嵌入局部聚类描述符的视频问答Transformer模型

郭 丹[1),2),3),4)] 姚沈涛[1)] 王 辉[1)] 汪 萌[1),2),3),4)]

[1)](合肥工业大学计算机与信息学院 合肥 230601)

[2)](合肥综合性国家科学中心人工智能研究院 合肥 230094)

[3)](大数据知识工程教育部重点实验室(合肥工业大学) 合肥 230601)

[4)](智能互联系统安徽省实验室(合肥工业大学) 合肥 230601)

摘 要 视频问答(Video Question Answering)是典型的跨模态理解任务,其目的是根据提问的文本对视频内容进行理解并推理正确的答案,如何有效地对多模态输入进行特征表示并建立跨模态间复杂的语义关联是解决这一任务的关键难点.为了正确地推理结果,模型首先必须捕获视频序列和复杂文本中包含的关键语义信息.本文提出了一种嵌入局部聚类描述符的视频问答Transformer模型,称为TVLAD-Net(Transformer Residual-less VLAD Network).TVLAD-Net主要包含一个端到端可训练的无残差局部聚合描述符模块(RVLAD, Residual-less Vector of Local Aggregated Descriptor),以及一个统一的语义转换模块(Transformer).具体来说,RVLAD通过设置多个不同的聚类中心将视频和文本特征分别聚合为少量紧凑的局部聚类描述符;每个聚类描述符从全局角度分配及汇总了序列上权重不一的语义信息,相比于聚合前的视频帧特征或文本词特征具有更丰富的表征能力.Transformer模块能够利用模态间的相互语义引导,实现多模态聚类描述符的语义交互,即采用多头注意力机制同时求解模态内和模态间的语义关联,进而避免了与所求解问题无关或者冗余的描述符语义单元的聚合.实验评估在TGIF-QA、MSVD-QA和MSRVTT-QA三个基准数据集上进行;实验结果表明本文方法能够实现先进的问答推理,在整体的评价指标上与现有方法相比有2%~5%的性能提升.

关键词 视频问答;多模态数据;聚类描述符;自注意力变换网络;深度学习

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2023.00671

Embedding VLAD in Transformer for Video Question Answering

GUO Dan[1),2),3),4)] YAO Shen-Tao[1)] WANG Hui[1)] WANG Meng[1),2),3),4)]

[1)](*School of Computer and Information Engineering*, *Hefei University of Technology*, *Hefei* 230601)

[2)](*Institute of Artificial Intelligence*, *Hefei Comprehensive National Science Center*, *Hefei* 230094)

[3)](*Key Laboratory of Knowledge Engineering with Big Data* (*Hefei University of Technology*), *Ministry of Education*, *Hefei* 230601)

[4)](*Intelligent Interconnected Systems Laboratory of Anhui Province* (*Hefei University of Technology*), *Hefei*, 230601)

Abstract Video question answering (VideoQA) is a typical cross-modal understanding task. Its challenge lies in how to learn appropriate multimodal representation and cross-modal correlation for answer inference. Most existing video question answering methods focus on the latter, *e. g.*, relationship learning between each video frame or clip and word. In this work, we devote to

收稿日期:2022-03-30;在线发布日期:2022-11-03. 本课题得到国家自然科学基金(Nos. 62272144,U20A20183,62020106007,72188101)、安徽省科技重大专项(No. 202203a05020011)资助.**郭 丹**,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为视频分析、模式识别、深度学习. E-mail: guodan@hfut.edu.cn. **姚沈涛**,硕士研究生,主要研究方向为计算机视觉、视频问答. **王 辉**,博士研究生,主要研究方向为计算机视觉、视频问答、视觉对话. **汪 萌**(通信作者),博士,教授,中国计算机学会(CCF)合肥会员活动中心主席、国家杰出青年科学基金入选者、IEEE Fellow,主要研究领域为机器视觉、深度学习、多媒体信息处理. E-mail: eric.mengwang@gmail.com.

advanced feature embedding of both video and query. We develop a clustering-based VLAD technique for VideoQA. The novelty of our work is the joint exploitation of temporal aggregation and correlation in multimodality. We propose an end-to-end trainable Transformed VLAD embedding network, named TVLAD-Net. TVLAD-Net constructs a differentiable aggregation network module (*i. e.*, convolutional Residual-less VLAD Block) to generate compact VLAD descriptors (transforming N frames, clips or words to compact K descriptors while $K < N$), and realizes multi-head attention to correlate multimodal RVLAD descriptors. The characteristics are to eliminate redundant and invalid clues in the feature sequence and ensure diversity with multiple to-be-learned descriptors (corresponding to multiple clustering cells). To be specific, at first, we argue that a suitable representation should effectively exhibit the potential core semantic clues of sequence data. Based on this rule, we focus on the temporal aggregation of multimodality to extract core descriptors of data. For either videos or questions, we develop a learnable clustering-based Residual VLAD encoder to summarize each entire feature sequence into compact descriptors, respectively. Each descriptor can be deemed as a weighted aggregation over the entire feature sequence (a global perspective of unimodality). Multiple descriptors mean viewing global sequence serval times. It ensures the rich perspectives of semantic summarization. In this work, we consider the summarization of visual frame features, clip features, the combined frame & clip features of video, and word features of question. Second, we construct a unified Transformed module to realize multimodal descriptor interaction. To avoid irrelevant or redundant semantics of both visual and textual descriptors, we leverage multi-head attention in the Transformer architecture to control informative flows from these descriptors. The proposed transformed VLAD embedding module performs the context correlation of both inter-modality and intra-modality. Finally, each answer inference decoder is constructed for specific question types. The questions in VideoQA can be divided into the following three types: 1) Multi-choice task, 2) Open counting task and 3) Open word task. We use the corresponding decoder for each specific question type to infer the final answer. We evaluated TVLAD-Net on three VideoQA benchmark datasets, TGIF-QA, MSVD-QA, and MSRVTT-QA. The experimental results show that the proposed method achieves high accuracy of answer reasoning. There is a performance improvement of 2% to 5% compared with the existing methods. To summarize, the main contributions are summarized as follows: 1) by introducing the clustering-based VLAD aggregation into the differentiable convolution network, we refine the original features and generate advanced multimodal descriptors for VideoQA; 2) the multi-head operation in transformed VLAD embedding ensures the context correlation of both inter-modality and intra-modality. Either visual or textual descriptors, descriptors with similar or consistent semantics gather round; 3) extensive experiments demonstrate the effectiveness of TVLAD-Net over other approaches on three benchmark datasets.

Keywords video question answering; multi-modal data; aggregated descriptors; transformer network; deep learning

1 引 言

近年来,多模态内容分析吸引了越来越多的研究人员关注,例如视频字幕任务[1-2]、图像问答任务[3]、视频问答任务[4-5]、跨模态图像检索[6-8]等.与基于图像的任务相比,以视频为中心的任务面临更大的挑战:图像关注于精细的静态信息,而视频包含复杂的时空变化,需要解决时序信息长距离依赖的问题.视频问答任务还需要解决视频和文本的多模态

语义表征学习.视频和文本具有不同的语义复杂性,分别来源于视频自身的动态性和时序依赖性,以及文本语义解析的多样性和差异化.本文关注如何从复杂的视频和文本数据中提取出有效的多模态语义表征并用于问答推理.

最新的视频问答方法多关注于序列学习,多模态交互,或者多模态关系推理.例如,Jang和Fan等人[4,9]利用递归神经网络(RNN)和记忆网络(Memory Network)对视频问答中的序列学习进行建模,但是这类方法面临着长距离信息传递的梯度消失问题.图神经网络(GNN)适合推理型任务,Jiang等人[10]构建了一个异构图网络来推理视频帧和文本词之间的多模态关系.还有一些研究者探索注意力和其他学习模型,Jiang等人[11]提出了一个问题引导的时空上下文注意网络.Le等人[12]提出了一个条件关系网络,以更精细的视觉粒度对视频进行编码.最近,受Transformer[13]架构在自然语言领域中出色性能的启发,Li等人[14]提出了基于Transformer的视频问答模型.上述方法均在获取视频和文本特征后,重点建模问答推理过程.本文认为视频和文本的特征优化也是解决视频问答的基础,是会影响后续答案推理的关键环节.

视频和问题中包含着大量与推理无关的冗余信息(例如,背景帧/连词).如图1所示,在问题"What does the woman do after push hand ?"中,"woman","push hand"和"after"这些关键词就能反映问题的核心语义.冗余的单词或背景帧无法为答案推理提供有用的信息,甚至有可能干扰模型对关键信息的感知,并且增加了模型计算量.基于以上分析,本文致力于提取视频和文本问题中的关键信息,获取紧凑的、有效的多模态语义特征.在早期的研究工作中,已经存在一些特征聚合的方法.例如,平均池化或最大池化方法[15],视觉词袋模型[16](BOVW),Fisher向量编码[17]和局部聚类描述符[18](VLAD,Vector of Locally Aggregated Descriptor)等,这些方法多用于基于图像或视频的单模态的任务.

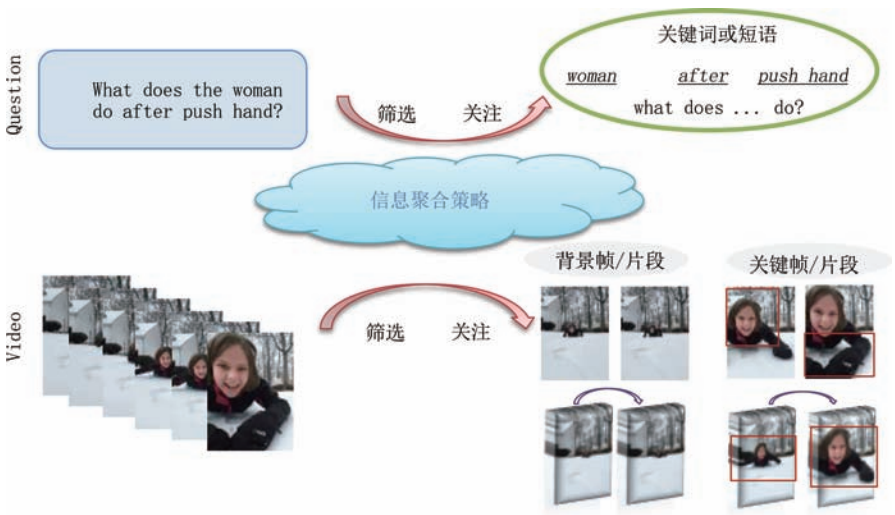

图1　视频问答任务中视频和文本的关键线索聚合

在本文中,我们开发了一种端到端可训练的嵌入局部聚类描述符的视频问答Transformer模型,称为TVLAD-Net(Transformer Residual-less VLAD Network).TVLAD-Net首先构建一个可学习的软分配聚合神经网络模块(称为RVLAD)来生成VLAD聚合描述符,然后利用Transformer方案[13]来学习多模态VLAD描述符之间的潜在语义关系,从而将原始的特征逐渐转化为嵌入描述符.最后,将变换后的VLAD描述符结合起来预测答案.本文的新颖之处在于解决了多模态视频理解中的时序聚合和时序关联问题:(1)如何有效聚合各模态的核心语义?本文设计了一种新的RVLAD(Residual-less VLAD)特征聚合模块,其包含K个可学习的聚类中心,以覆盖更为全面的语义信息.对于任一模态的特征序列(例如视频帧/视频片段/单词级特征),本文在K个RVLAD聚类中心上分别执行特征聚合,所有特征单元自适应地以不同的权重值分配到这些聚类中心上,并聚合起来得到RVLAD描述符(向量).聚类中心的个数K远小于特征序列的长度,因此,我们可以得到每个模态的紧凑的VLAD描述符.(2)如何有效地集成来自不同模态的VLAD描述符?本文利用Transformer[13]体系结构中的多头

注意力机制对来自视觉和文本模态的RVLAD描述符进行语义相关性学习，称之为Transformer跨模态交互模块．如图2所示，通过这两个模块，本文方法实现了模态间和模态内的语义聚合．

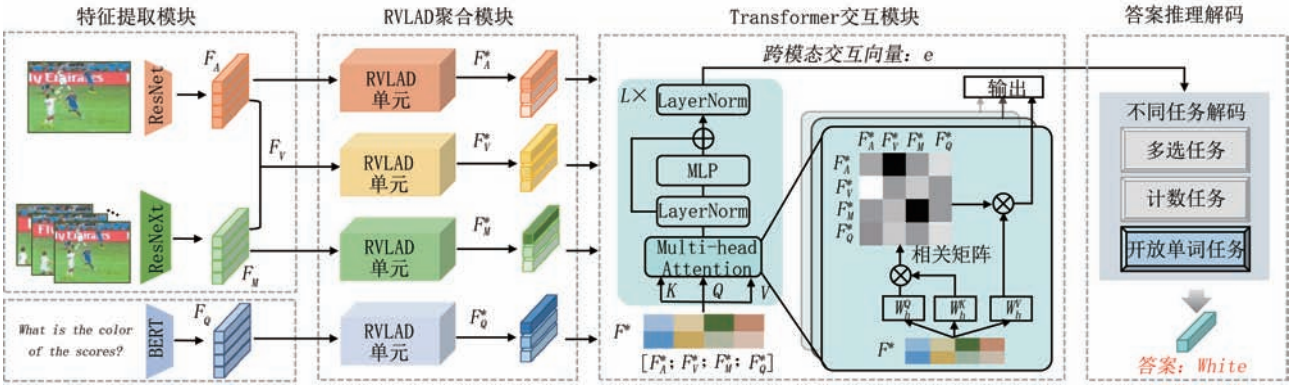

图2 嵌入局部聚类描述符的视频问答Transformer模型框架(TVLAD－Net)

本文的主要贡献可以总结为以下3个方面：

(1)本文提出一种新颖的端到端RVLAD聚合策略，用于提炼各个模态内的核心语义信息．RVLAD聚类描述符关注核心语义提取，去除视频与文本中的冗余信息．利用同种方式能够同时处理视频和文本的信息聚合表明了RVLAD策略在不同模态数据上的泛化性．

(2)本文提出基于Transformer跨模态交互模块融合多模态信息．对来自视频/问题的RVLAD描述符构建一个统一的多模态信息表示，用自注意力机制寻找视频与问题中的关联线索，引导视频与问题语义之间的交叉验证．

(3)在三个基准数据集TGIF-QA，MSVD-QA和MSRVTT-QA上的实验结果表明，本文提出的TVLAD-Net模型优于现有方法．消融实验证明了TVLAD-Net在增强特征表示和多模态信息融合上的有效性．

2 相关工作

2.1 视频表征

随着卷积神经网络(CNN)网络在图像领域的蓬勃发展[19-20]，CNN网络也逐渐拓展到视频任务上．应用在视频领域的视频表征框架主要有两大类：(1)基于2D网络的方法[15, 21-22]；(2)基于3D网络的方法[23-28]．前者可以用于提取视频帧的特征，后者常用于提取视频片段特征．C3D[25]在视频片段内提取时空间信息．I3D[23]以复制权重的方式将卷积核和池化核从二维拓展到三维．S3D[24]和S3D-G[24]提出将三维卷积分解为时间和空间卷积，其中S3D-G在S3D的基础上增加了门控操作，通过抑制不重要的信息来提高准确性．TSN[15]针对视频本身的时空特性构建了一个双路二维卷积网络以获取视频中的时序信息．TSM[26]通过一个简单的时间移位模块来捕获视频的时序依赖关系．SlowFast[27]提出了一种快慢结合的网络来用于视频分类，其中Fast网络用来捕获运动信息，Slow网络用来捕获空间信息．TR3D[28]将三维卷积模块嵌入在二维卷积神经网络(ResNet-50)中，以高效地提取视频的时空信息．Kensho等人[29]利用更深的图像分类预训练网络对视频帧进行表征，基于ResNet[30]设计出了ResNeXt[31]网络提取3D特征．本文兼顾了二维和三维网络提取视频的特征表达，分别采用ResNet和ResNeXt提取视频外观特征和片段动作特征．

2.2 视频问答的特征提取

视频不仅包含静态的帧级特征，还包含动态的片段级特征．目前视频问答[4-5, 10]工作大多同时采用两者以捕获更多类型的视觉信息．为了捕获视频中的运动信息，三维卷积神经网络被引入视频问答任务．C3D[25]网络在动作识别任务中表现出突出的捕获视频动态信息的能力，很多工作[4]利用预训练的C3D网络提取动态的视频片段级特征．例如，TGIF-QA[4]分别利用ResNet[30]网络和C3D网络获取视频的帧级特征和动态视频片段级特征．最近的HCRN[12]利用ResNet网络提取视频的帧级特征，和ResNeXt网络提取视频的动态特征．本文遵循现有工作HCRN[12]的方式，利用ResNet和ResNeXt网络提取视频特征．

对于文本特征，视频问答许多方法[4, 12]采

用GloVe[32]或BERT[33]提取文本的单词级特征. GloVe[32]单词嵌入是一个包含词汇表中所有单词的向量表示的特征矩阵，它将所有单词映射到低维嵌入空间，并计算单词之间的语义相似性. BERT[33]在海量的文本数据集上进行自监督学习取得巨大成功，使用预训练的BERT特征作为词嵌入表示是有效的. 本文主要使用BERT[33]提取文本的单词级特征，也同样在GloVe[32]嵌入基础上验证了本文方法的有效性.

最近，有些工作关注于大规模预训练模型来获取有效的视频-文本表征以促进视频语言任务. 例如，Wang等人[63]采用Transformer架构基于大规模视频-文本数据对进行模型预训练，而后在各种下游务上进行微调，包括文本-视频检索和视频问答任务等. 大规模预训练虽然有助于优化特征，但也面临着巨大的计算资源消耗.

2.3　视频问答的主流方法

视频问答任务具有挑战性，它不仅需要模型理解视频的空间线索，还需要理解视频中复杂的时间线索. 早期的视频问答方法[34-35]主要是基于循环神经网络模型构建的“编码器-解码器”结构，利用编码器分别编码文本和视频信息，之后融合两种模态的信息进行答案推理.

这种方法属于早期的基线模型，由于循环神经网络本身存在梯度消失问题，导致对长时序列的建模能力较差. 随后的研究[15, 19, 29, 36]考虑使用记忆网络模型存储不同时刻的上下文信息，从而保证模型能有效的利用长距离的时序上下文信息. 例如，Fan等人[9]构建了基于视觉、文本和多模态交互的三种记忆网络来捕获多模态中的复杂语义，然后更新记忆网络单元以推理正确答案. 注意力机制[11, 13-14, 37-39]也广泛应用在视频问答任务中，以解决循环神经网络的长时序依赖问题. Jin等人[39]提出一种基于自注意力机制(Self-attention)的交互网络，从对象和片段两种不同的特征粒度层面实现多模态交互. Gao等人[11]利用共同注意力机制建模每个单词和视频帧之间的关系以推断正确答案. Li等人[37]致力于改进注意力机制以使关系学习具有多样性，因此将单路共同注意力扩展为多路金字塔共同注意力. Jiang等人[38]进一步应用注意力机制分别指导空间和时间维度上的视觉编码，以捕获时空序列中的线索. 启发于BERT预训练模型在相关领域的成功，基于BERT的视频问答方法也逐渐被提出. Yang等人[40]提出利用BERT模块以统一的方式编码视频信息、文本信息和字幕信息，捕捉复杂视频场景的联合信息. 近年来，图神经网络(GNN)在各个任务上的广泛应用展示了其在关系学习上的优异能力. 由于视频问答任务需要捕获视频帧与单词、视频对象与单词，以及视频对象之间的空间时序关系，各种基于GNN的方法也被提出. Jiang等人[10]构建了一个异构图网络来建模视频帧和文本单词的之间的相互关系. 更进一步，Huang等人[36]提出了对象位置感知的图网络，利用更细粒度的帧级对象来建模不同视觉对象之间的空间关联. 上述方法均直接解决视频和文本之间的语义交互和推理. 不同地，本文首先致力于去除各模态的冗余信息，从而对视频和问题进行更好的语义表征，然后求解多模态语义交互，挖掘模态内和模态间的潜在丰富语义.

2.4　视频特征聚合

特征聚合的目的是将完整的视频特征序列映射为紧凑的语义单元. 最简单的聚合方式采用平均池化或者最大池化[15]，使用单个向量来代表完整的视频特征. 然而视频包含大量的视觉对象，动作信息和时序信息，利用单个向量表示整个视频可能是不够的. 有研究提出利用循环神经网络(包括RNN、LSTM和GRU[41-42])通过对视频特征进行时序上下文建模实现聚合，但此类方式在长视频序列上的建模效果不够理想. 还有研究人员尝试将传统的聚类方式(包括BOVW[16]，Fisher Vector[43]和VLAD[18])与深度学习相结合. 例如，Girdhar等人[44]针对视频动作分类任务提出了ActionVLAD，对视频的外观和运动特征进行时空聚合以捕获视频隐含的关键语义信息. Zhang等人[45]将VLAD与RNN相结合用于聚合视频中对象级的时序信息，并应用于视频字幕生成任务.

本文提出了一个RVLAD聚合模块，RVLAD的软分配策略摒弃了传统VLAD中将特征强制分配到固定的聚类中心的方式，而是动态地学习特征序列到各个聚类中心的分配权重，使得聚类后的特征能够自适应地探索整个序列的核心信息. 对于任一聚类中心，特征序列中各个特征单元自适应地分配不同的权重，并用加权求和的方式获取聚合信息，其中关键特征的聚合权重较大. 利用多个隐式聚类中心进行聚合，从不同的角度关注新的核心语义，保证了聚类信息的多样性. RVLAD聚合策略同时适用于视频和文本数据.

3 本文方法

视频问答任务是指根据视频 V 和文本问题 Q 推断出答案 A，答案可在预定义的答案空间或候选列表中找到. 图2展示了本文提出的TVLAD-Net的总体框架. TVLAD-Net主要由四个部分组成，特征提取模块，RVLAD聚合模块，Transformer交互模块，以及答案推理模块. 本章节安排如下：在3.1节，我们介绍了多模态特征提取的方式以及特征预处理. 在3.2节，RVLAD聚合模块对各模态各自特征序列进行聚合处理，分别聚合视频和文本中关键信息去除冗余信息. 3.3节描述对多模态间语义关系的建模，构建了统一的Transformer交互模块；在问题驱动的引导下，对模态间和模态内关系进行学习. 最后3.4节在视频问答任务中的不同问题类型上，利用不同的损失函数，构建独立的答案解码模块.

3.1 特征提取

视频存在静态和动态的语义信息，本文分别提取静态的外观特征 F_A(appearance feature)和动态的运动特征 F_M(motion feature). 由于 F_A 和 F_M 是分别独立提取的，在时序上的对齐较弱，本文构建了组合视觉特征 F_V 对齐运动特征 F_M 和外观特征 F_A. 本文将利用这三种视觉特征(即外观特征 F_A、运动特征 F_M 和组合视觉特征 F_V)进行后续的推理学习.

具体来说，给定视频 V，我们首先将 V 分成 N 个等长的视频片段. 然后，利用ResNeXt-101[31]作为运动特征提取器，获取视频的 N 个片段级特征 $F_M=(f_m^{(1)},\cdots,f_m^{(N)})\in\mathbb{R}^{d_m\times N}$，其中 d_m 表示运动特征 F_M 的特征维度. 同时，使用ResNet[15]作为提取外观特征的基准网络，对片段中的所有帧使用ResNet进行特征提取得到帧级特征. 为了在时序上对齐外观特征和运动特征，本文在每个片段内实施平均池化(mean-pooling)操作，将帧级特征压缩到视频外观特征 $F_A=(f_a^{(1)},\cdots,f_a^{(N)})\in\mathbb{R}^{d_a\times N}$，其中 d_a 为外观特征 F_A 的特征维度. 对于第 i 个视频片段，我们可以得到其视觉组合特征 $f_v^{(i)}=[f_a^{(i)};f_m^{(i)}]$. 因此，视频的组合特征为 $F_V=(f_v^{(1)},\cdots,f_v^{(N)})\in\mathbb{R}^{(d_a+d_m)\times N}$.

对于文本问题 Q，本文使用预训练的BERT[33]模型提取单词特征，将问题 Q 编码为一系列单词级特征 $F_Q=(f_q^{(1)},\cdots,f_q^{(W)})\in\mathbb{R}^{d_w\times W}$，其中 d_w 为问题特征的维度，W 为问题包含的单词数量.

3.2 语义发现I:RVLAD聚合

3.2.1 RVLAD时序聚合描述符

为了介绍各种模态的聚合过程，此处定义了一个通用序列 $\boldsymbol{X}=[x_1,x_2,\cdots,x_N]\in\mathbb{R}^{d_x\times N}$，其中 d_x 表示 x_i 的维度，N 为序列长度. 局部聚合描述符(VLAD，Vector of Local Aggregated Descriptors)最初被提出应用于图像检索任务[18]. 它利用K-means聚类在离散的图像数据库上生成 K 个聚类中心 c_k. 对每个聚类中心，计算图像的特征向量 x_i 与聚类中心 c_k 的残差，并根据权重 $r_k(x_i)$ 合并成新的聚类表达，即VLAD描述符. 具体如公式(1)所示：

$$VLAD(j,k)=\sum_{i=1}^{N}r_k(x_i)\cdot(x_i(j)-c_k(j)) \quad (1)$$

其中，K 表示聚类中心 $\{c_k\}$ 的个数，$x_i(j)$ 表示第 i 个特征向量 x_i 的第 j 维特征.

传统VLAD采用硬分配的聚合方式，即根据描述子 x_i 与聚类中心 c_k 的距离将 $r_k(x_i)$ 值设置为0或者1. 这种方式随后被替代为软分配方式[46]，构建端对端的可训练网络. 软分配是依据 c_k 与描述子 x_i 之间的距离分配[0, 1]的权重值. 相对于硬分配的0或1的权重值，软分配可以更好地描述 x_i 与所有聚类中心 c_k 之间的联系. 软分配的权重计算方式如公式(2)所示：

$$r_k(x_i)=\frac{e^{-\alpha\|x_i-c_k\|^2}}{\sum_{k=1}^{K}e^{-\alpha\|x_i-c_k\|^2}} \quad (2)$$

其中 α 是自定义的距离响应值.

与上述处理离散数据的聚合方式不同，本文要对时序序列 $\boldsymbol{X}=[x_1,x_2,\cdots,x_N]$ 进行聚合，得到新的表示形式 $\boldsymbol{X}'=[x'_1,x'_2,\cdots,x'_K]$. 本文提出了RVLAD(Residual-less VLAD)聚合策略，将 $\boldsymbol{X}=[x_1,x_2,\cdots,x_N]$ 中各元素映射到潜在的RVLAD描述符 $\boldsymbol{C}=\{c_k\}|_{k=1}^{K}\in\mathbb{R}^{d_c\times K}$. RVLAD去除了传统VLAD中的残差结构($-c_k$)的显示求解，将 c_k 转为隐式求解，如公式(3)所示：

$$\mathrm{RVLAD}(j,k)=\sum_{i=1}^{N}r_k(x_i)\cdot x_i(j) \quad (3)$$

其中 $k\in K$，N 表示序列 $\boldsymbol{X}$ 的长度，$j\in d_x$，$x_i(j)$ 表示第 i 个描述符的第 j 维特征.

本文利用软分配策略对特征序列进行加权. 原始特征 x_i 与第 k 个隐式语义单元 c_k 的相关权重 $r_k(x_i)$ 计算如下：

$$r_k(x_i)=\frac{e^{\boldsymbol{w}_k^T x_i+\boldsymbol{b}_k}}{\sum_{k=1}^{K}e^{\boldsymbol{w}_k^T x_i+\boldsymbol{b}_k}} \quad (4)$$

其中，$\boldsymbol{w}_k$，$\boldsymbol{b}_k$ 均为可训练参数.

因此，RVLAD将序列$\boldsymbol{X}=[x_1, x_2, \cdots, x_N]\in\mathbb{R}^{dx\times N}$经由$K$个隐藏的聚核中心$[c_1, c_2, \cdots, c_K]$转化为新的序列$\boldsymbol{X}'\in\mathbb{R}^{dx\times K}$. 如图3所示，由于软分配策略，$\boldsymbol{X}'$中单个元素可被视为对整体$\boldsymbol{X}$的聚合特征，即通过对时序序列$[x_1, x_2, \cdots, x_N]$分配不同权重累加而成的新的特征向量.

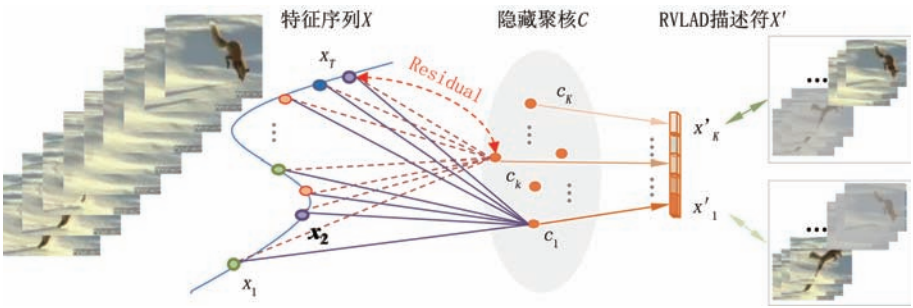

图3　RVLAD聚合特征描述符映射图

具体地，本文利用一维卷积神经网络$Conv_1D$对特征序列$\boldsymbol{X}$进行空间映射，使序列$\boldsymbol{X}$中元素x_i都关联至K个聚类中心，并利用softmax函数得到特征x_i在不同聚类中心上的权重：

$$\boldsymbol{R}=softmax\left(Conv\_1D(\boldsymbol{X})\right)\Big|_{col\_wise}\in\mathbb{R}^{N\times K} \quad (5)$$

根据分配权重$\boldsymbol{R}$，我们将序列$\boldsymbol{X}$中的所有特征x_i分配到不同的隐式聚类中心c_k，并根据公式(6)得到每个聚类中心c_k上的聚合特征，利用加权和的方式获得聚合序列$\boldsymbol{X}'$：

$$\boldsymbol{X}'=\boldsymbol{X}\times\boldsymbol{R}=[x'_1,\cdots,x'_K]\in\mathbb{R}^{d_c\times K} \quad (6)$$

其中，d_x为特征元素x_i的维度.

为了后续在同一联合语义空间进行多个模态的语义交互学习，本文对$\boldsymbol{X}'$进行映射，通过可学习参数$\boldsymbol{W}_k$将K个聚合描述符映射为d_c维向量$\boldsymbol{X}^*$，如公式(7)所示：

$$\boldsymbol{X}^*=Multi\_concat[\boldsymbol{W}_1x'_1,\cdots,\boldsymbol{W}_Kx'_K] \quad (7)$$

其中，$\boldsymbol{W}_1, \boldsymbol{W}_2, \cdots, \boldsymbol{W}_K\in\mathbb{R}^{dc\times dx}$均为可训练参数，$Multi_concat[\cdot,\cdot]$表示特征拼接操作.

本文将公式(5)～(7)记为RVLAD信息聚合模块$Aggregator(\cdot)$，如公式(8)所示. 本文提出的特征聚合思想以自适应方式学习K个核心语义的RVLAD聚合描述符. K个描述符代表从K个视角来表征整个序列. 其中每个描述符都包含了整个输入序列的不同权重响应的聚合. RVLAD描述符可视为原始特征的高级核心语义表示.

$$\boldsymbol{X}^*=Aggregator(\boldsymbol{X}),\boldsymbol{X}^*\in\mathbb{R}^{d_c\times K} \quad (8)$$

3.2.2　单模态语义发现

本文提出RVLAD模块用于时序信息聚合，$Aggregator(\cdot)$策略可以被拓展到视觉和文本多个模态的信息表征学习中. 由于视频的多种视觉特征与文本特征包含的核心语义不同，本文设置了四个独立的RVLAD模块分别对他们进行建模. 如公式(9)所示，基于$Aggregator(\cdot)$策略，我们可以得到运动特征F_M、外观特征F_A、组合视觉特征F_V以及问题特征F_Q各自的聚合特征F^*_M、F^*_A、F^*_V和F^*_Q.

$$\begin{cases} F^*_A=Aggregator(F_A),F^*_A\in\mathbb{R}^{d_c\times K}\\ F^*_M=Aggregator(F_M),F^*_M\in\mathbb{R}^{d_c\times K}\\ F^*_V=Aggregator(F_V),F^*_V\in\mathbb{R}^{d_c\times K}\\ F^*_Q=Aggregator(F_Q),F^*_Q\in\mathbb{R}^{d_c\times K} \end{cases} \quad (9)$$

其中，d_c为RVLAD描述符的维度，K表示描述符的数量. 为了在统一特征空间中学习多模态语义交互，各模态特征均被聚合到同一维度为d_c的特征空间中.

值得注意的是，在本文的方法中，聚合特征的维度远小于原始特征的维度，并且核心语义单元数量K远小于序列长度N，即$d_c\ll d_a$和$K\ll N$. 因此，无论从特征维度或数量上来说，获取的RVLAD特征描述符比原始特征序列更加紧凑，包含了核心语义信息.

3.3　语义发现II：Transformer语义交互

至此，优化各模态序列特征后，本文利用Transformer结构来构建模态内和模态间的语义关系，以进行答案推理. 首先，将外观聚合语义$F^*_A=(f^{*(1)}_a,\cdots,f^{*(K)}_a)\in\mathbb{R}^{dc\times K}$、运动聚合语义$F^*_M=(f^{*(1)}_m,\cdots,f^{*(K)}_m)\in\mathbb{R}^{dc\times K}$、组合视觉聚合语义$F^*_V=(f^{*(1)}_v,\cdots,f^{*(K)}_v)\in\mathbb{R}^{dc\times K}$和文本聚合语义$F^*_Q=(f^{*(1)}_q,\cdots,f^{*(K)}_q)\in\mathbb{R}^{dc\times K}$按列拼接得到特征表示$F^*$：

$$F^*=\left[F^*_A;F^*_M;F^*_V;F^*_Q\right]\in\mathbb{R}^{4K\times d_c} \quad (10)$$

然后，本文引入Transformer结构中的多头注意力机制，对特征表示F^*进行跨模态语义交互学习，

捕获特征单元间丰富的依赖关系.具体计算细节如公式(11)~(12)所示:

$$GuideAtt(F^*,F^*)=[head_1,head_2,\cdots,head_H]W^o \tag{11}$$

$$\begin{aligned}&head_h(F_h^*,F_h^*)\\&=Att(X_h^Q,Y_h^K,Y_h^V)|_{\{X_h^Q=W_h^QF_h^*,Y_h^K=W_h^KF_h^*,Y_h^V=W_h^VF_h^*\}}\\&=softmax\left(\frac{X_h^Q(Y_h^K)^\top}{\sqrt{d_k}}\right)\cdot Y_h^V\end{aligned} \tag{12}$$

其中,$W_h^Q,W_h^K,W_h^V\in\mathbb{R}^{d_c\times d_k}$,$W^o\in\mathbb{R}^{H*d_k\times d_c}$均为特征变换矩阵.$head_h(\cdot)$表示第$h$个注意力头的投影矩阵,$h=\{1,\cdots,H\}$,$H$为多头注意头的数量,$d_k=d_c/H$.

本文中,跨模态Transformer交互模块叠加了L层自注意力机制,即使用第l-1层的输出作为第l层的输入,如公式(13)所示:

$$F^{*(l)}=GuideAtt(F^{*(l-1)},F^{*(l-1)}) \tag{13}$$

出于简便,本文将完整的跨模态交互策略,即公式(11)~(13),概括为公式(14):

$$e=Transformer(F^*,F^*),e\in\mathbb{R}^{1\times(4K*d_c)} \tag{14}$$

其中,e来自Transformer模块最后一层的输出$F^{*(L)}$,$F^{*(L)}$经过变形操作展开成向量e.

Transformer结构的特性之一在于捕获长序列中丰富的依赖关系.本文的F^*包含了四种不同的语义单元F_A^*,F_M^*,F_V^*和F_Q^*,它们各自的内部关系($X\leftrightarrow X$,此处X表示F_A^*,F_M^*,F_V^*或F_Q^*),以及四个语义单元之间的相互关系($X\leftrightarrow Y$,此处X,Y为F_A^*,F_M^*,F_V^*及F_Q^*中任意不同的两个元素)都可以在Transformer模块中被捕获,实现语义关系计算.如图1所示,对于问题"What does the woman do after push hand ?",视频中包含hand的图像帧(F_A^*)要与问题中的文本"push hand"关联(F_Q^*);同时,视频中包含hand的图像帧(F_A^*)可以引导"push hand"行为的片段定位(F_M^*);另外,同时包含"woman","hand"和"push hand"的联合视觉语义(F_V^*)应该和(F_A^*),(F_M^*)都有关联.以F_Q^*为例,这种多模态语义的相互引导可表示为:$F_Q^*\in F^*$,$F_Q^*\leftrightarrow F_Q^*$,$F_Q^*\leftrightarrow\{F_M^*,F_V^*,F_A^*\}$.基于此准则,上述方式可视为四个单元模块间$\{F_M^*,F_V^*,F_A^*,F_Q^*\}$的语义交互.

3.4 答案推理解码

答案推理解码器是针对特定任务而构造的.视频问答任务包含以下三种问题类型:(1)多选择任务;(2)开放计数任务;(3)开放单词任务.针对特定的问题类型,采用相应的答案推理解码器[4].

3.4.1 多选择任务

对于多选择任务,本文将经过RVLAD聚合和Transformer交互捕获的特征向量e输入到一个全连接层进行特征映射,预测每个候选答案的分数s.损失函数如公式(15)所示:

$$\begin{cases}s=W_se+b_s,\\HingeLoss=max(0,1+s_n-s_p)\end{cases} \tag{15}$$

其中$W_s\in\mathbb{R}^{No\times d}$和偏置单元$b_s\in\mathbb{R}^{1\times1}$是可训练的参数.$N_o$为候选答案个数.$s_p$和$s_n$分别表示正确答案和错误答案上的预测分数.

3.4.2 开放计数任务

视频问答的开放计数任务被定义为计算一个动作的重复次数,其答案取值范围为[0, 10]中的整数,≥10次的答案均视为同一种分类结果[4, 12],即"10次".如公式(16)所示,本文利用线性回归函数根据特征向量e得到一个整数值的答案n,并利用MSE均方误差衡量预测值与真实值之间的距离:

$$\begin{cases}n=W_ne+b_n,\\MSELoss=|n-n_p|^2\end{cases} \tag{16}$$

其中,$W_n\in\mathbb{R}^{1\times d}$和偏置单元$b_n\in\mathbb{R}^{1\times1}$是可训练参数,$n_p$表示真实值.

3.4.3 开放单词任务

对于开放单词任务,本文使用softamax分类器进行概率预测,其中概率最高的候选答案被输出为最终答案.我们利用交叉熵损失(Cross Entropy Loss)对模型进行优化,如公式(17)所示:

$$\begin{cases}o=softmax(W_oe+b_o),\\CELoss=-\sum_{i=1}^{N_a}p_ilog(o_i)\end{cases} \tag{17}$$

其中,$W_o\in\mathbb{R}^{Na\times d}$和偏置单元$b_o\in\mathbb{R}^{Na\times1}$是可训练的参数.$N_a$是开放式答案集的大小,$p_i$表示第$i$个候选答案是否是正确答案,如果是则为1,否则为0;o_i表示模型预测的概率.

4 实验结果

4.1 实验设置

数据集.本文方法在三个基准数据集上进行验证.TGIF-QA[4]数据集由7.2万个动画GIF和16.5万个问答对组成.它分为四个子集:(1)Action子集,用于多选问答任务,选择具有指定重复数量的动作类别,其中每个问题对应5个可选答案;(2)Transition(缩写为Trans.)子集,同样是多选问答任务,有5个可选答案,问题有关视频中的面部表

情，动作，位置和对象属性的过渡状态；(3)Count子集，用于开放计数任务，计算指定动作的重复次数；(4)FrameQA子集，用于开放单词任务．答案可以从视频中的任何一帧推断出来，候选答案词典的大小为1，746. MSVD-QA[47]和MSRVTT-QA[48]数据集包含五种类型的问题，即What，Who，How，When和Where，所有问题均与开放式单词任务相对应，其中MSVD-QA包含1，970个视频和50，505个问答对，MSRVTT-QA包含10，000个视频和243，000个问答对．

评价指标．依照视频问答任务的惯例[12,38]，本文使用准确率(ACC)和均方误差(MSE)作为评价指标，其中MSE仅用于开放计数任务(如表2～6中的Count任务)的性能评估．ACC计算预测的正确答案数与预测的答案总数的比率，指标越高越好；MSE衡量预测值与真实值之间的均方误差，指标越低越好．

实验细节．对于视频特征的提取，我们将每个视频分为N=8个视频片段，每个视频片段包含16个视频帧．文本特征由预训练BERT[33]神经网络提取．特征维度设置为d_a=2048，d_m=2048，d_v=4096，d_w=768. 聚合后的RVLAD描述符维度设置为d_c=256. RVLAD模块中聚类中心的个数K将在图4和表1中进行讨论，Transformer交互模块的层数L和注意力头数H将在表2中进行讨论．训练时，使用Adam优化器进行参数优化，学习率设置为10^{-4}，每10次迭代后学习率下降一半．

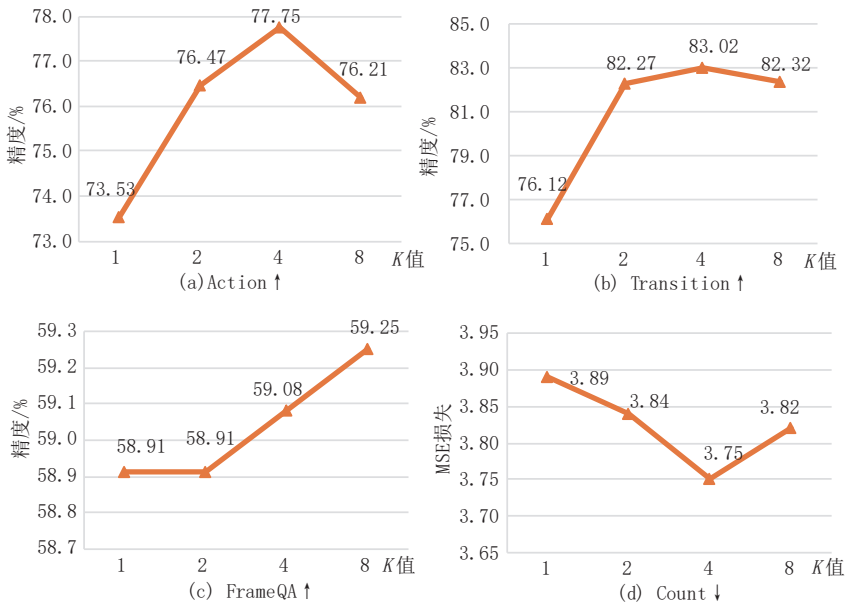

图4　在TGIF－QA数据集上，设置不同RVLAD聚类中心数K的性能变化

表1　在MSVD-QA和MSRVTT-QA数据集上，设置不同聚类中心数K的性能比较

数据集	K取值			
	1	2	4	8
MSVD-QA	37.4	38.8	**39.9**	39.0
MSRVTT-QA	35.4	37.0	**37.5**	36.7

4.2　消融实验

经验参数．为了评估RVLAD聚类中心数量K对模型性能的影响，我们测试了聚类中心个数K=[1，2，4，8]. 如图4所示，在TGIF-QA数据集上，K=4时模型在Action、Transition和Count三个任务上取得最好性能．当K=1时，描述符个数仅为1，此时描述符数量不足，导致方法精度不佳．增加RVLAD聚类中心数量使得描述符多样化，聚类特征的效果上升，K=4时取得了最佳性能．然而当K＞4时，描述符的数量过多，导致聚合描述符中出现语义冗余，实验精度反而下降．对于FrameQA任务，根据视频的每一帧都能推断出正确的答案，各帧包含的信息接近，因此，RVLAD在视觉序列上的聚合优势得不到发挥，反而对问题文本的依赖性很强．在FrameQA任务中，问题文本的单词数量W相对视频片段N较多，文本的描述符数量增加未达聚合上限，在K=8时性能仍持续上升．然而，本文为了在统一特征空间中进行Transformer交互模块的计算，将视觉描述符数量和文本描述符数量一致设置为K. 另外，如表1所示，在数据集MSVD-QA和MSRVTT-QA上K=4仍为最优设置．因此，本文

最终设置$K=4$.

另外，我们还测试了注意力头的数量H和Transformer层数L对问答精度的影响，如表2所示，本文测试了$H=[1, 2, 4, 8]$以及$L=[0, 1, 2, 3, 4]$. 实验结果表明，本文的方法在$H=4$和$L=1$达到最佳性能因此，本文在后续的实验设置中均采用$K=4$，$H=4$和$L=1$.

表2 TGIF-QA上经验参数H、L研究

参数	任务类型			
	Action ↑	Trans. ↑	Count ↓	FrameQA ↑
多头注意力头的数量H($L=1$)				
H=1	76.52	82.40	3.82	58.06
H=2	76.78	81.59	3.78	59.02
H=4	**77.75**	**83.02**	**3.75**	59.08
H=8	75.90	82.45	3.77	**59.10**
Transformer层数L($H=4$)				
L=0	76.39	82.71	3.80	58.72
L=1	**77.75**	**83.02**	**3.75**	59.08
L=2	77.48	82.01	3.79	58.58
L=3	75.79	82.00	3.81	58.48
L=4	75.79	83.04	3.84	**59.10**

主要模块消融分析. 表3测试了文本方法中RVLAD和Transformer模块的有效性. 如表3所示，两个主要模块的缺失会带来性能下降. (1)“w/o RVLAD”表示直接将原始特征输入Transformer交互模块. 可以看出，“w/o RVLAD”在所有任务上表现差于TVLAD-Net，它们之间的精度对比分别是：Action任务是75.79%和77.75%，Transition任务是82.37%和83.02%，Count任务是3.79和3.75，FrameQA任务是57.64%和59.08%. RVLAD模块用于聚合视频和文本问题中的核心语义，实验结果充分证明了RVLAD聚合核心语义信息的有效性. 本文还测试了其他的VLAD结构——NetVLAD[49]和NeXtVLAD[46]. 两者均遵循传统VLAD的硬分配策略，计算特征序列与聚类中心的特征差异来实现权重分配. 如表3所示，将本文中的RVLAD替换为NetVLAD[49]和NeXtVLAD[46]后性能均出现大幅下降. 实验结果证明了本文提出的RVLAD的有效性. (2)“w/o Transformer”直接将RVLAD提取的聚合描述符拼接进行答案推理. “w/o Transformer”与TVLAD-Net的精度对比为：Action任务是76.39%和77.75%，Transition任务是82.73%和83.02%，Count任务是3.80和3.75，以及FrameQA任务是58.72%和59.08%，相比之下本文方法均取得了更优效果. 这是因为缺少跨模态推理导致视频和文本中的线索不能有效关联. 为验证Transformer交互模块对视觉推理的有效性，如图5所示，我们探索了不同的Transformer交互形式Q2V和V2Q. 在Q2V中，我们设置Transformer模块的query值由拼接的视频特征得到$Q=\boldsymbol{W}^{Q}[F_A^*; F_M^*; F_V^*]$，key和value值由问题特征得到$K=\boldsymbol{W}^{K}F_Q^*$，$V=\boldsymbol{W}^{V}F_Q^*$，其中$\boldsymbol{W}^{Q}$，$\boldsymbol{W}^{K}$和$\boldsymbol{W}^{K}$均为特征变换矩阵；在V2Q中，则设置为$Q=\boldsymbol{W}^{Q}F_Q^*$，$K=\boldsymbol{W}^{K}[F_A^*; F_M^*; F_V^*]$，$V=\boldsymbol{W}^{V}[F_A^*; F_M^*; F_V^*]$. 不同的query，key和value值设置了视频和问题间不同的交互关系. 如表3所示，相比于

表3 TGIF-QA数据集上主要模块消融实验

对比模型	任务类型			
	Action ↑	Trans. ↑	Count ↓	FrameQA ↑
w/o RVLAD	75.8	82.4	3.79	57.6
NetVLAD[49]	72.1	79.5	4.13	56.6
NeXtVLAD[46]	75.5	77.5	4.13	53.9
TVLAD-Net	**77.7**	**83.0**	**3.75**	**59.1**
w/o Transformer	**76.3**	**82.7**	**3.80**	**58.7**
Q2V	76.9	82.4	3.84	58.1
V2Q	74.6	82.6	3.81	57.5
MFB[50]	76.2	82.0	3.89	52.5
MFH[51]	76.1	81.6	3.95	54.3
MUTAN[52]	77.0	81.9	3.80	58.0
TVLAD-Net	**77.7**	**83.0**	**3.75**	**59.1**

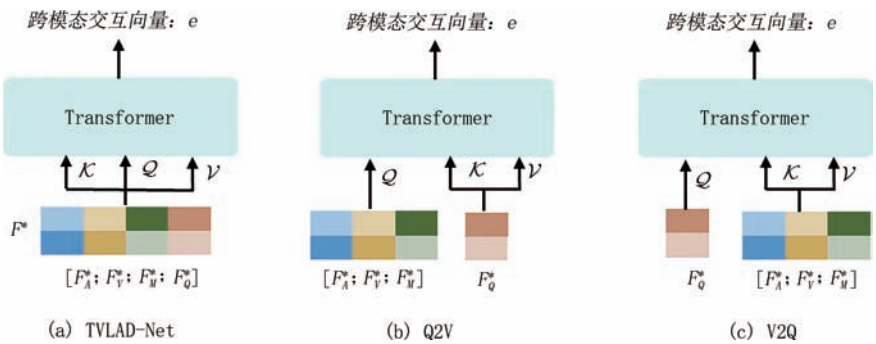

图5 不同的Transformer交互形式

TVLAD-Net，Q2V和V2Q的性能均有所下降. 这说明在TVLAD-Net中通过$\{Q,K,V\}=\boldsymbol{W}^X[\boldsymbol{F}_A^*;\boldsymbol{F}_M^*;\boldsymbol{F}_V^*;\boldsymbol{F}_Q^*]$，$X\in\{Q,K,V\}$的交互形式能够为模型提供更丰富的跨模态推理线索. 我们还测试了经典的跨模态交互模型，包括MFB[50]，MFH[51]，以及MUTAN[52]. 如表3所示，本文方法表现出了最好性能；在FrameQA任务上，准确率相比于MFB[50]，MFH[51]，以及MUTAN[52]分别提高了6.6%，4.8%和1.1%. 这三种方法都是简单地执行全局视频和文本特征的线性池化融合，忽略了多模态序列中丰富的时序信息. 在本文中，Transformer作用于聚合后的视频和文本特征序列，不仅能在视觉引导下寻找文本线索，同时也能利用文本寻找视觉线索，以此获得更丰富的推理能力. (3)另外，我们发现“w/o RVLAD”较“w/o Transformer”的性能下降更为明显，原因可能是RVLAD对特征的优化可以保证后续Transformer交互模块中语义交互的有效性.

视觉特征对方法的影响. 本文方法中引入了多种视觉特征，包括外观特征$\boldsymbol{F}_A$，运动特征$\boldsymbol{F}_M$以及组合视觉特征$\boldsymbol{F}_V$. 表4展示了它们各自在模型中的作用. (1)如表4所示，在仅使用外观特征$\boldsymbol{F}_A$时，“only $\boldsymbol{F}_A$”的表现相对完整模型较差. 在Action任务上精度由77.7%下降至72.5%，在Transition任务上精度由83.0%下降至81.7%，以及在Count任务上精度由3.75变化至4.33. 实验结果表明在需要理解动作或者视觉变化的内容时，仅使用外观特征不足以准确推理出正确的结果；而由于FrameQA任务类似于图片视觉问答，仅使用外观特征并不会带来精度的剧烈下降. (2)在仅利用运动特征$\boldsymbol{F}_M$时，“only $\boldsymbol{F}_M$”在Action任务上精度由完整模型的77.7%下降至73.9%，在Transition任务上精度由83.0%下降至80.0%，在FrameQA任务上精度由59.1%下降至52.2%，以及在Count任务上精度由3.75变化至4.03. 特别地，“only $\boldsymbol{F}_M$”在FrameQA任务上性能下降明显. 这是因为FrameQA任务对视频静态信息需求更高，主要利用外观特征$\boldsymbol{F}_A$进行答案推理.

表4　TGIF-QA数据集上视觉特征的消融实验

特征类型	任务类型			
	Action ↑	Trans. ↑	Count ↓	FrameQA ↑
only F_A	72.5	81.7	4.33	58.9
only F_M	73.9	80.0	4.03	52.2
only F_V	77.1	82.5	3.88	58.8
w/o F_V	76.2	82.1	3.78	58.8
TVLAD-Net	**77.7**	**83.0**	**3.75**	**59.1**

(3)与完整模型相比，“only $\boldsymbol{F}_V$”只使用$\boldsymbol{F}_V$和“w/o $\boldsymbol{F}_V$”不使用$\boldsymbol{F}_V$两种模型的性能都有下降. 而且，“only $\boldsymbol{F}_V$”相比“w/o $\boldsymbol{F}_V$”在除Count任务外的其它任务上性能都有微弱优势. 这证明了本文提出的融合特征$\boldsymbol{F}_V$的有效性. $\boldsymbol{F}_A$，$\boldsymbol{F}_M$和$\boldsymbol{F}_V$语义互补，联合使用三组视觉特征时性能最好.

文本特征对比分析. 我们在表5中对比了Glove和BERT文本特征对性能的影响. 可以看出，利用BERT特征相比Glove特征进一步提升了本文方法的性能，在后续实验中，默认采用BERT文本特征. 在相关工作中，已有LAD-Net[37]等工作使用了BERT特征. 另外，公平起见，我们对比了三种使用Glove特征的方法，即HGA[10]，LAG[36]和HCRN[12]. 如表5所示，本文方法即使使用Glove特征，仍然优于对比方法，在TGIF-QA数据集，MSRVTT-QA数据集以及MSVD-QA数据集的不同任务上精度均有明显提升.

表5　GloVe文本特征和BERT文本特征对比

模型	TGIF-QA				MSVD-QA	MSRVTT-QA
	Action ↑	Trans. ↑	Count ↓	FrameQA ↑		
HGA[10]	75.4	81.0	4.09	55.1	34.7	35.5
LAG[36]	74.3	81.1	3.95	56.3	—	—
HCRN[12]	75.0	81.4	3.82	55.9	36.1	35.6
Glove	75.8	81.7	3.83	57.5	37.5	36.5
BERT	**77.7**	**83.0**	**3.75**	**59.1**	**39.9**	**37.5**

4.3　主实验比较

TGIF-QA上的结果. 表6给出了在TGIF-QA数据集上本文方法与现有方法的性能对比. 根据模型的算法结构，我们将对比方法分为五类：基于循环神经网络和记忆网络的模型[15, 19, 29, 36]、基于注意力的模型[11, 13-14, 37-39]、基于图神经网络的模型[10]、条件关系模型[12]和基于聚合网络的模型[37]. 显然，本文方法的实验结果优于其他工作.

(1)在现有基于循环神经网络和记忆网络的方法中，FAM[53]取得了最好的性能. TVLAD-Net的表现好于它. TVLAD-Net学习了K个视频聚类描述符，能够更紧凑和全面的视频表示学习，因此提升了视频问答性能，特别是在与动作相关的任务中. 例如在Action任务上，本文精度为77.7%，而FAM[53]取得的精度仅为75.4%；(2)在基于注意力的模型中，具有代表性的模型是QueST[38]，它主要利用文本挖掘视觉线索，而本文方法关注视频和文

表6 TGIF-QA数据集上本文方法与其他方法的精度对比

模型	任务类型			
	Action↑	Trans.↑	Count↓	FrameQA↑
基于循环神经网络和记忆网络的模型				
FAM[53]	75.4	79.2	3.79	56.9
VQA-MCB[54]	58.9	24.3	5.17	25.7
VIS+LSTM[55]	46.8	56.9	5.09	34.6
CT-SAN[56]	56.1	64.0	5.13	39.6
Co-memory[57]	68.2	74.3	4.10	51.5
HME[9]	73.9	77.8	4.02	53.8
基于注意力的模型				
ST(R)[4]	59.0	65.5	4.55	45.6
ST(R+C)[4]	60.8	67.1	4.40	49.3
ST(R+F)[4]	62.9	69.4	4.32	49.5
STA[11]	72.3	79.0	4.25	56.6
ST-VQA[42]	73.5	79.7	4.22	52.0
PSAC[14]	70.4	76.9	4.27	55.7
MIN[39]	72.7	80.9	4.17	57.1
QueST[38]	75.9	81.0	4.19	**59.7**
基于图神经网络的模型				
LAG[36]	74.3	81.1	3.95	56.3
HGA[10]	75.4	81.0	4.09	55.1
条件关系模型				
HCRN[12]	75.0	81.4	3.82	55.9
基于聚合网络的模型				
LAD-Net[37]	72.0	80.7	4.24	58.2
TVLAD-Net	**77.7**	**83.0**	**3.75**	59.1

本的双向交互信息.相对于QueST[38],本文方法在需要复杂时序推理的任务上(例如Action,Transition和Count上)提升明显.在Action任务上,本文方法与QueST[38]的精度分别为77.7%和75.9%;在Transition任务上的对比为83.0%和81.0%,在Count(↓)任务上的对比为3.75和4.19.值得注意的是,在FrameQA任务上,本文方法与QueST[38]相比精度略有下降,分别为59.1%和59.7%.这是因为QueST[38]更侧重于问题语义建模,将问题语义细分为空间和时间维度,构建了更深层次的文本语义理解,而FrameQA任务类似图像问答,根据视频中任意一帧都能推理出答案,这种情况下问题文本提供的语义相对较多,更深层次地挖掘文本语义能带来性能提升;(3)对于基于图的模型,LAG[36]和HGA[10]都关注于多模态的关系学习,而忽略了视频中的时序信息.相比之下,TVLAD-Net利用时序聚合保留时序信息,在所有任务中均取得了显著效果.(4)对于条件关系模型,HCRN[12]致力于在不同的视觉粒度(帧级,片段级,视频级)下进行视频编码,但是忽略了对问题语义的深度挖掘.在Transition任务上,本文方法精度为83.0%,而HCRN[12]精度为81.4%.5)基于聚合的模型LAD-Net[37]仅利用时序注意力将特征序列聚合为单个特征向量.本文针对一个特征序列学习了K个RVLAD描述符以聚合视频和问题中的关键语义信息,取得了更好的效果.相比于LAD-Net[37],TVLAD-Net在Action,Transition和FrameQA任务上的精度分别提升了5.7%,2.3%和0.9%.

MSVD-QA和MSRVTT-QA上的结果.与TGIF-QA数据集相比,MSVD-QA和MSRVTT-QA数据集是具有更大挑战的基准数据集,它们的答案候选空间更大,且视频更加复杂(视频长度平均约为2分钟,而TGIF-QA中的视频长度平均约3秒).如表7和表8所示,在MSVD-QA和MSRVTT-QA数据集上,本文方法与最新工作相比均有显著的性能提升.在MSVD-QA和MSRVTT-QA数据集上,本文方法的总体性能"All"相比目前的最优工作HCRN[12]分别提高了3.8%和1.9%的精度.另外,MSRVTT-QA和MSVD-QA这两个数据集存在严重的数据不平衡问题,在MSVD-QA数据集上,How和Where的占比仅为2.8%和0.2%;在MSRVTT-QA

表7 MSVD-QA数据集上本文方法与其他方法的精度对比

问题类型	All	What	Who	When	How	Where
占总量比率	100%	62.7%	33.9%	0.4%	2.8%	0.2%
不同模型实验结果对比						
基于循环神经网络和记忆网络的模型						
E-VQA[58]	23.3	9.7	42.2	72.4	**83.8**	**53.6**
E-MN[58]	26.7	12.9	46.5	70.7	80.3	50.0
Co-memory[57]	31.7	19.6	48.7	74.1	81.6	31.7
HME[9]	33.7	22.4	50.1	70.7	73.0	42.9
FAM[53]	34.5	23.1	51.6	71.4	82.2	51.9
基于注意力的模型						
E-SA[58]	27.6	15.0	45.1	65.5	**83.8**	32.2
ST[4]	31.3	18.1	50.0	72.4	**83.8**	28.6
DLAN[59]	31.8	21.2	46.0	72.4	83.2	50.0
GRAAM[58]	32.0	20.6	47.5	72.4	83.5	**53.6**
AA-Net[60]	32.6	21.3	48.3	70.7	82.4	**53.6**
STCA[61]	35.0	24.3	49.6	74.1	83.0	**53.6**
MIN[39]	35.0	24.2	49.5	74.1	**83.8**	53.6
QueST[38]	36.1	24.5	52.9	72.4	79.1	50.0
基于图神经网络的模型						
LAG[36]	34.3	—	—	—	—	—
HGA[10]	34.7	23.5	50.4	72.4	83.0	46.4
条件关系模型						
HCRN[12]	36.1	—	—	—	—	—
基于聚合网络的模型						
TVLAD-Net	**39.9**	**29.8**	**54.5**	**74.1**	76.5	50.0

表8　MSRVTT-QA数据集上本文与其他方法精度对比

问题类型	All	What	Who	When	How	Where
占总量比率	100%	68.5%	27.7%	1%	2.5%	0.3%
不同模型实验结果对比						
基于循环神经网络和记忆网络的模型						
E-VQA[58]	26.4	18.9	38.7	70.5	83.5	29.2
E-MN[58]	30.4	23.4	41.8	70.8	**83.7**	27.6
Co-memory[57]	32.0	23.9	42.5	69.0	74.1	**42.9**
HME[9]	33.0	26.5	43.6	**76.0**	82.4	28.6
FAM[53]	33.2	26.9	43.9	70.6	82.8	31.1
基于注意力的模型						
E-SA[58]	29.3	22.0	41.6	73.1	79.6	33.2
DLAN[59]	32.0	25.4	42.8	72.1	81.0	31.2
GRAAM[58]	32.5	26.2	43.0	72.5	80.2	30.0
STCA[61]	34.2	27.4	45.4	74.0	83.7	33.2
MIN[39]	35.4	29.5	45.0	74.7	83.2	42.4
QueST[38]	34.6	27.9	45.6	75.7	83.0	31.6
基于图神经网络的模型						
HGA[10]	35.5	29.2	45.7	75.2	83.5	34.0
条件关系模型						
HCRN[12]	35.6	—	—	—	—	—
基于聚合网络的模型						
TVLAD-Net	**37.5**	**31.5**	**47.4**	74.6	81.9	37.2

数据集上，When，How和Where的占比分别为1%，2.5%和0.3%. 本文提出的TVLAD-Net是一种基于聚类的方法，训练样本不足可能会限制RVLAD模块的特征聚合能力，导致本文方法在问题类型“How”和“Where”上的表现不突出. 但是在聚合样本数量充足的“What”和“Who”问题上，本文方法的性能远远大于现有方法，证明了本文方法在聚合关键语义信息方面的有效性.

模型复杂度分析. 我们比较了本文方法与HCRN[12]之间的模型复杂性. 如表9所示，本文方法的模型参数量、GPU占用率和测试时间都大大优于HCRN；同时，本文方法在Action任务上相比于

表9　本文方法与HCRN模型的模型复杂度对比

Model	Params ↓	GPU (MB) ↓	Time(s) ↓	Action ↑
HCRN[12]	155.2 M	4069	227	75.0
TVLAD-Net	**55.4 M**	**2199**	**57**	**77.7**

HCRN取得了2.7%的性能增益.

4.4　定量分析

如图5所示，本文利用t-SNE[62](T-Distribution Stochastic Neighbour Embedding)可视化外观特征F_A、运动特征F_M、组合视觉特征F_V和问题特征F_Q分布的变化情况，其分别被表示为红色、黄色、蓝色和绿色的数据点. 在原始特征空间中距离相近的数据点投影到低维t-SNE空间中仍然相近. 图5(a)展示了所有未经处理的原始视觉特征分布. 可以看出，F_M、F_A和F_V汇集在一起，其中组合视觉特征F_V分布位于F_M和F_A之间，而问题特征F_Q则单独集中在一起，与视觉特征相分离. 图5(b)表示经过聚合RVLAD核心单元后，聚合后的问题特征F_Q^*出现在特征映射空间的中心位置，被聚合后的视觉特征(F_M^*、F_A^*和F_V^*)包裹. 语义相近的视觉特征(F_M^*、F_A^*和F_V^*)变得靠拢，彼此融入. 在图5(c)中，经过Transformer语义交互模块计算后各个模态的特征充分融合，多模态语义向量中语义一致的描述符彼此靠拢(不再呈现各自模态分布)，表明了问题驱动的相关视觉、文本语义聚合的一致性.

图6可视化了RVLAD聚合模块以及Transformer语义交互模块中的权重分布图，分别为“各模态聚合权重分布”和“多模态交互权重均值(多头注意力)”，其中$\boldsymbol{v}_i$，$\boldsymbol{m}_i$，$\boldsymbol{a}_i$，$\boldsymbol{q}_i$分别表示经过RVLAD聚合的F_V，F_M，F_A，F_Q中第i个聚类中心上的特征. (1)对于Action任务，引导多头注意力交互的权重响

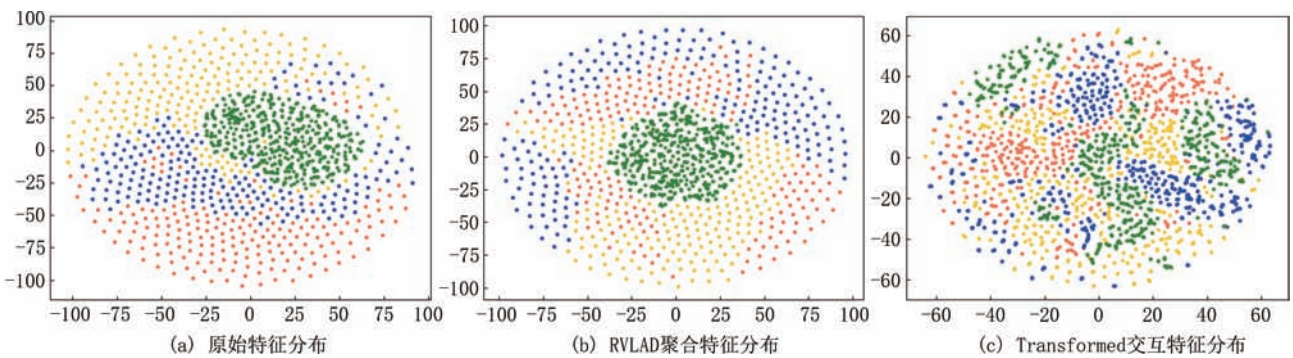

图5　t-SNE特征分布图. 红色、黄色、蓝色和绿色分别代表外观特征F_A、运动特征F_M、组合视觉特征F_V和问题特征F_Q. (a)和(b)均为各模态内特征学习，呈现出视觉和文本模态分布的彼此独立；(b)中，聚合的多模态RVLAD分布中，问题文本的分布位于中心，视觉围绕其展开. (c)所展示的特征分布是由文本和视觉交互后的特征表示，此时的语义出现了彼此交融；具有语义一致性的特征靠拢在了一起.

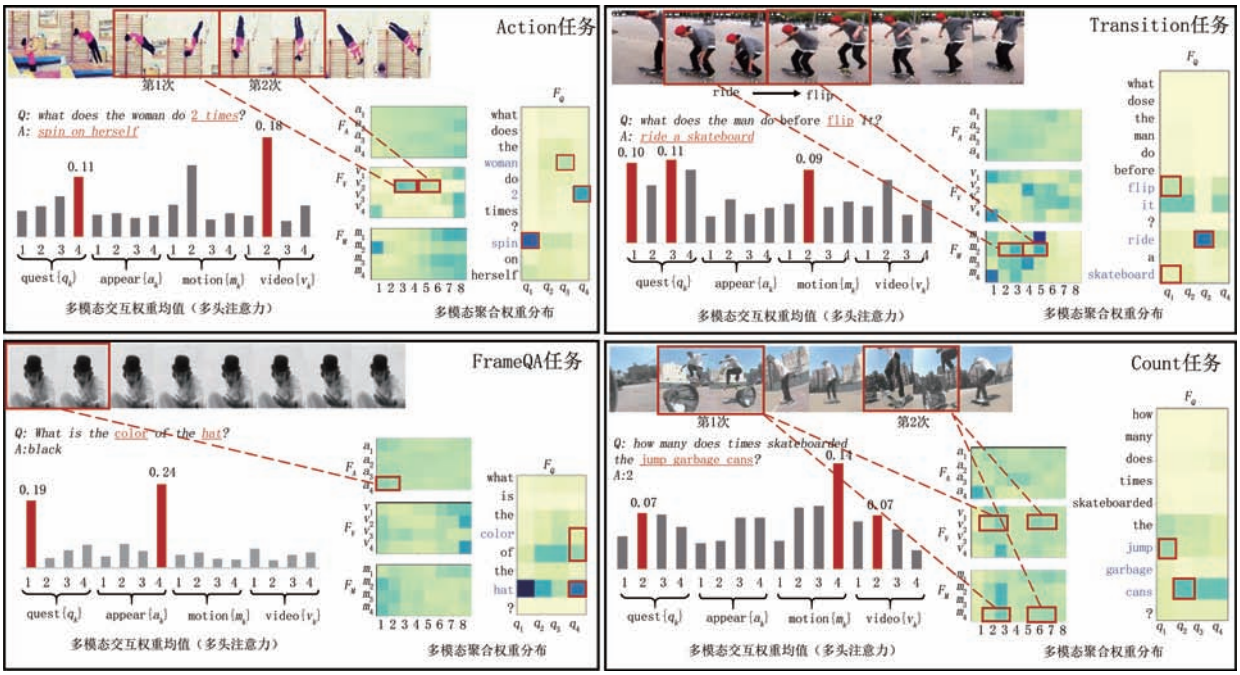

图6 TGIF-QA数据集四个任务的可视化示例

应最强烈的是q_4和v_2，表明了本文提出的组合视觉特征F_V的有效性. 问题特征F_Q经过RVLAD聚合模块后，聚集在q_1，q_3，q_4对应的“spin”，“2”，“woman”关键词上，而组合视觉特征F_V经过RVLAD聚合模块，核心语义集中在视频片段3至片段6中（即真实动作spin发生2次的片段）. (2) Transition任务考验模型对动作状态转换的识别能力，理解动作信息至关重要. 在该示例中，引导多头注意力分布权重响应最强烈的是q_3和m_2，这表明运动特征F_M很好地支持了模型对状态转换进行识别. 问题F_Q经过聚合后，聚集在q_1，q_3中的“flip ”，“ride”两个关键动作中. 运动特征F_M经过聚合后集中在m_2中的视频片段2至片段5中. (3) FrameQA任务中，对于问题“what is the color of the hat？ ”，引导多头注意力分布权重响应最强烈的是q_1和a_4，本文方法根据问题q_1的引导可以准确识别帽子的颜色外观a_4. F_Q经过聚合后，聚集在关键词“hat”. 外观特征F_A的权重分布均匀，因为利用任意一帧都可以回答. (4) Count任务中，引导多头注意力分布权重响应最强烈的是q_2，m_4和v_2，表明运动特征F_M和组合视觉特征F_V在问题特征F_Q的引导下准确定位了两次“jump”动作. 问题特征F_Q集中在q_1，q_2的“jump”以及“cans”上，同时运动特征F_M和组合视觉特征F_V分别响应在视频片段2到视频片段3，和视频片段6到视频片段7上，正好覆盖了两次“jump”行为定位. 本文方法依据聚合信息在不同情形下推理正确答案. 四个不同任务的可视化图表明了多路视觉特征（F_Q、F_M、F_A和F_V）的积极作用，也证明了RVLAD聚合模块的聚合能力和多样性以及在Transformer模块中语义交互的合理性.

在图7中，我们展示了一些本文方法和HCRN[12]的问答推理样例. 显然，本文方法相比HCRN[12]生成的答案更加准确. 如Count (a)，HCRN错误地将行为“pump hips”发生次数计为3次，本文正确识别出次数为2. FrameQA (a)中，HCRN错误地生成视觉显著性较强的“room”. 本文方法则正确地识别出“minor”. 正确答案“minor”的视觉显著性较弱，没有明显的形状等外观特征，考验模型的深度视觉理解和推理能力. Transition (b)中，后面的人的动作与“kiss”很相似，HCRN错误判断为“receive a kiss from another”，而本文更准确地识别到“stare”动作，表明本文方法更清晰准确地识别动作. 在例子Action(a)中，“lower hand”和“close leg”动作均在视频中出现，本文方法通过问题中的“2 times”引导成功捕获了正确的“close leg”语义.

另外，我们在图8中列举了三个非常具有挑战性的视频，根据视频标签，本文模型预测的答案被判定为失败案例. 然而，经过人工评价，有些失败案例中存在标签误差，预测的结果被认为是正确答案. 例如例子1属于开放式问答任务. 对于问题“How many times does the dog jump ?（这只狗跳跃了几次?）”，本文模型预测结果为3次；小狗第1次跳跃的幅度不明显，本文模型识别到精细的动作信息. 例

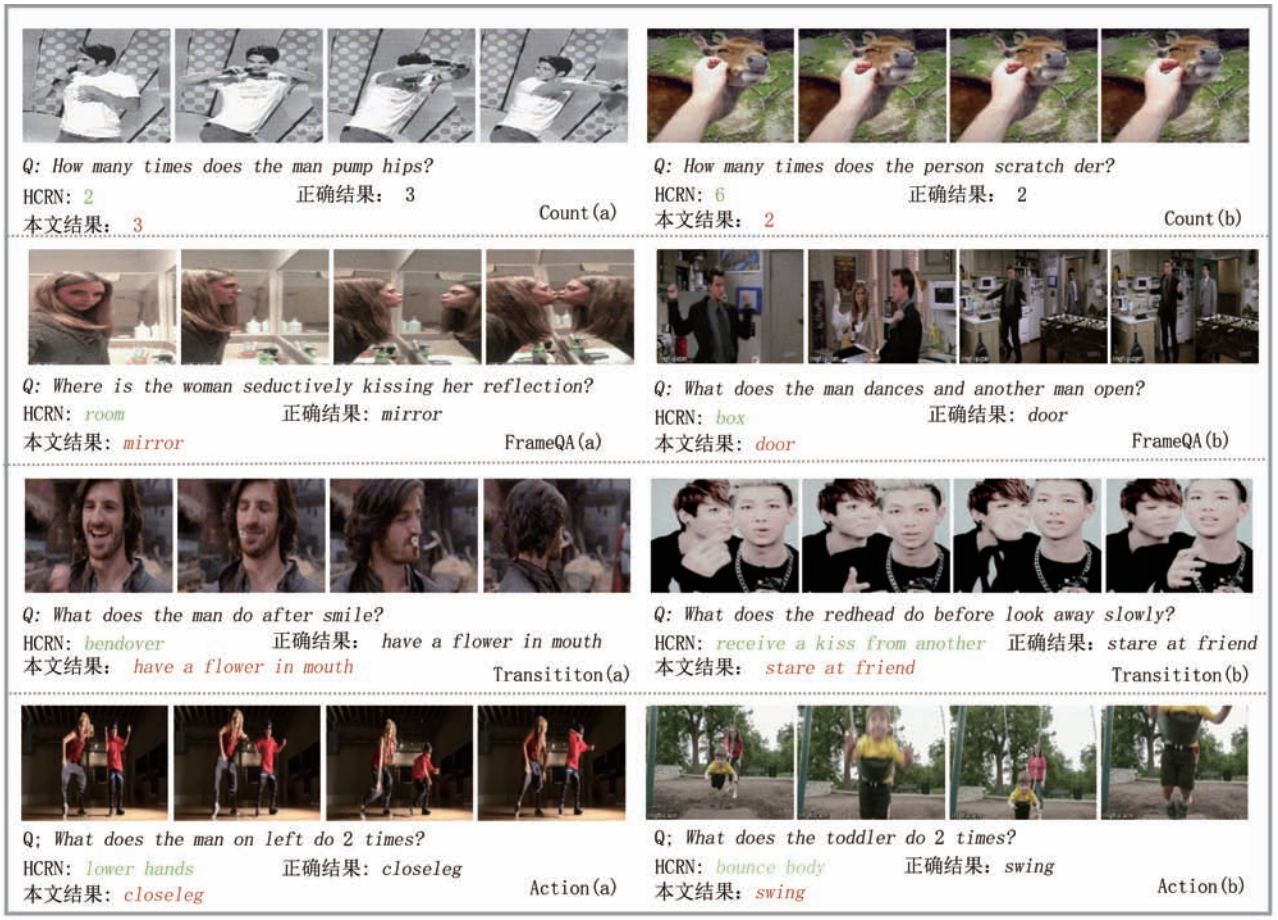

图7　TGIF-QA数据集上本文方法与HCRN方法的问答结果样例

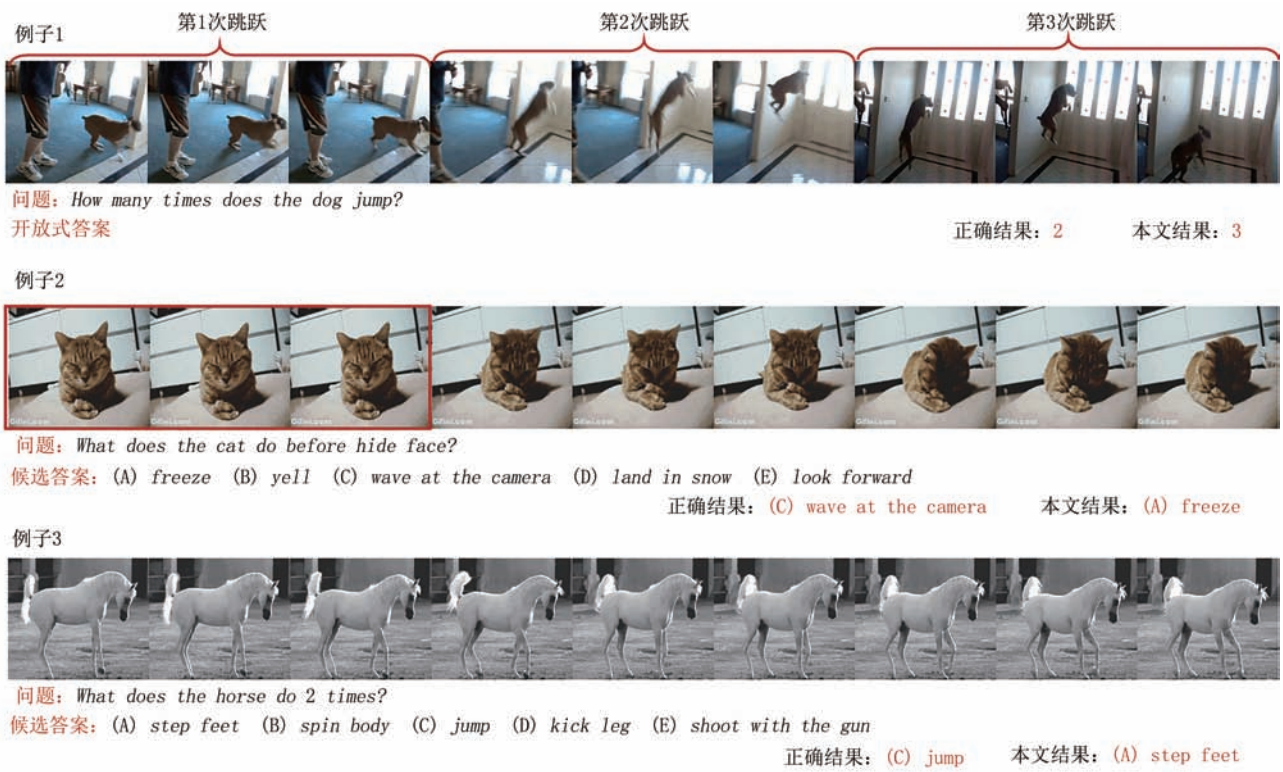

图8　TGIF-QA数据集上的失败案例

子2属于多选问答任务，问题是“What does the cat do before hide face?(这只猫在掩面之前在做什么?)”，模型需要从5个难以区分的候选答案中选出正确答案.模型通过分析视频内容推理出了“freeze(定格)”.经过人工校正，“freeze(定格)”是最合理的答案.案例1和2中，尽管存在标签误差，本文模型能够推理出正确答案.有些失败案例中不存在标签误差，但存在多个近似语义的答案难以区分.例如，例子3属于多选问答任务.真实标签为“jump(跳跃)”，本文模型选择了“step feet(踏脚)”.事实上，根据例子3的视频内容，由于运动幅度较小，人类也难以分辨“step feet(踏脚)”和“jump(跳跃)”的细微之处，似乎两者都是合理的.

5 结 论

本文提出了一种优化视频和文本时序特征的语义聚合方法.TVLAD-Net通过基于聚类的RVLAD模块提炼各个模态内的核心语义，并引入Transformer语义交互模块寻找视频与问题中的关键线索，引导视频与问题语义之间的交叉验证，最后通过答案解码模块获得问题答案.实验结果表明，本文方法在三个基准视频问答数据集上具有良好性能，在多个评价指标上优于现有方法.我们将在后续工作开展各模态RVLAD聚类中心数量的自适应研究.

参 考 文 献

[1] Pei W, Zhang J, Wang X, Ke L, Shen X and Tai Y-W. Memory-attended recurrent network for video captioning //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8347-8356

[2] Tang Peng-Jie and Wang Han-Li. From video to language: survey of video captioning and description. Acta Automatica Sinica, 2021, 47(x): 1-23 (in Chinese)
(汤鹏杰，王瀚漓.从视频到语言：视频标题生成与描述研究综述.自动化学报，2021，47(x)：1-23)

[3] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick C L and Parikh D. Vqa: Visual question answering //Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2425-2433

[4] Jang Y, Song Y, Yu Y, Kim Y and Kim G. TGIF-QA: toward spatio-temporal reasoning in visual question answering //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1359-1367

[5] Lei J, Yu L, Bansal M and Berg T L. TVQA: localized, compositional video question answering //Proceedings of the Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 1369-1379

[6] Dong Zhen and Pei Ming-Tao. Cross-modality face retrieval based on heterogeneous hashing network. Chinese Journal of Computers, 2019, 42(1): 75-86 (in Chinese)
(董震，裴明涛，基于异构哈希网络的跨模态人脸检索方法.计算机学报，2019，42(1)：75-86)

[7] Yan Shuang-Yong, Liu Chang-Hong, Jiang Ai-Wen, Ye Ji-Hua and Wang Ming-Wen. Discriminative cross-modal hashing with coupled semantic correlation. Chinese Journal of Computers, 2019, 42(1): 164-175 (in Chinese)
(严双咏，刘长红，江爱文，叶继华，王明文，语义耦合相关的判别式跨模态哈希学习算法.计算机学报，2019，42(1)：164-175)

[8] Zhao Qi-Lu and Li Zong-Min. Cross-modal social image clustering. Chinese Journal of Computers, 2018, 41(1): 100-113 (in Chinese)
(赵其鲁，李宗民，跨模态社交图像聚类.计算机学报，2018，41(1)：100-113)

[9] Fan C, Zhang X, Zhang S, Wang W, Zhang C and Huang H. Heterogeneous memory enhanced multimodal attention model for video question answering //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1999-2007

[10] Jiang P and Han Y. Reasoning with heterogeneous graph alignment for video question answering //Proceedings of the Association for the Advance of Artificial Intelligence. New York, USA, 2020: 11109-11116

[11] Gao L, Zeng P, Song J, Li Y-, Fang, Liu W, Mei T and Shen H T. Structured two-stream attention network for video question answering //Proceedings of the Association for the Advance of Artificial Intelligence. Honolulu, Hawaii, USA, 2019: 6391-6398

[12] Le T M, Le V, Venkatesh S and Tran T. Hierarchical conditional relation networks for video question answering //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9968-9978

[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. Attention is all you need //Proceedings of the International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[14] Li X, Song J, Gao L, Liu X, Huang W, He X and Gan C. Beyond rnns: positional self-attention with co-attention for video question answering //Proceedings of the Association for the Advance of Artificial Intelligence. Honolulu, USA, 2019: 8658-8665

[15] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X and Gool L V. Temporal segment networks: towards good practices for deep action recognition //Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands, 2016: 20-36

[16] Sivic J and Zisserman A. Video google: a text retrieval approach to object matching in videos //Proceedings of the IEEE International Conference on Computer Vision. Nice, France, 2003: 1470-1477

[17] Perronnin F and Dance C R. Fisher kernels on visual vocabularies for image categorization //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Minneapolis, Minnesota, USA, 2007: 1-8

[18] Jégou H, Douze M, Schmid C and Pérez P. Aggregating local descriptors into a compact image representation //Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 3304-3311

[19] Chollet and Francois. Xception: deep learning with depthwise separable convolutions //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1800-1807

[20] Tran D, Wang H, Torresani L, Ray J, LeCun Y and Paluri M. A closer look at spatiotemporal convolutions for action recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6450-6459

[21] Piergiovanni A, Angelova A and Ryoo M S. Tiny video networks. arXiv preprint arXiv:1910.06961, 2019

[22] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R and Fei-Fei L. Large-scale video classification with convolutional neural networks //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1725-1732

[23] Carreira J and Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6299-6308

[24] Xie S, Sun C, Huang J, Tu Z and Murphy K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification //Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018:305-321

[25] Tran D, Bourdev L D, Fergus R, Torresani L and Paluri M. Learning spatiotemporal features with 3d convolutional networks //Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497

[26] Lin J, Gan C and Han S. Tsm: Temporal shift module for efficient video understanding //Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 7083-7093

[27] Feichtenhofer C, Fan H, Malik J and He K. Slowfast networks for video recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 6202-6211

[28] Gao Z, Guo L, Guan W, Liu A-A, Ren T and Chen S. A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-R2. IEEE Transactions on Image Processing, 2020, 30: 767-782

[29] Hara K, Kataoka H and Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6546-6555

[30] He K, Zhang X, Ren S and Sun J. Deep residual learning for image recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[31] Xie S, Girshick R, Dollár P, Tu Z and He K. Aggregated residual transformations for deep neural networks //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1492-1500

[32] Pennington J, Socher R and Manning C D. Glove: global vectors for word representation //Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[33] Devlin J, Chang M-W, Lee K and Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding //Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 4171-4186

[34] Gu M, Zhao Z, Jin W, Hong R and Wu F. Graph-based multi-interaction network for video question answering. IEEE Transactions on Image Processing, 2021, 30: 2758-2770

[35] Guo D, Wang H and Wang M. Dual visual attention network for visual dialog //Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 4989-4995

[36] Huang D, Chen P, Zeng R, Du Q, Tan M and Gan C. Location-aware graph convolutional networks for video question answering //Proceedings of the Association for the Advance of Artificial Intelligence. New York, USA, 2020: 11021-11028

[37] Li X, Gao L, Wang X, Liu W, Xu X, Shen H T and Song J. Learnable aggregating net with diversity learning for video question answering //Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 1166-1174

[38] Jiang J, Chen Z, Lin H, Zhao X and Gao Y. Divide and conquer: question-guided spatio-temporal contextual attention for video question answering //Proceedings of the Association for the Advance of Artificial Intelligence. New York, USA, 2020: 11101-11108

[39] Jin W, Zhao Z, Gu M, Yu J, Xiao J and Zhuang Y. Multi-interaction network with object relation for video question answering //Proceedings of the ACM International Conference on Multimedia. Nice, France, 2019: 1193-1201

[40] Yang Z, Garcia N, Chu C, Otani M, Nakashima Y and Takemura H. Bert representations for video question answering //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 1556-1565

[41] Ibrahim M S, Muralidharan S, Deng Z, Vahdat A and Mori G. A hierarchical deep temporal model for group activity recognition //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1971-1980

[42] Jang Y, Song Y, Kim C D, Yu Y, Kim Y and Kim G. Video question answering with spatio-temporal reasoning. International Journal of Computer Vision, 2019, 127(10): 1385-1412

[43] Peng X, Zou C, Qiao Y and Peng Q. Action recognition with stacked fisher vectors //Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 581-595

[44] Girdhar R, Ramanan D, Gupta A, Sivic J and Russell B C. ActionVLAD: learning spatio-temporal aggregation for action classification //Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3165-3174

[45] Zhang J and Peng Y. Video captioning with object-aware spatio-temporal correlation and aggregation. IEEE Transactions on Image Processing, 2020, 29: 6209-6222

[46] Lin R, Xiao J and Fan J. Nextvlad: an efficient neural network to aggregate frame-level features for large-scale video classification // Proceedings of the European Conference on Computer Vision (ECCV) Workshops. Munich, Germany, 2018: 0-0

[47] Ye Y, Zhao Z, Li Y, Chen L, Xiao J and Zhuang Y. Video question answering via attribute-augmented attention network learning //Proceedings of the Special Interest Group on Information Retrieval. Tokyo, Japan 2017: 829-832

[48] Xu J, Mei T, Yao T and Rui Y. MSR-VTT: a large video description dataset for bridging video and language //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 5288-5296

[49] Arandjelovic R, Gronat P, Torii A, Pajdla T and Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition //Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307

[50] Yu Z, Yu J, Fan J and Tao D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering // Proceedings of the Proceedings of the IEEE International Conference on Computer Vision. 2017: 1821-1830

[51] Yu Z, Yu J, Xiang C, Fan J and Tao D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 5947-5959

[52] Ben-Younes H, Cadene R, Cord M and Thome N. Mutan: Multimodal tucker fusion for visual question answering // Proceedings of the Proceedings of the IEEE International Conference on Computer Vision. 2017: 2612-2620

[53] Cai J, Yuan C, Shi C, Li L, Cheng Y and Shan Y. Feature augmented memory with global attention network for videoqa // Proceedings of the International Joint Conference on Artificial Intelligence. Yokohama, Japan, 2020: 998-1004

[54] Fukui A, Park D H, Yang D, Rohrbach A, Darrell T and Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding //Proceedings of the Empirical Methods in Natural Language Processing. Austin, USA, 2016: 457-468

[55] Ren M, Kiros R and Zemel R S. Exploring models and data for image question answering //Proceedings of the International Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 2953-2961

[56] Yu Y, Ko H, Choi J and Kim G. End-to-end concept word detection for video captioning, retrieval, and question answering //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 3261-3269

[57] Gao J, Ge R, Chen K and Nevatia R. Motion-appearance co-memory networks for video question answering //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6576-6585

[58] Xu D, Zhao Z, Xiao J, Wu F, Zhang H, He X and Zhuang Y. Video question answering via gradually refined attention over appearance and motion //Proceedings of the ACM International Conference on Multimedia. Mountain View, USA, 2017: 1645-1653

[59] Zhao Z, Yang Q, Cai D, He X and Zhuang Y. Video question answering via hierarchical spatio-temporal attention networks // Proceedings of the International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 3518-3524

[60] Zhang W, Tang S, Cao Y, Pu S, Wu F and Zhuang Y. Frame augmented alternating attention network for video question answering. IEEE Transactions on Multimedia, 2020, 22(4): 1032-1041

[61] Zha Z-J, Liu J, Yang T and Zhang Y. Spatiotemporal-textual co-attention network for video question answering. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019, 15(2s): 1-18

[62] Van der Maaten L and Hinton G. Visualizing data using t-SNE. Journal of machine learning research, 2008, 9(11): 2579-2605

[63] Wang A J, Ge Y, Yan R, Ge Y, Lin X, Cai G, Wu J, Shan Y, Qie X and Shou M Z. All in One: Exploring Unified Video-Language Pre-training. arXiv preprint arXiv:2203.07303, 2022.

GUO Dan, Ph. D., professor. Her research interests include video analysis, pattern recognition and deep learning.

YAO Shen-Tao, M. S. candidate. His research interests include computer vision, video question answering.

WANG Hui, Ph.D. candidate. His research interests include computer vision, video question answering, visual dialogue.

WANG Meng, Ph. D., professor, IEEE Fellow. His research interests include machine vision, deep learning, multimedia information processing.

Background

In recent years, short video-based AI applications are popular. People share their daily lives conveniently. The amount of short videos on the Internet is huge, and how to effectively understand and summarize the core contents of these videos is an urgent issue to be solved. Video question answering (VideoQA) is a task to answer the question based on a video with rich visual content. It is an important and typical research topic in the fields of computer vision and natural language processing.

For VideoQA, current works usually adopt the encoder-decoder structure. Various encoders are used to extract the representations of video and question, *i. e.*, visual and textual feature sequences. Then, these visual and textual features are incorporated into a multimodal embedding vector for answer decoding via RNN cells, attention mechanisms, or graph-based modules. At last, according to each specific type of question-answer task, different decoders are employed to reason the correct answer. The existing works promise that they have obtained the feature sequences of frames, clips and words independently, and learned visual-textual correlation between video and question.

In this work, we argue that after feature extraction, merely modeling relationship between individual frames (vision), clips (vision) and words (language) is insufficient to infer the correct answer. Up to now, learning appropriate representation of video is still has a long way to develop, so is the representation of words in question too. For example, due to the dynamic and temporal nature of video, it tends to have redundant information, and directly performing cross-modal interaction will learn uninformative relations for VideoQA. Motived by this view, how to summarize core sequential clues in the video and question is a core technical point in this work. We hope and believe our work will insight other related video-based applications.

Our academic group devote to the research of visual & language understanding and reasoning, and has published some related works about visual dialog in the top conferences and journals such as IEEE TPAMI, IEEE TIP, CVPR, AAAI, ACM MM, and IJCAI. These works provide a preliminary foundation for the research on the VideoQA task.

This work was supported in part by the National Natural Science Foundation of China under Grant 62272144, Grant U20A20183, Grant 62020106007, and Grant 72188101, and in part by the Major Project of Anhui Province under Grant No. 202203a05020011.