计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100176

基于视觉和语言的跨媒体问答与推理研究综述

武阿明　姜　品　韩亚洪
天津大学智能与计算学部　天津 300350
(tjwam@tju.edu.cn)

摘　要　基于视觉和语言的跨媒体问答与推理是人工智能领域的研究热点之一,其目的是基于给定的视觉内容和相关问题,模型能够返回正确的答案。随着深度学习的飞速发展及其在计算机视觉和自然语言处理领域的广泛应用,基于视觉和语言的跨媒体问答与推理也取得了较快的发展。文中首先系统地梳理了当前基于视觉和语言的跨媒体问答与推理的相关工作,具体介绍了基于图像的视觉问答与推理、基于视频的视觉问答与推理以及基于视觉常识推理模型与算法的研究进展,并将基于图像的视觉问答与推理细分为基于多模态融合、基于注意力机制和基于推理3类,将基于视觉常识推理细分为基于推理和基于预训练2类;然后总结了目前常用的问答与推理数据集,以及代表性的问答与推理模型在这些数据集上的实验结果;最后展望了基于视觉和语言的跨媒体问答与推理的未来发展方向。

关键词:跨媒体问答与推理;图像问答与推理;视频问答与推理;视觉常识问答与推理;多模态融合;注意力机制;预训练
中图法分类号　TP391

Survey of Cross-media Question Answering and Reasoning Based on Vision and Language

WU A-ming,JIANG Pin and HAN Ya-hong
College of Intelligence and Computing,Tianjin University,Tianjin 300350,China

Abstract　Cross-media question answering and reasoning based on vision and language is one of the popular research hotspots of artificial intelligence. It aims to return a correct answer based on understanding of the given visual content and related questions. With the rapid development of deep learning and its wide application in computer vision and natural language processing,cross-media question answering and reasoning based on vision and language has also achieved rapid development. This paper systematically surveys the current researches on cross-media question answering and reasoning based on vision and language,and specifically introduces the research progress of image-based visual question answe-ring and reasoning,video-based visual question answering and reasoning,and visual commonsense reasoning. Particularly,image-based visual question answering and reasoning is subdivided into three categories,i.e.,multi-modal fusion,attention mechanism,and reasoning based methods. Meanwhile,visual commonsense reasoning is subdivided into reasoning and pre-training based methods. Moreover,this paper summarizes the commonly used datasets of question answering and reasoning,as well as the experimental results of representative methods. Finally,this paper looks forward to the future development direction of cross-media question answering and reasoning based on vision and language.

Keywords　Cross-media question answering and reasoning,Image-based question answering and reasoning,Video-based question answering and reasoning,Visual commonsense question answering and reasoning,Multi-modal fusion,Attention mechanism,Pre-training

1　引言

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,不同媒介的信息覆盖了人类生活的方方面面,如气味、语音、文本、图像、视频等,每一种模态都承载了不同的信息。跨媒体的信息交互往往能够传递更为丰富的信息,人类的生活也因为多种视听媒介信息的交互而变得绚丽多彩,其中尤其以视觉和语言的交互内容为主,如图像配以标题或文章、视频配以字幕等。随着计算机处理能力的提升和人工智能技术的进步,基于计算机视觉和自然语言处理的研究分

到稿日期:2020-10-25　返修日期:2021-01-01
基金项目:国家自然科学基金"重点项目"(61932009):跨媒体智能问答与推理关键理论与方法研究(2020/01-2024/12)
This work was supported by the National Natural Science Foundation of China Key Program (61932009):Research on Key Theories and Methods of Cross-media Intelligent Question Answering and Reasoning(2020/01-2024/12).
通信作者:韩亚洪(yahong@tju.edu.cn)

别帮助计算机学会了如何去“看”和如何去“读”。由于视觉内容细腻丰富，语言内容精炼准确，二者表达信息的能力各不相同，存在一定的模态差异，因此研究基于视觉和语言的跨媒体交互成为跨媒体智能的一个重要研究方向。

智能问答最早可追溯到人工智能诞生的时期。阿兰·图灵于1950年提出了著名的“图灵测试”，通过测试机器是否具备正确回答问题的能力，来验证机器是否具有人类智能[1]。基于视觉和语言的跨媒体问答与推理是智能问答系统的扩展，要求问答系统在回答问题的同时考虑视觉信息和语言信息，然后推理出正确答案。根据视觉媒介的不同，通常可分为基于图像的问答与推理[2-3]、基于视频的问答与推理[4]和基于视觉常识的问答与推理[5]3种常见的任务。

基于视觉和语言的问答与推理作为一个极具挑战性的研究方向，涉及了计算机视觉技术、自然语言处理技术以及视觉和语言的融合技术。对于基本的问答系统[6]，通常包括了对于视觉和语言两种模态的表征、跨模态融合和问答推理3个步骤。得益于深度学习的快速发展，单一模态表征技术不断更新完善[7-8]，跨媒体问答任务的核心挑战在于跨模态融合和问答推理两个部分。特别地，为了促进多模态特征间元素的充分交互，本文提出了一种双线性融合的方法[9]来得到视觉-语言的联合表达。相比常用的融合算法，即对应元素相加、对应元素相乘和特征拼接等算法，所提方法能够获得充分包含各模态信息的融合表示，从而提升了视觉问答任务的性能。另外，注意力机制[10-11]是一种常用的解决视觉问答的推理方法。通过捕捉与答案相关的视觉及语言信息，该方法提升了视觉问答的性能并提高了其可解释性。除了基于自然图像和视频的视觉问答与推理，目前已经开展了对于医疗图像问答系统[12]及视觉对话系统[13-14]的研究。

本文系统地梳理了当前基于视觉和语言的跨媒体问答与推理的相关工作，具体介绍了基于图像的视觉问答与推理、基于视频的视觉问答与推理以及基于视觉常识推理模型与算法的研究进展，同时总结了常用的视觉问答与推理的数据集，并给出了代表性的方法在这些数据集上的实验结果。最后，本文展望了基于视觉和语言的跨媒体问答与推理的未来发展方向。

2 图像的视觉问答与推理

如图1所示，给定一张图像和一个相关的问题(通常使用文本表示问题)，视觉问答(Visual Question Answering，VQA)任务要求模型给出正确的答案。通常情况下，基于一组给定的候选答案选项，模型需要从中选出正确的选项[15]。相比传统的计算机视觉问题，如图像识别、检测和分割，模型不仅需要理解图片以及对应问题的内容，还需要将视觉与问题内容进行准确的关联才能正确地回答问题。从这个意义上讲，该任务更能反映模型理解视觉内容的能力。近年来，随着深度学习的发展，视觉问题问答取得了很大的进展。特别地，由于视觉和语言属于不同的模态，第一类方法通过设计有效的融合方法来得到视觉和语言的联合表达，从而任务的性能；第二类方法通过借助注意力机制来捕捉与问题相关的视觉信息，从而使视觉与语言内容更加一致；第三类方法基于提取的视觉和语言表示，设计一种有效的推理方法来推出准确的答案选项。下文对这3类方法中典型的算法模型及研究进展进行介绍。

Question: What is the mustache made of?

Answer: Banana

图1 视觉问答示例

Fig. 1 Example of visual question answering

2.1 多模态融合的视觉问答方法

多模态融合的目的是得到一个能够尽可能包含各种模态信息的紧凑表示。对于视觉问答任务，早期的工作通常基于一阶交互的思想来促进视觉-语言两种模态信息的融合。Ren等[16]首次使用拼接的方法来得到视觉特征与问题表示的融合表达，把问题中所有单词嵌入特征的加和结果作为问题表示。Shih等[17]提出一个注意力机制的框架来求取包含视觉表示、问题表示和答案表示三元组的得分。基于视觉特征与文本特征的相似性，模型会输出一个得分，然后基于这个得分对多模态特征进行融合。与Shih等的工作相似，Lu等[10]提出一种相互注意力的方法，即分别求取基于视觉特征的文本注意力表示和基于文本特征的视觉注意力表示。最后对视觉注意力表示和文本注意力表示进行求和，从而得到最终的融合表示。

相比一阶交互融合的算法，用于建模视觉和文本表示的二阶交互是一种更有效的获取融合表示的方法。其中，文献[18-19]采用对应元素乘积的方式得到视觉与文本的融合表示。为了进一步促进多模态表示间的交互，Fukui等[9]提出了一种紧凑的多模态双线性融合方法(Multimodal Compact Bilinear pooling，MCB)。具体地，基于视觉和文本表示首先进行外积运算，然后通过一个非线性变换[20]，将外积运算的结果映射到一个低纬度的空间。尽管这种方法能够获得较好的融合表示，但通常只针对高维度的特征具有明显效果。为此，Kim等[21]提出了一种多模态低秩双线性融合的方法(Multimodal Low-rank Bilinear pooling，MLB)，通过分别使用一个低秩张量来表示视觉与文本表示，这种方法能够获得比MCB更加有效的融合表示且计算量更小。在前述研究工作的基础上，Ben等[22]提出了一种新的基于双线性交互的多模态融合方法，即MUTAN。具体地，通过控制模型的参数数量，MUTAN可以减小特征嵌入矩阵的尺寸；然后，基于双线性交互的机制获取准确的多模态融合表示。同样地，为了弥补MCB需要高维度特征的缺陷，Yu等[23]提出了一种可分解的双线性融合方法来获取基于视觉和语言的有效融合表示，并在视觉问答任务上证明了该方法的有效性。最后，Ben

等[24]提出了一种基于块超对角线张量分解的融合算法，进一步提升了视觉-语言融合的效果以及视觉问答任务的性能。

2.2 基于注意力机制的视觉问答方法

注意力机制[25]已经成为多模态系统的一种常用操作，使用注意力机制通常可以提升多模态任务的性能。Yang 等[26]堆叠了多个问题引导的注意力机制，可以有效地捕捉与问题相关的视觉信息并提升视觉问答的性能。Li 等[27]首先在图片上提取多个物体框，然后基于文本表示选出与文本表示相关的物体框。通过这种方式可以有效地捕捉到与问题相关的视觉信息。受这一工作的启发，Anderson 等[28]首先使用目标检测网络 Faster R-CNN[29]在图片上提取多个物体框，然后使用一个自下而上和自上而下的注意力机制来提取与问题相关的物体框，从而得到有效的视觉表示。为了让模型具有认知能力，Schwartz 等[30]提出了一种用于学习视觉-语言高阶相关性的注意力机制，通过这种机制可以有效地学到与文本内容对应的视觉内容，从而提升视觉问答任务的性能。此外，考虑到现有的注意力机制通常只能提取某个元素的相关表示，Li 等[31]提出一种区域注意力机制来捕捉与某个元素的区域相关的表示。Patro 等[32]通过一个或多个支持和反对示例来取得一个微分注意力区域。使用这种方法可以获得更加接近人类注意力的区域，从而提升视觉问答任务的性能。最后，考虑到现有的视觉问答模型很少从候选答案中捕获重要的信息，Guo 等[33]提出了一种从候选答案中捕获相关信息的注意力机制，进一步提升了视觉问答任务的性能。

2.3 视觉问答与推理

推理是操纵先前获得的知识以得出新颖的推理或提升回答新问题的能力，是智能思维的基本组成部分之一。前述方法本质上是获得有效的视觉-语言表示，并没有使模型具有推理能力。为此，最近提出的一些方法基于提取的视觉-语言表示，设计有效的推理方法来推出问题的答案。

2.3.1 基于模块化的推理

Andreas 等[34]设计一个由多个模块组成的推理网络。具体地，首先使用一个自然语言解析树模型对问题进行解析，从而得到不同的句子成分；然后，利用一种注意力机制来选取重要的部分进行答案的推理。尽管这个工作提升了视觉问答任务的性能，但这种方法需要自然语言解析树来生成问题的层次结构表示。为了解决这个问题，Hu 等[35]提出了一个端到端训练的模块网络。无需借助自然语言解析树，这个网络可以通过直接预测实例特定的网络布局来学习推理。Hudson 等[36]提出了一种用于推理的组合注意力网络。该网络通过将问题分解为一系列基于注意力的步骤来推导出最终的答案，且每个步骤均由包含记忆、注意力和整合 3 个部分的单元执行推理。在多个视觉问答数据集上，该网络均能获得优异的推理表现。

2.3.2 基于实体关联的推理

上文介绍的推理工作都是试图把模型分解成多个不同的模块，每个模块完成不同的子任务，最终实现推理。接下来介绍基于实体关联的推理模型，这种模型主要基于图片中各个实体间的关系来推导正确的答案。具体地，考虑到当前主要的方法通过捕捉单个视觉区域与单词之间的关系来促进推理，这样可能会不能准确地回答问题。为此，Gao 等[37]提出了一种多层交互的模型，通过使用一个分层的视觉-语言交互模型，可以更加有效地捕捉到视觉内容与对应文本内容之间的关系，从而进一步提升视觉问答任务的性能。此外，对于视觉问答任务，当前的注意力机制相当于在给定问题的前提下，对每个图像区域打分后做信息加权。由于这种方式忽略了图像区域间的空间与语义间的关联，因此不能有效地推理。为此，Cadene 等[38]提出了一个多模态关系推理模型。该模型主要由双线性融合模块以及关系建模模块组成，挖掘问题和图像区域间的细粒度关联，输出每个区域感知的上下文信息。Gao 等[39]首次在视觉问答任务中同时考虑了模态内部关系和跨模态关系，其中模态内部关系是跨模态关系的补充。具体地，每个图像区域不仅与问题是相关的，而且与其他图像区域也是关联的。因此，这个模型充分融合了模态内部关系与跨模态关系，从而进一步提升了视觉问答任务的性能。

2.3.3 基于图网络的推理

除了上文介绍的基于实体关联的方法外，最近，随着图神经网络的发展，一些工作开始尝试借助图网络[40-42]进行推理。由于图网络可以有效地捕捉局部以及全局的关系，因此图网络适用于视觉问答任务。具体地，Teney 等[43]在视觉问答任务上首次尝试利用图网络来捕捉场景与问题的结构化表示。受该思想的启发，Brown 等[44]提出一个用于学习条件化图结构表示的模型，该模型首先使用物体检测器在图片上提取多个物体框，然后以问题为条件动态地捕捉与问题相关的物体框之间的关系。由于这种方法不仅可以捕捉到与问题相关的物体，还可以捕捉到相关物体间的关系，因此这种方法提升了视觉问答任务的性能及可解释性。类似地，Hu 等[45]也使用图网络来捕捉物体间以及物体与语言间的关系。Khademi 等[46]提出了一个图记忆网络[47]模型，该模型首先使用图网络分别提取文本与视觉内容的结构化表示，然后使用一个序列模型来进一步整合结构化的表示，最后在多个视觉问答任务上进行实验，证明了该方法的有效性。Hudson 等[48]提出了一个神经状态机推理模型，该模型首先预测一个表示潜在语义关系的概率图，然后基于这个概率图进行序列化的推理，从而得出正确的答案，最后在多个视觉问答数据集上进行实验，证明了该方法的有效性。

3 基于视频的视觉问答与推理

基于图像的问答与推理通常可以回答“What”“Where”和“How many”等单一图像能传达的问答信息。对于更高级别的语义内容，如动作转换和动作意图推断等，视频的时序特性往往能提供更丰富的信息。因此，将基于视频的视觉问答和推理[49-50]的任务（见图 2）作为图像任务的延伸，得到了广泛的关注。由于视频包含多个连续帧的视觉信息，因此如何有效地存储这些信息并找到与其对应的答案成为了解决问题的关键。目前，主流的解决方案主要包括基于记忆网络的方法和基于协同注意力的方法。接下来将详细介绍这两种方法。

Question: What does Pat do with the money he got from Joey?

Answer: He buys a guitar

图 2 视频问答示例

Fig. 2 Example of video question answering

3.1 基于记忆网络的方法

视频问答涉及的视频序列通常较长，为了正确回答与长视频相关的问题，模型需要具备对长视频的感知能力和记忆能力。传统的基于循环神经网络的模型[51]通常使用隐藏层单元和注意力机制作为记忆手段，然而这种模型存储记忆的能力有限。因此，一些研究工作使用了具备可读写的外部记忆模块的记忆网络[47]。特别地，Tapaswi 等[52]构建了一个基于电影视频的问答数据集，同时使用一个端到端训练的记忆网络模型来分割电影的帧，并创建多个用于存储视频和字幕特征的记忆集合，最后该模型可以从记忆集合中选出与答案相关的信息来推导正确的答案。Gao 等[53]提出了一种用于视频问答的运动特征和外观特征协同网络，通过有效地捕捉视频帧间的信息，这种方法进一步提升了视频问答的性能。Kim 等[54]提出一种渐进式注意力记忆网络用于电影视频问答。借助渐进式注意力机制，可以利用问题和答案中的线索逐步删减内存中无关的时间片段，从而提升任务的性能。

3.2 基于协同注意力的方法

协同注意力机制是一种对称的注意力机制，即同时捕获文本相关的视频信息与视频相关的文本信息，然后进一步融合这两种信息来提升视频问答的性能。特别地，Li 等[55]提出了一个位置自注意模块，其通过自注意力机制来更新序列中每个位置的响应，可以有效地捕捉相关的时序信息，从而提升任务的性能。此外，Kim 等[56]提出了一个模态转移注意网络。这个网络将视频问答任务分解为两个子任务，即定位与问题相关的时间段和基于局部时间段进行答案预测。通过这种方法可以更加有效地捕捉到相关的视频信息，从而提升任务的性能。

4 视觉常识的推理

最近有关视觉理解的进展主要是关于视觉内容的感知（如目标检测和语义分割）或者基于图像区域的视觉概念理解（如视频描述生成[57-59]和视觉问答[60-61]）。为了实现全面的视觉理解，模型必须从感知转向推理，其中包括具有场景相关细节和相关常识的认知推理。作为实现全面视觉理解的关键步骤，Rowan 等[5]提出了一个视觉常识推理任务以及一个新的数据集（见图 3）。对于这个任务，给定一张图片，模型不仅需要回答有关视觉内容的问题，而且还需要提供一个解释以说明选择这个答案的原因。相比前文介绍的视觉问答任务，视觉常识推理的问题与答案都比视觉问答任务更复杂，仅仅基于图片内容很难选出正确的答案及对应的正确解释。最近研究人员提出了一些解决视觉常识推理任务的方法，这些方法可以归结为两类：基于推理的方法和基于预训练的方法。下面将详细介绍这两类方法。

Question: Why is [Person4] pointing at [Person1]?

Answer: He is telling [Person3] that [Person1] ordered the pancakes.

Rationale: [Person3] is delivering food to the table, and she might not know whose order is whose.

图 3 视觉常识推理示例

Fig. 3 Example of visual commonsense reasoning

4.1 基于推理的方法

基于推理的方法本质上是设计一个推理模型来有效地整合视觉区域特征间的关系以及视觉区域与文本特征之间的关系，从而推出正确的答案及其对应的解释。特别地，作为缩小感知和认知层面视觉理解差距的首次尝试，Zellers 等[5]提出了一个识别-认知网络。这个模型首先把自然语言的语义表示与相应参考物体的语义表示进行关联；然后，基于答案的语义表示，模型分别对问题以及视觉表示进行处理；最后，模型在之前处理的基础上进行推理以得到正确的答案。然而，由于视觉常识推理与人类大脑的认知方式存在巨大差异，识别-认知网络并不能达到人类的认知高度。受关于大脑认知思想的启发，Wu 等[58]提出了一个面向视觉常识推理的有向视觉连接网络，其主要过程是基于当前推理任务中的问题与候选答案的语义表示，动态地整合视觉神经元的连接。具体地，连接网络包括视觉神经元连接、情景化连接以及面向推理的有向连接。其中，对于视觉神经元连接，Wu 等[62]提出了一个条件化的 GraphVLAD 模块，其可以根据视觉和问题的内容动态地捕捉高层次视觉语义信息，从而提升视觉理解的水平。在视觉常识推理数据集上，该方法的性能明显优于识别-认知网络模型。同样地，Yu 等[63]提出了一个异质图网络模型来实现视觉常识推理，该模型通过构建语言与视觉内容的关联图，有效地提升了视觉常识推理任务的性能。不同于上述两种方法，Lin 等[64]提出了一个结构简单的推理模型，该模型通过融合属性信息，可以有效地提升视觉常识推理任务的性能。

4.2 基于预训练的方法

由于结合自然语言的相关任务都需要一个 Word2Vec 模块来把单词转变为相应的表示，因此，一个表示准确的 Word2Vec 模块对提升结合自然语言的视觉问答与推理任务的性能非常重要。最近，谷歌团队提出了一个 BERT 模型[65]，用于提取准确单词的表示。具体地，BERT 在海量语料的基础上运行自监督学习来为单词学习一个好的特征表示。在以后特定的自然语言任务中，可以直接使用 BERT 的

特征表示作为该任务的词嵌入特征。受这种思想的启发，一些工作开始尝试使用预训练的方法来完成视觉常识推理任务。与 BERT 的思想一致，这些方法通过预训练来完成视觉-语言多模态的任务。特别地，Lu 等[66]提出了一个 ViLBERT (Vision-and-Language BERT) 模型，用于提取与任务无关的视觉-语言联合表示。ViLBERT 提取的表示有效提升了多个视觉-语言任务的性能，如视觉常识推理。Su 等[67]提出了一个简单且有效的 VL-BERT 模型，通过扩展 Transformer，其使得 Transformer 能够同时使用文本和视觉内容作为输入。最后，通过在视觉-语言数据集上进行预训练，该模型可以有效提升多个视觉-语言任务的性能。尽管 BERT 可以提取准确的单词嵌入表示并提升相关任务的性能，但这种方法需要基于大量的视觉-语言数据来训练模型，增加了训练成本。

5 数据集与评估

对于基于视觉和语言的跨媒体问答与推理，一些标准的数据集被广泛使用。接下来将分别介绍数据集的详情及相关方法的实验结果。

5.1 图像的视觉问答与推理

针对基于图像的视觉问答与推理，VQA v2.0[68]是一个常用的标准数据集。该数据集分为训练集、验证集和测试集 3 个部分，其中，训练集包括80 000 张图片和 444 000 个问题，验证集包括 40 000 张图片和 214 000 个问题，测试集包括 80 000 张图片和 448 000 个问题。表 1 列出了不同方法的性能对比结果。

表 1 面向视觉问答的不同方法的性能对比

Table 1 Performance comparison of different methods for video question-answering

Method	Overall	Yes/No	Number	Other
BottomUp[28]	65.32	81.82	44.21	56.05
Counter[69]	68.09	83.14	51.62	58.97
Murel[38]	68.03	84.77	49.84	57.85
MFH[70]	68.76	84.27	49.56	59.89
DFAF[39]	70.22	86.09	53.32	60.49
BAN[71]	69.66	85.46	50.66	60.60
MCAN[11]	70.63	86.82	53.26	60.72
MLIN[37]	70.18	85.96	52.93	60.40

从表 1 可以看出，基于注意力机制的方法 MCAN[11]获得了最好的性能。这表明同时捕捉与答案相关的视觉与语言信息可以明显提升视觉问答任务的性能。

5.2 视频的视觉问答与推理

对于基于视频的视觉问答与推理，TGIF-QA[72]是一个常用的标准数据集。TGIF-QA 数据集包含了多个子任务，包括动作计数(Count)、重复动作识别(Action)、动作转换问答(Transition)和视频帧内容问答(FrameQA)。其中，Count 是开放式计数任务，结果由 L2 损失来表示。Action 和 Transition 对于每个问题提供了 5 个候选答案，结果由答案的准确率来衡量。FrameQA 被形式化为多分类问题并提供了答案词表，结果表示为分类的准确率。表 2 列出了不同方法的对比结果。

表 2 面向视频问答的不同方法的性能对比

Table 2 Performance comparison of different methods

Method	Count	Action	Transition	FrameQA
ST-VQA[72]	4.28	60.8	67.1	49.3
Comem[53]	4.10	68.2	74.3	51.5
PSAC[55]	4.27	70.4	76.9	55.7
HME[73]	4.02	73.9	77.8	53.8

从表 2 可以看出，基于注意力机制的方法 HME[73]获得了最好的性能。这表明捕捉相关的视频信息有助于提升视频问答任务的性能。

5.3 视觉常识的推理

对于基于视觉常识的推理，VCR[5]是一个标准数据集。这个数据集包含 290 000 对问题-答案-理由。同时，视觉常识推理任务包括 3 个子任务，即 Q2A(给定问题，选择正确的答案)、QA2R(给定问题和正确答案，选择正确的解释)和 Q2AR(给定问题，选择正确的答案及对应的正确解释)。表 3 列出了不同方法的性能对比结果。

表 3 面向视觉常识推理的不同方法的性能对比

Table 3 Performamce comparison of different methods visual commonsense reasoning

Method	Q2A	QA2R	Q2AR
R2C[5]	65.1	67.3	44.0
CCN[62]	68.5	70.5	48.4
HGL[63]	70.1	70.8	49.8
TAB[64]	70.4	71.7	50.5
Vilbert[66]	73.3	74.6	54.8
Vlbert[67]	75.8	78.4	59.7

从表 3 可以看出，使用预训练的方法(如 Vlbert[67])，能够获得最优的性能。这表明利用预训练的方法获得一个较好的单词嵌入表示能够明显提升视觉常识推理任务的性能。

6 跨媒体问答和推理研究展望

本文分别总结了基于图像的问答与推理、基于视频的问答与推理和视觉常识推理的相关研究。结合目前已有的方案，本节总结了一些值得深入讨论的问题，可作为未来研究的方向。

(1)高效的多模态融合算法。多模态融合是跨模态的基础研究问题。目前常用的跨模态方法的目的仍然是让多模态特征元素之间充分交互。然而，随着特征维度的增加，这种方法的计算量也会增大。同时，这种方法并不能保证所得到的融合特征包含与任务相关的信息。因此，设计一种让模型主动学习如何有效地融合多模态特征是一种解决方案。通过与任务一起训练，可以让学到的融合表示包含与任务相关的信息，从而提升性能。

(2)准确的语言单词嵌入和视觉特征提取。对于基于视觉和语言的跨媒体问答和推理，提出单词嵌入特征和视觉特征是解决这个问题的第一步。对于单词嵌入，目前常用的方法是使用一个预训练的单词嵌入矩阵(如 Bert[65]或 Glove[74])，来提取单词的嵌入特征。同样地，对于视觉特征提取，目前常用的方法是使用一个在 ImageNet[75]数据集上预训练的分类网络(如 ResNet[76])来提取视觉特征。然而，这些基于预训练的方法并不能提取与任务相关的表示，从而影

响了任务的性能。把单词嵌入矩阵和视觉特征提取网络与任务一起训练，可以得到与任务相关的单词嵌入表示与视觉特征，从而提升性能。因此，如何把单词嵌入矩阵和视觉特征提取网络与任务一起训练是一个值得研究的问题。

(3)对于基于视频的视觉问答与推理，较长的视频内容(例如电影视频)对视频问答模型的记忆和推理能力提出了更大的挑战。一些研究工作设计了外部记忆模块来存储长序列信息，并根据问题到记忆库中检索相关内容。当前视频问答技术正从简单的感知层面向更进一步的认知推理层面迈进，对于正确答案的预测不能仅仅依赖视觉或者文本内容的共现性，而应该具备对视频内容的细粒度理解。此外，如何在视频问答时结合外部知识并提供先验信息，也是理解视频内容的一个重要研究思路。

(4)高效的推理算法。基于视觉与语言的跨媒体问答和推理的目的是得到与问题对应的正确答案。因此，设计一种能够有效利用视觉和语言内容的推理网络是一个重要的研究方向。目前，常用的解决方案仍然是基于注意力机制的思路，即捕捉与答案相关的视觉内容以及语言内容来推导正确的答案。然而，这种方法本质上还是为了提取有效的表示，并没有考虑实体间的关系。最近研究人员提出的一些方法开始使用图神经网络来捕捉实体间的关系，从而提升问答和推理的性能。然而，这些方法通常采用原始的全连接网络来捕捉实体间的所有关系，增加了计算量。另外，推理往往是具有方向性的，而这类方法使用的是无向网络。因此，如何设计一种高效的有向图神经网络来解决跨媒体问答与推理是一个值得研究的问题。

(5)可解释的推理算法。目前主流的解决视觉问答任务的方法都是得到输出表示，然后利用这个输出表示与候选答案表示来计算匹配分数，最后选出分数最高的对应选项，即为最终答案。然而，这类方法的一个缺陷是模型不具有可解释性，即人们不知道模型选出答案的原因。因此，设计一种可解释的模型，即在输出正确的答案选项的同时能够给出选择这个选项的原因，是一个值得研究的问题。

结束语 当前是一个多媒体的时代，海量的文本、音频、视频等多媒体数据快速涌现，多媒体信息的使用不仅有助于人与人之间更好的交流，而且有助于人们更加完整地了解某一事物。通常，多媒体数据类型多样、关系复杂、呈现跨模态特性，因此基于多模态数据的理解和推理成为了一个极具挑战性的问题。解决这一问题有助于不同行业领域的信息交互，从而推动了各领域的发展。基于此背景，本文对基于视觉和语言的跨媒体问答与推理的研究进展进行了详细阐述，并分别从基于图像的视觉问答与推理、基于视频的视觉问答与推理和基于视觉常识的推理3个方面进行了总结，同时介绍了常用的数据集及代表方法的实验结果，最后对跨媒体问题与推理进行了前景展望。

参 考 文 献

[1] TURING A M. Computing machinery and intelligence [J]. Mind, 1950, 59(236): 433-460.

[2] TENEY D, ANDERSON P, HE X, et al. Tips and tricks for visual question answering: Learnings from the 2017 challenge [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4223-4232.

[3] JABRI A, JOULIN A, VAN DER MAATEN L. Revisiting visual question answering baselines[C]// European Conference on Computer Vision. Springer, Cham, 2016: 727-739.

[4] ZHU L, XU Z, YANG Y, et al. Uncovering the temporal context for video question answering[J]. International Journal of Computer Vision, 2017, 124(3): 409-421.

[5] 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.

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

[7] DRUZHKOV P N, KUSTIKOVA V D. A survey of deep learning methods and software tools for image classification and object detection [J]. Pattern Recognition and Image Analysis, 2016, 26(1): 9-15.

[8] YANG S, WANG Y, CHU X. A Survey of Deep Learning Techniques for Neural Machine Translation[J]. arXiv: 2002. 07526, 2020.

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

[10] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]// Advances in Neural Information Processing Systems. 2016: 289-297.

[11] 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.

[12] NGUYEN B D, DO T T, NGUYEN B X, et al. Overcoming data limitation in medical visual question answering[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 522-530.

[13] 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.

[14] 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.

[15] ANTOL S, AGRAWAL A, LU J, et al. VQA: Visual question answering [C]// Proceedings of IEEE Conference on Computer Vision. New York: IEEE Press, 2015: 2425-2433.

[16] REN M, KIROS R, ZEMEL R. Exploring models and data for image question answering[C]// Advances in Neural Information Processing Systems. 2015: 2953-2961.

[17] 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.

[18] 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.

[19] LI R，JIA J. Visual question answering with question representation update (qru)[C]// Advances in Neural Information Processing Systems. 2016：4655-4663.

[20] CHARIKAR M，CHEN K，FARACH-COLTON M. Finding frequent items in data streams[C]// International Colloquium on Automata，Languages，and Programming. Berlin，Heidelberg：Springer，2002：693-703.

[21] KIM J H，ON K W，LIM W，et al. Hadamard Product for Low-rank Bilinear Pooling[C]// In ICLR. 2016.

[22] 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.

[23] 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.

[24] 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. 2019，33：8102-8109.

[25] XU K，BA J，KIROS R，et al. Show，attend and tell：Neural image caption generation with visual attention[C]// International Conference on Machine Learning. 2015：2048-2057.

[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. 2016：21-29.

[27] LI R，JIA J. Visual question answering with question representation update (qru)[C]// Advances in Neural Information Processing Systems. 2016：4655-4663.

[28] 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.

[29] REN S，HE K，GIRSHICK R，et al. Faster r-cnn：Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015：91-99.

[30] SCHWARTZ I，SCHWING A，HAZAN T. High-order attention models for visual question answering[C]// Advances in Neural Information Processing Systems. 2017：3664-3674.

[31] LI Y，KAISER L，BENGIO S，et al. Area attention[C]// International Conference on Machine Learning. PMLR，2019：3846-3855.

[32] PATRO B，NAMBOODIRI V P. Differential attention for visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7680-7688.

[33] GUO W，ZHANG Y，WU X，et al. Re-Attention for Visual Question Answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：91-98.

[34] 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.

[35] 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.

[36] HUDSON D A，MANNING C D. Compositional Attention Networks for Machine Reasoning[C]// International Conference on Learning Representations. 2018.

[37] GAO P，YOU H，ZHANG Z，et al. Multi-modality latent interaction network for visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019：5825-5835.

[38] 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.

[39] 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 Conference on Computer Vision and Pattern Recognition. 2019：6639-6648.

[40] KIPF T N，WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]// International Conference on Learning Representations. 2016.

[41] VELIKOVI P，CUCURULL G，CASANOVA A，et al. Graph Attention Networks[C]// International Conference on Learning Representations. 2018.

[42] MONTI F，BOSCAINI D，MASCI J，et al. Geometric deep learning on graphs and manifolds using mixture model cnns[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：5115-5124.

[43] 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. 2017：1-9.

[44] NORCLIFFE-BROWN W，VAFEIAS S，PARISOT S. Learning conditioned graph structures for interpretable visual question answering[C]// Advances in Neural Information Processing Systems. 2018：8334-8343.

[45] 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.

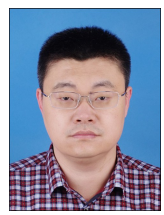
[46] KHADEMI M. Multimodal Neural Graph Memory Networks for Visual Question Answering[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：7177-7188.

[47] SUKHBAATAR S，WESTON J，FERGUS R. End-to-end memory networks[C]// Advances in Neural Information Processing Systems. 2015：2440-2448.

[48] HUDSON D，MANNING C D. Learning by abstraction：The neural state machine[C]// Advances in Neural Information Processing Systems. 2019：5903-5916.

[49] HAN Y，WANG B，HONG R，et al. Movie question answering via textual memory and plot graph[J]. IEEE Transactions on Circuits and Systems for Video Technology，2019，30(3)：875-887.

[50] WANG B, XU Y, HAN Y, et al. Movie question answering: Remembering the textual cues for layered visual contents[J]. arXiv: 1804.09412, 2018.
[51] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[52] TAPASWI M, ZHU Y, STIEFELHAGEN R, et al. Movieqa: Understanding stories in movies through question-answering [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4631-4640.
[53] GAO J, GE R, CHEN K, et al. Motion-appearance co-memory networks for video question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6576-6585.
[54] KIM J, MA M, KIM K, et al. Progressive attention memory network for movie story question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 8337-8346.
[55] LI X, SONG J, GAO L, et al. Beyond rnns: Positional self-attention with co-attention for video question answering[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 8658-8665.
[56] KIM J, MA M, PHAM T, et al. Modality Shifting Attention Network for Multi-Modal Video Question Answering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10106-10115.
[57] GAN Z, GAN C, HE X, et al. Semantic compositional networks for visual captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5630-5639.
[58] YAO T, PAN Y, LI Y, et al. Exploring visual relationship for image captioning[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 684-699.
[59] CHEN L, ZHANG H, XIAO J, et al. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5659-5667.
[60] JIANG P, HAN Y. Reasoning with Heterogeneous Graph Alignment for Video Question Answering[C]// AAAI. 2020: 11109-11116.
[61] SONG X, SHI Y, CHEN X, et al. Explore multi-step reasoning in video question answering[C]// Proceedings of the 26th ACM International Conference on Multimedia. 2018: 239-247.
[62] WU A, ZHU L, HAN Y, et al. Connective Cognition Network for Directional Visual Commonsense Reasoning[C]// Advances in Neural Information Processing Systems. 2019: 5669-5679.
[63] YU W, ZHOU J, YU W, et al. Heterogeneous Graph Learning for Visual Commonsense Reasoning[C]// Advances in Neural Information Processing Systems. 2019: 2769-2779.
[64] LIN J, JAIN U, SCHWING A G. TAB-VCR: Tags and Attributes based Visual Commonsense Reasoning Baselines[C]// Advances in Neural Information Processing Systems. 2019.
[65] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2018: 4171-4186.
[66] 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.
[67] SU W, ZHU X, CAO Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations[C]// International Conference on Learning Representations. 2020.
[68] 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.
[69] ZHANG Y, HARE J, PRÜGEL-BENNETT A. Learning to count objects in natural images for visual question answering [J]. arXiv: 1802.05766, 2018.
[70] YU Z, YU J, XIANG C, et al. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 5947-5959.
[71] KIM J H, JUN J, ZHANG B T. Bilinear attention networks [C]// Advances in Neural Information Processing Systems. 2018: 1564-1574.
[72] JANG Y, SONG Y, YU Y, et al. Tgif-qa: Toward spatio-temporal reasoning in visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2758-2766.
[73] FAN C, ZHANG X, ZHANG S, et al. Heterogeneous memory enhanced multimodal attention model for video question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1999-2007.
[74] 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.
[75] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[76] 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.



WU A-ming, born in 1987, Ph. D. His main research interests include multimedia analysis and machine learning.

HAN Ya-hong, born in 1977, Ph. D, professor. His main research interests include multimedia analysis, computer vision and machine learning.