计算机科学
COMPUTER SCIENCE

基于深度学习的视觉问答研究综述

李祥, 范志广, 李学相, 张卫星, 杨聪, 曹仰杰

引用本文

李祥, 范志广, 李学相, 张卫星, 杨聪, 曹仰杰基于深度学习的视觉问答研究综述[J]. 计算机科学, 2023, 50(5): 177-188.

LI Xiang, FAN Zhiguang, LI Xuexiang, ZHANG Weixing, YANG Cong, CAO Yangjie. Survey of Visual Question Answering Based on Deep Learning [J]. Computer Science, 2023, 50(5): 177-188.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500124

基于深度学习的视觉问答研究综述

李 祥[1] 范志广[2] 李学相[1] 张卫星[1] 杨 聪[1] 曹仰杰[1]
1 郑州大学网络空间安全学院 郑州 450000
2 郑州大学河南先进技术研究院 郑州 450000
(lixiang.zg@qq.com)

摘 要 视觉问答是计算机视觉和自然语言处理的交叉领域。在视觉问答的任务中,机器首先需要对图像、文本这两种模态数据进行编码,进而学习这两种模态之间的映射,实现图像特征和文本特征的融合,最后给出答案。视觉问答任务考验模型对图像的理解能力以及对答案的推理能力。视觉问答是实现跨模态人机交互的重要途径,具有广阔的应用前景。最近相继涌现出了众多新兴技术,如基于场景推理的方法、基于对比学习的方法和基于三维点云的方法。但是,视觉问答模型普遍存在推理能力不足、缺乏可解释性等问题,值得进一步地探索与研究。文中对视觉问答领域的相关研究和新颖方法进行了深入的调研和总结。首先介绍了视觉问答的背景;其次分析了视觉问答的研究现状并对相关算法的和数据集进行了归纳总结;最后根据当前模型存在的问题对视觉问答的未来研究方向进行了展望。

关键词:视觉问答;跨模态;人机交互;推理能力;可解释性
中图法分类号 TP181

Survey of Visual Question Answering Based on Deep Learning

LI Xiang[1],FAN Zhiguang[2],LI Xuexiang[1],ZHANG Weixing[1],YANG Cong[1] and CAO Yangjie[1]
1 School of Cyber science and Engineering,Zhengzhou University,Zhengzhou 450000,China
2 Henan Institute of Advanced Technology,Zhengzhou University,Zhengzhou 450000,China

Abstract Visual question answering(VQA) is an interdisciplinary research paradigm that involves computer vision and natural language processing. VQA generally requires both image and text data to be encoded,their mappings learned,and their features fused,before finally generating an appropriate answer. Image understanding and result reasoning are therefore vital to the performance of VQA. With its importance in realizing cross-modal human-computer interaction and its promising applications,a number of emerging techniques for VQA,including scene-reasoning based methods,contrastive-learning based methods,and 3D-point-cloud based methods,have been recently proposed. These methods,while achieving notable performances,have revealed issues such as insufficient inferential capability and interpretability,which demand further exploration. We hence present in this paper an in-depth survey and summary of related research and proposals in the field of VQA. The essential background of VQA is first introduced,followed by the analysis and summarization of state-of-art approaches and datasets. Last but not least,with the insight of current issues,future research directions in the field of VQA are prospected.

Keywords Visual question answering,Cross-modal,Human-Computer interaction,Reasoning ability,Interpretability

1 引言

随着计算机技术不断地进步,视频、图像、语音、字幕等信息无处不在,其中每一种信息的形式可以称作一种模态。为了处理多源模态信息,多模态学习应运而生,成为了热门的研究方向。多模态学习包含图像-文本匹配[1]、图片描述[2]和视觉问答(Visual Question Answering,VQA)等常见任务。与其他多模态学习任务相比,视觉问答是一个更具挑战性的任务,它结合了计算机视觉和自然语言处理两个领域,需要对图像和文本进行细致的语义理解[3]。

此前人工智能领域的两大分支,即计算机视觉和自然语言处理,几乎没有任何交集。近几十年来,可获取和可处理的可视化数据和文本数据的爆炸式增长使得这两个领域迅速发展。计算机视觉任务研究如何让机器处理和理解图像中的内容,其主要研究内容包括图像分类[4-6]、图像分割[7-11]、图像生成[12-15]和目标检测[16-21]等。自然语言处理是一个机器分析

到稿日期:2022-05-16 返修日期:2022-09-05
基金项目:国家自然科学基金面上项目(61972092);郑州市协同创新重大专项(20XTZX06013)
This work was supported by the General Project of National Natural Science Foundation of China(61972092) and Collaborative Innovation Major Project of Zhengzhou(20XTZX06013).
通信作者:李学相(lxx@zzu.edu.cn)

和理解人类语言的过程，包括文章、句子和情感。它的主要研究任务包括情感识别[22-24]、意图识别[25-27]和机器翻译[28-30]等。研究者不再满足于计算机对图像进行基础的感知，而是希望其能够对图像进行全局的理解并进一步具有推理能力，最终以友好的人机交互的方式呈现出来。

在常见的视觉问答任务中，将图像和文本这两种跨模态的数据输入计算机，计算机对这两种模态的数据进行感知和理解，给出问题对应的正确答案[31]。答案通常是一个单词、数字或者由几个单词组成的短语。视觉问答任务的例子如图1所示。视觉问答任务的形式可以分为开放式和多项选择式两大类。对于开放式的视觉问答任务，问题的答案是不确定的，计算机没有任何参考，只是根据图像和问题生成自然语言作为答案。相反，多项选择式的视觉问答任务比较简单，机器已知问题的几个候选答案，在对图像和问题进行推理之后选择出最正确的答案。

图1 视觉问答的样本

Fig. 1 Samples of visual question answering

视觉问答任务比其他的计算机视觉任务更具挑战性。常规的计算机视觉任务的形式单一，并且回答单个问题是预先确定的，只有输入图像发生了变化。然而视觉问答任务需要处理不同模态之间的信息，解决不同模态之间的“语义鸿沟”问题，通过跨模态数据之间的交互，建立统一的语义表达[32]。VQA任务更符合人工智能的真实形式，可以帮助模型更深入地理解视觉和语言。视觉问答的问题类型是多种多样的，问题的主要类型如下：

(1)二元问题——图像中是否有小球？

(2)计数问题——图像中共有多少个球？

(3)物体识别——图像中有什么？

(4)物体检测——图像中存在马吗？

(5)属性分类——图像中的猫是什么颜色？

(6)地点问题——图像中的男孩在哪儿？

(7)原因问题——为什么右边的男孩被吓坏了？

视觉问答还有很多潜在的应用。视觉问答任务可以帮助盲人和视障人士获得更多的信息，实现智能化的人机交互。视觉问答系统也能应用于医学领域，协助医生阅读医学影像，随时随地提供答案。视觉问答还可以应用于图像检索领域，在不使用标签的情况下根据问题搜索相关的图像。视觉问答任务涉及计算机视觉领域许多基础的研究，其发展必将提高机器对图像深层次的理解能力。

视觉问答任务自2014年以来取得了快速的发展。对于早期的视觉问答，其图像文本特征的交互方式较为简单，主要采用联合嵌入的方法。后来出现了基于注意力机制的跨模态交互方法，根据问题着重关注与问题关键词相关的区域。随着研究的深入，NLP领域的Transformer结构[33]被引入到视觉问答任务中，Transformer中的多头注意力起到了关键的作用。通过Transformer的多头注意力，加强了图像和文本特征的细粒度交互，使融合更加充分，取得了不错的性能。但是，有些问题往往需要借助额外知识才能给出正确答案。研究人员往往会引入外部知识库，如知识图谱或者维基百科等，根据问题在外部知识库中搜索答案。最近，对比学习的思想被引入到视觉问答任务中，使得不同嵌入空间的图像特征和文本特征更容易交互。此外，基于三维点云的方法最近被提出，该模型有很大的改进空间，是未来重要的研究趋势。

到目前为止，已有多篇综述详细地介绍了视觉问答任务中的经典方法，但是它们很少涉及基于场景推理的方法、基于对比学习的方法和基于三维点云的方法。这些方法近年来引起了较多的关注，具有一定的研究意义。本文重点介绍了这3类新方法，并对以前的综述做了进一步的扩充。

本文综述了基于深度学习的视觉问答算法。第1节介绍了视觉问答研究的背景；第2节对视觉问答的算法进行了分类总结；第3节归纳整理了视觉问答主要的数据集和评价标准；第4节结合视觉问答算法存在的不足对未来的发展趋势进行了讨论；最后总结全文。

2 模型介绍

自视觉问答任务被提出以来，国内外研究人员提出了各种性能不错的模型。大部分模型都可以总结为一个框架。该框架包括4个部分：图像特征化模块、问题特征化模块、特征融合模块、得出答案模块。图像特征化模块主要采用VGG-Net[34]，ResNet[5]和GoogLeNet[35]提取图像特征。随着目标检测的不断发展，利用Faster R-CNN[18]提取图像特征成为了主流。问题特征化模块主要利用LSTM[36]，GRU[37]，Transformer[33]和BERT[38]等语言编码模型提取问题特征。特征融合模块的作用是把图像特征和文本特征映射到同一特征空间并进行交互融合。该模块是VQA模型的核心部分。得出答案模块可以采用分类方法或者生成方法。对于多项选择式的任务，融合特征被送入分类器，得到每个候选答案的可能性分数，选出分数最高的候选答案作为正确答案；对于开放式任务，融合后的特征被送入RNN或LSTM等模型中生成答案。视觉问答模型的大概框架如图2所示。

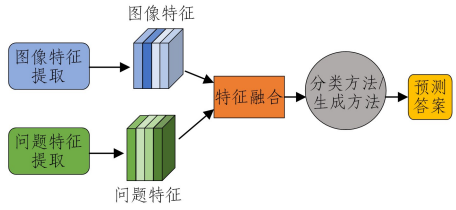

图2 视觉问答总体框架图

Fig. 2 Overall framework of visual question answering

本节按照模型的特点将其分为 6 个类别：基于联合嵌入的方法、基于注意力机制的方法、基于场景推理的方法、基于外部知识的方法、基于对比学习的方法和基于三维点云的方法。针对这 6 类模型，详细描述了模型提出的原因、模型的思想、模型之间的联系以及模型存在的问题。

2.1　基于联合嵌入的方法

为了完成视觉问答任务，研究人员首先探索了图像与文本联合嵌入的概念。联合嵌入的方法往往通过暴力操作将两种特征整合，如串联、逐元素乘法或逐元素加法等。基于联合嵌入的方法大致流程如图 3 所示。

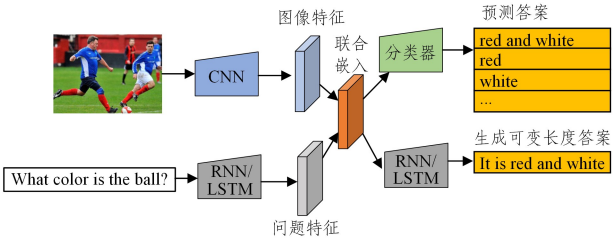

图 3　基于联合嵌入的方法流程图

Fig. 3　Flowchart of joint-embedding based methods

在视觉问答模型中，最早研究联合嵌入方法的是 Malinowski 等[39]。他们通过结合图像表示和自然语言处理的最新进展，提出了 Neural-Image-QA，用一种新的基于循环神经网络的方法来解决图像问题中的挑战性任务。该模型将 CNN 与 LSTM 结合到一个端到端的架构中，预测问题和图像的答案。Neural-Image-QA 模型通过生成的方式得到答案。然而 Gao 等[40]提出了 mQA 模型，他们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答案。

受深度残差结构的启发，Kim 等[41]提出了一种用于视觉问答多模态残差学习的多模态残差网络（Multimodal Residual Networks，MRN），扩展了深度残差学习的思想。与深度残差学习不同，MRN 能有效地从视觉和语言信息中学习联合表示，其主要思想是使用元素乘法用于联合残差映射。与大多数模型同时使用图像编码器和问题编码器不同，Ma 等[42]提出的模型由 3 个 CNN 组成：一个图像 CNN 对图像内容进行编码；一个句子 CNN 对问题进行编码；一个多模态卷积层学习它们的联合表示，用于在候选答案单词的空间中分类。

基于联合嵌入的方法只是对图像特征和文本特征进行简单地拼接，并没有将图像区域与问题关键词对应起来。基于联合嵌入的方法直接利用全部的视觉信息和文本信息得到答案。但是，视觉特征和文本特征中有很大部分信息都是无用信息，会干扰最终的答案分类或答案生成。这种融合方法比较粗糙，没有根据问题做出推理，因此还有较大的改进空间。

2.2　基于注意力机制的方法

深度学习中的注意力方法模仿了人类的注意力机制。人类通过快速扫描图像或者文本，关注重点区域和关键词，从而快速理解主要信息。注意力机制快速迁移到人工智能的各个领域，并取得了不错的性能。在视觉问答任务中，注意力方法增强了模型对图像和语义的理解能力。近年来，注意力的方法大致可以分为问题引导的注意力方法、协同注意力方法和多粒度注意力方法。下面主要围绕这 3 类方法展开介绍。

2.2.1　问题引导的注意力方法

早期的注意力方法是利用问题来计算图像的各个区域的重要性，找到与问题联系密切的区域。例如，Shih 等[43]通过选择与基于文本的查询相关的图像区域来学习回答视觉问题，将视觉特征与文本特征简单相乘得到注意力权重，注意力权重与视觉特征相乘后更新视觉特征。Kazemi 等[44]使用基于 ResNet 的卷积神经网络来提取图像特征，输入问题被送入多层 LSTM 中。然后，利用拼接后的图像特征和 LSTM 的最终状态，计算图像特征上的多个注意力分布。已有方法只计算了一次视觉注意力分布，为了进一步提升跨模态特征融合效果，Yang 等[45]提出了堆叠注意力网络模型（Stacked Attention Network，SAN）。该模型含有多个步骤的推理过程，建立了多层注意力机制，根据问题的语义对一个图像进行多次查询，以逐步推断出答案。

2.2.2　协同注意力方法

协同注意力方法对早期的注意力做了进一步完善。协同注意力方法不仅考虑问题引导获得图像特征的注意力，还考虑利用图像特征得到问题的注意力。

为了提升模型性能，视觉问答需要同时对视觉内容和文本内容有精细的理解。具体来说，将问题中的关键词与图像中的关键物体联系起来成为了一个挑战。Yu 等[46]提出深度模块化协同注意力网络，网络框架借鉴了 Transformer 模型，编码器部分由 6 个自注意力单元堆叠而成，主要用于文本内部的交互；解码器部分由自注意力单元和引导注意力单元组合而成，目的是实现文本特征和图像特征这两种模态之间的交互。后期有很多研究者基于 Transformer 结构做出了改进。这类模型的大概流程图如图 4 所示。Rahman 等[47]提出了一种改进的基于注意力的网络结构。在编码器-解码器框架中加入了 AoA 模块，该模块能够确定注意力结果和查询之间的关系，注意模块为每个查询生成加权平均。另外，他们还提出了多模态融合模块，将视觉信息和文本信息结合起来。该融合模块的目标是动态地决定考虑多少视觉信息和文本信息。在视觉问答任务中，多模态预测往往需要从宏观到微观的视觉信息。因此，如何在 Transformer 中动态调度全局和局部依赖关系成为了一个新问题。Zhou 等[48]提出了一个依赖于实例的路由机制——TRAR，来解决这个问题。在 TRAR 中，每个可视化 Transformer 层都配备了具有不同注意力广度的路由模块。该模型可以根据前一个推理步骤的输出动态地选择相应的注意范围，从而为每个实例制定出最优的路由路径。

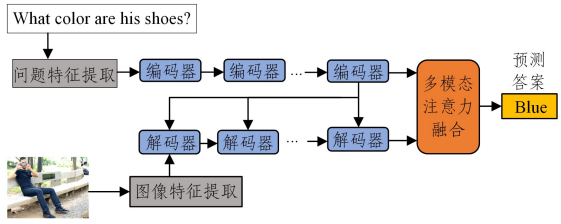

图 4　协同注意力方法的流程图

Fig. 4　Flowchart of co-attention methods

基于 Transformer 结构的方法在视觉问答中取得了巨大的成功。然而,这些模型往往具有更深的网络和更广的嵌入维度,这使得其很难被部署在资源受限的平台上。设计支持运行时自适应剪枝的 VQA 模型以满足不同平台的效率约束,是一项有价值的任务。Yu 等[49]提出了双重轻量级 Transformer(DST),这是一个通用框架,可以与任何基于 Transformer 的 VQA 模型无缝集成。DST 在宽度和深度上简化模型,一次训练一个单一的模型,并获得了多种可自适应不同平台的高效子模型。此外,Wang 等[50]提出了跨模态注意力蒸馏框架来训练双编码器模型,从而完成视觉语言理解任务。该框架采用融合编码器模型(教师)中的图像-文本和文本-图像注意力分布来进行蒸馏,从而指导双编码器模型的训练。

2.2.3 多粒度注意力方法

现有的视觉语言模型要么采用细粒度的以对象为中心的图像特征对齐文本,要么采用粗粒度的整体图像特征对齐文本。这两种方法虽然有效,但仍存在一些不足。细粒度检测可以识别图像中所有可能的对象,但是其中一些对象可能与文本无关。以对象为中心的特征不能很容易地表示多个对象之间的关系。另一方面,粗粒度方法不能有效地学习视觉和语言之间的细粒度对齐。

Zeng 等[51]提出了一种新的方法 X-VLM 来进行多粒度视觉语言预训练。他们将现有的数据集重构为视觉概念和对应的文本。视觉概念可以是一个物体、一个区域或图像本身。模型将文本与相关的视觉概念对齐,对齐方式是多粒度的。由于图像特征具有高度的多样性,缺乏语言的结构和语法规则,因此语言特征很可能丢失细节信息。为了更好地学习视觉和文本之间的注意力,Xiong 等[52]提出了一种新的多粒度对齐架构,该架构可以在概念-实体水平、区域-名词短语水平和空间-句子水平 3 个不同层次上联合学习模态内和模态间的相关性,然后构造了一个决策融合模块来合并来自不同粒度 Transformer 模块的输出。

为了同时预训练用于多模态表示提取的编码器和用于句子生成的语言解码,Li 等[53]提出了一个预训练的通用编码器-解码器网络(Uni-EDEN),以促进视觉语言感知和生成。该模型通过多粒度的视觉语言代理任务对 Uni-EDEN 进行预训练:屏蔽物体分类(MOC),屏蔽区域短语生成(MRPG),图像-句子匹配(ISM),屏蔽句子生成(MSG)。多粒度的视觉语言代理任务旨在更好地将视觉内容与不同粒度的语言表征(从单个标签、短语到自然句子)进行对齐。

注意力方法是视觉问答任务中的主流方法,受到大多数研究者的关注。基于注意力方法的模型不断被改进,取得了卓越的性能。但是,注意力方法只是更加关注图像区域和文本关键词,没有感知到图像中物体间的关系,对于需要推理的问题几乎没有帮助。如何在视觉问答任务中融入推理链,并准确地定位与答案有关的图像区域需要进一步地探索。

2.3 基于场景推理的方法

场景图是对场景的结构化表示,能清晰地表达场景中的对象、属性以及对象之间的关系[54]。目前,人们已不再满足于简单地检测和识别图像中的物体,而是需要更高层次的视觉理解和推理任务来捕捉场景中物体之间的关系。场景图是理解场景的强大工具。因此,场景图引起了大量研究者的关注,而相关研究往往是跨模态的、复杂的,并且发展迅速。

基于场景图的视觉问答模型的主要思想是利用问题的语义线索引导视觉内容进行推理。该方法的大概框架图如图 5 所示。Yang 等[55]提出了一种利用先验视觉关系学习解决关系推理问题的新方法——场景图卷积网络(SceneGCN),该模型通过预训练的目标检测器和视觉关系编码器对场景图中的物体和关系进行向量化表示,然后使用了场景图卷积,场景图卷积运算利用物体和关系的信息更新每个节点的隐藏状态。与 SceneGCN 模型不同,Liang 等[56]提出了一种语言引导的图神经网络框架 GraphVQA,该框架首先将问题转换为 M 个指令向量,然后通过图神经网络对每个指令向量进行消息传递,最后对消息传递后的最终状态进行汇总并预测答案。例如,给出问题“拿着汉堡的女孩左边的红色物体是什么”。GraphVQA 把问题通过下面的形式进行传递“汉堡包→小女孩→红色托盘”,从而回答问题。

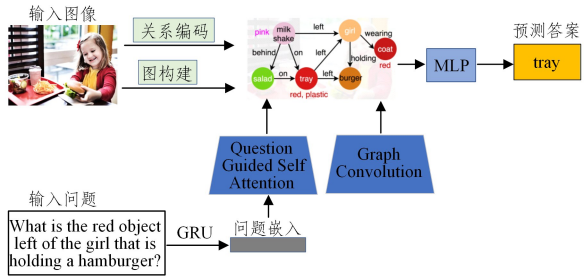

图 5 基于场景推理的方法框架

Fig. 5 Framework of scene-reasoning based methods

因为场景图具有实体以及语义和空间关系,所以将 VQA 任务建模为场景图上寻找路径的问题是可行的。Koner 等[57]提出了一种新的方法 Graphhopper,Graphhopper 是第一个将强化学习用于场景图的多跳推理的 VQA 方法。具体来说,Graphhopper 思想是训练强化学习 agent 根据问题内容在场景图上进行多跳自主导航生成推理路径,推理路径是得到答案的基础。与纯粹的基于嵌入的方法相比,Graphhopper 提供了明确的推理链,可以通过推理链引导得到答案。

在之前的研究中,图像可以通过场景图恰当地表示,但是问题总是简单地嵌入,不能很好地表示完整的语义。为此,Cao 等[58]提出了一种图匹配注意(Graph Matching Attention,GMA)网络,该算法不仅从物体的外观、几何特征、空间关系的角度构造图像的场景图,而且利用从问题中提取的语法树和语言特征对问题进行图的构造。首先,通过一个双阶段图编码器来获取模内关系,然后用双向跨模态图匹配注意力来推断图像和问题之间的关系,并相互传播跨模态信息。现有的视觉问答模型嵌入了各种信息,却没有细粒度地搜索答案,可能会引入额外的噪声数据,从而干扰模型给出正确答案。如何准确地获得面向问题的支持证据是一个关键性挑战。Zhu 等[59]用多模态异构图来描述图像,该图比普通的场景图具有更丰富的信息,包含与图像的视觉特征、语义特征和事实特征相关的多层信息,然后构建了一种模态感知的异构图卷积网络来迭代地选择和收集模态内和跨模态的证据信息。该方法在得出答案的过程中提供了良好的可解释性。

与传统的 VQA 方法相比，场景图可以以图结构的形式捕获图像的基本信息，这使得基于场景图的 VQA 方法优于传统算法。但是基于场景图的视觉问答算法仍然不够完善。由于模型可以根据问题在场景图中搜索出答案，该模型在关系推理的问题中有良好的表现，但在计数、原因、时间等问题上的效果并不理想。此外，基于场景图的视觉问答推理出答案的过程不够透明，其可解释性方面需要进一步研究。

2.4　基于外部知识的方法

视觉问答任务的问题通常是复杂多样的，仅仅依靠有限的视觉信息无法给出答案。此时，视觉问答模型需要从外部知识库中获得信息作为回答问题的支持证据。例如，对于“在这幅图像中哪个物体可以被用来保护头部”这个问题，必须理解保护头部是一种作用，然后搜索哪个物体有这个用处。基于外部知识的视觉问答方法是未来研究的趋势，尤其是在特定的专业领域具有一定的应用价值。

2.4.1　知识推理方法

Wang 等[60]提出了一种在大规模知识库中对图像内容进行推理的视觉问答方法 Ahab。Ahab 首先检测图像中的相关内容，并将其与知识库中可用的信息联系起来。将自然语言问题处理成合适的查询，并将图像和知识库信息结合起来进行查询。此查询可能需要多个推理步骤才能完成，根据查询的反馈形成最终答案。之前提出的方法只能向系统提出一组特定的问题，而所采用的查询生成方法要求使用非常特定的问题形式。为了适用于一般的知识库并回答广泛的问题，Wu 等[61]提出了一种视觉问答方法，该方法构建图像语义内容的文本表示，并将其与来自知识库的文本信息进行合并，以加深对所查看场景的理解，使得模型可以回答比以前更广泛、更复杂的问题。目前已有的方法都依赖于针对问题检索的基本事实，而真实世界的应用中可能会对知识图谱中不存在的事实提出问题。Ramnath 等[62]开发了一种新的问答架构，即使知识图谱中缺少所需边也能解决 FVQA 任务。在这个过程中，该方法结合了互补的词汇特征和知识图谱语义特征，提高了答案检索的准确性。

这些方法通常会以管道的方式构建模型，在知识图谱中进行查询，经常会导致错误级联。其次，VQA 推理能力较弱，不能预测训练集中不存在的答案，因此 Chen 等[63]提出了 Zero-shot 视觉问答算法。为了更好地整合外部知识，他们将 VQA 从传统的分类任务转换为基于映射的对齐任务，以实现对未知答案的预测。其中，图像/问题和知识图谱之间的对齐是隐式地通过多个特征空间来完成的。在答案预测模块，使用基于掩码的方法调整答案预测得分。这种软/硬掩码方法可以有效地增强对准过程，同时减少错误级联。

场景文本识别正逐渐地从实验室走向工业应用[64-66]。在基于知识推理的视觉问答模型中融入文本识别还没有被探索。Singh 等[67]提出了一个 VQA 模型，可以阅读场景文本，并在知识图谱上进行推理，以得到准确的答案。该模型无缝集成了视觉内容、识别词、问题和知识事实，并使用门控图神经网络（Gated Graph Neural Networks，GGNN）对多关系图进行推理。

知识推理方法在需要外部知识的复杂问题中取得了不错的效果，其大概流程图如图 6 所示。但是，基于知识推理的方法不能自适应地从大型知识图谱中选择特定领域的知识，推理过程相对简单。在未来的工作中，开发能够在知识图谱中选择小范围的知识从而进行多跳和更复杂推理的模型值得进一步研究。

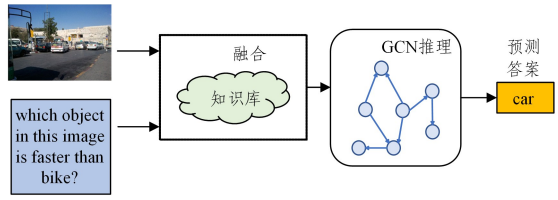

图 6　知识推理方法的框架

Fig. 6　Framework of knowledge-reasoning based methods

2.4.2　知识搜索方法

最近的研究开始着眼于如何将知识搜索方法整合到 VQA 中。这些方法研究了将知识库和检索方法与 VQA 数据集相结合，并为每个问题提供了一组相关的事实，该方法的大致流程如图 7 所示。Marino 等[68]提出了基于知识的基线 ArticleNet。首先，通过将问题中的单词与经过训练的图像和场景分类器识别出的单词相结合，为每个问题收集所有可能的查询。其次，使用 Wikipedia 搜索 API 为每个查询获取最热门的文章。然后，根据这些查询词在句子中的出现频率，在文章中选择最符合的句子。为了找到问题的答案，从检索的句子中挑选得分最高的单词。

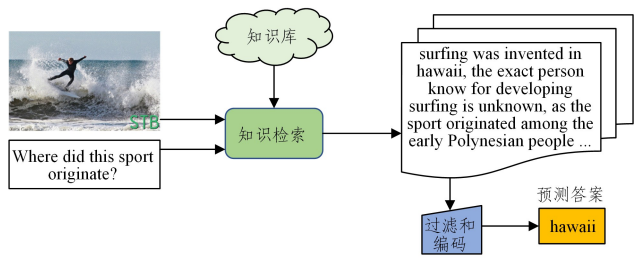

图 7　知识搜索方法的框架

Fig. 7　Framework of knowledge-searching based methods

基于知识搜索的视觉问答使用庞大的知识库，很有可能检索到不相关或者嘈杂的知识，使理解事实和找到答案变得困难。为了解决这个问题，Wu 等[69]引入了一种新的方法 MAVEx，考虑了 3 种知识来源，即 Wikipedia，ConceptNet 和 Google images，分别提供事实知识、常识知识和视觉知识。利用视觉知识和文本知识进行多模态知识检索，在开放领域的 VQA 中使用候选答案指导知识检索，根据检索到的知识，学习验证一组候选答案的有效性并决定每个候选答案的可信来源。不同于以往基于知识搜索的视觉问答方法，Qu 等[70]研究了适用于 OK-VQA 数据集的文章检索，可以应用于更广泛的非结构化知识资源。文章检索方法包括稀疏检索和密集检索。首先使用 BM25 进行稀疏检索，并用物体名称和图像描述扩展问题。然后构造了一个双编码密集检索器，查询编码器为 LXMERT[71]，这是一个多模态的预训练 Transformer，它被用来学习图像和问题的交互。现有的工作利用不同的知识库来获取外部知识。由于知识库不同，因此很难对模型的性能进行公平的比较。为了解决这个问题，Luo 等[72]收集了

一个可用于任何 VQA 系统的自然语言知识库并提出了一个可视化的检索器-阅读器管道结构。视觉检索器的目的是检索相关知识，而视觉阅读器则是根据已知知识预测答案。检索器和阅读器都是在弱监督的情况下训练的。

基于知识搜索的视觉问答引起了越来越多研究者的关注。然而，大部分问题需要知识库少量的知识。如何排除嘈杂的信息干扰以及准确地抽取相关知识是一个需要克服的挑战。

2.5 基于对比学习的方法

自监督学习是无监督学习范式的一种，它不需要人工标注的类别标签信息，而是利用数据本身提供的监督信息来学习样本数据的特征表达，并用于下游任务。对比学习是自监督学习中的一类重要的方法。在视觉语言表示学习中，通过对比学习实现图像-文本对齐，这种对齐策略能够获得成功是由于它能够最大化图像和匹配文本之间的互信息（Mutual Information，MI）。互信息是一种衡量变量之间相互依赖的方法，通过区分正样本对和负样本对来衡量图像和问题之间的关系。

多模态编码器学习图像文本的交互具有挑战性。为了应对这个问题，Li 等[73]提出了 ALBEF 模型，引入图像文本对比学习，利用图像编码器、文本编码器和多模态编码器进行预训练，预训练的目标是使图像文本的互信息最大化、图像文本进行细粒度地交互，以及图像文本配对。Wang 等[74]提出了一个统一的视觉语言预训练模型 VLMo，它联合学习一个双编码器和一个共享 MoME Transformer 网络的融合编码器。MoME 引入一个模态专家池来编码模态特定信息，并使用共享的自注意力模块来对齐不同的模态。通过 MoME 进行统一的预训练，模型参数在图像-文本对比学习、屏蔽语言模型和图像-文本匹配任务中共享。大多数模型的编码器主要从某些不相关/有噪声的图像块或文本分词中提取信息。为此，Yang 等[75]提出了一种新的视觉语言预训练框架 TCL。与以往通过交叉模态对比损失简单地对齐图像和文本表示的研究不同，TCL 进一步考虑了模态内监督，这反过来有利于交叉模态对齐和联合多模态嵌入学习。为了将局部信息和结构信息结合到表示学习中，TCL 进一步引入了局部互信息，最大限度地利用全局表示与图像块或文本词语的局部信息之间的互信息。

对比学习能够使匹配的图像-文本对尽可能接近，同时使未匹配的图像-文本对相互远离。对比学习的目的是让融合编码器更容易学习多模态交互。但是，视觉问答中的对比学习还存在一定的局限性，它增强了图像和文本的全局互信息，忽略了输入中的局部信息和结构信息。此外，某些噪声可能会主导 MI，导致模型倾向于学习不相关的特征。

2.6 基于三维点云的方法

视觉问答近年来取得了巨大的进步。然而，目前的研究主要集中在二维图像问答任务上。研究人员尝试将 VQA 扩展到 3D 领域，这可以促进人工智能对 3D 真实场景的感知，从而模拟现实世界的场景，并有利于广泛的应用。与基于图像的二维视觉问答不同，3D 问答以点云为输入，在回答与 3D 场景相关的问题时需要语言处理和 3D 场景理解，模型的大致框架如图 8 所示。

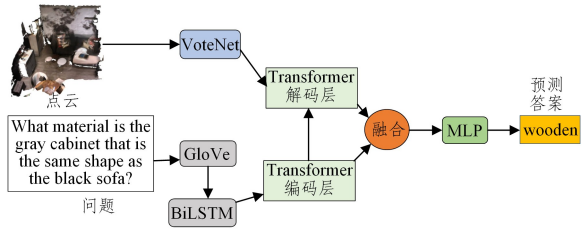

图 8 基于三维点云的方法流程图

Fig. 8 Flowchart of 3D-point-clouds based methods

现有的基于 2D 图像的模型在准确理解 3D 世界方面存在一些挑战。例如，二维图像缺乏对三维场景中相对方向和距离的准确感知，一些物体在重叠时被其他物体隐藏。为了解决这些问题，Azuma 等[76]提出了一个三维问答的基线模型，称为 ScanQA。ScanQA 模型包括 3D 和语言编码器、3D 和语言融合模块、物体定位和 QA 层。3D 和语言编码器层将问题转换为特征向量表示，并将点云转换为物体候选框。3D 和语言融合层使用基于 Transformer 的编码器层和解码器层将语言信息引导的多个 3D 物体特征以及文本信息融合在一起。物体定位和 QA 层评估目标对象框和对象标签，并预测与问题和场景内容相关的答案。与其他的 3D 场景理解任务相比，3D 问答对三维几何理解的要求明显更高。它不仅需要理解物体的外观和几何结构，甚至需要理解不同物体之间的空间关系。Ye 等[77]提出了一种新的基于 Transformer 的 3D 问答框架“3DQA-TR”，它利用一个语言分词器进行问题嵌入，利用两个编码器分别提取外观和几何信息，然后使用 3D-L BERT 将外观、几何和语言问题的多模态信息相互关联，以预测目标答案。传统的三维场景理解工作更多地关注单个物体，而忽略了物体之间的关系。Yan 等[78]引入了 3D 真实场景中的视觉问答任务，它旨在回答给定的 3D 场景中所有可能的问题。他们设计了 TransVQA3D。TransVQA3D 首先使用一个跨模态 Transformer 来融合问题和物体的特征。然后，通过应用场景图初始化，取场景图的附加边来进行场景图感知注意，从而获得物体之间的关系并推断出答案。

三维场景理解是一个比较新兴的研究领域。与基于二维图像的推理相比，在真实的三维场景中进行推理可以避免二维数据中的空间模糊性，从而获取真实的几何信息和物体间关系。同时，3D 场景通常包含更多的物体，涉及更复杂的对象间关系。尽管研究者们在探索空间表征以增强场景理解方面做出了大量努力，但目前的研究在三维感知（如计数、验证和存在）和获取对象属性（如大小、材质和结构）方面仍存在不足，具有进一步提升的空间。

3 数据集和评价指标

3.1 数据集介绍

为了追求高性能，大多数视觉问答算法需要在大规模的数据集上进行训练，丰富多样的数据集是视觉问答任务学习的基础。视觉问答数据集的主要形式为三元组，包含图像、问题和对应的答案。表 1 列出了主要的数据集，下文对常见的视觉问答数据集展开详细的介绍。

表 1　经典的视觉问答数据集

Table 1　Typical datasets for visual question answering

数据集	图像的数量	问题的数量	平均每张图像对应的问题数目	平均问题长度	平均答案长度	是否人工注释问答对
DAQUAR	1449	12468	8.60	11.5	1.2	Yes
Visual7W	47300	327939	6.93	6.9	2.0	Yes
Visual genome	108000	1445322	13.4	5.7	1.8	Yes
VQAv1	204000	614000	—	—	—	Yes
VQAv2	204000	1100000	—	—	—	Yes
VQA-CP v2	219000	603000	—	—	—	Yes
FVQA	1906	4608	2.5	9.7	1.2	Yes
KB-VQA	700	2402	3.4	6.8	2	Yes
CLEVR	100000	999000	—	—	—	No
OK-VQA	14031	14055	—	8.1	1.3	Yes
TextVQA	28408	45336	1.5	7.18	1.58	Yes
GQA	113000	22000000	—	—	—	No

(1)DAQUAR 数据集

DAQUAR[79]是第一个针对视觉问答任务发布的数据集。DAQUAR 数据集的图像内容主要是户内场景，来源于 NYU Depth Dataset V2 数据集。基于图像的问题与答案组合由两种方式生成：1)定义若干问题模板，根据图像标签自动生成问答对；2)使用人工标注，由志愿者回答自动生成的问题[79]。DAQUAR 数据集的缺点是数据量较少，问题语句不够清晰明确，难以回答。

(2)VQA 数据集

VQA 数据集既包含真实的图像又包含抽象场景图像[80]。真实图像来源于 MSCOCO 数据集；抽象场景图像是用剪切画来描述户内和户外的场景。数据集中的问题和答案均由人工生成，每张图片对应多个问答对。VQA v1 存在语言偏见，答案分布不均匀，比如对于 yes/no 问题，如果永远回答 yes，计算机就能答对大部分问题。为了弥补 VQA v1 的不足，学者在 VQA v1 的基础上，采集了新的数据，发布了 VQA v2 的版本，该版本比之前的版本大了一倍。基本上对于每个问题，都能给出两张相似的图像，但答案不同。与 VQA v1 相比，VQA v2 数据集更加平衡。该数据集能够全面地评估模型的性能，经常被用来作为 VQA 挑战赛的官方数据集。

(3)VQA-CP 数据集

现有的大多数 VQA 模型过度利用问题和答案之间的表面相关性来生成答案，导致在现实世界的场景中表现不佳。针对视觉问答中的语言偏见问题，Agrawal 等[81]重组数据集 VQA v1 和 VQA v2，分别得到了 VQA-CP v1 和 VQA-CP v2 数据集。重新划分数据集的目的是迫使训练集和测试集中每个问题的答案分布不一致。许多依赖语言偏见的模型在 VQA-CP 数据集上的性能显著下降，充分证明了 VQA-CP 数据集的难度。

(4)TextVQA 数据集

视力受损的用户对阅读周围图像中的文字有着较大的需求。为了满足此类需要，Singh 等[82]引入了 TextVQA 数据集。TextVQA 数据集包含了来自 Open images 数据集[83]中的 28408 张图片以及人们提出的 45336 个问题，这些图片往往包含文本，如“广告牌”“交通标志”等。问题经常询问“时间”“姓名”“品牌”或“作者”。回答问题需要阅读图像中的文本并结合视觉内容进行推理。与 VQA v2 相比，在 TextVQA 数据集上，人和机器的表现差距非常大，这表明 TextVQA 数据集是对 VQA v2 的进一步补充。

(5)CLEVR 数据集

CLEVR 数据集[84]比较特殊，它被用来测试模型的视觉推理能力。它有详细的注释，用于描述每个问题需要的推理类型。CLEVR 图像是由随机采样场景图生成的，并使用 Blender 来渲染。每个场景包含 3～10 个几何体，这些几何体的形状、大小、材质、颜色和位置都是随机的。问题总共包含 90 个系列，问题类型主要包括计数、比较、逻辑推理等。CLEVR 数据集能够展现模型的推理能力以及目前的局限性。

(6)GQA 数据集

以前的数据集答案分布不均匀并且缺乏关于问题内容的注释。为了弥补以前 VQA 数据集的关键缺陷，Hudson 等[85]引入了 GQA 数据集。GQA 数据集能衡量模型的一系列推理性能，如对象和属性识别、空间推理、逻辑推理和比较等。图像、问题和相应的答案都与匹配的语义表示相结合：每幅图像都有一个密集的场景图来标注，表示它所包含的对象、属性和关系[85]。每个问题都与一个功能程序相关联，该程序列出了为了得到答案需要执行的一系列推理步骤。每个答案都被添加了文本和视觉上的解释，指向图像中的相关区域。目前来看，GQA 数据集能够满足真实视觉场景理解的需要。

(7)包含外部知识的数据集

上述数据集几乎没有涉及外部知识库，为了解决基于知识的视觉问答任务，研究人员提出了 KB-VQA 数据集[60]、FVQA 数据集[86]和 OK-VQA 数据集[68]。KB-VQA 数据集、FVQA 数据集提供大量的支持事实(常识知识)，这些常识知识与视觉概念相联系。FVQA 数据集是从一个固定的知识库中选择一个知识三元组(如“狗是哺乳动物”)来注释问题，但问题是三元组不足以代表一般的知识。OK-VQA 数据集只包括需要外部知识来回答的问题，它由 14000 多个问题组成，涵盖了各种知识类别，如科学技术、历史和体育，测试模型从网站等知识库中检索相关事实的能力。

(8)3D 问答的数据集

3D 问答受到了越来越多的关注。为了支持 3D 问答任务，ScanQA 数据集[77]和 CLEVR3D 数据集[78]被相继提出。ScanQA 数据集是建立在 ScanNet 数据集的基础之上，它在

自由视角的3D扫描中提供了自然、自由形式和开放式的问题和答案，包含了806个场景的约6 000个问题，约30 000个答案。CLEVR3D数据集包含1 129个真实场景和60 000个问题。通过基于三维场景图的问题引擎来生成各种各样的关于对象属性和空间关系的推理问题。

3.2 评价标准

在视觉问答任务中要考虑答案和问题之间的相关性。根据答题的形式，视觉问答任务可以分为两种类型：多项选择形式的视觉问答任务和开放式的视觉问答任务。前者从多个候选答案中选择答案，后者生成答案，答案可能为单词、短语或句子，需要比较生成的字符串答案和真实的答案。多项选择问题准确度的度量标准由式(1)确定。

$$Accuracy=\frac{Correctanswers}{Totalnumberofquestions} \tag{1}$$

开放式问题的答案通常为一个或多个单词，评估标准比较复杂。若算法得出的答案与真实的答案不一致，直接判断该答案为错误答案显得过于苛刻，因为有的错误答案与真实答案接近，有的错误答案与问题没有关系。同一个问题可能有多个表达同样意思的答案，比如问题“这是哪个地方”，正确答案为“shop”，而回答“store”或“market”与正确答案的语义一致。此时，如果将上述情况得到的答案与得出的与题目完全不符的答案都给予同样的惩罚程度显然是不合理的。对于开放式的问题，为了合理地比较预测答案和真实答案，研究者提供了多种准确性评估方法。

WUPS[87]根据语义相似度度量预测值与真实值之间的相似度，并将权重从0分配到1。值越小，相似度越低。同时设置阈值，高于给定阈值的答案都将被认为是正确的。WUPS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。

在DAQUAR数据集[77]中，问题和答案的比例是1∶5。在VQA数据集[78]中，问题和答案的比例是1∶10。对于每一个预测答案而言，其准确率的计算方式为：

$$Accuracy=\min\left(\frac{n}{3},1\right) \tag{2}$$

其中，n为模型预测出的答案与标注者给出答案一致的数量。总之，如果至少3个标注者认为某个答案为正确答案，那么就认为这个答案是100%正确的。

BLEU(Bilingual Evaluation Understudy)和METEOR(Metric for Evaluation of Translation with Explicit Ordering)作为VQA的评价指标，在VizWiz数据集上进行测试。

$$BLEU^{g,h,i}=BP\exp\left(\sum_{n=1}^{N}W_n\lg P_n\right) \tag{3}$$

其中，BP表示短句惩罚因子；W_n表示当前单词元组所占的权重，所有的权重比相加应当为1；P_n表示整个语料库的准确度分数。

$$METEOR^j=(1-pen)F_{\text{mean}} \tag{4}$$

其中，j用于pen和F_{mean}的计算。

在视觉问答任务中，Hit@n是在可能性得分前n名的候选答案中匹配到正确答案的占比，Hit@n越大，效果越好。n通常取1，3，10。

3.3 模型性能展示

表2—表5列出了近年来最先进的模型在各个数据集上的性能，说明了各模型所使用的方法和准确率。这里对问题的类型进行细化，分别列出了模型在是否、计数、其他3类问题的准确率。表2中，VQA v2数据集包含验证测试集和标准测试集。在基于注意力机制的方法中，采用协同注意力方法的模型取得了较高的准确度，表明利用Transformer结构对视觉、语言信息进行交互能够明显提升模型的性能。结合预训练和对比学习的方法准确率普遍高于基于注意力机制的方法，这是因为在大规模的数据集上预训练之后模型学到了大量图像和问题之间的关联，具有较强的泛化性。基于对比学习的方法VLMo[74]获得了最好的性能，这表明对比学习更有利于图像文本特征的交互，是未来研究的热点。表3列出了模型在FVQA数据集上的表现。ZS-VQA[63]取得了最优的性能，表明利用GCN在知识图谱上进行推理，能显著提升模型的性能。表4中，BoUp[88]在VQA-CP v2数据集上的性能下降明显，这表明模型需要克服语言偏见问题，模型在VQA-CP v2数据集上的准确度有很大的进步空间。表5列出了模型在GQA数据集上的表现，GQA是面向场景图的数据集，适合模型利用图结构做出推理。基于场景推理的方法GraphVQA[56]获得了最优的性能，这表明结合图神经网络对场景图进行推理能够显著提升模型性能，是未来研究的趋势。

表2 最先进模型在数据集VQA v2上的比较

Table 2 State-of-the-art comparison on VQA v2 dataset

模型算法	是/否问题准确率 test-dev/%	计数问题准确率 test-dev/%	其他问题准确率 test-dev/%	总体准确率 test-dev/%	总体准确率 test-std/%	方法类别
BoUp[88]	81.82	44.21	57.26	65.32	65.67	注意力机制
BLOCK[89]	83.60	47.33	58.51	67.58	67.92	注意力机制
MuRel[90]	84.77	49.84	57.85	68.03	68.41	注意力机制
BAN[1]	85.42	54.04	60.52	70.04	70.35	注意力机制
MCAN[46]	86.82	53.26	60.72	70.63	70.90	注意力机制
MCAoAN[47]	86.96	53.45	61.01	70.84	71.16	注意力机制
DFAF[91]	86.09	53.32	60.49	70.22	70.34	注意力机制
TRAR[48]	88.11	55.33	63.31	72.62	72.93	注意力机制
$MCAN_{DST}$[49]	87.39	52.96	61.19	71.05	71.28	注意力机制
ALBEF[73]	—	—	—	74.54	74.70	对比学习
TCL[75]	—	—	—	74.90	74.92	对比学习
VLMo[74]	—	—	—	79.94	79.98	对比学习

表 3　最先进模型在数据集 FVQA 上的比较

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

模型方法	*Hit*@1	*Hit*@3	*Hit*@10	方法类别
BoUp[88]	34.81	50.13	64.37	注意力机制
SAN[45]	41.62	58.17	72.69	注意力机制
BAN[1]	44.02	58.92	71.34	注意力机制
top-3-QQmaping[86]	56.91	64.65	65.54	外部知识
ZS-VQA[63]	58.27	75.20	86.40	外部知识

表 4　最先进模型在数据集 VQA-CP v2 上的比较

Table 4　State-of-the-art comparison on VQA-CP v2 dataset

模型方法	是/否问题准确率/%	计数问题准确率/%	其他问题准确率/%	总体准确率/%	方法类别
BoUp[88]	41.96	12.36	46.26	39.84	注意力机制
CSS[92]	84.37	49.42	48.21	58.95	注意力机制
Rubi[93]	67.05	17.48	39.61	44.23	注意力机制
SSL[94]	86.53	29.87	50.03	57.59	注意力机制
CCB[95]	89.12	51.04	45.62	59.12	注意力机制

表 5　最先进模型在数据集 GQA 上的比较

Table 5　State-of-the-art comparison on GQA dataset

模型方法	总体准确率/%	方法类别
SceneGCN[55]	54.56	场景推理
GraphVQA[56]	94.78	场景推理
GMA[58]	57.26	场景推理
MGA-VQA[52]	65.93	场景推理

4　未来研究方向展望

视觉问答任务是多模态领域一个非常艰巨的任务，它是人类探索人机对话过程中的关键一步。虽然近年来视觉问答模型层出不穷，但是目前的视觉问答模型有很大的局限性，不能够与人类进行自然地沟通交流，仍然需要进行改进与研究。总的来说，视觉问答的研究不够成熟，还存在着以下问题和挑战。

(1)数据集不够丰富

目前，各种通用领域数据集或针对某一应用领域的数据集不断被构建出来，但是数据集仍然不够多样，应考虑更多的数据源，以提供更多的语料。VQA 的问题必须与图像内容相关是一种限制。在现实的应用场景中，对话内容往往超过了呈现的图像内容，例如在临床医学中预测疾病或者解释疾病的原因。总体而言，在种类丰富的 VQA 数据集上训练的 VQA 模型具有更好的泛化能力，更适合真实的生活场景。

(2)语言偏见问题

在视觉问答领域，如何解决语言偏见问题是近年来的一个热点研究问题。许多视觉问答模型倾向于选择训练集中经常出现的答案作为测试集中的正确答案。模型在回答问题时依赖于问题与答案之间的表面相关性，而忽略了图像信息。比如“图中的狗是什么颜色的”，虽然图中的狗是黑色，但模型仍倾向于依赖训练集中的高频答案而预测狗为“白色”。此外，在一些情况下模型可能关注突出的视觉信息，而忽视文本内容。语言偏差给 VQA 的应用带来了很大的困难。语言偏差可以分解为由模型产生的语言捷径偏差或由数据分布引起的语言数据偏差。由于语言偏差的存在，模型在应用中鲁棒性较差，可解释性很低。例如，在社交平台等公共场所部署相关 VQA 模型，可能会因为偏见而产生一些误导，从而影响对技术的合理使用。

(3)模型的推理能力不足

研究人员希望视觉问答能够模仿人脑的推理过程，在回答问题时，机器会寻找用于推理的证据，然后根据证据得到推理过程。但是，视觉问答模型在推理方面仍然不尽人意。虽然注意力机制有将图像关键区域和文本关键词进行对齐的效果，但是这只是浅层次的交互，模型并不理解文本语义与图像内容之间的联系，在推理方面缺乏可解释性。基于场景图的视觉问答尝试在答题过程中进行推理。然而，它只是依赖问题进行链式推理，推理过程相对简单，遇到逻辑复杂的问题时会束手无策。

综合视觉问答研究现状以及存在的问题，未来的研究方向可以围绕以下几个方面展开。

(1)构建丰富平衡的数据集

目前视觉问答的数据集类别不够丰富，可以从更多的数据源采集数据，如维基百科、教科书等。另外，数据集的形式过于单一，通常一张图片对应一个问题和一个答案。为了更加接近现实中的应用场景，可以建立多张图像对应一个问题或者一张图像对应多个问题的数据集。单一图像的信息有限，若算法结合多张图像的内容，回答问题会更加准确。

(2)克服语言偏见问题

由于语言偏见的存在，视觉问答的应用存在很大的局限性。语言偏见的来源之一是问题和答案的不均匀分布，因此，应该构建更加均衡的数据集，迫使模型在回答问题的同时兼顾图像和文本信息。目前 VQA 的任务一般是从候选答案中找到正确答案。使用自适应的框架动态结合外部知识来回答问题可能是减少语言偏差的有效手段。在数据集分布均衡的情况下，因果关系方法可以帮助模型更好地分析和利用图像与问题的核心信息，并将两者关联起来。如何让机器学习数据之间深刻的因果关系，学习人类的思维方式，是解决语言偏见的有效途径。

(3)提高模型的推理能力和可解释性

目前视觉问答算法存在推理能力不足的缺陷。将场景图和知识图相结合，开发能够对图进行多跳和更复杂推理的模型是未来的研究趋势。随着 3D 问答被提出，模型从一个丰富的 RGB-D 室内扫描的整个 3D 场景中接收视觉信息，并回答关于 3D 场景的文本问题，能够解决物体的空间理解问题以及根据文本进行物体定位的问题。3D 问答任务的天然优势有利于推理的进一步发展，需要进一步研究。基于注意力机制的方法缺乏可解释性，比如 Transformer 模型仅仅关注图像的关键区域和问题的关键词。研究模型在回答问题时提供支持证据以及模型给出的解释是否合理是未来热门的方向。

结束语　视觉问答任务属于高层次的计算机视觉任务，在人工智能领域具有较高的热度。除了视觉问答中的经典方法，本文重点介绍了基于场景推理、基于对比学习和基于三维点云这 3 种新方法，并对视觉问答的研究现状进行了分析，从而总结提炼出视觉问答遇到的主要挑战是数据集不够丰富、存在语言偏见、模型推理能力不足。未来的研究可以从构建丰富平衡的数据集、克服语言偏见问题、提高模型的推理能力和可解释性入手，进一步提升模型的综合性能。随着计算机视觉领域相关技术的不断

发展，相信视觉问答任务一定会走向成熟。

参 考 文 献

[1] KIM J H, JUN J, ZHANG B T. Bilinear attention networks [C]// Advances in Neural Information Processing Systems. 2018: 1571-1581.

[2] DONAHUE J, ANNE HENDRICKS L, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.

[3] YUAN D. Language bias in Visual Question Answering: A Survey and Taxonomy[J]. arXiv: 2111.08531, 2021.

[4] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.

[5] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[6] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704.04861, 2017.

[7] SHI H, LI H, WU Q, et al. Query reconstruction network for referring expression image segmentation[J]. IEEE Transactions on Multimedia, 2020, 23: 995-1007.

[8] YANG L, LI H, WU Q, et al. Mono is enough: Instance segmentation from single annotated sample[C]// 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2020: 120-123.

[9] MENG F, GUO L, WU Q, et al. A new deep segmentation quality assessment network for refining bounding box based segmentation[J]. IEEE Access, 2019, 7: 59514-59523.

[10] XU X, MENG F, LI H, et al. Bounding box based annotation generation for semantic segmentation by boundary detection [C]// 2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS). IEEE, 2019: 1-2.

[11] SHI H, LI H, MENG F, et al. Key-word-aware network for referring expression image segmentation[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 38-54.

[12] KINGMA D P, WELLING M. Auto-encoding variational bayes [J]. arXiv: 1312.6114, 2013.

[13] VAN DEN OORD A, KALCHBRENNER N, ESPEHOLT L, et al. Conditional image generation with pixelcnn decoders[C]// Advances in Neural Information Processing Systems. 2016: 4790-4798.

[14] KINGMA D P, DHARIWAL P. Glow: Generative flow with invertible 1x1 convolutions[C]// Advances inNeural Information Processing Systems. 2018: 10236-10245.

[15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014: 2672-2680.

[16] CHEN X, LI H, WU Q, et al. Bal-r^2cnn: High quality recurrent object detection with balance optimization[J]. IEEE Transactions on Multimedia, 2021, 24: 1558-1569.

[17] CHEN X, LI H, WU Q, et al. High-quality R-CNN object detection using multi-path detection calibration network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(2): 715-727.

[18] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[19] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[20] REDMON J, FARHADI A. Yolov3: An incremental improvement[J]. arXiv: 1804.02767, 2018.

[21] GE Z, LIU S, WANG F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv: 2107.08430, 2021.

[22] XIA R, DING Z. Emotion-cause pair extraction: A new task to emotion analysis in texts[J]. arXiv: 1906.01267, 2019.

[23] CALEFATO F, LANUBILE F, NOVIELLI N. EmoTxt: a toolkit for emotion recognition from text[C]// 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2017: 79-80.

[24] GHOSAL D, MAJUMDER N, PORIA S, et al. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation[J]. arXiv: 1908.11540, 2019.

[25] LIU B, LANE I. Attention-based recurrent neural network models for joint intent detection and slot filling[J]. arXiv: 1609.01454, 2016.

[26] NIU P, CHEN Z, SONG M. A novel bi-directional interrelated model for joint intent detection and slot filling[J]. arXiv: 1907.00390, 2019.

[27] ZHANG H, LI X, XU H, et al. TEXTOIR: An Integrated and Visualized Platform for Text Open Intent Recognition[J]. arXiv: 2110.15063, 2021.

[28] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv: 1406.1078, 2014.

[29] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. arXiv: 1609.08144, 2016.

[30] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[J]. arXiv: 1508.04025, 2015.

[31] NIU Y L, ZHANG H W. A survey of visual question answering and dialogue [J]. Computer Science, 2021, 48(3): 87-96.

[32] YU J, WANG L, YU Z. Research on visual question answering techniques[J]. Journal of Computer Research and Development, 2018, 55(9): 1946-1958.

[33] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[34] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556,

2014.
[35] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. 2015：1-9.
[36] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.
[37] CHO K，VAN MERRIËNBOER B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv：1406. 1078，2014.
[38] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810. 04805，2018.
[39] MALINOWSKI M，ROHRBACH M，FRITZ M. Ask your neurons：A neural-based approach to answering questions about images[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015：1-9.
[40] GAO H，MAO J，ZHOU J，et al. Are you talking to a machine? dataset and methods for multilingual image question[C]// Advances in Neural Information Processing Systems. 2015：2296-2304.
[41] KIM J H，LEE S W，KWAK D，et al. Multimodal residual learning for visual. qa[C]// Advances in Neural Information Processing Systems. 2016：361-369.
[42] MA L，LU Z，LI H. Learning to answer questions from image using convolutional neural network[C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016：3567-3573.
[43] SHIH K J，SINGH S，HOIEM D. Where to look：Focus regions for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：4613-4621.
[44] KAZEMI V，ELQURSH A. Show，ask，attend，and answer：A strong baseline for visual question answering[J]. arXiv：1704. 03162，2017.
[45] YANG Z，HE X，GAO J，et al. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：21-29.
[46] YU Z，YU J，CUI Y，et al. Deep modular co-attention networks for visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：6281-6290.
[47] RAHMAN T，CHOU S H，SIGAL L，et al. An Improved Attention for Visual Question Answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：1653-1662.
[48] ZHOU Y，REN T，ZHU C，et al. TRAR：Routing the Attention Spans in Transformer for Visual Question Answering[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：2074-2084.
[49] YU Z，JIN Z，YU J，et al. Towards Efficient and Elastic Visual Question Answering with Doubly Slimmable Transformer[J]. arXiv：2203. 12814，2022.
[50] WANG Z，WANG W，ZHU H，et al. Distilled Dual-Encoder Model for Vision-Language Understanding[J]. arXiv：2112. 08723，2021.
[51] ZENG Y，ZHANG X，LI H. Multi-Grained Vision Language Pre-Training：Aligning Texts with Visual Concepts[J]. arXiv：2111. 08276，2021.
[52] XIONG P，SHEN Y，JIN H. MGA-VQA：Multi-Granularity Alignment for Visual Question Answering[J]. arXiv：2201. 10656，2022.
[53] LI Y，FAN J，PAN Y，et al. Uni-EDEN：Universal Encoder-Decoder Network by Multi-Granular Vision-Language Pre-training[J]. arXiv：2201. 04026，2022.
[54] CHANG X，REN P，XU P，et al. Scene graphs：A survey of generations and applications[J]. arXiv：2104. 01111，2021.
[55] YANG Z，QIN Z，YU J，et al. Scene graph reasoning with prior visual relationship for visual question answe- ring[J]. arXiv：1812. 09681，2018.
[56] LIANG W，JIANG Y，LIU Z. GraphVQA：Language-guided graph neural networks for graph-based visual question answering[J]. arXiv：2104. 10283，2021.
[57] KONER R，LI H，HILDEBRANDT M，et al. Graphhopper：Multi-hop Scene Graph Reasoning for Visual Question Answering[C]// International Semantic Web Conference. Cham：Springer，2021：111-127.
[58] CAO J，QIN X，ZHAO S，et al. Bilateral Cross-Modality Graph Matching Attention for Feature Fusion in Visual Question Answering[J]. arXiv：2112. 07270，2021.
[59] ZHU Z，YU J，WANG Y，et al. Mucko：multi-layer cross-modal knowledge reasoning for fact-based visual question answering[J]. arXiv：2006. 09073，2020.
[60] WANG P，WU Q，SHEN C，et al. Explicit knowledge-based reasoning for visual question answering[J]. arXiv：1511. 02570，2015.
[61] WU Q，WANG P，SHEN C，et al. Ask me anything：Free-form visual question answering based on knowledge from external sources[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：4622-4630.
[62] RAMNATH K，HASEGAWA-JOHNSON M. Seeing is knowing! fact-based visual question answering using knowledge graph embeddings[J]. arXiv：2012. 15484，2020.
[63] CHEN Z，CHEN J，GENG Y，et al. Zero-shot visual question answering using knowledge graph[C]// International Semantic Web Conference. Cham：Springer，2021：146-162.
[64] MISHRA A，ALAHARI K，JAWAHAR C V. Top-down and bottom-up cues for scene text recognition[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2012：2687-2694.
[65] NEUMANN L，MATAS J. Real-time scene text localization and recognition[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2012：3538-3545.
[66] WANG K，BABENKO B，BELONGIE S. End-to-end scene text recognition[C]// 2011 International Conferenceon Computer Vision. IEEE，2011：1457-1464.
[67] SINGH A K，MISHRA A，SHEKHAR S，et al. From strings to things：Knowledge-enabled vqa model that can read and reason[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：4602-4612.
[68] MARINO K，RASTEGARI M，FARHADI A，et al. Ok-vqa：A visual question answering benchmark requiring external know-

ledge[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:3195-3204.
[69] WU J,LU J,SABHARWAL A,et al. Multi-modal answer validation for knowledge-based vqa[J]. arXiv:2103.12248,2021.
[70] QU C,ZAMANI H,YANG L,et al. Passage Retrieval for Outside-Knowledge Visual Question Answering[C]// Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021:1753-1757.
[71] TAN H,BANSAL M. Lxmert:Learning cross-modality encoder representations from transformers[J]. arXiv:1908.07490,2019.
[72] LUO M,ZENG Y,BANERJEE P,et al. Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering[J]. arXiv:2109.04014,2021.
[73] LI J,SELVARAJU R,GOTMARE A,et al. Align before fuse:Vision and language representation learning with momentum distillation[C]// Advances in Neural Information Processing Systems. 2021:9694-9705.
[74] WANG W,BAO H,DONG L,et al. VLMo:Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts[J]. arXiv:2111.02358,2021.
[75] YANG J,DUAN J,TRAN S,et al. Vision-Language Pre-Training with Triple Contrastive Learning[J]. arXiv:2202.10401,2022.
[76] AZUMA D,MIYANISHI T,KURITA S,et al. ScanQA:3D Question Answering for Spatial Scene Understanding[J]. arXiv:2112.10482,2021.
[77] YE S,CHEN D,HAN S,et al. 3D Question Answering[J]. arXiv:2112.08359,2021.
[78] YAN X,YUAN Z,DU Y,et al. CLEVR3D:Compositional Language and Elementary Visual Reasoning for Question Answering in 3D Real-World Scenes[J]. arXiv:2112.11691,2021.
[79] MALINOWSKI M,FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input [C]// Advances in Neural Information Processing Systems. 2014:1682-1690.
[80] ANTOL S,AGRAWAL A,LU J,et al. Vqa:Visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:2425-2433.
[81] AGRAWAL A,BATRA D,PARIKH D,et al. Don't just assume; look and answer:Overcoming priors for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4971-4980.
[82] SINGH A,NATARAJAN V,SHAH M,et al. Towards vqa models that can read[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:8317-8326.
[83] KRASIN I,DUERIG T,ALLDRIN N,et al. Openimages:A public dataset for large-scale multi-label and multi-class image classification[EB/OL]. https://github.com/openimages.
[84] JOHNSON J,HARIHARAN B,VAN DER MAATEN L,et al. Clevr:A diagnostic dataset for compositional language and elementary visual reasoning[C]// Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. 2017:2901-2910.
[85] HUDSON D A,MANNING C D. Gqa:A new dataset for real-world visual reasoning and compositional question answering [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:6700-6709.
[86] WANG P,WU Q,SHEN C,et al. FVQA:Fact-based visual question answe-ring [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence,2018,40(10):2413-2427.
[87] MANMADHAN S,KOVOOR B C. Visual question answering:a state-of-the-art review[J]. Artificial Intelligence Review,2020,53(8):5705-5745.
[88] ANDERSON P,HE X,BUEHLER C,et al. Bottom-up and top-down attention for image captioning and visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:6077-6086.
[89] BEN-YOUNES H,CADENE R,THOME N,et al. Block:Bilinear superdiagonal fusion for visual question answering and visual relationship detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019,33(1):8102-8109.
[90] CADENE R,BEN-YOUNES H,CORD M,et al. Murel:Multimodal relational reasoning for visual question answering [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:1989-1998.
[91] GAO P,JIANG Z,YOU H,et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering [C]// Proceedings of the IEEE/CVF Conferenceon Computer Vision and Pattern Recognition. 2019:6639-6648.
[92] CHEN L,YAN X,XIAO J,et al. Counterfactual samples synthesizing for robust visual question answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:10800-10809.
[93] CADENE R,DANCETTE C,CORD M,et al. Rubi:Reducing unimodal biases for visual question answering[C]// Advances in Neural Information Processing Systems. 2019:839-850.
[94] ZHU X,MAO Z,LIU C,et al. Overcoming language priors with self-supervised learning for visual question answering[J]. arXiv:2012.11528,2020.
[95] YANG C,FENG S,LI D,et al. Learning content and context with language bias for visual question answering[C]// 2021 IEEE International Conference on Multimedia and Expo(ICME). IEEE,2021:1-6.

LI Xiang, born in 1997, postgraduate. His main research interests include visual question answering and so on.

LI Xuexiang, born in 1965, professor, master supervisor. His main research interests include high performance computing and cloud computing.

(责任编辑:杨雪敏)