计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200174

视觉问答与对话综述

牛玉磊　张含望
南洋理工大学计算机科学与工程学院　新加坡 639798

摘　要　视觉问答与对话是人工智能领域的重要研究任务,是计算机视觉与自然语言处理交叉领域的代表性问题之一。视觉问答与对话任务要求机器根据指定的视觉图像内容,对单轮或多轮的自然语言问题进行作答。视觉问答与对话对机器的感知能力、认知能力和推理能力均提出了较高的要求,在跨模态人机交互应用中具有实用前景。文中对近年来视觉问答与对话的研究进展进行了综述,对数据集和算法进行了归纳,对研究挑战和问题进行了总结,最后对视觉问答与对话的未来发展趋势进行了讨论。

关键词:视觉问答;视觉对话;视觉语言;视觉推理;深度学习
中图法分类号　TP391

Survey on Visual Question Answering and Dialogue

NIU Yu-lei and ZHANG Han-wang
School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore

Abstract　Visual question answering and dialogue are important research tasks in artificial intelligence, and the representative problems in the intersection of computer vision and natural language processing. Visual question answering and dialogue tasks require the machine to answer single-round or multi-round questions based on the specified visual content. Visual question answering and dialogue require the machine's abilities of perception, cognition and reasoning, and have application prospects in cross-modal human-computer interaction applications. This paper reviews recent research progress of visual question answering and dialogue, and summarizes datasets, algorithms, challenges, and problems. Finally, this paper discusses the future research trend of visual question answering and dialogue.

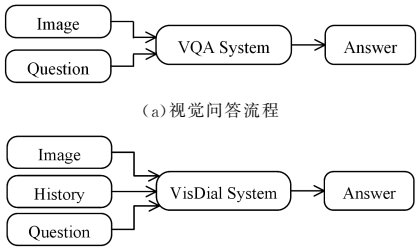
Keywords　Visual question answering, Visual dialogue, Vision and language, Visual reasoning, Deep learning

1　引言

随着深度学习技术和计算资源的发展,计算机视觉和自然语言处理领域中的众多基础任务得到了广泛关注,并取得了一定的研究进展,如图像分类、目标检测、问答系统、对话系统等。随着基础任务的突破以及计算机感知能力的提高,研究者们开始聚焦于计算机视觉和自然语言处理的交叉领域。从研究的角度出发,研究者们希望计算机不仅具有基础的感知与认知能力,而且进一步具备多模态推理能力。从应用的角度考虑,用户希望能够以更加友好的交互方式使用计算机视觉系统。因此,视觉语言领域(Vision and Language)的相关任务在近年来得到了广泛关注。

视觉问答与对话是视觉语言领域中基础的研究任务之一,是实现面向用户的人机交互式视觉系统的重要途径。视觉问答与对话任务流程如图 1 所示。在视觉问答(Visual Question Answering, VQA)任务中,计算机以图像和问题为输入,通过对视觉(即图像)和自然语言(即问题)两个模态进行理解与推理,输出自然语言作为对应的答案。视觉对话(Visual Dialogue)可以看作视觉问答的一般形式,即多轮视觉问答。在视觉对话任务中,计算机除了对图像和问题进行理解之外,还要结合由问答对组成的对话历史上下文信息进行推理,并输出自然语言作为相应的回答。视觉问答与对话是众多复杂的人工智能系统与应用中不可或缺的组成部分。例如,在火灾等救援场景中,救援人员为了保障自身安全和了解场景情况,不便直接深入救援现场。在这种情况下,救援人员希望以交互式的方式,通过直观的自然语言与救援机器人进行沟通。救援机器人根据用户的自然语言指令,对现场进行探查反馈,并根据后续指令进行下一步操作。在该过程中,救援机器人需要对自然语言进行理解,并根据指令对其所见的视觉信息进行感知与推理,最终将相应的视觉内容反馈给救援人员。这一基于视觉的交互形式,可以通过视觉问答与对话来实现。

到稿日期:2020-12-20　返修日期:2021-01-28
基金项目:阿里巴巴-南洋理工大学新加坡联合研究所
This work was supported by the NTU-Alibaba JRI.
通信作者:牛玉磊(yn.yuleiniu@gmail.com)



(a)视觉问答流程

(b)视觉对话流程

图 1 视觉问答与对话流程

Fig. 1 Example for visual QA and dialogue

本文对视觉问答与对话的研究进展进行了梳理与总结，详细分析了现有研究方法及面临的挑战，同时对未来的研究方向和问题进行了讨论。Yu 等[1]对早期的视觉问答工作进行了总结。随后，研究人员在视觉问答领域取得了一系列的研究进展，同时视觉对话任务也被提出并获得关注。本文将对近两年的视觉问答工作进行归纳，并对视觉对话任务的进展进行总结。

2 问题与挑战

2.1 研究挑战

2.1.1 多模态感知与理解

视觉问答与对话作为视觉语言领域的经典问题，需要同时处理来自视觉和语言两个模态的输入。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。

具体到视觉问答与对话，视觉端的输入为图像，语言端的输入为一段自然语言描述。视觉问答任务的自然语言描述的输入为一个问题。视觉对话任务的自然语言描述的输入为一个问题和由若干个问答对组成的对话历史。多模态输入的处理可以分为感知与推理两个部分。感知强调单模态特征提取，如通过卷积神经网络提取视觉特征，通过词向量模型和长短时神经网络等提取文本特征。理解强调视觉特征和文本特征的进一步交互和关联，通过理解问题的意图并能正确回应，进而得到多模态联合特征表示。

2.1.2 数据集偏差

数据集偏差(Dataset Bias)是计算机视觉领域和自然语言领域中普遍面临的挑战。作为视觉语言领域的代表性问题，视觉问答和对话任务由于同时具有视觉和语言两种输入，因此也面临着数据集偏差的问题。

视觉问答任务中的数据集偏差主要是语言偏差(Language Bias)。语言偏差问题具体体现在视觉问答数据集中的问题与答案存在的强相关关系。例如，对于所有的以“你是否看见”为开头的问题而言，接近 90%的问题的答案为“是”。又如，对于与运动相关的问题而言，接近 40%的答案为“网球”。在训练过程中，视觉问答模型会过于依赖问题与答案之间的相关性，通过记住问题与答案之间的匹配模式进行学习，从而忽略了对图像内容的探索。因此，视觉问答模型的通用性和鲁棒性受到了极大的限制。如何解决语言偏差、提高模型的通用性和鲁棒性，已经成为视觉问答任务的重要挑战。

与视觉问答类似，视觉对话任务也面临着数据集偏差的问题。视觉对话任务中的偏差问题主要体现在以下两个方面。1)模型没有充分利用多源、多模态的输入，而是将视觉对话简化为视觉问答或自然语言问答问题。由于视觉对话是视觉问答的一般形式，理想的视觉对话模型除了根据当前问题和图像进行作答外，还需要结合对话历史信息对问题的语义进行理解。如果模型没有对历史信息进行充分理解与应用，那么视觉对话问题将被简化为视觉问答问题进行处理。类似地，视觉对话问题也有可能被进一步简化为问答问题。2)模型过分依赖对话历史中体现的语言偏好与习惯。Qi 等[2]发现，传统的视觉对话模型会将对话历史中的答案长度与当前问题的答案长度进行关联。在进行回答时，模型会倾向于根据历史答案的平均长度决定回复的长短。这一现象说明，模型可以通过记忆历史答案进行作答。对历史答案的记忆，可能会导致作答时过多关注答案长度，而忽视了答案内容。

2.1.3 视觉指代消解

视觉指代消解(Visual Co-reference Resolution)是计算机视觉领域和自然语言领域中普遍面临的挑战。指代消解(Co-reference Resolution)又称为共指消解，是自然语言处理、信息检索、机器翻译等自然语言相关领域中的重要技术。在自然语言中，人们为了避免用词重复，往往会通过代词、缩略语等形式指代同一用词。指代导致语句表意不够明确、语句结构不完整，给机器理解自然语言带来了障碍。指代广泛存在于人的对话中，尤其在人们讨论同一个人或物体时。视觉对话可以看作带有一定目的性的对话。视觉对话围绕给定的视觉场景(即图像)展开，提问者对其看不见的视觉场景进行询问，回答者根据提问者的问题与视觉场景进行作答。由于指代的存在，回答者不仅需要将指代词进行消解，而且需要进一步将指代词与视觉场景中的目标物体关联起来，从而准确理解提问者的意图并作答。因此，视觉指代消解是机器实现视觉对话和复杂视觉推理的重要手段之一。

2.2 数据集与评测

2.2.1 视觉问答

目前视觉问答任务最常用的数据集为 VQA 数据集[3-4]，又称为 VQA-real。VQA 数据集的图像来自 MSCOCO 数据集[5]。MSCOCO 数据集的图像包括多个物体与丰富的视觉背景，为视觉问答提供了丰富的视觉复杂度。VQA 数据集共分为 v1 和 v2 两个版本。VQA v1[3]共包含 12 万张训练集图片和 8 万张测试集图片。在问答收集过程中，标注者被要求提供多样性的问题，包括开放式问题和封闭式问题。每个问题由 10 个标注者独立地进行答案的标注，共得到 10 个标注答案。通过上述标注流程，VQA v1 数据集共得到 61 万条问题和 614 个标注答案。然而，VQA v1 数据集的问题和答案存在强相关性。在不考虑任何视觉信息的情况下，简单地利用语言偏差进行作答，选择问题所属类型最高频的回答，在运动和计数类问题上可以取得约 40%的准确率。为了消除语言偏差的影响，让视觉问答模型能够更好地利用视觉信息进行作答，研究者对 v1 版本进行了完善，进一步提出了 VQA v2 数据集[4]，以实现准确的评估。VQA v2 数据集的图片同样来自 MSCOCO 数据集。对于数据集中的每一个图像、问题和答案三元组，标注平台提供了与图像相似的 24 张图片。标注者需要在这些候选图片中选择与问题所问内容相匹配但

答案不同的图片,以构成新的三元组。通过上述方式,每个问题都匹配了至少两个不同的答案。这种方式有效地将答案分布进行了平衡,从而消除了语言偏差。通过这一标注流程,在图像规模不变的基础上,VQA v2 共包含了 110 万条问题和 1105 万个答案。与 VQA v1 相比,VQA v2 的问答数据规模扩充了近一倍。

2.2.2 视觉对话

视觉对话任务的常用数据集为 VisDial[6]。VisDial 共分为 v0.9 和 v1.0 两个版本。VisDial v0.9 基于 MSCOCO 数据集的图像和标题进行收集。对于一个图片上的对话而言,两名标注人员通过交互游戏来实现标注。在游戏中,一名标注者扮演提问者,另一名标注人员扮演回答者。提问者只能看到标题和对话历史,看不到图像;而回答者可以同时看到标题、对话历史和图像。为了解图像内容,提问者对看不见的图像进行连续提问。回答者根据提问者的问题,综合图像和对话历史进行作答。通过这一数据收集的流程,每张图片匹配了 10 轮问答的对话。VisDial v0.9 分为训练集和验证集两个子集。其中,训练集包含 8.3 万个对话,验证集包含 4 万个对话。VisDial v1.0 的收集过程与 VisDialv0.9 相同。VisDial v1.0 共分为训练集、验证集和测试集 3 个子集。其中,VisDial v1.0 的训练集由 VisDial v0.9 的全部数据组成,图像与对话均基于 MSCOCO 数据集得到。VisDial v1.0 的验证集和测试集则基于 Flickr 的图像生成。其中,VisDial v1.0 的验证集包含 2000 个对话,测试集包含 8000 个对话。

2.2.3 评测指标

VQA 数据集提供了两种评测任务。第一种评测任务为开放式问答(open-ended),评测指标为准确率。对于每一个预测答案而言,其准确率的计算方式为:

$$\text{准确率}=\min\left(\frac{\text{标注此答案的标注者人数}}{3},1\right)$$

换言之,如果至少 3 个标注者将一个答案标注为正确答案,那么这个答案就认为是 100%正确的。第二种评测任务为多选式问答(multiple-choice)。在多选式问答中,每个问题分配了 18 个候选答案,候选答案由正确答案和错误答案构成。在候选答案中,正确答案为被最多数标注者标注的答案。错误答案由 3 部分组成,分别为合理答案、常见答案和随机答案。合理答案为在不给出图像的情况下给出的答案,即依据常识作答。常见答案为数据集中最高频的答案,包括“是”“否”“2”等。随机答案为从其他问题的答案中随机选取的答案。多选式问答的评测指标与开放式问答相同。

视觉对话的评测方法借鉴了检索问题中的评测方式。在评测过程中,每个待回答问题分配了一个包含 100 个答案的候选集,模型要求返回这 100 个答案的排序。与视觉问答类似,视觉对话的答案候选集也包括了标准答案、常见答案、随机答案等。根据标注答案数量的不同,视觉对话的评测指标可以分为两大类。第一类评测指标基于稀疏的答案标注,即 100 个答案中有且只有一个正确答案。这一类评测指标包括标准答案的平均排序(Mean)、标准答案在前 k 个排序中的比例(Recall@k)和平均倒数排序(Mean Reciprocal Rank,MRR)。第二类评测指标基于密集的答案标注,即候选集中有若干个正确的标准答案,正确程度通过 0 到 1 之间的相关性得分进行表示。这一评测指标为归一化折损累计增益(Normalized Discounted Cumulative Gain,NDCG)。

2.2.4 其他相关数据集

除了 VQA 数据集以外,一些其他的相关研究也围绕视觉问答任务提出了不同的数据集。如表 1 所列,具有代表性的数据集包括 DAQUAR,COCO-QA,FM-IQA,Visual Genome QA,Visual7W,Visual Madlibs 等。DAQUAR 数据集是由 Malinowski 等[7]于 2014 年提出的,其基于 NYU-Depth v2 图像数据集构建,包含了 1449 张图片和 12468 个问答对。DAQUAR 数据集的答案限定在预先定义的 16 种颜色和 894 个物体类别中。COCO-QA 数据集是由 Ren 等[8]于 2015 年提出的,其基于 MSCOCO 图像数据集构建,包含了近 12 万的图像和问题。在数据集构建过程中,问题将 MSCOCO 的图像描述进行自动转换生成,并分别归入物体、数量、颜色和地点等类别。FM-IQA 数据集[9]同样是基于 MSCOCO 数据集构建的,共包含 12 万张图像和 25 万个问答对。Visual Genome QA[10]和 Visual7W[11]是基于 Visual Genome 数据集构建的。其中除问答对外,每张图片还附带了场景图(Scene Graph)标注的结构信息。这一结构信息有助于提高视觉问答系统的推理能力。Visual Madlibs[12]数据集采用了完形填空式的任务,要求计算机对挖空的句子进行补充。Visual Madlibs 数据集包含了 1 万余张 MSCOCO 数据集的图像与 36 万个自然语言描述。挖空的句子通过模板进行自动生成。

表 1 视觉问答数据集

Table 1 VQA Datasets

Dataset	Source of images	Number of images	Number of questions	Average question length	Annotation
DAQUAR	NYU-Depth v2	1449	12468	11.5	Human
COCO-QA	MSCOCO	117684	117684	9.7	Automatic
FM-IQA	MSCOCO	158392	316193	7.4	Human
Visual Genome QA	MSCOCO	108000	1445322	5.7	Human
Visual7W	MSCOCO	47300	327939	6.9	Human
Visual Madlibs	MSCOCO	10738	360001	6.9	Human
VQA-real v1	MSCOCO	240721	614163	6.2	Human
VQA-real v2	MSCOCO	240721	1105904	6.4	Human

对于视觉对话任务而言,除了 VisDial 之外,“GuessWhat?!”数据集[13](下文称为“GuessWhat”)也探索了视觉与对话相结合的问题。GuessWhat 属于任务型导向的视觉对话,通过两个玩家之间的游戏进行数据标注。在游戏中,提问者和回答者均能看到给定的图像,同时回答者被随机告知了图像中的一个目标物体。提问者需要向回答者提出一系列以“是”或“不是”为答案的问题,从图像中锁定目标物体。与 VisDial 不同的是,GuessWhat 中的问题均为封闭式问题,答案选项限制在“是”“否”“不适用”之中。GuessWhat 任务的目的导向性更强,而 VisDial 更接近实际应用场景。

3 研究现状分析

3.1 视觉问答算法

3.1.1 基于注意力机制的方法

注意力机制是视觉问答任务中的主流技术之一。基于注

意力机制的方法通过对问题或图像进行注意力加权，增强视觉与语言之间的交互，来对问题和图像的主体信息进行准确捕捉。

早期，最具代表性的方法为基于问题的图像注意力，包括SMem[14]和SAN[15]。其中，SMem计算每个单词的词向量与图像区域特征的相关性，得到基于问题的图像注意力。SAN使用CNN或LSTM计算整个问题的特征表示，并将注意力模块进行堆叠以实现多步迭代。上述方法的注意力机制均是单向的，即先计算语言特征，再计算视觉区域关于语言特征的注意力。为了更好地进行视觉语言交互，后续工作提出使用协同注意力的模式。例如，分层注意力模型(Hierarchical Question-Image Co-Attention，HieCoAtt)[16]并行地计算关于问题的图像注意力特征与关于图像的问题注意力特征，从而得到最终的视觉表示与文本表示。考虑到协同注意力机制在浅层网络中的成功应用，模块协同注意力网络(Modular Co-Attention Networks，MCAN)[17]进一步将协同注意力拓展到深层模型中。具体而言，MCAN借鉴了Transformer[18]模型的结构，利用自注意力单元对单词之间和区域之间的交互关系进行建模，并引导注意力单元对单词与区域之间的关系进行建模。MCAN进一步对上述两个注意力单元进行模块化组合，最终得到级联的模块协同注意力网络。双线性注意力网络(Bilinear Attention Networks，BAN)[19]则将低秩双线性池化(Low-rank Bilinear Pooling)应用到注意力计算上，以增强视觉与文本的注意力表示。

3.1.2 基于特征融合的方法

由于视觉语言领域的任务需要综合处理来自视觉和语言两个模态的输入，因此如何将多模态的特征表示进行融合也是视觉问答任务需要考虑的问题。常用的特征融合方法包括将视觉特征与语言特征进行拼接、按位相加和按位点乘等，处理方式较为简单，特征表示能力有提升的空间。研究者们考虑到基于双线性池化(Bilinear Pooling)的方法在细粒度图像分类中得到了成功应用，进而将双线性池化方法应用到视觉问答任务中。一些视觉问答的研究工作将重点放在如何更好地将视觉特征与语言特征的不同通道进行交互和融合。其经典方法包括多模态压缩双线性融合(Multimodal Compact Bilinear Pooling，MCB)[20]、多模态低秩双线性融合注意力网络(Multimodal Low-rank Bilinear Attention Networks，MLB)[21]、多模态因子分解双线性融合(Multi-modal Factorized Bilinear Pooling，MFB)[22]等。其中，MCB采用张量算法进行特征融合，但具有参数量过大、实用性低的缺点。针对这些缺点，MLB提出使用Hadamard product的方式对特征进行融合，从而降低输出特征的维数。然而，MLB对超参数较为敏感，且收敛速度慢。MFB进一步采用了矩阵分解的方式，并结合了协同注意力机制，兼具低维特征和强表达能力的优点，并可以将MLB看作其特例。此外，MUTAN框架[23]采用了Tucker分解的方法来减少参数，可以将MCB和MLB融入到框架中进行实现，具备良好的通用性。BLOCK[24]则基于块项张量分解(Block-term Tensor Decomposition)进行多模态特征融合，并将CP分解与Tucker分解进行泛化。

3.1.3 基于模块网络的方法

尽管深度神经网络在众多机器学习任务中取得了一系列的突破与进展，网络的可解释性一直是研究人员关注的问题。考虑到视觉问答任务可以被拆解为若干步骤，研究者们提出通过将神经网络模块化的方式，使神经网络具有序列化视觉推理的能力。其中，神经模块网络(Neural Modular Network，NMN)[25]是最具代表性的方法。NMN通过浅层神经网络将“寻找(find)”“转换(transform)”“组合(combine)”“计数(count)”等基本功能实现为若干子模块，并通过解析器得到问题中单词之间的语义联系。通过组合模块，NMN可以根据不同问题动态地生成不同的推理序列。在NMN的基础上，研究者们进一步对模块网络进行了改进。例如，N2NMN[26]利用了端到端的训练方式，通过布局策略(Layout Policy)同步学习如何将语言解析成语义结构及如何将语义结构组合成若干函数，从而使得NMN不再依赖于预先准备的解析器。上述NMN方法的训练均需要推理步骤的强监督信息。与上述工作相比，Stack-NMN[27]则无需推理步骤的监督信息。Stack-NMN采用栈存储每一步子模块的输出，并将离散的模块选择改为连续的模块选择，使得网络可以直接通过梯度下降进行训练。此外，PG+EE[28]和TbD[29]方法提出的程序生成器与布局策略相似，其执行模块也采用了神经模块网络。NS-VQA[30]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。NS-VQA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。XMN[31]则将图片解析为场景图，并提出了场景图上的元模块操作，在场景图上执行程序，大大增强了模型的可解释性。Prob-NMN[32]类似地考虑了将神经网络与符号计算相结合的方法，将NMN扩展为概率模型，实现了更具解释性的推理和更少执行程序的标注。MMN方法[33]针对NMN存在的可扩展性与通用性的瓶颈，借助元学习的思想，通过学习元函数的方式来解决具体的子任务，使每个实例模块实现参数共享，从而增强了可扩展性和通用性。上述方法在视觉推理数据集CLEVR[34]上取得了一系列性能上的突破。

3.1.4 基于视觉关系建模的方法

在视觉问答任务中，高阶的视觉语义信息对于回答涉及多个物体的复杂问题以及实现复杂的视觉推理具有重要意义。这一语义信息可以通过视觉关系(Visual Relation)与场景图(Scene Graph)来表示。视觉关系可以表示为“〈物体1，谓词，物体2〉”的三元组，刻画了两个视觉物体之间的关系属性。场景图可以看作视觉关系的集合，将所有视觉物体的两两之间的关系表示为图结构的形式。通过对视觉关系的刻画，视觉特征有了更丰富的语义信息。基于视觉关系和场景图的代表性方法包括多模态关系推理模型(Multimodal Relational Reasoning，MuRel)[35]和关系感知图注意力网络(Relation-Aware Graph Attention Network，ReGAT)[36]。具体而言，MuRel提出了一个可以迭代的单元模块，该单元模块采用双线性融合的方式得到每个视觉物体的多模态特征，并通过对空间特征和多模态特征进行建模，来得到视觉物体的关

系向量表示,并用关系特征表示更新每个物体的多模态特征表示。上述过程可以迭代多次,从而实现多步推理。ReGAT通过图注意力网络(Graph Attention Network)探索了显式和隐式的视觉关系。其中,显式关系通过预训练的分类器来识别,将物体之间的空间和语义联系表示为稀疏的图结构。隐式关系考虑了两两物体之间的隐式联系,并通过全连接图和注意力机制进行捕捉。此外,LCGN[37]通过迭代消息传播机制来提取视觉场景的上下文特征,丰富了图上每个节点(即视觉物体)的特征表达。

3.1.5 基于数据增强的方法

基于数据增强(Data Argumentation)的方法在众多计算机视觉和自然语言处理的任务中取得了显著的性能提升。视觉问答中的数据增强技术主要通过自动生成问题或问答对来实现。Kafle等的早期工作[38]提出了两种数据增强方式,分别为根据图像语义标注生成模板化问题和根据循环神经网络生成问题。Ray等[39]提出通过生成意图相似的问答对进行数据增强。Shah等[40]提出了循环一致性的问答生成策略,在训练过程中,根据答案还原出与之对应的原始问题,并根据生成的问题再次作答。

3.1.6 基于预训练的方法

预训练是计算机视觉和自然语言领域的重要技术。随着迁移学习、自监督学习、无监督学习等领域的发展,利用外源数据集进行有效的预训练成为了计算机视觉与自然语言处理领域的重要研究问题与技术手段。在大规模视觉数据集和文本数据集上,对基础模型进行预训练并将其迁移到下游任务上,模型的性能可以得到显著提升。早期,视觉问答模型的视觉分支采用VGG[41]或ResNet[42]结构,在大规模图像数据集ImageNet[43]上进行预训练。文本分支通常采用GloVe[44]对词表示进行初始化。

视觉分支的预训练经历了从网格特征到区域特征,再回到网格特征的探索过程。Anderson等[45]提出使用目标检测的模型作为视觉分支的基础模型,借助Visual Genome数据集[10]上的区域标注、类别标注和属性标注等对目标检测模型进行预训练。这一视觉特征在视觉问答、看图说话等一系列视觉语言任务上取得了明显的突破,在后续的研究方法中得到了广泛应用。最近,Jiang等[46]发现,通过对模型稍加修改,可以使网格特征得到比区域特征更好的性能,且计算速度得到了大幅提高。Jiang等[46]指出,预训练数据集Visual Genome中的目标及属性标注和图像的分辨率对预训练具有重要作用。

早期的视觉对话方法中,视觉模块与语言模块独立进行预训练,因此在预训练过程中,视觉语言之间的联系没有得到很好的捕捉。近年来,研究人员考虑到基于Transformer结构的预训练模型在自然语言处理领域取得了突破性的性能提升,进而将其应用到视觉语言领域。其中,具有代表性的工作包括ViLBERT[47],VLBERT[48],LXMERT[49],UNITER[50],OSCAR[51],12-in-1[52],VisualBERT[53]等,上述方法大多采用了BERT的结构。预训练任务包括掩码语言建模(Masked Language Modeling)和跨模态匹配(Cross-modality Matching)等。预训练数据集的选择包括Conceptual Captions[54],SBU captions[55],MS COCO,Visual Genome等。借助在大规模数据集上进行跨模态的自监督预训练任务,上述方法在包括视觉问答在内的多个下游任务中取得了最佳性能。

3.1.7 提升鲁棒性的方法

研究发现,视觉语言任务广泛受到了数据集偏差的影响,进而影响了模型的鲁棒性和泛化性。近期,视觉问答工作主要从改进语言分支、增强视觉注意力、减弱语言先验、增强训练数据4个方面展开。1)基于改进语言分支的方法提出,将文本概念进行分解[56]或生成视觉指引的问题特征表示[57]。2)基于增强视觉注意力的方法[58-59]探索如何通过使用额外的视觉[60]或文本[61]标注来增强训练。3)基于减弱语言先验的方法借助一个独立的语言分支,显式地对训练数据中的语言先验进行建模,并去除语言先验的影响[62-64]。4)基于增强训练数据的方法认为,语言偏差的来源是训练数据的不平衡分布,其通过生成额外的训练数据或对训练集进行重新划分的数据增强方式,让模型进行无偏训练[65-71]。

3.2 视觉对话算法

3.2.1 基于特征融合的基准方法

基准视觉对话算法为基于特征融合的方法,代表方法有后期融合(Late Fusion,LF)和层次递归编码(Hierarchical Recurrent Encoder,HRE)[6]。这两种方法的共同点是,并行独立地分别提取图像、问题和历史的特征,并将3个特征进行拼接或通过层次LSTM进行融合,最终将融合特征作为多模态特征。基于特征融合的方法的设计较为简单,没有深入考虑多源输入之间的交互联系。

3.2.2 基于注意力机制的方法

在视觉对话任务中,注意力机制体现在视觉和语言的细粒度交互。基于注意力机制的方法包括基于注意力的层次递归编码模型(Hierarchical Recurrent Encoder with Attention,HREA)[6]、记忆网络(Memory Network,MN)[6]、基于历史的图像注意力编码模型(History-Conditioned Image Attentive Encoder,HCIAE)[72]、联合注意力模型(Co-Attention,CoAtt)[73]和协同网络(Synergistic Network)[74]。其中,HREA在提取图像和问题的特征后,对对话历史中的每个单词进行注意力加权计算,提取对话历史的特征。MN根据图像和问题的特征,对每条对话历史进行注意力加权计算。HCIAE首先提取问题特征,然后根据问题对每个图像区域进行注意力加权计算,最后根据问题和图像对每条对话历史进行注意力加权计算。CoAtt采取序列化的方式,对于3个不同来源的输入,依次有序地根据其他2个输入计算第3个输入的注意力。Synergistic在CoAtt的基础上,将推理过程分为两个阶段,首先对候选答案进行初排序,然后根据第一阶段的排序结果进行优化。

3.2.3 基于视觉指代消解的方法

基于视觉指代消解的方法显式或隐式地对指代词进行消解,并将其与相应的视觉区域进行关联。具有代表性的基于视觉指代消解的方法有注意力记忆模型(Attention Memory,AMEM)[75]、神经模块网络(CorefNMN)[76]、递归视觉注意力模型(Recursive Visual Attention,RvA)[77]、对偶注意力网络(Dual Attention Networks,DAN)[78]和循环对偶注意力网络

(Recurrent Dual Attention Network, ReDAN)[79]。其中,AMEM 借助记忆网络,将每个根据历史问答对计算的视觉注意力进行记忆存储,并根据当前问题对存储的视觉注意力进行加权,在句子层面进行视觉指代消解。CorefNMN 结合了符号化计算与神经网络,将视觉推理过程分解为若干个基本操作,通过指代池将对话历史中出现的实体进行存储。在遇到指代词时,CorefNMN 通过查询模块将指代词与目标视觉物体相关联,从而实现了单词级别的视觉指代消解。RvA 采取了递归策略,在作答之前,首先判断当前问题是否表意清晰。如果问题表意不清,则回溯到与当前问题话题最匹配的问题,并递归地重复上述过程,直到问题表意明确,递归终止。通过递归回溯的过程,RvA 显式地在单词级别实现视觉指代消解。DAN 将视觉指代消解分为指代(REFER)和查询(FIND)两个步骤,并提出了相应的模块实现推理过程。ReDAN 将推理过程表示为多步推理过程,通过多步循环操作来优化视觉注意力与多模态特征。

3.2.4 基于图的方法

近年来,图结构和图神经网络在深度学习领域得到了广泛关注和快速发展。近期的视觉对话相关工作也将图像和文本等多源输入表示为图结构,并采用图神经网络提取多模态特征。基于图的代表性方法包括因子图注意力模型(Factor Graph Attention, FGA)[80]、图神经网络(GraphNeural Network, GNN)[81]、对偶视觉对话模型(DualVD)[82]等。具体而言,FGA 将图像、标题、历史问题和历史答案看作图上的实体,并对两两实体建立关系,提取细粒度的实体级交互特征。GNN 将多轮对话表示为图结构,依托图结构计算当前问答的特征表示,通过最大期望算法(EM 算法)并借助消息传递机制来更新图上节点的特征表示。DualVD 将图片表示为视觉场景图,对视觉物体之间的关系进行表达,提取物体-关系视觉特征。基于图的方法很好地刻画了视觉物体之间与视觉文本之间的关系,在视觉对话任务上取得了很好的性能。

3.2.5 基于预训练的方法

随着预训练方法在其他视觉语言任务上的成功应用,近期的视觉对话模型也开始应用于预训练,包括视觉对话 VisDial-BERT[83]和 VD-BERT[84]。这两种方法均采用 Transformer 作为基础结构,并采用掩码语言建模(Masked Language Modeling)作为预训练任务。两者在预训练数据集和模型上有两点主要差别。1) VisDial-BERT 在 Conceptual Captions[54]和 VQA[4]数据集上进行预训练,并在 VisDial 数据集上进行微调,而 VD-BERT 则不需要在上述两个大规模数据集上进行预训练。2) VD-BERT 支持判别式和生成式两种设定,而 VisDial-BERT 仅支持判别式设定。

3.2.6 生成式对话

上述方法主要聚焦在如何提取更好的多模态特征表示,即如何建立更有效的编码器。此外,生成式解码器也是视觉对话中的重要研究方向。生成式对话对于视觉对话的实际应用具有更重要的意义。最简单的生成式对话采用长短时记忆网络作为模型,通过最大似然估计进行训练。在此基础上,Lu 等[72]提出通过对抗学习的方式,使生成模型尽可能地生成无法与人类标注的标准答案相区分的答案。Wu 等[73]在应用对抗学习的同时,考虑了使用强化学习来训练生成模型,并将分类器对生成答案的置信度输出作为奖励函数。

4 未来研究方向展望

4.1 生成式视觉问答与对话系统

目前,相关工作多是将视觉问答与对话任务定义为分类问题或检索问题进行处理,即在预先给定候选答案集的情况下,从候选集中选取答案。这一设定也可以称为判别式设定。这种处理方式的优点是容易评测,通过准确率或召回率即可对模型的性能进行评价。这一评测方式更加强调编码器对视觉和文本信息的提取能力。然而,在实际问答与对话应用中,限定范围的答案集合难以覆盖所有的可能答案,从而难以满足实际场景中用户对问答与对话系统的需求。对于个性化的问答与对话系统而言,能够面对不同的问题和场景进行准确作答,是系统所需具备的能力。与判别式设定相比,生成式设定能产生更多样化的答案,也更容易实现对模型的迁移。生成式视觉问答与对话系统仍具有研究空间。

4.2 鲁棒的视觉问答系统

如前文所述,视觉问答系统容易受到数据集偏差的影响,尤其是语言偏差。这一偏差问题可以看作长尾分布问题。目前,提高视觉问答系统鲁棒性的研究能够在答案分布不一致的情况下提高模型的准确率。然而,这些研究在训练和测试分布一致的情况下,准确率会有所下降。因此,在保证当训练与测试分布一致时性能不变的情况下,如何提高模型在分布不一致情况下的鲁棒性,是未来视觉问答系统仍需解决的问题。近期,Niu 等[85]提出通过对因果关系建模和反事实推理的方法,来去除传统模型过于依赖相关关系的弊端。在其他计算机视觉领域中,基于因果关系的研究工作[86-91]也广泛体现出因果推理对鲁棒的人工智能系统的重要性。如何使模型通过探索因果关系进行鲁棒的推理,也将是未来的研究方向之一。

4.3 视觉对话数据集与评测

一方面,对于视觉对话任务而言,如何提高数据集的质量、建立更科学的评测任务,仍是一个开放性的研究问题。一些研究[92-93]认为,简单的对话模型在不关注视觉图像内容的情况下,仍可以在视觉对话任务中取得不错的性能。类似的结论在简单的问答模型不关注对话历史的情况下也可以得出。如何将视觉对话与对话任务和视觉问答任务区分开,使数据集能够更好地刻画视觉与对话两方面的特性,将成为视觉对话任务及其相关问题所要考虑的首要问题。

另一方面,如何对视觉对话任务进行评测仍是具有挑战性的问题。早期的视觉任务评测指标依赖于唯一的标注答案,采用了 Recall@k, MRR 等为代表的评测指标。这样的评价方式仅仅关注了单一标注,忽视了其他具有相同语义的正确答案,使得评测存在偏差。为了消除这一偏差,VisDial 官方数据集对小部分对话进行了额外的标注。具体而言,官方通过对每个答案与对话内容的相关性进行逐个打分,得到了密集注释(Dense Annotations),并根据密集注释使用 NDCG 评价指标进行评测。Qi 等[1]发现,通过在小规模的密集注释上进行微调,可以轻易地使众多基线模型在 NDCG 指标上得

到大幅提升，这反映了使用 NDCG 和密集注释进行评测仍存在不能很好评价模型性能的风险。此外，Massiceti 等[94]指出，视觉对话的排序式的评价指标与视觉对话任务的目的不完全匹配，并提出使用 CIDEr，METEOR 等自然语言处理领域常用的生成式评价指标进行评测，通过半监督自动标注的方式和典型相关分析（Canonical Correlation Analysis）方法在整个数据集上建立了答案标注集。综上，如何构建数据集和定义评价指标仍是视觉对话任务中有待继续研究的问题。

4.4 视觉推理

视觉推理是视觉语言交互系统所需具备的重要能力。未来视觉问答中的视觉推理研究可以概括为两个方面。1）模型方面。随着注意力机制在自然语言处理领域的成功应用，视觉语言系统的视觉推理能力也可以通过注意力机制来体现。而符号计算和神经网络的融合，也促进了离散化视觉推理的发展。随着大规模预训练和基于 Transformer 结构系统的兴起，视觉问答系统的性能有了进一步的提升。如何更好地将性能与推理能力相结合，尤其是将符号计算与神经网络相结合，使视觉问答系统兼具高性能与可解释性，仍是未来需要关注的问题。2）任务与数据集方面。传统视觉问答任务的目标相对简单，近期一些研究工作从不同角度对视觉问答任务进行了拓展。表 2 列出了近期视觉问答延伸的数据集与任务。例如，TextVQA[95]和 ST-VQA[96]研究如何在视觉场景中对文本内容进行理解与推理；VizWiz[97]根据真实的盲人用户体验收集数据，其图像和问答更接近实际应用场景。另外，一些工作对视觉推理的复杂度进行研究，如 GQA 数据集[98]在收集过程中利用了场景图的结构化信息，在对真实场景图片生成组合式问题的同时，尽可能消除了语言偏差的影响；视觉常识推理（Visual Commonsense Reasoning，VCR）任务[99]要求模型在进行视觉问答的同时，给出作答的依据和理由。对视觉问答中推理能力的深层考察，将成为未来研究的重点之一。

表 2 视觉问答拓展任务数据集

Table 2 Beyond-VQA datasets

Dataset	Goal	Number of images	Number of questions
TextVQA	To recognize text in visual scenes	28408	45336
ST-VQA	To recognize text in visual scenes	23038	31791
VizWiz	To collect data from blind people	20523	20523
GQA	To conduct visual reasoning in real visual scenes	113018	22669678
VCR	To provide reasons when answering the questions	99904	264720

结束语 随着计算机视觉和自然语言处理领域的发展，越来越多的研究人员对视觉与语言交叉问题开展研究。其中，视觉问答与对话是视觉语言交叉领域的代表性任务，对机器的多模态理解与推理能力具有较高的要求，也是实现满足人机交互要求的视觉智能系统的重要技术。

本文阐述了视觉问答与对话的概念与应用，介绍了视觉问答与对话的相关数据集，重点梳理了视觉问答与对话算法在多模态理解与推理方面的研究现状，并对未来的研究方向进行了展望。

参 考 文 献

[1] YU J，WANG L，YU Z. Research on Visual Question Answering Techniques[J]. Journal of Computer Research and Development，2018，55(9)：1946-1958.

[2] QI J，NIU Y，HUANG J，et al. Two causal principles for improving visual dialog[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10860-10869.

[3] 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.

[4] 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 Recognition. 2017：6904-6913.

[5] LIN T Y，MAIRE M，BELONGIE S，et al. Microsoft coco：Common objects in context[C]// European Conference on Computer Vision. Springer，Cham，2014：740-755.

[6] DAS A，KOTTUR S，GUPTA K，et al. Visual dialog[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：326-335.

[7] MALINOWSKI M，FRITZ M. A Multi-world Approach to Question Answering about Real-world Scenes based on Uncertain Input[C]// Twenty-Eighth Annual Conference on Neural Information Processing Systems. Curran，2014：1682-1690.

[8] REN M，KIROS R，ZEMEL R S. Exploring models and data for image question answering[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015：2953-2961.

[9] GAO H，MAO J，ZHOU J，et al. Are you talking to a machine? Dataset and methods for multilingual image question answering[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015：2296-2304.

[10] 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.

[11] 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. 2016：4995-5004.

[12] YU L，PARK E，BERG A C，et al. Visual madlibs：Fill in the blank image generation and question answering[J]. arXiv：1506.00278，2015.

[13] DE VRIES H，STRUB F，CHANDAR S，et al. Guesswhat?! visual object discovery through multi-modal dialogue[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：5503-5512.

[14] XU H，SAENKO K. Ask，attend and answer：Exploring question-guided spatial attention for visual question answering[C]// European Conference on Computer Vision. Springer，Cham，2016：451-466.

[15] 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.

[16] 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. 2016: 289-297.

[17] YU Z, YU J, CUI Y, et al. Deep modular co-attention networks for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6281-6290.

[18] 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. 2017: 6000-6010.

[19] KIM J H, JUN J, ZHANG B T. Bilinear attention networks [C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 1571-1581.

[20] FUKUI A, PARK D H, YANG D, et al. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016: 457-468.

[21] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[J]. arXiv: 1610. 04325, 2016.

[22] YU Z, YU J, FAN J, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 1821-1830.

[23] 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. 2017: 2612-2620.

[24] BEN-YOUNES H, CADENE R, THOME N, et al. Block: Bilinear superdiagonal fusion for vi-sual question answering and visual relationship detection [C] // Proceedings of the AAAI Confe-rence on Artificial Intelligence. 2019, 33: 8102-8109.

[25] ANDREAS J, ROHRBACH M, DARRELL T, et al. Neural module networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 39-48.

[26] HU R, ANDREAS J, ROHRBACH M, et al. Learning to reason: End-to-end module networks for visual question answering [C] // Proceedings of the IEEE International Conference on Computer Vision. 2017: 804-813.

[27] HU R, ANDREAS J, DARRELL T, et al. Explainable neural computation via stack neural module networks [C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 53-69.

[28] JOHNSON J, HARIHARAN B, VAN DER MAATEN L, et al. Inferring and executing programs for visual reasoning[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2989-2998.

[29] MASCHARKA D, TRAN P, SOKLASKI R, et al. Transparency by design: Closing the gap between performance and interpretability in visual reasoning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4942-4950.

[30] 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. 2018: 1039-1050.

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

[32] VEDANTAM R, DESAI K, LEE S, et al. Probabilistic Neural Symbolic Models for Interpretable Visual Question Answering [C] // International Conference on Machine Learning. 2019: 6428-6437.

[33] CHEN W, GAN Z, LI L, et al. Meta module network for compositional visual reasoning[J]. arXiv: 1910. 03230, 2019.

[34] 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. 2017: 2901-2910.

[35] CADENE R, BEN-YOUNES H, CORD M, et al. Murel: Multimodal relational reasoning for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1989-1998.

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

[37] HU R, ROHRBACH A, DARRELL T, et al. Language-conditioned graph networks for relational reasoning[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 10294-10303.

[38] KAFLE K, YOUSEFHUSSIEN M, KANAN C. Data augmentation for visual question answering[C]// Proceedings of the 10th International Conference on Natural Language Generation. 2017: 198-202.

[39] RAY A, SIKKA K, DIVAKARAN A, et al. Sunny and Dark Outside?! Improving Answer Consistency in VQA through Entailed Question Generation[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). 2019: 5863-5868.

[40] SHAH M, CHEN X, ROHRBACH M, et al. Cycle-consistency for robust visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6649-6658.

[41] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.

[42] 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.

[43] 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.

[44] PENNINGTON J，SOCHER R，MANNING C D. Glove：Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2014：1532-1543.

[45] 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.

[46] JIANG H，MISRA I，ROHRBACH M，et al. In Defense of Grid Features for Visual Question Answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10267-10276.

[47] LU J，BATRA D，PARIKH D，et al. Vilbert：Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]// Advances in Neural Information Processing Systems. 2019：13-23.

[48] SU W，ZHU X，CAO Y，et al. V-bert：Pre-training of generic visual-linguistic representations[J]. arXiv：1908. 08530，2019.

[49] TAN H，BANSAL M. LXMERT：Learning Cross-Modality Encoder Representations from Transformers[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019：5103-5114.

[50] CHEN Y C，LI L，YU L，et al. Uniter：Learning universal image-text representations[J]. arXiv：1909. 11740，2019.

[51] LI X，YIN X，LI C，et al. Oscar：Object-semantics aligned pre-training for vision-language tasks[C]// European Conference on Computer Vision. Springer，Cham，2020：121-137.

[52] LU J，GOSWAMI V，ROHRBACH M，et al. 12-in-1：Multi-task vision and language representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10437-10446.

[53] LI L H，YATSKAR M，YIN D，et al. Visualbert：A simple and performant baseline for vision and language[J]. arXiv：1908. 03557，2019.

[54] SHARMA P，DING N，GOODMAN S，et al. Conceptual captions：A cleaned，hypernymed，image alt-text dataset for automatic image captioning[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018：2556-2565.

[55] ORDONEZ V，KULKARNI G，BERG T L. Im2Text：describing images using 1 million captioned photographs[C]// Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011：1143-1151.

[56] JING C，WU Y，ZHANG X，et al. Overcoming Language Priors in VQA via Decomposed Linguistic Representations[C]// AAAI. 2020：11181-11188.

[57] KV G，MITTAL A. Reducing Language Biases in Visual Question Answering with Visually-Grounded Question Encoder[J]. arXiv：2007. 06198，2020.

[58] 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 International Conference on Computer Vision. 2019：2591-2600.

[59] WU J，MOONEY R. Self-critical reasoning for robust visual question answering[C]// Advances in Neural Information Processing Systems. 2019：8604-8614.

[60] DAS A，AGRAWAL H，ZITNICK L，et al. Human Attention in Visual Question Answering：Do Humans and Deep Networks look at the same regions? [C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016：932-937.

[61] HUK PARK D，ANNE HENDRICKS L，AKATA Z，et al. Multimodal explanations：Justifying decisions and pointing to the evidence[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：8779-8788.

[62] RAMAKRISHNAN S，AGRAWAL A，LEE S. Overcoming language priors in visual question answering with adversarial regularization[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018：1548-1558.

[63] CADENE R，DANCETTE C，CORD M，et al. Rubi：Reducing unimodal biases for visual question answering[C]// Advances in neural information processing systems. 2019：841-852.

[64] CLARK C，YATSKAR M，ZETTLEMOYER L. Don't Take the Easy Way Out：Ensemble Based Methods for Avoiding Known Dataset Biases[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). 2019：4060-4073.

[65] ABBASNEJAD E，TENEY D，PARVANEH A，et al. Counterfactual vision and language learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：10044-10054.

[66] TENEY D，ABBASNEJAD E，HENGEL A. Unshuffling data for improved generalization[J]. arXiv：2002. 11894，2020.

[67] TENEY D，KAFLE K，SHRESTHA R，et al. On the Value of Out-of-Distribution Testing：An Example of Goodhart's Law [J]. arXiv：2005. 09241，2020.

[68] ZHU X，MAO Z，LIU C，et al. Overcoming Language Priors with Self-supervised Learning for Visual Question Answering [J]. arXiv：2012. 11528，2020.

[69] 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.

[70] LIANG Z，JIANG W，HU H，et al. Learning to Contrast the Counterfactual Samples for Robust Visual Question Answering [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020：3285-3292.

[71] GOKHALE T，BANERJEE P，BARAL C，et al. MUTANT：A Training Paradigm for Out-of-Distribution Generalization in Visual Question Answering[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020：878-892.

[72] LU J, KANNAN A, YANG J, et al. Best of both worlds: transferring knowledge from discriminative learning to a generative visual dialog model[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 313-323.

[73] WU Q, WANG P, SHEN C, et al. Are you talking to me? reasoned visual dialog generation through adversarial learning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6106-6115.

[74] GUO D, XU C, TAO D. Image-question-answer synergistic network for visual dialog[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10434-10443.

[75] SEO P H, LEHRMANN A, HAN B, et al. Visual reference resolution using attention memory for visual dialog [C] // Advances in Neural Information Processing Systems. 2017: 3719-3729.

[76] KOTTUR S, MOURA J M F, PARIKH D, et al. Visual coreference resolution in visual dialog using neural module networks [C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 153-169.

[77] NIU Y, ZHANG H, ZHANG M, et al. Recursive visual attention in visual dialog[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6679-6688.

[78] KANG G C, LIM J, ZHANG B T. Dual Attention Networks for Visual Reference Resolution in Visual Dialog[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). 2019: 2024-2033.

[79] GAN Z, CHENG Y, KHOLY A, et al. Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 6463-6474.

[80] SCHWARTZ I, YU S, HAZAN T, et al. Factor graph attention [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2039-2048.

[81] ZHENG Z, WANG W, QI S, et al. Reasoning visual dialogs with structural and partial observations [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6669-6678.

[82] JIANG X, YU J, QIN Z, et al. DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue [C]// AAAI. 2020, 1(3): 5.

[83] MURAHARI V, BATRA D, PARIKH D, et al. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline[J]. arXiv: 1912. 02379, 2019.

[84] WANG Y, JOTY S, LYU M R, et al. Vd-bert: A unified vision and dialog transformer with bert[J]. arXiv: 2004. 13278, 2020.

[85] NIU Y, TANG K, ZHANG H, et al. Counterfactual VQA: A Cause-Effect Look at Language Bias [J]. arXiv: 2006. 04315, 2020.

[86] TANG K, NIU Y, HUANG J, et al. Unbiased scene graph generation from biased training[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3716-3725.

[87] YANG X, ZHANG H, CAI J. Deconfounded image captioning: A causal retrospect[J]. arXiv: 2003. 03923, 2020.

[88] 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. 2020: 10760-10770.

[89] TANG K, HUANG J, ZHANG H. Long-tailed classification by keeping the good and removing the bad momentum causal effect [J]. arXiv: 2009. 12991, 2020.

[90] YUE Z, ZHANG H, SUN Q, et al. Interventional few-shot learning[J]. arXiv: 2009. 13000, 2020.

[91] ZHANG D, ZHANG H, TANG J, et al. Causal intervention for weakly-supervised semantic segmentation [J]. arXiv: 2009. 12547, 2020.

[92] MASSICETI D, DOKANIA P K, SIDDHARTH N, et al. Visual dialogue without vision or dialogue [J]. arXiv: 1812. 06417, 2018.

[93] AGARWAL S, BUI T, LEE J Y, et al. History for Visual Dialog: Do we really need it? [J]. arXiv: 2005. 07493, 2020.

[94] MASSICETI D, KULHARIA V, DOKANIA P K, et al. A Revised Generative Evaluation of Visual Dialogue[J]. arXiv: 2004. 09272, 2020.

[95] SINGH A, NATARAJAN V, SHAH M, et al. Towards vqa models that can read[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8317-8326.

[96] BITEN A F, TITO R, MAFLA A, et al. Scene text visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 4291-4301.

[97] GURARI D, LI Q, STANGL A J, et al. Vizwiz grand challenge: Answering visual questions from blind people[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3608-3617.

[98] HUDSON D A, MANNING C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6700-6709.

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

NIU Yu-lei, born in 1992, Ph. D. His main research interests include vision-language and visual reasoning.