面向跨模态数据协同分析的视觉问答方法综述

崔 政, 胡永利, 孙艳丰, 尹宝才
(北京工业大学信息学部, 北京 100124)

摘 要: 协同分析和处理跨模态数据一直是现代人工智能领域的难点和热点,其主要挑战是跨模态数据具有语义和异构鸿沟. 近年来,随着深度学习理论和技术的快速发展,基于深度学习的算法在图像和文本处理领域取得了极大的进步,进而产生了视觉问答(visual question answering, VQA)这一课题. VQA 系统利用视觉信息和文本形式的问题作为输入,得出对应的答案,核心在于协同理解和处理视觉、文本信息. 因此,对 VQA 方法进行了详细综述,按照方法原理将现有的 VQA 方法分为数据融合、跨模态注意力和知识推理3类方法,全面总结分析了 VQA 方法的最新进展,介绍了常用的 VQA 数据集,并对未来的研究方向进行了展望.

关键词: 跨模态数据; 深度学习; 视觉问答; 数据融合; 跨模态注意力; 知识推理
中图分类号: U 461; TP 308 **文献标志码**: A **文章编号**: 0254 - 0037(2022)10 - 1088 - 12
doi: 10.11936/bjutxb2021040030

Visual Question Answering Methods of Cross-modal Data Collaborative Analysis: a Survey

CUI Zheng, HU Yongli, SUN Yanfeng, YIN Baocai
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Collaborative analysis and processing of cross-modal data are always difficult and hot topics in the field of modern artificial intelligence. The main challenge is the semantic and heterogeneous gap of cross-modal data. Recently, with the rapid development of deep learning theory and technology, algorithms based on deep learning have made great progress in the field of image and text processing, and then the research topic of visual question answering (VQA) has emerged. VQA system uses visual information and text questions as input to get corresponding answers. The core of the system is to understand and process visual and text information cooperatively. Therefore, VQA methods were reviewed in detail. According to the principle of methods, the existing VQA methods were divided into three categories including data fusion, cross-modal attention and knowledge reasoning. The latest development of VQA methods was comprehensively summarized and analyzed, the commonly used VQA data sets were introduced and prospects for future research direction were suggested.

Key words: cross-modal data; deep learning; visual question answering (VQA); data fusion; cross-modal attention; knowledge reasoning

收稿日期: 2021-04-28; 修回日期: 2021-06-07
基金项目: 国家自然科学基金资助项目(61672071, U1811463, U19B2039)
作者简介: 崔 政(1996—), 男, 博士研究生, 主要从事多模态智能、深度学习方面的研究, E-mail: CuiZ@emails.bjut.edu.cn

如何使算法可以像人类一样同时理解和利用多种模态数据是人工智能领域中的一个重要研究课题. 随着深度学习技术的成熟，基于深度学习的计算机视觉和自然语言处理技术飞速发展，在此基础上视觉问答(visual question answering，VQA)这一涉及图像理解和自然语言处理 2 个领域的研究课题受到越来越多的关注. 虽然人工智能领域的学者已经提出了多种基于深度学习的 VQA 模型，但是如何准确地学习跨模态数据特征，目前还没有一个完整的解决方案.

1　VQA 简介

随着大数据时代的到来，全球的数据量正在呈指数级增长. 每个用户都在社交媒体和互联网应用上产生大量的数据，这些数据包括图片、文本、声音、视频和浏览记录等，具有明显的跨模态性质. 面对庞大的跨模态数据，如何提取有效的信息和进行准确的分析成为了一个研究难点和热点. 在此背景下，VQA 这一研究课题被提出. 如图 1 所示，当给定一张图片和一个对应的问题，VQA 系统需要根据问题来提取图片上的有效信息，进而得出正确的答案. 这就要求算法能够对图像和问题的语义信息具有高层次的理解，并且能够同时处理和分析图像和文本 2 种模态的数据.

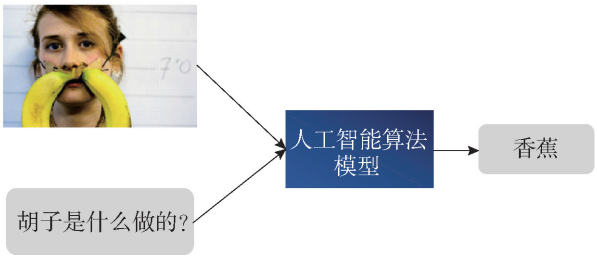

图 1　VQA 示意图

Fig. 1　Schematic diagram of VQA

近年来，许多基于深度学习的方法被提出以解决 VQA 任务[1-8]，为了更加清晰地阐述不同方法的研究思路和便于学者参考，本文按照原理的不同将这些方法分为数据融合、跨模态注意力和知识推理 3 类，介绍了每一类方法的相关工作和常用的 VQA 数据集，并对最新出现的基于视频和文本问题的 VQA 任务进行了介绍. 最后，对每一类方法做出总结并对未来的研究方向进行了展望.

2　VQA 研究现状及方法

首先，给出 VQA 系统的定义，给定一个图像 v 和一个问题 q，VQA 系统的目的是预测一个与真实标签 a^* 相匹配的答案 $\hat{a}$，目前 VQA 中常用的方法通过分类器 $f_\theta(\cdot)$ 的得分来获得正确答案的预测，即

$$\hat{a} = \arg\max_{a \in A} f_\theta(a \mid v, q) \tag{1}$$

一个完整的 VQA 系统通常由 4 个部分组成：图像特征提取器、文本特征提取器、跨模态特征学习模块和答案分类器.

最初各种卷积神经网络被用来作为图像特征提取器，包括亚历克斯网络(Alex network，AlexNet)[9]、谷歌网络(Google network，GoogLeNet)[10]、视觉几何组网络(visual geometry group network，VGGNet)[11]和残差网络(residual network，ResNet)[12]. AlexNet 是一个具有 5 个卷积层的深层网络，是第 1 个大幅度提高分类精度的深度卷积网络，并获得了 2012 年的 ImageNet 数据集大规模视觉识别挑战赛冠军. 在 2014 年的挑战赛中，GoogLeNet 获得了第 1 名、VGGNet 获得了第 2 名，这 2 类模型结构的共同特点是层次更深了. VGGNet 采用连续的几个 3 × 3 的卷积核代替 AlexNet 中的较大卷积核，在保证具有相同感知野的条件下提升了网络的深度，在一定程度上提升了神经网络的效果. GoogLeNet 使用 1 × 1 的卷积来进行降维，并且在多个尺寸上同时进行不同尺度的卷积，然后再进行聚合，最终取得了更加优越的性能. ResNet 有效地解决了深层神经网络难以训练的问题，是卷积网络发展史上具有里程碑意义的工作. 采用卷积网络作为图片的特征提取器可以得到包含丰富语义信息的优质的图像特征表示，这也推动了 VQA 这一课题的发展. 虽然这些卷积网络通常能够提取具有概括性的全局图像特征描述，但是也丢失了大量有用的细粒度信息，这些细粒度的信息可以帮助算法得到精准的图像理解. 因此，最近的研究工作探讨了目标检测器提取的区域级特征的可用性. Anderson 等[13]提出了自下而上的注意力机制来提取图像的特征，这一方法类似于人类视觉系统中的注意力机制，可以过滤掉不重要信息的特征，最终通过在视觉基因数据库[14]上预训练的快速目标检测模型[15]得到区域级的图像特征. 这些区域特征包含了丰富的细粒度语义信息，非常有利于图像的细粒度理解和跨模态特征的学习.

文本特征提取器被用来抽取文本问题的特征，通常首先利用文本特征提取方法[16-22]将每个单词或整个问题嵌入到问题的文本语义空间，然后通过递归神经网络(recurrent neural network，RNN)来得

到序列化的特征. 长短时记忆网络(long and short term memory network, LSTM)、门循环单元(gate recurrent unit, GRU)常被用作文本特征编码器,因为它们对于序列数据的处理非常有效.

跨模态特征学习模块是整个VQA系统的核心,这一模块的主要目的是综合分析和利用2种模态的数据,挖掘2种模态数据之间的关联关系,通过数据融合、跨模态注意力、知识推理等方法学习一个对于输入数据的跨模态特征表示.

答案分类器通常由一个多层全连接神经网络组成,输入是图片和问题的跨模态特征表示,其最终输出维度是预选答案的个数. 通过这一模块可以得到每个预选答案的置信度得分,从而选择得分最高的答案作为预测的正确答案.

2.1 数据融合

在VQA算法中,核心在于文本和视觉这2种模态数据的联合表示. 基于数据融合的方法将图像和文本模态的特征向量进行数据融合,从而得到跨模态特征表示.

2.1.1 多模态紧凑双线性池化(multimodal compact bilinear pooling, MCB)模型

Fukui等[23]提出了MCB模型,这一模型利用MCB得到一个特征的联合表示. 双线性池化方法是计算2个向量之间的外积,与元素积不同,它允许2个向量的所有元素之间的乘法交互. 当特征向量的维度较大时会导致学习参数的激增,因此,MCB模型使用了Count Sketch函数将外积投影到低维空间,避免了直接计算外积.

MCB方法使用152层的ResNet作为图像特征提取器、LSTM循环神经网络作为文本特征提取器,然后计算问题特征向量和每个图像网格特征向量之间的融合表示和每个融合向量的权重,最后将融合向量按照权重求和,这样就得到了经过2个模态交互的加权图像特征表示. 接着将文本向量和加权后的视觉向量再进行一次数据融合,得到跨模态的特征表示. 最终以跨模态特征作为输入,使用一个全连接网络计算每个候选问题的得分.

MCB方法的主要特点是降低了双线性池化的参数量,实现了文本和图像2种模态数据的交互,并进行了深度的数据融合.

2.1.2 基于Hadamard积的多模态低秩双线性池化(multimodal low-rank bilinear pooling, MLB)模型

与线性模型相比,双线性模型提供了更丰富的信息,也被应用于各种视觉任务,如对象识别、分割和VQA,并且也获得了优良的性能. 然而,由于特征的维度往往很高,导致了双线性表示的计算复杂性较高,这也限制了该模型的适用性. Kim等[24]提出了一种基于Hadamard积的MLB模型来实现有效的多模态注意力机制学习和数据融合.

MLB将双线性池化中的三维权重张量分解为3个二维权重矩阵,使权重张量变为低秩张量. 模型首先计算经过2个权重矩阵线性投影的2个输入特征向量的Hadamard积,并且使用非线性函数进行激活,添加了残差连接. 在得到融合向量后,使用MLB方法得到了一个有效的面向VQA任务的视觉特征注意力机制. 最后,通过另一个MLB融合文本特征和注意力加权的视觉特征,得到跨模态特征表示.

MLB模型利用Hadamard积来降低计算的复杂性,得到了更加紧凑的特征表示,也实现了跨模态数据间的深度融合.

2.1.3 多级注意力网络(multi-level attention networks, MLAN)模型

许多VQA的方法主要从抽象的低级视觉特征推断答案,而忽略了图像高层语义和丰富的文本语义空间的建模. Yu等[25]提出了一种MLAN,这一网络通过语义注意力机制缩小不同模态数据之间的语义鸿沟,通过视觉注意力增强细粒度图像特征的空间推理.

MLAN模型包括3个部分,分别是语义注意力、上下文意识的视觉注意力和联合注意力. 语义注意力模块的目的是从图像中挖掘出对于回答问题更重要的概念. 上下文意识的视觉注意力模块把图片进行卷积计算后的特征按区域输入到双向GRU中,将每一步GRU中的前向和后向隐层向量组合起来,为每个区域形成一个新的特征向量. 新的特征向量不仅包含了对应区域的视觉信息,而且还包含了来自周边区域的上下文信息. 然后,将每个包含上下文信息的图像特征加权求和. 联合注意力模块将问题向量和学习到的视觉向量进行融合,最终得到了跨模态的特征表示.

MLAN模型在数据融合的过程中考虑了不同视觉特征的重要性和视觉特征的上下文语境,得到了更加优良的数据融合特征表示.

2.1.4 多模态塔克融合模型

双线性模型是VQA任务中信息融合的一种有效的方法. 它有助于学习问题意义和图像中视觉概念之间的高级关联,但也始终面临着数据维度太大

的问题. 为了解决这一问题，Ben-Younes 等[26]提出了多模态塔克融合模型 MUTAN，这一模型通过多模态张量的塔克分解有效地实现了视觉和文本特征表示之间的双线性交互.

双线性模型是对数据融合问题有效的解决方案，它对矢量 $\boldsymbol{q}$ 和 $\boldsymbol{v}$ 之间的双线性相互作用进行了编码，即

$$\boldsymbol{y}=(\boldsymbol{E}\times_1\boldsymbol{q})\times_2\boldsymbol{v} \tag{2}$$

式中 $\boldsymbol{E}$ 为约束张量. 尽管双线性模型有很强的建模能力，但完全参数化的数据双线性交互在 VQA 中很难实现，因为文本、视觉和输出特征向量使用相同的维度，使得参数量变得非常庞大. 因此，MUTAN 使用塔克分解将式(2)重写为

$$\boldsymbol{y}=((\boldsymbol{E}\times_1(\boldsymbol{q}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{q}}))\times_2(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{v}}))\times_3\boldsymbol{W}_{\mathrm{o}} \tag{3}$$

式中 $\boldsymbol{W}_{\mathrm{q}}$、$\boldsymbol{W}_{\mathrm{v}}$ 和 $\boldsymbol{W}_{\mathrm{o}}$ 为可学习的投影矩阵. 这一方法对 $\boldsymbol{q}$ 和 $\boldsymbol{v}$ 的投影进行双线性相互作用编码. MUTAN 模型在降低了计算复杂性的基础上实现了更强的表现力，得到了较优的预测准确性.

MCB 模型和 MLB 模型在双线性池化的基础上进行了改良，实现了跨模态数据之间的交互，计算了数据之间的高级关联. MUTAN 利用塔克分解得到了表现力更强的跨模态特征表示. MLAN 模型创新地考虑了视觉向量的上下文语境信息.

以上几种模型对 VQA 任务进行了初步的探索，通过池化和矩阵分解的方式融合图像和文本特征，从而得到可以预测答案的跨模态特征表示. 然而，数据融合的方法缺乏对图像和文本特征之间关联关系的深度挖掘，缺乏对特征的精细化计算，得到的跨模态特征中冗余数据和噪声较多.

2.2　跨模态注意力

视觉场景往往包含大量信息，如何利用有限的感知和计算资源从大量信息中筛选出高价值的信息是计算机视觉中的核心问题. 在长期进化中，人类形成了一种特有的大脑信号处理机制——视觉注意力机制. 这一机制极大地提高了视觉信息处理的效率与准确性. 具体而言，当看到一张图片时，人类视觉系统可以快速扫描整个图片并获得需要重点关注的目标区域，形成注意力焦点，然后对目标区域投入较多的感知和计算资源，从而获取更多关注区域的细节信息，同时抑制其他无用信息[27].

在 VQA 任务中，跨模态注意力是一种非常高效的方法. 通过注意力机制，可以得到跨模态数据之间准确的关联关系和语义理解. 最初，研究者利用视觉注意力机制[13,28-41]得到图像中与问题相关的区域. 之后，考虑到单向注意力机制没有有效利用文本信息，研究者提出了基于跨模态协同注意力的方法[42-53]，利用图像和文本的双向注意力信息挖掘出有效知识. 下面就典型方法进行介绍.

2.2.1　堆叠注意力网络(stacked attention networks，SAN)模型

Yang 等[41]提出了 SAN 模型，这一模型根据问题特征在图像上进行多步推理，最终得到图像上的关键特征.

SAN 模型利用 VGGNet 提取图像的特征，并利用文本卷积网络或 LSTM 提取问题特征，得到图像特征矩阵 $\boldsymbol{V}$ 和问题特征向量 $\boldsymbol{Q}$. SAN 模型通过多步迭代计算的方式预测答案. 首先计算以问题特征为查询，每个视觉向量的权重公式为

$$\boldsymbol{h}_1=\tanh(\boldsymbol{W}_{\mathrm{v}}\boldsymbol{V}\oplus(\boldsymbol{w}_{\mathrm{q}}\boldsymbol{Q}+\boldsymbol{b})) \tag{4}$$

$$\boldsymbol{p}=\mathrm{softmax}(\boldsymbol{W}_{\mathrm{p}}\boldsymbol{h}_1+\boldsymbol{b}_{\mathrm{p}}) \tag{5}$$

式中：$\boldsymbol{W}_{\mathrm{q}}$ 和 $\boldsymbol{W}_{\mathrm{p}}$ 为可学习的投影矩阵；$\boldsymbol{b}$ 和 $\boldsymbol{b}_{\mathrm{p}}$ 为偏执向量. 基于第 1 次得到的视觉向量的注意力分布 $\boldsymbol{p}$，将视觉向量的权重求和，并加上文本特征形成新的查询向量 $\boldsymbol{u}$，公式为

$$\tilde{\boldsymbol{v}}=\sum_i\boldsymbol{p}_i\boldsymbol{v}_i \tag{6}$$

$$\boldsymbol{u}=\tilde{\boldsymbol{v}}+\boldsymbol{v} \tag{7}$$

然后，可以根据新的查询向量进行下一步的注意力权重分布计算，并延续到第 k 次，即

$$\boldsymbol{h}_1^k=\tanh(\boldsymbol{W}_{\mathrm{v}}^k\boldsymbol{V}\oplus(\boldsymbol{W}_{\mathrm{q}}^k\boldsymbol{u}^{k-1}+\boldsymbol{b}^k)) \tag{8}$$

$$\boldsymbol{p}^k=\mathrm{softmax}(\boldsymbol{W}_{\mathrm{p}}^k\boldsymbol{h}_1+\boldsymbol{b}_{\mathrm{p}}^k) \tag{9}$$

式中：$\boldsymbol{W}_{\mathrm{v}}^k$、$\boldsymbol{W}_{\mathrm{q}}^k$ 和 $\boldsymbol{W}_{\mathrm{p}}^k$ 为第 k 次计算的投影矩阵；$\boldsymbol{b}^k$ 和 $\boldsymbol{b}_{\mathrm{p}}^k$ 为第 k 次计算的偏执向量. 在经过多次推理计算后，将最终得到的 $\boldsymbol{u}^k$ 作为跨模态特征表示，并通过分类器计算每个问题的得分.

2.2.2　由下到上和由上到下的注意力模型

Anderson 等[13]提出了由下到上和由上到下的注意力模型，由下到上注意力模块相当于对整个图片上的所有像素点进行了注意力分布的计算，最终得到了包含丰富语义特征目标级别的视觉特征. 如果输入是一张厨房的图片，那么这一模块可以得到很多显著性区域，包括食物、人、汤勺、平底锅等. 以显著性区域特征作为跨模态特征学习模块的输入，算法可以精确地找到视觉特征和问题特征之间的对应关系. 由上到下的注意力模块以文本特征为查询向量找到图像上的关键区域，甚至是答案所对应的区域.

由下到上和由上到下的注意力模型是一个在 VQA 领域具有里程碑意义的工作，大幅提高了 VQA

的准确性,同时,其提出的目标级别的视觉特征也让各种任务受益.

2.2.3 双线性注意力网络(bilinear attention networks,BAN)模型

Kim 等[49]提出了 BAN 模型. 这一模型首先将图像编码为显著性区域特征,并提取问题中每个单词的特征. 在得到图像和文本的特征后,计算 2 种模态特征之间的双线性注意力,也就是计算 2 组特征中两两之间的相似性. BAN 模型通过多个双线性特征图按相关性的大小融合 2 种模态的数据,在每一次融合后都添加了残差连接. 这一模型考虑了模态之间双向的高级关联,实现了多模态数据之间细粒度的交互.

2.2.4 密集的对称协同注意力网络(dense symmetric co-attention network,DCN)模型

Nguyen 等[50]提出了 DCN 模型,这一模型利用协同注意力机制以改善视觉特征与文本特征的融合. 得到图像和问题后,首先计算每个单词的特征和图像的卷积特征,然后在 DCN 中执行 3 种计算:1)注意力特征图的计算;2)多模态特征的拼接;3)残差连接的整流线性单元(rectified linear unit,ReLU)映射. 这些计算被封装成一个复合的计算模块,被称为密集协同注意力模块,因为它考虑了任何图像区域和任何问题词之间的每一次交互. 该模块在 2 种模态的计算之间具有完全对称的架构,并且可以堆叠,形成一个层次结构,使得图像和问题数据对之间能够进行多步交互.

2.2.5 动态融合的模态内和模态间注意力流(dynamic fusion with itra-and inter-modality attention flow,DFAF)模型

Gao 等[51]提出了 DFAF 模型. 如图 2 所示,DFAF 模型整合了跨模态的自注意力和协同注意力来实现视觉和文本 2 种模态内部和之间的有效信息流. DFAF 模型首先通过模态间注意力模块生成模态间的注意力信息流来实现信息的交互,在模态间注意力模块中,视觉和语言特征生成一个联合模态协同注意力矩阵. 每个视觉区域和文本单词根据联合模态协同注意力矩阵选择特征,模态间注意力模块根据来自另一模态的注意加权信息流融合和更新每个图像区域和每个单词的特征. 在这一模块之后,DFAF 计算动态的模态内注意力信息流,用于在每个模态中传递信息以捕获复杂的模态内关系. 视觉区域和单词产生自注意力权重,并从其他实例中按照注意力权重整合信息. 在动态的模态内注意力模块中,虽然信息流只在相同的模态中传播,但是另一个模态的信息被考虑并用于调节模态内注意力权重和信息流.

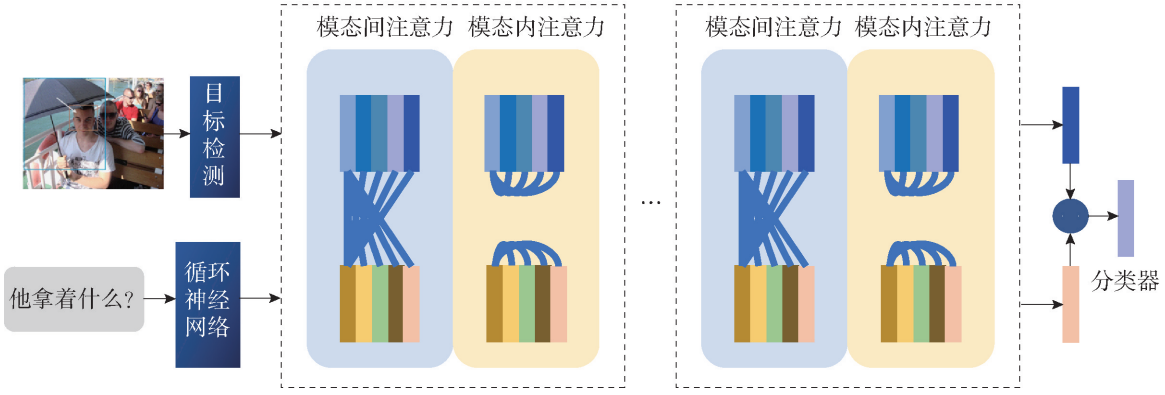

图 2 DFAF 示意图[51]

Fig. 2 Schematic diagram of DFAF[51]

DFAF 模型多次堆叠模态间注意力模块和动态的模态内注意力模块,实现了模态间和模态内的注意力信息流的深度交互.

2.2.6 多模态潜在交互(multi-modality latent interaction,MLI)模型

Gao 等[52]提出了 MLI 模型,这一模型由一系列叠加的多模态潜在交互模块组成,其目的是将输入的视觉区域和问题词信息汇总为每个模态的少量潜在具有高级语义的摘要向量. 其核心思想是在潜在摘要向量之间传播视觉和语言信息,从全局角度对复杂的跨模态交互进行建模. 在潜在交互摘要向量之间进行信息传播后,视觉区域和单词特征将整合跨域摘要中的信息以更新其特征. MLI 模块的输入和输出具有相同的维度,整个网络将 MLI 模块分多个阶段堆叠,逐步精炼视觉和语言特性. 最后,将视觉区域和问题词的平均池化特征进行元素相乘后作

为跨模态特征来预测最终答案.

注意力模型在 VQA 任务中取得了极大的成功，大幅度提高了答案预测的准确性，促进了这一领域的发展. 基于注意力机制的方法，通过计算模态内数据和模态间数据的关联关系，对数据进行了细粒度的关联建模，成功提取了有效信息，抑制了冗余数据.

相比于特征融合的方式，基于注意力模型的方法同时考虑了模态内和模态间的信息流，利用多层的神经网络对信息进行深度建模，实现了数据间的深度交互，得到了拥有更高级语义信息的较为精炼的特征表示. 因此，基于注意力机制的方法获得了较高的预测准确性.

2.3　知识推理

在逻辑学中，推理是一种思维的基本形式，是由一个或几个已知的判断（前提）推出新判断（结论）的过程，包含直接推理、间接推理等. 人类具有强大的推理能力，在面对一些问题时，通过深度的思考和多步的推理使问题得以解决. 在人工智能领域，如何让算法具有推理能力是一个核心课题.

在 VQA 中，一个问题往往无法直接得出答案，问题中描述了场景和不同物体之间的联系，因此，算法必须具备推理能力，可以根据问题描述推理判断物体之间和物体与所处场景之间的关系.

2.3.1　基于图表示的 VQA 模型

Teney 等[54]提出了一种基于场景内容和问题的结构化表示的 VQA 系统模型 Graph VQA. VQA 中的一个关键挑战是需要在视觉域和文本域上进行联合推理.

针对每一对图片和问题数据，Graph VQA 生成一个视觉场景图和一个文本问题图. 视觉场景图以每一个视觉向量作为节点，2 个特征之间的空间关系作为它们的连接边；文本问题图以每个单词作为节点，单词之间的语法关系作为连接边. GRU 被用来编码 2 个图上的节点，在多次迭代中，GRU 更新每个节点的表示，该节点集成了图中相邻节点的上下文语境信息. 所有图像目标和所有单词的特征被成对地组合，并以注意力的形式对它们进行加权，有效地匹配了问题和场景之间的元素. 经过注意力加权的特征通过最终的分类器得到每个固定候选答案的预测分数.

2.3.2　复合关系注意力网络（composed relation attention network，CRA-Net）模型

现有的 VQA 模型一部分利用注意力机制来定位相关的目标区域，另一部分利用关系推理的方法来检测目标关系. 然而，这些模型大多对简单的关系进行编码，不能为回答复杂的视觉问题提供足够的复杂知识，也很少组合、利用目标视觉特征和对象间的关系特征.

Peng 等[55]提出了 CRA-Net 模型，这一模型包括 2 个问题自适应关系注意力模块，不仅可以提取细粒度和精确的二元关系，而且可以提取更复杂的三元关系. 这 2 种与问题相关联的目标关系都能揭示更深层次的语义，从而提高问答的推理能力. 此外，CRA-Net 在相应问题的指导下，将目标的视觉特征与关系特征相结合，有效地融合了这 2 类特征，得到了拥有丰富知识的特征表示.

在得到图片目标区域特征和问题中每个单词的特征后，CRA-Net 首先利用单词自注意力计算每个单词的权重，然后把所有问题单词进行加权求和，得到问题的向量表示. 接着，CRA-Net 以问题向量作为语境学习目标之间的细粒度、精确的二元关系和更复杂的三元关系. 这 2 种问题的相关关系都能揭示更深层次的语义，提高推理能力. 此外，推导出的三元关系将多个重要对象联系起来，提供了一种更全面的视觉关系表示，弥补了二元关系对复杂关系表达的局限性. 最后，融合了问题特征的单目标注意力特征、二元关系特征和三元关系特征通过元素级的点积得到用于预测答案的跨模态特征.

2.3.3　深度模块化协同注意力网络（modular co-attention networks，MCAN）模型

VQA 要求对图像的视觉内容和问题的文本内容同时进行精细的理解. 因此，设计一个有效的协同注意力模型，将问题中的关键词与图像中的关键对象联系起来是 VQA 系统具有良好性能的核心. 到目前为止，大多数成功的协同注意力学习尝试都是通过浅层模型实现的，而深度协同注意力模型与浅层模型相比几乎没有改善.

Yu 等[56]提出了 MCAN 模型，这一模型的灵感来自于 Transformer 模型[57]. Transformer 是第一个只用注意力机制搭建的自然语言处理模型，不仅计算速度更快，在翻译任务上也获得了更好的结果. MCAN 模型是由多个协同注意力模块组成的具有编码和解码两部分的深度模块化网络. 每个协同注意力模块由 2 个基础的注意力单元组成，这 2 个单元对问题和图像的自注意力以及图像的引导注意力进行建模. 协同注意力的基础注意力计算由多头点积注意力机制组成，在给定查询 $\boldsymbol{q}$、键值 $\boldsymbol{k}$ 和特征值 $\boldsymbol{v}$ 对后，可以得到经过注意力加权

后的特征值

$$\mathrm{Att}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}) = \mathrm{softmax}(\boldsymbol{q}\boldsymbol{k}^{\mathrm{T}}/\sqrt{d_{\mathrm{k}}})\boldsymbol{v} \quad (10)$$

式中 d_{k} 为特征向量的维度,然后将不同通道拼接,公式为

$$\boldsymbol{h}_i = \mathrm{Att}(\boldsymbol{q}\boldsymbol{W}_{\mathrm{q}},\boldsymbol{k}\boldsymbol{W}_{\mathrm{k}},\boldsymbol{v}\boldsymbol{W}_{\mathrm{v}}) \quad (11)$$

$$\mathrm{MHead}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}) = \mathrm{Concat}(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n)\boldsymbol{W}_{\mathrm{o}} \quad (12)$$

式中:$\boldsymbol{W}_{\mathrm{q}}$、$\boldsymbol{W}_{\mathrm{k}}$ 和 $\boldsymbol{W}_{\mathrm{v}}$ 为注意力计算中的特征投影矩阵;Concat( )为特征拼接函数;$\boldsymbol{W}_{\mathrm{o}}$ 为投影矩阵.

MCAN 在编码和解码的框架下对跨模态数据进行了深度的注意力编码,取得了较高的预测精度.

2.3.4 关系感知的图注意力网络模型

为了回答与图像相关的具有复杂语义的问题,VQA 模型需要充分理解图像中的视觉场景,尤其是不同对象之间的动态交互. Li 等[58]提出了关系感知的图注意力网络模型 ReGAT,将每个图像编码成一个图,通过图注意力机制建立多类型的对象间关系模型,学习基于问题特征的图像自适应关系表示.

ReGAT 建模了 2 类视觉对象关系:1) 表示对象之间几何位置和语义交互的显式关系;2) 捕捉图像区域间隐藏的动态隐式关系. 在得到问题特征和图像上的目标区域特征后,ReGAT 首先将问题特征和每个目标的特征进行融合,得到了包含问题特征的目标特征. 利用新的目标特征,ReGAT 构建了一个目标之间的关系图,并在 3 种尺度上学习目标间的高级关联关系,分别是语义关系、空间位置关系和隐藏关系. 在对目标特征进行图关联学习后,融合视觉特征和问题特征进行答案预测.

2.3.5 多模态关系推理模型

Cadene 等[59]提出了多模态关系推理模型 MUREL,这一多模态关系推理模型在问题和图像的推理学习方面取得了领先的效果. MUREL 由多个多模态关系单元组成,它能够表示问题和图像区域之间丰富的交互作用,并显式地为区域之间的关系建模. 整个模型将多模态关系单元嵌入一个迭代推理过程中,该过程逐步精炼内部的知识表示来回答问题. 通过迭代推理计算,图像中与问题相符的二元组关系被准确提取,进而得到问题的准确答案.

2.3.6 基于线性调制的视觉推理模型

Perez 等[60]提出了一种基于线性调制模块的视觉推理模型,利用包含调制模块的残差单元进行迭代推理,实现对视觉信息的深度理解.

对于给定的图片和问题,首先提取问题向量和图片的卷积特征,然后利用问题向量中的信息对视觉特征中不同通道的数据进行线性映射,进而调整卷积特征. 多次使用这一调制方法,可以学习到图像中与问题相关的特征信息.

2.3.7 组合注意力网络

神经网络已在图像识别、语音识别等感知层面取得巨大成功,但是在更进一步的推理层面仍有欠缺. 为解决这一问题,Hudson 等[61]提出了一种记忆、关注和组合(memory, attention, and composition, MAC)网络架构.

MAC 网络由一个输入神经元、一个核心的循环神经网络以及一个输出神经元组成. 输入神经元将原始图像和问题转化为分布式向量表征. 核心的循环神经网络将问题分解为一系列运算(也叫控制),它们可以从图像(知识库)中检索信息,并将结果聚合为循环记忆. 通过这些运算,网络按照序列推理问题. 答案分类器使用问题特征和最终记忆状态特征得出最终答案.

2.3.8 基于隐式信息和符号重表示的知识推理模型

Marino 等[62]提出了一种基于隐式信息和符号重表示的知识推理模型 KRISP,如图 3 所示. 这一模型在知识库上集成了隐式知识和基于显式图的推理. 隐式知识模型接受视觉特征和问题编码,而显式知识模型处理图像和问题符号.

KRISP 首先对自然语言处理算法无监督学习得到的隐式知识进行预训练,并利用基于 Transformer 的模型进行监督训练;然后利用知识图谱对符号知识进行编码;最后对这 2 种知识进行知识推理计算和融合学习.

基于知识推理的方法在 VQA 任务中取得了突破性的进展,这一类方法结合跨模态注意力机制和推理学习的思路对图像和问题的联合输入数据进行推理学习,进而取得了较高的准确率.

基于注意力机制的方法注重于对数据的关联关系进行建模,面对较为复杂的场景,答案通常无法直接得出,必须根据多组特征之间的关联信息推理得出,因此,由单纯的关联建模得到的特征仍然包含较多的冗余数据.

基于知识推理的方法通过推理计算在提炼有效信息的基础上大大减少了特征中的冗余数据,同时,可以对特征之间的多元关系进行建模学习. 这类方法通常对输入数据进行多步迭代计算,对信息进行逐步地建模和推理学习,进而得到较优的跨模态特

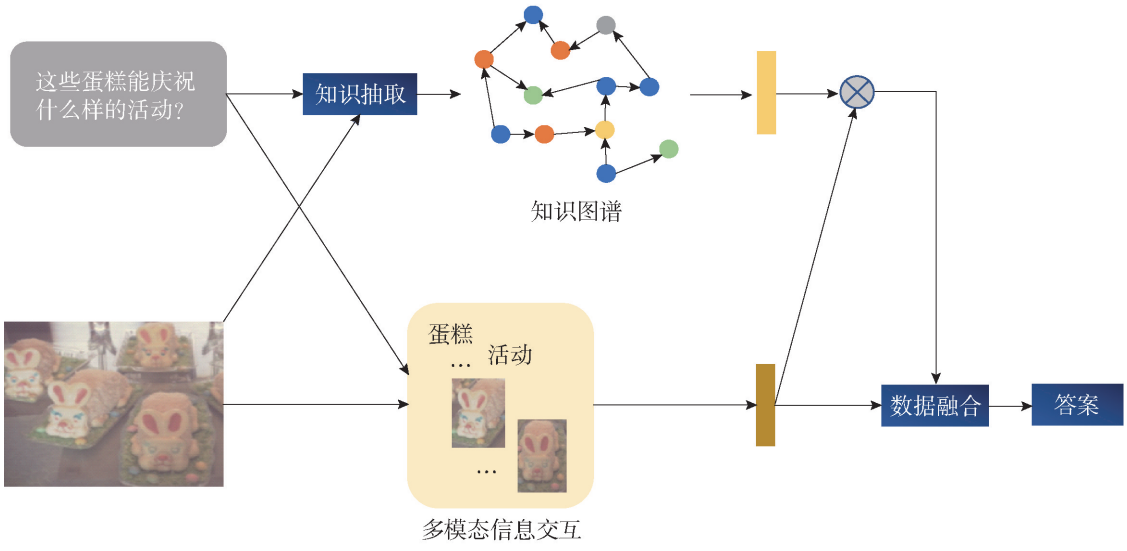

图 3　KRISP 示意图[62]

Fig. 3　Schematic diagram of KRISP[62]

征表示.

2.4　基于视频的 VQA

视频问答是 VQA 领域的一个新兴课题,由于其在人工问答系统、机器人对话、视频检索等方面的广泛应用,近年来受到越来越多的关注. 与基于图像的问答任务不同,视频问答更加实用,因为输入的视觉信息经常动态变化.

与图像问答相比,视频问答更具有挑战性. 视频中的视觉内容更为复杂,一个视频可能包含数千帧. 视频中经常包含多种动作,但只有一部分动作是关注者感兴趣的. 视频问答任务中的问题往往包含着与时间线索有关的信息,这意味着在进行答案推理时,既要考虑目标的时间位置,又要考虑目标之间的复杂交互作用.

Huang 等[63]提出了位置意识的图卷积网络模型来完成视频问答任务. 这一模型整合视频中目标的位置信息,构建具有位置意识的图,在图中每个节点都由其特征向量和位置特征进行联合表示. 基于所构造的图,这一模型使用图卷积来推断动作的类别和时间位置. 由于图形是建立在对象上的,因此,该方法能够聚焦于前景的动作内容,以便更好地进行视频问答.

Jiang 等[64]提出了一种问题引导的时空上下文注意力网络模型. 这一模型将问题产生的语义特征分为两部分:空间部分和时间部分,分别从空间和时间 2 个维度指导语境注意力的构建过程. 在相应的语境注意力的引导下,视觉特征可以在空间和时间维度上得到更好的利用.

3　VQA 数据集

1) Visual Genome[14]:该数据集包含 108 077 张图片、1 445 233 个图片和问题的数据对,图像来源为 YFCC100M 和 COCO 数据集,共有约 540 万张图像中的区域描述信息,这些信息能够达到精细的语义层次,问题类型是 6W(what、where、how、when、who、why).

2) VQA-v1[65]:训练集包含 82 783 张图片、248 349 个问题和 2 483 490 个答案. 验证集包含 40 504 张图片、121 512 个问题和 1 215 120 个答案. 测试集包含 81 434 张图片和 244 302 个问题. 数据集中的图片来源于 COCO 数据集.

3) VQA-v2[66]:训练集包含 82 783 张图片、443 757 个问题和 4 437 570 个答案. 验证集包含 40 504 张图片、214 354 个问题和 2 143 540 个答案. 测试集包含 81 434 张图片和 447 793 个问题. 数据集中的图片来源于 COCO 数据集.

4) CLEVR[67]:该数据集包含 10 万张经过渲染的图像和大约 100 万个自动生成的问题,其中有 85.3 万个问题是互不相同的. 其中包含了测试计数、比较、逻辑推理和在记忆中存储信息等视觉推理能力的图像和问题. 尽管 CLEVR 中的图像可能看起来很简单,但它的问题却很复杂,需要一系列的推理能力. 例如:归纳未见过的物体和属性的组合可能需要分解表征;计数或比较这样的任务可能需要短期记忆或关注特定的物体;以多种方式结合多个子任务的问题可能需要组合式系统来回答.

5）TGIF-QA[68]：该数据集包含 72 000 个的动画 GIF 文件和 165 000 个的问答对. 这个数据集提供了 4 种任务来处理视频的独特属性. 重复计数是检索一个动作的出现次数. 重复动作是一项任务，用于识别在多项选择中重复给定次数的动作. 状态转换是一项多项选择任务，用于根据动作状态的时间顺序确定动作. 帧定位是在视频中找到一个能回答问题的特定帧.

6）MSRVTT-QA[69]：该数据集包含 10 000 个视频和 243 000 个问答对. 这些问题由 5 种类型组成，包括 what、who、how、when 和 where. 视频的长度为 10 ~ 30 s.

4 方法对比

在表 1 和表 2 中分别对多种方法在 VQA-v1 和 VQA-v2 数据库上的准确率进行对比. 可以看出，数据融合的方法取得了初步的结果，基于跨模态的注意力的方法可以学习到更加精确的数据间的关联关系，准确率高于数据融合的方法. 基于知识推理的方法利用了推理的思路，经过多次迭代的推理计算来学习更加有效的信息，也取得了最好的结果.

表 1 VQA-v1 数据库上的准确率对比

Table 1 Accuracy comparison on VQA-v1 database

%

模型	验证集				测试集
	是/否	数量	其他	全部	全部
MCB[23]	83.40	39.80	58.50	66.70	66.50
MLB[24]	84.57	39.21	57.81	66.77	66.89
MUTAN[26]	85.14	39.81	58.52	67.42	67.36
DCN[50]	84.61	42.35	57.31	66.89	67.02

表 2 VQA-v2 数据库上的准确率对比

Table 2 Accuracy comparison on VQA-v2 database

%

模型	验证集				测试集
	是/否	数量	其他	全部	全部
MCB[23]					62.27
Bot-up[13]	81.82	44.21	56.05	65.32	65.67
DCN[50]	83.70	46.65	56.77	66.72	67.04
DFAF[51]	86.09	53.32	60.49	70.22	70.34
CRA-N[55]	84.87	49.46	59.08	68.61	68.92
BAN[49]	85.31	50.93	60.26	69.52	
MCAN[56]	86.82	53.26	60.72	70.63	70.90

5 结论

综上所述，目前 VQA 方法研究的核心问题有 2 点：视觉和文本数据的特征表示、多模态特征联合学习. 由于细粒度的特征表示可以提供丰富的细节语义信息，这一表示方法也取得了较好的效果. 然而，对于图像的特征表示还有不足之处，目前，还没有找到能够准确提取和表示图像语义信息的方法. 在多模态特征联合学习中，注意力机制发挥了重要作用，这一机制可以深度挖掘模态间和模态内信息之间的关联关系，因此，取得了较好的效果. 但是，注意力机制缺乏推理学习的能力，对于包含复杂语义信息的图像和文本信息其无法有效学习 2 种跨模态数据之间的语义关联关系. 面对这一问题，知识推理的方法通过多步迭代的推理学习对多模态的信息进行语义学习和关联学习，可以挖掘出深层次的关联信息.

本文结合现有的 VQA 方法，对未来的有潜力的研究方向进行展望.

1）在特征表示方面，研究者一直在探索图像的特征表示方法，在 VQA 中，图像的特征提取也是一个重要环节. 目前，基于卷积网络的网格特征和基于目标检测方法的区域特征均有所不足，这 2 种特征都无法充分保留全局语义信息和细粒度语义信息，在如何提取适用于 VQA 任务、精度高并包含细粒度语义信息的图像特征方面具有较大的研究价值. 在图像的特征表示过程中结合知识图谱进行结构化的特征提取和表示是一个值得探索的方向.

2）在跨模态特征学习方面，知识推理的方向具有较大的研究价值. 多年来，研究者都在探索知识的表示和推理学习的方法，人类面对复杂问题展现出强大的推理能力，通过推理分析得到解决办法. 在 VQA 中推理也非常重要，推理的方法可以对特征之间的复杂关系进行提取和建模. 结合知识图谱中的先验知识来解答真实场景中的 VQA 任务是一个有价值的研究方向. 如何利用跨模态的知识图谱对视觉特征和文本问题进行有效的推理计算具有较大的研究潜力.

参考文献：

[1] JIANG A, WANG F, PORIKLI F, et al. Compositional memory for visual question answering[EB/OL]. [2021-

03-18]. https://arxiv. org/abs/1511. 05676.
[2] NOH H, SEO P H, HAN B, et al. Image question answering using convolutional neural network with dynamic parameter prediction [C] // International Conference on Computer Vision. Piscataway: IEEE, 2016: 30-38.
[3] SAITO K, SHIN A, USHIKU Y, et al. DualNet: domain-invariant network for visual question answering [C] // International Conference on Multimedia and Expo. Piscataway: IEEE, 2017: 829-834.
[4] KAFLE K, KANAN C. Answer-type prediction for visual question answering [C] // Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 4976-4984.
[5] WANG P, WU Q, SHEN C, et al. Explicit knowledge-based reasoning for visual question answering [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1511. 02570.
[6] WANG P, WU Q, SHEN C, et al. FVQA: fact-based visual question answering [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1606. 05433.
[7] BORDES A, USUNIER N, CHOPRA S, et al. Large-scale simple question answering with memory networks [EB/OL]. [2021-03-18]. https://arxiv. org/abs/1506. 02075.
[8] XIONG C, MERITY S, SOCHER R, et al. Dynamic memory networks for visual and textual question answering [C] // International Conference on Machine Learning. Pittsburg: IMLS, 2016: 2397-2406.
[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances In Neural Information Processing Systems, 2012, 25: 1097-1105.
[10] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] // Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1-9.
[11] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1409. 1556.
[12] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] // Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.
[13] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C] // Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6077-6086.
[14] 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.
[15] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1506. 01497.
[16] GOLDBERG Y, LEVY O. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1402. 3722.
[17] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1301. 3781.
[18] WU S, MANBER U. Fast text searching: allowing errors [J]. Communications of the ACM, 1992, 35(10): 83-91.
[19] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation [C] // Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.
[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1810. 04805.
[21] ETHAYARAJH K. How contextual are contextualized word representations? comparing the geometry of BE-RT, ELMo, and GPT-2 embeddings [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1909. 00512.
[22] WARSTADT A, CAO Y, GROSU I, et al. Investigating BERT's knowledge of language: five analysis methods with NPIs [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1909. 02597.
[23] FUKUI A, PARK D H, YANG D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1606. 01847.
[24] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling [EB/OL]. [2021-03-18]. https: //arxiv. org/abs/1610. 04325.
[25] YU D, FU J, MEI T, et al. Multi-level attention networks for visual question answering [C] // Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 4709-4717.
[26] BEN-YOUNES H, CADENE R, CORD M, et al. MUTAN: multimodal tucker fusion for visual question answering [C] // International Conference on Computer Vision. Piscataway: IEEE, 2017: 2612-2620.
[27] 王俊毅，周小兵. 一种基于深度神经网络的停车位检测算法开发[J]. 上海汽车，2021(6)：20-27.
WANG J Y, ZHOU X B. A parking spot detection

algorithm based on deep neural network[J]. Shanghai Auto, 2021(6): 20-27. (in Chinese)

[28] ZHU Y, GROTH O, BERNSTEIN M, et al. Visual7W: grounded question answering in images[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 4995-5004.

[29] SHIH K J, SINGH S, HOIEM D. Where to look: focus regions for visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 4613-4621.

[30] CHEN K, WANG J, CHEN L C, et al. ABC-CNN: an attention based convolutional neural network for visual question answering[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1511.05960.

[31] ILIEVSKI I, YAN S, FENG J. A focused dynamic attention model for visual question answering[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1604.01485.

[32] ILIEVSKI I, FENG J. Generative attention model with adversarial self-learning for visual question answering[C]// ACM International Conference on Multimedia. New York: ACM, 2017: 415-423.

[33] PATRO B, NAMBOODIRI V P. Differential attention for visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7680-7688.

[34] KIM J H, LEE S W, KWAK D H, et al. Multimodal residual learning for visual QA[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1606.01455.

[35] SCHWARTZ I, SCHWING A G, HAZAN T. High order attention models for visual question answering[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1711.04323.

[36] XU H, ASK S K. Attend and answer: exploring question-guided spatial attention for visual question answering[C]// European Conference on Computer Vision. Zurich: ECVA, 2016: 451-466.

[37] KAZEMI V, ELQURSH A. Show, ask, attend, and answer: a strong baseline for visual question answering[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1704.03162.

[38] ZHAO Z, JIANG X, CAI D, et al. Multi-turn video question answering via multi-stream hierarchical attention context network[C]// International Joint Conference on Artificial Intelligence. Palo Alto: AAAI, 2018: 3690-3696.

[39] MUN J, SEO P H, JUNG I, et al. MarioQA: answering questions by watching gameplay videos[C]// International Conference on Computer Vision. Piscataway: IEEE, 2017: 2867-2875.

[40] YE Y, ZHAO Z, LI Y, et al. Video question answering via attribute-augmented attention network learning[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2017: 829-832.

[41] YANG Z, HE X, GAO J, et al. Stacked attention networks for image question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 21-29.

[42] OSMAN A, SAMEK W. Dual recurrent attention units for visual question answering[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1802.00209.

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

[44] MA C, SHEN C, DICK A, et al. Visual question answering with memory-augmented networks[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6975-6984.

[45] 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.

[46] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification[C]// International Conference on Computer Vision. Piscataway: IEEE, 2015: 1026-1034.

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

[48] LAO M, GUO Y, WANG H, et al. Cross-modal multistep fusion network with co-attention for visual question answering[J]. IEEE Access, 2018, 6: 31516-31524.

[49] KIM J H, JUN J, ZHANG B T. Bilinear attention networks[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1805.07932.

[50] NGUYEN D K, OKATANI T. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6087-6096.

[51] GAO P, JIANG Z, YOU H, et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 6639-6648.

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

[53] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1606.00061.

[54] TENEY D, LIU L, VAN DEN HENGEL A. Graph-structured representations for visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1-9.

[55] PENG L, YANG Y, WANG Z, et al. CRA-Net: composed relationattention network for visual question answering[C]// ACM International Conference on Multimedia. New York: ACM, 2019: 1202-1210.

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

[57] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1706.03762.

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

[59] CADENE R, BEN-YOUNES H, CORD M, et al. MUREL: multimodal relational reasoning for visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 1989-1998.

[60] PEREZ E, STRUB F, DE VRIES H, et al. FiLM: visual reasoning with a general conditioning layer[C]// AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2018, 32(1): 3942-3952.

[61] HUDSON D A, MANNING C D. Compositional attention networks for machine reasoning[EB/OL]. [2021-03-18]. https://arxiv.org/abs/1803.03067.

[62] MARINO K, CHEN X, PARIKH D, et al. KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA[EB/OL]. [2021-03-18]. https://arxiv.org/abs/2012.11014.

[63] HUANG D, CHEN P, ZENG R, et al. Location-aware graph convolutional networks for video question answering[C]// AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020, 34(7): 11021-11028.

[64] JIANG J, CHEN Z, LIN H, et al. Divide and conquer: question-guided spatio-temporal contextual attention for video question answering[C]// AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020, 34(7): 11101-11108.

[65] ANTOL S, AGRAWAL A, LU J, et al. VQA: visual question answering[C]// International Conference on Computer Vision. Piscataway: IEEE, 2015: 2425-2433.

[66] 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]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6904-6913.

[67] JOHNSON J, HARIHARAN B, VAN DER MAATEN L, et al. CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2901-2910.

[68] JANG Y, SONG Y, YU Y, et al. TGIF-QA: toward spatio-temporal reasoning in visual question answering[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2758-2766.

[69] XU J, MEI T, YAO T, et al. MSR-VTT: a large video description dataset for bridging video and language[C]// Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 5288-5296.

（责任编辑　梁　洁）