计算机科学
COMPUTER SCIENCE

基于层次化视觉注意力的富语义视频对话生成

赵倩，郭斌，刘宇博，孙卓，王豪，陈梦琦

引用本文

赵倩，郭斌，刘宇博，孙卓，王豪，陈梦琦. 基于层次化视觉注意力的富语义视频对话生成[J]. 计算机科学, 2025, 52(1): 315-322.

ZHAO Qian, GUO Bin, LIU Yubo, SUN Zhuo, WANG Hao, CHEN Mengqi. Generation of Enrich Semantic Video Dialogue Based on Hierarchical Visual Attention [J]. Computer Science, 2025, 52(1): 315-322.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100107

基于层次化视觉注意力的富语义视频对话生成

赵 倩 郭 斌 刘宇博 孙 卓 王 豪 陈梦琦
西北工业大学计算机学院 西安 710129
(qzhao@mail.nwpu.edu.cn)

摘 要 视频对话是多模态人机交互领域中的重要内容。视频对话中包含大量时空视觉信息和复杂的多模态关系,这给相关研究带来了巨大的挑战。现有的视频对话模型利用跨模态注意力机制或图结构捕捉视频语义和对话上下文之间的相关性,然而,所有视觉信息均是在单一粗粒度下处理的,这导致模型容易忽略一些细粒度时空信息,如同一物体在时间上的持续运动或图像不显著位置的物体信息,从而降低了视频对话性能。同时,细粒度处理全部视觉信息又将增加处理时延,降低视频对话的流畅性。因此,提出了一种层次化视觉注意力的富语义视频对话生成方法。首先根据对话上下文,利用全局视觉注意力捕捉全局视觉语义信息,并定位到对话输入关注的视频时间序列/空间范围,其次利用局部注意力机制进一步捕捉细粒度视觉信息,结合多任务学习方法,生成对话回复。在DSTC7 AVSD数据集上的实验结果表明,相比现有基准方法,所提方法生成的对话具备更高的准确性和多样性,其中METEOR指标提高了23.24%。

关键词:多模态人机交互;层次化注意力机制;多任务学习;场景感知
中图分类号 TP391

Generation of Enrich Semantic Video Dialogue Based on Hierarchical Visual Attention

ZHAO Qian,GUO Bin,LIU Yubo,SUN Zhuo,WANG Hao and CHEN Mengqi
School of Computer Science,Northwestern Polytechnical University,Xi'an 710129,China

Abstract As an important research direction in the field of multimodal human-computer interaction,video dialogue emerges. The large amount of temporal and spatial visual information and complex multimodal relationships makes it challenging to design efficient video dialogue systems. Existing video dialogue systems utilize cross-modal attention mechanisms or graph structures to capture the correlation between video semantics and dialogue context. However,all visual information is processed with a single coarse granularity. It results in a loss of some fine-grained temporal and spatial information,such as the continuous motion of the same object and the insignificant position information of an image. Moreover,the fine-grained process of all visual information increases the delay and degrades the dialogue fluency. Therefore,we propose a hierarchical visual attention-based semantic-rich video dialogue generation method in this paper. Firstly,according to the dialogue context,global visual semantic information is captured by using global visual attention and located to the time sequence/spatial scope of the video associated with the dialogue input. Secondly,the local attention mechanism is used to further capture fine-grained visual information in the localized area,and to generate the dialogue response by exploiting the multi-task learning method. Experimental results on DSTC7 AVSD datasets show that the dialogue generated by the proposed method has higher accuracy and variety,and its METEOR index improves by 23.24%.

Keywords Multi-modal human-computer interaction,Hierarchical attention mechanism,Multi-task learning,Scene perception

1 引言

随着多媒体技术的发展,问答系统已经从传统的纯自然语言领域扩展到了多模态领域。多模态人机交互作为人-机-物的技术载体,旨在通过整合多种输入、输出通道(如视频、图像、语音等)来提高用户的交互效率和体验,实现人与计算机之间自然、流畅的信息交换。随着人工智能技术的发展,以人为中心,更加细致地理解和模拟人类的行为和语言表达,成为未来智能化、自然化人机交互的重要发展方向[1]。

视觉问答(VQA)任务的目标是基于图像中存在的视觉信息来回答用户提出的问题[2]。视频对话(Video Dialogue)任务可被视为视觉问答任务的扩展,如图1所示,可以直观地

到稿日期:2023-11-19 返修日期:2024-05-06
基金项目:国家杰出青年科学基金(62025205);国家自然科学基金(62032020,62102322)
This work was supported by the National Science Fundation for Distinguished Young Scholars of China(62025205) and National Natural Science Foundation of China(62032020,62102322).
通信作者:郭斌(guob@nwpu.edu.cn)

看出两者之间的差异。

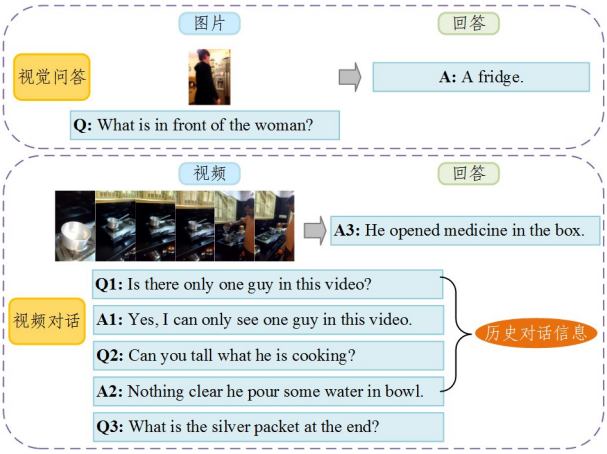

图 1 视觉问答与视频对话任务对比图

Fig. 1 Comparison of VQA and video dialogue

视频对话生成任务指在特定场景下，根据视频内容、历史对话和具体问题，通过推理生成流畅的自然语言，以实现与现实世界的交互[3]。与视觉问答任务相似，视频对话需要深入理解视频中的视觉概念和关系，并根据当前的问题进行推理和回答。

但是，相比视觉问答，视频对话不仅要处理图像特征提取本身的复杂性，还需要考虑视频帧之间的时空关联信息[4]，因此面临着特有的挑战：1）采用全局处理视频信息的方式，忽略了视频特有的时间和空间特征，将所有视觉信息等权重处理，导致无法有效地捕捉细粒度的时空信息；2）视频对话所涉及的场景复杂多样，现有的视频对话方法在不同场景下的对话质量和效果稳定性差异较大；3）视频对话需从语法、语义、逻辑等多个层面对多模态信息进行深度理解与表达推理，对生成内容的连贯性和自然度要求较高。

本文旨在利用层次化视觉注意力机制解决视听场景感知对话任务中的富语义视频对话生成问题。为此，首先对层次化注意力进行建模；然后根据对话上下文，利用全局视觉注意力捕捉全局视觉语义信息，通过定位与输入对话相关性高的视频时间序列和空间范围，利用局部注意力机制进一步捕捉细粒度视觉信息；最后结合多任务学习方法，生成对话回复。本文的主要贡献包括以下 3 个方面：

（1）提出了一种基于层次化视觉注意力的视频特征提取模型。该模型从视频的画面特征、运动特征中提取信息；同时，考虑问题的粗粒度特征和细粒度单词特征，并逐步改进其注意力机制。该模型可扩展性较强，后续可应用到包含文本、音频等多种模态的信息渠道中。

（2）提出了一种基于多任务学习的富语义视频对话生成方法，主要包括视频场景描述、音视频序列建模和对话响应生成三大模块，旨在进一步适应多样化的应用场景。通过共享模型参数和学习多个任务之间的相关性，以提高模型的效率和准确性。

（3）为证明所提模型的有效性，本文在 DSTC7 AVSD 数据集上进行实验。实验结果表明，该方法在 BLEU，METEOR 等多个评价指标上均优于现有的基准方法。与此前该任务的 SOTA 模型相比，本文模型的 METEOR 指标提高了 23.24%。

2 相关工作

2.1 对话生成技术

对话生成任务旨在提供与给定对话上下文相关的响应回复[5]。Serban 等[6]对分层递归编码器-解码器神经网络进行了扩展，用于对话系统中的响应学习。Weston[7]提出了一种基于记忆网络的对话式语言学习方法，其中监督信号来自对话伙伴的响应，实现了自然隐式学习。

随着注意力机制的发展，Xing 等[8]提出了一种分层递归注意力网络，用于多轮响应生成。为模拟句子和其多样化响应之间的关系，Zhou 等[9]假设存在一些潜在的响应机制，并构建了一个具有不同响应机制的编码器-分流器-解码器框架，用于对话响应生成。Wu 等[10]提出了一种基于检索的多轮响应选择任务的顺序匹配网络，并采用多层匹配机制来提取多粒度匹配信息。

近年来，以 GPT[11]系列为代表的开放领域对话模型取得了显著进展。其中，ChatGPT[12]是一种由 OpenAI 开发的生成式人工智能（AIGC）大语言模型。该模型基于 Transformer 神经网络架构进行了大规模的预训练和微调，并结合基于人类反馈的强化学习（RLHF），实现了与人类的高质量对话交互。然而，该模型目前仍存在可解释性差、训练成本高等问题[13]。

2.2 视频对话系统

在多模态研究领域，视频对话作为一项新兴任务备受关注。与以往图像问答、图像摘要等任务的研究不同，视频对话任务需要基于视频内容和对话上下文语境生成响应回复。这意味着视频对话需要处理更为复杂的时间结构和空间信息，因此更具有挑战性。

为推动视频对话研究的发展，Yu 等[14]于 2017 年提出了一种语义注意力机制，将视频中检测到的概念与文本解码相结合生成对话回复。与图像相比，视频具有独特的时间域特性，因此需要通过时间分析来获得正确回复。为此，Jang 等[15]提出了时序注意力机制，以便选择性关注视频中的一个或多个时段。Garcia 等[16]同时结合了视频场景描述和外部知识推理。最近，Le 等[17]使用基于视频的神经网络提取视觉线索，并对视频对话任务中的信息检索过程进行建模。然而，这些模型大多关注对视频内容的理解，缺少对视频中运动信息的分析。

本文提出了一种基于层次化视觉注意力的特征提取方法，用于捕捉视频画面特征和运动特征之间的关系，从而生成语义丰富、逻辑连贯的对话响应。整体框架如图 2 所示。

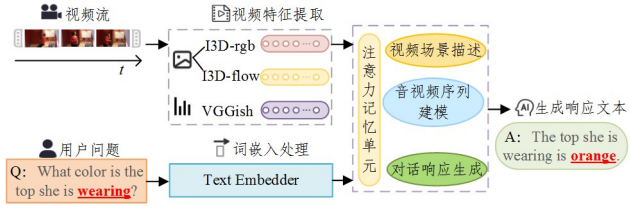

图 2 系统架构示意图

Fig. 2 Schematic diagram of system architecture

3　层次化视觉注意力多模态特征提取

在视频特征提取的过程中，以相同粒度对所有视频信息进行分析计算，往往会导致模型忽略部分视频所独有的细粒度时空特征，从而丢失一些重要信息。因此，本文提出了一种基于层次化视觉注意力机制的多模态特征提取方法，旨在根据用户所提出的对话问题，针对性选择并提取不同层次的时空序列特征，以获取更加有价值的关键性信息。图 3 为模型实例化图示。

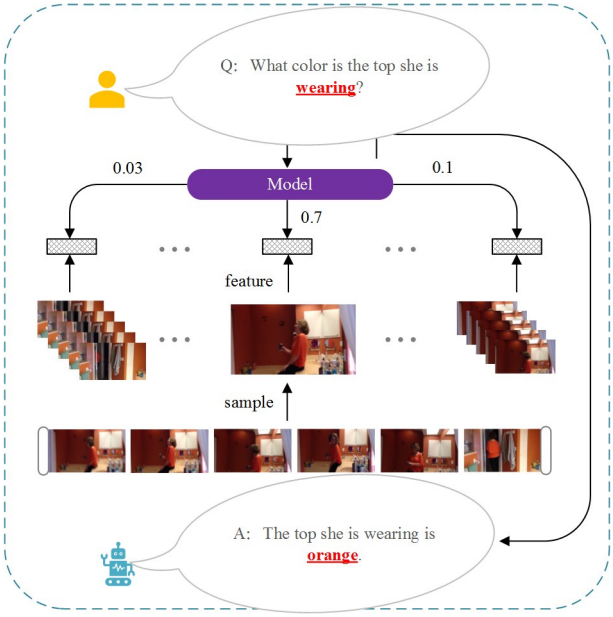

图 3　基于层次化视觉注意力的视频特征提取实例图

Fig. 3　Video feature extraction diagram based on hierarchical visual attention

视频帧指视频中的静态图像，它是由一系列在不同时间捕捉到的连续图像帧组成的。在每个图像帧中，对象的形状、纹理、尺度等信息主要由视频的画面特征来体现。因此，通过从视频帧中提取画面特征可以完成部分相关任务。此外，运动特征是视频中包含的另一类重要信息，它可以提供对象随时间、空间变化的运动轨迹和运动模式，是产生问题回复的重要信息源。

在本文中，对于给定视频，模型首先提取视频的画面特征和运动特征，然后对用户所提出的问题逐个单词进行分析，并在每个时间步中通过注意力记忆单元对这些特征进行精细化调整。在处理完问题的最后一个单词之后，模型生成最相关和最具价值的视觉注意力以回答具体问题并给出答案。同时，模型还会关注历史对话上下文并将其作为对话生成的参考，以充分融合画面、运动特征，获得视频表示。下文将详细阐述本文方法的多模态特征提取和注意力记忆单元部分。

3.1　多模态特征提取

本节提出了在帧级别和片段级别分别提取多模态特征的方法，即从包含 16 个连续帧的视频片段中提取运动特征，以获得视频的一系列向量表示。由于画面特征和运动特征在视频中具有较强的通用性，且足以解释此模型，因此本文主要考虑这两个特征。下面将从画面特征、运动特征和所提问题这 3 个角度来介绍多模态特征提取的详细过程。

(1)视频画面特征

本文使用 VGG 网络[18]作为帧级画面特征提取器。首先，将视频中的图像帧作为输入，经过 VGG 网络的卷积层和池化层，得到图像的画面特征表示。VGG16 的网络结构如图 4 所示，其中卷积层可以提取图像的局部特征，池化层可以对特征进行降维和平移不变性处理。对于给定视频，其画面特征可表示为 $F_a=[f_1^a,f_2^a,\cdots,f_N^a]$，其中，$N$ 表示视频帧的采样数量，上标 a 表示画面特征。

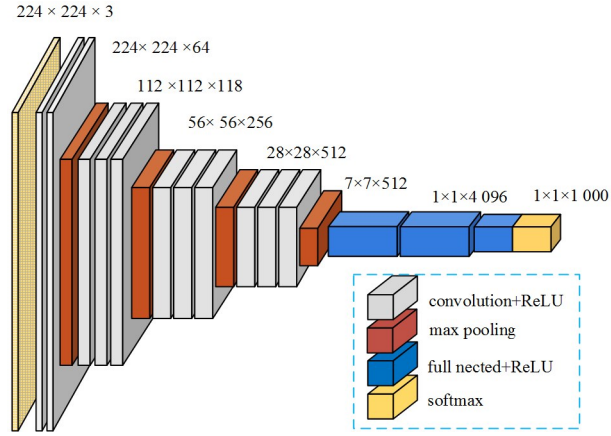

图 4　VGG16 的网络结构示意图

Fig. 4　Schematic diagram of network structure of VGG16

(2)视频运动特征

I3D 网络具有较强的视频动态信息捕获能力，因此，本文采用 I3D 网络[19]作为片段级运动特征的提取器。与 VGG 网络的处理过程类似，首先将视频片段中的视频帧作为输入，经过 I3D 网络的卷积层和池化层，得到视频片段的特征表示。与 2D 卷积神经网络不同，I3D 网络的卷积核在时间维度上也进行了卷积计算，通过将不同时间点的视频片段特征融合，能够更全面地捕捉到视频片段中的运动特征信息。对于给定的视频片段，从中提取的运动特征可表示为 $F_m=[f_1^m,f_2^m,\cdots,f_N^m]$，其中 N 表示视频帧的采样数量，上标 m 表示运动特征。

(3)用户问题理解

在问题理解方面，由于 Word2Vec 模型[20]能够捕获单词的上下文语境信息和单词之间的语义关系，且具有较高的计算效率，因此，本文采用 Word2Vec 模型作为词嵌入方法。用户问题可以表示为一系列单词，即 $Q=[q_1,q_2,\cdots,q_T]$，逐一顺序提取单词 q_t 并将其转换为语义嵌入向量 $\boldsymbol{x}_t$，从而获取整句问题的语义信息。

经过上述视频及问题特征提取之后，本文采用了一种新型层次化视觉注意力机制，来产生基于所提问题的富语义视频特征注意力。具体框架如图 5 所示，其中语义嵌入 $\boldsymbol{x}_t$ 作为 $LSTM_q$ 的输入信息，$LSTM_q$ 的隐藏层状态 h_t^q 为针对问题语义特征的记忆信息。

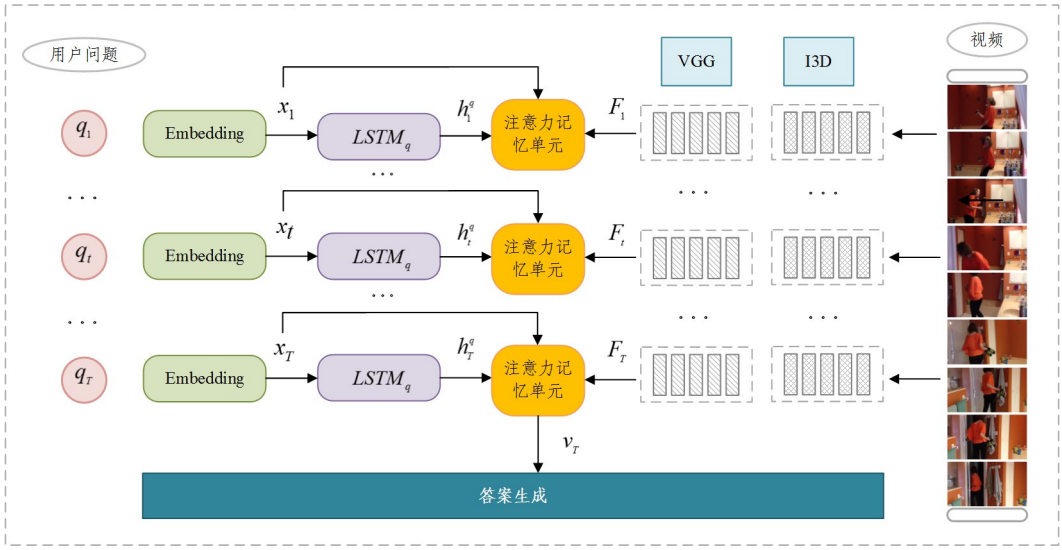

图 5 视频特征提取机制框架图

Fig. 5 Framework of video feature extraction mechanism

3.2 注意力记忆单元

注意力记忆单元主要包括 4 个操作模块：注意力模块(ATT)、通道融合模块(CF)、记忆模块($LSTM_a$)和优化模块(REF)。在执行的过程中，将当前的词嵌入、问题信息以及视频特征作为输入，以实现逐步细化视频特征注意力的目标。图 6 给出了 AMU 中各操作块的执行过程示意图。

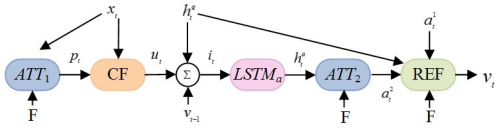

图 6 注意力记忆单元内部操作块执行过程示意图

Fig. 6 Schematic diagram of internal operation block execution process of attention and memory unit

首先，ATT_1 基于当前词嵌入 x_t 对 F 执行注意力机制进行关注。然后，将视频画面特征 p_t^a 与运动特征 p_t^m 取加权和得到 p_t，再与 $\boldsymbol{x}_t$ 相融合，以提取 F 中与当前词相关联的视频特征，从而为每个通道分配权重得分，并得到视频的中间融合表示 u_t。此时，将之前得到的 $LSTM_q$ 的隐藏层状态 h_t^q、上一个视频表示 v_{t-1} 和中间视频表示 u_t 三者相加，作为 $LSTM_a$ 的输入，从而获取并记忆全部关联用户问题的视频特征信息。之后，ATT_2 使用 h_t^a 再次对 F 执行注意力机制，在 REF 中细化注意力权重 a_t^1 和 a_t^2，生成的视频表示 v_t 可以在下一个时间步长中被继续使用。下面将分别详细介绍每个操作模块。

(1)注意力模块(ATT)

针对一个与视频内容相关的问题，其对话生成所参考的关键信息通常只对应于视频中部分帧或片段所包含的特征。注意力模块(ATT)旨在分别给视频的静态特征和动态特征分配权重，并根据两者的权重组合关注最有用的信息。在 AMU 中，有两个注意力模块 ATT_1 和 ATT_2。接下来以 ATT_1 为例来解释 ATT 的操作流程。由图 5 所示，ATT_1 在每个特征通道上利用词嵌入 $\boldsymbol{x}_t$ 在视频特征 F 上执行注意力机制，其注意力模块的计算过程表述如下：

$$e_t=\tanh(\boldsymbol{W}_f f_i+b_f)^{\mathrm{T}}\tanh(\boldsymbol{W}_x\boldsymbol{x}_t+b_x) \tag{1}$$

$$a_i=\frac{\exp(e_i)}{\sum_{i=1}^{N}\exp(e_i)} \tag{2}$$

其中，权重 a_i 反映当前单词与第 i 个视频特征间的相关性；$\boldsymbol{W}_f$ 和 $\boldsymbol{W}_x$ 用于将词嵌入和视频特征转换到相同的底层嵌入空间；融合特征 p_t 表示当前单词所对应的视频特征，结合注意力权重 a_i，其计算过程如下：

$$p_t=\sum_{i=1}^{N}a_i\tanh(\boldsymbol{W}_f f_i+b_f) \tag{3}$$

ATT_1 模块的应用，增强了当前单词在生成回答时的影响力。

(2)通道融合模块(CF)

特征 p_t 可以看作是由特征 p_t^a 和 p_t^m 组成的。由于不同单词针对视频画面、运动特征的强度不同，因此，基于当前单词，模型对两个特征通道分别进行强度评分并通过 CF 模块融合得到中间视频表征 u_t，融合过程表示如下：

$$s_t^a,s_t^m=\mathrm{softmax}(\boldsymbol{W}_m\boldsymbol{x}_t+b_m) \tag{4}$$

$$u_t=s_t^a p_t^a+s_t^m p_t^m \tag{5}$$

其中，s_t^a 和 s_t^m 分别代表不同画面和不同运动特征的强度评分。CF 会基于当前单词的强度评分分别从视频的画面、运动通道中提取信息，融合后得到中间视频表征 u_t。

(3)记忆模块($LSTM_a$)

在每个时间步内，模型同时处理问题的一个单词和两个注意力操作 ATT_1 和 ATT_2。首先，通过 ATT_1 获得中间视频表征 u_t，结合上一时间步得到的隐藏层状态 v_{t-1} 作为 $LSTM_a$ 的输入，输出的 h_t^a 用于执行注意力操作 ATT_2。

(4)优化模块(REF)

在执行完 ATT_2 之后，模型在 F 处计算生成注意力权重 a_t^2。a_t^1 和 a_t^2 均用于注意力机制的优化。REF 的内部工作机制可以用以下公式表示：

$$a_t=(a_t^1+a_t^2)/2 \tag{6}$$

$$g_t=\sum_{i=1}^{N}a_t^i\tanh(\boldsymbol{W}_f f_i+b_f) \tag{7}$$

$$v_t=CF(h_t^g,g_t) \tag{8}$$

其中，g_t 包含优化后的画面、运动特征 g_t^a 和 g_t^m，v_t 是时间步长 t

处视频的最终融合表征。

4　多任务学习富语义视频对话生成

4.1　视频对话模型架构设计

本文提出的视频对话生成模型是一个基于 GPT2 架构[21]的多层 Transformer 编码器，如图 7 所示。

针对用户问题的文本特征信息，模型遵循 GPT2 的处理方法，将其标记为文本块。视频、音频类型的数据也采用类似的处理方法。对于给定的视频V_k，首先将V_k分割成T_k个视频片段，每个视频片段的滑动窗口中包含 N 个视频帧 f。

对于每一个视频片段$S_t=\{s_1,s_2,\cdots,s_t\}$，采用 I3D-rgb 和 I3D-flow 模型分别提取其多维视频特征信息V_{rgb}和V_{flow}。考虑到音频与对应视频的同步性，本文使用 VGG 模型提取来自同一段视频的多维音频特征A_{vgg}，如图 8 所示。

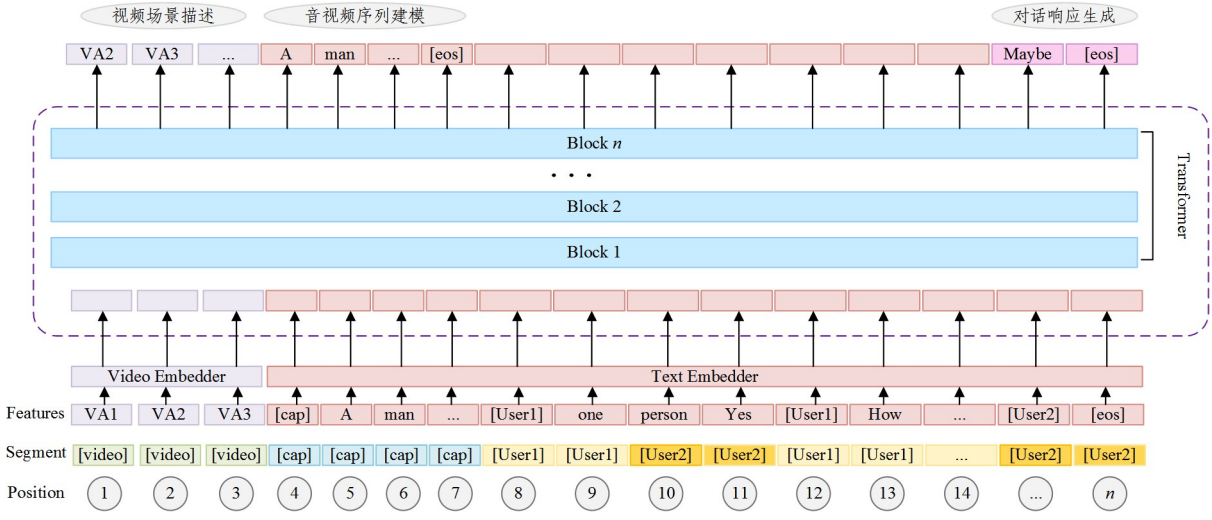

图 7　视频对话生成模型架构示意图

Fig. 7　Schematic diagram of architecture of video dialogue generation model

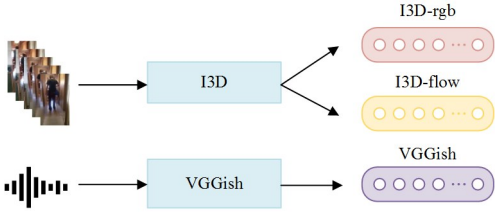

图 8　多模态视频特征提取示意图

Fig. 8　Schematic diagram of multimodal video feature extraction

将上述处理得到的V_{rgb}，V_{flow}和A_{vgg}进行特征拼接，即可得到视频-音频特征VA_t，拼接后的VA_t的维度为 $2d_v+d_a$。

$$VA_t=[V_{\text{rgb}},V_{\text{flow}},A_{\text{vgg}}] \tag{9}$$

将所得到的视频-音频特征 VA 输入全连接层，并映射到与文本特征维度相同的词嵌入空间。最终，每个单词标记都以其词嵌入、位置编码和片段嵌入三者相加来表示，使得模型能够有效区分视频信息(Video)、视频描述信息(Video Captioning)和用户信息(User)等不同的输入。

4.2　多任务学习的实现

为进一步适应多样化的应用场景，本文利用多任务学习方法对模型进行调整。多任务学习指在一个模型中同时学习多个相关任务的过程，其核心思想在于让模型共享不同任务之间的底层表示，以提高模型的泛化能力和学习效率[22]。图 9 为多任务学习模型示意图。

本文的多任务学习方法主要包括 3 个任务，分别为：视频场景描述、音视频序列建模和对话响应生成。其目的在于提高模型在语音与视频处理方面的能力，进一步拓宽其潜在应用场景。

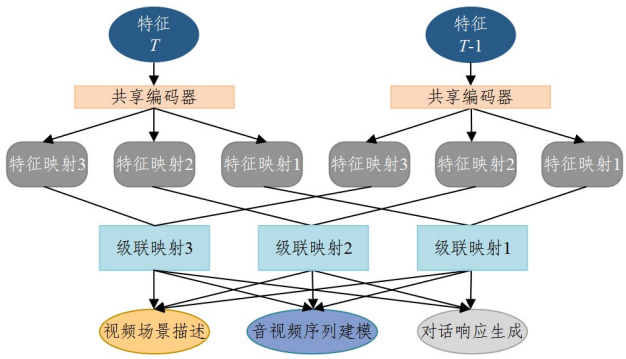

图 9　多任务学习模型示意图

Fig. 9　Schematic diagram of multi-task learning model

(1)视频场景描述

该模块主要利用给定视频-音频特征 VA 生成的视频描述$C=\{c_1,c_2,\cdots,c_l\}$以训练模型。

$$L(\theta)=-E_{(VA,C)\sim D} \tag{10}$$

$$\log\prod_{i=0}^{l}P(c_i\,|\,VA,c_{<i}) \tag{11}$$

其中，θ是(VA,C,U,Q)对整个训练集 D 采样所得。

(2)音视频序列建模

该模块主要根据给定的视频描述信息和历史对话上下文信息来预测视频-音频特征。首先，采用视频-音频特征回归方法，将视频-音频特征对应 Transformer 层的输出o_t回归到下一个视频-音频特征VA_{t+1}中。然后，使用全连接层将输出转换为与VA_{t+1}同维度的向量$\boldsymbol{g}_\theta(o_t)$。可利用最小化 L2 损失函数进行训练。

$$L(\theta)=-E_{(VA,C,U)\sim D} \tag{12}$$

$$\frac{1}{T}\sum_{t=1}^{T}\| g_\theta(o_t)-VA_{(t+1)}\|_2^2 \tag{13}$$

其中，$o_t=f_\theta(VA_{t+1},C,U)$，$f_\theta$代表此模型。

(3)对话响应生成

该任务的目标是基于所获得的视频-音频特征 VA、视频描述 C、历史对话上下文 H 和所提问题 Q 生成对话响应$R_n=\{r_{n1},r_{n2},\cdots,r_{nm}\}$。

5 实验验证

5.1 实验数据

为了更好地探究视听场景理解与时空推理的问题，本文采用数据集 DSTC7 AVSD。相较于 AVSD 数据集，该数据集具有数据规模巨大、注重常识性问题、涵盖对话内容广泛且质量较高等特点[23]。表 1 列出了 DSTC7 AVSD 数据集的相关统计数据。

表 1 DSTC7 AVSD 数据集的统计数据

Table 1 Statistics of DSTC7 AVSD dataset

	训练集	验证集	测试集
对话数	7659	1787	1710
对话轮数	153180	35740	13490
总单词数	1450754	339006	110252

该数据集具备多轮次、多用户、多层次文本信息和多样对话风格等特征，每个视频场景都包含多轮对话，涉及多个说话者，同时包含视频描述、问题、回答等多个层次的信息，话题范围广泛。此外，该数据集中的视频和音频数据均采集于真实场景，涵盖多种自然环境下的噪声，使得模型具备更强的泛化能力。

5.2 实验设置

首先，基于 GPT2 模型初始化本模型的预训练权重。在预训练的过程中，采用了三轮历史对话信息，模型隐藏层的大小设置为 768，batch 大小设置为 4，使用了学习率为 6.25×10^{-5}的 Adam 优化器。在解码的过程中，主要使用了波束搜索算法(Beam Search)，其波束大小设置为 5，最大长度设置为 20，长度惩罚项设置为 0.3。

5.3 对比方法分析

本节主要选取了 Naive Fusion，MTN，BiST 和 STSGR 这 4 种相关的方法与本模型的实验结果进行比较。

(1)Naive Fusion[24]：一种用于多轮对话系统的简单模型。其主要将各模态信息与投影矩阵相结合，将历史对话文本串联并输入一个单一的神经网络模型中进行处理，以实现对话的生成。

(2)MTN[25]：一种基于多头注意力机制的神经网络模型。其可以在多模态设置下生成良好的对话响应，主要包括编码器、解码器和自动编码器层。

(3)BiST[26]：一种基于文本线索的视觉语言神经框架，用于视频高分辨率查询。其通过时间与空间的双向推理，学习两个特征空间之间的动态信息扩散。

(4)STSGR[27]：一个基于语义控制的多模态置换推理框架，由一系列 Transformer 模块组成。每个模块接受一种模态作为输入，并根据输入的问题生成表征。

5.4 实验结果

针对视频对话生成任务，常用的评估指标主要包括人工评估和自动评估两种。由于人工评估需要耗费大量的时间和成本，同时存在主观性和不可重复性等问题，因此本文采用自动评估指标度量最终生成的对话响应的质量。常用的自动评估指标主要包括 BLEU，ROUGE，METEOR 和 CIDEr 等。

(1)BLEU[28]：一种基于准确率的相似度量指标。其主要衡量生成文本与参考文本之间的 n-gram 重叠情况，根据单个词到多个词组的匹配程度，给出一个介于 0～1 的分数。

(2)ROUGE[29]：一种基于召回率的相似度量指标，其衡量生成文本中出现的参考文本词组数。与 BLEU 的计算方式相似，ROUGE 主要分为 ROUGE-N，ROUGE-L 和 ROUGE-W 这 3 种，可以针对不同评估需求进行度量。

(3)METEOR[30]：一种综合考虑准确率和召回率的相似度量指标。其不仅衡量词序的重叠情况，同时关注同义词、词形变化、单复数等语义和语法信息的匹配度，提供了更加全面的评估。

(4)CIDEr[31]：一种用于评估图像描述生成质量的指标，侧重于考虑描述内容的多样性和详细程度，通过分析生成内容是否包含视觉关键信息及语言丰富性来评估其质量。

利用上述评估指标将本文模型与各基线方法进行对比，最终得到的实验结果如表 2 所列。

表 2 实验结果

Table 2 Experimental results

指标	NF	MTN	BiST	STSGR	Ours
BLEU-1	—	0.731	**0.753**	—	0.660
BLEU-2	—	0.587	0.518	—	**0.590**
BLEU-3	—	0.493	0.510	—	**0.580**
BLEU-4	0.310	0.390	0.329	0.130	**0.370**
ROUGE	0.451	0.569	**0.581**	0.362	0.576
METEOR	0.215	0.269	0.284	0.165	**0.350**
CIDEr	1.127	1.129	1.192	1.272	**1.275**

实验结果显示，与此前该任务的 SOTA 模型 BiST 相比，本文设计的模型实现了极大的提升，其中 METEOR 指标提升最为显著，从 0.284 提升到了 0.350，提高了 23.24%。相较于纯文本任务 NF 和 STSGR，本文模型的性能增益更加明显，7 项指标均高于 NF 和 STSGR。这表明该层次化视觉注意力机制的视频理解方法能够有效地应用于此类任务，生成的文本具有较高的准确性和多样性。与 MTN 相比，本文模型的 BLEU-4 指标略低，但在综合性能表现上取得了较大的改进。

5.5 消融实验

为验证本文模型在组成模块及实验设置方面的有效性和科学性，进行了以下两组消融实验：

(1)使用贪心搜索/核采样/波束搜索解码器

为寻求一种最为有效的多模态对话生成解码方法，本文尝试了贪婪搜索、波束搜索和核采样 3 种典型解码方法，实验结果如表 3 所列。

表 3　3 种解码器的评估指标结果对比

Table 3　Comparison of evaluation index results of three decoders

指标	贪心搜索	波束搜索	核采样
BLEU-1	0.660	0.660	**0.680**
BLEU-2	0.580	**0.590**	0.520
BLEU-3	0.470	**0.580**	0.410
BLEU-4	0.350	**0.370**	0.320
ROUGE	0.488	**0.576**	0.453
METEOR	0.340	**0.350**	0.250
CIDEr	1.191	1.275	**1.127**

可以发现，在这 3 种解码方法中，波束搜索算法的表现最佳。这一结果表明，波束搜索算法在生成文本时具有更好的语法正确性、流畅性和相关性，并且在生成对话响应的结果中更注重长文本序列的生成质量，因此在 BLEU-3 和 BLEU-4 等与长文本序列相关的评估指标上表现更好。

相比其他评估指标，BLEU-1 更注重单个词的选择，却忽略了上下文信息的影响。因此，本文模型的在 BLEU-1 上的表现略差。这意味着该算法在生成文本时更注重上下文信息，而不是仅仅依赖于对逐个单词的匹配程度，因此，本文算法对于对话响应中单个词的选择和运用会更加灵活。

本文算法在 METEOR 指标上表现良好，表明其能够有效地处理单词形态变化和同义词的问题，从而提高了对话生成的整体质量。

总的来说，使用波束搜索算法的优点在于其能够生成更加流畅和准确的长序列对话结果，并且能够处理一些单词形态变化和同义词的问题。因此，本文使用波束搜索算法作为解码算法，获得了最佳的性能表现。

(2)使用不同的历史对话回合长度

本实验通过设置视频回归损失来探究不同轮数的历史对话信息对模型评估结果的影响。在对话生成系统中，不同的对话轮数对系统性能的影响差异较大。较少的对话轮数会导致系统无法捕捉到充足的上下文信息，而较多的对话轮数可能会使系统过度依赖历史对话，导致回答不准确或不相关。通过设置不同的对话轮数，可以确定对话轮数对系统性能的影响，从而确定适当的对话轮数范围，并选择最佳的对话轮数设置。

表 4 所列为 DSTC 7-AVSD 测试集的客观评估结果。其中，最大历史对话回合长度选取的范围为 3～8，每个指标的最佳结果以粗体突出显示。可以看出，当历史对话的回合长度参数设置为 5 时，各项评估指标的结果表现整体最优，模型获得最佳性能。

表 4　不同历史对话回合长度下的评估结果对比

Table 4　Comparison of evaluation results with different historical dialogue session lengths

指标	3	4	5	6	7	8
BLEU-1	0.64	0.660	**0.660**	0.660	0.660	0.650
BLEU-2	0.550	0.590	**0.590**	0.580	0.590	0.560
BLEU-3	0.430	0.480	**0.580**	0.480	0.500	0.460
BLEU-4	0.300	0.360	**0.370**	0.350	0.370	0.320
ROUGE	0.475	0.559	**0.576**	0.562	0.562	0.420
METEOR	0.307	0.348	**0.350**	0.349	0.349	0.309
CIDEr	1.156	1.270	1.274	1.274	**1.275**	1.169

结束语　面向多模态人机交互领域，本文提出了基于层次化视觉注意力的富语义视频对话生成算法，通过将视觉和语言的注意力机制引入到对话生成过程中，实现了对视频和语言的联合建模。针对用户所提问题，对视频的不同区域进行不同程度的局部注意力分配。同时，结合历史对话上下文信息，可以生成语义更加丰富、更具连贯性的对话内容。此外，本文还进行了大量的实验验证和实验结果分析工作，对算法的实际性能和生成文本的质量进行了有效评估。

经过上述的工作积淀，该算法目前已经取得了可观的评估结果。但是在研究过程的中仍然存在不足之处，例如数据集适用场景有限，只针对特定的视频场景和语境，缺乏对更广泛对话场景的研究等，后续需要进行进一步优化和改进。

参 考 文 献

[1] XU W, DAINOFF M J, GE L, et al. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI[J]. International Journal of Human-Computer Interaction, 2023, 39(3): 494-518.

[2] YUSUF A A, FENG C, MAO X L. An analysis of graph convolutional networks and recent datasets for visual question answering[J]. Artificial Intelligence Review, 2022, 55(8): 6277-6300.

[3] LIN X, BERTASIUS G, WANG J, et al. Vx2text: End-to-end learning of video-based text generation from multimodal inputs[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE Press, 2021: 7005-7015.

[4] WANG H Y, HUANG J Y, LEE W P. Integrating Scene Image and Conversational Text to Develop Human-Machine Dialogue[J]. International Journal of Semantic Computing, 2022, 16(3): 425-447.

[5] SERBAN I V, SORDONI A, LOWE R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 3295-3301.

[6] SERBAN I V, SORDONI A, BENGIO Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016: 3776-3783.

[7] WESTON J. Dialog-based language learning[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: AAAI Press, 2016: 829-837.

[8] XING C, WU Y, WU W, et al. Hierarchical recurrent attention network for response generation[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 5610-5617.

[9] ZHOU G, LUO P, CAO R, et al. Mechanism-aware neural machine for dialogue response generation[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 3400-3406.

[10] WU Y, WU W, XING C, et al. Sequential Matching Network: A

New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: Association for Computational Linguistics, 2017: 496-505.
[11] LIU X, ZHENG Y, DU Z, et al. GPT understands, too[J]. arXiv: 2103.10385, 2023.
[12] NAZIR A, WANG Z. A Comprehensive Survey of ChatGPT: Advancements, Applications, Prospects, and Challenges [J]. Metaradiology, 2023, 1(4): 100022.
[13] WU T, HE S, LIU J, et al. A brief overview of ChatGPT: The history, status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122-1136.
[14] YU Y, KO H, CHOI J, et al. End-to-end concept word detection for video captioning, retrieval, and question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 3165-3173.
[15] JANG Y, SONG Y, YU Y, et al. Tgif-qa: Toward spatio-temporal reasoning in visual question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 2758-2766.
[16] GARCIA N, NAKASHIMA Y. Knowledge-based video question answering with unsupervised scene descriptions[C]// European Conference on Computer Vision. Glasgow: Springer, 2020: 581-598.
[17] LE H, CHEN N, HOI S. Vgnmn: Video-grounded neural module networks for video-grounded dialogue systems[C]// Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022: 3377-3393.
[18] HAQUE M F, LIM H Y, KANG D S. Object detection based on VGG with ResNet network[C]// International Conference on Electronics, Information, and Communication (ICEIC). Auckland: IEEE Press, 2019: 1-3.
[19] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE Press, 2017: 6299-6308.
[20] CHURCH K W. Word2Vec[J]. Natural Language Engineering, 2017, 23(1): 155-162.
[21] LAGLER K, SCHINDELEGGER M, BÖHM J, et al. GPT2: Empirical slant delay model for radio space geodetic techniques [J]. Geophysical Research Letters, 2013, 40(6): 1069-1073.
[22] BHATTACHARJEE D, ZHANG T, SÜSSTRUNK S, et al. Mult: an end-to-end multitask learning transformer [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE Press, 2022: 12031-12041.
[23] YE M, YOU Q, MA F. QUALIFIER: Question-Guided Self-Attentive Multimodal Fusion Network for Audio Visual Scene-Aware Dialog[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visison. Waikoloa: IEEE Press, 2022: 248-256.
[24] HORI C, ALAMRI H, WANG J, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features[C]// ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE Press, 2019: 2352-2356.
[25] LE H, SAHOO D, CHEN N F, et al. Multimodal transformer networks for end-to-end video-grounded dialogue systems[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019: 5612-5623.
[26] LE H, SAHOO D, CHEN N, et al. BiST: Bi-directional Spatio-Temporal Reasoning for Video-Grounded Dialogues[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 1846-1859.
[27] GENG S, GAO P, CHATTERJEE M, et al. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers[C]// Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI Press, 2021: 1415-1423.
[28] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics, 2002: 311-318.
[29] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Proceedings of theWorkshop on Text Summarization Branches Out. Barcelona: Springer, 2004: 74-81.
[30] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]// Proceedings of the acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor: Association for Computational Linguistics, 2005: 65-72.
[31] VEDANTAM R, LAWRENCE ZITNICK C, PARIKH D. Cider: Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE Press, 2015: 4566-4575.

ZHAO Qian, born in 2001, postgraduate, is a member of CCF(No. P2226G). Her main research interest is visual human-computer dialogue.

GUO Bin, born in 1980, Ph. D, professor, doctoral supervisor. His main research interests include ubiquitous computing, mobile crowd sensing, big data intelligence and so on.

(责任编辑:何杨)