ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 1,Jan. 2023,pp. 1－20
DOI:10. 16337/j. 1004-9037. 2023. 01. 001

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

跨模态视觉问答与推理研究进展

张飞飞[1]，张建庆[1]，屈思佳[1]，周琬婷[2]

(1. 天津理工大学计算机科学与工程学院，天津 300384;2. 北京邮电大学人工智能学院，北京 100876)

摘　要：随着社交媒体和人机交互技术的快速发展,视频、图像以及文本等多模态数据在互联网中呈爆炸式增长,因此多模态智能研究受到关注。其中,视觉问答与推理任务是跨模态智能研究的一个重要组成部分,也是人类实现人工智能的重要基础,已成功应用于人机交互、智能医疗以及无人驾驶等领域。本文对视觉问答与推理的相关算法进行了全面概括和归类分析。首先,介绍了视觉问答与推理的定义,并简述了当前该任务面临的挑战;其次,从基于注意力机制、基于图网络、基于预训练、基于外部知识库和基于可解释推理机制5个方面对现有方法进行总结和归纳;然后,全面介绍了视觉问答与推理常用公开数据集,并对相关数据集上的已有算法进行详细分析;最后,对视觉问答与推理任务的未来方向进行了展望。

关键词：视觉问答;视觉常识推理;可解释推理;语义对齐

中图分类号：TP391.4　　**文献标志码**：A

Recent Advances in Visual Question Answering and Reasoning

ZHANG Feifei[1], ZHANG Jianqing[1], QU Sijia[1], ZHOU Wanting[2]

(1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China;2. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: With the rapid development of the social media and human-computer interaction, the volume of multimedia data, such as video, image and text, has grown tremendously. Therefore, researchers have focused their attention on the multi-modal intelligence research. As an essential and fundamental research topic in the multi-modal intelligence and artificial intelligence area, some scientific research results on the visual question answering and reasoning task have been successfully implemented in the fields of human-computer interaction, intelligent medical care, and unmanned driving. This paper makes a comprehensive overview of the related algorithms of visual question answering and reasoning, meanwhile classifies and analyzes the existing methods. Firstly, we introduce the definition of the visual question answering and reasoning task, and briefly describe the main challenges of this task. Then, we summarize the existing methods that focus on attention mechanism, graph network, model pretraining, external knowledge and explainable reasoning mechanism. After that, we comprehensively introduce the common visual question answering and reasoning benchmarks and discuss the existing methods on these benchmarks in detail. Finally, we prospect future directions of the visual question answering and reasoning task.

基金项目:国家重点研发计划(2018AAA0102200);国家自然科学基金(62036012, 62002355, 61832002, 62072455, 62102415, 62106262, 62006227);北京自然科学基金(L201001)。
收稿日期:2022-10-28;**修订日期**:2022-12-09

Key words: visual question answering; visual commonsense reasoning; explainable reasoning; semantic alignment

引　言

随着计算机视觉和自然语言处理的蓬勃发展,图像处理、目标检测以及行为识别等计算机视觉任务以及机器翻译和常识问答等自然语言处理任务都取得了巨大的成功。但是这些任务都只涉及对单一模态数据的分析,例如:对于目标检测和图像分割等纯视觉任务只需要对图像中的物体进行识别即可;对于机器翻译和词性解析等纯语言任务只需要理解文本中每个单词的含义,不需要对其中更深层次的上下文相关信息进行挖掘和推理。随着现代科技的发展以及多媒体传感器的相继出现和大规模运用,文本、图像以及视频等多模态数据在网络世界中大量涌现。然而,传统的单模态方法不能对多模态数据进行有效的组织和管理,因此跨模态任务应运而生。其中,视觉问答(Visual question answering, VQA)[1-3]和视觉常识推理(Visual commonsense reasoning, VCR)[4-7]等作为多模态领域的重要组成部分,受到了研究人员的广泛关注。

2014年,VQA被首次提出,给定一幅图片和一个自然语言问题,要求模型选出一个符合上下文语境的自然语言答案,如图1所示。VQA是一项值得研究的任务,可以将之与图像字幕、视觉问题生成和视觉对话等多模态任务相结合,从而创建一个智能代理实现与人类进行交流。它还可以应用到许多具体的领域,比如帮助情报分析人员和视障人士从网络或生活中获取图像信息等。

当人们看到一幅图像时,可以很快推断出图像中人物的心理状态和行为目的,但这对于VQA系统而言却是一个极其困难的任务。为了让VQA系统拥有更高阶的认知能力和更强大的推理能力,2018年VCR任务作为VQA的子任务被提出,如图2所示。该任务设定为给定一幅图像和一个问题,VCR模型不仅需要选出正确的答案,更要给出一个合理的理由来解释所选答案的正确性。VCR任务提出了2个四项选择问题:问题回答(Q→A)和答案合理性(QA→R)。整体设置(Q→AR)要求模型首先在Q→A中选择正确的答案,然后在QA→R中选择正确的理由。因此只需要训练Q→A和QA→R的模型即可。可以通过从给定查询的4个选项中选择1个响应来进一步统一这2个子任务。对于Q→A,查询是一个问题,选项是候选答案。对于QA→R,查询是问题和正确答案的组合,而选项是候选理由。想要教会机器像人类一样能够正确地回答问题,并且实现进一步的推理,不仅需要识别层面的特征提取,即对图像和文本特征进行提取;更重要的是认知层面的语义挖掘,即学习到图像和文本的跨模态特征。

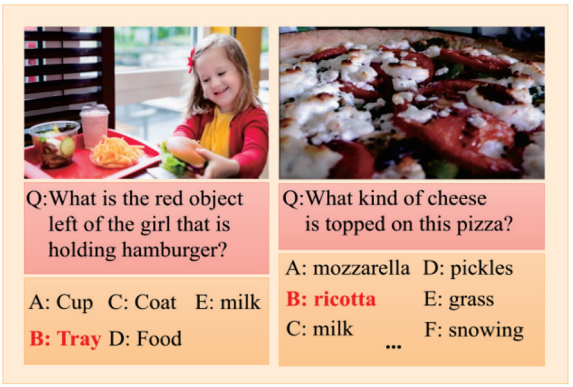

图1　视觉问答任务示例

Fig.1　Example of VQA task

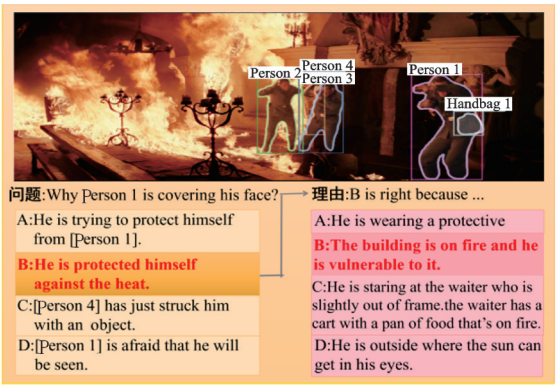

图2　视觉常识推理任务示例

Fig.2　Example of VCR task

这样才能更好地提升模型的通用性和计算效率。

VQA与VCR作为跨模态任务的重要组成部分,已有很多文献对其进行综述和研究。文献[8]先从注意力机制、模块化组合模型以及鲁棒性模型等多方面介绍了当前视觉问答任务的进展,然后指出了现有方法存在的挑战,最后对视觉问答技术的未来发展进行了展望。文献[9]详细介绍了VQA相关数据集,并将视觉问答方法细分为联合嵌入方法、注意力方法、基于神经模块网络的组合方法以及基于外部知识的方法,同时介绍了每种方法的动机、细节及其缺点。文献[10]先从视觉问答与对话的多模态感知困难、数据集偏差、视觉对象与文本中代词难以对齐这3个难点出发,通过基于注意力机制、基于外部知识、基于预训练的方法、基于数据增强的方法以及基于特征模型鲁棒性的方法对视觉问答与对话任务进行了介绍。文献[11]主要对知识型视觉问答任务进行了详细阐述。通过如何得到规模化的数据集和知识库、怎样实现视觉场景拓展延申以及如何去除数据集偏差引起的过拟合问题等挑战作为切入点,将知识细分为图像知识、内部推理知识、外部知识以及未知知识4类对现有模型进行了详细介绍。文献[12]根据如何有效连接视觉和文本模态的方法进行分类,特别研究了将卷积神经网络和递归神经网络相结合,以将图像和问题映射到公共特征空间的常见方法,讨论了如何将具有结构化知识库接口的记忆增强和模块化架构应用到视觉问答任务中。

尽管文献[9-12]从不同角度对视觉问答任务进行了较全面总结,但仍存在以下不足:(1)仅侧重于对视觉问答现有方法的总结和分类,没有考虑到视觉常识推理任务是特殊的视觉问答子任务,忽略了对视觉常识推理任务的相关介绍,从而导致内容不够全面;(2)大多方法的推理过程是基于隐式的黑盒推理,未考虑模型的可解释性推理机制,即没有详细分析如何对跨模态特征进行显式推理,进而为每一步推理预测提供可解释性依据;(3)在介绍各数据集方法对比部分,仅以表格的形式列出了每种方法的性能,均未对各个对比方法进行简要介绍,也未对其进行详细对比分析。因此,本文将视觉常识推理作为特殊的视觉问答任务进行考量,全面对视觉问答与推理的研究进展进行了系统性的总结与分析。

1 视觉问答与推理相关挑战

视觉问答与推理任务在拥有无限应用前景的同时,也面临着诸多挑战。现有的视觉问答与推理系统大多是基于多模态特征融合的框架,即先分别提取视觉特征和文本嵌入,再将两种特征映射到同一个特征子空间中,然后将融合后的特征应用于后续的答案预测。在此过程中,为了更好地提升模型的性能,研究人员需考虑缓解数据集偏差、缩小视觉语言之间的语义鸿沟以及实现显式推理等问题。

1.1 缓解数据集偏差

在计算机视觉和自然语言处理领域,数据集偏差普遍存在,作为跨模态领域的象征性问题,视觉问答与推理毫无疑问也面临着同样的挑战。视觉问答与推理任务的数据集偏差主要来自语言部分,文本和答案(和理由)之间有着强相关性,模型在回答问题时可能忽略图片中的重要信息,只关注文本信息之间的虚假相关性进行推理。例如,在视觉问答经典的VQA-v1[1]数据集中,所有的以一般疑问句为问题的实例中大约90%的答案都为“是”,以致于模型利用这种统计偏差而非利用跨模态内容本身的特征信息进行预测[13];由于数据集中这类偏差的存在,导致模型忽略视觉内容只根据偏差做出选择,从而选择一个错误的答案。在VCR数据集中也存在数据集偏差,由于该数据集中问题和答案中常出现重复单词,模型很容易利用问题和选项之间重复单词数过多就认定其为正确答案,而非经过正确的推理判断过程得到答案[14]。在模型训练过程中模型过分依赖问题和答案之间模态内的关系,使其忽略与图片内容的模态间关系。常用的解决方法包括生成反事实样本平衡偏差[14-15]以及额外添加监督信息[16]等。因此,如何有效解决这样的数据集偏差成为了视觉问答与推理任务的研究热点。

1.2 缩小语义鸿沟

视觉常识与推理任务需要同时处理来自视觉与语言两个模态的输入，这两个模态的底层特征异构，高层语义相关。如何从多模态输入中提取有效信息是视觉语言任务的共同挑战。由于表现形式的不同，文本通常展现的是一种离散的高层语义信息，而图像则由连续的底层像素特征来表示，这导致底层的视觉表示和高层的文本表示之间存在着巨大的语义鸿沟，例如有些模型[17]只探索视觉对象的内部关系，而忽略了视觉概念和文本内容之间的跨模态语义相关性，此时则需要把单一模态的两种特征映射到同一个特征子空间进行特征融合，以得到富含更多的语义信息的高级特征，使模型有更强的表达能力。缩小语义鸿沟常用的方法主要有向量操作[18]、构建异构图[6]和注意力机制[19]等。因此，怎样有效缩小语义鸿沟以得到更丰富的语义关联表示，还有待研究人员进一步探索。

1.3 实现显式推理

视觉问答与推理任务旨在引导研究领域解决认知层面的挑战，不仅希望模型实现正确预测，还希望模型提供令人信服的推理路径。然而现有的研究大多都是隐式推理过程，例如依靠强大的端到端网络[20-21]使模型性能得到大幅提升，这种方式无法产生可解释的显式推理路径。即使一些研究[22-24]利用图网络中结点和边的关系对跨模态内容进行建模，让模型学习结构化的跨模态特征，根据追踪图结点每一层的更新执行显式推理过程，但如何通过挖掘模态内和模态间的相关性来连接视觉和语言领域，进而得到分步骤推理过程，实现该任务的显示推理仍是视觉问答与推理的一项难题。

2 视觉问答与推理相关研究

如图3所示，给定一幅图片以及对应的问题，经过视觉编码器以及文本编码器分别对其进行特征提取。为了使模型能够得到令人满意的答案预测结果，现有的研究方法主要包括：(1)通过注意力机制使模型更关注重点区域或者重点单词，从而实现跨模态特征的细粒度对齐；(2)通过构建图结构不仅可以探索底层视觉特征和高层文本特征之间的语义相关性，还能让模型拥有显式的推理路径；(3)使用预训练模型学习到泛化性更好的多模态融合特征，从而缩小跨模态语义鸿沟；(4)通过引入外部知识使模型不仅能提取到跨模态内容本身具备的知识，还能挖掘更多外部世界的常识知识来辅助推理预测；(5)通过因果推理和构建反事实样本等可解释性机制使模型提供显式的推理路径，并消除数据集偏差。

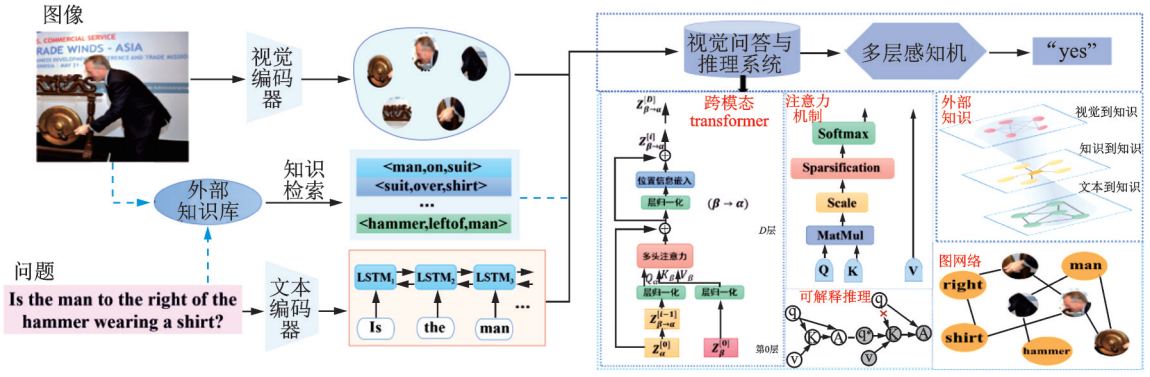

图3 视觉问答与推理任务常用方法

Fig.3 Common methods for VQA and VCR

2.1 基于注意力机制

在视觉问答与推理任务中,图片中总有一些与问题无关的区域,问题中也有一些单词是不具备实体含义的无关信息,因此需要智能系统更关注重点区域和重点单词,而不是给予每个区域或者每个单词同样的关注度。注意力机制[25]是一种在更局部的级别上表示与图像中每个区域相对应的特征向量,再根据其与问题特征的相似性进行排序的算法。仅仅提取图像的全局特征来回答只关注图像特定区域的问题可能会限制视觉问答与推理系统,因此注意力机制的思想恰好可以应用到视觉问答与推理任务中,以实现将文本模态和视觉模态的注意力引导到产生正确答案所必需图像的相关区域上,而非均匀关注每个文本单词和每个图像区域。

自提出注意力机制概念以来,每个研究领域都试图引入注意力机制来提升模型性能,视觉问答与推理任务也不例外。例如,Zhu等[3]提出将注意力机制和长短时记忆网络进行结合,将每一步的视觉特征和单词特征作为输入得到的注意力图和视觉特征进行矩阵相乘进而得到更新的视觉特征。Yang等[26]提出了堆叠注意力网络,主要利用问题的文本特征的引导得到视觉区域特征的注意力权重使模型关注到图片中与问题相关的局部重点区域。Lu等[19]提出分层共同注意力机制,将“视觉引导的问题注意力”和“文本引导的视觉注意力”进行结合,从单词级、短语级和句子级学习文本的细粒度特征实现更准确的答案预测。Wu等[27]提出关系注意力机制用于建模模态内的关系,并利用一个相互关系单元来学习视觉和文本之间的细粒度相关性,最后输出图像中各区域上下文感知的特征信息。Peng等[28]提出单词到区域注意网络,只需要使用问题的关键字来识别高注意力值的图像区域或特征,该注意力网络填补了问题关键字和图像区域之间的现有差距。Cai等[29]提出了一种基于多模态特征融合和多级注意机制的视觉问答方法,该方法主要利用多模态因子双线性池方法对图像特征和文本特征进行融合,并将自注意机制、引导注意机制和多头注意机制相结合形成一个多层注意网络,从而能够准确地从图像中获得所需的目标特征。Guo等[30]提出了一个再关注机制(Re-attention),该框架首先通过计算特征空间中每个单词-对象对的相似度来学习对象的初始注意权重。然后根据答案,通过重新关注图像中的对象来重构视觉注意图,从而达到利用答案引导模型推理的目的,较大地提升了该模型的性能。

同时,在VCR任务中注意力机制也常被用于关注重点区域或重点实体单词。例如,Yu等[31]提出了一个基于注意力网络的视觉推理模块和一个视觉关系推理模块,以捕获丰富的视觉语义并帮助增强视觉表示进而改进跨模态分析。其中,基于注意力网络的视觉推理模块是利用双线性视觉注意力模块来识别关键对象,视觉关系推理模块是用于推理由文本信息引导的对象之间的成对和组内视觉关系,这两个模块增强了关系级别和对象级别的视觉特征,从而大大提升了模型的视觉推理能力。Yang等[32]提出了一种新的注意机制:因果注意,以消除现有基于注意力机制的视觉语言模型中训练集和测试集分布不一致的问题。这种数据分布不一致会导致有害的数据偏差,误导注意力模块将注意力集中在训练数据中的虚假相关性上,从而破坏模型的泛化能力。由于模型无法观察到混淆因子(即导致某种结果发生的虚假原因),于是使用前门调节进行因果干预,即通过切断混淆因子与原因之间的联系,从而保证只有真正的原因才会导致结果的发生。Ding等[33]提出了一种更通用的基于神经网络的动态视觉推理方法,此方法结合了以对象为中心的表示、自注意力机制和自监督的动态化学习3个要素,使模型性能显著提升。

由于基于注意力机制的方法能够使模型更好地关注到与答案相关的视觉区域或者文本单词,并且能够将具有相同语义不同模态的特征信息进行有效增强与融合,从而学习到更具判别性的数据表征关系,以达到联合推理,提升模型准确率的目的,因此该类方法在视觉常识与推理任务中已被研究人员广泛应用。但由于基于注意力的方法需要较多的矩阵计算,因此在一定程度上增加了模型的计算复杂度,并且有时模型只聚焦在重点内容会忽略其他有用信息,因此想要对跨模态内容有更充分的理解还

须进一步研究。

2.2 基于图网络

图神经网络[34]是一种通过将对象视为图域上的结点来捕获对象之间依赖关系的神经网络。由于其高解释性和有效提升模型性能的特点,近年来被广泛应用于跨模态任务中。研究人员通常用结点对视觉实体和文本实体进行建模,用边来建模实体之间的关系。因此使用图结构来对视觉问答与推理任务建模成为了研究人员对关系建模的优先选择之一[34-37]。

由于图结构不仅能清晰地对视觉实体和文本实体进行建模,也可以利用图结构表示去执行推理过程,所以其在视觉问答中有诸多应用[38-46]。例如,Teney等[41]提出使用图卷积网络(Graph convolutional network,GCN)生成捕获语法关系的文本图像表示,这种表示使用预训练的词嵌入来形成结点特征,并将单词之间的句法依赖编码为边缘特征,从而实现对文本模态单词之间的关系建模。Li等[42]提出了一种关系感知图注意网络,研究了两种类型的视觉对象关系:视觉对象和语义交互的显式关系以及图像区域之间的动态隐式关系。此方法将每个图像编码为一个图,并通过图注意机制对图像中每个区域对之间的关系进行建模,来引导模型自适应地学习到更好的文本表示。Shi等[43]将复杂视觉推理任务中经常使用的黑盒神经架构分解为可解释的和显性的神经模块,使用场景图(对象作为结点,成对关系作为边)进行结构化知识的显式推理。Zhu等[44]提出用一个多模态异构图来描述图像,该图包含了对应于视觉、语义和事实特征的多层信息。还提出了一个模态感知的异构图卷积网络,从与给定问题最相关的不同层捕获证据,即模态内图卷积从每个模态中选择证据,跨模态图卷积在不同模态中聚合相关信息。通过多次叠加这个过程,使模型进行迭代推理,并通过分析所有面向问题的证据预测最佳答案,从而提升了模型性能。

由于图网络的可解释推理步骤的有效性,使其在VCR任务中也有广泛的应用[46-50]。例如,Yu等[6]提出了一种新的异构图学习框架,通过构建图像-问题的异构图、图像-答案的异构图和问题-答案的同构图,去指导模型做出正确的选择和推理,以达到图内和图间推理。Wu等[46]提出了一种多层次语义增强方向图网络。该网络设计了一个模态交互单元模块,主要通过聚合多层次视觉语言关系来实现高阶跨模态对齐;然后提出了一个方向线索感知图推理模块,该模块可以根据视觉实体和文本实体的重要性,在每个推理步骤中动态选择有价值的实体,从而得到更高级的跨模态融合特征,以缩小语义鸿沟。Zhang等[47]提出了一种新的显式跨模态表示学习网络,将句法信息融入到视觉推理和自然语言理解中。首先,该方法基于双分支神经模块网络,在语言表达的高级句法结构指导下进行显式跨模推理,然后将语言表达的语义结构整合到句法GCN中,以促进语言理解。

通过使用图网络建立结点之间的关系,在结点更新过程中使相邻结点具有了语义丰富的上下文相关性,为答案和理由提供了可见的细粒度证据,也为视觉问答与推理模型提供了可跟踪的推理流程,连接了视觉和语言领域,缩小了语义鸿沟。但是,图网络往往构建的是所有区域对、单词对以及区域-单词对之间的关系,这种全连接图如何自适应地选择一个恰当的阈值来作为筛选条件,选出与上下文语境最相关的图结构仍是一个有待解决的问题。

2.3 基于预训练

在多模态机器学习领域,为特定任务而制作的人工标注数据昂贵,且不同任务难以进行迁移,从而需要大量重新训练来重新适应不同的任务,导致训练多个任务时效率低下以及资源浪费。预训练模型通过以自监督为代表的方式进行大规模数据训练,对数据集中不同模态的信息进行提取和融合,以学习其中蕴涵的通用知识表征,从而服务于广泛的下游视觉语言多模态任务。由于使用预训练会显著提升模型的性能,所以这一方法逐渐成为人工智能各领域的主流方法[51-58]。

预训练模型在自然语言处理领域已经取得了巨大的成功,目前已有很多研究人员将预训练技术引入到视觉问答与推理任务中。例如,Lu等[50]提出了一个通用预训练框架。该框架先通过大型概念性字幕数据集上的重建任务和匹配任务对模型进行预训练,然后将双向Transformer的Encoder(BERT)架构扩展为多模式双流模型,再通过共同注意力Transformer层交互中的视觉和文本输入到预训练后的模型中得到泛化性更好的跨模态特征。Zhong等[51]提出了一种新的神经网络模型,称为自适应神经模块Transformer。该模型通过考虑中间问答结果自适应地调整问题特征编码和位置信息解码,通过一个新的Transformer模块对具有给定问题特征的中间结果进行编码,以生成经过推理步骤演化的动态问题特征嵌入。Li等[52]提出了一种统一的模型预训练体系结构。该结构通过利用大规模自由文本语料库和图像集合来提高文本和视觉理解能力,并利用跨模态对比学习将文本和视觉信息对齐到一个统一的语义空间中。该作者将该方法用于视觉问答任务中,取得了当时最先进的效果。Kamath等[21]提出了一种端到端的基于Transformer的检测框架,该框架是先在短语和图像中的对象之间具有明确的对齐关系的大型多模态数据集上进行预训练,然后在下游跨模态任务上进行微调。该框架主要用于检测图像中以原始文本查询为条件的对象,如标题或问题。通过在模型早期进行跨模态特征融合实现文本和图像联合推理,并在视觉问答任务上取得了有竞争力的效果。Ma等[53]使用视觉Transformer作为视觉推理的基本模型,并更好地利用定义为对象实体的概念及其关系来提高其推理能力。该方法引入了一个新的概念特征字典,允许在训练时使用概念关键字进行灵活的图像特征检索,从而提取到细粒度的视觉特征以辅助模型选出正确答案。Wang等[54]提出了一个支持任务全面性、对特定任务和特定模态都无关的框架。该框架基于预训练和微调且不需要为下游任务设置额外的任务特定层,并将其应用于视觉问答任务上显著提升了模型性能。

在视觉常识推理中,预训练的方式也深受研究人员青睐。例如,Su等[55]提出基于多层双向Transformer编码器的VL-BERT框架,该框架可对所有输入元素之间的依赖进行建模。VL-BERT的输入是图像中感兴趣区域上的特征和输入句子中的子单词。此感兴趣区域可以是目标探测器生成的边界框,也可以是特定任务中的注释框。为了更好地利用通用表示,该框架先在大规模概念型字幕数据集和纯文本语料库上预训练,实现了视觉语言线索对齐,使其更适应视觉常识推理和视觉问答等下游任务。Marasovic等[22]认为多模态语义准确合理化的关键挑战是各个层面的全面图像理解,而不仅仅是停留在像素级别的理解。于是提出了基于生成式的Rationalevt Transformer视觉问答模型,它通过将预训练语言模型和视觉常识图相结合来生成用自然语言描述的理由。Wang等[37]提出了一个场景图增强图像-文本学习框架。为了开发场景图结构,在模型结构层面提出了一种多跳图Transformer,用于规范化多跳之间的注意力交互,可显式地对视觉场景图进行多跳推理。Yang等[58]提出了一个通过利用跨模态和模态内自监督进行视觉语言预训练的三重对比学习框架。该框架通过模态间对比学习和模态内对比学习在表征学习中提供互补优势,不仅充分利用了来自图像和文本输入的局部化和结构化信息,并且进一步最大化了图像/文本局部区域与其全局特征之间的相关性。

预训练这种以自监督的方式先在大规模图文对中进行预先处理,再将其应用到下游跨模态任务中的方法,帮助下游任务学习到了更加细粒度的跨模态特征,更好地实现了视觉模态和文本模态之间的信息交互,在很大程度上打破了不同视觉语言任务之间的鸿沟,并提升了视觉问答与推理任务的性能表现。但是基于预训练的模型需要的训练数据量更大,对实验所需的硬件要求也更高。

2.4 基于外部知识

当人类在看到一幅图像时,看到的不仅仅是图像表面的物体,还能进一步推测出图像所表达的深层含义,比如图中某个人的心理状态和行为目的等。这主要是因为人是一个“智能系统”,不仅停留在识别层面,更能结合实际生活中的常识或者经验做出合理的推测。由此科研人员开始考虑一个人工智

能模型在做出选择时是否也可以引用外部世界的额外知识以达到辅助预测的目的,近几年基于外部知识的视觉问答与推理十分流行[59-68]。

在视觉问答与推理任务中,仅考虑图像或者文本实体的特征表示并不能实现很好的推理效果,通过加入外部知识作为辅助可以实现更好的推理。例如,Yu等[61]将基于知识的视觉问答任务重新表述为从多模态信息中获取补充证据的循环推理过程,从视觉、语义和事实的角度,通过多个知识图来描述图像,其中多个知识图由视觉图、语义图以及由图像引导的外部知识图构成。主要通过将视觉问答模型分解为一系列基于图结构的推理步骤,实现视觉和语义信息的并行推理来预测全局最优答案。Zhang等[62]提出了一个基于知识的增强网络,通过引入与对象相关的开放域知识来帮助回答问题。该方法还设计了一个注意模块,可以根据具体问题进行自我调整来衡量外部知识对检测到的对象的重要性。因此该网络从图像中提取了更多的视觉信息,并引入知识图为推理过程提供了必要的常识,从而实现了较好的视觉问答效果。Wu等[46]设计了一个视觉问答模型。该模型将图像区域表示与知识库中提取的信息相结合来执行视觉问答与推理任务。Marino等[63]提出利用两种类型的知识进行表示学习和推理。首先使用基于Transformer模型从无监督语言预训练和有监督训练中有效地学习隐含知识,然后从知识库中提取符号化的外部知识,使模型不仅不会因隐式嵌入而失去显式语义,还能将各种知识来源结合起来,从而得到了知识增强的跨模态融合特征。Dai等[64]提出通过视觉语言知识提取,用文本预训练语言模型来增强双流视觉语言模型,从而增强模型的泛化能力。

由于VCR任务在正确回答问题的基础之上还需给出恰当的理由,而且QA→R这一子任务往往是需要进一步推理才能选出正确的理由,所以在视觉常识推理中引入外部知识很有必要。Whitehead等[66]提出了一种新的学习组合技能和概念的方法(技能指的是视觉对象之间的关系或者文本中的动词,概念指的是视觉对象或者文本实体)。该方法通过将技能的编码与概念的编码分离开来,并将这两个因素隐含在模型中以学习到鲁棒的概念表示和技能表示。再通过一种新的对比学习方法来强化这两种特性,使模型能更精准地定位到视觉实体和文本中的动词及名词,从而提升视觉常识推理任务的准确率。Wen等[47]提出了基于常识的推理模型,主要通过细胞级、层次级和注意力级的多层次知识转移网络注入外部常识知识。该模型可以从不同的角度有效地捕获知识,提前感知人类的常识。为了进一步促进认知层面的图像理解,该方法可以将转移的知识与视觉内容相关联形成推理线索以得出最终答案。为了解决视觉推理方法与现实世界图像的语义复杂性之间的知识差距,Zhang等[67]提出了一种显式的视觉推理方法,该方法提出了一个知识整合网络,显式地为外部知识库中的名词实体和谓词进行连接形成图结构,以丰富显式推理中使用的场景图语义信息。然后创建一个新的图相关模块,对丰富的场景图利用GCN执行高阶关系关注,以提取细粒度的结点特征。Yu等[24]提出了一种知识增强的方法,该方法通过从场景图中获得的结构化知识来学习视觉语言的联合表示,试图在视觉和语言之间建立详细的语义连接(对象、对象属性和对象之间的关系)以缩小跨模态语义鸿沟。Zellers等[68]提出了一种完全无标签、自监督的学习多模态知识的方法,该方法先在YouTube视频和转录的语音库上混合利用帧级(空间)和视频级(时间)目标进行预训练,这样不仅可以学习将图像与局部单词进行匹配,还可以实现全局信息的上下文语境化,很大程度上缩小了视觉问答任务的语义鸿沟。

引入外部知识的方式不仅可以学习到跨视觉和语言的详细语义对齐的联合表示,而且为推理提供了更坚实的事实基础。但是基于外部知识的视觉问答与推理任务仍存在一些局限性:模型联合嵌入了视觉、文本以及外部知识多种信息,而没有细粒度的选择哪些信息是与当前上下文强相关的,这可能会为推理正确答案带来意外的噪音从而导致模型性能不升反降。

2.5 基于可解释推理机制

人工智能技术大多属于黑盒模式,这造成人类无法知晓模型是如何思考的,从而导致模型缺乏可

解释性。研究人员常常利用因果推理[13]和胶囊网络[70]等方法说明模型的可解释性。所以研究人员尝试将可解释性融入到算法中,通过学习显式的可解释的知识构建完成相关跨模态任务。

视觉问答任务中,不仅要求选出与跨模态上下文最相关的答案,还希望模型具备一定可解释性,从而显示展现其推理过程。于是,Cao等[69]提出解析树引导推理网络模型。该网络由3个协作模块组成:1个注意模块可利用问题中解析出的每个单词的局部视觉证据;1个门控残余合成模块,可组成先前挖掘的证据;1个解析树引导传播模块,可沿解析树传递挖掘到的证据。该模型利用从问题中解析出来的单词所构成的树结构进行全局推理,在问题驱动的解析树推理之后逐步导出图像线索,从而能够构建一个可解释的视觉问答系统。Chen等[13]提出了一种模型无关的反事实样本合成(Counterfactual samples synthesizing,CSS)方法,该方法通过在问题中用图像或单词遮挡关键对象,并为其分配不同的基本事实答案,生成大量反事实训练样本。在使用补充样本(即原始样本和生成的样本)进行训练后,VQA模型可关注到所有关键对象和单词上,这大大提高了模型对视觉解释和问题敏感的能力。Urooj等[70]以弱监督的方式建立关于视觉实体的基准,提出了一个视觉胶囊模块,该模块提供具有胶囊特征的问题选择机制,允许模型根据问题中文本线索关注视觉的相关区域,从而增强了模型的可解释性并提升了模型性能。Han等[71]提出了一个新的去偏框架(Greedy gradient ensemble, GGE),它结合了多个有偏模型进行无偏学习,有偏模型即利用数据集虚假相关性进行预测的模型。通过强制有偏模型优先过拟合有偏数据分布,从而使该模型更加关注有偏模型难以求解的示例,以达到增强模型可解释性的目的。Cao等[72]提出了一个知识路由模块化网络的高阶视觉问答方法,该方法中将问题建模为一系列三元组,由问题三元组引导图像场景图和常识知识库构建多步骤推理过程,从而实现显式推理增强模型的可解释性。Yi等[73]则将视觉语言理解与推理进行完全地解耦,考虑了将神经网络与符号计算相结合的方式。首先通过场景解析器将图片转换为结构化场景特征,再通过问题解析器将问题转化为多层级的序列程序,最后将程序执行器作用于结构化场景特征得到最终答案。这种将符号计算与神经网络相结合的分步骤推理过程大大增强了视觉问答模型的可解释性。

在VCR子任务中,为了得到正确答案和正确理由,也要求模型具备一定的可解释性。Wang等[74]提出了一种新的无监督特征表示学习方法,即基于视觉常识区域的卷积神经网络。该模型的训练目标是通过使用因果干预预测区域的上下文对象。Zhang等[75]通过联合建模视觉内容的层次结构以及视觉域和语言域模态间关系,提出了一种多级反事实对比学习网络。首先通过实例级、图像级和语义级对比学习,模型可对图像和语言表达进行全面理解;其次利用反事实思维,生成信息丰富的事实和反事实样本,从而增强模型的感知能力;最后加入了一个辅助对比度模块,以直接优化VCR中的答案预测,从而提升了模型的准确率。

引入因果推理机制和生成反事实样本的方式可以平衡数据集偏差,让模型进行无偏推理。使用胶囊网络和构建分步骤推理过程等方式为模型提供了显式的推理路径,增强了模型的可解释性和泛化能力。尽管上述方法一定程度上增强了视觉常识推理模型的可解释性使其能够稳定学习,但是也加重了模型的计算负担,如何在增强其可解释性的同时保持原有的计算效率值得深入思考。

3 数据集相关研究

对于跨模态任务而言,设计一个强大的模型固然重要,但收集一个符合现实生活观念和实际应用需求的数据集也是相关任务不可缺少的一部分。自视觉问答与推理任务提出以来,大量数据集随之出现。该任务数据集的一般表示形式为<图像、问题、答案>的三元组集合,也有部分数据集有对图像的注释或者额外的标签等,例如视觉常识推理的VCR数据集[4]中包括了每幅图像检测到的视觉对象列表以及每个视觉对象的类别等信息。

3.1 数据集介绍

早期构建的数据集有严重的偏差存在，如VQA-v1数据集[1]中问“是否……”的问题90%的回答都是“yes”，这导致模型往往使用惯性思维进行学习，于是为了平衡这类偏见研究人员又构建了VQA-v2数据集[76]，即针对不同图像提问同一个问题，但答案要代表相反的含义。然而该数据集仍然存在问题，如训练集和测试集的答案分布太过于类似，模型仍可以通过答案分布存在的偏差进行预测。此外，研究人员希望模型不仅具备识别图像文本信息的能力，更要具备挖掘其深层次跨模态知识的认知能力。基于此人们构建了CLEVR[77]、FVQA[78]、OK-VQA[63]、GQA[79]、VQA-CP v2[80]、VCR[4]等数据集。其中，CLEVR数据集使用的图片是经过3D渲染的几何图形，更考验模型的空间结构和逻辑关系的推理能力；而FVQA在构建数据集时引入了3个外部知识库，要求模型使用外部知识才能做出正确回答。同样OK-VQA数据集也要用外部知识才能正确回答问题；GQA数据集是围绕真实世界推理的视觉问答数据集。除了让模型具备认知能力以外，更要考虑的是正确的可解释的推理过程，因此产生了VQA-CP v2数据集，该数据集的训练集和测试集的答案分布不同，可以使模型减少使用语言依赖进行推理判断的情况。对于视觉问答的子任务视觉常识推理来说，VCR数据集和VCR$^+$数据集的场景信息更加丰富，需要识别图像中的物体更多，类型更广泛，问题和答案的长度也更长，更加考验模型的推理认知能力。视觉问答与推理常用数据集如表1所示。

表1 常用视觉问答与推理公开数据集

Table 1 Common datasets in visual question answering and reasoning

数据集	适用任务	图片数量/幅	问题数量/个
VQA-v1[1]	视觉问答	204 721	614 163
VQA-v2[76]	视觉问答	204 721	1 105 904
CLEVR[77]	视觉问答	100 000	1 000 000
FVQA[78]	视觉问答	2 190	5 826
VQA-CP v2[80]	视觉问答	219 000	638 000
OK-VQA[63]	视觉问答	14 031	14 055
GQA[79]	视觉问答	113 000	22 000 000
VCR[4]	视觉常识推理	99 000	290 000
VCR$^+$[81]	视觉常识推理	99 000	290 000

(1)VQA-v1[1]：该数据集由MS COCO数据集的真实场景下的204 721幅图像的614 163个问题和7 984 119个答案，以及50 000个抽象场景的150 000个问题和1 950 000个答案组成。其中，真实场景部分使用123 287幅图片作为训练集和验证集图片、81 434幅图片作为测试集图片，抽象场景部分分别使用20 000、10 000、20 000幅图片作为训练集、验证集、测试集图片。

(2)VQA-v2[76]：该数据集由204 721幅图片、614 163个自然语言问题和超过600万个答案构成。问题类型大约20种，问题的平均长度约为6.2个单词，答案的平均长度约为1.1个单词。训练集包括82 783幅图片，验证集包括40 504幅图片，测试集包括81 434幅图片，每幅图片对应至少3个问题，对于每个问题都标记10个正确答案。其中图片均来自COCO数据集。

(3)CLEVR[77]：该数据集由合成场景下的10万幅图片和大概100万个自动生成的问题答案对构成。图像主要由圆柱体和正方体等各种立体图形经过3D渲染组成，问题中由85.3个互不相同的提问类型构成，包括与属性、材料以及形状等相关的问题、计数问题、比较问题和需要模型进行推理才能得到正确答案的问题，因此该数据集更加考验模型对图像区域空间特征的感知能力。

(4)FVQA数据集[78]：该数据集不仅包括2 190幅图片和5 826个问题答案对，还为每个视觉对象提供了额外的知识。其中，问题类型大约32种，平均问题长度约为9.5个单词，答案的平均长度约为1.2个单词。图片来自MSCOCO验证集和ImageNet的测试集，与每幅图片相关的常识知识来自DBpedia、ConceptNet和WebChild知识库，因此至少参考上述3个知识库的知识，才能正确回答FVQA数据集中

的问题。

(5)VQA-CP v2数据集[80]:该数据集由VQA-v2重构而成,主要通过改变不同类型问题在训练集和测试集中答案的分布以减少答案中存在的统计偏差问题,从而解决VQA-v2数据集中的语言先验问题。其中训练集由121 000幅图片、438 000个问题和4 400 000个答案组成,测试集由98 000幅图片、220 000个问题和2 200 000个答案组成。该数据集的特点是训练集数据和测试集数据分布不同,所以要求模型学习到更鲁棒、更泛化的特征,才能在VQA-CP v2训练集和测试集上都表现良好。

(6)OK-VQA数据集[63]:该数据集由14 031幅图片和14 055个问题答案对构成。其中,图像来自COCO数据集的随机图像,每个问题的平均长度约为8.1个单词,候选答案的平均长度约为1.3个单词。训练集由大约8万幅图像和9 009个问题构成;测试集由大约6万幅图像和5 046个问题构成。该数据集中的问题须使用外部知识才能进行正确推理从而得到正确答案,外部知识涵盖人文地理、日常生活、地方习俗以及烹饪等10个类型。

(7)GQA数据集[79]:该数据集包括113 000幅真实世界的图片和22 000 000个不同类型问题以及每张图像的视觉特征、场景图信息以及空间特征信息。其中,问题设置涉及物体和物体属性的识别、交互关系的跟踪、空间推理和逻辑关系推理等多个方面。其中每幅图片都是丰富的场景图,问题相当于一个复合函数,需要逐步推理才能得到正确答案。

(8)VCR数据集[4]:该数据集包含290 000个多选择QA问题,这些问题源自110 000个电影场景。VCR数据集中的答案和理由通常比VQA数据集更复杂。据统计,答案平均长度超过7.5个单词,而理由的长度最长超过16个单词。训练集包含80 418幅图像和212 923个问题,验证集包含9 929幅图像和26 534个问题,测试集包含9 557个图像和25 263个问题。每个问题都有4个答案,每个问题联合正确答案都有4种基本理由,即每个图像问题对都包含1个正确答案和3个错误的反事实答案,以及1个正确的理由和3个不正确的理由。

(9)VCR$^+$数据集[81]:该数据集是在VCR数据集[4]基础之上进行了去除语言偏差。由于VCR数据集中问题和答案中常出现重复单词,模型很容易利用问题和选项之间重复单词数过多就认定其为正确答案,而非经过正确的推理判断过程得到答案。所以VCR$^+$数据集通过基于对抗和基于规则2种方法对验证集部分答案进行修正以去除数据集偏差。

3.2 现有方法和结果对比

为了方便研究人员更好地了解视觉问答与推理的发展现状,首先对上述所提数据集进行分类,VQA-v1和VQA-v2数据集是常规的视觉问答数据集;FVQA、OK-VQA和GQA都是涉及外部知识的数据集;VQA-CP v2是消除语言先验的数据集;CLEVR是合成场景下的抽象数据集;VCR和VCR$^+$是视觉常识推理相关数据集。由于VQA-v1存在的数据集偏差过于严重,所以现有的方法极少在该数据集上进行训练和测试。因此常规数据集以VQA-v2为例说明现有方法的最新效果。VQA-CP v2、FVQA、OK-VQA和GQA都是知识推理的数据集,由于FVQA在2020年文献[82]中的效果已与人类结果相近,所以近期该数据集的结果未得到更新。GQA是基于现实场景构建的数据集,但其蕴含的常识知识呈隐式状态存在。在OK-VQA数据集中包含的外部知识种类更加丰富,包括车辆交通、公司产品、品牌、食物烹饪、人文地理以及语言文化等10类,因此基于外部知识的数据集以OK-VQA数据集为代表说明现有方法的最新效果。在视觉常识推理子任务中,VCR数据集是普遍使用的数据集。VCR$^+$数据集在2021年被提出,因此目前暂未被研究人员广泛使用。接下来将基于VCR数据集对现有方法进行详细分析。由于CLEVR数据集在近几年出现许多变体,导致该数据集不常被使用,因此不针对该数据集的方法进行分析对比。

3.2.1 VQA-v2数据集上相关方法比较

表2给出了较为先进的VQA模型在VQA-v2数据集上的相关结果,其中Test-dev表示在该数据集的验证测试集上的准确率,Test-std表示在该数据集的标准测试集上的准确率。VILLA方法[82]是先在预训练阶段利用与任务无关的正负样本进行对抗训练,然后再在下游任务中进行对抗微调,从而学习到更泛化的特征表示;UNIMO方法[52]提出一种统一的模型预训练体系结构使文本知识和视觉知识在统一的语义空间中相互增强来学习更具泛化性的表示;VinVL方法[83]在更大的训练语料库上进行了预训练,该语料库结合了多个有标注的对象检测数据集,因此可以生成更丰富的视觉和语言信息;ALBEF方法[84]引入了一种对比损失,在特征融合前利用跨模态注意力对齐图像和文本表示,实现更扎实的视觉和语言表示学习;CLIP-ViL方法[85]使用CLIP[64]作为各种跨模态模型中的视觉编码器,以学习更鲁棒的视觉特征表示;VLMo方法[86]通过模块化Transformer网络来联合学习双编码器和融合编码器更好的学习跨模态特征;SimVLM方法[87]利用大规模的弱监督来降低训练复杂性,并使用单个前缀语言建模目标进行端到端训练;Florence方法[88]将特征表示从粗略场景迁移到精细对象,从静态图像迁移到动态视频,学习到了细粒度的视觉特征;OFA方法[54]获得了当前在VQA-v2数据集上最好的性能,该方式使用了一个统一的基于编码器-解码器的结构,并且在无需引入额外的特定任务层进行微调的情况下,在多种跨模态任务上都取得了很好的效果。

表2 VQA-v2数据集的最新比较

Table 2 Comparison results on VQA-v2 dataset

方法	准确率/%	
	Test-dev	Test-std
VILLA[82]	74.70	74.9
UNIMO[52]	75.10	75.3
VinVL[83]	76.50	76.6
ALBEF[84]	75.80	76.0
CLIP-ViL[85]	76.50	76.7
VLMo[86]	79.90	80.0
SimVLM[87]	80.03	80.3
Florence[88]	80.00	80.4
OFA[54]	**82.00**	**82.0**

以上方法都是基于预训练的方法,但是只有OFA方法[54]具有3大特性:(1)任务无关性:建立了一个支持分类任务和生成任务等任务的通用预训练模型,应用到下游任务时无需微调;(2)模态无关性:在所有任务中共享统一的输入输出表示以处理不同的模态;(3)任务综合性:通过将模型在多种任务上进行预训练增强了模型的泛化能力。然而,VILLA方法[82]和ALBEF方法[84]均需要微调以适应下游任务,无法做到任务无关性;VinVL方法[83]和SimVLM方法[87]需要额外的视觉编码器或者语言编码器进行特征提取,无法做到任务无关性;UNIMO方法[52]和VLMo方法[86]等只能适用于多模态任务,无法做到模态无关性和任务综合性。OFA方法得到了最佳效果。

3.2.2 OK-VQA数据集上相关方法比较

表3给出了较为先进的VQA模型在OK-VQA数据集上的相关结果,并提供了每个方法使用的外部知识库。ArticleNet方法[63]使用基于Transformer的模型从无监督语言预训练和有监督训练中有效地学习隐含知识,再将外部知识与跨模态特征融合执行推理过程;BAN方法[89]在低秩双线性池化的基础上使用双线性注意力机制学习多模态的联合特征,从而实现跨模态语义对齐;GRUC方法[61]从视觉、语义和事实的角度通过多个知识图谱来描绘图像,从局部和全局的角度提取视觉图像的细粒度特征去辅助模型预测;Mucko方法[44]用一个多模态异构图来描述图像,该图包含了对应于视觉、语义和事实特征的多层信息,以学习多层级视觉表示;KM方法[60]通过学习由键-值对记忆网络派生的外部知识嵌入,使文本和视觉信息更加完整,还利用先验知识来提高模型性能;KRISP方法[57]先从预训练模型中学习有效的隐式知识,然后在知识库中提取显式外部知识,将隐式和显式知识结合去预测正确答案;ConceptBert模型[58]是基于BERT的联合概念-视觉-语言嵌入的学习框架,主要通过构建图像的视觉元素和

表3 OK-VQA数据集的最新比较

Table 3 Comparison results on OK-VQA dataset

方法	外部知识来源	准确率/%
ArticleNet(AN)[63]	维基百科	5.28
BAN[89]	维基百科、ConceptNet	25.17
GRUC[61]	ConceptNet	29.87
Mucko[44]	ConceptNet	29.20
KM[60]	来自OK-VQA的跨模态知识	31.32
KRISP[57]	DBpedia+ConceptNet+VisualGenome+haspartHB	38.90
ConceptBert[58]	ConceptNet	33.66
Knowlegde is Power[69]	YAGO3	39.24
MuKEA[90]	来自VQA-v2和OK-VQA的跨模态知识	**42.59**

知识图的多模态表示去推理答案;Knowlegde is Power方法[59]是一个分层知识嵌入式元学习框架,主要解决艺术领域视觉推理的关键问题;MuKEA方法[90]则通过结合多个有偏模型进行无偏学习的方法取得了最好的效果,这种消除语言偏差的方法进一步增强了模型的可解释性,使模型更趋近于真正的无偏推理。

在OK-VQA数据集上进行实验的方法大多需关注如何从结构化或者半结构化知识获取有效的知识。尽管常用的维基百科、ConceptNet以及DBpedia等知识库通过大规模的人类标注提供了高质量的知识,但这些常识知识都仅限于三元组表示、自然语言描述或者低阶谓词表达的事实,ArticleNet[63]、BAN[89]和GRUC[61]等方法仅利用这些低级表示挖掘与跨模态上下文相关的外部知识;MUTAN方法[44]未使用外部知识库,仅从数据集内挖掘内部跨模态知识辅助模型进行预测。而MuKEA方法[90]不仅通过预训练-微调策略来积累数据集内部知识和知识库的外部知识,还考虑了如何表示高阶谓词和多模态知识去帮助模型回答复杂问题。因此该方法得到了最佳性能。

3.2.3 VCR数据集上相关方法分析与比较

表4给出了较为先进的视觉常识推理模型在VCR数据集上的相关结果。其中,Q→A表示给出问题和图像,要求模型选出正确答案;QA→R表示给出图像、问题和正确答案,要求模型选出正确理由;Q→AR表示给出图像和问题,模型不仅要选出正确答案还要选出正确理由来证明所选答案的合理性。R2C模型[4]先用双向长短时记忆网络提取视觉和文本特征,再利用注意力机制进行跨模态特征融合去执行推理;TAB-VCR方法[91]以R2C模型作为基准,进一步提取了图像对象的属性特征,以挖掘细粒度的特征提升模型预测的准确率;HGL模型[6]通过构建模态间和模态内图结构去缩小跨模态语义鸿沟;VisualBERT模型[92]以无监督的方式通过问题和选项中的动词跟踪视觉图像区域之间的关系以完成视觉语言任务;ViLBERT模型[50]将提取到的视觉文本特征输入到共同注意力Transformer中以学习联合映射用于多种跨模态任务中;HSDGN方法则提出一个多层语义增强方向图网络实现高阶跨模态对齐。VL-BERT模型[55]将视觉单词或感兴趣区域和文本中每个单词整体输入到模型中,从而构建了一个通用的模型可学习多种跨模态任务;UNITER模型[93]通过在4个图像-文本大规模数据集的预训练学习,使其可更稳定地应用于具有联合多模态嵌入的异构下游跨模态任务中;ERNIE-ViL方法[24]在预训练阶段从句子中解析出不同类型的节点生成场景图去执行预测任务,使其学习到细粒度的跨模态联合表示;SGEITL方法[37]利用场景图增强图像-文本学习框架取得了最好的性能,这说明构建场景图可以更好地对齐视觉文本特征,缩小视觉文本之间的语义鸿沟。

表4 VCR数据集的最新比较

Table 4 Comparison results on VCA dataset

方法	是否需要预训练	Q→A准确率/%		QA→R准确率/%		Q→AR准确率/%	
		验证	测试	验证	测试	验证	测试
R2C[4]	否	63.8	65.1	67.2	67.3	43.1	44.0
TAB-VCR[100]	否	69.4	70.2	71.9	71.2	50.3	50.2
HGL[6]	否	69.1	69.6	70.5	70.4	49.0	49.8
VisualBERT[61]	否	70.3	71.4	73.0	73.1	51.9	52.2
ViLBERT[50]	否	72.3	72.9	74.6	74.0	54.1	53.9
HSDGN[45]	否	72.9	**73.5**	74.5	**74.2**	54.4	**54.3**
ERNIE-ViL[23]	是	74.1	—	76.9	—	56.9	—
UNITER[60]	是	73.4	—	76.0	—	55.8	—
VL-BERT[56]	是	72.9	—	75.3	—	56.9	—
SGEITL[36]	是	**74.9**	—	**77.2**	—	**57.8**	—

R2C[4]、TAB-VCR[92]以及HGL[6]等方法仅仅考虑了视觉区域-文本单词之间的跨域语义对齐，忽略了视觉概念和文本中短语等内容的隐式对应关系。而HSDGN方法[45]全面地考虑了视觉场景-文本单词之间、视觉区域-文本短语之间以及视觉场景-文本短语之间的3种对应关系，并利用单元交互模块聚合3个层次的高阶特征实现跨模态对齐，还使用了线索感知模块在每个推理步骤中动态选择重要的实体以增强其可解释性。因此，在未经过预训练的情况下HSDGN方法取得了最佳效果。近年来，预训练-微调这一架构在跨模态任务中有广泛应用，SGEITL方法[37]将视觉场景图和单词作为模型输入构建了一个多跳Transformer，并在文本的弱监督下可生成语义丰富的视觉场景图。而与之对比的其他预训练方法均未考虑将图结构引入预训练模型中，未能利用恰当的知识结构去辅助模型推理预测，所以整体而言SGEITL方法取得了最先进的效果。由此可见，基于预训练的方法整体要比未使用预训练的方法性能高很多，也足以说明在未来视觉问答与推理任务的研究中，基于预训练的方法更加具有挖掘潜力。

4 总结与展望

4.1 高层次知识形态表示

除了可以引入外部知识以外，还可以引入更高层级的知识：未知知识。未知知识即无法利用过去经验回答的问题，而是需要通过当前跨模态内容来预测才能得到的知识。例如，文献[94]认为现有的知识型视觉问答在利用外部知识时可能检索到无关知识或引入有害噪声，因此该文献不仅检索到外部知识库的常识知识，还进一步挖掘了视觉文本上下文预测得到的知识，即未知知识，进行多种知识源融合。除此之外，该方法还旨在了解每个候选答案应该信任哪个知识源，并考虑如何让利用该知识源去验证候选答案的正确性。然而，现有的基于知识的框架基本都是利用先验知识来实现推理预测的，基于未知知识去完成跨模态视觉问答与推理任务的方法仍处于发展匮乏的阶段。所以如何构建更高层级的知识形态来促进模型达到更好的效果是个值得思考和探究的问题。

4.2 更通用的预训练方法

基于Transformer框架的模型在自然语言处理和计算机视觉等各个领域取得了极大的进展，这也预示着用一个统一的Transformer架构学习不同模态知识解决不同领域任务的可能性。然而，现有的模型

在迁移到下游时需要进行微调才能显著地提升下游任务的性能。目前已有一些通用的预训练模型[54,86]可以应用到视觉问答、视觉常识推理以及跨模态检索等多模态任务中,如文献[52]提出了一种统一的模型预训练体系结构,该结构通过利用大规模的自由文本语料库和图像集合来提高文本和视觉理解能力,并利用跨模态对比学习将文本和视觉信息对齐到一个统一的语义空间中。尽管该预训练模型实现了跨模态任务上的通用,但是如何构建一个无须进行微调即可迁移到纯视觉任务、纯文本任务以及视觉文本跨模态任务中的通用预训练方法值得研究者进一步深入思考。此外,研究者也可挖掘具备知识类型更复杂多样的大规模预训练数据集以学习到更泛化的特征。

4.3 生成式视觉问答与推理

目前,视觉问答与推理任务大多被定义为分类问题,但在实际应用中,针对一个问题可能存在很多正确答案,所以分类式视觉问答与推理系统一定程度上限制了模型的想象力。因此根据不同场景和不同问题,视觉问答与推理系统理应具备直接给出正确的、精准的答案的能力。尽管已有一些生成式的视觉问答与推理相关任务,例如文献[95]提出了一个基于生成式视觉问答任务框架,该框架将图像和问题作为输入,利用人工标注者的有效标注信息和网络数据噪声的辅助去生成一个合理准确的答案。由于生成式问答任务在视觉问答领域中刚刚崭露头角,所以现有方法生成的答案仍不够精准,且通常只能利用视觉图像中表面的信息进行推理,而无法综合常识知识等深层信息。因此,可以通过构建图像问题种类更丰富的数据集,让模型可以生成更加多样的答案。此外,生成式的设定更有利于将模型迁移到其他任务中,如视觉对话任务。所以,生成式设定具有广阔的研究空间,值得进一步探索。

4.4 新型视觉推理数据集构建

目前,对于视觉问答任务而言,已有很多不同种类且丰富多样的数据集,如OK-VQA数据集考察模型汲取外部知识的能力;VQA-CP v2数据集更考察模型对数据在训练和测试阶段分布不同时的泛化能力。但是对于视觉常识推理任务而言,现有的主流数据集仍是2018年提出视觉常识推理任务的VCR数据集,尽管也有研究人员提出了新数据集,例如VCR$^+$数据集[81]、GD-VCR数据集[96],但是只有提出该数据集的作者进行了相应的性能验证,其他领域并未使用该数据集。因此,研究人员可以考虑以下关于数据集方面的创新:(1)构建考验模型推理能力的、需要知识种类更丰富的以及图像种类问题设置更多样的数据集;(2)实现数据集评价指标多元化,除了准确率以外,也可考虑加入召回率等指标全面地评估模型的推理能力;(3)现有的视觉常识推理任务都是基于图像的,可以构建关于视频的常识推理数据集,这样可以更考验模型对跨模态特征的提取能力及模型的推理预测能力。

参考文献:

[1] ANTOL S, AGRAWAL A, LU J, et al. VQA: Visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2015: 2425-2433.

[2] 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 (CVPR). [S.l.]: IEEE, 2018: 6077-6086.

[3] ZHU Y, GROTH O, BERNSTEIN M, et al. Visual7W: Grounded question answering in images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2016: 4995-5004.

[4] ZELLERS R, BISK Y, FARHADI A, et al. From recognition to cognition: Visual commonsense reasoning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019: 6720-6731.

[5] YANG X, ZHANG H, QI G, et al. Causal attention for vision language tasks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021: 9842-9852.

[6] YU W, ZHOU J, YU W, et al. Heterogeneous graph learning for visual commonsense reasoning[J]. Advances in Neural

Information Processing Systems (NIPS), 2019, 249: 2769-2779.

[7] SONG D, MA S, SUN Z, et al. KVL-BERT: Knowledge enhanced visual-and-linguistic BERT for visual commonsense reasoning[J]. Knowledge-Based Systems, 2021, 230: 107408.

[8] 王雪枫,张雪松,王峰,等.视觉问答中的模型分析与展望[J].阜阳师范大学学报(自然科学版), 2022, 39(2): 76-84.
WANG Xuefeng, ZHANG Xuesong, WANG Feng, et al. Model analysis and prospects in visual question answering[J]. Journal of Fuyang Normal University (Natural Science Edition), 2022, 39(2): 76-84.

[9] 包希港,周春来,肖克晶,等.视觉问答研究综述[J].软件学报, 2021, 32(8): 2522-2544.
BAO Xigang, ZHOU Chunlai, XIAO Kejing, et al. Review of visual question answering[J]. Journal of Software, 2021, 32(8): 2522-2544.

[10] 牛玉磊, 张含望.视觉问答与对话综述[J].计算机科学, 2021, 48(3): 87-96.
NIU Yulei, ZHANG Hanwang. Review of visual question and answer and dialogue[J]. Computer Science, 2021, 48(3): 87-96.

[11] 王瑞平,吴士泓,张美航,等.知识型视觉问答研究综述[J].计算机科学, 2022(8): 2522-2544.
WANG Ruiping, WU Shihong, ZHANG Meihang, et al. A review of knowledge-based visual question answering[J]. Computer Science, 2022(8): 2522-2544.

[12] WU Q, TENEY D, WANG P, et al. Visual question answering: A survey of methods and datasets[J]. Computer Vision and Image Understanding (CVIU), 2017, 163: 21-40.

[13] 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 (CVPR). [S.l.]: IEEE, 2020: 10800-10809.

[14] ZHANG X, ZHANG F, XU C. Explicit cross-modal representation learning for visual commonsense reasoning[J]. IEEE Transactions on Multimedia, 2021, 24: 2986-2997.

[15] NIU Y, TANG K, ZHANG H, et al. Counterfactual VQA: A cause-effect look at language bias[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021: 12700-12710.

[16] SELVARAJU R R, LEE S, SHEN Y, et al. Taking a hint: Leveraging explanations to make vision and language models more grounded[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2019: 2591-2600.

[17] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2017: 299-307.

[18] BEN-YOUNES H, CADENE R, CORD M, et al. Mutan: Multimodal tucker fusion for visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision (CVPR). [S.l.]: IEEE, 2017: 2612-2620.

[19] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: [s.n.], 2016: 289-297.

[20] XU H, YAN M, LI C, et al. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning[EB/OL]. (2021-06-04)[2022-10-10]. https://arxiv.org/abs/2106.01804v2.

[21] KAMATH A, SINGH M, LECUN Y, et al. MDETR-modulated detection for end-to-end multi-modal understanding[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2021: 1780-1790.

[22] MARASOVIĆ A, BHAGAVATULA C, PARK J S, et al. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs[EB/OL]. (2020-10-15)[2022-10-10]. https://arxiv.org/abs/2010.07526v1.

[23] CHEN Z, CHEN J, GENG Y, et al. Zero-shot visual question answering using knowledge graph[C]//Proceedings of the Semantic Web-ISWC 2021. Cham: Springer International Publishing, 2021: 146-162.

[24] YU F, TANG J, YIN W, et al. Ernie-ViL: Knowledge enhanced vision-language representations through scene graphs[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). [S.l.]: AAAI, 2021: 3208-3216.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2017: 6000-6010.

[26] 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(CVPR). [S.l.]: IEEE, 2016: 21-29.

[27] WU C, LIU J, WANG X, et al. Object-difference attention: A simple relational attention for visual question answering[C]//

Proceedings of the 26th ACM International Conference on Multimedia. [S.l.]: ACM, 2018: 519-527.

[28] PENG L, YANG Y, BIN Y, et al. Word-to-region attention network for visual question answering[J]. Multimedia Tools and Applications, 2019, 78(3): 3843-3858.

[29] CAI Linqin, LIAO Zhongxu, ZHOU Sitong, et al. Visual question answering combining multi-modal feature fusion and multi-attention mechanism[C]//Proceedings of 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). [S.l.]: IEEE, 2021: 1035-1039.

[30] GUO W, ZHANG Y, YANG J, et al. Re-attention for visual question answering[J]. IEEE Transactions on Image Processing, 2021, 30: 6730-6743.

[31] YU J, ZHANG W, LU Y, et al. Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval[J]. IEEE Transactions on Multimedia, 2020, 22(12): 3196-3209.

[32] YANG X, ZHANG H, QI G, et al. Causal attention for visionlanguage tasks[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021: 9842-9852.

[33] DING D, HILL F, SANTORO A, et al. Attention over learned object embeddings enables complex visual reasoning[EB/OL]. (2021-10-26)[2022-10-10]. https://arxiv.org/abs/2012.08508v3.

[34] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.

[35] LI C, CAO Y, HOU L, et al. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). [S.l.]: Association for Computational Linguistics, 2019: 2723-2732.

[36] WANG J, BAO B K, XU C. DualVGR: A dual-visual graph reasoning unit for video question answering[J]. IEEE Transactions on Multimedia, 2021, 24: 3369-3380.

[37] WANG Z, YOU H, LI L H, et al. SGEITL: Scene graph enhanced image-text learning for visual commonsense reasoning [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022: 5914-5922.

[38] ANDREAS J, ROHRBACH M, DARRELL T, et al. Learning to compose neural networks for question answering[J].(2016-06-07)[2022-10-10]. https://arxiv.org/abs/1601.01705.

[39] KRISHNA R, ZHU Y, GROTH O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[40] ANDERSON P, WU Q, TENEY D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2018: 3674-3683.

[41] TENEY D, LIU L, VAN DEN HENGEL A. Graph-structured representations for visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2017: 1-9.

[42] LI L, GAN Z, CHENG Y, et al. Relation-aware graph attention network for visual question answering[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). [S.l.]: IEEE, 2019: 10313-10322.

[43] SHI J, ZHANG H, LI J. Explainable and explicit visual reasoning over scene graphs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019: 8376-8384.

[44] ZHU Z, YU J, WANG Y, et al. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering [EB/OL]. (2021-11-04)[2022-10-10]. https://arxiv.org/abs/2006.09073v3.

[45] WU M, QI S, RAO J, et al. Hierarchical semantic enhanced directional graph network for visual commonsense reasoning[C]//Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing. New York, NY, United States: ACM, 2021: 27-36.

[46] WU Q, SHEN C, WANG P, et al. Image captioning and visual question answering based on attributes and external knowledge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1367-1381.

[47] WEN Z, PENG Y. Multi-level knowledge injecting for visual commonsense reasoning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(3): 1042-1054.

[48] ZAREIAN A, WANG Z, YOU H, et al. Learning visual commonsense for robust scene graph generation[C]//Proceedings of European Conference on Computer Vision (ECCV). Cham: Springer, 2020: 642-657.

[49] WU A, ZHU L, HAN Y, et al. Connective cognition network for directional visual commonsense reasoning[C]// Proceedings of Neural Information Processing Systems. Vancouver, Canada: [s.n.], 2019.

[50] LU J, BATRA D, PARIKH D, et al. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2019: 13-23.

[51] ZHONG H, CHEN J, SHEN C, et al. Self-adaptive neural module transformer for visual question answering[J]. IEEE Transactions on Multimedia, 2020, 23: 1264-1273.

[52] LI W, GAO C, NIU G, et al. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning[C]//Proceedings of Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2020.

[53] MA X, NIE W, YU Z, et al. RelViT: Concept-guided vision transformer for visual relational reasoning[J]. (2022-06-11)[2022-10-20]. https://arxiv.org/abs/2204.11167.

[54] WANG P, YANG A, MEN R, et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//Proceedings of International Conference on Machine Learning (PMLR). [S.l.]: IEEE, 2022: 23318-23340.

[55] SU W, ZHU X, CAO Y, et al. VL-BERT: Pre-training of generic visual-linguistic representations[J]. (2021-02-18)[2022-10-10]. https://arxiv.org/abs/1908.08530v3.

[56] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of International Conference on Machine Learning (ICML). [S.l.]: IEEE, 2021: 8748-8763.

[57] MARINO K, CHEN X, PARIKH D, et al. KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021: 14111-14121.

[58] GARDÈRES F, ZIAEEFARD M, ABELOOS B, et al. ConceptBERT: Concept-aware representation for visual question answering[C]//Findings of the Association for Computational Linguistics: (EMNLP). [S.l.]: Association for Computational Linguistics, 2020: 489-498.

[59] ZHENG W, YAN L, GOU C, et al. Knowledge is power: Hierarchical-knowledge embedded meta-learning for visual reasoning in artistic domains[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. [S.l.]: ACM, 2021: 2360-2368.

[60] ZHENG W, YAN L, GOU C, et al. KM4: Visual reasoning via knowledge embedding memory model with mutual modulation [J]. Information Fusion, 2021, 67: 14-28.

[61] YU J, ZHU Z, WANG Y, et al. Cross-modal knowledge reasoning for knowledge-based visual question answering[J]. Pattern Recognition, 2020, 108: 107563.

[62] ZHANG L, LIU S, LIU D, et al. Rich visual knowledge-based augmentation network for visual question answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4362-4373.

[63] MARINO K, RASTEGARI M, FARHADI A, et al. Ok-VQA: A visual question answering benchmark requiring external knowledge[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019: 3195-3204.

[64] DAI W, HOU L, SHANG L, et al. Enabling multimodal generation on CLIP via vision-language knowledge distillation[J]. (2022-03-30)[2022-10-10]. https://arxiv.org/abs/2203.06386.

[65] CHAE J, KIM J. Uncertainty-based visual question answering: Estimating semantic inconsistency between image and knowledge base[C]//Proceedings of 2022 International Joint Conference on Neural Networks (IJCNN). [S.l.]: IEEE, 2022: 1-9.

[66] WHITEHEAD S, WU H, JI H, et al. Separating skills and concepts for novel visual question answering[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021.

[67] ZHANG Y, JIANG M, ZHAO Q. Explicit knowledge incorporation for visual reasoning[C]// Proceedings of 2021 IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021: 1356-1365.

[68] ZELLERS R, LU X, HESSEL J, et al. MERLOT: Multimodal neural script knowledge models[J]. Advances in Neural Information Processing Systems, 2021, 34: 23634-23651.

[69] CAO Q, LIANG X, LI B, et al. Interpretable visual question answering by reasoning on dependency trees[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(3): 887-901.

[70] UROOJ A, KUEHNE H, DUARTE K, et al. Found a reason for me? Weakly-supervised grounded visual question answering using capsules[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2021: 8465-8474.

[71] HAN X, WANG S, SU C, et al. Greedy gradient ensemble for robust visual question answering[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). [S.l.]: IEEE, 2021: 1584-1593.

[72] CAO Q, LI B, LIANG X, et al. Explainable high-order visual question reasoning: A new benchmark and knowledge-routed network[J]. (2019-09-23)[2022-10-10]. https://arxiv.org/abs/1909.10128.

[73] YI K, WU J, GAN C, et al. Neural-symbolic VQA: Disentangling reasoning from vision and language understanding[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. [S.l.]: ACM, 2018: 1039-1050.

[74] WANG T, HUANG J, ZHANG H, et al. Visual commonsense R-CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). [S.l.]: IEEE, 2020: 10760-10770.

[75] ZHANG X, ZHANG F, XU C. Multi-level counterfactual contrast for visual commonsense reasoning[C]//Proceedings of the 29th ACM International Conference on Multimedia. [S.l.]: ACM, 2021: 1793-1802.

[76] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the V in VQA matter: Elevating the role of image understanding in visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Eecognition (CVPR). [S.l.]: IEEE, 2017: 6904-6913.

[77] 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 Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2017: 2901-2910.

[78] WANG P, WU Q, SHEN C, et al. FVQA: Fact-based visual question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(10): 2413-2427.

[79] 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 (CVPR). [S.l.]: IEEE, 2019: 6700-6709.

[80] 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 (CVPR). [S.l.]: IEEE, 2018: 4971-4980.

[81] YE K, KOVASHKA A. A case study of the shortcut effects in visual commonsense reasoning[C]//Proceedings of the Association for the Advance of Artificial Intelligence (AAAI). [S.l.]: AAAI, 2021: 3181-3189.

[82] GAN Z, CHEN Y C, LI L, et al. Large-scale adversarial training for vision-and-language representation learning[J]. Advances in Neural Information Processing Systems (NIPS), 2020, 33: 6616-6628.

[83] LI J, SELVARAJU R, GOTMARE A, et al. Align before fuse: Vision and language representation learning with momentum distillation[J]. Advances in Neural Information Processing Systems, 2021, 34: 9694-9705.

[84] DOU Z Y, XU Y, GAN Z, et al. An empirical study of training end-to-end vision-and-language transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). [S.l.]: IEEE, 2022: 18166-18176.

[85] SHEN S, LI L H, TAN H, et al. How much can CLIP benefit vision-and-language tasks? [EB/OL]. (2021-07-13)[2022-10-10]. https://arxiv.org/abs/2107.06383v1.

[86] WANG W, BAO H, DONG L, et al. VLMO: Unified vision-language pre-training with mixture-of-modality-experts[J]. (2022-05-27)[2022-10-10]. https://arxiv.org/abs/2111.02358v1.

[87] WANG Z, YU J, YU A W, et al. SimVLM: Simple visual language model pretraining with weak supervision[EB/OL]. (2022-05-15)[2022-10-10]. https://arxiv.org/abs/2108.10904v1.

[88] YUAN L, CHEN D, CHEN Y L, et al. Florence: A new foundation model for computer vision[J]. (2022-05-27)[2022-10-10]. https://arxiv.org/abs/2111.02358v1.

[89] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[EB/OL]. (2018-10-19)[2022-10-10]. https://arxiv.org/abs/1805.07932v1.

[90] DING Y, YU J, LIU B, et al. MuKEA: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). [S.l.]: IEEE, 2022: 5089-5098.

[91] LIN J, JAIN U, SCHWING A. TAB-VCR: Tags and attributes based VCR baselines[EB/OL]. (2020-01-09)[2022-10-10]. https://arxiv.org/abs/1910.14671

[92] LI L H, YATSKAR M, YIN D, et al. VisualBERT: A simple and performant baseline for vision and language[EB/OL]. (2019-08-09)[2022-10-10]. https://arxiv.org/abs/1908.03557.

[93] CHEN Y C, LI L, YU L, et al. UNITER: Universal image-text representation learning[C]//Proceedings of European Conference on Computer Vision (ECCV). Cham: Springer, 2020: 104-120.

[94] LI J, LI D, XIONG C, et al. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation[EB/OL]. (2022-02-15)[2022-10-10]. https://arxiv.org/abs/2201.12086v2.

[95] CHEN C, ANJUM S, GURARI C. Grounding answers for visual questions asked by visually impaired people[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).[S.l.]: IEEE, 2022: 19076-19085.

[96] YIN D, LI L H, HU Z, et al. Broaden the vision: Geo-diverse visual commonsense reasoning[EB/OL]. (2021-09-14)[2022-10-10]. https://arxiv.org/abs/2109.06860.

作者简介:

张飞飞(1989-),**通信作者**,女,教授,硕士生导师,研究方向:多媒体计算、计算机视觉、模式识别、图像处理等,E-mail: feifeizhang@email.tjut.edu.cn。

张建庆(1997-),女,硕士研究生,研究方向:视觉问答、视觉常识推理。

屈思佳(1998-),女,硕士研究生,研究方向:跨模态检索。

周琬婷(1991-),女,副教授,博士生导师,研究方向:人工智能、机器学习和模式识别等,E-mail: wanting.zhou@bupt.edu.cn。

(编辑:张黄群)