动态视音场景下问答模型研究

段毛毛,连培榆,史海涛
(中国石油大学(北京)克拉玛依校区 石油学院,新疆 克拉玛依 834000)

摘 要:现实世界由大量不同模态内容构建而成,各种模态的信息相互关联和互补,充分挖掘不同模态之间的关系和特性能够有效弥补单一模态信息的局限性。动态视音场景下的问答模型研究,旨在通过视频中多模态信息回答不同视觉物体、声音及其相互联系的问题,使人工智能获得场景感知和时空推理能力。针对视音问答不准确的问题,提出了一种空间时序问答模型,该模型通过空间融合建模和时序融合建模对多模态特征进行融合,从而提高问答准确率。首先,分别使用 Resnet_18,VGGish 和 Bi-LST 对音频、视频和文字进行特征提取;其次,根据声音和视频的关系,在特征融合时对声音和视频两种模态进行早期的空间融合,并使用联合注意力机制在相互辅助学习后进行特征融合,增强特征互补性;最后,在特征融合后添加注意力机制以增强融合特征与文字的相关性。基于 MUSIC-AVQA 数据集的实验准确率达 73.49%,实现了场景感知和时空推理能力的提升。

关键词:视音问答;多模态融合;联合注意力机制;Bi-LSTM;MUSIC-AVQA
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2024)03-0163-07
doi:10.3969/j.issn.1673-629X.2024.03.024

Research on Question and Answer Models in Dynamic Audio-visual Scenarios

DUAN Mao-mao,LIAN Pei-yu,SHI Hai-tao
(School of Petroleum,China University of Petroleum-Beijing at Karamay,Karamay 834000,China)

Abstract:The real world consists of a variety of modalities,and information from different modalities is interrelated and complementary. Exploring the relationships and characteristics between different modalities can effectively compensate for the limitations of individual modalities. The research on dynamic audiovisual question answering (QA) models aims to use multimodal information from videos to answer questions about visual objects,sounds,and their relationships,enabling artificial intelligence to achieve scene understanding and spatio-temporal reasoning capabilities. To address the problem of imprecise audiovisual QA,a spatio-temporal question answering model is proposed. This model combines spatial fusion modelling and temporal fusion modelling to integrate multimodal features and improve the accuracy of QA. Firstly,audio,video and text features are extracted using ResNet-18,VGGish and Bi-LSTM respectively. Secondly,an early fusion approach is applied to spatially fuse the audio and video modalities based on their relationship. Then,a joint attention mechanism is applied to fuse the features after mutual learning to enhance their complementarity. Finally,a post-fusion attention mechanism is added to enhance the correlation between the fused features and the text. Experimental results on the MUSIC-AVQA dataset show an accuracy of 73.49%,indicating the improvement in scene understanding and spatio-temporal reasoning capabilities achieved by the proposed model.

Key words:audio-visual question and answer;multimodal fusion;joint attention mechanism;Bi-directional Long Short-Term Memory;MUSIC-AVQA

0 引 言

日常生活中,人类无时无刻处于视音场所中。但在复杂的视音场景下,跨视音视图关联对象或声音事件,其中大多数仍然仅具有有限的跨模态推理能力。因此,整合多模态信息以获得类人的客观场景感知和理解能力是一个有趣而有价值的研究课题。

近年来,深度学习的逐步发展使其被应用于多模态特征融合。深度学习模型可以处理高维复杂的多模态信息,而多模态深度学习具有在人类层面处理多模态数据的潜力。

收稿日期:2023-06-15 **修回日期**:2023-10-18
基金项目:克拉玛依市创新人才专项(XQZX20220047)
作者简介:段毛毛(1990-),女,工程师,硕士,研究方向为虚拟仿真、多模态;通信作者:连培榆(2001-),男,研究方向为软件工程。

早期的问答任务主要以自然语言的形式进行提问和回答,后来随着人工智能、深度学习、多模态等技术发展逐渐演变为文字、图片、音频、视频等多种模态间信息查询的广义问答系统。

Heeseung Yun 等人[1]提出了一种新的基于 360 度空间和视音问答的基准任务,结果表明球形空间嵌入和多模态训练目标有助于更好地理解数据集上的全景环境,正确率达到 68.93%。Jing L 等人提出的自监督[2]学习方法为多模态视音场景分析模型的建立提供了新的思路,通过将视频信号中的视音信息进行融合,成功地解决了视音信息能否在同一时间上[3]的问题。Hori 等人[4]提出了一种对话问答系统,使用注意力机制对视音场景进行感知。Li Guangyao 等人[5]提出了基于动态视音场景下的问答模型,正确率达到了 71.5%。

视音问答模型仍存在以下问题:(1)一些与问题无关的视觉物体或声源均参与单模态编码,需寻找更适合后续融合的特征提取方式。(2)在融合学习的过程中,需寻找其他多模态融合方式,以提高模态间的互补性,进而提高问答模型的正确率。

为解决上述问题,该文通过分析现有视音问答模型得知文本所含信息量最大,因此先对视频和音频信息进行融合获得更多信息以支撑文本特征信息。首先,分别使用 Resnet_18,VGGish 和 Bi-LST 对音频、视频和文字进行单模态特征提取;然后,通过空间融合模块对视频和音频特征进行融合,将复杂的场景分解为具体的视音关联;最后,通过联合注意力机制对文字、视频和音频进行混合学习,实现视频特征、音频特征和文字特征的融合,增强三种模态之间的关联关系。基于联合注意力建立空间时序模型,进一步提高了动态视音场景下问答的准确率,提升了模型场景感知和时空推理能力。

1 相关工作

对多模态问题进行深入研究,充分利用多种模态之间的互补性和冗余性,是推动人工智能更好地了解和认知周围世界的关键。多模态学习中的两个重点问题是多模态数据的异质性差距[6]和模态间的信息融合[7]。在研究多模态问题时,如何充分挖掘模态之间的信息和消除数据异构,一直是多模态任务的主要挑战[8]。

在多模态问答任务中主要以视频问答(Video Question Answer)[9]和视觉问答(Visual Question Answer)[10]为主,视频问答早于视觉问答。视频问答根据序列的图像信息和时间线索,针对不同问题提取不同数量的帧求取答案,多采用融入注意力机制[11-12]和融入记忆网络的思路构建问答模型[13-14]。视觉问答基本分为联合嵌入、注意力机制、神经网络架构和知识库增强[15],以一幅图像和一个问题作为输入,通过对多模态信息进行融合与推理,以自然语言的形式给出问题的答案[16]。

在视音场景的研究中,合适的多模态融合方式尤为重要。曲志等人[17]提出了一项裂纹检测方法,使用多尺度卷积特征融合模块进行裂纹信息提取实现特征的全面融合。李钊[18]提出一项基于深度学习的跨模态检索方法,实现更好的跨模态相似度度量。Fu 等人[19]提出了一种基于双注意网络场景分割结果,该方法将局部特征与其全局依赖性相结合,并且可以捕获丰富的上下文信息。Peng 等人[20]提出了一种基于注意力引导的多视图融合网络来解决三维物体识别问题,可以在更为复杂的场景下取得更好的效果。

注意力机制在多模态融合任务中也展现出了其重要性。Schwartz 等人[21]提出的高阶注意力模型是一种将文本、图像和文本答案三种不同的模态信息进行多模态融合的模型,得到更准确的答案。Chen 等人[22]提出了一种条件注意力融合策略在连续维度情绪预测中的应用,提高预测的准确性和可靠性。Li 等人[23]提出了一种针对多模型多标签分类任务的通用策略,通过选择更好的特征组合来实现更精确的分类。杨清溪等人[24]提出了一种基于注意力机制的场景识别模型,大大提高了场景识别的准确度。该文亦将采用注意力机制实现各模态特征的提取与融合。

2 基于联合注意力机制的视音问答模型

模型使用补充约束的特征来应对单模态数据不足,同时采用多模态联合表示的思想,使得不同模态信息相互补充、相互学习、相互制约。通过拼接和矩阵点乘法将三个模态特征混合为一个特征,建立视音字空间时序模型。

2.1 整体结构

注意力机制能够对输入数据的每个部分提供不同的权重信息,从而抽取出重要关键的特征信息,使得模型获得更加准确的捕捉判断。该文所提出的基于联合注意力机制的视音场景下的问答模型结构如图 1 所示。模型主要包括三个组成部分:

(1)视音字单模块:含视频模块、音频模块、和文字模块,分别采用适合各个模态的网络结构,对视频、音频、文字三种模态数据进行特征提取,以便后续进行融合学习。

(2)空间融合模块:采用多模态联合表示的思想,使用矩阵叉乘法对视频特征和音频特征进行融合学习得到混合特征,完成视音融合,建立空间模型。

(3)空间时序融合模块:采用一种新的多模态协同表示方法,通过联合注意力机制对文字、视频和音频进行混合学习,实现视频特征、音频特征和文字特征的融合。

视频模块、音频模块和文字模块分别选用视频、音频和文字嵌入向量作为各自模块分支的输出数据,并将其作为空间融合模块中视频分支的输入数据。

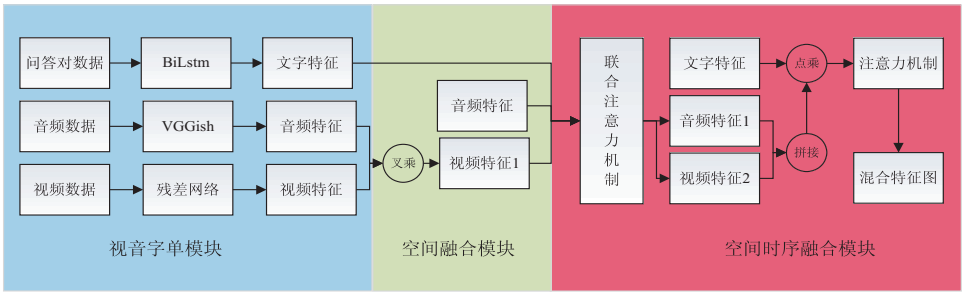

图 1　整体结构

2.2　视频模块

视频模块使用 Resnet_18 进行视频特征提取。为了保持所有视频的信息完整,对所有视频片段采样固定数量的帧。

Resnet_18 的具体结构如图 2 所示,包含 17 个卷积层和 1 个全连接层,使用两种不同大小的卷积核(7×7 和 3×3)提取视频图像特征。在卷积层中,采用了非线性的 Relu 激活函数,最终输出的数据维度设置为 320,512,14,14。视频特征提取的具体步骤如下:

(1)提取全局特征:使用 7×7 卷积核提取特征后,采用最大池化层降维。

(2)提取细节特征:使用 3×3 卷积核提取特征后,为获得更完整的图像信息加入平均池化层。

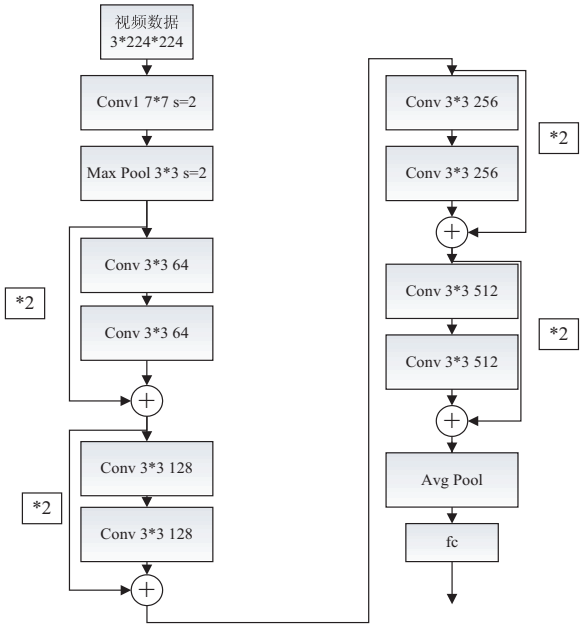

图 2　Resnet_18 结构

2.3　音频模块

音频模块使用 VGGish 进行音频特征提取,对每个音频片段进行编码,为音频信息赋予语义信息。VGGish 的具体网络结构如图 3 所示。

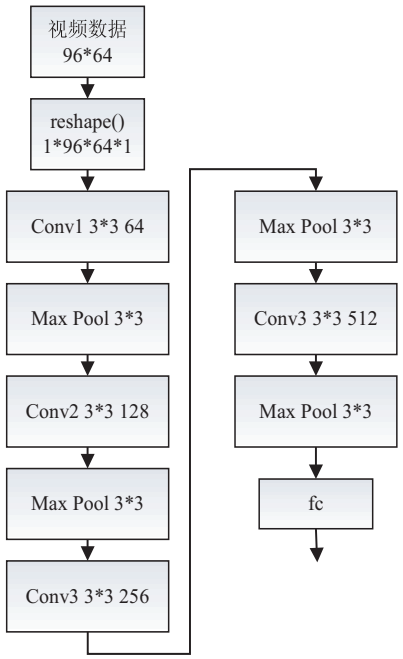

图 3　VGGish 结构

VGGish 由四个卷积层、四个池化层和一个全连接层组成,卷积核大小为 3×3,池化层选择最大池化法,卷积层采用了非线性激活函数 Relu,方便进行卷积处理,最终输出的数据维度为 64,10,512。音频特征提取的具体步骤如下:

(1)将数据重塑为 4 维,便于进行卷积运算。

(2)使用 3×3 卷积核和最大池化层得到音频特征。

2.4　文字模块

文字模块使用 Bi-LSTM 进行文字特征提取,对文字特征进行更细粒度的分类。该文对每个问答对进行词嵌入,具体网络结构如图 4 所示。

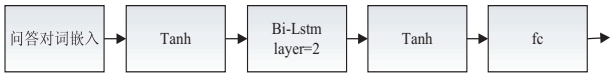

图 4　Bi-LSTM 结构

该模块中 Bi-LSTM 模型由两层双向 LSTM 和一

个全连接层组成，词嵌入后先选用非线性激活函数 Tanh 激活函数，再通过含有一个隐藏层的 Bi-LSTM 提取文字特征，最终输出数据维度为 64，512。

2.5 空间融合模块

声音及其视源的位置能够反映视音模态之间的空间关联，因此引入基于声源定位的空间融合模型，将复杂的场景分解为具体的视音关联[5]。具体模型结构如图 5 所示。

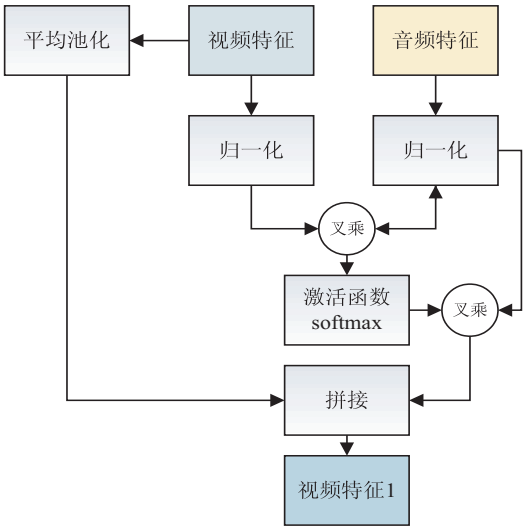

图 5 空间融合网络结构

该模块具体处理步骤如下：

(1)为了方便后续的联合表示，先使用线性变换将之前得到的配对视频音频特征转换为 512，512。

(2)为了防止视频信息丢失，先对视频特征进行平均池化，得到全局视频特征。

(3)将视频特征和音频特征进行归一化处理后进行矩阵叉乘，使用 softmax 激活函数再与音频特征进行矩阵叉乘，得到混合视频特征维度为 512，512。

(4)拼接两个视频特征，得到最终混合的视频特征 1，建立空间模型。

2.6 空间时序融合模块

为了突出与问题密切相关的关键时间戳，使用联合注意力机制进行视听特征与文字特征的协同表示。首先加入 relu 激活函数和 dropout 层，得到关键时间戳下的视听特征；然后拼接视听特征，将其联合表示为一个混合特征，具体网络结构如图 6 所示。

空间时序融合模块数据处理步骤如下：

(1)以文字特征为查询增强关键时间戳下的视频特征 1，得到视频特征 2。

(2)以文字特征为查询增强关键时间戳下的音频特征，得到音频特征 1。

(3)以文字特征为查询，以音频特征 1 为键值，增强关键时间戳下视频特征 2 中与音频相关的信息，得到视频特征 3。

(4)以文字特征为查询，以视频特征 3 为键值，增强关键时间戳下音频特征 1 中与视频相关的信息，得到音频特征 2。

(5)将协同表示后的视频特征 3 和音频特征 2 进行归一化处理，再与空间融合后的视频特征 1 和音频特征相加，以防止信息丢失。

(6)将视频特征 3 和音频特征 2 进行拼接，再与文字特征点乘得到联合表示的混合特征 1。

(7)以文字特征为查询增强混合特征中的文字信息，得到混合特征 2。

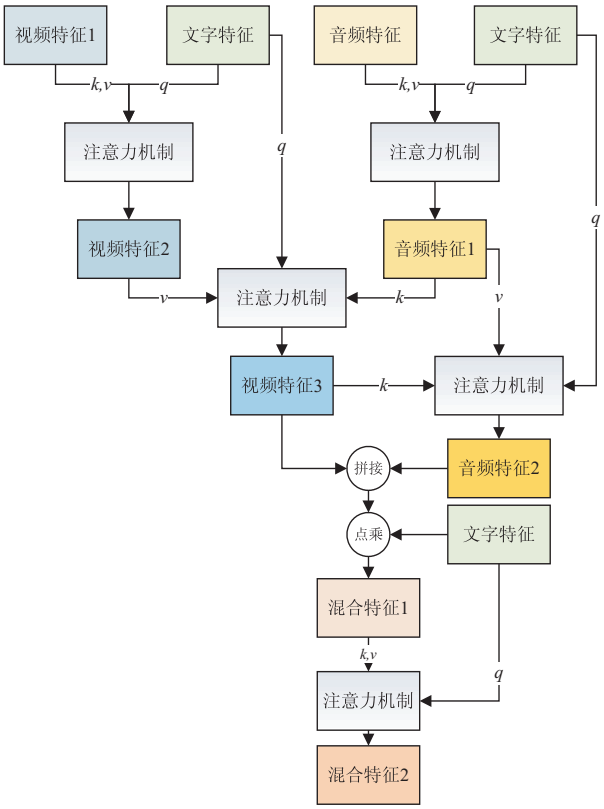

图 6 空间时序融合结构

该文采用两阶段训练策略，先建立空间模型，损失函数使用交叉熵损失函数 L_s，公式如下：

$$L_s(p,q) = -\sum_{i=1}^{n} p_i \log q_i \tag{1}$$

其中，p_i 为真实值，q_i 为预测值。

第二阶段建立空间时序模型，损失函数 L 的计算公式如下：

$$L = L_{qa} + 0.5L_s \tag{2}$$

其中，L_s 为第一阶段的交叉熵损失，L_{qa} 为第二阶段的交叉熵损失。

3 实验和模型优化

3.1 实验环境

实验采用了 Pytorch 框架，具体实验环境如表 1 所示。

表 1　实验环境

软硬件设备名称	内容
操作系统	Linux
深度学习框架	TensorFlowTorch1.13
CPU	15 vCPU Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60 GHz
GPU	RTX3080(10 GB)
显存	80 GB
开发平台	Python3.9
开发工具	Pycharm2022.2.4

3.2　数据集及其预处理

MUSIC-AVQA(Spatial-Temporal Music AVQA)大规模视听数据集总计包含 9 288 个演奏视频,真实视频和合成视频分别占 79.9% 和 20.1%。其中,真实视频由 14.8% 的单人视频、71.7% 的二重唱视频和 13.5% 的其他合奏视频组成。数据集涵盖 22 种不同乐器(如吉他、钢琴、二胡、唢呐等),总时长超过 150 小时。

此外,MUSIC-AVQA 具有 45 867 个问答对,平均每个视频约 5 个问答对,这些问答对涵盖了 3 种不同场景(声音(Audio)、视觉(Visual)和视音(Audio-Visual))下的 9 类问题类型(以存在 Existential、计数 Counting、位置 Location、比较 Comparative 和事件 Temporal 为主)以及 33 个不同的问题模板。3 种场景及对应的问题类型如表 2 所示。

表 2　MUSIC-AVQA 数据集问答对类型划分

场景	问答类型	占比/%	举例
视音	存在	56.7	Is there a voiceover?
	计数		How many instruments in the video did not sound?
	位置		Which is the instrument on the right of piano?
	比较		Is the violin on the left more rhythmic than the electric_bass on the middle?
	事件		What is the third instrument that comes in?
视觉	计数	26.6	How many types of musical instruments appeared in the entire video?
	位置		Where is the performance?
声音	计数	16.7	How many musical instruments were heard throughout the video?
	比较		Are there saxophone and accordion sound?

将数据集随机分为训练集、验证集和测试集,分别包含 32 087,4 595 和 9 185 对问答对。在特征提取前,对每段视频中的声音和视频帧进行采样,采样率分别为 16 kHz 和 1 fps,将其分为长度相等、互不重叠的 1 帧视频段。此外,使用 6 s 取 1 s 的方式对视频进行采样,使用 normalize 函数对图像进行标准化。

3.3　模型训练

(1)音频特征提取:将音频文件载入 VGGish 网络,得到 64,10,512 维音频特征,完成音频嵌入。

(2)视频特征提取:将视频文件载入 Resnet_18 网络,先得到 16,20,3,244,244 维特征,后重组维度为 320,512,14,14,以便后续统一为 512-特征。

(3)文字特征提取:将问答对文件载入 Bi-LSTM 网络,先得到 64,512 维特征,完成词嵌入,后通过词特征组成问答对特征,完成问答嵌入。

(4)建立空间融合模型:构建空间融合模型网络结构,将视频特征和音频特征调整为相同维度,训练得到维度 512,512 混合特征。

(5)建立空间时序模型:构建空间时序模型网络结构,设置批大小和轮次分别为 64 和 30。学习率初始为 e^{-4},每训练十轮乘 0.1,学习率下降,使用 Adam 优化器,得到最终的混合特征,完成模型建立。

3.4　模型评估与测试

使用答案预测精度作为度量标准,评估模型在回答不同类型问题时的表现。答案词汇表由 42 个可能的答案(22 个对象,12 个计数选择,6 个未知类型,是/否)组成,用于回答数据集中不同类型的问题。

评估测试步骤具体如下:

(1)在每轮训练完毕进行模型评估,对评估集中的问题进行预测,得到评估结果。

(2)在完成所有训练后进行模型测试,对测试集中的问题进行预测,得到测试结果。

(3)设置基线。在相同环境下训练 MUSIC-AVQA 的网络结构并评估测试,以便与文中模型进行比对。

3.5 不同模态消融实验

如表 3 所示,实验结果表明 V+Q 比 A+Q 效果更好,表明视频特征是问答模型中的强信号。加入视音字空间时序融合模块(CTG)后,虽然独立的音频问答和视频问答准确率稍差,但是视音问答的准确率最佳,证明了视音字空间时序融合模块的有效性。

表 3 不同模态与不同模块消融研究

方法	声音问答	视觉问答	视音问答	全部
Q	65.59	44.42	54.98	53.98
A+Q	67.15	61.81	63.67	64.06
V+Q	68.22	67.01	62.97	64.89
AV+Q	70.24	69.54	65.67	67.22
AV+Q+TG	72.81	71.64	68.01	70.05
AV+Q+TG+SG	**73.81**	**73.60**	69.14	71.10
AV+Q+CTG+SG	73.62	72.88	**73.67**	**73.49**

(Q:问答文字,A:音频模态,V:视频模态,AV:视音模态,TG:时序融合模块,SG:空间融合模块,CTG:视音字空间时序融合模块)

3.6 不同方法对比实验

为了验证模型的有效性,分别从声音问答、视觉问答和视音问答出发,对比了各种方法在计数、比较或空间等类型问题回答的准确率,结果如表 4 所示。其中文献[5]为 MUSIC-AVQA 数据集官方空间时序问答模型。进一步优化官方模型,该文在文本特征融入后,加入关键时间戳下的视频和音频特征的辅助学习,从而提高三种模态之间的关联程度。

实验结果表明,文中方法的性能在音频问答和视频问答上相比文献[5]的略次,但在视音问答上,除存在类问题准确率较差,其余类型问题的准确率均高,视音问答的平均准确率达 73.67%,为最佳。

该文建立的空间时序模型更有效地互补了单个模态的信息缺失,加强了通过问题查找关键图像和声音能力,从而增强了模型的时空推理的能力,提升了视音问答的准确率。然而由于模型过于关注三种模态的关联关系,对于不需要三种模态的问答增加了干扰信息,导致单模态音频问答和视频问答准确率降低,仍需进一步完善。

表 4 问答方法对比

任务	方法	声音问答			视觉问答			视音问答						总平均
		计数	比较	平均	计数	空间	平均	存在	空间	计数	比较	时间	平均	
音频	FCNLSTM[25]	70.32	66.17	68.43	63.57	46.58	54.96	81.84	46.05	59.11	62.02	47.17	59.89	60.12
问答	CONVLSTM[25]	73.91	68.74	71.89	67.27	54.43	60.81	**82.31**	50.59	62.89	60.15	51.21	62.08	63.28
视觉	GRU[10]	72.14	66.71	70.14	67.52	70.01	68.78	81.54	59.25	62.54	61.68	59.79	64.91	66.89
问答	MCAN[26]	77.35	54.79	69.11	71.32	70.88	71.10	80.14	54.19	64.76	57.04	47.24	61.24	65.18
视频	PSAC[27]	75.55	66.01	71.99	68.45	69.62	68.92	77.23	54.86	63.25	61.08	59.31	63.38	66.23
问答	HME[28]	74.49	63.13	70.56	67.84	69.32	68.54	80.12	53.04	63.01	62.59	59.74	63.84	66.14
视音	文献[5]	**77.94**	66.95	**73.81**	71.16	**76.08**	**73.60**	80.98	64.14	69.84	65.46	63.10	69.14	71.10
问答	文中	73.94	**70.49**	73.62	**72.89**	72.87	72.88	66.92	**73.04**	**75.85**	**75.69**	**76.4**	**73.67**	**73.49**

4 结束语

该文主要使用联合注意力机制对多模态信息进行融合,建立动态视音场景下的空间时序问答模型。实验结果表明该模型性能较好,有助于视音问答的准确率。文中工作存在以下缺陷和改进空间:

(1)实验仅使用 MUSIC-AVQA 数据集,在其它数据集和场景上的应用效果不得而知,需增加数据集测试以验证模型的性能。

(2)单模态特征提取方式有待改进,后期可以选取更匹配的特征提取方式来进行特征提取,以提高后续多模态特征融合的效率。

(3)使用联合注意力机制进行多模态特征融合,后续可以改进多模态融合方式,进一步提高多模态视频问答任务的性能。另外,选用集成模型策略也可作为该模型的一个发展方向,以进一步提高性能并拓展应用范围。

参考文献:

[1] YUN H,YU Y,YANG W,et al. Pano-AVQA:grounded audio-visual question answering on 360 circ videos[J]. arXiv:2110.05122,2021.

[2] JING L,TIAN Y. Self-supervised visual feature learning with deep neural networks:a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,43(11):4037-4058.

[3] OWENS A,EFROS A A. Audio-visual scene analysis with self-supervised multisensory features[J]. arXiv:1804.03641,2018.

[4] HORI C. End-to-end audio visual scene-aware dialog using multimodal attention-based video features[C]//IEEE international conference on acoustics, speech and signal process-

ing (ICASSP). Brighton:IEEE,2019:2352-2356.

[5] LI Guangyao, WEI Yake, TIAN Yapeng, et al. Learning to answer questions in dynamic audio-visual scenarios[C]// 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). New Orleans: IEEE, 2022: 19086-19096.

[6] PENG Y, QI J. CM-GANs: cross-modal generative adversarial networks for common representation learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15(1): 1-24.

[7] GUO W, WANG J, WANG S. Deep multimodal representation learning: a survey[J]. IEEE Access, 2019, 7: 63373-63394.

[8] 杜鹏飞,李小勇,高雅丽. 多模态视觉语言表征学习研究综述[J]. 软件学报,2021,32(2):327-348.

[9] TU K W, MENG M, LEE M W, et al. Joint video and text parsing for understanding events and answering queries[J]. IEEE MultiMedia, 2014, 21(2): 42-70.

[10] AGRAWAL A, LU J S, ANTOL S, et al. VQA: visual question answering[J]. International Journal of Computer Vision, 2017, 123(1): 4-31.

[11] 梁丽丽,刘昕雨,孙广路,等. MSAM:针对视频问答的多阶段注意力模型[J]. 哈尔滨理工大学学报,2022,27(4):107-117.

[12] GAO D F, WANG R P, BAI Z Y, et al. Env-QA: a video question answering benchmark for comprehensive understanding of dynamic environments[C]//Proceedings of the 2021 IEEE/ CVF international conference on computer vision. Montreal: IEEE, 2021: 1655-1665.

[13] GE Y Y, XU Y J, HAN Y H. Video question answering using a forget memory network[C]//Proceedings of the 2nd CCF Chinese conference on computer vision. Tianjin: Springer, 2017: 404-415.

[14] KIM J, MA M, KIM K, et al. Progressive attention memory network for movie story question answering[C]//Proceedings of the 2019 IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 8329-8338.

[15] 闫　悦,郭晓然,王铁君,等. 问答系统研究综述[J]. 计算机系统应用, 2023: 1-18. https://doi.org/10.15888/j.cnki.csa.009208.

[16] 余　宙,俞　俊,朱俊杰,等. 融合知识表征的多模态 Transformer 场景文本视觉问答[J]. 中国图象图形学报, 2022,27(9):2761-2774.

[17] QU Z, CAO C, LIU L, et al. A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4890-4899.

[18] 李　钊. 多模态数据分类与检索的关键技术研究[D]. 北京:北京交通大学,2018.

[19] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[J]. arXiv: 1809.02983, 2018.

[20] PENG B, YU Z, LEI J, et al. Attention-guided fusion network of point cloud and multiple views for 3D shape recognition[C]//2020 IEEE international conference on visual communications and image processing (VCIP). Macau, China: IEEE, 2020: 185-188.

[21] SCHWARTZ I, SCHWING A G, HAZAN T. High-order attention models for visual question answering[J]. arXiv: 1711.04323, 2017.

[22] CHEN S, JIN Q. Multi-modal conditional attention fusion for dimensional emotion prediction[J]. arXiv: 1709.02251, 2017.

[23] LI P, LI X. Multimodal fusion with co-attention mechanism[C]//2020 IEEE 23rd international conference on information fusion (FUSION). Sun City: IEEE, 2020: 1-8.

[24] 杨清溪,张丽红. 基于语义信息的场景识别方法研究[J]. 测试技术学报,2021,35(6):521-528.

[25] FAYEK H M, JOHNSON J. Temporal reasoning via audio question answering[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2283-2294.

[26] YU Zhou, YU Jun, CUI Yuhao, et al. Deep modular co-attention networks for visual question answering[C]//2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach: IEEE, 2019: 6274-6283.

[27] LI X, SONG J, GAO L, et al. Beyond RNNs: positional self-attention with co-attention for video question answering[C]//Proceedings of the AAAI conference on artificial intelligence. [s.l.]: AAAI, 2019: 8658-8665.

[28] FAN Chenyou, ZHANG Xiaofan, ZHANG Shu, et al. Heterogeneous memory enhanced multimodal attention model for video question answering[C]//2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Long Beach: IEEE, 2019: 1999-2007.