准确性;第 2 种是使用 WUPS[108]计算预测答案与正确答案在分类树中公共子序列之间的相似性,当两者的相似度超过一定的阈值后,可以判定为正确.比如“秃鹰”和“鹰”的相似度为 0.96,而“秃鹰”和“鸟”的相似度为 0.88.若设定阈值为 0.85,则上述答案均可视为正确答案.WUPS 度量的方法是评估 DAQUAR 数据集和 CoCo-QA 数据集的标准度量,但是 WUPS 度量对于某些词在词汇上非常相似,但含义却大相径庭给出相似的分数,并且其只适用于严格的语义概念,这些概念几乎都是单个单词,不能评价短语或句子答案.VQA 数据集[78]中的答案由注释者给出 10 个答案,VQA 数据集的准确性度量标准由下式确定:

$$Accuracy_{VQA}=\min\left(\frac{n}{3},1\right),$$

其中,n 为预测答案与注释者给出答案相同的数量.换言之,如果预测答案至少与 3 个注释者提供的答案相同,则认为预测答案是 100%准确的.这种度量方式为大部分研究者所采用,但是其仍有其局限性,注释者针对同一问题给出的答案不尽相同,甚至有的答案含义相反,COCO-VQA 数据集中的注释者拥有共识的问题占比仅为 83.3%.其中,超过 59%的问题中,只有不到 3 个注释者给出完全相同的答案,这使得无法在这些问题上获得满分.并且当遇到答案为单个单词时,正确答案的可能性会大大增加.注释者对答案的描述,同样影响最终的准确率.

在 VQA 数据集中,问题类型和答案的分布偏斜.比如在“是/否”问题中,71%的问题的答案为“是”,如果每个测试问题都得到同等对待,则很难评估在较罕见的问题类型上的表现并弥补偏差.Kafle 等人[14]提出了多种措施来补偿偏差和偏斜分布.由于 TDIUC 数据集[14]的问题分为 12 种类型,分别计算了 12 种问题类型的准确性.目前,大部分研究将问题类型分为计数、是/否以及其他这 3 类.总体指标是每个问题类型的所有准确性的算术或调和均值,调和均值衡量标准具有在所有问题类型上均具有高分并偏向性能最低的类别的能力.使用归一化的指标对问题类型内答案分布不平衡补偿偏差,计算每个唯一答案的准确性,然后将其平均化为问题类型的准确率.若模型未归一化的分数与归一化的分数之间存在巨大差异,说明该模型无法推广到更稀有的答案.

4 挑战和展望

视觉问答任务是计算机视觉领域一个非常严峻的挑战,其拥有非常广泛的应用前景.尽管近几年视觉问答任务发展迅速,各种通用数据集或某一特定问题的数据集被不断提出,然而目前的视觉问答模型尚不能实现真正意义上的问答,不能够与人类进行良好的互动,其仍需要不断地进行研究.总的来说,目前的视觉问答任务仍处于一个起步阶段,各个方面还存在着诸多问题和挑战.比如:

(1) 特征表示能力不足

视觉问答模型的输入特征在提取的过程将图像和文本信息的部分信息丢失,目前的视觉特征和文本特征不足以进行问题回答的推理,这依赖于日后得到更好的特征提取和特征表示方法的出现.目前,传统的特征融合方法过于简单,日后需研究如何将视觉特征和文本特征更好地进行融合,使得融合后的特征含有更丰富的信息.目前的特征融合后得到的特征一般用来作为分类器的输入,日后的工作应更好地建立融合后特征与答案之间的关联.

(2) 模型评估能力不足

当前,大部分研究将视觉问答任务视为多分类任务,但多分类任务只能得到训练集中出现过的答案,这不符合人工智能的最终目标.生成式答案则更符合正常的逻辑,但其受限于答案的评估,目前的方法尚不能准确地评估预测答案是否与地标答案一致.其中,句子答案中存在语义、语法等问题,需要更加准确的评价标准对生成式视觉问答任务进行评估.

(3) 模型推理能力不足

当前,大部分视觉问答模型着力在得到更好的视觉和文本特征,缺乏根据问题对图片内容进行推理的能力,组合式模型在自然图像上表现仍不尽人意,不能将自然图像转化成推理的过程.虽然注意力机制能使模型更加关注某一重要区域或单词,但是模型在推理方面仍缺乏可解释性.

(4) 模型的鲁棒性与泛化能力不足

近年来,许多研究集中在如何消除视觉问答模型的语言相关性,消融研究[11,92]显示,仅问题模型的性能比仅图像模型好得多.这表明模型更倾向于利用文本信息回答问题.由于视觉问答数据集存在偏见,模型会利用数据集分布偏见达到很好的效果,但这导致训练集与测试集的结果有很大差异,模型的鲁棒性和泛化能力需要进一步提高,消除模型的表面相关性是实现这一目标的重要步骤.

因此,未来的研究工作可以从以下方面展开.

(1) 构建更全面均衡的数据集

当前的通用数据集在衡量各项能力时并不均衡,比如有关于图像中文本、计数等问题在通用数据集中的比例不高.不均衡的数据集并不能准确地衡量视觉问答模型的能力.同时,当前针对模型的评价标准仍需要提高,进一步研究对于生成式答案的视觉问答模型的评价标准.

(2) 提高模型的可解释性

当人类回答问题时,会根据问题进行推理,寻找可以支持答案的证据.在构建数据集时加入支持证据,让模型在每一次预测时提供回答问题的支持证据,基于VQA-X数据集进一步提高模型的可解释性,将目前注意力方法中的注意力权重可以着重表示重要区域的方式与文本解释相结合,研究模型给出更合理的解释方式,这也是

未来的研究方向.

(3) 提高模型的鲁棒性和泛化能力

首先应尽力消减数据集中存在的各种偏见问题,答案分布应更加合理,使得模型无法利用数据集中的偏见不经过推理得到问题的答案.在模型方面,多种方法应结合发展,将组合式方法和注意力方法结合应用.若视觉问答模型需要回答全部的问题,视觉回答模型必然要考虑利用外部知识.

5 结束语

本文总结了视觉问答的研究现状,介绍了当前主要的数据集,分析了目前数据集存在的偏见.总结目前主流的模型方法,联合嵌入方法几乎是所有模型方法的基础,注意力方法帮助模型更加关注图像中某部分区域或问题中重要的单词.组合方法和图结构使模型更加注重推理的过程,符合人类回答问题的逻辑.外部知识使得模型能够回答更加复杂的问题.部分研究针对模型存在的各种鲁棒性问题,如语言偏见、软注意力导致计数困难、有关图片中的文本问题回答困难等.除此之外,我们认为,目前的视觉问答模型的瓶颈在于提取的特征不足以回答问题.相信:随着各个计算机视觉任务的不断发展,视觉问答任务的目标一定会实现.

References:

[1] Szegedy C, Vanhoucke V, Ioffe S, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 2818−2826. [doi: 10.1109/CVPR.2016.308]

[2] Huang G, Liu Z, Van Der Maaten LQ, Weinberger K. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4700−4708. [doi: 10.1109/CVPR.2017.243]

[3] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 779−788. [doi: 10.1109/CVPR.2016.91]

[4] Lin T, Dollár P, Girshick R, He KM, Harharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2117−2125. [doi: 10.1109/CVPR.2017.106]

[5] Zhu G, Zhang L, Shen P, Shen PY, Song J. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 2017,5:4517−4524.

[6] Narayana P, Beveridge R, Draper BA. Gesture recognition: Focus on the hands. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 5235−5244. [doi: 10.1109/CVPR.2018.00549]

[7] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2625−2634. [doi: 10.1109/CVPR.2015.7298878]

[8] Karpathy A, Joulin A, Li FF. Deep fragment embeddings for bidirectional image sentence mapping. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 1889−1897.

[9] Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A. Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proc. of the Int'l Conf. on Learning Representations. 2015.

[10] Bajaj P, Campos D, Craswell N. Ms Marco: A human-generated machine reading comprehension dataset. arXiv preprint arXiv: 1611.09268, 2016.

[11] Hu M, Peng Y, Huang Z, Qiu X, Wei F, Zhou M. Reinforced mnemonic reader for machine reading comprehension. arXiv preprint arXiv:1705.02798, 2017.

[12] Xian GJ, Huang YZ. A review of research on visual question-answering technology based on neural network. Network Security Technologies and Applications, 2018(1):42−47 (in Chinese with English abstract).

[13] Yu J, Wang L, Yu Z. Research on visual question answering techniques. Journal of Computer Research and Development, 2018, 55(9):1946−1958 (in Chinese with English abstract).

[14] Kafle K, Kanan C. An analysis of visual question answering algorithms. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1965−1973. [doi: 10.1109/ICCV.2017.217]

[15] Wu Q, Teney D, Wang P, Shen CH, Dick A, Van Den Hengel A. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 2017,163:21−40.

[16] Kafle K, Kanan C. Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 2017,163:3−20.

[17] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6904−6913. [doi: 10.1109/CVPR.2017.670]

[18] Ramakrishnan S, Agrawal A, Lee S. Overcoming language priors in visual question answering with adversarial regularization. In: Proc. of the Advances in Neural Information Processing Systems. 2018. 1541−1551.

[19] Yang Z, He X, Gao J, Deng L, Smola A. Stacked attention networks for image question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 21−29. [doi: 10.1109/CVPR.2016.10]

[20] Deng J, Dong W, Socher R, Li L, Li K, Li F. Imagenet: A large-scale hierarchical image database. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2009. 248−255. [doi: 10.1109/CVPR.2009.5206848]

[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the Int'l Conf on Learning Representations. 2015.

[22] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 770−778. [doi: 10.1109/CVPR.2016.90]

[23] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 1−9. [doi: 10.1109/CVPR.2015.7298594]

[24] Anderson P, He X, Buehler C, Teney D, Johnson M, Dould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6077−6086. [doi: 10.1109/CVPR.2018.00636]

[25] Ren S, He K, Girshick R, Sun J. Faster *r*-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 91−99.

[26] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735−1780.

[27] Cho K, Van Merriënboer B, Gulcehre C, Bahdnau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1724−1734.

[28] Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S. Skip-thought vectors. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 3294−3302.

[29] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images. In: Proc. of the IEEE Int'l Conf on Computer Vision. 2015. 1−9. [doi: 10.1109/ICCV.2015.9]

[30] Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W. Are you talking to a machine? Dataset and methods for multilingual image question. In: Proc. of the Advances in Neural Information Processing Systems. 2015. 2296−2304.

[31] Noh H, Hongsuck Seo P, Han B. Image question answering using convolutional neural network with dynamic parameter prediction. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 30−38. [doi: 10.1109/CVPR.2016.11]

[32] Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 457−468.

[33] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li J, Shamma DA, Bernstein MS, Li F. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123(1):32−73.

[34] Kim JH, On KW, Lim W, Kim J, Ha J, Zhang B. Hadamard product for low-rank bilinear pooling. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[35] Yu Z, Yu J, Fan J, Tao P. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proc. of the IEEE Int'l Conf on Computer Vision. 2017. 1821−1830. [doi: 10.1109/ICCV.2017.202]

[36] Yu Z, Yu J, Xiang C, Fan J, Tao D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans. on Neural Networks and Learning Systems, 2018,29(12):5947−5959.

[37] Ben-Younes H, Cadene R, Cord M, Thome N. Mutan: Multimodal tucker fusion for visual question answering. In: Proc. of the IEEE Int'l Conf on Computer Vision. 2017. 2612−2620. [doi: 10.1109/ICCV.2017.285]

[38] Ben-Younes H, Cadene R, Thome N, Cord M. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 8102−8109.

[39] Kim JH, Lee SW, Kwak D, Caramanis C. Multimodal residual learning for visual QA. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 361−369.

[40] Saito K, Shin A, Ushiku Y, Harada T. Dualnet: Domain-invariant network for visual question answering. In: Proc. of the IEEE Int'l Conf. on Multimedia and Expo (ICME). IEEE, 2017. 829−834.

[41] Gao P, You H, Zhang Z, Wang X, Li H. Multi-modality latent interaction network for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 5825−5835. [doi: 10.1109/ICCV.2019.00592]

[42] Do T, Do TT, Tran H, Tjiputra E, Tran QD. Compact trilinear interaction for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 392−401. [doi: 10.1109/ICCV.2019.00048]

[43] Bro R, Harshman RA, Sidiropoulos ND, Lundy ME. Modeling multiway data with linearly dependent loadings. Journal of Chemometrics: A Journal of the Chemometrics Society, 2009,23(7-8):324−340.

[44] Wang W, Shen J, Dong X, Borji A. Salient object detection driven by fixation prediction. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 1711−1720. [doi: 10.1109/CVPR.2018.00184]

[45] Ke L, Pei W, Li R, Shen X, Tai Y. Reflective decoding network for image captioning. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 8888−8897. [doi: 10.1109/ICCV.2019.00898]

[46] Xiao T, Li Y, Zhu J, Yu Z, Liu T. Sharing attention weights for fast transformer. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. 2019. 5292−5298.

[47] Xu K, Ba J, Kiros R, Cho K, Courvile A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. Int'l Conf. on Machine Learning. 2015. 2048−2057.

[48] Zhu Y, Groth O, Bernstein M, Li F. Visual7w: Grounded question answering in images. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4995−5004. [doi: 10.1109/CVPR.2016.540]

[49] Shih KJ, Singh S, Hoiem D. Where to look: Focus regions for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4613−4621. [doi: 10.1109/CVPR.2016.499]

[50] Patro B, Namboodiri VP. Differential attention for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7680−7688. [doi: 10.1109/CVPR.2018.00801]

[51] Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. In: Proc. of the Advances in Neural Information Processing Systems. 2016. 289−297.

[52] Nam H, Ha JW, Kim J. Dual attention networks for multimodal reasoning and matching. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 299−307. [doi: 10.1109/CVPR.2017.232]

[53] Nguyen DK, Okatani T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6087−6096. [doi: 10.1109/CVPR.2018.00637]

[54] Yu D, Fu J, Mei T, Rui Y. Multi-level attention networks for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 4709−4717. [doi: 10.1109/CVPR.2017.446]

[55] Wang P, Wu Q, Shen C, Van Den Hengel A. The VQA-machine: Learning how to use existing vision algorithms to answer new questions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1173−1182. [doi: 10.1109/CVPR.2017.416]

[56] Wu Q, Wang P, Shen C, Reid I, Van Den Hengel A. Are you talking to me? Reasoned visual dialog generation through adversarial learning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6106−6115. [doi: 10.1109/CVPR.2018.00639]

[57] Yu Z, Yu J, Cui Y, Tao D, Tian Q. Deep modular co-attention networks for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 6281−6290. [doi: 10.1109/CVPR.2019.00644]

[58] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Lukasz K, Polosukhin I. Attention is all you need. In: Proc. of the Advances in Neural Information Processing Systems. 2017. 5998−6008.

[59] Gao P, Jiang Z, You H, Lu P, Hoi S, Wang X, Li H. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 6639−6648. [doi: 10.1109/CVPR.2019.00680]

[60] Teney D, Anderson P, He X, Van Den Hengel A. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4223−4232. [doi: 10.1109/CVPR.2018.00444]

[61] Lu P, Li H, Zhang W, Wang J, Wang X. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. arXiv preprint arXiv:1711.06794, 2017.

[62] Wu C, Liu J, Wang X, Dong X. Object-difference attention: A simple relational attention for visual question answering. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. 2018. 519−527.

[63] Cadene R, Ben-Younes H, Cord M, Thome N. Murel: Multimodal relational reasoning for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 1989−1998. [doi: 10.1109/CVPR.2019.00209]

[64] Li L, Gan Z, Cheng Y, Liu J. Relation-aware graph attention network for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2019. 10313−10322. [doi: 10.1109/ICCV.2019.01041]

[65] Andreas J, Rohrbach M, Darrell T, Klein D. Neural module networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 39−48.

[66] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics. 2003. 423−430.

[67] De Marneffe MC, Manning CD. The Stanford typed dependencies representation. In: Proc. of the Workshop on Cross-framework and Cross-domain Parser Evaluation. 2008. 1−8.

[68] Andreas J, Rohrbach M, Darrell T, Klein D. Learning to compose neural networks for question answering. In: Proc. of the Annual Conf. of the North American Chapter of the Association for Computational Linguistics. 2016. 1545-1554.

[69] Hu R, Andreas J, Rohrbach M, Darrell T, Saenko K. Learning to reason: End-to-end module networks for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 804−813.

[70] Kumar A, Irsoy O, Ondruska P, Iyyer M, Bradbury J, Gulrajani I, Zhong V, Paulus R, Socher R. Ask me anything: Dynamic memory networks for natural language processing. In: Proc. of the Int'l Conf. on Machine Learning. 2016. 1378−1387.

[71] Xiong C, Merity S, Socher R. Dynamic memory networks for visual and textual question answering. In: Proc. of the Int'l Conf. on Machine Learning. 2016. 2397−2406.

[72] Noh H, Han B. Training recurrent answering units with joint loss minimization for VQA. arXiv preprint arXiv:1606.03647, 2016.

[73] Wang P, Wu Q, Shen C, Van Den Hengel A, Dick A. Explicit knowledge-based reasoning for visual question answering. arXiv preprint arXiv:1511.02570, 2015.

[74] Wang P, Wu Q, Shen C, Dick A, Van Den Hengel A. FVQA: Fact-based visual question answering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018,40(10):2413−2427.

[75] Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. Dbpedia: A nucleus for a Web of open data. In: Proc. of the Semantic Web. Berlin, Heidelberg: Springer-Verlag, 2007. 722−735.

[76] Wu Q, Wang P, Shen C, Dick A, Van Den Hengel A. Ask me anything: Free-form visual question answering based on knowledge from external sources. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4622−4630. [doi: 10.1109/CVPR.2016.500]

[77] Wu Q, Shen C, Wang P, Dick A, Van Den Hengel A. Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,40(6):1367−1381.

[78] Agrawal A, Batra D, Parikh D. Analyzing the behavior of visual question answering models. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. 2016. 1955−1960.

[79] Zhang P, Goyal Y, Summers-Stay D, Batra D, Parikh D. Yin and yang: Balancing and answering binary visual questions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5014−5022. [doi: 10.1109/CVPR.2016.542]

[80] Shah M, Chen X, Rohrbach M, Parikh D. Cycle-consistency for robust visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 6649−6658.

[81] Xu X, Chen X, Liu C, Rohrbach A, Darrell T, Song D. Fooling vision and language models despite localization and attention mechanism. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4951−4961. [doi: 10.1109/CVPR.2018.00520]

[82] Agrawal A, Batra D, Parikh D, Kembhavi A. Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 4971−4980. [doi: 10.1109/CVPR.2018.00520]

[83] Chen L, Yan X, Xiao J, Zhang H, Pu S, Zhuang Y. Counterfactual samples synthesizing for robust visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2020. 10800−10809.

[84] Grand G, Belinkov Y. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In: Proc. of the 57th Conf. on Computational Natural Language Learning. ACL, 2019. 1−13.

[85] Belinkov Y, Poliak A, Shieber SM, Durme BV, Rush AM. Don't take the premise for granted: Mitigating artifacts in natural language inference. In: Proc. of the 57th Conf. on Computational Natural Language Learning. ACL, 2019. 877−891.

[86] Cadene R, Dancette C, Cord M, Parikh D. Rubi: Reducing unimodal biases for visual question answering. In: Proc. of the Advances in Neural Information Processing Systems. 2019. 841−852.

[87] Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing. 2019. 4069−4082.

[88] Mahabadi RK, Henderson J. Simple but effective techniques to reduce biases. arXiv preprint arXiv:1909.06321, 2019.

[89] Wu J, Mooney R. Self-critical reasoning for robust visual question answering. In: Proc. of the Advances in Neural Information Processing Systems. 2019. 8604−8614.

[90] Singh A, Natarajan V, Shah M, Jiang Y, Chen X. Towards VQA models that can read. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 8317−8326. [doi: 10.1109/CVPR.2019.00851]

[91] Biten AF, Tito R, Mafla A, Gomez L, Rusinol M, Valveny E, Jawahar CV, Karatzas D. Scene text visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 4291−4301. [doi: 10.1109/CVPR.2019.00851]

[92] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. of the IEEE Int'l Conf on Computer Vision. 2017. 2223−2232. [doi: 10.1109/ICCV.2017.244]

[93] Zhang Y, Hare J, Prügel-Bennett A. Learning to count objects in natural images for visual question answering. In: Proc. of the Int'l Conf. on Learning Representations. 2018.

[94] Acharya M, Kafle K, Kanan C. TallyQA: Answering complex counting questions. In: Proc. of the AAAI Conf. on Artificial Intelligence, Vol.33. 2019. 8076−8084.

[95] Shrestha R, Kafle K, Kanan C. Answer them all! Toward universal visual question answering models. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2019. 10472−10481. [doi: 10.1109/CVPR.2019.01072]

[96] Hudson D, Manning CD. Learning by abstraction: The neural state machine. In: Proc. of the Advances in Neural Information Processing Systems. 2019. 5903−5916.

[97] Shi Y, Furlanello T, Zha S, Anandkumar A. Question type guided attention in visual question answering. In: Proc. of the European Conf. on Computer Vision (ECCV). 2018. 151−166.

[98] Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 1682−1690.

[99] Ren M, Kiros R, Zemel R. Image question answering: A visual semantic embedding model and a new dataset. arXiv preprint arXiv:1505.02074, 2015.

[100] Antol S, Agrawal A, Lu J, Antol S, Mitchell M, Zitnick L, Batra D, Parikh D. VQA: Visual question answering. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2425−2433.

[101] Huk Park D, Anne Hendricks L, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M. Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 8779−8788. [doi: 10.1109/CVPR.2018.00915]

[102] Johnson J, Hariharan B, van der Maaten L, Li F, Zitnick CL, Girshick R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 2901−2910. [doi: 10.1109/CVPR.2017.215]

[103] Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollar P. Microsoft coco: Common objects in context. In: Proc. of the European Conf. on Computer Vision (ECCV). 2014. 740−755.

[104] Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. 2018. 71−79.

[105] Chattopadhyay P, Vedantam R, Selvaraju RR, Batra D, Parikh D. Counting everyday objects in everyday scenes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1135−1144. [doi: 10.1109/CVPR.2017.471]

[106] Trott A, Xiong C, Socher R. Interpretable counting for visual question answering. In: Proc. of the Int'l Conf. on Learning Representations. 2017. 133−138.

[107] Zitnick CL, Agrawal A, Antol S, Mitchell M, Batra D, Parikh D. Measuring machine intelligence through visual question answering. AI Magazine, 2016,37(1):63−72.

[108] Wu Z, Palmer M. Verb semantics and lexical selection. In: Proc. of the Conf. on Association for Computational Linguistics. 1994.

附中文参考文献:

[12] 鲜光靖,黄永忠.基于神经网络的视觉问答技术研究综述.网络安全技术与应用,2018(1):42−47.

[13] 俞俊,汪亮,余宙.视觉问答技术研究.计算机研究与发展,2018,55(9):1946−1958.

包希港(1997－),男,博士生,主要研究领域为视觉问答,知识库问答.

肖克晶(1991－),女,博士生,主要研究领域为自然语言处理,深度学习,数据挖掘.

周春来(1976－),男,博士,副教授,CCF 专业会员,主要研究领域为人工智能不确定性.

覃飙(1972－),男,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为人工智能,因果分析和不确定数据库.