计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220294
61(3): 639−673, 2024

视频问答技术研究进展

包翠竹[1,3] 丁 凯[2] 董建峰[1,3] 杨 勋[4] 谢满德[2] 王 勋[1,3]
[1](浙江工商大学计算机科学与技术学院 杭州 310018)
[2](浙江工商大学信息与电子工程学院 杭州 310018)
[3](浙江省电子商务与物流信息技术研究重点实验室(浙江工商大学) 杭州 310018)
[4](中国科学技术大学信息科学技术学院 合肥 230026)
(baocuizhu@ zjgsu.edu.cn)

Research Progress of Video Question Answering Technologies

Bao Cuizhu[1,3], Ding Kai[2], Dong Jianfeng[1,3], Yang Xun[4], Xie Mande[2], and Wang Xun[1,3]
[1](*School of Computer Science and Technology*, *Zhejiang Gongshang University*, *Hangzhou* 310018)
[2](*School of Information & Electronic Engineering*, *Zhejiang Gongshang University*, *Hangzhou* 310018)
[3](*Zhejiang E-Commerce Key Lab*(*Zhejiang Gongshang University*), *Hangzhou* 310018)
[4](*School of Information Science and Technology*, *University of Science and Technology of China*, *Hefei* 230026)

Abstract VideoQA (video question answering), which automatically answers natural language question according to the content of videos, is a relatively new research direction in the field of visual language and has attracted extensive attention in recent years. The solution of videoQA task is of great significance for human-computer interaction, intelligent education, intelligent transportation, scenario analysis, video retrieval, and other fields. VideoQA is a challenging task because it requires a model to understand semantic information of the video and the question to generate the answer. In this work, we analyze the difference between VideoQA and ImageQA (image question answering), and summarize four challenges faced by VideoQA relative to ImageQA. Then, the existing VideoQA models are carefully classified according to the research method around these challenges. Following the classifications, we introduce the generation background and focus on the implementation of models and the relationship between different models. After that, the benchmark datasets commonly used in VideoQA are summarized, the performances of current mainstream algorithms on some datasets are introduced in detail, and the comparison, analysis and summary are carried out. Finally, the future challenges and research trends in this field are discussed, which will provide some ideas for further research in the future.

Key words video question answering; attention; memory network; recurrent neural network; graph network model; pre-training model

摘 要 视频问答(video question answering, VideoQA)根据视频内容自动回答自然语言问题,是视觉语言领域较为新兴的一个研究方向,近年来引起了广泛关注. VideoQA 问题的解决对于人机交互、智慧教育、智能交通、场景分析以及视频检索等各个领域都有着重大意义. VideoQA 是一项具有挑战性的任务,因为它需要模型同时理解视频与文本内容来生成问题的答案. 首先,分析了 VideoQA 与图像问答(image question

收稿日期: 2022-04-12; **修回日期**: 2023-06-07
基金项目: 国家自然科学基金项目(61972352, 61902347, 61976188, 62272435, U22A2094);浙江省重点研发计划项目(2021C03150);浙江省省属高校基本科研业务费专项
This work was supported by the National Natural Science Foundation of China (61972352, 61902347, 61976188, 62272435, U22A2094), the Key Research and Development Program of Zhejiang Province (2021C03150), and the Fundamental Research Funds for the Provincial Universities of Zhejiang.
通信作者: 谢满德(xiemd@zjgsu.edu.cn)

answering, ImageQA) 的区别, 总结了当下 VideoQA 相对于 ImageQA 所面临的 4 个挑战; 然后, 围绕着这些挑战对目前现有 VideoQA 模型进行了细致的分类, 并重点介绍了模型的实现及不同模型之间的关联; 接着详细介绍了在 VideoQA 中常用的基准数据集及目前主流算法在部分数据集上的性能, 并进行了对比与分析; 最后, 讨论了该领域未来面临的挑战和研究趋势, 为未来进一步研究提供一些思路.

关键词 视频问答; 注意力; 记忆网络; 循环神经网络; 图网络模型; 预训练模型

中图法分类号 TP391

图灵测试至今依然是评判机器是否具有人类智慧的重要手段, 它代表着机器对人类知识体系或交互的理解程度, 是人工智能的一个充分条件. 近些年来, 为了让机器像人类一样思考与交互, 研究者们一直热衷于类似图灵测试相关人工智能系统的研究, 如问答系统. 随着问答系统在自然语言处理的成功, 人们开始考虑将这种交互式的问答方式引入计算机视觉领域, 对视觉对象进行交互式理解[1]. 在此背景下, 基于图像的 ImageQA 在 2015 年被提出[2], 并受到广泛关注[3-5]. 而 VideoQA 则是 ImageQA 的自然扩展, 即将 ImageQA 的单幅图像延伸为图像序列, VideoQA 任务应运而生. VideoQA 可以被认为是一种视觉图灵测试, 它也需要理解问题的能力, 但不一定需要更复杂的自然语言处理. 一个优秀的 VideoQA 模型能够解决诸多计算机视觉相关方面的问题, 因此它可以作为图灵测试的重要组成部分. 由于生活中大多数数据均为动态视觉信息, VideoQA 系统在实际场景中的应用更为广泛, 基于视频的对话问答已逐渐成为人机交互的重要方式. 这项技术的发展对智慧教育、智能交通以及视频取证等方面均有着积极的影响, 尤其能极大程度地帮助视障人士在网络与现实世界获取有用的视觉信息[6].

据当前调研所知, 国内单纯的 VideoQA 方向的综述文章未见发表, 与之相关的 ImageQA 综述如文献 [1] 对 VideoQA 任务的难点只做了简单分析, 文献 [7] 针对 ImageQA 任务提出的部分挑战同样适用于 VideoQA 任务. 国外鲜有报道, 在目前发表的几篇综述论文中, Patel 等人[8] 对目前常用的数据集和主流方法进行分类与介绍, 但仅仅是简单罗列, 并没有分析各方法之间的关系. Khurana 等人[9] 对主流方法的分类总结比文献 [8] 更为全面, 然而该文提到的方法与数据集不够完善. Sun 等人[10] 对现有 VideoQA 任务的数据集与方法的归纳与总结相对完善, 对模型分析比较透彻, 但近 2 年新出现的方法与数据集也有待完善. 相对于国外的这 3 篇综述, 本文除了回顾基于注意力机制[11] 的方法与记忆网络 (memory networks, MemNN)[12] 的方法之外, 还将近 2 年新出现的基于图网络, 如图神经网络[13](graph neural network, GNN) 与图卷积神经网络[14] (graph convolutional network, GCN) 的方法、基于预训练的方法以及基于 Transformer[15] 与 BERT[16] 的方法进行了总结与分析; 同时, 将当下绝大多数用于 VideoQA 数据集各项指标进行了收集, 并对常用数据集的模型性能进行汇总与分析. 通过表 1 进行对比, 可以看出本文无论是数据集还是方法介绍均是目前最全面的.

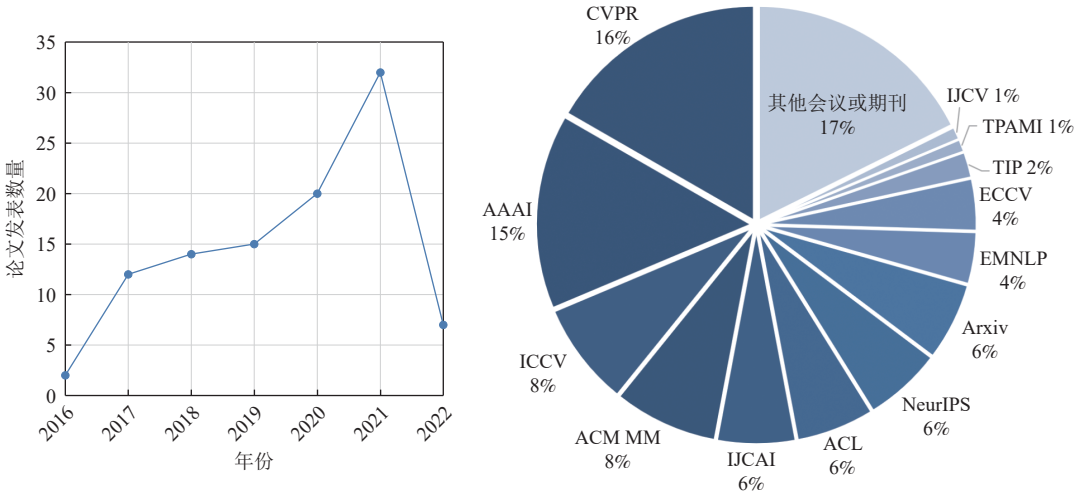
除此之外, 本文的 VideoQA 研究进展调研力求提供一个全面且系统的综述工作, 收集了计算机视觉、自然语言处理、多媒体和机器学习等诸多领域的知名会议与期刊论文, 如 CVPR, ICCV, ECCV, AAAI, NeurIPS, IJCAI, ACM MM, ACL, EMNLP, TPAMI, TIP, IJCV 等. 这些论文主要发表于 2017—2021 年, 对于 2022 年 1~3 月发表的论文也进行了统计, 如图 1 所示. 本综述具体行文的组织结构如图 2 所示.

Table 1 Comparison of VideoQA Survey Works

表 1 VideoQA 综述工作对比

综述工作	数据集		方法						
	最新年份	个数	最新年份	个数	注意力	记忆网络	图网络	Transformer/BERT	预训练
Patel 等人[8]	2020	18	2020	26	√	√	×	×	×
Khurana 等人[9]	2019	11	2020	22	√	√	√	×	×
Sun 等人[10]	2019	11	2020	41	√	√	×	×	×
本文	2022	30	2022	83	√	√	√	√	√

注: “√”表示使用; “×”表示未使用.



(a) 每年发表的论文数量 (b) 按会议/期刊的论文分布

Fig. 1 Paper statistics

图 1 论文统计

1 VideoQA 概述

本节主要介绍 VideoQA 问题定义、VideoQA 相较与 ImageQA 的区别及挑战以及近几年 VideoQA 的发展趋势.

1.1 问题定义

VideoQA 的目标是根据一个视频$\mathcal{V}$和与之相关的问题q推断出问题的答案$\tilde{a}$. VideoQA 模型可以表述为:

$$\tilde{a}=\arg\max_{a\in\mathcal{A}}\mathcal{F}_\theta(a|q,\mathcal{V}), \tag{1}$$

其中$\mathcal{F}$为评分函数, θ为模型参数, $\mathcal{A}$是一个可能存在答案的集合.

VideoQA 任务中的问题可以分为开放式问题、选择题和填空题 3 种类型. 开放式问题类型如图 3(a)所示, 问题没有候选答案; 选择题类型如图 3(b) 所示, 1 个问题对应多个候选答案, 候选答案中只有 1 个是正确的; 填空题类型如图 3(c) 所示, 1 句话中缺少 1 个词, 需要模型进行填充. 其中, 填空题可以视为一种特殊的开放式问题, 在后文 2.3 节将其视为开放式问题.

1.2 VideQA 与 ImageQA 的区别及挑战

VideoQA 是一项结合计算机视觉与自然语言处理的任务, 该任务根据问题来分析视频内容并得出答案. VideoQA 由 ImageQA 发展而来, ImageQA 任务中的模型根据给定的一幅图片及自然语言的问题, 以自然语言的形式给出答案, 如图 4 所示. 相较于 ImageQA, VideoQA 将图片替换成视频, 从静态的图片变成了动态连续的帧序列. 所以, 不能简单应用原有的 ImageQA 模型, 而是需要在原有的基础之上对帧序列进行动作分析和时序建模, 并同时考虑到视频中的多模态信息, 因此 VideoQA 相对而言更具有挑战性.

具体来讲, VideoQA 相对于 ImageQA 主要有 2 方面的区别: 一方面, 对于 ImageQA 来讲, 图片是静态的, 包含的信息 (颜色、对象及其空间位置等) 更容易通过模型进行提取与分析, 所以模型只需要充分理解单幅图片上的所有信息, 外加少量辅助知识, 就能够在回答针对单幅图像的问题上取得具有竞争性的结果. 相比于 ImageQA, VideoQA 模型处理的是连续变化的动态视频帧信息, 包含更丰富多变的信息 (外观信息、音频信息、动作与状态转换等). 另一方面, ImageQA 大多是关于对象外观的帧级问题, 主要关注图片的对象属性, 例如颜色或者空间位置等, 推理部分相对较少. 然而 VideoQA 中大多是关于对象行为与因果关系的问题, 此类问题的回答就需要模型具有更强的对上下文建模和因果推理的能力. 同时, 视频中包含许多时间线索, 大部分问题也是关于视频的时序推理, 比如状态转换、动作计数等, 所以 VideoQA 模型又需要良好的时序建模能力.

从上述 VideoQA 与 ImageQA 的对比分析中, 可以总结 VideoQA 任务主要面临 4 方面挑战:

1) 基于问题的关键帧定位

视频通常每秒包含多帧图片, 必然存在大量问题无关的冗余信息, 这种冗余会干扰模型的推理过程. 如何保证从大量复杂信息中根据问题定位到预测答案所需的关键时刻信息是至关重要的. 突出与

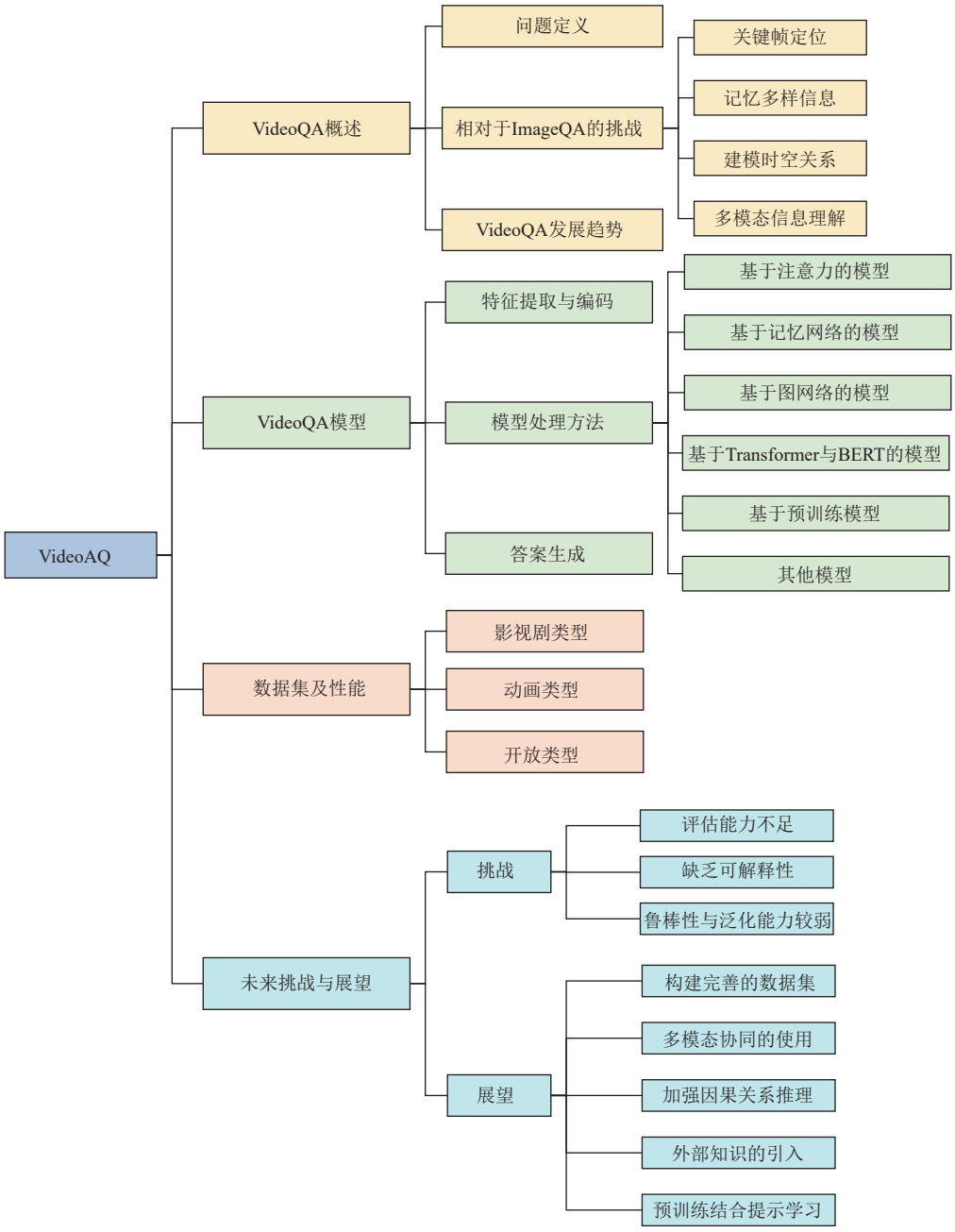


Fig. 2 The overview of our paper

图 2 本文的概述

问题相关的重要信息是进行准确推理的关键.

2)丰富多样的信息需要记忆

为了更准确地回答有关动作、因果等相关的复杂问题,模型无论在数量上还是在多样性上都需要记忆更长的信息,而信息的完整性是进行推理的前提.

3)基于帧序列的时空建模

视频作为图像序列,不仅包含静态属性,如颜色、位置,还包含更多时空相关的动态属性,如动作和状态转换. 丰富多样化的信息意味着推理过程更加复杂,需要根据问题整合时空 2 个维度的信息进行综合推理. 多维度信息可以互相增强,对于时空推理的问题起到很大帮助.

4)多模态信息的语义理解

多模态信息体现在视频包含多个类型的媒体数据,如图像、语音、标题和字幕等. 这些数据的交叉使得视频具有更复杂的语义信息,也对模型获取更好的多模特征表示提出了更高的要求.

应对这 4 个挑战的关键在于如何建立视频和问题之间的语义联系及其在时序上的连续性与关联性,其关键是将视频特征和文本特征进行处理与融合.



1) 开放式问题
问题: what is a man with long hair and a beard is playing ?
答案: guitar
2) 选择题
问题: what is a man with long hair and a beard is playing ?
答案: a0 piano
a1 guitar √
a2 Harmonica
a3 violin
3) 填空题
问题: A man with long hair and a beard is playing the _____.
答案: guitar

Fig. 3 Examples of various types of questions

图 3 各类型问题示例

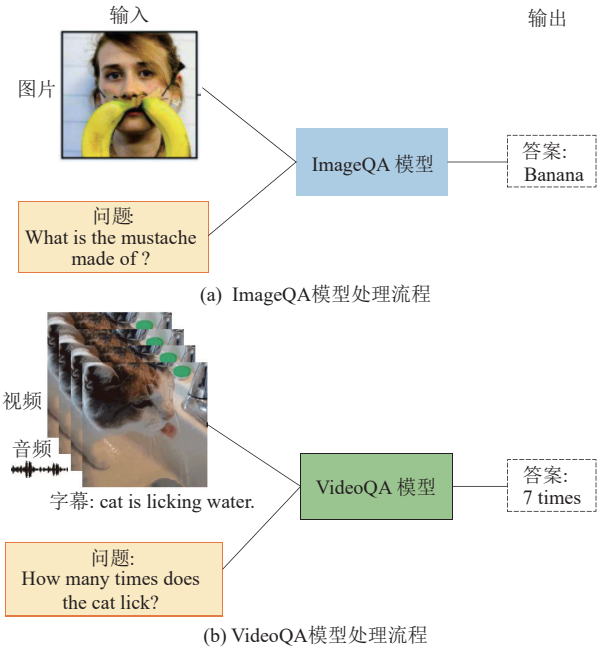


Fig. 4 Comparison of VideoQA and ImageQA models

图 4 VideoQA 与 ImageQA 模型对比

根据模型处理方法的不同，本文将 VideoQA 模型分为基于注意力的模型、基于记忆网络的模型、基于图网络的模型、基于预训练的模型、基于 Transformer 与 BERT 的模型及其他模型.

1.3 VideoQA 发展趋势

总体来讲，VideoQA 模型由最初的注意力和记忆网络发展到目前流行的图网络、预训练及 Transformer 与 BERT，代表性模型的年历表概览如图 5 所示. 早期的 VideoQA 模型 (2017—2019) 为了去除视频中大量问题无关的冗余信息，多数采用了注意力机制，此类方法根据问题有效地提取视频的关键信息以用于答案推理. 同时，为了保证模型能够从视频中更好地挖掘多个时间帧信息的关联性，保证信息的完整性，许多模型引入了记忆网络. 2020 至今，鉴于图神经网络在关系建模和推理方面的优异表现，部分研究者开始将其应用于 VideoQA 任务中以更好地建模对象关系信息，并且挖掘丰富的时-空相关的动态属性.



Fig. 5 Overview of the mainstream VideoQA model almanacs

图 5 主流的 VideoQA 模型年历表概览

除此之外，随着 Transformer 与 BERT 模型在自然语言处理领域所展现的出色性能，越来越多的研究者尝试将其引入自己的 VideoQA 模型 (包括预训练模型) 中去. Transformer 与 BERT 改进了循环神经网络 (recurrent neural network, RNN) 训练慢的缺点，利用自注意力机制实现快速并行，并且可以增加到非常深的深度，充分挖掘深度神经网络模型的特性，提升模型准确率. 与此同时，视觉语言预训练模型在近

2年也展现了在海量互联网数据中学习视觉-文本联合表征的强大能力，这种由数据驱动的预训练模型的性能在下游的VideoQA任务中也崭露头角.

2 VideoQA模型

图6为VideoQA模型的详细处理流程. VideoQA模型处理主要分为视频与文本的特征提取与编码、多模态特征处理与融合以及答案生成3个部分. 特征提取与编码又分为视频和文本的特征提取与编码，主要介绍目前通用的技术，该部分内容将在2.1节进行介绍；对特征的处理与融合是VideoQA的核心与关键，也是研究者们围绕该任务展开研究的主要方面，每一个方法的不同之处也均体现在这一部分，该内容将在2.2节进行重点介绍；答案生成部分对于不同的问答任务已经形成了相对固定的答案解码方式，该部分将在2.3节进行综合介绍.

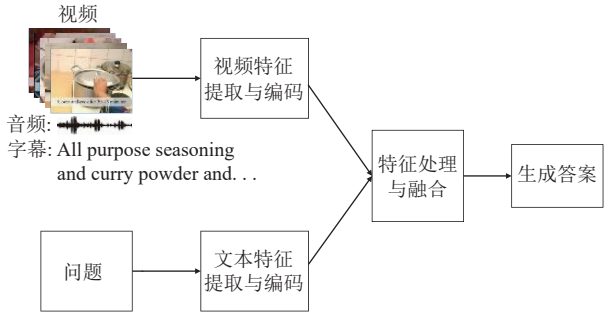


Fig. 6 VideoQA model processing flow

图6 VideoQA模型处理流程

2.1 特征提取与编码

2.1.1 视频特征

视频的视觉特征提取不仅包含静态的区域级(对象)特征和帧级特征提取，还包含动态的片段级特征提取. 区域级特征是对局部信息的细粒度表示，可以表示为对象特征及其标签，此类特征一般使用目标检测网络进行提取，如Faster R-CNN[17]. 帧级视频特征是对全局视觉信息的粗粒度表示，较之于区域级特征可以获取更丰富的信息，如场景信息等. 目前帧级特征常直接应用ImageQA模型中的图片特征提取方法，即在ImageNet[18]上预训练的2D卷积神经网络，如VGGNet[19], GoogLeNe[20], ResNet[21]. 片段级视频特征，是顺序的和动态的特征表示，如动作. 片段级特征早期采用用于动作识别的3D卷积网络，从空间和时间维度中提取特征，从而捕获在多个相邻帧中编码的动作信息. 随着C3D[22]网络在动作识别和捕捉视频动态信息中展现出优异性能，C3D成为主流的动作特征提取方法，后续也有方法在C3D的基础上进一步地优化，如双流膨胀3D卷积网络I3D[23]等与基于时域和空间域分离的S3D[24]网络. 除此之外，ResNeXt[25]与SlowFast[26]等也用于3D动作识别. 除了视觉特征以外，视频中音频模态信息也被用来作为视觉特征的补充[27-28]. 比如对于唱歌、演讲类型的视频，音频信息对于视频中事件的理解可以起到很好的补充作用. 针对VideoQA中的音频特征的提取，常见的有手工设计的传统特征梅尔频谱以及基于深度神经网络的SoundNet[29]与WALNet[30]. 此外，视频中的标题和字幕包含了丰富的语义信息，通常也用作为视频的文本模态信息的建模. 由于标题或字幕特征提取与问题特征提取类似，其特征提取方式将在2.1.2节文本特征提取中介绍.

视频在输入时包含视觉、字幕与音频多种模态信息，模型对其进行融合方式也不尽相同. 简单地操作实现不同模态的特征向量的整合，例如部分模型使用点乘、点加、拼接和加权求和[31]，然而简单操作使得参数之间几乎没有联系，这种方式的联合向量表达能力明显不足. 所以有模型使用双线性池化融合视频多模态特征向量来获得一个联合表征空间[32-33]，其方法是计算两者的外积，通过将外积生成的矩阵线性化成一个向量表示，这意味着该方法更具有表现力. 目前，很多模型利用注意力机制来融合视频多模态信息[34-35]，多模态特征向量通过注意力操作可以动态产生求和时要用到的权重，特征融合时可以保存额外的权重信息，起到信息互补的作用. 此外，由于Transformer全自注意力的结构天生就具备处理不同模态数据的能力，逐渐成为主流的多模态融合的方法. 其主要分为单流模型和多流模型，单流模型将视觉和文本的嵌入拼接到一起，输入到一个Transformer中；而多流模型让视觉和文本使用2个或多个独立的Transformer分别编码[36-37]，并且可在中间层加入视觉和文本2个模态之间互注意力来融合多模态信息.

2.1.2 文本特征

文本特征提取包括视频中字幕的文本特征提取与问题的文本特征提取，文本特征提取又可以分为单词级特征与句子级特征. 对于单词级特征，主要采用包括Word2Vec[38]和GloVe[39]的词嵌入来提取单词级特征. 而对于句子级特征，主要采用Skip-Thought[40]和BERT来提取句子级特征. 其中BERT是一种经过微调的基于Transformer的语言模型，它能够捕捉双向上下文信息，用以在不同的句子级别任务中预测句子，目前已成为主流的文本处理方案.

视频中的视觉和文本都是序列数据，因此在VideoQA任务中，许多研究者使用基于RNN的模型来编码视频的帧级特征与文本的单词级特征，来获取视频片段级特征和问题级特征. 常被使用的经典RNN结构包括长短期记忆 (long short term memory, LSTM)[41] 编码器和门控递归单元 (gated recurrent unit, GRU)[42]，双向LSTM(Bi-LSTM)和双向GRU(Bi-GRU)均为前两者的变体.

2.2 模型处理方法

2.2.1 基于注意力的模型

注意力机制广泛应用于自然语言处理和计算机视觉领域，目前已经成为神经网络结构的重要组成部分. 常见的注意力机制包括点积模型、双线性模型以及近几年流行的自注意力模型和多头注意力模型等. 在VideoQA任务中可将问题作为查询，将视频作为源，那么注意力机制就是用来定位视频中与问题相关的信息. 由此，注意力的实质可以看作一个查询到一系列键值对的映射. 如图7所示，注意力的计算主要分为3阶段：第1阶段是将查询和每个键进行相似度计算得到权重，常用的相似度函数有点积、拼接、感知机等；第2阶段一般是使用一个softmax函数对这些权重进行归一化；第3阶段将权重和相应的键值进行加权求和得到最终的注意力.

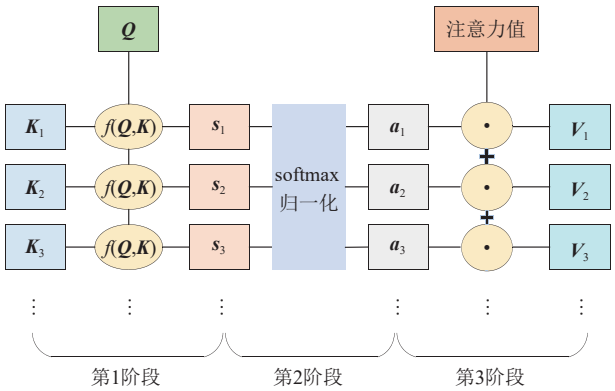


Fig. 7 Three stages of attention calculation

图7 注意力计算的3个阶段

在VideoQA任务中，键和值通常是同一个. 因此，常见的注意力计算过程可以描述为3种计算方式：

$$s_i = f(\boldsymbol{Q}, \boldsymbol{K}_i) = \begin{cases} \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}_i, & \text{点积} \\ \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{a}}\boldsymbol{K}_i, & \text{双线性} \\ \boldsymbol{W}_{\mathrm{a}}[\boldsymbol{Q};\boldsymbol{K}_i], & \text{连接} \\ \dfrac{\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}_i}{\sqrt{d_k}}, & \text{缩放点积} \\ \boldsymbol{v}_{\mathrm{a}}^{\mathrm{T}}\tanh(\boldsymbol{W}_{\mathrm{a}}\boldsymbol{Q}+\boldsymbol{U}_{\mathrm{a}}\boldsymbol{K}_i), & \text{感知机} \end{cases} \tag{2}$$

$$\boldsymbol{a}_i = softmax(\boldsymbol{s}_i) = \frac{\exp(\boldsymbol{s}_i)}{\sum_i \exp(\boldsymbol{s}_i)}, \tag{3}$$

$$\boldsymbol{c} = Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \sum_i \boldsymbol{a}_i\boldsymbol{V}_i. \tag{4}$$

其中$\boldsymbol{s}_i$是计算的注意力得分，代表查询与键之间的相似性；$\boldsymbol{a}_i$是$\boldsymbol{s}_i$经过归一化得到的注意力分布，实质为概率分布；$\boldsymbol{c}$是由权重与值进行加权求和得到的最终注意力向量. 从注意力计算过程可以看出，注意力机制本质相当于资源再分配机制，对原资源根据对象重要程度重新分配资源，所分配的资源其实就是权重. 本文将基于注意力机制的模型分为：单跳注意力[43-54]、多跳注意力[33, 55-61] 和多模态注意力[62-74].

1）单跳注意力

单跳注意力模型是指以问题为查询，对视频的视觉特征只执行1次注意力计算，视频的视觉特征包含区域级、帧级和片段级特征.

Zhao等人[43] 提出了双层注意力网络 (dual-level attention network, DLAN) 模型. 该网络基于帧级与片段级的视频特征分别利用词级与问题级注意力机制来学习问题的联合视频表征. 尽管DLAN利用细粒度词级注意力来增强视频表示，然而它忽略了词级语义，不同的单词需要不同程度的注意力，甚至有些词并不需要注意. Xue等人[44] 提出异构树型网络，该网络通过问题中的词来构建语义树，并根据单词词性对树中的词进行处理，区分视觉词和语言词，使注意力计算更为合理. 此外，与DLAN模型分层思想不同，Jang等人[45] 提出的时空视频问答 (spatio-temporal VQA, ST-VQA) 模型基于时空注意力机制来突出重要的区域与重要的帧，使用2个双层LSTM来挖掘视频视觉内容与问答文本内容之间的关系. Falcon等人[46] 对ST-VQA模型的帧特征提取做了微调并设计使用了3种数据增强技术，分别为重采样、镜像和水平翻转. Mazaheri等人[47] 提出基于分层时空注意的模型更加关注文本编码的网络，使用2个独立的LSTM分别对填空题句子空缺处的左右片段进行并行编码及反编码. Xu等人[48] 提出利用粗粒度问题特征和细粒度词特征来逐步细化注意力的方法. 对于给定视频，该模型以问题的词级特征作为引导，在每个时间步上通过设计的注意力单元 (AMU) 对帧级外观特征与片段级动作特征进行关注. 除了利用分层机制对视频与问题进行建模，Chao等人[49] 认为视频中的对话具有多层上下文关系，从分层注意力角度对视频与文本进行时空注意力机制学习. 而Zhao等人[50] 从自适

应分层增强编解码网络 (AHRN) 学习的角度来对视频内容进行建模. 自适应编码网络根据其设计的二进制门函数对视频进行分割, 然后利用注意力机制在问题的引导下学习相关帧与片段的联合表示, 生成问题感知视频表示.

Kim 等人[51] 从多任务学习的角度来解决 VideoQA 任务, 提出问题引导下的视频与字幕匹配任务和时间定位任务作为 VideoQA 的辅助任务. 与文献 [51] 设计思想相似, Lei 等人[52] 提出基于证据的时空答案 (spatio-temporal answerer with grounded evidence, STAGE) 模型在空间与时间维度上进行监督训练, 用于辅助主任务. 这种额外的辅助监督学习在一定程度上可以弥补模型在小规模数据集上监督学习的不足. 然而 STAGE 模型主要是在时间维度上提取具有预设间隔的网格级特征. 相比之下, 为了更好地执行多事件时间推理, Gao 等人[53] 提出了时序分割与事件注意力网络模型, 该模型利用设计的算法将视频分割成事件级片段表示, 然后利用注意力机制来定位给定问题的关键事件并输出答案. 事件级片段表示使得模型在多变环境中更容易定位到关键事件.

2) 多跳注意力

多跳注意力机制在视频上进行迭代注意力计算, 本次注意力计算结果作为下次注意力计算的输入. 相对于单跳注意力计算, 多跳注意力计算可以逐渐细化问题对于视频的注意, 以逐步引导注意到答案的正确位置. 假设 k–1 次的注意力计算结果为 $h_{z_{k-1}}(\boldsymbol{Q},\boldsymbol{V})$, 多跳计算过程可以表示为

$$y_k = y_{k-1} + h_{z_{k-1}}(\boldsymbol{Q},\boldsymbol{V}), \tag{5}$$

$$y_0 = q. \tag{6}$$

使用问题 q 初始化 y_0, 问题与视频产生的注意力与问题级联, 作为新的问题特征与视频特征再次进行注意力计算, 如此迭代计算以更新 y_k. 使用最终更新后的问题特征与视频特征产生最后的问题引导的视频注意力.

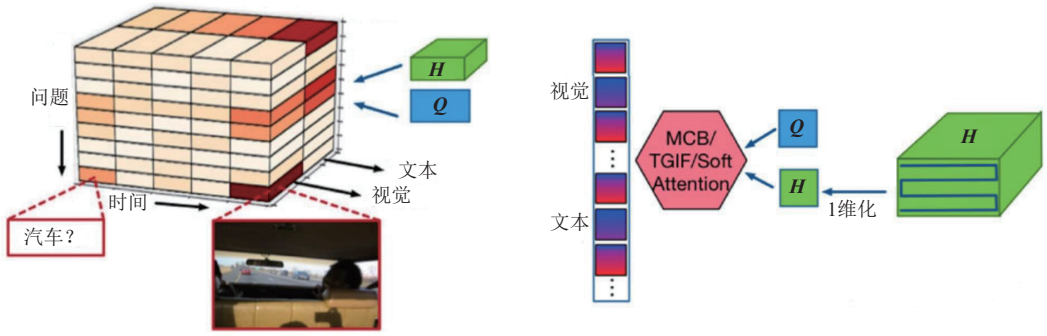
基于对象属性在视觉理解任务中的有效应用, Ye 等人[55] 提出了一种基于属性增强的注意网络模型. 该模型利用对象属性来增强视频表示, 然后引入多步推理过程, 对视频进行多跳注意力计算. 在文献 [55] 中强调了对象属性对于 VideoQA 任务的重要性, 但没有很好地利用问题中的每个词与视频每个部分的关系. 为了能够根据问题找到视频不同部分之间的上下文关系, Chowdhury 等人[56] 提出分层关系注意力模型. 该模型在每个时间步上以问题的每个词嵌入和视频的动作与外观特征作为注意力模块的输入, 注意力模块的输出与问题编码后的特征一块传递到关系模块, 其注意力模块借鉴于文献 [48]. Zhao 等人[57] 提出了多流分层注意力上下文网络, 和文献 [56] 中的问题与视频不同部分具有上下文关系类似, Zhao 等人认为对话具有双层顺序关系, 所以使用层次注意力上下文网络对其进行分层建模, 与问题特征结合生成上下文感知问题表征.

为了同时利用视频的空间特征和时序信息, 部分方法采用注意力机制对时-空信息进行建模. Zhao 等人[58] 提出了一个分层时空注意网络模型 r-STAN, 该模型根据目标对象与问题分别从空间层次与时间层次联合学习关键帧的视觉特征表示, 并且在网络中加入了多步推理过程来进一步提升模型性能. Song 等人[59] 利用空间注意力完成多重逻辑推理操作, 利用时间注意力捕捉长时间依赖并收集完整的视觉线索. 其中时间注意力模块使用的是经过改良的 GRU, 称为 ta-GRU(temporal-attention GRU), 将时间注意力与其隐藏的状态转移过程关联起来, 通过捕捉长时间依赖性, 获取更完整的时序视觉线索. 此外, Jiang 等人[60] 提出问题引导时空上下文注意的网络 (question-guided spatio-temporal contextual attention network, QueST) 模型. 该模型从时间和空间 2 个维度引入视觉信息对问题信息进行协同建模, 然后从 2 个维度挖掘与问题相关的视觉信息.

区别于以往文献 [45, 48] 中提出的方法, 将外观与动作分别做单独处理, Yang 等人[33] 提出了问题感知管道交换网络 (tube-switch network, TSN), 其注意力模块是基于多模态分解双线性池 (multi-modal factorized bilinear pooling, MFB)[62], 可以对外观和动作进行同步注意力机制, 而更新模块可以逐步细化多层 TSN 中的关注点, 交换模块则根据问题在每个推理步骤中自适应地选择外观或运动管道作为主特征, 另一个特征作为支撑特征用于丰富主特征, 指导多步推理过程中的注意力细化.

现有模型对于 VideoQA 任务只提供答案, 并未针对答案提供可解释的依据. Liang 等人[61] 提出了一种聚焦视觉与文本注意力 (focal visual-text attention, FVTA) 模型, 该模型可以预测答案的同时, 给出视觉和文本证据以解释推理过程. 图 8 展示了 FVTA 与传统注意力的区别, 模型从问题、文本与视觉 3 个维度进行关注, 应用于 3 维张量, 而一般的注意力模型应用于矢量或矩阵. 该模型的核心在于视觉与文本的注意力层, 其在每个时间步上进行多跳注意, 跨越多

个序列，充分利用了多时间步、多序列的特质. FVTA 注意力机制特有的性质使其既考虑了视觉与文本序列的内相关性，又考虑了交叉序列的相互作用，保留了序列数据中的多模态特征表示而不丢失重要信息.



(a) FVTA (b) MCB/TGIF/Soft注意力

Fig. 8 Comparison of FVTA and traditional attention[61]

图 8 FVTA 和传统注意力的比较[61]

3）多模态注意力

VideoQA 模型需要处理的数据包含多种模态，模态内与模态间往往存在很多关联信息. 对于模态内的关系，使用由注意力机制演变而来的自注意力来挖掘自身内部信息特征得到相关性；对于模态间的关系，通常使用共同注意力机制来获取，例如问题引导视频注意力与视频引导问题注意力. 此外，对于包含字幕的视频，还将存在字幕引导的问题注意力与问题引导的字幕注意力. 模型通常包含但不仅限于这 2 种注意力机制，对于注意力模块的输入可以按需调整.

Xue 等人[63] 提出了一种基于视频与问题共同注意力机制的方法. 该方法提出的注意力机制分为问题引导的视频注意力与视频引导的问题注意力，以及将两者进行整合的统一注意力. 与文献 [63] 类似，Chu 等人[64] 进一步提出了重看与重读机制，实质上也是视频与问题的共同注意力. 两者组合的遗忘观察模型为共同注意力模型，更好地利用了视频的时间信息和答案的短语信息. Gao 等人[65] 提出了一种结构化双流注意力网络 (structured two-stream attention network, STA) 的模型，不同的是该模型由多层共同注意力网络组成.

然而由于视频的复杂性与时序性，仅仅将共同注意力机制应用到 VideoQA 中往往效果很差. 针对这一问题，Li 等人[66] 提出具有多样性学习的可学习聚合网络 (learnable aggregating net with diversity learning, LAD-Net)，该网络使用独特的多路径金字塔式共同注意力机制. 多样性学习是为了处理视频复杂的特征，将视频特征与问题特征以不同维度的特征表示进行多次共同注意力学习，再利用惩罚机制进行多样性学习.

共同注意力机制能够较好地捕获了模态之间的关系，而对于模态内关系的内部依赖刻画不足. Li 等人[67] 将自注意力机制与共同注意力机制结合，提出了位置自注意力和共同注意力 (positional self-attention with co-attention, PSAC) 模型，PSAC 模型结构如图 9 所示. 位置自注意力通过关注同一序列中的所有位置，然后添加绝对位置的表示来计算每个位置的响应. 同时利用共同注意力机制以使模型能够同时考虑相关的视频和文本特征，从而消除了不相关的视频和文字信息，确保了正确答案的生成.

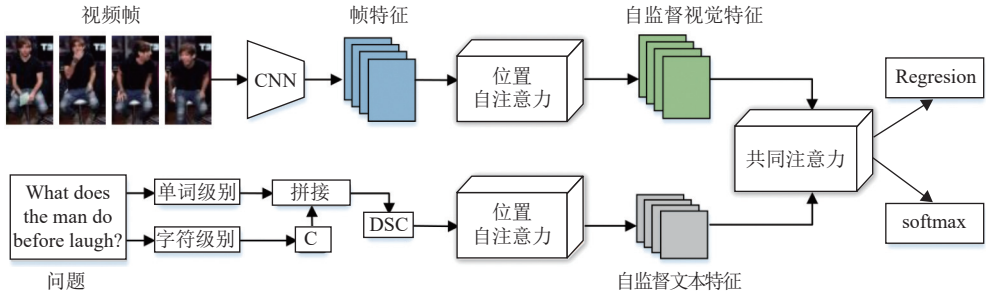


Fig. 9 The structure of PSAC model [67]

图 9 PSAC 模型结构[67]

部分方法不仅限于将注意力机制应用到视频与问题2种模态, 而且考虑了字幕与问题之间的注意力. Kim等人[68]提出了一种多模态双重注意力记忆(multimodal dual attention memory, MDAM)模型. MDAM中的双重注意力与多模态融合是其关键所在. 双重注意力机制的设计思想来源于Transformer. 自注意力模块用于学习预处理帧与字幕潜在的可变信息. 多头注意模块在给定问题下根据自注意力模块的输出来找出与问题相关的潜在信息. 与文献[68]相比, Lei等人[69]提出的网络增加了一个视觉概念特征. 该网络分为3流进行独立处理, 将区域视觉特征、视觉概念特征和字幕特征分别与问答对进行基于注意力机制的特征融合, 每个处理流均使用上下文匹配模块[75-76]和Bi-LSTM策略. 该方法并没有充分地考虑到视频和字幕之间的交互以及视频中的对象关系, 对于多模态推理的能力有所欠缺. 针对文献[69]中的方法不足, Li等人[70]提出关系感知分层注意力网络, 在引入视觉概念的同时充分考虑了对象之间的动态关系和交互理解. 该网络使用基于GAT[77]编码器建模对象之间的空间和语义关系, 并采用问题引导的层次注意力模块捕捉多模态对象的静态和动态关系, 最后利用自注意力机制进行多模态融合以突出各模态本身的重要性.

利用多模态注意力机制来定位问题相关的关键时刻, 然而关键时刻定位所需的模态可能与答案预测所需要的模态不同. Kim等人[71]提出模态转移注意力网络(modality shifting attention network, MSAN), 它很好地解决了模态转换问题. MSAN包含2个组件, 即时刻提议网络(如图10中的②)与异构推理网络(如图10中的③). 前者用来定位具体时刻, 利用注意力机制对上下文与假设进行联合建模; 后者使用多模态注意力机制来预测答案, 它引入异构注意力机制来考虑模态间和模态内的相互作用. 同时也提出了模态重要性调制(如图10中的①)来给定2个组件中每个模态的权重.

与传统的注意力机制不同, Jin等人[72]提出了一种新的多交互注意力机制. 多交互在该模型中指的是视觉信息与文本信息的交互, 以及多模态中多层次交互, 即帧级和片段级2种类型的交互, 其与Transformer模型结构相似. 该模型既考虑了视频中的动态特征, 又考虑了不同级别的句子表示, 这对模型的推理起到关键作用.

Kim等人[73]提出双级注意力机制, 分别是词/对象级与帧级. 然后以自注意力和交叉注意力机制融合视频与密集字幕来进一步改进模型的时间定位, 最后通过门控机制选择信息量最大的帧. 视频中密集字幕的引入相比于单个图像字幕能够提供更有用的线索来回答问题. 基于文献[73]提出的模型, Chadha等人[74]在其输入端加入了常识知识库, 从输入视频中的帧特征来生成对应常识性特征与原视频特征连接. 这些常识性特征可以帮助模型更好地感知视频中事件之间的关系, 从而提高模型在常识推理的视频任务中的表现.

本节从单跳注意力、多跳注意力与多模态注意力3个互相独立又存在包含关系的方面对基于注意力的模型进行了详细介绍. 单跳注意力仅能突出视频与问题的浅层关系, 深层关系则需要视频与问题的多跳注意力递进挖掘. 而多模态注意力除了关注视频与问题的交互权重之外, 同时考虑各模态的自注意力以及视频与问题, 字幕与问题之间的共同注意力, 充分探索了多模态中模内关系与模间关系. 整体来讲, 基于注意力的模型有着低复杂度且与时序

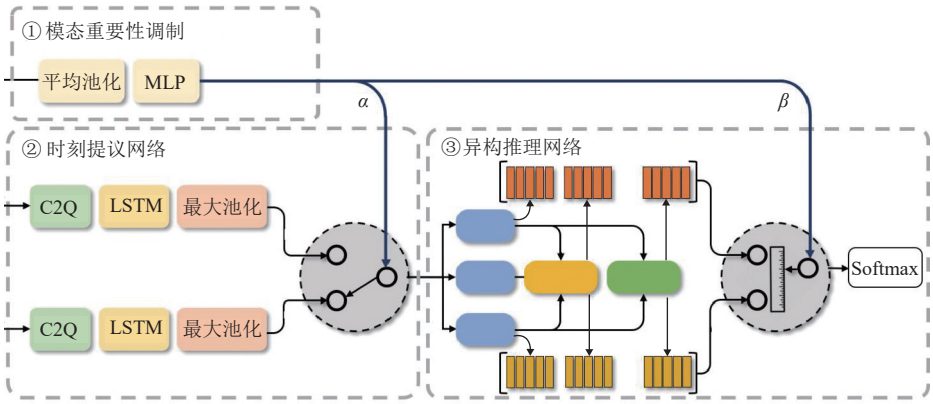


Fig. 10 Key modules of MSAN model[71]

图10 MSAN模型的关键模块[71]

无关的特点，也说明该类模型无法捕捉位置信息，即不能学习视频与问题序列中的顺序关系.

2.2.2 基于记忆网络的模型

在 VideoQA 这一长序列学习任务中，模型需要记忆更多的视频内容，并在答案推理时，根据问题在记忆的多个时间帧信息中进行准确定位. 现有的大多数机器学习模型都缺乏能够与推理无缝结合的长期记忆单元. 虽然基于 RNN 的方法可以进行序列学习，但受其记忆单元本身的特性影响，不能准确完整记忆长序列内容. 为了解决这一问题，研究人员探索使用记忆网络来进行长序列学习和推理. 本文将基于记忆网络的模型分为静态记忆网络[31-32,78-81]和动态记忆网络[34-35, 82-86].

1）静态记忆网络

MemNN 由 Weston 等人[12]提出，被用于文本问答. 该网络的核心思想是构建一个可以读写的记忆组件，同时建立故事、问题与答案之间的关系模型. MemNN 模型需要支持答案的事实进行监督训练，然而现有数据集并不包含答案对应的事实支撑. 因此，Sukhbaatar 等人[78]提出了端到端记忆网络 (end-to-end memory network, MemN2N)，同样用于文本问答. 与 MemNN 不同的是，该网络使用输入输出对进行端到端的训练，所以 MemN2N 仅仅需要弱监督训练模式，更普遍地适用于现实环境.

受 ImageQA 任务的启发，Zeng 等人[79]进一步对 MemN2N 模型进行扩展，提出了拓展型端到端记忆网络模型 E-MN，将其原始输入修改为由帧序列组成的视频，使用双向 LSTM 对帧表示序列进行编码，捕获连续帧中动作之间的时序关系，提高了模型对时间信息的感知能力. 但是，由于视频包含了丰富多样的数据，简单的扩展模型并不能很好地利用它们. 与文献 [79] 类似，Tapaswi 等人[80]基于 MemN2N 模型进行修改. 为了应用于更大规模的 MovieQA 数据集，他们将原模型的词嵌入替换为 Word2Vec 预训练的词嵌入来减少训练参数，并学习一个共享的线性投影层将视频和问题映射到一个低维的公共空间.

Kim 等人[31]提出了一种深度嵌入记忆网络 (deep embedded memory network, DEMN) 用来解决视频故事问答任务，该模型对记忆网络的泛化成分进行了优化. 该网络将视频的场景与对话作为重点，将两者组合成视觉语言特征对，由其学习场景嵌入与对话嵌入. 然后以句子形式将场景和对话结合起来，从视频场景与对话的联合流中重新构建视频故事，将其储存在长期记忆组件中.

由于 DEMN 记忆网络模型是将每个记忆槽视为独立的内存块，因此忽略了相邻记忆块之间的相关性. Na 等人[32]提出了一种用于电影故事问答的可读写记忆网络 (read-write memory network, RWMN)，该网络的卷积分层网络由多个更高容量和更具灵活性的读写内存构成，并采用连续的方式存储，增强了存储单元之间的关联性进而使得后续推理更加准确. RWMN 的输入与 DEMN 相似，推理和回答预测部分与 MemNN 相似. 与其他记忆网络的不同之处在于该模型记忆阵列的维数在处理过程中减小，而其他模型的维数是不变的.

以上基于静态记忆网络的方法将不同模态信息保存于不同的记忆块中，并没有对需要记忆的特征做预处理或增强处理. Cai 等人[81]提出了一种基于递归神经网络和自注意力模块的模型. 其核心是特征增强模块与注意力机制. 特征增强利用视觉特征与问题特征通过记忆机制相互增强，两者进行细粒度的模态交互后，再记忆到内存中. 从 2 个记忆模块输出问题引导的视觉特征与视觉引导的问题特征，作为自注意力的输入，来捕获序列的全局上下文. 然后再使用互注意力机制进行 2 种特征的互相关注. 这种跨模态的特征增强记忆方法，可以实现在没有冗余信息的情况下有效记忆.

2）动态记忆网络

为了能够解决长序列的动态记忆和推理问题，Kumar 等人[82]提出动态记忆网络 (dynamic memory network, DMN)，用于解决基于文本问答问题. 其核心处理模块为情景记忆模块，它由注意力模块与循环网络组成，以问题、上一次记忆的内容和事实表征作为当前迭代的输入，用来更新情景记忆内容，并通过多次迭代更新得到最终的答案预测. Xiong 等人[83]在 DMN 模型之上做了进一步优化，提出了动态记忆网络优化模型 DMN+. 该优化模型将 DMN 中单向的 GRU 换成了双向 GRU，将原来记忆更新使用的 GRU 替换成 ReLU，不但简化了模型，还提高了模型的准确率. 与静态记忆网络相比，动态记忆网络能够通过注意力机制来迭代更新记忆内容，过滤掉不相关的记忆内容.

考虑到 DMN/DMN+缺乏动作分析与时序建模，Gao 等人[34]提出了一种基于 DMN/DMN+的动作与外观共同记忆网络模型. 具体来讲，将视频的动作特征和外观特征输入时间卷积和反卷积神经网络，生成多级上下文事实. 这些上下文事实被用作记忆网络的输入，共同记忆网络拥有 2 种独立的记忆状态，一种用于动作，另一种用于外观. 最后，使用共同注意

力机制解决动作与外观信息的交互和联合建模. 与文献 [34] 相同, Fan 等人[84] 提出的异构记忆增强多模态注意力模型同样考虑了视频的动作特征与外观特征, 不同之处在于该模型能充分地利用视觉特征和问题特征与记忆内容的相互作用来学习全局上下文感知表征, 模型架构如图 11 所示. 模型第 1 部分将外观特征和运动特征融合起来, 同时学习时空注意力, 解决了多数方法未能正确识别注意力的问题. 第 2 部分设计了新的网络结构, 将问题编码器和问题记忆网络整合起来, 主要是为了解决有较为复杂语义且需要推理的问题. 最后一部分, 设计了一个多模态融合层, 可以有效地将视觉特征和问题特征与注意力权重结合起来, 并支持多步推理.

针对电影类的 VideoQA 任务, 视频中包含大量的字幕信息, 这些字幕对问答尤为重要. Wang 等人[35] 提出了一个由静态词记忆模块和动态字幕记忆模块组成的分层记忆网络 (layered memory network, LMN), 能够学习电影内容的分级 (帧级与片段级) 表示. 首先, 静态词记忆模块利用帧区域视觉特征映射到单词空间, 得到帧级语义表示. 然后, 通过动态字幕记忆获得电影片段中特定帧的语义表示, 即片段级表示. 片段级表示是从词空间到句子空间的转换, 从而可获得大量的语义信息. Wang 等人[35] 同时提出了静态词记忆的多跳机制、动态字幕记忆的更新机制以及问题引导下的字幕表示机制, 这 3 个机制对模型的延伸方法去除无关信息起到非常大的作用, 也大大提高了模型的推理能力. 此外, 一般的多模态融合方法[87-88] 只关注于建模不同模态的交互特征, 这些方法对问题是不知晓的, 因为在模态融合过程中问题与答案是不参与其中的, 所以 Kim 等人[85] 提出了渐进注意力记忆网络 (progressive attention memory network, PAMN). PAMN 包含 3 个主要功能模块: ①递进注意力机制, 找出与回答问题有关的时间部分; ②动态模态融合, 自适应地确定每个模态的贡献来聚合每个记忆模块的输出; ③信念修正答案方案, 该方案基于已有的问题和注意力对每个候选答案的预测分数进行连续修正.

Yu 等人[86] 提出了由 2 个不同的记忆网络组成的多模态分层记忆注意力网络框架. 该网络又可以称为一种粗粒度到细粒度的记忆体系结构, 它能完成从粗粒度到细粒度的推理过程. 第 1 层为顶部引导记忆网络, 浅层次上过滤与问题不相关的信息. 第 2 层为底部增强的多模态记忆注意力网络, 该网络负责进行深度推理. 双层记忆网络的共同协作能够学习到视频帧之间的细粒度注意力, 从而提升问答的质量.

总的来看, 记忆网络的引入主要解决了模型对视频长序列建模的问题, 在基于注意力突出重要内容的基础之上又保证模型不遗漏其他用于推理的必

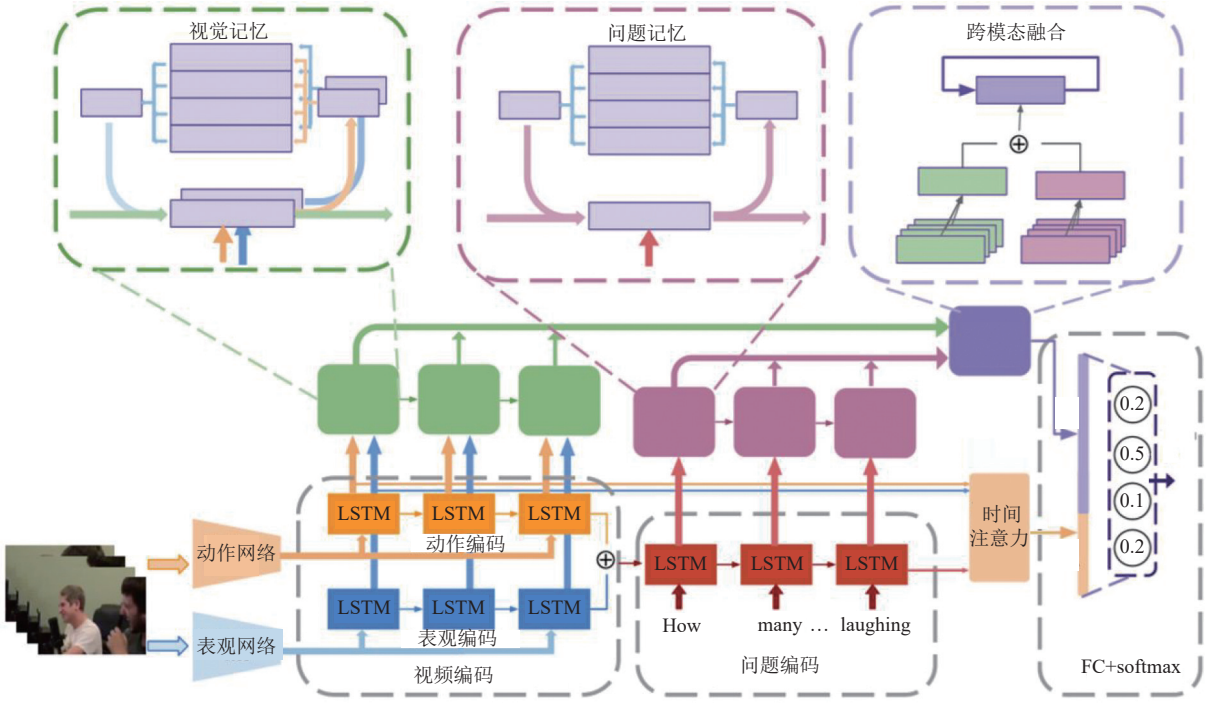


Fig. 11 Heterogeneous memory enhanced multimodal attention model[84]

图 11 异构记忆增强多模态注意力模型[84]

要信息. 静态记忆网络虽然能够完成对长序列信息的记忆, 但并未对其记忆内存 2 次或多次加工, 即其保存的信息往往是粗粒度的. 动态记忆网络则是对静态记忆网络的优化, 通过多次注意力使得记忆内容更加精细, 更有助于模型推理. 但是记忆网络需要较大存储空间与读写开销, 因此模型计算量也相对很大. 同时, 该类方法建模视频复杂的时空结构的能力较弱, 在处理时空推理的问题上表现不佳.

2.2.3 基于图网络的模型

近年来, 图神经网络在知识图谱与社交网络等领域取得了重大突破[89-90]. 图神经网络可将数据看作图中节点, 节点与节点间通过边进行连接, 并通过消息传播对节点间的依赖关系进行建模. 同时, 图神经网络不仅能很好地处理结构化数据, 而且能够处理像视频这样的非结构化数据. 视频包含丰富的时空相关的动态属性, 在一个视频中, 帧级信息能够提供视频的空间结构, 例如对象的位置信息与动作信息, 而片段级信息能够提供视频的时间结构, 例如发生动作的序列以及状态的转换. 为了能够更进一步地提高 VideoQA 的性能表现, 对视频时空结构的联合推理十分必要. 因此, 研究者通过利用图神经网络来挖掘视频中的时空依赖关系, 更好地建模对象间的关系信息.

Huang 等人[91] 提出了一个位置感知图卷积网络 (location-aware graph convolutional network, L-GCN) 来建模视频中被检测对象之间的关系. 视频编码流利用对象位置感知图来理解视频内容. 基于对象的位置感知图既考虑了对象之间的交互, 又考虑了对象的时间位置信息. 但是该方法比较依赖于学习模态之间的位置关系, 并没有挖掘到问题的深层次语义关系.

与文献 [91] 提出的方法不同, Jiang 等人[92] 提出了异构图对齐 (heterogeneous graph alignment, HGA) 网络, 把问题与视频特征融合形成一个异构图, 再送入 GCN 中. 该网络将多模态因素视为统一的异构图节点, 并通过对齐策略来生成加权邻接矩阵, 构造多层图卷积网络进行多模态交叉推理. HGA 通过建模多种模态之间的复杂相关性, 促进了模态间和模态内的相互作用以及跨模态推理.

虽然 HGA 同时利用了视频的外观特征与动作特征, 但是并没有充分挖掘两者分别与问题的深层关系. Seo 等人[93] 提出了动作-外观协同网络 (motion-appearance synergistic network, MASN), 将 GCN 与注意力机制相结合. 前期利用 GCN 分别对外观特征与动作特征进行时空建模, 后期以问题为引导的视觉表示为输入, 利用注意力机制生成分别以外观和动作为中心的特征及混合特征, 最后通过融合模块来调节 3 种信息的权重.

现有模型 HGA 与 MASN 直接将外观特征与动作特征进行交互, 没有充分利用异构模态的交互. Park 等人[94] 提出了一种称为 Bridge2Answer 的方法, Bridge2Answer 模型核心组件如图 12 所示. 该方法采用 GCN 思想构建表观图、运动图与问题图, 并充分利用它们之间的交叉关系来推断答案. 其中以问题

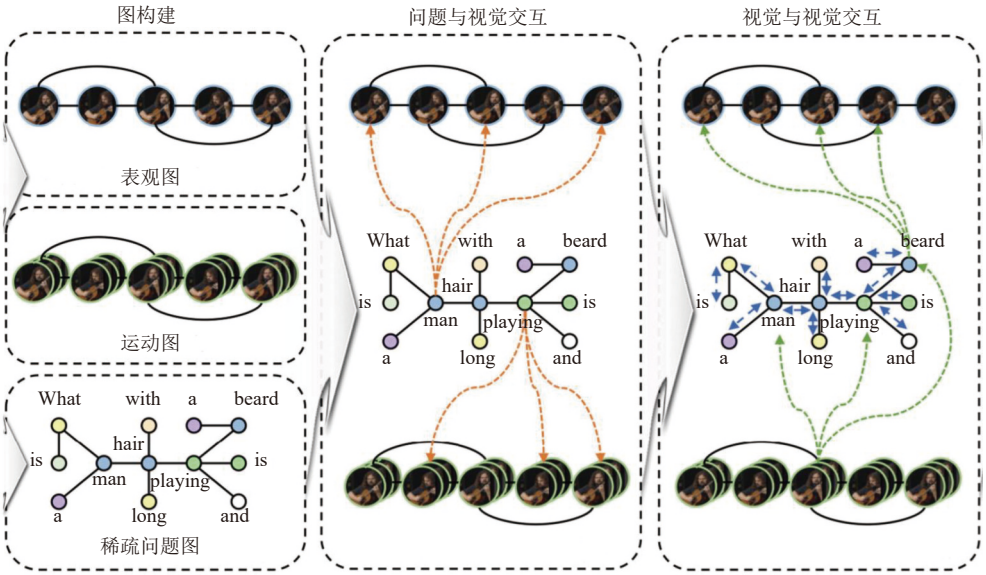


Fig. 12 The graph interaction part of Bridge2Answer method[94]

图 12 Bridge2Answer 方法的图交互部分[94]

图作为外观图到动作图和动作图到外观图的交互桥梁, 利用问题的合成语义以调节外观图与动作图之间的关系. 由于问题图的结构可以反映单词之间的语义依赖关系, 因此问题条件的视觉节点能够有效地传递到问题边缘的相关视觉节点.

此外, 为了在执行推理时进一步挖掘外观特征和动作特征关联和互补的关系, Wang 等人[95] 提出了双视觉图推理单元 (dual-visual graph reasoning unit, DualVGR), 模型以迭代方式堆叠该单元执行多步推理. DualVGR 通过查询惩罚模块过滤掉不相关片段的特征, 使用多视图图网络提供上下文感知的特征表示. 多视图图网络分别处理外观、动作及两者之间关系的特征图, 对于外观图与动作图, 通过自注意力机制更新其邻居节点的表示, 对于两者关系图则基于 AM-GCN[96], 通过执行图卷积操作为各自寻求一个特定的嵌入和一个公共嵌入.

为了更关注于视频中的对象及其交互, Dang 等人[97] 提出以对象为中心的视频表示作为构建视频时空结构基础的方法, 该方法的重点是将视频抽象为时空中存在的动态交互对象. 问题条件下的对象特征通过 GCN 与上下文对象特征进行交互, 整合动态对象图的时间维度信息, 创建一个由 N 个对象组成的无序集合. 最终, 视频被抽象为一个时空图, 其空间和时间依赖性取决于问题. 以对象为中心的视频表示的输出用作通用关系推理引擎的知识库, 并应用于提取问题的相关视觉信息.

与 MASN, HGA, DualVGR 相比, Jiang 等人[98] 提出的轻量级视觉语言推理 (lightweight visual-linguistic reasoning, LiVLR) 模型, 在同一数据集上较大程度地减少了模型参数的同时又提升了模型的性能. 该模型主要由基于 GCN 的视觉编码器、语言编码器与多样性感知视觉语言推理模块 (diversity-aware visual-linguistic reasoning module, DaVL) 组成. 视觉和语言编码器最终生成多粒度的视觉和语言表示, 由于模型考虑了视觉表征和语言表征在不同语义层次上的多样性, 所以使用基于 GCN 的 DaVL 模块进一步编码和捕获节点之间的关系, 并输出联合问题相关表征.

现有基于 GCN 的方法均在相同尺度的视频片段中寻找答案, 然而这些方法往往会导致获取的信息不足或冗余的问题. Jiao 等人[99] 提出了一种多尺度递进注意力网络 (multi-scale progressive attention network, MSPAN), 将 GCN 与注意力结合来实现跨尺度视频信息之间的关系推理, MSPAN 网络结构如图 13 所示. 通过不同核大小的最大池化得到多尺度图, 多尺度图中的每个节点通过 GCN 进行节点更新, 再利用逐步注意力机制来实现跨尺度图交互过程中多尺度特征的融合. 这种跨尺度特征交互能够挖掘不同尺度视频片段中对象之间的深层次关系.

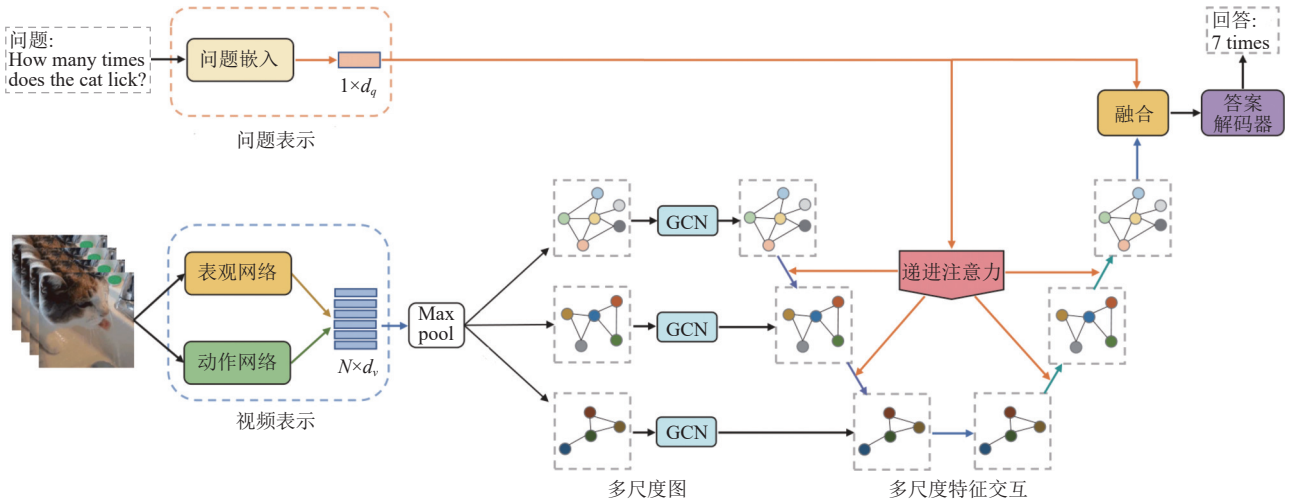


Fig. 13 MSPAN network structure[99]

图 13 MSPAN 网络结构[99]

以往方法一般只研究对象间或帧间的单一交互, 不足以理解视频中复杂的场景. Peng 等人[100] 提出了一种递进图注意网络模型 (progressive graph attention network, PGAT), 它通过图注意网络以渐进方式探索视频的对象级、帧级和片段级的多重关系. 这些不同级别的图以循序渐进的方式连接起来, 以理解从低级到高级的视觉关系. Liu 等人[101] 将记忆机制结合到图网络中, 提出了视觉图记忆与语义图记忆, 并认为语义关系与视觉关系对于推理一样重要. 这 2 种图记忆机制通过可学习的视觉到语义和语义到视觉的节点

映射相互协作和交互. 最后, 构建了从对象级到帧级的层次结构, 从而实现了层次的视觉语义关系推理.

本节主要介绍了基于图网络的模型, 该类模型之所以能够达到较好效果的原因在于它能够直接对视频内容结构进行时间和空间的统一建模, 较容易捕获到视频中各对象之间的关系, 能够学习到更好的节点表示, 对后续时空相关的推理问题起到较大作用. 由于 GCN 需要将整个图放到内存和显存, 多层 GCN 将会有很大开销, 模型训练耗时也会很久.

2.2.4 基于 Transformer 和 BERT 的模型

针对 RNN 等序列模型不适合处理序列的长期依赖以及不易于并行化数据处理的问题, 研究人员提出了 Transformer[15]. Transformer 通过其内部自注意力机制能以有限的层数建模长期依赖关系, 而且相比于 RNN, Transformer 能够利用分布式 GPU 进行并行训练, 提升模型训练效率. BERT 实际是 Transformer 的复合体, 其最早被应用于自然语言处理领域[16]. 随着 Transformer 与 BERT 的流行, 越来越多的模型开始将二者引入到各个领域, 并取得了令人惊叹的结果. 当前存在的大部分用于 VideoQA 的模型都是基于 RNN 的模型, 如 LSTM, 然而类似这样的模型可能无法捕获长序列之间的关系. 因此研究者尝试将 Transformer 和 BERT 引入到 VideoQA 任务上, 并取得了显著的效果.

Yang 等人[102] 提出使用 BERT 对视频中的视觉概念与文本内容进行编码来获得视频场景的视觉信息与文本信息. 同时, Urooj 等人[103] 提出了 MMFT-BERT 模型, 采用 BERT 单独处理多模态中的每一个模态, 然后使用一个新设计的基于 Transformer 的融合方法进行后期融合. 该方法考虑到了早期对不同模态的单独处理, 将模态融合放在后期, 这样处理使得模态更能友好交互, 突出关键信息. 并且该方法也是第一个使用 Transformer 进行模态融合的方法.

文献 [104] 提出的 ROLL 模型的 3 个独立分支 read, observe, recall 均是使用 Transformer 提取语言特征建模, 但是该模型依赖于知识库. 与之相比, Engin 等人[105] 提出的 DialogSummary 方法则不需要这些外部知识. 该方法视频描述的生成借鉴于 ROLL, 其核心思想是从视频中的原始数据提取所需知识, 将以往人工生成知识的过程替换为从视频任务原始对话中自动生成情节摘要. 模态处理与 MMFT-BERT 类似, 每个模态由 BERT 进行独立编码, 不同的是该方法采用一个相对简单的模态融合方法, 而 MMFT-BERT 则采用了一种基于 Transformer 的多模态融合方法.

VideoQA 评估任务大多仅限于单个单词的开放式答案或从多个短语中选择一个短语, 限制了模型的应用场景. Sadhu 等人[106] 将 VideoQA 任务作为填充短语任务, 为了能够评估短语式答案, 模型计算预测答案对比空字符串的相对改进. 基于此任务, 提出了 5 个基准模型, 其中 VOG-QAP 与 MTX-QAP 综合表现突出. VOG-QAP 使用了额外的短语编码器并在多模态特征上应用 Transformer. MTX-QAP 与 ActBert 具有类似的架构, 但它用一个普通的 Transformer 替换了 ActBert 的 TNT, 在一个 Transformer 中联合编码语言和视觉特征.

与文献 [106] 动机类似, Castro 等人[36] 提出的 T5+I3D 模型同样是以生成式答案解决填词或短语的任务. 该模型属于早期融合模型, 模型的编码与解码均基于 Transformer. 此外, Castro 等人[36] 使用 T5(编解码 Transformer 网络) 对模型进行初始化, 并结合 I3D 提取的视频特征使得模型性能略优于后期融合模型.

文献 [107] 中基于 2D 的场景图忽略了视频本质是发生在 3D 空间中的事件, Cherian 等人[108] 提出基于 Transformer 的 (2.5+1)D 时刻场景图的方法, 该方法的处理流程如图 14 所示. 他们将视频帧 2D 画面转换成 2.5D(伪 3D) 场景图, 然后构造一个包含静态与动态子图的 (2.5+1)D 时空场景图表示, 以更好地捕捉视频中的时空信息流. Transformer 将场景图嵌入到时空分层潜在空间中, 以不同的粒度捕获子图及其交互, 其核心思想是使用图节点的时空接近度来定义相似性.

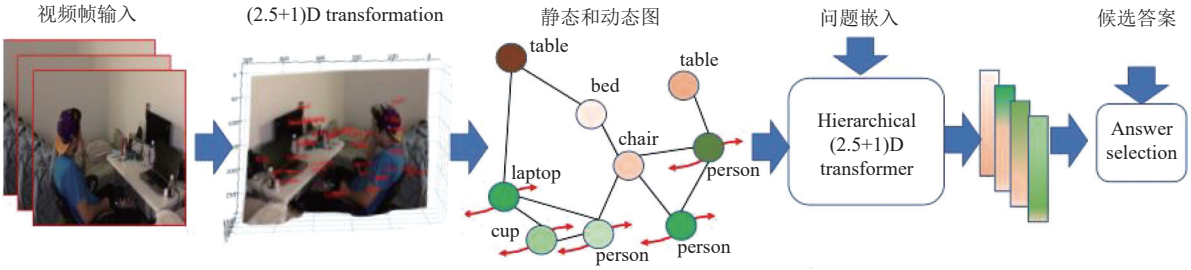


Fig. 14 The schematic illustration of (2.5+1)D VideoQA reasoning pipeline[108]

图 14 (2.5+1)D 视频问答推理流程示意图[108]

与文献 [99] 的思想有些相似, Peng 等人[109] 提出的 PTP 同样利用视频中的多尺度信息. 将视频按不同级别构建时间金字塔, 高层级比低层级具有更丰富的局部信息, 低层级比高层级具有更完整的全局信息. 该模型包括问题 Transformer 和视觉推理 2 个模块, 两者均在 Transformer 上进行了改进, 在每个模块中引入了一种多模态注意力机制来辅助问题与视频交互, 并在不同层次的信息传递中采用残差连接. 问题 Transformer 用来构建从粗粒度到细粒度的问题词与视觉内容之间的多模态语义信息, 在问题特定语义的指导下, 视觉推理模块从问题与视频之间局部到全局的多级交互中推断出视觉线索.

以上基于 Transformer 的部分模型在 VideoQA 任务上实现了最优性能比基于图网络模型更优的性能, 这归因于其自注意力结构的设计. Transformer 主要由多头注意力机制组成, 且相较于传统 RNN, CNN, Transformer 在大模型和大数据方面具有强大的可扩展性且架构灵活. 然而正因其对大数据训练的依赖, 使其在小规模数据集上泛化性与自适性较弱.

2.2.5 基于预训练的模型

预训练模型最早是在自然语言处理和计算机视觉等单模态领域崭露头角, 并在许多下游单模态任务中也被证实它的有效性. 后来, 研究者们将预训练模型应用于多模态任务, 并取得了重大进展[110-111]. 目前主流的多模态预训练模型是视觉-语言预训练模型, 其通常利用辅助任务从大规模未标注或弱标注数据中自动挖掘监督信号来训练模型, 从而学习通用表示. 这些预训练模型通过在下游任务上使用少量人工标记数据进行微调就能实现令人惊讶的效果.

最近一些方法使用带有图像字幕的数据集 (如 COCO[112] 和 Visual Genome[113]) 或视频字幕的数据集 (如 HowTo100M[114]) 来预训练多模态视觉语言表示. 这些方法绝大部分是基于 Transformer 之上在大数据集上进行预训练, 它们通常使用通用目标进行优化, 例如掩码语言损失、图像-文本匹配损失以及图像标题生成损失等. 以下介绍的预训练模型部分是针对于特定 VideoQA 任务的, 其余则是与下游任务无关的预训练模型.

Kim 等人[115] 提出了自监督预训练方法, 有效地利用了数据集的额外优势以及学习更好的特征表示. 自监督预训练阶段不需要额外的数据或注释, 在给定视频与字幕的条件下来预测相关问题而非预测答案, 这样使得模型能够学习到较好的权重. Yang 等人[116] 提出针对特定任务的预训练模型 VQA-T(VideoQA-Transformer), 对于目标 VideoQA 效果的提升有更大帮助. 该模型的 2 个分支均是基于 Transformer, 可以很容易地对不同的下游 VideoQA 数据集进行微调, 这些数据集可能包含训练中没有出现的新答案.

相比于特定任务的预训练模型, 下游无关的预训练模型更加灵活、应用更广. Zhu 等人[37] 提出了用于多种视频和语言任务的预训练模型 ActBERT, 该模型从无标记数据中进行联合视频与文本表示的自监督学习. ActBERT 模型的核心为 TNT(TaNgled Transformer block), 其包含 3 个 Transformer 来编码 3 个来源特征, 即全局动作特征、区域对象特征和语言特征. 为了增强视觉特征和语言特征之间的相互作用, ActBERT 在语言 Transformer 中注入视觉信息的同时, 在视觉 Transformer 中加入语言信息. 通过跨模态的交互作用, TNT 可以动态地选择有用的线索进行目标预测. 基于此模型, Zhu 等人[37] 提出了 4 个预训练任务: 掩码语言建模、掩码动作分类、掩码目标分类和跨模态匹配.

文献 [37] 的模型设计是对 BERT 的直接改编, 简单地将视觉和文本特征拼接作为输入, 而失去了视频和文本模式之间的时间对齐. Li 等人[117] 提出的 HERO 模型以一种分层的方式对多模态输入进行编码, 其包含 2 层 Transformer. 第 1 层为跨模态 Transformer, 用于融合字幕与其对应的局部视频帧; 第 2 层为时序 Transformer, 用于获取视频每一个片段的全局上下文嵌入. 该分层模型首先在帧级层面挖掘视觉和文本局部上下文, 然后将其转化为全局视频级时间上下文. 基于此模型提出了 4 个预训练任务, 相对于常见的掩码语言建模与掩码帧建模增加了视频与字幕匹配和帧顺序建模.

与文献 [117] 提出的预训练任务相似, Zellers 等人[118] 提出了一个通过大规模无标签的视频片段以自监督方式训练基于 Transformer 的预训模型 MERLOT. 视觉与语言特征均加入了位置嵌入, 然后由基于 RoBERTa[119] 结构的 Transformer 对视觉和语言进行联合编码, 并设计了帧与字幕匹配、掩码语言建模和帧顺序建模 3 个预训练任务.

现有的部分工作如文献 [37] 提出模型离线提取密集的视频特征和文本特征, 然而从视频帧的全部序列中提取特征会导致对内存和计算的过多需求. Lei 等人[120] 提出了一个通用的预训练模型 CLIPBERT, 其核心思想为稀疏采样与密集推理, 图 15 为常见的视频-语言学习方法和 CLIPBERT 的比较. CLIPBERT 将来自同一视频的不同片段子集用于不同的训练步

骤，因而其在一定程度上提高了模型的泛化能力. 此外与文献 [37, 117] 不同的是，该模型使用的是图像文本数据集进行的预训练，实验结果表明图像文本预训练同样有益于视频-文本任务.

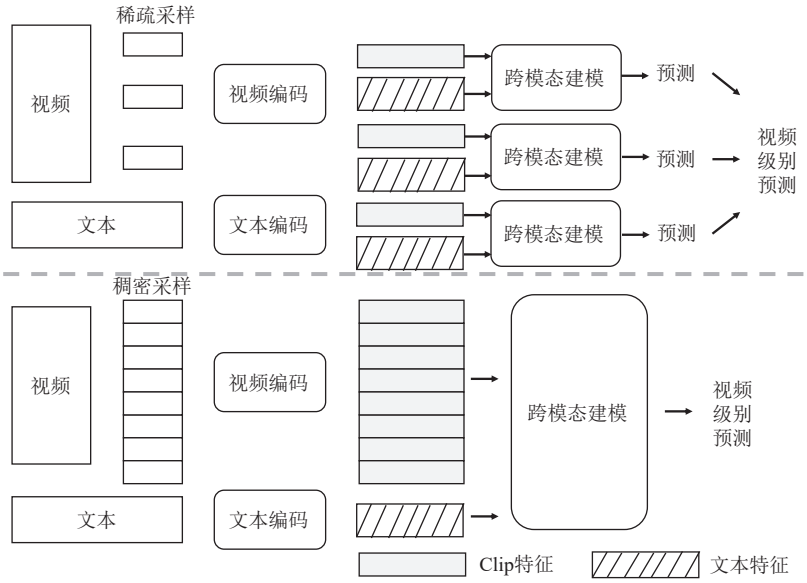


Fig. 15 Comparison between popular video-and-language learning paradigm and CLIPBERT[120]

图 15 流行的视频和语言学习范式和 CLIPBERT 之间的比较[120]

受文献 [120] 中稀疏采样策略的启发，Yu 等人[121] 提出了基于 CLIPBERT 的孪生采样与推理的方法 (siamese sampling and reasoning, SiaSamRea). SiaSamRea 的思想为多个片段应该相互依赖，应将上下文片段之间的相互依赖知识融于网络推理中，以在同一视频中捕获相似的视觉和关键语义信息. 所以该方法在稀疏采样的基础之上进一步采样了多个相似的片段，来学习片段之间的相互关系. 不同于以往的相关工作如 CLIPBERT，在同一个视频中挖掘它们的上下文知识，SiaSamRea 模型中的每个视频-文本对都被独立地编码到网络中. 充分地利用了片段之间丰富的上下文信息，可以进一步提升模型推理的准确性.

现有多模态学习任务中，通常会因存在噪声而使模型无法达到预期效果. Amrani 等人[122] 提出了一种去噪声的方法，使用自监督方式去训练一个去噪模块. 在多模态数据中，当 2 个或多个模态不具有相同语义含义时样本认定包含噪声. 因此，该模型将噪声估计简化为多模态密度估计任务，利用多模态密度估计，又提出了一种用于多模态表示学习的噪声估计组件，该组件严格基于不同模态之间的内在相关性. 该方法从去噪声的角度，在多模态任务中一定程度上提高升了性能.

此外，Luo 等人[123] 认为掩码输入将不可避免地为掩码建模与跨模态匹配等任务引入噪声，所以提出了对比跨模态匹配和去噪的方法 CoCo-BERT. 该方法包含 2 个耦合的视频/句子编码器，同时利用屏蔽和非屏蔽的多模态输入，从多模态对比学习的角度加强跨模态关联. 模型的核心是通过对比方式同时追求模态间匹配和模态内去噪，并利用掩码和非掩码输入来加强跨模态推理.

Seo 等人[124] 提出了一个多任务的预训练模型 CoMVT，该模型的目标主要是基于当前的视频片段和对应字幕来预测下一段话语，微调后的模型在下游 VideoQA 上取得了具有竞争性的结果. CoMVT 有 2 个关键点：1）虽然该模型没有像 CLIPBERT 稀疏采样，但其利用注意力机制聚合冗余特征，从而构造出更紧凑的视觉特征；2）使用一个共同注意力 Transformer CoTRM[125] 进行跨模态融合. CoTRM 由双流组成，每一个流由 2 个 TRM 组成，其中一个用于模态间特征交互，另一个用于模态内特征交互. 2 个流本质上分别处理每个模态，允许通过每个流中 TRM 的不同模态特征进行特定的操作和表示.

与先前相关工作提出的预训练任务不同，Fu 等人[126] 提出了掩码视觉标识建模的预训练任务. 视频帧被“标记”为离散的视觉标识，用于重建原始视频帧. 在预训练期间，沿空间和时间维度屏蔽了部分视频输入，模型学习恢复这些屏蔽部分的离散视觉标记. 相对于掩码语言/帧建模，模型需要在离散空间上

进行预测, 这避免了与文献 [117] 中类似的特征维度的过度训练问题. 此外, 所提出的 VIOLET(video-language Transformer) 模型并不是简单地均值池化或对一系列单个帧特征进行连接, 而是包含 Video Swin Transformer[127], 它可以显式地为视频语言学习建模视频时间.

如果预训练数据集和下游数据集之间存在领域差距, 当前流行的“先训练后微调”的视觉和语言模型泛化能力就会变弱. Zhou 等人[128] 系统地研究了视频语言预训练与微调模型中的领域差距问题, 并提出了一个任务自适应的视频语言预训练模型, 通过过滤和调整源数据到目标数据, 然后进行领域聚焦的预训练, 这有效地缩小了源数据 (用于预训练) 和目标 (用于微调) 数据之间的领域差距.

本节主要介绍了预训练模型在 VideoQA 任务中的应用, 它们主要通过基于大规模数据进行预训练来学习不同模态之间的语义对应关系. 目前, 基于 Transformer 的预训练模型取得了 VideoQA 任务的最佳性能. 这归因于预训练模型不仅能够充分利用广泛的网络资源, 而且还能完美地解决人工标记数据较为复杂的问题. 预训练模型通常是通过微调将知识转移到下游任务, 随着模型规模的不断增加, 每个下游任务均有不同的微调参数, 将导致参数学习效率低下, 同时多种下游任务也使得预训练和微调阶段的设计变得繁琐.

2.2.6 其他模型

除 2.2.1~2.2.5 节所述的 5 种 VideoQA 任务的解决方法外, 还有许多研究者们提出了不同于上述方法的模型来解决该问题, 同时也达到了具有竞争性的表现. 例如基于基础构建单元的模型、基于神经符号的推理模型、基于强化学习的模型、引入外部知识的模型、引入音频信息的模型等.

1) 基于基础构建单元的模型

当前 VideoQA 任务中的问题类型不受限制, 许多模型根据数据集特性来进行设计, 导致其在数据形态改变或视频长度改变的数据集上的性能表现不升反降. 为了缓解这一问题, 模型需要具备对视觉信息与文本信息深厚的建模能力, 学习时空中跨模态信息以对对象、关系和事件进行推理.

Le 等人[129] 提出了一种分层次条件关系网络 (hierarchical conditional relation network, HCRN). 条件关系网络 (conditional relation network, CRN) 是 HCRN 的基础可重用构建块, 该单元计算输入对象之间的稀疏高阶关系, 然后通过指定的上下文调制编码. 然而, CRN 只是专注于单个对象动作的时间推理, 不能很好地推广到时空中多个物体相互作用的情景. 与 HCRN 分层推理结构的设计理念相似, Dang 等人[130] 提出了一种由 OSTR 基础单元构建的面向对象时空推理层次 (hierarchical object-oriented spatio-temporal reasoning, HOSTR) 模型. HOSTR 的特点是对象内时间聚集和对象间空间相互交互的划分, 从而提高推理过程的效率. HCRN 与 HOSTR 都以通用的可视化推理为目标, 两者都忽略了问题的不同部分可能会需要不同粒度级别的视觉信息. Xiao 等人[131] 设计了分层问题引导图注意网络 (HQGA), 基于问题条件的图注意力单元 (QGA) 通过图的聚合和池化将低层次的视觉信息聚合为高层次的视频元素, 并通过堆叠 QGA 单元在每层注入问题, 从而实现多粒度级别的视觉-文本匹配.

2) 基于神经符号的推理模型

部分模型侧重于对复杂的视觉与语言的模式识别能力, 而忽略了蕴含于视频结构中的时序与因果关系. Yi 等人[132] 提出了基于碰撞事件的视频推理数据集 CLEVRER, 同时又提出针对于该数据集的模型-结合神经网络和符号的动态推理 (neuro-symbolic dynamic reasoning, NS-DR) 模型, 该模型结合了用于模式识别和动力学预测的神经网络, 以及用于因果推理的符号逻辑. NS-DR 将动态规划纳入视觉推理任务中, 能够直接对未观察到的运动和事件进行预测, 并能够对预测性和反事实性任务进行建模, 这将对 VideoQA 任务有着积极的影响. 然而 NS-DR 模型需要对视频视觉属性和物理事件进行密集注释, 这在真实场景中是不切实际的. Chen 等人[133] 提出了一个统一的神经符号框架, 即动态概念学习器 (dynamic concept learner, DCL), 它基于对象追踪和语言建模来识别视频中的对象与事件并分析其时间和因果结构, 而无需对视觉属性和物理事件 (如训练期间的碰撞) 进行注释.

在 NS-DR 与 DCL 基础之上, Ding 等人[134] 提出了基于可微物理模型的神经符号视觉推理框架 VRDP (visual reasoning with differentiable physics), 它通过从视频和问题对中学习物理模型, 并利用显式的物理模型对物体动力学进行建模, 基于准确的动力学预测来回答长期和反事实预测问题. VRDP 由视觉感知、概念学习器和可微物理模型 3 个模块组成. 视觉感知模块用于得到物体及其轨迹; 概念学习器借鉴于 NS-CL[135], 负责从物体的轨迹信息和问题对中学习物体的属性; 根据物体的轨迹和属性, 通过可微物理模拟

学习相关物理参数，得到较为准确的物理模型. 模型的神经符号执行器利用了 NS-DR 和 DCL 中的方案，通过预测出的物体轨迹和碰撞事件进行逐步显式的符号推理，使得模型具有良好的解释性.

3）基于强化学习的模型

迄今为止，用于 VideoQA 的方法在现实生活中应用性非常弱，原因之一就是应用性强的数据集非常少. Xu 等人[136]提出了应用性较强的交通问答数据集 TrafficQA，并基于该数据集设计了一种基于动态推理的高效一瞥网络. Xu 等人[136]通过 6 个具有挑战性的任务来训练该网络模型，与现有的 VideoQA 模型不同，为了减少视频帧之间冗余信息的影响，该模型自适应地确定每一步跳过的帧数和选择的帧位置，以及对选择帧需要分配的计算粒度. 该方法避免了对视频中不相关的片段进行特征提取，从而大大降低了整体的计算成本，实现了推理的可靠和高效.

4）引入外部知识的模型

除了利用数据集本身信息之外，数据集外部的知识对问答推理也有极大的帮助. Garcia 等人[137]提出的模型 ROCK (retrieval over collected knowledge) 通过知识检索模块来获取与问题最相关的知识. 知识源来自于他们自己构建的数据集，知识类型为人工注释的句子，知识检索模块通过计算问题与知识的相似性分数来获取有用的信息. Han 等人[138]提出了一个利用电影片段、字幕和基于图像的外部知识库来回答问题模型. 图像的外部知识库是他们设计的一个 PlotGraphs 的数据集，该数据集以图像形式提供回答问题的额外信息. 与文献 [137−138] 不同，Garcia 等人[104]提出了 ROLL 模型，其获取的外部知识来源于在线的外部知识. 不同于以往人工手动生成场景描述或者故事摘要等，该模型使用无监督方式生成视频场景描述，并且以弱监督方式获取外部知识.

5）引入音频信息的模型

以往研究忽略了利用视频中的音频信息，虽然有相关工作利用语音转换字幕系统，但仅限于提取其中的文字信息. Le 等人[27]提出的模型 VGNMN (video-grounded neural module network) 尝试将音频模态加入推理过程，其分为对话理解与视频理解 2 部分. VGNMN 模型由多个负责不同功能的神经网络块组成，形成复合推理结构，实现逐步检索语言和视觉信息的显示推理过程，这种模块化方法可以实现模型更好的性能和透明度. Shah 等人[28]提出了三重注意力网络模型，同样也将音频信息整合到 VideoQA 任务中. 模型利用 Mel Spectrograms, SoundNet 与 WALNet 提取 3 种音频特征，与视频和字幕形成异构信息源，音频、视频、字幕三者分别与问题使用注意力机制来不断更新内存向量. 该模型通过消融实验证明了音频信息的加入有利于 VideoQA 模型性能提升.

2.3 答案生成

在 1.1 节提到，问题大致可以分为开放式问题与选择题 2 种类型. 开放式问题可以分为开放式单词问题与开放式数字问题，即开放式问题对应的答案是单词或者数字，所以此类问题需要 2 种解码器. 而选择题则只需要 1 种解码器.

对于开放式单词问题的任务，一般视为多标签分类任务，因此使用 softmax 回归函数. 定义一个 softmax 分类器，该分类器以多模态融合表示 $\boldsymbol{O}$ 为输入，通过计算置信度向量 $\boldsymbol{s}$ 从词汇表中选择答案，计算形式表述为

$$\boldsymbol{s} = softmax(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{O} + \boldsymbol{b}), \tag{7}$$

其中 $\boldsymbol{W}^{\mathrm{T}}$ 与 $\boldsymbol{b}$ 是模型参数. 一般通过交叉熵损失函数或 softmax 损失函数来训练该解码器，通过 $\tilde{a} = \arg\max_{a \in \mathcal{A}}(\boldsymbol{s})$ 得到预测答案.

对于开放式数字问题的任务，与选择题任务类似，将上下文表示 $\boldsymbol{O}$ 输入一个线性回归函数中，与之不同的是，通过舍入函数（舍入到最近整数）输出的是一个整数值答案. 计算形式表述为

$$\boldsymbol{s} = round(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{O} + \boldsymbol{b}), \tag{8}$$

其中 $\boldsymbol{W}^{\mathrm{T}}$ 与 $\boldsymbol{b}$ 是模型参数. 通过均方差损失函数来训练该解码器.

对于选择题任务，1 个问题对应多个候选答案，只有 1 个选项为正确答案. 每一个候选答案将与给定的问题以相同的方式进行处理，最后将得到的融合表示 $\boldsymbol{O}$ 送入一个定义的线性回归函数，并为每个候选答案输出一个真实分数. 其计算形式表述为

$$\boldsymbol{s} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{O} + \boldsymbol{b}, \tag{9}$$

其中 $\boldsymbol{W}^{\mathrm{T}}$ 与 $\boldsymbol{b}$ 是模型参数. 通常训练模型的方法都是最小化预测答案与正确答案之间的损失，所以通过最小化成对比较的铰链损失来训练解码器 $\max(0, 1+s_{\mathrm{n}}-s_{\mathrm{p}})$，其中 s_{n} 和 s_{p} 分别是由错误答案和正确答案计算的分数.

3 数据集及性能

随着越来越多的研究者们关注 VideoQA 领域，用于解决此任务的数据集也越来越丰富. 例如以电影与电视剧为视频源的数据集 MovieQA[80], TVQA[69],

MovieFIB[139], KnowIT VQA[137] 等，这些数据集更加注重评测模型对视频与文本的理解能力，以及对故事情节的推理能力. 基于动画类型的数据集 SVAQ[59], MarioQA[140], PororoQA[31], Env-QA[53], CLEVRER[132], CRAFT[141] 也相继被提出. 该类型视频中的场景相对简单，故事线也较为清晰. 此外，以开放类视频为视频源的数据集有 MSRVTT-QA[48], MSVD-QA[48], YouTube2Text-QA[55], TGIF-QA[45], Activitynet-QA[142] 等，它们的数据大多来自于 YouTube 或其他在线网络视频. 这一类数据集更注重于生活场景，对实际应用来讲更具有意义. 各数据集的详细指标如表 2 所示，部分数据集示例如图 16 所示，图 16 中仅展示了视频的 1 帧，但是所有的问题和答案都属于视频中的一个片段. 对于每个数据集，我们只展示 1 个问题和相应的正确答案. 下面将对每个数据集进行详细介绍，同时统计了频繁使用的数据集对应的模型实验结果，并进行了对比与分析.

3.1 影视剧类型

1）MovieQA[80]

MovieQA 是应用广泛的大规模数据集之一，旨在评估从视频和文本中自动理解故事的能力. 为了更好地理解视频内容，数据集还提供了与电影视频、

Table 2 Comparison of Indicators of Each Data Set

表 2 各数据集指标对比

数据集	年份	数据源	视频数	片段数	平均长度/s	问答对	问答类型	问答生成
MovieQA[80]	2016	电影	408	6 771	202	14 944	选择题	人工
LSMDC-QA[143]	2017	M-VAD/MPII-MD	202	118 081	200	118 114	选择题	人工
MovieFIB[139]	2017	LSMDC2016	180	118 507	4.1	348 998	填空题	自动
YouTube2Text-QA[55]	2017	YouTube2Text	1 987	1 987	40	99 421	选择题/开放问题	自动
TGIF-QA[45]	2017	TGIF		71 741	3	165 165	选择题/开放问题	人工/自动
MSRVTT-QA[48]	2017	MSRVTT	7 000	10 000	15	243 000	开放问题	自动
MSVD-QA[48]	2017	MSVD	1 970	1 970	10	50 505	开放问题	自动
Video-QA[79]	2017	在线网络视频	18 100	18 100	90	175 076	开放问题	自动
MarioQA[140]	2017	游戏视频				187 757	选择题	自动
PororoQA[31]	2017	卡通视频	171	16 066		8 913	选择题	人工
TVQA[69]	2018	电视剧	925	21 793	76	152 545	选择题	人工
SVQA[59]	2018	Unity3D 生成	12 000	12 000		118 680	开放问题	自动
TVQA+[52]	2019	TVQA	279	4 198	60~90	29 383	选择题	人工
KnowIT VQA[137]	2019	电视剧	207	12 087	20	24 000	选择题	人工
Activitynet-QA[142]	2019	ActivityNet	5 800	5 800	180	58 000	开放问题	人工
EgoVQA[144]	2019	IU Multiview	16	520	20~100	600	选择题/开放问题	人工
Social-IQ[145]	2019	YouTube	1 250	1 250		7 500	选择题	人工
DramaQA[146]	2020	电视剧	18	23 928	3.6	17 983	选择题	人工
LifeQA[147]	2020	YouTube	59	275	74	2 326	选择题	人工
Tutorial-VQA[148]	2020	网络教学视频	76	408		6 195	开放问题	人工
How2QA[117]	2020	HowTo100M/TV	9 035	22 000	60	44 007	选择题	人工
Env-QA[53]	2021	模拟器生成		23 261		85 072	选择题	自动
CLEVRER[132]	2021	模拟器生成	20 000	20 000	5	300 000	选择题/开放问题	自动
TrafficQA[136]	2021	交通视频	10 080	10 080		62 535	选择题	人工
NExT-QA[149]	2021	YFCC-100M	5 440	5 440	44	52 044	选择题/开放问题	人工
AGQA[150]	2021	Action genome	9 601	9 601	30	36 000 000	选择题/开放问题	自动
STAR[151]	2021	Charades		22 000	30	60 000	选择题	自动
Fill-in-the-Blank[36]	2022	VaTeX	28 000	28 000	10	28 000	填空题	人工
CRAFT[141]	2022	模拟器生成	9 917	57 524	10	57 524	选择题	自动
EgoTaskQA[152]	2022	LEMMA	2 000	2 000	45	40 000	选择题/开放问题	自动



Q: Why do Joy and Jack get married that first night they meet in Las Vegas ?
A: They are both vulnerable and totally drunk.

(a) MovieQA



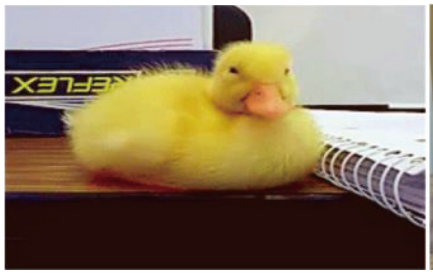
Q: He opens the ____?
A: door.

(b) MovieFIB



Q: What is on the couch behind Joey when he is at the counter ?
A: A soccer ball.

(c) TVQA



Q: What does the duck do 3 times ?
A: Shake head.

(d) TGIF-QA



Q: How many people are playing ?
A: 3.

(e) LifeQA



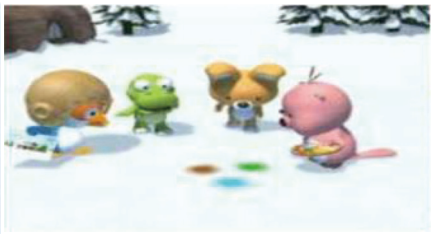
Q: What could have been done to prevent this accident from happening ?
A: The accident could have been avoided if the white sedan had stayed on its lane.

(f) TrafficQA



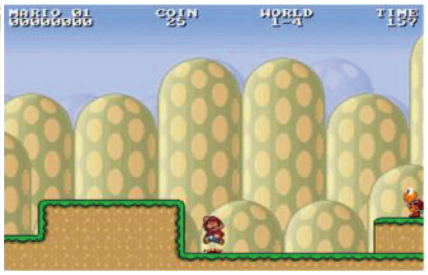
Q: How many cylinders behind the big cylinder that moves letf ?
A: 5.

(g) SVQA



Q: What do they think when they see the paint in the snow ?
A: They think it is beautiful.

(h) PororoQA



Q: What enemy came in before a Red Koopa Troopa appears ?
A: Spiky.

(i) MarioQA

Fig. 16 Some examples of datasets

图 16 部分数据集示例

字幕、描述性视频服务、脚本和情节概要相关的 5 种故事来源. 基于这些来源的不同组合, 该数据集包括 6 个子任务: 视频加字幕、仅字幕、仅描述性视频服务、仅剧本、仅情节概要和开放式. 由于影片片段的长度、不断变化的背景和故事情节, MovieQA 更注重用抽象和高层次的信息来理解故事.

表 3 统计了主流模型在该数据集上的性能表现, 根据测试集准确率对实验结果进行了排序. 将 DEMN 作为基准模型, 其在验证集与测试集的实验结果为 44.7% 和 30.0%. RWMN 在测试集的性能明显优于 DEMN, 表明相邻记忆块关联性在提高电影理解能力方面起着至关重要的作用. FVTA 测试集准确度相对于 RWMN 提升 1%, 在一定程度上体现了引入多跳注意机制对答案推理的必要性. LMN 采用了更新机制和问题引导注意力模型, 使字幕记忆与问题更加相关, 并且 LMN 具有良好的可扩展性. MDAM 通过后期融合避免了早期融合的过拟合现象, 并利用自注意力模型使得性能有所提升. 与 LMN 相比, PAMN 模型将多跳注意力与记忆网络相结合来动态推理, 使生成的答案更加准确可靠. Jasani 等人[153] 使用在维基百科上的电影情节训练的 WikiWord Embedding 词嵌入模型进行推理时仅使用问题和答案, 而忽略任何参考字幕或视频. 该模型性能之所以最优, 是因为该数据集中的问题存在语言偏见或问题较为简单.

Table 3 Performance of Mainstream Models on MovieQA

表 3 主流模型在 MovieQA 上的性能表现 %

模型	验证集准确率	测试集准确率
DEMN[31]	44.7	30.0
RWMN[32]	38.7	36.3
FVTA[61]	41.0	37.3
LMN[35]	42.5	39.0
MDAM[68]		41.4
PAMN[85]	43.3	42.5
WikiWord Embedding[153]	50.0	47.0

总的来讲, 由于 MovieQA 数据集本身长视频及多模态特性 (包含字幕), 使其在 VideoQA 这一任务中更具挑战性. 由表 3 可以看出, 基于注意力与记忆网络的模型更适用于此类长视频数据集, 注意力机制可以从复杂的故事情节中寻找关键信息, 记忆网络则用来保证电影情节的完整性. PAMN 模型正是由于巧妙地整合了两者从而达到了次优的性能. 然而诸如此类的的模型严重依赖语言提示, 由于未充分利用视觉特征而更容易出现语言偏见, WikiWord Embedding 模型就是通过使用经过适当训练的词嵌入, 利用数据集的偏好达到性能最优.

2) TVQA[69]

TVQA 数据集中的视频来源于 3 种类型的 6 部经典美剧. 数据集中的问题采用了组合式的设计, 包含问答和定位 2 个部分, 并且每个问题都带有时序定位. 也就是说问题模板首先使用开始时间戳和结束时间戳, 根据“when / before / after”来定位与问题相关的视频片段中的相关时刻, 然后构成与视频和问题理解相关的“what / how / where / why”问题. 回答这样的问题需要模型具有一定的时序定位、理解对话和视频的能力.

表 4 统计了主流模型在该数据集上的性能表现, 根据准确率对实验结果进行了排序. 作为基准网络, 文献 [69] 中的模型在验证集与测试集上的实验结果分别是 65.85% 与 66.64%. STAGE[52] 相对于原有模型同时考虑了时间与空间信息, 联合定位时刻与对象位置, 大幅度提升了模型的准确率, 表明时空信息对于回答问题的重要性. 其次, MSAN 模型性能的提升从回答问题需要不同模态的角度证明了模态转移的必要性. 文献 [115] 提出使用预训练模型, 将文本进行掩码并与原始文本形成对比学习, 从而学习到更好的特征表示, 模型准确率相对于基准模型提升近 10%.

Table 4 Performance of Mainstream Models on TVQA

表 4 主流模型在 TVQA 上的性能表现 %

模型	验证集准确率	测试集准确率
文献 [69]	65.85	66.46
PAMN[85]	66.38	66.77
Multi-task[51]	66.22	67.05
STAGE[52]	70.50	70.23
MSAN[71]	70.79	71.13
BERT Rep[102]	72.35	73.06
DCM&FSG[73]	74.20	74.09
iPerceive[74]	76.97	75.15
SP&CRL[115]	76.23	76.15

与 MovieQA 数据集相似, TVQA 同为带有字幕的长视频. 从性能表现上来看, STAGE 虽然引入时空位置信息, 但在建模对象交互上的欠缺导致其性能不佳. 文献 [102] 提出基于 BERT 的模型, 结构设计较为简单也达到了具有竞争性的性能, 足以证明此类模型在 VideoQA 任务上的巨大潜力. 其次文献 [115] 提出的预训练模型实现了当前最优性能, 但是其只是在已有数据集上进行自监督预训练, 有限的数据集导致模型不能学习得到更好的权重. 根据 TVQA 数据集上的对比结果可以预见 BERT 与预训练模型将会是下一步的发展趋势.

3) LSMDC-QA[143]

LSMDC-QA 数据集是来源于大规模电影描述挑战 LSMDC16[154], 该数据集由 M-VAD 和 MPII-MD 数据集融合而成. 该数据集针对模型对电影与字幕的理解提出了单项选择与视频检索 2 个任务. 单项选择任务中的正确答案来自真实字幕, 而其他候选答案是从其他字幕中随机选择的. 与其他数据集相比, 该数据集具有更多的视频片段, 更关注电影本身.

4) 其他

电影与电视剧类型数据集还包括 MovieFIB, TVQA+, KnowIT VQA, DramaQA. MovieFIB 是为视障人士提供的一个基于描述性视频注释的填空问答数据集, 其拥有超过 30 万条的问答与视频对. TVQA+ 数据集是来源于 TVQA 中的一个电视剧《生活大爆炸》, 在此基础上为问题的相关视频的帧上添加目标边框注释, 使其具有更多的时空关系. KnowIT VQA 数据集来自《生活大爆炸》的前 9 季, 该数据集试图通过整合外部知识来解决之前数据集有限的推理能力, 是最大的基于知识的人工生成 VideoQA 数据集之一. DramaQA 数据集来源于韩剧《又是吴海英》, 提

供217 308张以字符为中心的注释图像，该数据集着重于以角色为中心的表示形式，注释考虑了角色的行为和情感方面.

3.2 动画类型

1）MarioQA[140]

MarioQA数据集中的视频源于一款《无限马里奥兄弟》的游戏视频，每个视频片段都带有事件记录，其基于手工构建的模板与不同的事件. 事件的类型包括吃、举、敲和扔等. 数据集由3个子集组成，包含不同的时间关系特征问题：没有时间关系问题、简单时间关系问题与复杂时间关系问题. MarioQA数据集的特征是具有时间依赖性与多个事件的大量视频，视频中事件的发生是清晰的，所以很容易在游戏视频中学习完整的语义信息.

2）PororoQA[31]

PororoQA数据集源于儿童卡通视频. 该视频共有171集，每集有一个不同的故事，平均长度为7.2 min，总时长为20.5 h，共16 066对场景对话和27 328个人工生成的细粒度场景描述语句. 卡通视频相对于电影、电视剧等其他视频来说，视频简单明了，故事结构连贯，人物和背景的数量较少.

3）SVQA[59]

SVAQ数据集是由Unity3D生成的关于几何变化的视频组成. 数据集中的每个视频片段包含了3~8个静态或动态3D几何图形. 每个几何图形都有3个基本属性：形状、大小和颜色. 其中动态几何图形具有额外的动作类型和动作方向的属性. 基于这些属性，可以根据对象之间特定的时空关系、相对位置和动作顺序来构造问题. 与其他VideoQA数据集相比，合成视频包含了真实视频中难以收集的对象之间清晰的时空关系，这也导致该数据集的视频内容不够丰富，只包含具有对象之间各种时空关系的长结构化的问题. 此外，SVQA中的问题需要多步推理，它可以分解为可读的逻辑树或链布局，每个节点表示需要进行比较或算术等推理操作的子任务.

4）Env-QA[53]

Env-QA数据集的提出旨在评估模型理解动态环境的能力. 通过最近发布的AI2-THOR[155]模拟器生成以自我为中心的关于在环境中探索和互动的视频，这些视频共涉及15种基本动作、115种物体和120种室内模拟环境. Env-QA提供了5种类型的问题，从不同的方面评估对环境的动态理解，包括查询对象属性、对象状态、事件、事件的时间顺序、事件或对象的计数数量. 与MovieQA和TVQA这类影视数据集相比，Env-QA更加关注于环境的交互.

5）CLEVRER[132]

CLEVRER数据集中的每个视频都展示了一个简单的玩具物体场景，它们模拟物理中的相互碰撞. 该数据集的任务设计侧重于时序和因果的逻辑推理，因其有着较好的注释，可为复杂推理任务的模型提供有效评估. 该数据集中的问题分为描述性、解释性、预测性和反事实4种类型，从互补的角度研究了视频中的时序和因果推理问题.

6）CRAFT[141]

该数据集由Box2D模拟器创建，旨在评估模型对2D模拟视频相关问题的时间和因果推理能力. 数据集中的视频包含各种运动物体，它们彼此和场景相互作用. 问题类别包括以前研究过的描述性问题和反事实问题，同时引入了一个新的因果问题类别，通过因果、使能、预防概念来理解物体之间的因果交互作用.

3.3 开放类型

1）TGIF-QA[45]

TGIF-QA数据集来源于TGIF数据集(Tumblr GIF)[156]，以GIF动态图作为视频源. 该数据集基于TGIF数据集提出了4种类型的任务：重复计数、重复动作、状态转换和帧问答. 重复计数任务是关于计算某一动作重复次数的开放式问题. 重复动作任务定义为识别视频中重复动作的单项选择问题. 状态转换任务也是一个单项选择问题，是关于识别另一种状态之前或之后的状态，包括面部表情(如从悲伤到快乐)、动作(如从站立到坐)、位置(如从卧室到客厅)、物体属性(如从空到满). 帧问答任务是一个开放式问题，主要是基于视频中的某一帧，类似于图像问答. 与其他数据集相比，TGIF-QA包含了更多的动词形式，理解视频片段的内容需要丰富的时空推理.

表5统计了现有主流模型在该数据集上的实验结果，对重复动作、状态转换与帧问答这3个任务使用准确率进行评估，对计数任务使用损失进行评估. 一些基于注意力机制的方法虽然尝试从时空角度处理视频特征，并取得一些成效，但其并未真正理解对象的空间交互，以至于在时空问题上表现一般. L-GCN通过位置感知图建模对象位置信息与空间关系；HGA则构建视频与问题的异构图，侧重于模态的对齐与推理过程. L-GCN与HGA的模型性能均高于一般的注意力模型，但是它们没有利用或利用视频帧级的信息不够充分，缺乏对视频细粒度的理解. MSPAN通过多尺度视频特征交互来挖掘视频中对

象的高层次关系，由于其限于帧级与片段级的单一交互，这不足以处理视频中的复杂场景. PGAT 同时探索了对象、帧和片段之间的多个模态内交互，以更全面地理解视觉内容，在图网络模型中达到最优性能. 引入音频信息的 VGNMN 模型在各个任务上都取得了仅次于预训练模型的较优性能，同时 Le 等人[27]通过消融实验证明了音频模态在 VideoQA 任务中的重要性. 预训练模型在该数据集上表现最为突出，例如 MERLOT 与 VIOLET 等，它们在动作与状态类问题上的准确率高达 90% 以上. 尽管预训练在各种视频语言任务上都有明显的改善，但 Transformer 在视频语言上的潜力并未得到充分挖掘，一方面是缺少标准的预训练数据集，另一方面则是 Transformer 的效率问题，包括内存占用量和计算量.

Table 5 Performance of Mainstream Models on TGIF-QA

表 5 主流模型在 TGIF-QA 上的性能表现

模型	重复动作 /%	状态转换 /%	帧问答 /%	计数 损失
ST-VQA[45]	60.8	67.1	49.3	4.28
Co-Mem[34]	68.2	74.3	51.5	4.10
PSAC[67]	70.4	76.9	55.7	4.27
LAD-Net[66]	69.9	78.4	57.5	4.32
STA[65]	72.3	79.0	56.6	4.25
Jin 等人[72]	72.7	80.9	57.1	4.17
HME[84]	73.9	77.8	53.8	4.02
L-GCN[91]	74.3	81.1	56.3	3.95
HGA[92]	75.4	81.0	55.1	4.09
HCRN[129]	75.0	81.4	55.9	3.82
HOSTR[130]	75.0	83.0	58.0	3.65
FAM[81]	75.4	79.2	56.9	3.79
QueST[60]	75.9	81.0	59.7	4.19
Bridge2Answer[94]	75.9	82.6	57.5	3.71
TPT[109]	76.6	81.6	57.8	3.63
HAIR[101]	77.8	82.3	60.0	3.88
MSPAN[99]	78.4	83.3	59.7	3.57
HQGA[131]	76.9	85.6	61.3	
CoCo-BERT[123]	78.3	85.6	61.1	3.78
SiaSamRea[121]	79.7	85.3	60.2	3.61
PGAT[100]	80.6	85.7	61.1	3.96
CLIPBERT[120]	82.8	87.8	60.3	
VGNMN[27]	84.5	88.7	74.7	2.65
VIOLET[126]	92.5	95.7	68.9	
MERLOT[118]	94.0	96.2	69.5	

2）MSRVTT-QA 和 MSVD-QA[48]

MSRVTT-QA 和 MSVD-QA 数据集分别来源于 MSRVTT 和 MSVD 视频数据集. MSRVTT-QA 数据集更大且具有更复杂的场景. 数据集包含约 1 万个视频片段和约 24.3 万个问答对，问题有 5 种类型，包括 what, who, how, when, where, 该数据集中的视频相对较长，长度为 10~30s 不等，相当于每个视频 300~900 帧. MSVD-QA 数据集共有 1 970 个视频片段和 50 505 个问题答案对. 与 MSRVTT-QA 类似，问题有 5 种类型，主要用于视频字幕实验，但由于其数据量较大，因此也用于 VideoQA 任务.

表 6 统计了主流模型在该数据集上的性能表现，表中值均为准确率，其与表 5 如出一辙，性能更优的模型都是基于图网络和预训练模型. 基于图结构的

模型 LiVLR 在 MSRVTT-QA 取得了最佳的实验结果，其中 GCN 对视觉与语言的多粒度信息进行时空建模起到关键性作用. 基于大规模数据集的预训练模型在下游 VideoQA 任务上性能表现出众，MSRVTT-QA 和 MSVD-QA 数据集上的实验结果也印证了这一结论. 但此类数据驱动的模型，可解释性相对较弱.

Table 6 Performance of Mainstream Models on MSRVTT-QA and MSVD-QA

表 6 主流模型在 MSRVTT-QA 和 MSVD-QA 上的性能表现 %

模型	MSRVTT-QA						MSVD-QA					
	What	Who	How	When	Where	All	What	Who	How	When	Where	All
E-VQA[48]	18.9	38.7	83.5	70.5	29.2	26.4	9.7	42.2	83.8	72.4	53.6	23.3
E-SA[48]	22.0	41.6	79.6	73.1	33.2	29.3	15.0	45.1	83.8	65.5	32.2	27.6
E-MN[48]	23.4	41.8	83.7	70.8	27.6	30.4	12.9	46.5	80.3	70.7	50.0	26.7
AMU[48]	26.2	43.0	80.2	72.5	30.0	32.5	20.6	47.5	83.5	72.4	53.6	32.0
HRA[56]						35.1						34.4
HME[84]	26.5	43.6	82.4	76	28.6	33	22.4	50.1	73	70.7	42.9	33.7
L-GCN[91]												34.3
Jin 等人[72]	29.5	45.0	83.2	74.7	42.4	35.4	24.2	49.5	83.8	74.1	53.6	35.0
QueST[60]	27.9	45.6	83	75.7	31.6	34.6	24.5	52.9	79.1	72.4	50	36.1
FAM[81]	26.9	43.9	82.8	70.6	31.1	33.2	23.1	51.6	82.2	71.4	51.9	34.5
SSML[122]						35.1						35.1
TSN[33]	27.9	46.1	84.1	77.8	37.6	35.4	25.0	51.3	83.8	78.4	59.1	36.7
HGA[92]	29.2	45.7	83.5	75.2	34.0	35.5	23.5	50.4	83.0	72.4	46.4	34.7
MHMAN[86]	28.7	47.1	85.1	77.1	35.2	35.6	23.3	50.7	84.1	72.4	53.6	34.6
HCRN[129]						35.6						36.1
ActBERT[37]	29.4	45.6	79.8	76.7	36.4	35.5	28.7	53.8	80.0	70.7	46.4	39.0
HOSTR[130]						35.9						39.4
Bridge2Answer[94]						36.9						37.2
OCRL+LOGNet[97]						36.0						38.2
HAIR[101]						37.5						36.9
CLIPBERT[120]						37.4						
PGAT[100]						38.1						39.0
TPT[109]						38.5						37.7
MSPAN[99]	31.9	47.2	83.2	77.5	38.4	37.8	31.0	53.8	77.0	72.1	53.6	40.3
HQGA[131]	32.5	48.9	81.5	78.3	38.4	38.6	30.4	57.2	76.2	75.9	32.1	41.2
CoMVT[124]						39.5						42.6
SiaSamRea[121]						41.6						45.5
VQA-T[116]						41.5						46.3
VIOLET[126]						43.9						47.9
LiVLR[98]	50.3	77.1	94.2	81.3	48.4	59.4						

由 TGIF-QA, MSRVTT-QA, MSVD-QA 上统计的实验数据不难发现，用于 VideoQA 上的方法逐渐由常规注意力和记忆网络转向图神经网络与基于 Transformer 的预训练模型. 此外，也有研究者提出其他的算法模型同样有着出色的表现，如基于基础构建单元的 HCRN 与 HOSTR，它们在某种程度上为后续模型设计提供了另一种设计思路，如 HQGA.

3）Activitynet-QA[142]

Activitynet-QA 数据集来源于 Activitynet 视频数据集，由 YouTube 短片组成. 其中问题的类型分为 3 种：基于动作的问题，检测模型对粗略时序动作的理解；基于空间关系的问题，测试模型对静态帧的空

间推理能力；基于时序关系的问题，考察模型从一个序列的帧中推理对象的时序关系的能力. 答案类型分为 6 类: 是/否、数字、颜色、对象、位置和其他. 为了确保问题的质量，数据集限制问题和答案的长度，题目最多 20 个词，答案最多 5 个词. 与其他 VideoQA 数据集相比，Activitynet-QA 具有大规模、全人工注释的长视频，而且来源于生活，更加贴合实际，因此相对其他数据集更有意义.

4）YouTube2Text-QA[55]

YouTube2Text-QA 数据集是提出比较早的一个 VideoQA 数据集，来源于 YouTube2Text[157] 数据集. 其数据取自 YouTube 短视频，对于每一个视频片段，均由人工手动生成自然语言描述. 问题的类型为单项选择与开放式，问题的提问方式分为 who、what 和其他. YouTube2Text-QA 数据集规模相对较大，而且带人工注释的视频描述对于问答对的产生很方便，对模型训练也有很好的作用.

5）Video-QA[79]

Video-QA 数据集来源于互联网上的在线视频库. 每个视频通常有 3~5 个描述句子，描述性句子由视频提交者制作. 描述包含场景细节、演员、动作以及可能的非视觉信息. 问答类型也较为丰富. Video-QA 数据集是以最少的人力生成的一个较大规模的基于视频的问答数据集.

6）AGQA[150]

AGQA 源于 Action genome[158]，是一种新的组合时空推理的数据集，提供了一个评估视觉模型中各种维度的组合时空推理的基准. 其包含约 1.92 亿个不平衡问答对，这种不平衡的问答对将引起模型偏见，Grunde-McLaughlin 等人[150] 通过平衡答案分布和问题结构的类型来最大限度地减少这种偏见，将原始不平衡问答对集合减少为 390 万个问答对的平衡子集. 该数据集的语料库纯粹是基于视觉的，比现有的基准测试集大 3 个数量级，适用于评估模型复杂的多步推理能力.

7）其他

数据源为开放类型的数据集还包括 EgoVQA, Social-IQ, LifeQA, Tutorial-VQA, How2QA, TrafficQA, NExT-QA, STAR, Fill-in-the-Blank. EgoVQA 数据集是一个新颖的以自我为中心视角的 VideoQA 数据集，视频都是第一人称，其视频源是公共 IU Multiview 数据集，为多视图自中心视频研究而收集的. Social-IQ 数据集来自于 YouTube 上各种各样的视频，视频中的情景与事件贴近生活，其目标是分析由自然互动组成的非约定俗成的社交情境，是一个开拓性的真实世界无约束数据集，旨在评估现在和未来人工智能技术的社交智能. LifeQA 数据集来源于 YouTube 上人们在不同场景下的日常生活视频，这些视频均为在自然环境下带语音互动的视频；因其问答均与生活息息相关，非常有助于真实的问答系统. Tutorial-VQA 数据集由 76 个教程网站上的视频组成，视频均经过预处理，包含文本及每句话的时间戳信息，是一种用于在教程视频中寻找答案范围的新型数据集. How-2QA 数据集来源于 HowTo100M 和电视剧，视频类型具有多样性特点. 与 TVQA 类似，该数据集也为每个问题提供了相关时刻的开始点和结束点. TrafficQA 数据集通过在线和离线获取相结合的方式收集了覆盖各种真实的交通场景的视频，非常有助于交通场景中的辅助驾驶、违章检测等应用. NExT-QA 中的视频源于 YFCC-100M[159]，视频主要关注于现实的生活场景，内容具有丰富的对象交互. 该数据集的问题分为因果关系问题、时序性问题与描述性问题 3 种类型，旨在评估模型的因果动作推理与时间动作推理的能力. STAR[151] 源于 Charades[160] 数据集，它与 AGQA 同为真实世界场景的数据集. AGQA 中的任务设计侧重于时空关系，而 STAR 更关注基于现实情景的推理，侧重于人与对象交互、时间序列分析、动作预测和可行性推理. Fill-in-the-Blank 源于一个多语言视频字幕数据集 VaTeX[161]，通过掩码视频英文字幕中的名词性单词或短语得到具有多个正确答案的填空题，其余正确答案均由人工生成. EgoTaskQA[152] 在 LEMMA[162] 数据集基础上进行数据增强，与 EgoVQA 类似，也是一个以自我为中心的 VideoQA 数据集. 采用与 AGQA 同样的方式平衡答案分布和问题结构的类型来减少数据偏见，其问题设计借鉴 CLEVRER，旨在评估模型的时空和因果推理能力.

4 未来挑战与展望

4.1 挑 战

VideoQA 任务涉及计算机视觉与自然语言处理 2 个领域，面临着更为严峻的挑战，相对于 ImageQA 而言其应用前景更为广泛. 随着近几年研究者们在该任务上的不断改进与创新，众多模型被提出并在特定的数据集上有着出色表现. 同时，受到该任务的启发，应用于各种场景的数据集也层出不穷. 尽管目前有很多优秀的模型与数据集，但是能够真正应用于实际生活中的少之又少. 这意味着目前的模型大

多只是为了提升对特定数据集的性能，并不能在现实中实现人机互动，所以该任务还需要进行不断探索与研究. 总之，VideoQA 仍处于一个发展阶段，也必然存在着诸多问题与挑战.

4.1.1 模型的评估能力不足

目前开放式问答模型相对较少，其主要原因是对于该类型的问答难以制定合适的评估标准，其次真正意义上的开放式问答的数据集几乎没有. 传统的开放式问答任务是预定义一个答案集合，从中选择一个答案，其被视为一个多分类任务，与选择题任务类似. 即当前开放式问答任务的答案都属于一个潜在的答案集合，这并不符合人工智能的发展目标，所以生成式答案更适于人们的逻辑. 文献 [36, 63, 106] 提出的模型解决开放式问题，该模型根据视频与问题来生成一个自由形式的答案. 这种类型的开放式问答能够像人一样回答问题，所以其应用范围更广，更符合实际需求，但模型的准确率评估依然是个难题.

4.1.2 模型缺乏可解释性

目前，大多数模型不能够对其问答过程进行充分的可视可解释性分析，因而 VideoQA 的结果一直很难被完全信服. 因此，如何利用可视化的工具分析解释模型的内部机理尤为重要. 一旦能够进行可视化的机理解释，就可以给出一个通用的 VideoQA 模型范式，在各种不同类型的数据源间进行迁移学习.

4.1.3 模型的鲁棒性与泛化能力较弱

数据集缺陷是导致模型鲁棒性与泛化能力不足的重要原因. 部分数据集规模较小，学习样本不足将导致模型欠拟合，测试样本不足也将导致预测结果可靠性较低. 几乎近一半数据集中的问答对是通过固定问题模板结合程序自动生成，这使得问答类型缺乏多样性，很容易导致模型训练过拟合. 部分数据集存在偏见，这些数据集收集的问题有的更侧重于视觉信息，有的更侧重于文本信息，而有的不需要视觉与文本信息就能够正确回答. 这种数据集偏见可能会使得模型达到很好的训练效果，但在测试集表现出很大差异，难以泛化到其他数据集.

4.2 展 望

基于 4.1 节所述的 VideoQA 问题与挑战，未来的研究工作可以从以下 5 个方面进行开展.

4.2.1 构建更加完善的数据集

好的数据集是训练优秀模型的前提. 首先数据集规模不能太小，可以在必要时通过数据增强来扩充数据集，由此达到增强模型的鲁棒性、提升模型泛化能力的目的. 其次可以加入对抗样本训练，提升模型在对抗样本的鲁棒性，但是若模型过于鲁棒，其泛化能力就会下降. 然后，数据集中的各种类型的问题需要均衡，比如根据单一模态回答的问题与需要根据融合模态回答的问题比例不要相差太大，只需单一模态就能回答的问题的数量最好也不要偏向某一模态信息. 对于具体问题，例如涉及到状态转移、计数等时序性问题的比例也需要提高. 只有均衡的数据集才能够正确地评价 VideoQA 模型的能力.

4.2.2 多模态协同学习

相比于图像，视频具有更加丰富的多模态信息，包括音频、字幕、光流等. 当前绝大多数 VideoQA 模型在处理视频数据时，以视频的视觉模态为主；也有一些方法同时利用视频的视觉模态和字幕模态，但通常只是对多模态特征进行简单地融合，如简单的点乘、拼接或双线性池化等[31-33]，这种融合方式很难充分利用模态之间的互补信息. 为了更加充分地利用视频的多模态信息：一方面可以同时引入更多不同模态信息而不只是 2 种模态；另一方面可以通过不同模态之间的协同学习而不只是简单的模态融合来挖掘多模态信息的潜力，比如通过不同模态分支之间的互学习来提升对视频的表示能力. 此外，并不是所有视频都存在各个模态的信息，比如某个视频可能没有音频信号，如何处理某些模态丢失的情况是十分有价值的研究问题；同时，对多模态信息的建模会增加模型的复杂度，如何构建更加轻量的多模态协同使用的模型也是未来值得研究的问题.

4.2.3 加强因果关系推理

在 ImageQA 中，Wang 等人[163] 提出了 QA R-CNN 的模型，同时构建了 EST-VQA 数据集. 该数据集加入了支撑答案的证据，模型在推理答案时会提供预测该答案的支撑证据. 文献 [164−165] 也提出了使用显式知识进行因果推理的方法，可以使模型预测的答案更具有可解释性. 与 ImageQA 相同，VideoQA 也更加需要使预测的答案具有可解释性，因为大多数模型依赖于预定义的答案池，无法处理词汇表之外的问题答案. 因此这些模型是否真正具有理解和推理问题的能力，还是仅仅对固定答案空间的过度拟合很难知晓. 所以在构建视频数据集时加入支撑答案的证据也是未来的必要工作，让 VideoQA 模型进行答案推理，并提供支撑证据保证了推理过程的因果关系，也使得预测答案更加合理.

4.2.4 外部知识的引入

目前，大部分 VideoQA 模型只关注数据集中可

利用的视觉与文本信息,然而忽略了并不是所有的问题都能够凭借数据集提供的信息进行回答. 由于数据集本身提供的信息有限,无论是在 ImageQA 还是在 VideoQA 中,仅仅利用给定的视觉与文本信息来回答问题往往是不充分的,VideoQA 任务中的部分问题更需要结合先验知识进行推理. 这些知识包括常识知识、关系知识等,它们一般通过在线获取或者手动构建. 将与问题相关的实体对象与外部知识进行关联,从而提升模型对视频和问题的理解程度以达到知识推理的准确性.

如图 17 所示,这是一个涉及视觉与常识知识的问题[166]. 要正确回答“地面上的红色物体能用来做什么?”,模型所凭借的不仅源于图像上所识别的“消防栓”这单一信息,而且必须依靠来自外部的常识知识,即“消防栓能灭火”作为支撑的事实,才能正确给出“灭火”这一答案. 这是一个 ImageQA 模型上利用外部知识的场景,此外文献 [167−172] 提出 ImageQA 的模型均结合了外部知识进行推理,均取得了优异的实验结果. 同理,VideoQA 包含更丰富的信息,推理过程中更加需要外部知识的支撑. 目前已有少数工作如文献 [104, 137−138] 在 VideoQA 模型推理时合理地查询外部知识,进一步提升了模型回答的准确率. 同时,外部知识也可以解决现有方法在基于特定数据集训练的模型泛化能力弱的问题,所以,如何将外部知识与 VideoQA 模型结合起来也是未来要讨论的重点.



Fig. 17 Examples of questions that require visual and common sense knowledge to answer[166]

图 17 需要视觉和常识知识来回答的问题示例[166]

4.2.5 预训练与提示的结合

近 2 年来,视觉语言预训练模型在从大规模数据中学习联合视觉-文本表示方面取得了巨大成功. 预训练模型能够流行起来的一个重要原因是用于训练这些强大的视觉语言模型的大规模数据可以很容易地从互联网上获取到,而无需任何耗时费力的手动注释. 因此,有理由相信,随着数据集规模的不断增大,在不久的将来会训练出用于下游任务更强大的模型.

将预训练模型用于特定的下游任务,比较流行的方法是微调 (Fine-tuning). 而现在有研究者希望用提示 (Prompting) 来代替原来的 Fine-tuning 方法,它不同于 Fine-tuning 改造原有模型参数的方式,Prompting 则是将下游任务的输入输出形式改造成预训练任务中的形式. Radford 等人[173] 提出 CLIP 模型,给定其适当手工设计的 Prompting,使模型有效地缩小预训练和下游任务之间的差距. Ju 等人[174] 在 CLIP 基础上提出通过学习特定任务的提示向量来实现高效和轻量级的模型适应,并将 CLIP 的图像理解扩展到视频理解,并增加了对时间维度的处理. 其在动作识别与文本视频检索任务中的性能均优于现有方法,相信未来将其应用到 VideoQA 任务中也将会取得优异的性能表现.

5 总 结

本文主要对该领域的发展现状、各种模型框架以及不同的基准数据集进行了回顾. 对比分析了 VideoQA 任务与 ImageQA 任务两者的重要区别与挑战,主要区别是视频相对于图片具有更复杂的语义信息. 同时对用于该任务的各种模型进行了详细的分析与讨论:注意力机制的应用能够像人一样关注到视频与问题中的有效信息;记忆机制加上注意力能够对关键信息不断更新与保存;利用图网络建模视频时空结构进行联合推理更能准确地解决时空相关问题;利用海量数据进行预训练基于 Transformer 或 BERT 的模型具有较好的鲁棒性. 本文对用于该任务的数据集也进行了细致介绍,并分析了部分模型在数据集上的实验结果,最后指出了目前数据集与模型存在的一些不足,数据集存在的局限性导致模型鲁棒性与泛化能力不足,以及模型本身缺乏可解释性. 虽然 VideoQA 被提出已有五六年之久,但该领域仍处于发展阶段,很少有实际落地的应用. 随着越来越多的研究者们的加入,相信在不远的将来,VideoQA 技术一定会应用于现实生活中.

作者贡献声明: 包翠竹负责课题设计、文献归纳、论文撰写与修改;丁凯负责论文撰写、文献整理与数

据收集；董建峰负责课题构思、论文修改与结构设计；杨勋负责论文指导与修改以及提供材料支持；谢满德负责规划论文整体结构、提出论文修改意见；王勋负责论文指导与修改、论文审阅.

参 考 文 献

[1] Yu Jun, Wang Liang, Yu Zhou. Research on visual question answering technology[J]. Journal of Computer Research and Development, 2018, 55(9): 1946−1958 (in Chinese)
(俞俊, 汪亮, 余宙. 视觉问答技术研究[J]. 计算机研究与发展, 2018, 55(9): 1946−1958)

[2] Antol S, Agrawal A, Lu Jiasen, et al. VQA: Visual question answering[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 2425−2433

[3] Yang Zichao, He Xiaodong, Gao Jianfeng, et al. Stacked attention networks for image question answering[C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 21−29

[4] Qiao Tingting, Dong Jianfeng, Xu Duanqing. Exploring human-like attention supervision in visual question answering[C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 7300−7307

[5] Yu Zhou, Yu Jun, Cui Yuhao, et al. Deep modular co-attention networks for visual question answering[C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 6281−6290

[6] Luo Haozheng, Qin Ruiyang. Open-ended multi-modal relational reason for video question answering[J]. arXiv preprint, arXiv: 2012. 00822, 2020

[7] Bao Xigang, Zhou Chunlai, Xiao Kejing, et al. Review of visual question answering research[J]. Journal of Software, 2021, 32(8): 2522−2544 (in Chinese)
(包希港, 周春来, 肖克晶, 等. 视觉问答研究综述[J]. 软件学报, 2021, 32(8): 2522−2544)

[8] Patel D, Parikh R, Shastri Y. Recent advances in video question answering: A review of datasets and methods[C]//Proc of the Int Conf on Pattern Recognition. Berlin: Springer, 2021: 339−356

[9] Khurana K, Deshpande U. Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey[J]. IEEE Access, 2021, 9: 43799−43823

[10] Sun Guanglu, Liang Lili, Li Tianlin, et al. Video question answering: A survey of models and datasets[J]. Mobile Networks and Applications, 2021, 26(5): 1904−1937

[11] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]// Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 2204−2212

[12] Weston J, Chopra S, Bordes A. Memory networks[J]. arXiv preprint, arXiv: 1410. 3916, 2014

[13] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61−80

[14] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint, arXiv: 1609. 02907, 2016

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Proc of the 30th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 5998−6008

[16] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the 17th Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[17] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137−1149

[18] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proc of the 22nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248−255

[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint, arXiv: 1409. 1556, 2014

[20] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions[C]//Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1−9

[21] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[22] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 4489−4497

[23] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6299−6308

[24] Xie Saining, Sun Chen, Huang Jonathan, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 305−321

[25] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6546−6555

[26] Feichtenhofer C, Fan Haoqi, Malik J, et al. SlowFast networks for video recognition[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6202−6211

[27] Le H, Chen N F, Hoi S C H. VGNMN: Video-grounded neural module network to video-grounded language tasks[J]. arXiv preprint, arXiv: 2104. 07921, 2021

[28] Shah A, Lin T H, Wu Shijie. Triple attention network architecture for

MovieQA[J]. arXiv preprint, arXiv: 2111. 09531, 2021

[29] Aytar Y, Vondrick C, Torralba A. SoundNet: Learning sound representations from unlabeled video [C]//Proc of the 29th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2016: 892−900

[30] Kumar A, Khadkevich M, Fügen C. Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes [C]//Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 326−330

[31] Kim K M, Heo M O, Choi S H, et al. Deepstory: Video story QA by deep embedded memory networks [C]//Proc of the 26th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 2016−2022

[32] Na S, Lee S, Kim J, et al. A read-write memory network for movie story understanding [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 677−685

[33] Yang Tianhao, Zha Zhengjun, Xie Hongtao, et al. Question-aware tube-switch network for video question answering [C]//Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1184−1192

[34] Gao Jiyang, Ge Runzhou, Chen Kan, et al. Motion-appearance co-memory networks for video question answering [C]//Proc of the 31st IEEE CONF on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6576−6585

[35] Wang Bo, Xu Youjiang, Han Yahong, et al. Movie question answering: Remembering the textual cues for layered visual contents [C]//Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 7380−7387

[36] Castro S, Wang Ruoyao, Huang Pingxuan, et al. FIBER: Fill-in-the-blanks as a challenging video understanding evaluation framework [C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 2925−2940

[37] Zhu Linchao, Yang Yi. Actbert: Learning global-local video-text representations [C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 8746−8755

[38] Mikolov T, Chen Kai, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, arXiv: 1301. 3781, 2013

[39] Pennington J, Socher R, Manning C D. GloVe: Global vectors for word representation [C]//Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1532−1543

[40] Kiros R, Zhu Yukun, Salakhutdinov R R, et al. Skip-Thought vectors [C]//Proc of the 28th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2015: 3294−3302

[41] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735−1780

[42] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]//Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1724−1734

[43] Zhao Zhou, Lin Jinghao, Jiang Xinghua, et al. Video question answering via hierarchical dual-level attention network learning [C]//Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1050−1058

[44] Xue Hongyang, Chu Wenqing, Zhao Zhou, et al. A better way to attend: Attention with trees for video question answering[J]. IEEE Transactions on Image Processing, 2018, 27(11): 5563−5574

[45] Jang Y, Song Y, Yu Y, et al. TGIF-QA: Toward spatio-temporal reasoning in visual question answering [C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2758−2766

[46] Falcon A, Lanz O, Serra G. Data augmentation techniques for the video question answering task [C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 511−525

[47] Mazaheri A, Zhang Dong, Shah M. Video fill in the blank using LR/RL LSTMs with spatial-temporal attentions [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1407−1416

[48] Xu Dejing, Zhao Zhou, Xiao Jun, et al. Video question answering via gradually refined attention over appearance and motion [C]//Proc of the 25th ACM Int Conf on Multimedia. New York: ACM, 2017: 1645−1653

[49] Chao Guanlin, Rastogi A, Yavuz S, et al. Learning question-guided video representation for multi-turn video question answering [C]//Proc of the 20th Annual SIGDIAL Meeting on Discourse and Dialogue. Stroudsburg, PA: ACL, 2019: 215−225

[50] Zhao Zhou, Zhang Zhu, Xiao Shuwen, et al. Open-ended long-form video question answering via adaptive hierarchical reinforced networks [C]//Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 3683−3689

[51] Kim J, Ma M, Kim K, et al. Gaining extra supervision via multi-task learning for multi-modal video question answering [C]//Proc of the 2019 Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2019: 1−8

[52] Lei Jie, Yu Licheng, Berg T L, et al. TVQA+: Spatio-temporal grounding for video question answering [C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 8211−8225

[53] Gao Difei, Wang Ruiping, Bai Ziyi, et al. Env-QA: A video question answering benchmark for comprehensive understanding of dynamic environments [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 1675−1685

[54] Yu Y, Kim J, Kim G. A joint sequence fusion model for video question answering and retrieval [C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 471−487

[55] Ye Yunan, Zhao Zhou, Li Yimeng, et al. Video question answering via attribute-augmented attention network learning [C]//Proc of the 40th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2017: 829−832

[56] Chowdhury M I H, Nguyen K, Sridharan S, et al. Hierarchical relational attention for video question answering [C]//Proc of the 25th IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 2018: 599−603

[57] Zhao Zhou, Jiang Xinghua, Cai Deng, et al. Multi-turn video question answering via multi-stream hierarchical attention context network [C]//Proc of the 27th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 3690−3696

[58] Zhao Zhou, Yang Qifan, Cai Deng, et al. Video question answering via hierarchical spatio-temporal attention networks [C]//Proc of the 26th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 3518−3524

[59] Song Xiaomeng, Shi Yucheng, Chen Xin, et al. Explore multi-step reasoning in video question answering [C]//Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 239−247

[60] Jiang Jianwen, Chen Ziqiang, Lin Haojie, et al. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering [C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11101−11108

[61] Liang Junwei, Jiang Lu, Cao Liangliang, et al. Focal visual-text attention for visual question answering [C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6135−6143

[62] Yu Zhou, Yu Jun, Fan Jianping, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1821−1830

[63] Xue Hongyang, Zhao Zhou, Cai Deng. Unifying the video and question attentions for open-ended video question answering[J]. IEEE Transactions on Image Processing, 2017, 26(12): 5656−5666

[64] Chu Wenqing, Xue Hongyang, Zhao Zhou, et al. The forgettable-watcher model for video question answering[J]. Neurocomputing, 2018, 314: 386−393

[65] Gao Lianli, Zeng Pengpeng, Song Jingkuan, et al. Structured two-stream attention network for video question answering [C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6391−6398

[66] Li Xiangpeng, Gao Lianli, Wang Xuanhan, et al. Learnable aggregating net with diversity learning for video question answering [C]//Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1166−1174

[67] Li Xiangpeng, Song Jingkuan, Gao Lianli, et al. Beyond RNNs: Positional self-attention with co-attention for video question answering [C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 8658−8665

[68] Kim K M, Choi S H, Kim J H, et al. Multimodal dual attention memory for video story question answering [C]//Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 673−688

[69] Lei Jie, Yu Licheng, Bansal M, et al. TVQA: Localized, compositional video question answering [C]//Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 1369−1379

[70] Li Fangtao, Bai Ting, Cao Chenyu, et al. Relation-aware hierarchical attention framework for video question answering[C]//Proc of the 2021 Int Conf on Multimedia Retrieval. New York: ACM, 2021: 164−172

[71] Kim J, Ma M, Pham T, et al. Modality shifting attention network for multi-modal video question answering [C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10106−10115

[72] Jin Weike, Zhao Zhou, Gu Mao, et al. Multi-interaction network with object relation for video question answering [C]//Proc of the 27th ACM Int Conf on Multimedia. New York: ACM, 2019: 1193−1201

[73] Kim H, Tang Zineng, Bansal M. Dense-caption matching and frame-selection gating for temporal localization in VideoQA [C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4812−4822

[74] Chadha A, Arora G, Kaloty N. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering [C]//Proc of the 2021 IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2021: 1−13

[75] Seo M, Kembhavi A, Farhadi A, et al. Bidirectional attention flow for machine comprehension [C/OL]//Proc of the Int Conf on Learning Representations. 2017 [2022-01-10].https://openreview.net/forum? id=HJ0UKP9ge

[76] Yu A W, Dohan D, Luong M T, et al. Qanet: Combining local convolution with global self-attention for reading comprehension [C/OL]//Proc of the Int Conf on Learning Representations. 2018 [2022-01-10].https://openreview.net/forum?id=B14TlG-RW

[77] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks [C/OL]//Proc of the Int Conf on Learning Representations. 2018 [2022-01-12].https://openreview.net/forum?id=rJXMpikCZ

[78] Sukhbaatar S, Weston J, Fergus R. End-to-end memory networks [C]//Proc of the 28th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2015: 2440−2448

[79] Zeng K H, Chen T H, Chuang C Y, et al. Leveraging video descriptions to learn video question answering [C]//Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 4334−4340

[80] Tapaswi M, Zhu Yukun, Stiefelhagen R, et al. MovieQA: Understanding stories in movies through question-answering [C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4631−4640

[81] Cai Jiayin, Yuan Chun, Shi Cheng, et al. Feature augmented memory with global attention network for VideoQA [C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2021: 998−1004

[82] Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: Dynamic memory networks for natural language processing [C]//Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 1378−1387

[83] Xiong Caiming, Merity S, Socher R. Dynamic memory networks for

visual and textual question answering［C］//Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 2397−2406

［84］ Fan Chenyou, Zhang Xiaofan, Zhang Shu, et al. Heterogeneous memory enhanced multimodal attention model for video question answering［C］//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 1999−2007

［85］ Kim J, Ma M, Kim K, et al. Progressive attention memory network for movie story question answering［C］//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8337−8346

［86］ Yu Ting, Yu Jun, Yu Zhou, et al. Long-term video question answering via multimodal hierarchical memory attentive networks［J］. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(3): 931−944

［87］ Fukui A, Park D H, Yang D, et al. Multimodal compact bilinear pooling for visual question answering and visual grounding［C］//Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 457−468

［88］ Kim J H, On K W, Lim W, et al. Hadamard product for low-rank bilinear pooling［C/OL］//Proc of the Int Conf on Learning Representations.2017［2022-01-20］.https://openreview.net/forum?id=r1rhWnZkg

［89］ Wang Zhichun, Lv Qingsong, Lan Xiaohan, et al. Cross-lingual knowledge graph alignment via graph convolutional networks［C］//Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 349−357

［90］ Fan Wenqi, Ma Yao, Li Qing, et al. Graph neural networks for social recommendation［C］//Proc of the 2019 World Wide Web Conf. New York: ACM, 2019: 417−426

［91］ Huang Deng, Chen Peihao, Zeng Runhao, et al. Location-aware graph convolutional networks for video question answering［C］//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11021−11028

［92］ Jiang Pin, Han Yahong. Reasoning with heterogeneous graph alignment for video question answering［C］//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11109−11116

［93］ Seo A, Kang G C, Park J, et al. Attend what you need: Motion-appearance synergistic networks for video question answering［C］//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2021: 6167–6177

［94］ Park J, Lee J, Sohn K. Bridge to Answer: Structure-aware graph interaction network for video question answering［C］//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 15526−15535

［95］ Wang Jianyu, Bao Bingkun, Xu Changsheng. DualVGR: A dual-visual graph reasoning unit for video question answering［J］. IEEE Transactions on Multimedia, 2021, 24: 3369−3380

［96］ Wang Xiao, Zhu Meiqi, Bo Deyu, et al. AM-GCN: Adaptive multi-channel graph convolutional networks［C］//Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2020: 1243−1253

［97］ Dang L H, Le T M, Le V, et al. Object-centric representation learning for video question answering［C/OL］//Proc of the 2021 Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2021［2022-01-22］. https://arxiv.org/abs/2104.05166

［98］ Jiang Jingjing, Liu Ziyi, Zheng Nanning, et al. LiVLR: A lightweight visual-linguistic reasoning framework for video question answering［J/OL］. IEEE Transactions on Multimedia, 2022［2022-02-22］.https://github.com/jingjing12110/LiVLR-VideoQA

［99］ Guo Zhicheng, Zhao Jiaxuan, Jiao Licheng, et al. Multi-scale progressive attention network for video question answering［C］//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing (Volume 2: Short Papers). Stroudsburg, PA: ACL, 2021: 973−978

［100］ Peng Liang, Yang Shuangji, Bin Yi, et al. Progressive graph attention network for video question answering［C］//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 2871−2879

［101］ Liu Fei, Liu Jing, Wang Weining, et al. HAIR: Hierarchical visual-semantic relational reasoning for video question answering［C］//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 1698−1707

［102］ Yang Zekun, Garcia N, Chu Chenhui, et al. Bert representations for video question answering［C］//Proc of the 2020 IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2020: 1556−1565

［103］ Urooj Khan A, Mazaheri A, da Vitoria Lobo N, et al. MMFT-BERT: Multimodal fusion transformer with BERT encodings for visual question answering［C］//Proc of the 2020 Conf on Empirical Methods in Natural Language Processing(Findings). Stroudsburg, PA: ACL, 2020: 4648−4660

［104］ Garcia N, Nakashima Y. Knowledge-based video question answering with unsupervised scene descriptions［C］//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 581−598

［105］ Engin D, Schnitzler F, Duong N Q K, et al. On the hidden treasure of dialog in video question answering［C］//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 2064−2073

［106］ Sadhu A, Chen Kan, Nevatia R. Video question answering with phrases via semantic roles［C］//Proc of the 19th Int Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2021: 2460−2478

［107］ Ganesan A, Pal D, Muthuraman K, et al. Video based contextual question answering［J］. arXiv preprint, arXiv: 1804. 07399, 2018

［108］ Cherian A, Hori C, Marks T K, et al. (2.5+ 1) D spatio-temporal scene graphs for video question answering［C］//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 444−453

［109］ Peng Min, Wang Chongyang, Gao Yuan, et al. Temporal pyramid transformer with multimodal interaction for video question answering［J］. arXiv preprint, arXiv: 2109. 04735, 2021

[110] Tan Hao, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers [C]//Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 5100−5111

[111] Sun Chen, Myers A, Vondrick C, et al. Videobert: A joint model for video and language representation learning [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 7464−7473

[112] Chen Xinlei, Fang Hao, Lin T Y, et al. Microsoft COCO captions: Data collection and evaluation server[J]. arXiv preprint, arXiv: 1504. 00325, 2015

[113] Krishna R, Zhu Yuke, 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

[114] Miech A, Zhukov D, Alayrac J B, et al. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 2630−2640

[115] Kim S, Jeong S, Kim E, et al. Self-supervised pre-training and contrastive representation learning for multiple-choice video QA[J]. arXiv preprint, arXiv: 2009. 08043, 2020

[116] Yang A, Miech A, Sivic J, et al. Just ask: Learning to answer questions from millions of narrated videos [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 1686−1697

[117] Li Linjie, Chen Y C, Cheng Yu, et al. HERO: Hierarchical encoder for video+ language omni-representation pre-training [C]//Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 2046−2065

[118] Zellers R, Lu Ximing, Hessel J, et al. MERLOT: Multimodal neural script knowledge models [C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 23634−23651

[119] Liu Yinhan, Ott M, Goyal N, et al. RoBERTa: A robustly optimized bert pretraining approach [C/OL]//Proc of the Int Conf on Learning Representations.2020[2022-01-25].https://openreview.net/forum?id=SyxS0T4tvS

[120] Lei Jie, Li Linjie, Zhou Luowei, et al. Less is more: Clipbert for video-and-language learning via sparse sampling [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 7331−7341

[121] Yu Weijiang, Zheng Haoteng, Li Mengfei, et al. Learning from Inside: Self-driven siamese sampling and reasoning for video question answering [C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 26462−26474

[122] Amrani E, Ben-Ari R, Rotman D, et al. Noise estimation using density estimation for self-supervised multimodal learning [C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 6644−6652

[123] Luo Jianjie, Li Yehao, Pan Yingwei, et al. CoCo-BERT: Improving video-language pre-training with contrastive cross-modal matching and denoising [C]//Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 5600−5608

[124] Seo P H, Nagrani A, Schmid C. Look before you speak: Visually contextualized utterances [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 16877−16887

[125] Lu Jiasen, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [C]// Proc of the 32nd Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2019: 13−23

[126] Fu T J, Li Linjie, Gan Zhe, et al. VIOLET: End-to-end video-language transformers with masked visual-token modeling[J]. arXiv preprint, arXiv: 2111. 12681, 2021

[127] Liu Ze, Ning Jia, Cao Yue, et al. Video swin transformer [C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 3202−3211

[128] Zhou Luowei, Liu Jingjing, Cheng Yu, et al. Cupid: Adaptive curation of pre-training data for video-and-language representation learning[J]. arXiv preprint, arXiv: 2104. 00285, 2021

[129] Le T M, Le V, Venkatesh S, et al. Hierarchical conditional relation networks for video question answering [C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 9972−9981

[130] Dang L H, Le T M, Le V, et al. Hierarchical object-oriented spatio-temporal reasoning for video question answering [C]//Proc of the 30th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2021: 636−642

[131] Xiao Junbin, Yao A, Liu Zhiyuan, et al. Video as conditional graph hierarchy for multi-granular question answering [C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 2804−2812

[132] Yi Kexin, Gan Chuang, Li Yunzhu, et al. CLEVRER: Collision events for video representation and reasoning [C/OL]//Proc of the Int Conf on Learning Representations. 2020 [2022-09-03].https://openreview.net /forum?id=HkxYzANYDB

[133] Chen Zhenfang, Mao Jiayuan, Wu Jiajun, et al. Grounding physical concepts of objects and events through dynamic visual reasoning [C/OL]//Proc of the Int Conf on Learning Representations. 2021 [2022-09-03].https://openreview.net/pdf?id=bhCDO_cEGCz

[134] Ding Mingyu, Chen Zhenfang, Du Tao, et al. Dynamic visual reasoning by learning differentiable physics models from video and language [C]//Proc of the 34th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021, 34: 887−899

[135] Mao Jiayuan, Gan Chuang, Kohli P, et al. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision [C/OL]//Proc of the Int Conf on Learning Representations. 2019 [2022-09-04].https://research.ibm.com/publications/the-neuro-symbolic-concept-learner-interpreting-scenes-words-and-sentences-from-natural-supervision

[136] Xu Li, Huang He, Liu Jun. SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9878−9888

[137] Garcia N, Otani M, Chu Chenhui, et al. KnowIT VQA: Answering knowledge-based questions about videos [C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 10826−10834

[138] Han Yahong, Wang Bo, Hong Richang, 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

[139] Maharaj T, Ballas N, Rohrbach A, et al. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering [C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6884−6893

[140] Mun J, Hongsuck Seo P, Jung I, et al. MarioQA: Answering questions by watching gameplay videos [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2867−2875

[141] Ates T, Atesoglu M S, Yigit C, et al. CRAFT: A benchmark for causal reasoning about forces and interactions [C]//Proc of the 2022 Findings of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 2602–2627

[142] Yu Zhou, Xu Dejing, Yu Jun, et al. Activitynet-QA: A dataset for understanding complex web videos via question answering [C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 9127−9134

[143] Torabi A, Tandon N, Sigal L. Learning language-visual embedding for movie understanding with natural-language[J]. arXiv preprint, arXiv: 1609. 08124, 2016

[144] Fan Chenyou. EgoVQA-an egocentric video question answering benchmark dataset [C]//Proc of the 2019 IEEE Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2019: 4359−4366

[145] Zadeh A, Chan M, Liang P P, et al. Social-IQ: A question answering benchmark for artificial social intelligence [C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8807−8817

[146] Choi S, On K W, Heo Y J, et al. DramaQA: Character-centered video story understanding with hierarchical QA [C]//Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI 2021: 1166−1174

[147] Castro S, Azab M, Stroud J, et al. LifeQA: A real-life dataset for video question answering [C]//Proc of the 12th Language Resources and Evaluation Conf. Marseille: European Language Resources Association (ELRA), 2020: 4352−4358

[148] Colas A, Kim S, Dernoncourt F, et al. Tutorial-VQA: Question answering dataset for tutorial videos [C]//Proc of the 12th Language Resources and Evaluation Conf. Marseille: European Language Resources Association (ELRA), 2020: 5450–5455

[149] Xiao Junbin, Shang Xindi, Yao A, et al. NExT-QA: Next phase of question-answering to explaining temporal actions [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9777−9786

[150] Grunde-McLaughlin M, Krishna R, Agrawala M. AGQA: A benchmark for compositional spatio-temporal reasoning [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 11287−11297

[151] Wu Bo, Yu Shoubin, Chen Zhenfang, et al. STAR: A benchmark for situated reasoning in real-world videos [C/OL]//Proc of the 35th Conf on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2). Cambridge, MA: MIT Press, 2021 [2022-01-25].https://openreview.net/forum?id=EfgNF5-ZAjM

[152] Jia Baoxiong, Lei Ting, Zhu Songchun, et al. EgoTaskQA: Understanding human tasks in egocentric videos [C/OL]//Proc of the 36th Advances in Neural Information Processing Systems Datasets and Benchmarks Track. Cambridge, MA: MIT, 2022 [2022-01-25].https://openreview.net/forum?id=ttxAvIQA4i_

[153] Jasani B, Girdhar R, Ramanan D. Are we asking the right questions in MovieQA? [C]//Proc of the IEEE Int Conf on Computer Vision Workshops. Piscataway, NJ: IEEE, 2019: 1879−1882

[154] Rohrbach A, Torabi A, Rohrbach M, et al. Movie description[J]. International Journal of Computer Vision, 2017, 123(1): 94−120

[155] Kolve E, Mottaghi R, Han W, et al. AI2-THOR: An interactive 3D environment for visual AI[J]. arXiv preprint, arXiv: 1712. 05474, 2017

[156] Li Yuncheng, Song Yale, Cao Liangliang, et al. TGIF: A new dataset and benchmark on animated Gif description [C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4641−4650

[157] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2013: 2712−2719

[158] Ji Jingwei, Krishna R, Fei-Fei L, et al. Action genome: Actions as compositions of spatio-temporal scene graphs [C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 10236−10247

[159] Thomee B, Shamma D A, Friedland G, et al. YFCC-100M: The new data in multimedia research[J]. Communications of the ACM, 2016, 59(2): 64−73

[160] Sigurdsson G A, Russakovsky O, Gupta A. What actions are needed for understanding human actions in videos? [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 2137−2146

[161] Wang Xin, Wu Jiawei, Chen Junkun, et al. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research [C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 4581−4591

[162] Jia Baoxiong, Chen Yixin, Huang Siyuan, et al. LEMMA: A multi-view dataset for learning multi-agent multi-task activities [C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 767−786

[163] Wang Xinyu, Liu Yuliang, Shen Chunhua, et al. On the general value of evidence, and bilingual scene-text visual question answering [C]//Proc of the 33rd IEEE Conf on Computer Vision and Pattern

Recognition. Piscataway, NJ: IEEE, 2020: 10126−10135

[164] Marino K, Chen Xinlei, Parikh D, et al. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14111−14121

[165] Zhang Yifeng, Jiang Ming, Zhao Qi. Explicit knowledge incorporation for visual reasoning [C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1356−1365

[166] Wang Peng, Wu Qi, Shen Chunhua, et al. FVQA: Fact-based visual question answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(10): 2413−2427

[167] Wu Qi, Wang Peng, Shen Chunhua, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources [C]//Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4622−4630

[168] Marino K, Rastegari M, Farhadi A, et al. Ok-VQA: A visual question answering benchmark requiring external knowledge [C]//Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3195−3204

[169] Shah S, Mishra A, Yadati N, et al. KVQA: Knowledge-aware visual question answering [C]//Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 8876−8884

[170] Chen Zhou, Chen Jiaoyan, Geng Yuxia, et al. Zero-shot visual question answering using knowledge graph [C]//Proc of the 20th Int Semantic Web Conf. Berlin: Springer, 2021: 146−162

[171] Wu Jialin, Lu Jiasen, Sabharwal A, et al. Multi-modal answer validation for knowledge-based VQA [C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022: 2712−2721

[172] Wu Qi, Shen Chunhua, Wang Peng, et al. Image captioning and visual question answering based on attributes and external knowledge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1367−1381

[173] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[174] Ju Chen, Han Tengda, Zheng Kunhao, et al. Prompting visual-language models for efficient video understanding [C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2022: 105−124



Bao Cuizhu, born in 1995. PhD, lecturer. Member of CCF. Her main research interests include computer vision, intelligent traffic control, and smart city.

包翠竹, 1990 年生. 博士, 讲师.CCF 会员. 主要研究方向为计算机视觉、智能交通控制、智慧城市.



Ding Kai, born in 1995. Master candidate. Student member of CCF. His main research interests include computer vision, and visual question answering, and video question answering.

丁　凯, 1995 年生. 硕士研究生.CCF 学生会员. 主要研究方向为计算机视觉、视觉问答、视频问答.



Dong Jianfeng, born in 1991. PhD, professor. Member of CCF. His main research interests include multimedia understanding, computer vision, machine learning, and artificial intelligence.

董建峰, 1991 年生.博士, 教授.CCF 会员. 主要研究方向为多媒体理解、计算机视觉、机器学习、人工智能.



Yang Xun, born in 1989. PhD, professor. Member of CCF. His main research interests include cross-media analysis and reasoning, structured understanding of multimedia content, and visual media correlation modeling.

杨　勋, 1989 年生.博士, 教授.CCF 会员. 主要研究方向为跨媒体分析与推理、多媒体内容结构化理解、视觉媒体相关性关联建模.



Xie Mande, born in 1977. PhD, professor. Member of CCF. His main research interests include wireless sensor networks, cloud computing, network security, and swarm intelligence perception.

谢满德, 1977 年生.博士. 教授. CCF 会员.主要研究方向为无线传感器网络、云计算、网络安全、群智感知.



Wang Xun, born in 1967. PhD, professor. Distinguished member of CCF. His main research interests include visual media big data technology, mobile graphics computing, computer vision, and intelligent information processing and visual analysis.

王　勋, 1967 年生.博士, 教授.CCF 杰出会员. 主要研究方向为可视媒体大数据技术、移动图形计算、计算机视觉、智能信息处理与可视分析.