基于层次化一致性语义学习的多模态意图识别

彭俊杰[1],李铮一[1],张换香[1,2],王 兰[1]
(1. 上海大学计算机工程与科学学院,上海 200444;2. 内蒙古科技大学创新创业教育学院,内蒙古包头 014010)

摘 要: 多模态意图识别(Multimodal Intent Recognition,MIR)是在现实世界中理解人类意图的重要研究方向,旨在通过融合语言、视觉和音频等多种模态信息来准确判断说话人的意图. 然而,现有的MIR研究大多集中在如何为文本模态构建多模态语义环境,对视觉和音频模态中蕴含的大量语义信息(如动作和情感语义)的利用则不够深入. 尽管视觉和音频模态富含与意图相关的信息,但其固有的冗余信息和噪声却制约了模型对这些模态特征的有效利用. 为解决上述问题,本文提出了一种能够有效利用音频模态语义关系,同时有效抑制冗余信息的MIR模型. 该模型通过构建抑制冗余信息的初级语义特征,引导学习不同尺度的模态内与模态间语义关联,以理解说话人的意图. 在此基础之上,模型利用不同模态特征间潜在的意图一致性,将提取到的音视频语义特征与具有明确意图语义的文本特征进行配对,从而过滤掉那些单独通过意图识别任务无法消除的无关语义信息. 此外,模型采用多模态融合门控机制,整合来自不同模态的意图语义. 在多个意图理解任务的数据集上的实验表明:所提出的方法能够有效提取音视频模态语义并滤除意图识别无关语义,且在性能上优于现有的MIR方法. 具体而言,在准确率(ACCuracy,ACC)值、精确度(Precision,P)值、召回率(Recall,R)值和F_1值(F_1 score,F_1)上均取得了0.7~1.8个百分点的提升.

关键词: 意图识别;多模态融合;多模态语义学习;多任务学习;跨模态注意力
基金项目: 上海市服务业发展引导资金项目(No.06162021592)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)06-2007-15
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250009

Multimodal Intent Recognition Based on Hierarchical Semantic-Consistency Learning

PENG Jun-jie[1], LI Zheng-yi[1], ZHANG Huan-xiang[1,2], WANG Lan[1]
(1. *School of Computer Engineering and Science*, *Shanghai University*, *Shanghai* 200444, *China*;
2. *School of Innovation and Entrepreneurship Education*, *Inner Mongolia University of Science and Technology*, *Baotou*, *Inner Mongolia* 014010, *China*)

Abstract: Multimodal intent recognition (MIR) is a critical research for understanding human intent in the real world. It aims to judge the speaker's intent through multiple modalities including language, visual and audio modalities. However, existing studies in MIR primarily focus on constructing multimodal semantic environments for textual data, while the utilization of rich semantic information in visual and audio modalities, such as action and emotional semantics, remains insufficiently explored. Despite the visual and audio modalities carrying intents-related semantics, their inherent redundant information and noise hinder the effective use of these modalities. To address these challenges, this paper proposes a more effective MIR model that better leverages audio and visual information while suppressing redundant information. The proposed model understands the speaker's intent by constructing primary semantic features that suppress redundant information and guiding the learning of intra-modality and inter-modality semantic associations at different scales. Based on this, the model leverages the potential intent consistency across different modalities and pair audio and visual representations with textual features, which contain more explicit intent-related semantics, to filter out irrelevant semantics that cannot be eliminated by intent recognition tasks. Furthermore, the model uses multi-modal fusion gating mechanism to integrate intent semantics from different modalities. Experiments on several datasets of intents understanding tasks show that the proposed method can effectively extract the modal semantics of audio and video and filter out the irrelevant semantics of intent recog-

收稿日期:2025-01-02;修回日期:2025-05-21;责任编辑:覃怀银

nition, and outperforms the existing MIR methods, achieving 0.7 to 1.8 percentage points improvement in accuracy (ACC), precision (P), recall (R) and F_1 score (F_1).

Key words: intent recognition; multimodal fusion; multimodal semantic learning; multi-task learning; cross-modal attention

Foundation Item(s): Shanghai Service Industry Development Fund (No.06162021592)

1 引言

意图识别的目标是正确判断说话人的意图,从而为人机交互提供高质量的服务. 此前的研究[1~3]探讨并确认了文本在意图识别任务中的重要作用. 然而,在复杂的现实场景中,仅依赖文本往往难以正确判断说话人的意图[4],因为意图不仅通过文本信息表达,还通过说话人的表情、语调以及动作等方式传递. 通过融合音频、视频和文本等多模态信息进行意图识别,能够显著提升在复杂场景中的识别准确度. 因此,多模态意图识别(Multimodal Intent Recognition,MIR)逐渐成为研究者关注的重点. 为了有效地从多模态信息中识别人类意图,已有不少研究取得重要进展. 例如,文献[5]在多模态情感对话数据集(Multimodal Emotion Lines Dataset,MELD)的基础上,以人工标注对话行为(dialogue action),为使用文本、音频和图像数据的MIR数据集奠定基础. 文献[6]则将说话人的目标和行动作为意图标注,同时将对说话人情感的感知纳入意图的范畴,推出一个具有细粒度意图分类的MIR基准数据集MIntRec(Multimodal Intent Recognition corpus),并在后续的研究[7]中推出了规模更大、意图分类更丰富的MIntRec2.0数据集. 这些多模态意图数据集为研究者提供了重要的资源,推动MIR领域的进一步发展. 在这些新的数据集上,许多研究者也从热门的研究方法和框架的角度开展了大量研究,以提升MIR的效果. 文献[8]提出通过构建不同模态的共享特征和私有特征来学习更加丰富的多模态表征;文献[9]提出一种模态感知的提示词创建方法,旨在增强文本模态在多模态语境中的表达能力;文献[10]开发了一种跨视频库检索相似视频的技术,通过建立更丰富的全局上下文环境,用于减少因单个视频信息不足而导致的意图偏差. 这些创新性的工作不断推动着MIR技术的发展.

尽管现有研究在MIR方面取得了显著进展,但仍存在一些亟待解决的问题. 现有方法在利用音视频模态信息增强文本模态上已有较为深入的探讨,但在音视频模态特征的语义关系研究上仍不充分. 无法很好地捕捉音视频模态局部特征之间的相互关系,导致难以正确地理解说话人的意图[11,12]. 例如,如图1所示,文本句子"Guys, I'm gonna be a little busy around 2:00 p.m. Don't even ask me what I'm doing. It's private."乍一看可能会让人误以为说话人在告知对方勿要打扰,而仅仅依靠视觉上的局部信息也难以作出准确的判断. 然而,通过捕捉说话人的表情、手势等局部视觉信息在时间序列上的关联,便能更为精确地理解其实际意图——她其实是在炫耀. 不过,在音视频模态中,除了与意图相关的语义信息之外,还存在着大量的冗余信息和噪声. 以图1为例,与说话人意图相关的视觉信息可能会被其他无关画面所干扰. 虽然已经有一些方法尝试去除噪声和冗余信息[13,14],但利用门控机制或瓶颈模块来过滤噪声容易造成关键信息的丢失且不够灵活. 利用额外的标记学习跨模态语义关系,从而减少噪声和冗余信息的干扰,是一个尚未被充分探索的方向.

图1 通过多模态信息理解说话人意图的挑战示例

为解决上述问题,本文提出了一种基于层次化一致性语义学习的MIR模型. 该模型能够有效捕捉多模态语义关系,同时降低噪声和冗余信息的干扰,进而提升意图识别的准确性. 具体而言,模型首先将不同模态的特征转化为相同序列长度和相同特征维度的表示形式,从而实现不同模态信息的对齐. 其次,在层次化语义学习模块的第一层中,模型结合可学习标记与原始序列特征,将模态的基本信息转移到初级语义特征中,以平衡抑制噪声与保留关键信息. 在此基础上,该模块通过跨模态交互进一步增强不同尺度特征的语义学习能力. 然而,音视频模态中的无关语义仅依靠意图识别任务仍难以完全滤除. 为此,模型利用模态间的意图一致性,以意图语义更明确的文本模态为锚点,将文本模态特征分别与提取的音视频语义进行配对. 这一策略为层次化语义学习提供了额外的监督信号,从而进一步提高意图识别的准确性.

本文的主要贡献可以概括如下:

(1)提出了一种基于层次化一致性语义学习的MIR模型. 该模型以抑制噪声后的初始化特征为基础,针对音视频模态提取不同尺度的语义特征. 同时,利用模态语义一致性进一步提升模型的性能和意图识别的

准确性.

(2)设计了一个层次化语义学习模块,结合初始化标记和原始序列特征构建层次化语义特征,有效平衡了模态特征提取与冗余信息过滤. 以意图语义更明确的文本模态为锚点,通过文本-音视频特征配对任务过滤意图无关语义.

(3)在多个与意图相关的多模态任务数据集上进行了大量实验. 实验结果表明:提出的模型在多项指标上均超越了现有的意图识别模型.

2　相关工作

2.1　意图识别

意图识别是人机交互系统中一个重要组成部分. 传统的意图识别依赖人工制定规则,在复杂的语言环境中缺乏灵活性. 随着深度学习在自然语言领域的成功应用,研究者们开始在深度学习方法上探究意图识别. 基于卷积神经网络(Convolutional Neural Network, CNN)[15]、循环神经网络(Recurrent Neural Network, RNN)[16]和大规模语言预训练模型[17,18]的意图识别方法相继出现. 这些方法在基于文本的意图识别任务中表现出令人印象深刻的性能. 然而,现实世界中的人类意图通过包括情绪[19]、动作和语调[20]的多模态语言表达,基于文本模态的意图识别方法不能准确识别现实场景中的多模态意图.

最近,随着多模态语言理解逐渐得到更多的研究者关注,研究者[6]将说话人的情感评价和人类行为作为人类意图研究进行研究. 文献[21]提出利用单模态、双模态和三模态的7种情感标签,从文本、音频、图像及其组合中提取不同粒度的情感特征,捕捉多模态特征中的异质信息. 文献[22]提出一种分层跨模态交互框架,通过逐层融合单模态、双模态及三模态的语义信息,动态捕捉模态间的共享语义特征与情感可变性,旨在捕捉人类情感表达的多模态复杂特征. 文献[23]提出一种基于文本引导的交叉注意力机制,用于筛选音频和视频模态中信息质量较差的特征并进行替换,进而促进意图识别中的多模态融合. 由于情感和对话行为都是人类意图的重要组成,因此MIntRec和MIntRec2.0数据集中,提出了通过整合文本、音频和视频,识别包含行为目标和情感在内的新的MIR任务. 在构建数据集的过程中,文献[24]设计了一种质心引导的聚类机制,用于解决聚类分配不一致的问题,从而提供高质量的意图分类. 在这些新的数据集上,研究者迅速针对意图理解任务提出一系列应用多模态领域的热门技术的研究. 文献[10]通过构建相似视频库,基于局部信息和全局信息进行对比学习,从而更好地理解人类意图. 文献[25]通过在各模态特征中加入明确的模态标识,使得模型能够区分并有效融合不同模态的信息,从而增强整体理解和决策能力. 文献[26]利用大语言模型增强文本数据与其他模态配对,由浅入深地捕捉不同模态中的细节线索.

2.2　多模态融合

多模态融合技术致力于通过有效的融合过程实现高质量的多模态表示. 传统方法[27~29]依赖于张量对多模态数据的表征能力,通过如图神经网络、注意力门控等结构对多模态数据融合. 然而,这类方法在更复杂的特征融合场景中逐渐陷入瓶颈.

随着Transformer对不同领域单一模态序列建模的成功应用,越来越多的研究开始基于Transformer网络探索多模态融合的方法. 文献[30]通过多个Transformer网络进行不同模态间的翻译,应对非对齐多模态融合问题. 文献[31]设计了一种多模态自适应融合门控机制,以极小的代价使得预训练语言模型,例如,基于Transformer的双向编码器表示(Bidirectional Encoder Representations from Transformers, BERT)、XL-Net(eXtreme Language model pre-training),能够适应多模态数据. 文献[32]采用不同的Transformer网络分别学习模态共享特征和模态私有特征,从而提高多模态表征的性能. 在这些研究基础上,研究者们逐渐意识到文本模态在自然语言处理扩展的多模态任务中占有极其重要的地位. 文献[33]利用跨模态增强(Cross-modal Enhancement, CE)模块,根据音视频数据中隐含的长期情感线索来增强文本模态单词的表示,从而减少不同模态之间的分布差异. 文献[34]使用多阶段文本来引导提取图像中的关键信息特征,通过交叉模态注意力机制促进多模态特征之间的交互. 文献[35]通过基于文本的多头注意力机制,将文本信息融入学习情感相关的音视频表示学习中,从跨模态成对映射中获取有效的统一多模态表示. 文献[36]构建意图模板以增强文本表示,并利用跨模态对齐,从音视频模态中挖掘一致的隐藏意图信息.

然而,现有的大多数多模态融合方法对于音视频模态中的语义特征利用研究仍显不足,尤其是未能充分利用跨模态交互来理解音视频模态中复杂特征之间的相互关系. 考虑到音视频模态本身固有的噪声和冗余信息,直接融合反而会阻碍意图识别效果的提升. 文献[11]提出了一个门控函数,通过根据多模态特征提供模态级或融合级的实时决策,来过滤音视频模态中的噪声和冗余信息. 文献[37]则采用瓶颈机制压缩不同模态中的特征过滤噪声,并通过互信息最大化模块来调节过滤模块. 文献[12]将瓶颈机制和最优运输思想相结合,以缓解复杂场景中模型难以训练的困境. 然而,门控机制在捕捉动态变化信息方面的能力有限,而瓶颈机制则依

赖于事先定义的函数,这使其难以适应复杂的多模态数据.

针对这些问题,本文提出一种基于层次化一致性语义学习的意图识别模型. 所提出的模型在初级模态特征中加入可学习标记的音视频特征,用于学习不同尺度的语义关系,从而能够有效地捕捉音视频特征中的上下文联系. 这些额外标记学习到的语义关系相比原始音视频序列中的信息更精确,帮助模型在融合不同模态信息的过程中抑制冗余信息. 此外,模型利用文本模态作为锚点,通过意图一致性线索对文本-音频和文本-视觉进行配对,引导过滤意图无关的语义,从而提升MIR的准确率(ACCuracy,ACC)值.

3 层次化一致性语义学习

本节提出一个基于层次化一致性语义学习的MIR的模型. 如图2所示,该模型主要包括多模态特征提取、层次化音视频语义学习、多模态配对学习和多模态特征融合4个部分.

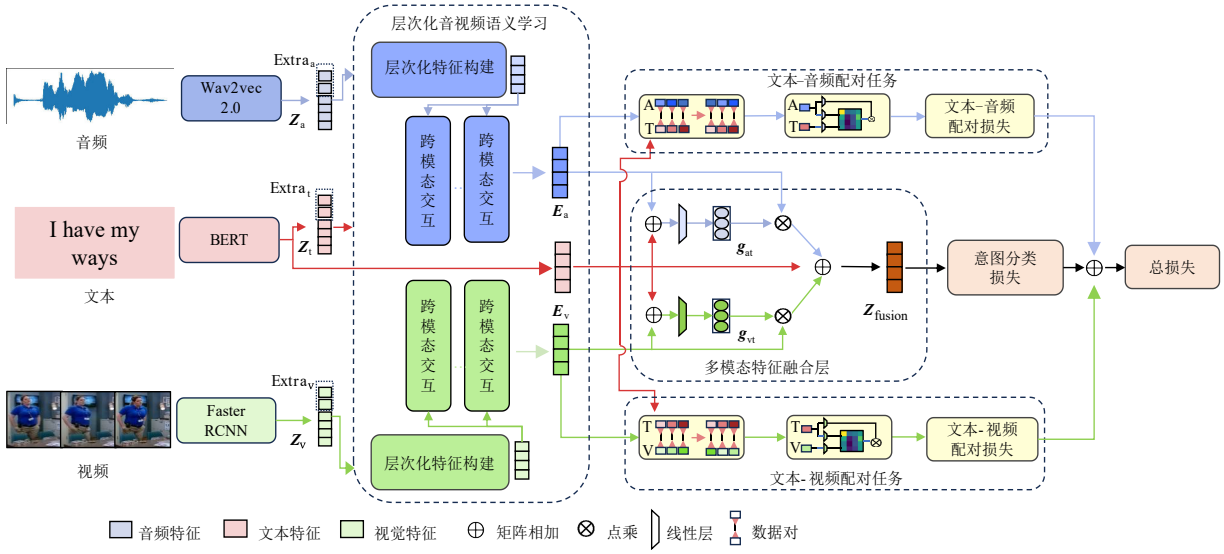


图2 模型架构

3.1 特征提取

本文遵循之前工作[6]的设定提取不同模态的特征. 对于文本语料x_t,音频片段x_a和视频片段x_v,分别使用BERT[38]、wave2vec[39]和Faster-RCNN[40]进行特征提取.

$$\boldsymbol{m}_t=\mathrm{BERT}(x_t) \tag{1}$$

$$\boldsymbol{m}_a=\mathrm{Wav2Vec}(x_a) \tag{2}$$

$$\boldsymbol{m}_v=\mathrm{AvgPool}\left(\mathrm{RoIAlign}(x_v,B)\right) \tag{3}$$

其中,RoIAlign(·)表示根据说话人标注框B提取视觉特征;AvgPool(·)用于将长度和宽度统一;$\boldsymbol{m}_t\in\mathbb{R}^{L_T\times H_T}$,$\boldsymbol{m}_a\in\mathbb{R}^{L_A\times H_A}$和$\boldsymbol{m}_v\in\mathbb{R}^{L_V\times H_V}$分别表示文本、音频和视频模态特征编码. H_T、H_A和H_V分别是特征维度;而L_T、L_A和L_V是初始特征序列长度,各不相同. 为了形成统一的多模态特征表示,引入CTC(Connectionist Temporal Classification)模块对词级序列进行对齐[41]. 计算式为

$$\boldsymbol{z}_t,\boldsymbol{z}_a,\boldsymbol{z}_v=\mathrm{SeqAligned}\left(\boldsymbol{m}_t,\boldsymbol{m}_a,\boldsymbol{m}_v\right) \tag{4}$$

其中,$\boldsymbol{z}_t\in\mathbb{R}^{L\times D_T}$、$\boldsymbol{z}_a\in\mathbb{R}^{L\times D_A}$和$\boldsymbol{z}_v\in\mathbb{R}^{L\times D_V}$分别表示文本、音频和视频模态对齐后的特征表示;SeqAligned(·)表示CTC对齐模块.

3.2 层次化音视频语义学习

已有的工作关注为文本模态建立多模态语义环境[9,26,31],但视觉和音频模态中包含大量的语义信息同样需要关注. 在与文本对齐的音视频数据序列中,结合不同层级的音视频特征能够精准地表达人类意图,例如,人类意图表达通常涉及整段视频或音频的情感信息和动作信息,且某些模态内和模态间的特征之间的联系(如特定的面部表情、肢体语言与语调相关)也可以提供重要的意图线索. 音视频层次化语义学习的整体过程在图3中展示.

3.2.1 层次化特征构建

原始的音视频数据通常包含大量冗余信息和噪声,这会影响Transformer网络对模态语义的学习效果. 为了解决这一问题,在Transformer的输入中添加不含噪声的可学习额外标记是一种有效的方法. 这些标记能够通过注意力机制自适应地优化模型对重要特征的识别能力,迫使模型同时处理原始输入和新增标记之间的关系,从而避免Transformer网络对局部噪声的过度依赖. 此前的研究[9,30,42]表明:Transformer能够将不

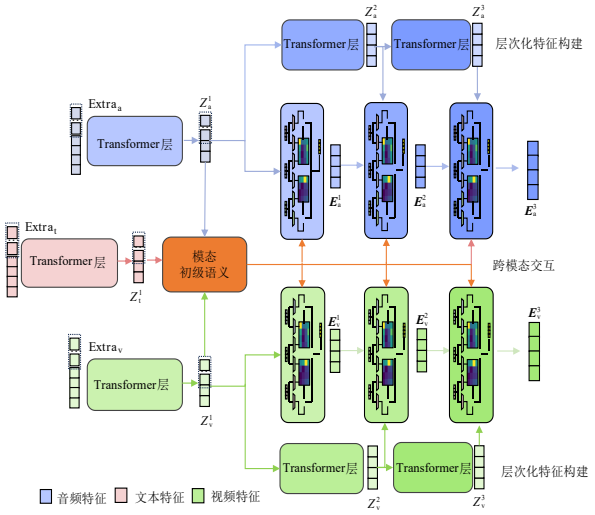


图3　层次化特征进行跨模态交互学习多尺度语义

同模态的基本信息有效地转移到这些标记中,从而避免噪声的干扰.

然而,单纯依赖可学习标记可能会导致信息瓶颈,从而损失意图识别所需的重要信息.为了解决这一问题,首先对多模态特征进行线性投影对齐,映射到统一的语义空间;然后将对齐后的特征与可学习标记进行拼接,形成增强的输入序列,同时保持原始序列的长度不变.这种方法平衡可学习标记的噪声抑制能力与通过保留多模态上下文的细节特征能力,为深层网络的语义学习提供模态基本信息:

$$Z_m = W_m z_m \tag{5}$$

$$Z_m^0 = \mathrm{Concat}(\mathrm{Extra}_m, Z_m) \tag{6}$$

$$Z_m^1 = \mathrm{Transformer}_m\left(Z_m^0, \theta_m\right) \tag{7}$$

其中,$m \in \{\mathrm{t}, \mathrm{a}, \mathrm{v}\}$;$W_m$ 表示可学习的线性变换参数;Extra_m 表示额外标记;$\mathrm{Concat}(\cdot)$ 表示拼接操作;$\mathrm{Transformer}_m$ 和 θ_m 分别表示模态 m 的 Transformer 编码器[43]和其对应的参数.

为了学习到音频和视觉模态中不同层级的语义特征,引入多层 Transformer 编码器以学习深层次的音频和视觉语义信息:

$$Z_m^i = T_m^i\left(Z_m^{i-1}, \theta_m^i\right) \tag{8}$$

其中,$m \in \{\mathrm{a}, \mathrm{v}\}$;$i \in \{2, 3\}$;$T_m^i$ 和 θ_m^i 分别表示第 i 层级模态 m 的 Transformer 编码器和其对应的参数.

3.2.2　层次化跨模态交互

由于意图通过多模态的方式表达,模态间互补的语义交互能帮助模型获取全面的意图线索.经过 Transformer 网络处理的模态特征会损失信息,因此选择与模态基本信息相关的初级特征用于跨模态的语义学习.针对音频和视觉模态的不同层级特征,模型通过计算当前层级模态特征与其他模态初级特征的互补关系,来增强不同尺度的语义表达.

为音频模态和视觉模态初始化特征维度与音频和视觉模态输入形状相同的特征向量 $\boldsymbol{E}_{\mathrm{a}}^0$ 和 $\boldsymbol{E}_{\mathrm{v}}^0$,通过模态间的交叉注意力计算逐步更新 $\boldsymbol{E}_{\mathrm{a}}^0$ 和 $\boldsymbol{E}_{\mathrm{v}}^0$ 中的语义信息.如图4所示,以音频模态层级语义学习为例来说明层次化跨模态交互.以音频模态的不同尺度特征 Z_{a}^i 为查询,从文本模态初级特征 Z_{t}^1 和视觉初级特征 Z_{v}^1 中学习语义信息补充更高层级的音频模态特征.

$$\boldsymbol{Q}_{\mathrm{at}}^i = \boldsymbol{W}_{\boldsymbol{Q}_{\mathrm{at}}} Z_{\mathrm{a}}^i \tag{9}$$

$$\boldsymbol{K}_{\mathrm{at}}^i = \boldsymbol{W}_{\boldsymbol{K}_{\mathrm{at}}} Z_{\mathrm{t}}^1 \tag{10}$$

$$\boldsymbol{V}_{\mathrm{at}}^i = \boldsymbol{W}_{\boldsymbol{V}_{\mathrm{at}}} Z_{\mathrm{v}}^1 \tag{11}$$

$$\mathrm{at\_head} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{at}}^i \left(\boldsymbol{K}_{\mathrm{at}}^i\right)^{\mathrm{T}}}{\sqrt{d_k}} \boldsymbol{V}_{\mathrm{at}}^i\right) \tag{12}$$

$$\boldsymbol{E}_{\mathrm{at}}^i = \mathrm{Concat}(\mathrm{at\_head}_1, \mathrm{at\_head}_2, \cdots, \mathrm{at\_head}_h) \tag{13}$$

$$\mathrm{av\_head} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{av}}^i \left(\boldsymbol{K}_{\mathrm{av}}^i\right)^{\mathrm{T}}}{\sqrt{d_k}} \boldsymbol{V}_{\mathrm{av}}^i\right) \tag{14}$$

$$\boldsymbol{E}_{\mathrm{av}}^i = \mathrm{Concat}(\mathrm{av\_head}_1, \mathrm{av\_head}_2, \cdots, \mathrm{av\_head}_h) \tag{15}$$

其中,$\boldsymbol{W}_{\boldsymbol{Q}_{\mathrm{at}}}$、$\boldsymbol{W}_{\boldsymbol{K}_{\mathrm{at}}}$ 和 $\boldsymbol{W}_{\boldsymbol{V}_{\mathrm{at}}}$ 表示可训练的参数矩阵;softmax 表示权重归一化函数;Concat 表示将两个特征向量进行拼接;d_k 用以保持数值的稳定;与 $\boldsymbol{E}_{\mathrm{at}}^i$ 的计算过程类似,$\boldsymbol{E}_{\mathrm{av}}^i$ 表示第 i 层音频特征 Z_{a}^i 和初级视觉模态特征 Z_{v}^1 之间的语义关系.

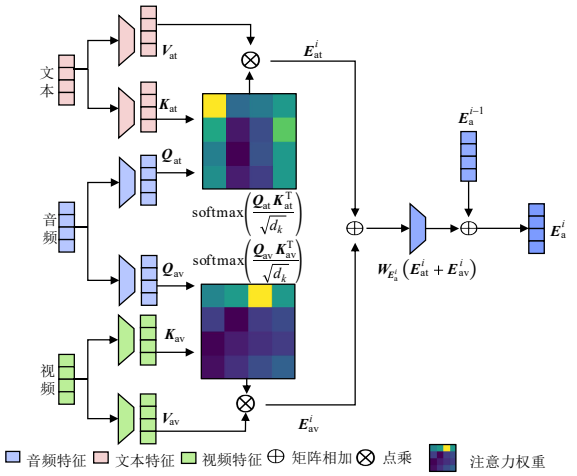


图4　音频-文本和音频-视觉的跨模态交互细节

为了增强视频模态的语义表达,类似地分别计算了第 i 层视觉特征和初级文本模态特征之间的语义关系,以及第 i 层视觉特征和初级音频模态特征之间的语义关系.接着通过加权计算,将音频和视频模态的跨模态语义关系逐层传递,并将其整合到特征向量 $\boldsymbol{E}_{\mathrm{a}}^0$ 和 $\boldsymbol{E}_{\mathrm{v}}^0$ 中:

$$\boldsymbol{E}_{\mathrm{a}}^{i}=\boldsymbol{E}_{\mathrm{a}}^{i-1}+\boldsymbol{W}_{\boldsymbol{E}_{\mathrm{a}}^{i}}(\boldsymbol{E}_{\mathrm{at}}^{i}+\boldsymbol{E}_{\mathrm{av}}^{i}) \tag{16}$$

$$\boldsymbol{E}_{\mathrm{v}}^{i}=\boldsymbol{E}_{\mathrm{v}}^{i-1}+\boldsymbol{W}_{\boldsymbol{E}_{\mathrm{v}}^{i}}(\boldsymbol{E}_{\mathrm{vt}}^{i}+\boldsymbol{E}_{\mathrm{va}}^{i}) \tag{17}$$

其中，$i\in\{1,2,3\}$；$\boldsymbol{E}_{\mathrm{a}}^{i}$和$\boldsymbol{E}_{\mathrm{v}}^{i}$表示第$i$个层次化语义学习层输出的音视频特征；第3层的输出是最终的音视频语义特征$\boldsymbol{E}_{\mathrm{a}}$和$\boldsymbol{E}_{\mathrm{v}}$；$\boldsymbol{W}_{\boldsymbol{E}_{\mathrm{a}}^{i}}$和$\boldsymbol{W}_{\boldsymbol{E}_{\mathrm{v}}^{i}}$表示可学习参数.

3.3 多模态一致性匹配

尽管模型通过层次化的音视频特征学习获得了包含多尺度语义信息的音频和视觉特征，由此增强模态表示，但仅依靠意图识别任务仍难以消除音频和视觉模态中的意图无关语义信息.

在对话型多模态任务中，尽管各模态对识别结果的贡献存在差异，但仍存在潜在的一致性[8,30]. 通过模态间潜在的一致性，可以引导模型过滤掉与意图无关的语义特征. 而文本模态有更加明确的意图语义特征[35,42]，因此以文本模态为锚点设计文本-音频配对和文本-视觉配对任务.

具体而言，类似于BERT中的核心训练任务“Next Sentence Prediction”，让模型学习预测，给定一对句子，第二个句子是否自然地接在第一个句子后面. 通过这种方式，BERT模型能够捕捉到句子之间的上下文关系，从而增强其对语言的理解能力. 论文为提出的模型设计了相似的任务，以引导模型学习跨模态的上下文关系，从而增强其模态一致性语义的理解能力. 首先，将所有标签初始化为0，然后以50%的概率随机替换与文本同一数据的音频特征，并将这些替换后的样本标签标记为1，其目的是制造带有干扰的信息（当前音视频特征与当前意图无关）；其次，通过使用以文本为查询的注意力机制，捕捉音视频特征和文本内容之间潜在的一致性语义关联；最后，将得到的配对表征输入分类器进行判断，分类器的任务是判断该样本是否经过特征替换，即是否包含当前意图无关的音视频语义. 以文本-音频配对任务损失$\mathrm{Loss}_{\mathrm{TApair}}$的计算为例，通过相似的方法可以得到$\mathrm{Loss}_{\mathrm{TApair}}$，并使用交叉熵损失函数来优化配对任务：

$$\boldsymbol{Q}_{\mathrm{TA}}=\boldsymbol{W}_{\boldsymbol{Q}_{\mathrm{TA}}}\boldsymbol{Z}_{\mathrm{t}} \tag{18}$$

$$\boldsymbol{K}_{\mathrm{TA}}=\boldsymbol{W}_{\boldsymbol{K}_{\mathrm{TA}}}\boldsymbol{E}_{\mathrm{a}} \tag{19}$$

$$\boldsymbol{V}_{\mathrm{TA}}=\boldsymbol{W}_{\boldsymbol{V}_{\mathrm{TA}}}\boldsymbol{E}_{\mathrm{a}} \tag{20}$$

$$\boldsymbol{E}_{\mathrm{TApair}}=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{TA}}\boldsymbol{K}_{\mathrm{TA}}^{\mathrm{T}}}{\sqrt{d_k}}\boldsymbol{V}_{\mathrm{TA}}\right) \tag{21}$$

$$y_{\mathrm{TApair}}=f_{\mathrm{TA}}(\boldsymbol{E}_{\mathrm{TApair}}) \tag{22}$$

$$\mathrm{Loss}_{\mathrm{TApair}}=-\sum_{1}^{N}y_{\mathrm{TApair}}\ln\left(\frac{\exp(\hat{y}_m)}{\sum_{m}(\hat{y}_m)}\right) \tag{23}$$

3.4 多模态特征融合层

在提取多层次音频和视觉模态语义之后，模型引入多模态融合门来融合3种模态信息. 已有的研究[13,31]认为音视频模态的输入会影响预训练模型生成的文本表征在语义空间中的位置. 这种影响受到音频和图像数据的共同控制，因此模型通过一种位移融合门机制向文本嵌入向量中加入音视频模态语义. 模型将文本特征与视觉语义特征或音频语义特征拼接，并通过线性层和非线性激活函数映射到多模态语义空间来生成门控向量：

$$\boldsymbol{g}_m=\mathrm{ReLU}\left(\boldsymbol{W}_m\left(\mathrm{Concat}(\boldsymbol{Z}_{\mathrm{t}},\boldsymbol{E}_m)+\boldsymbol{b}_m\right)\right) \tag{24}$$

其中，$m\in\{\mathrm{a},\mathrm{v}\}$，a和v分别表示音频模态与视觉模态；$\boldsymbol{W}_m$、$\boldsymbol{b}_m$表示可学习线性权重矩阵和偏置项. 模型通过将音频特征向量、文本特征向量及其对应的门控向量相乘，计算不同模态对文本语义偏移的贡献，从而创建一个受音视频模态控制的位移向量$\boldsymbol{\Delta}$：

$$\boldsymbol{\Delta}=\boldsymbol{g}_{\mathrm{a}}\boldsymbol{E}_{\mathrm{a}}+\boldsymbol{g}_{\mathrm{v}}\boldsymbol{E}_{\mathrm{v}} \tag{25}$$

将原文本特征$\boldsymbol{Z}_{\mathrm{t}}$与受音视频模态控制的位移向量$\boldsymbol{\Delta}$求和，生成多模态融合表示向量$\boldsymbol{Z}_{\mathrm{fusion}}$. 为防止音视频模态的过度影响，采用文本模态特征和位移向量的L2范数确保表示$\boldsymbol{Z}_{\mathrm{fusion}}$保持在合理范围内：

$$\omega=\min\left(\frac{\|\boldsymbol{Z}_{\mathrm{t}}\|_2}{\|\boldsymbol{\Delta}\|_2},1\right) \tag{26}$$

$$\boldsymbol{Z}_{\mathrm{fusion}}=\boldsymbol{Z}_{\mathrm{t}}+\omega\boldsymbol{\Delta} \tag{27}$$

最后，产生的融合特征$\boldsymbol{Z}_{\mathrm{fusion}}$被送入BERT模型编码器，接着经全连接层处理后用于MIR任务：

$$\hat{y}=f_c\left(\mathrm{BERT\_Encoder}(\boldsymbol{Z}_{\mathrm{fusion}})\right) \tag{28}$$

$$\mathrm{Loss}_{\mathrm{intent}}=-\sum_{1}^{N}\mathrm{y}_{\mathrm{intent}}\ln\left(\frac{\exp(\hat{y}_m)}{\sum_{m}(\hat{y}_m)}\right) \tag{29}$$

其中，$\mathrm{Loss}_{\mathrm{intent}}$表示意图预测的损失；$f_c(\cdot)$表示分类函数. 将意图预测损失和多模态配对学习的损失相加一起来优化整个模型：

$$\mathrm{Loss}_{\mathrm{total}}=\mathrm{Loss}_{\mathrm{intent}}+\mathrm{Loss}_{\mathrm{TApair}}+\mathrm{Loss}_{\mathrm{TVpair}} \tag{30}$$

4 实验结果与分析

4.1 数据集

MIntRec[6]是一个包含2 224个高质量样本的多模态细粒度意图识别数据集，含有20个意图类别的文本、音频和视频模态数据，详细的意图类别见表1，MIntRec数据集的统计数据见表2. 训练集、验证集和测试的比例为3∶1∶1.

MIntRec2.0[7]是一个大规模的MIR数据集，包括1 245个对话和15 040个多模态数据. 其中的数据和

表1　MIntRec数据集上的细粒度意图分类和解释

项目	意图种类	解释
表达情感或者态度	Complain	表达对某人或某事的不满
	Praise	表达对某人或某事的钦佩
	Apologize	对做错的事情表示后悔
	Thank	对他人给予或提供的便利或善意,在言语或行为上表示感谢
	Criticize	严厉地指出某人的错误
	Care	关心某人或对某事感到好奇
	Agree	对某事有相同的态度
	Taunt	使用隐喻和夸张来指责和嘲笑
	Flaunt	夸耀自己以获得钦佩、嫉妒或赞扬
	Oppose	对某事态度不一致
	Joke	说些逗笑的话
完成目标	Inform	告诉某人让他们意识到某事
	Advise	提供建议供考虑
	Arrange	计划或组织某事
	Introduce	让某人认识另一个人或推荐某物
	Comfort	用鼓励或同情来减轻痛苦
	Leave	离开某个地方
	Prevent	使某人不能做某事
	Greet	在相遇时表达彼此的善意或认可
	Ask for help	请求别人帮助

表2　MIntRec数据集的统计信息

项目	统计数据
粗粒度意图/类	2
细粒度意图/类	20
视频片段的数量/段	2 224
文本话语中的单词数量/词	15 658
文本话语中独特单词的数量/词	2 562
文本话语的平均长度/s	7.04
视频片段的平均长度/s	2.38

MIntRec中的数据不相重合,并且相较于MIntRec数据集,MIntRec2.0多达30个意图类别. 训练集、验证集和测试集的比例是7∶1∶2.

MOSI(Multimodal Opinion-level Sentiment Intensity dataset)[44]是卡耐基梅隆大学提出的一个多模态情感分析数据集,该数据集从YouTube上93个现实人物独白视频上获取视频数据. 共含有2 199个视频片段,其中训练集、验证集和测试集分别含有1 284、229和686个视频片段.

MELD-DA[5]是MELD语料库的DA(Dialogue Action)适应版本,是一个为对话行为分类设计的大规模数据集. 该数据集包括9 988个多模态样本,在12个常见的对话行为标签中进行注释,训练集、验证集和测试集的比例是7∶1∶2.

4.2　对比模型

为了具体地展现所提出的模型特点,本节模型与一系列具有竞争力的MIR模型进行对比,如基于时间对比学习与多注意力池化的模型(Temporal Contrastive Learning with Multi-Attention Pooling, TCL-MAP)、基于语义距离的对话行为交互框架(Semantic Distance-based Interaction Framework for Dialogue Act, SDIF-DA)、交叉注意力图卷积模型(Cross-Attention Graph Convolution, CAGC). 除此之外,还和从开源代码中复现的多模态融合方法跨模态嵌入网络(Cross-modal Embedding NETwork, CENET)和多模态任务自适应学习(Adaptive Learning for Multimodal Tasks, ALMT)模型的结果进行对比. 对比模型介绍如下.

多模态Transformer(Multimodal Transformer, MulT, 2019 ACL)[30]利用跨模态注意力机制捕捉不同模态之间的重要信息,并使用Transformer整合特征.

多模态注意力门控BERT(Multimodal Attention-based Gating BERT, MAG-BERT, 2020 ACL)[31]通过采用门控机制将多种模态结合在一起,使得BERT模型能够接收多模态输入.

多模态信息子空间对齐(Multimodal Information Subspace Alignment, MISA, 2020 ACM Multimedia)[32]将多模态数据映射到不同的特征子空间,并通过结合各种损失函数(如模态相似度损失和模态重建损失)以学

习模态不变和模态特定的特征.

CENET(2022 TMM)[33]嵌入CE模块,通过将视觉和音频信息集成到语言模型中来增强文本表示.

ALMT(2023 EMNLP)[42]引入自适应超模态学习(Adaptive Hyper-modality Learning,AHL)模块,在语言特征的指导下从视觉和音频特征中学习超模态特征.

文本增强型Transformer融合网络(Text Enhanced Transformer Fusion Network,TETFN,2023 PR)[35]学习面向文本的成对跨模态映射,以获得有效的统一多模态表示.

有效的多模态表示与融合方法(Effective Multimodal Representation and Fusion Method,EMRFM,2023 NeuroComputing)[8]将模态投射到特定子空间和共享子空间,并通过注意力机制决定不同模态的权重.

TCL-MAP(2023 AAAI)[9]引入模态感知标记来学习音视频模态的潜在语义特征,通过基于相似度的模态对齐和跨模态注意力机制融合多模态特征.

基于模态对齐感知的多模态意图识别提示学习(Prompt learning for Multimodal intent recognition with modal Alignment Perception,PMAP,2024 Cognitive Computation)[36]构建用于提示学习的模板,以增强文本表示,并使用跨模态对齐感知来解决模态间的不一致问题.

CAGC(2024 CVPR)[10]提出一个上下文增强的全局对比方法,用于捕捉丰富的全局上下文特征,识别多模态意图.

SDIF-DA(2024 ICASSP)[26]利用Chat-GPT对文本模态数据进行增强,并提出由浅入深的跨模态交互结构聚合多模态信息.

4.3 评价指标

在MIntRec数据集上参考之前工作[6]的指标来展示结果.通过以下指标评估结果:意图识别的F_1值(F_1 Score,F_1)、召回率(Recall,R)值、精确度(Precision,P)值和ACC值.在MIntRec2.0数据集上,参考之前工作[7]的指标:计算意图识别的ACC值、R值、加权F_1值(Weighted F_1 Score,WF_1)和加权精确度(Weighted Precision,WP)值对比实验结果.

对于MOSI数据集,本文遵循之前的工作[44],计算二分类准确率(ACCuracy for binary classification,ACC-2)值、F_1、七分类准确率(ACCuracy for seven-class classification,ACC-7)值、皮尔逊相关系数(Pearson Correlation coefficient,Corr)和平均绝对误差(Mean Absolute Error,MAE)指标来展示结果.对于MELD-DA数据集,遵循之前工作[9]采用的协议,使用ACC值、R值、WF_1值和WP值来对比模型结果.

4.4 实验设置

在所提出的模型的层次化语义学习模块中,注意力头数量设置为8,稳定数值的参数d_k被设置为16.初级语义特征中的额外标记长度为8,层次化语义学习层深度设置为3.

为了公平地对比实验结果,模型在MIntRec数据集上使用Hugging-face库中的bert-base-uncased模型提取文本特征,音视频特征提取的模型同样和之前的工作[6]保持一致.设置训练批次大小为64,验证和测试集批次大小为32,使用AdamW优化器训练,并设置学习率为2.5×10^{-5},权重衰减大小为0.03.

在MIntRec2.0数据集上遵循此前工作[7]的设置,模型使用bert-large-uncased模型提取文本特征,且音视频特征提取的模型也保持一致.设置训练批次大小为64,验证和测试集批次大小为32,使用AdamW优化器训练,并设置学习率为5×10^{-6},权重衰减大小为0.03.对于开源代码模型的实验结果,实验都遵守其论文中关键参数设置.

4.5 实验结果与分析

现有的工作共同使用的数据集是MIntRec,因此本文在这个数据集上对比最多的意图识别模型与开源多模态融合模型.此外,本文在意图任务相关的MELD-DA、MOSI和MIntRec2.0数据集上与这些模型进行了全面对比.

MIntRec数据集的结果如表3所示,加粗字体代表最佳结果,下划线代表次优结果.所提出的模型在所有指标的结果上都达到最佳.相较于次优结果,所提出的模型在ACC值上提高0.90个百分点,在F_1上提高0.78个百分点,在P值上提高了1.84个百分点,在R值上提高0.81个百分点.

表3 MIntRec数据集上MIR的整体结果 单位:%

模型	ACC值↑	F_1↑	P值↑	R值↑
TEXT	70.88	67.40	68.07	67.44
MulT	72.52	69.25	70.25	69.24
MISA	72.29	69.32	70.85	69.24
MAG-BERT	72.65	68.64	69.08	69.28
ALMT	71.01	68.77	69.54	69.90
CENET	73.48	69.88	70.48	69.99
PMAP	72.94	69.64	69.66	70.39
EMRFM	72.58	70.46	71.90	70.45
TCL-MAP	73.62	—	—	70.50
CAGC	73.39	70.09	71.21	70.39
SDIF-DA	<u>73.71</u>	<u>71.58</u>	<u>72.43</u>	<u>71.21</u>
OURS	**74.61**	**72.36**	**74.27**	**72.02**

从表3可知,CAGC和SDIF-DA是除本文所提出的模型外表现突出的两个模型.MISA和EMRFM通过不同的长短期记忆(Long Short-Term Memory,LSTM)网络编码器在共享子空间和私有子空间中对不同模态进行

建模,学习多模态表征. 此外,EMRFM提出一种将文本模态作为主要特征的注意力融合机制,能够有效地区分不同模态的贡献,从而在整体效果上取得更好的表现. 尽管这些方法在模态表征方面有研究进展,但在语义关系方面尚未进行深入探索. PMAP通过设计文本提示学习模板,以增强预训练模型生成的文本表示,用于文本表示在多模态语义环境中的对齐问题. SDIF-DA则提出基于大语言模型的数据增强方法,用于增加文本模态的语料,并以文本为中心进行模态对齐与融合. 这些模型在构建多模态语境时,往往关注复杂的模态对齐与融合机制,但未能充分利用音视频模态中的丰富语义信息. CAGC注意到,视觉和音频模态提供的意图语义存在不足,因此,其通过构建具有相似背景的视频库,以提供更长上下文信息,从而减少意图偏差. 然而,尽管不同视频来源的音视频在场景上相似,但由于说话人的表达方式不同,导致音视频模态的语义表达存在差异,这在一定程度上限制了意图识别的效果.

现有的MIR研究虽然注重基于文本模态的对齐与特征融合,但忽略了理解音视频模态中的语义关系的重要性;而音视频模态中的冗余信息和噪声阻碍有效的语义表达以及多模态表示进一步融合与对齐. 所提出的模型为不同模态构建抑制冗余特征的初级特征,通过层次化语义学习挖掘音视频模态内部的深层语义关系;以说话人意图更为明确的文本模态为锚点,过滤冗余语义和噪声,提升多模态融合效果. 因此,在MIntRec数据集上,所提出的模型取得最佳效果.

在更大规模的意图识别数据集MIntRec2.0上的对比实验结果如表4所示,加粗字体代表最佳结果,下划线代表次优结果. 本文提出的模型在绝大多数指标上表现优异. 例如,在ACC值上相比次优模型提高0.34个百分点,在WF_1值上提高0.87个百分点,在WP值上提高1.00个百分点,在R值上提高0.90个百分点. 在MIntRec和MIntRec2.0数据集上稳定的实验结果说明:提出的模型在MIR任务中,相较于其他模型具有更好的泛化性. 由于MIntRec2.0是一个相对较新的大规模数据集,现有的MIR研究在该数据集上开展的实验结果较少. 因此,本文主要与数据集作者提出的基线模型MulT、MAG-BERT以及开源的MISA、CENET和ALMT等模型进行对比.

MulT模型通过Transformer关注不同时间步长的多模态序列间的相互作用,将一种模态转换为另一种模态,对齐模态特征. 然而,MulT未进一步提取模态语义关系,造成其性能在大数据集上相对下降. 而MAG-BERT计算多模态位移门控向量微调文本模态语义,挖掘预训练模型中与意图有关的多模态知识,从而在多个数据集上的效果更加稳定. ALMT模型通过构造以文本模态特征为主导的精炼超模态表征. 然而,在以文本为主导的超模态表征编码模态基本信息时,虽然过滤了噪声,但也损失了模态上下文信息,导致在关注细节特征的意图识别任务中的表现下滑. 在所提出的模型中,通过结合可学习标记和原始序列构造多尺度特征,既学习了精炼的信息,又保留了模态上下文信息. CENET模型通过减小文本和音视频模态之间的初始分布差异,促进了多模态融合,并基于文本模态增强音视频模态的表征,从而在不同数据集上表现出色.

表4　MIntRec2.0数据集上MIR的整体结果　　单位:%

模型	ACC值↑	WF_1值↑	WP值↑	R值↑
TEXT	59.30	58.01	58.85	51.31
MulT	60.18	58.82	59.38	52.56
MISA	<u>60.90</u>	<u>59.47</u>	59.20	51.87
MAG-BERT	60.45	59.36	<u>60.49</u>	<u>54.07</u>
ALMT	58.09	57.67	57.59	52.23
CENET	60.65	59.40	60.47	53.55
OURS	**61.24**	**60.34**	**61.49**	**54.97**

对话行为分类任务的目的是捕捉说话人在对话交流时的意图. 为了进一步确认提出的模型的有效性,在MELD-DA数据集上和现有的意图识别模型进行了比较. 结果如表5所示,加粗字体代表最佳结果,下划线代表次优结果. 本文提出的模型在ACC值和R值上比次优结果高0.76个百分点和1.51个百分点,在WF_1值和WP值上分别比次优结果高1.00个百分点和1.20个百分点.

表5　MELD-DA数据集上的整体结果　　单位:%

模型	ACC值↑	WF_1值↑	WP值↑	R值↑
MulT	60.36	59.01	59.44	49.93
MAG-BERT	60.63	59.36	59.80	50.01
MISA	59.98	58.52	59.28	48.75
ALMT	59.71	57.90	57.90	<u>50.93</u>
CENET	61.26	58.74	58.91	46.24
TCL-MAP	<u>61.75</u>	<u>59.77</u>	<u>60.33</u>	50.14
OURS	**62.51**	**60.77**	**60.53**	**52.44**

TCL-MAP模型在MELD-DA数据集上表现优异,该模型通过将提示学习融入对比学习的监督信号,从而增强了文本表示能力. 该模型利用音视频特征训练可学习的额外标记来构建文本提示词,相比于手工设计的提示词,具有更高的灵活性和有效性. 然而,TCL-MAP仍然缺乏对音视频特征内部语义的深入提取,导致在多个指标上的表现不及提出的模型.

为了说明模型对人类情感意图数据识别的有效性,在CMU-MOSI上与意图识别模型以及情感分析模型进行对比. HyCon、ConFEDE和self-MM模型实验结

果源自文献[10]. 结果如表6所示,粗体表示最优的结果,下划线代表次优结果. 相比现有的意图识别模型,本文提出的模型在情感分析任务上也能取得优秀的效果. 这得益于所提出的模型对音视频模态特征多尺度语义关系的提取和对多模态信息的有效整合. 除提出的模型外,self-MM模型作为具有代表性的多模态情感分析模型,在多项指标上取得仅次优效果,然而,self-MM模型依赖特定的情感极性标签,因此无法应用于意图识别任务上.

表6 MOSI数据集上的整体结果 单位:%

模型	ACC-2值↑	F_1↑	ACC-7值↑	Corr↑	MAE↓
MulT	83.0	82.8	40.0	69.8	0.871
MAG-BERT	83.5	83.5	42.9	76.9	0.790
MISA	83.4	83.6	42.3	76.1	0.783
self-MM	85.5	85.4	<u>46.6</u>	<u>79.6</u>	<u>0.708</u>
HyCon	85.2	85.1	<u>46.6</u>	79.0	0.713
ConFEDE	85.5	85.5	42.3	78.4	0.742
EMRFM	84.7	84.8	46.1	78.5	0.722
CAGC	<u>85.7</u>	<u>85.6</u>	44.8	77.4	0.775
Ours	**86.0**	**85.9**	**46.7**	**79.7**	**0.703**

为了进一步分析现有的MIR模型和所提出的模型,在表7上展示对比模型的计算复杂度和参数数量. 本文使用每秒浮点运算次数(FLoating point Operations Per second,FLOPs)来衡量计算复杂度. 由于大多数意图识别模型采用BERT模型进行文本特征提取,因此将纯文本BERT模型作为基线对比. 值得注意的是,TCL-MAP模型显著增加了计算资源的消耗. 这是由于其利用两个BERT模型分别对文本数据和提示进行编码,以进行对比学习. 相比之下,MAG-BERT引入了轻量的门控机制,可以保持较低的计算成本. 类似地,ALMT仅使用少量参数来学习超模态表示,但过滤了太多信息,导致在意图识别任务中性能损失较大. 由于在音频和视频模态中进行了层次化语义学习,本文提出的模型引入了额外的参数量,但并未大幅增加计算复杂度. 这种效率来源于层次化语义学习中的额外标记的优化,使得所提出的模型能够在不显著增加计算负担的情况下,获得最佳的分类性能.

综合以上实验结果表明:所提出的模型在多个数据集上表现优异,体现出其在意图相关任务中的良好泛化能力. 同时,实验表明音视频模态信息未能充分发挥作用是制约意图识别模型性能提升的关键因素.

4.6 细粒度意图分类效果

为了更好地分析所提出的模型在细粒度意图分类上的表现,本节将其与一些基线意图识别模型进行对比,评估这些模型在MIntRec数据集上每个意图分类的F_1值. 实验结果如表8所示,其中粗体表示最优结果.

表7 在MIntRec上的计算复杂度和模型参数对比

模型	FlOPs/G	参数量/M
纯文本BERT	40.8	85.6
MAG-BERT	43.1	88.6
MISA	51.1	115.9
MulT	115.6	106.1
ALMT	42.9	88.4
CENET	67.3	92.9
TCL-MAP	106.5	204.2
Ours	54.4	109.7

表8 在MIntRec上的细粒度类别的F_1值对比 单位:%

类别	MAG-BERT	MulT	MISA	PMAP	EMRFM	TCL-MAP	OURS
Complain	67.65	65.48	63.91	67.12	61.54	**68.70**	66.67
Praise	86.03	84.72	86.63	88.85	**89.16**	87.20	88.37
Apologize	97.76	97.93	97.78	96.87	96.30	97.70	**100.00**
Thank	96.52	96.83	98.03	97.66	96.15	97.00	**98.04**
Criticize	49.02	49.72	53.44	50.28	**62.50**	51.30	47.83
Care	85.59	88.12	87.14	87.51	**89.47**	86.80	85.00
Agree	91.60	92.23	92.05	94.80	96.00	93.10	**100.00**
Taunt	15.78	26.12	22.15	21.49	15.38	17.20	**32.26**
Flaunt	47.09	48.91	46.44	44.31	47.06	50.80	**58.82**
Joke	37.54	33.95	38.74	36.06	37.50	29.00	**50.00**
Oppose	33.97	34.68	36.15	38.09	30.00	35.90	**42.11**
Comfort	76.43	76.44	78.78	78.27	**83.33**	79.80	81.08
Inform	71.00	70.85	70.18	68.90	67.83	72.80	**74.07**
Advise	69.30	69.43	69.56	70.43	**75.86**	68.90	70.00
Arrange	63.82	65.44	67.32	**67.38**	57.14	65.40	61.54
Introduce	67.42	71.19	67.22	66.53	70.00	68.40	**72.00**
Leave	75.77	75.58	77.23	79.59	**83.87**	83.40	**83.87**
Prevent	85.07	81.68	83.30	86.15	86.67	83.60	**89.66**
Greet	**91.06**	86.65	82.71	83.95	85.71	90.10	85.71
Ask for Help	64.44	69.12	67.57	64.97	**77.78**	66.40	56.00

所提出的模型在20个类别中有11个类别上达到最优结果. 特别是在“Apologize”和“Agree”等类别上,该模型展现出良好的效果,这得益于模型通过可学习的额外标记以及一致性语义学习,有效地过滤了音视频模态中与意图无关的语义信息. 从实验结果可以发现,现有的MIR模型在“Taunt”“Oppose”“Flaunt”和“Joke”等类别上表现较差,这是由于这些意图在现实世界中表现形式较为复杂;而本文提出的模型在这些类别上取得更好的结果,这主要得益于模型在音视频模态中对不同尺度语义的深入挖掘,从而使得该模型通过音视频与文本语义的互补关系更好地理解说话人意图. 但同时可以观察到,所提出模型在“Criticize”和“Ask for Help”类别上表现较差,这是由于一些带有对比性或反差的音视频语义在基于文本模态的配对任务

中被过滤所导致的.

4.7 消融实验与分析

为了更好地分析每个组成部分的合理性及其作用,在公开数据集 MIntRec 和 MIntRec2.0 上进行消融实验.

如表9所示,加粗字体代表最佳结果. 在相同的参数设置与环境下进行消融实验. 在进行音视频模态的消融实验中可以发现,消融音视频模态使得意图识别效果都产生明显的降低,说明结合音视频模态特征能更有效地识别说话人的意图. 此外对模型中关键组件做了如下消融实验分析其有效性.

(1)去除多模态配对任务,层次化语义学习中的语义特征学习只通过意图识别任务监督. 在 MIntRec 数据集上 ACC 值下降 0.90 个百分点,在 MIntRec2.0 数据集上 ACC 值下降 0.79 个百分点,这说明用配对任务损失来引导学习模态一致性语义的有效性.

(2)去除层次化音视频语义学习(Hierarchical Semantics Learning,HSL),将音视频模态特征直接用于下游任务. 在 MIntRec 数据集上 ACC 值下降 1.80 个百分点,在 MIntRec2.0 数据集上下降 0.54 个百分点. 这显示通过不同层级的音视频语义学习提取的语义特征与说话人意图有较强相关性.

(3)同时去除层次化语义学习模块与多模态配对损失,将音视频特征直接输入到多模态融合层,可以发现性能下降比例更大,在 MIntRec 数据集上 ACC 值下降 3.15 个百分点,在 MIntRec2.0 数据集上 ACC 值下降 1.62 个百分点. 这一结果体现出层次化语义学习模块和多模态配对损失的协同作用对提升模型性能的重要性.

(4)去除学习的音视频门控向量,音视频特征对融合特征贡献固定为相同. 在 MIntRec 数据集上 ACC 值下降了 1.58 个百分点,在 MIntRec2.0 数据集上 ACC 值下降了 0.79 个百分点. 这显示不同模态对意图识别的贡献存在差异,音视频门控向量通过动态调整不同模态的权重,使模型能有效地融合模态特征.

为了展示层次化语义学习模块和多模态配对任务提取音视频模态语义的效果,在表10中详细展示这些模块对音视频特征在意图识别任务中的影响,加粗字体代表最佳结果. 具体而言,将未经层次化语义学习模块处理的音视频特征进行序列对齐,通过3层 Transformer 网络提取特征,并将最后1层的最终时间步状态通过全连接层后用于 MIR. 实验结果表明:在冗余信息和噪声干扰的音视频模态中,Transformer 网络难以有效学习到有用的知识.

表9 MIntRec 和 MIntRec2.0 数据集上的消融实验结果　单位:%

模型	MIntRec ↑		MIntRec2.0 ↑	
	ACC 值	F_1	ACC 值	WF_1 值
w/o audio	71.91	69.31	60.80	59.28
w/o video	71.46	68.00	60.85	59.23
w/o Loss_pair	73.71	70.32	60.45	59.20
w/o HSL	72.81	68.72	60.7	59.42
w/o Loss_pair & HSL	71.46	68.12	59.62	57.94
w/o fusion gate	73.03	69.54	60.45	59.22
Overall	**74.61**	**72.36**	**61.24**	**60.34**

表10 音视频模态特征的意图识别实验结果　单位:%

模型	ACC 值 ↑	F_1 ↑	P 值 ↑	R 值 ↑
单音频模态+Transformer	26.52	21.35	24.70	20.02
单音频模态+HSL	66.07	63.60	65.50	65.34
单音频模态+HSL+Loss_pair	69.66	66.65	67.43	67.89
单视频模态+Transformer	14.61	9.65	9.80	10.03
单视频模态+HSL	65.17	62.42	65.2	64.17
单视频模态+HSL+Loss_pair	**69.89**	**67.2**	**70.57**	**68.02**

作为对比,用本文中提出的层次化语义学习模块替代 Transformer 网络. 通过层次化的特征表示与跨模态交互,模型能够更好地捕捉语义信息,从而提升意图识别的效果. 随后加入多模态配对任务,帮助过滤掉意图无关的语义. 实验结果表明:音视频单模态特征的识别效果也得到进一步的提升.

4.8 可视化分析

为了更直观地展示各模块的作用,对关键模块中的数据进行可视化. 图5以视频模态为例,展示不同层级特征 Z_v^1、Z_v^2 和 Z_v^3 的可视化结果. 从图5可得,不同层级特征在空间中表现出明显独立的分布差异. 从初级特征到高级特征,数据的分布从分散逐渐聚集. 这是由于随着特征关注语义的层级提升,关注的范围由局部信息扩展到整体内容,数据分布也从局部过渡到整体聚集. 因此模型能通过不同层级的特征更好地理解音视频模态语义.

为了说明多模态配对任务对模型的影响,向视频特征序列中添加了随机噪声,对比提出的模型添加噪声前后提取音视频特征的过程. 图6展示配对任务中的文本-视频注意力. 可以观察到,用红色框标记的加入噪声的位置对应的注意力明显下降,模型对随机噪声位置的信息关注减少,这表明多模态配对任务能辅助模型过滤意图无关的语义信息.

在 MIntRec 数据集上,对关键超参数(额外标记长度和层次化语义学习层数)对模型影响进行可视化,以探究这些参数取值对模型性能的影响. 在图7中展示了初级语义特征中额外标记长度对模型性能的影响. 当将额外标记初始长度设置为30时,初始特征中的信息完全由额外标记中的表示所覆盖,这导致模态信息

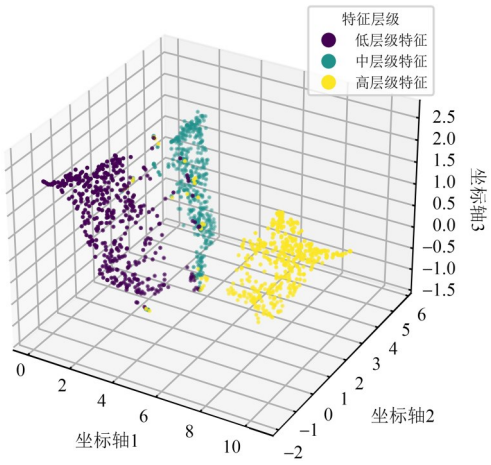
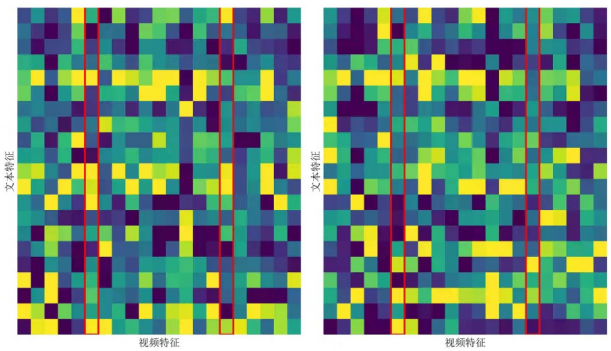


图5 不同层级特征在三维空间中的分布



(*a*) 加入随机噪声前 (*b*) 加入随机噪声后

图6 视频特征中加入随机噪声前后的配对任务注意力对比

的损失影响了深层语义提取. 通过逐步减少额外标记的长度来观察其变化. 实验结果表明:当额外标记长度为8时,模型性能最佳,ACC值达到了最高值. 当进一步减小额外标记长度时,音视频初级特征中的冗余信息开始积累,从而导致模型性能下降. 这表明过长或过短的标记长度都可能对模型性能产生负面影响.

在图8中展示了音视频层次化语义学习的层数对模型性能的影响. 由图8可知,最优的层数为3层,这时

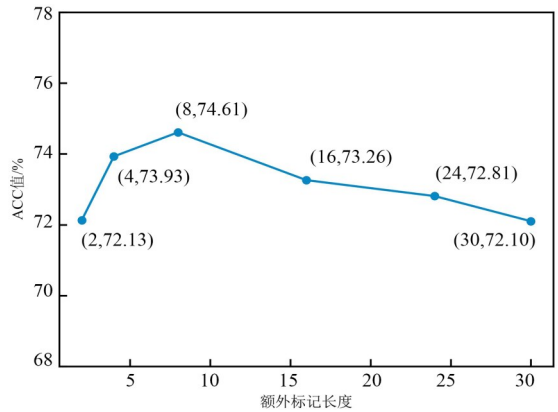


图7 初级语义特征中额外标记长度对模型性能的影响

模型能够最大限度地捕捉音视频数据的层次化语义信息. 过少的层数无法有效挖掘数据的深层特征,而过多的层数则可能引起过拟合或添加冗余信息,从而导致性能下降.

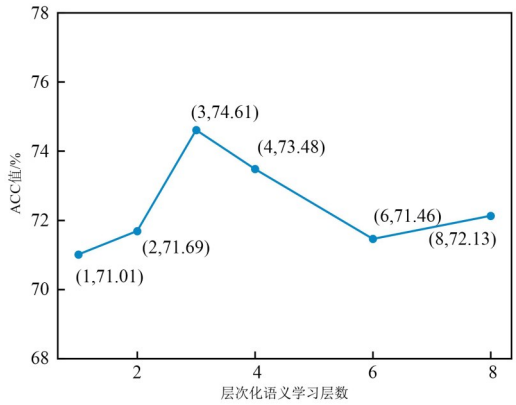


图8 层次化语义学习层深度对模型性能的影响

5 结论

针对MIR中存在的音视频模态语义利用不足、冗余信息和噪声干扰问题,本文提出了一种能够提取音视频模态语义关系并抑制冗余信息的MIR模型. 该模型通过层次化语义学习挖掘音视频模态中的语义关联,增强音视频模态特征. 为避免在语义学习过程中受到冗余信息和噪声的干扰,模型结合可学习的额外标记与原始序列,构成模态初级特征,从而抑制冗余信息并提供更全面的上下文信息. 为了提升音视频模态的语义表达能力,模型引入跨模态交互机制,通过提供互补的模态信息以帮助理解说话人的意图. 此外,模型利用不同模态特征之间潜在的意图一致性,将包含明确意图语义的文本特征与音视频模态表征进行匹配,过滤掉仅通过意图识别任务难以消除的无关语义信息. 在多模态融合阶段,模型采用自适应门控机制,动态地整合多模态的语义信息. 通过对比实验、消融研究以及可视化分析,验证了所提出的方法在提高MIR性能方面的有效性.

参考文献

[1] 杨帆, 饶元, 丁毅, 等. 面向任务型的对话系统研究进展[J]. 中文信息学报, 2021, 35(10): 1-20.
YANG F, RAO Y, DING Y, et al. Progress in task-oriented dialogue system[J]. Journal of Chinese Information Processing, 2021, 35(10): 1-20. (in Chinese)

[2] MEI J, WANG Y F, TU X H, et al. Incorporating BERT with probability-aware gate for spoken language understanding[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 826-834.

[3] ZHANG F, CHEN W, DING F, et al. Dual class knowledge propagation network for multi-label few-shot intent detection[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2023: 8605-8618.

[4] WEI Y W, YUAN S Z, YANG R S, et al. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2023: 5240-5252.

[5] SAHA T, PATRA A, SAHA S, et al. Towards emotion-aided multi-modal dialogue act classification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 4361-4372.

[6] ZHANG H L, XU H, WANG X, et al. MIntRec: A new dataset for multimodal intent recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 1688-1697.

[7] ZHANG H, WANG X, XU H, et al. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations[C]//The 12th International Conference on Learning Representations. Washington DC: ICLR, 2024:1-16.

[8] HUANG X J, MA T H, JIA L, et al. An effective multimodal representation and fusion method for multimodal intent recognition[J]. Neurocomputing, 2023, 548: 126373.

[9] ZHOU Q R, XU H, LI H, et al. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(15): 17114-17122.

[10] SUN K L, XIE Z W, YE M, et al. Contextual augmented global contrast for multimodal intent recognition[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2024: 26953-26963.

[11] ZHENG C Z, PENG J J, WANG L, et al. Frame-level nonverbal feature enhancement based sentiment analysis[J]. Expert Systems with Applications, 2024, 258: 125148.

[12] ZHENG C Z, PENG J J, CAI Z S. Extracting method for fine-grained emotional features in videos[J]. Knowledge-Based Systems, 2024, 302: 112382.

[13] XUE Z H, MARCULESCU R. Dynamic multimodal fusion[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2023: 2575-2584.

[14] XIE Y, ZHU Z, LU X, et al. InfoEnh: Towards multimodal sentiment analysis via information bottleneck filter and optimal transport alignment[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics. Marseille: LREC. 2024: 9073-9083.

[15] SU B Y, WANG J, LIU S Q, et al. A CNN-based method for intent recognition using inertial measurement units and intelligent lower limb prosthesis[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2019, 27(5): 1032-1042.

[16] MENSIO M, RIZZO G, MORISIO M. Multi-turn QA: A RNN contextual approach to intent classification for goal-oriented systems[C]//Companion of the Web Conference 2018 on the Web Conference 2018-WWW'18. New York: ACM, 2018: 1075-1080.

[17] HU J X, PENG J J, ZHANG W Q, et al. An intention multiple-representation model with expanded information[J]. Computer Speech & Language, 2021, 68: 101196.

[18] XU Q Q, PENG J J, ZHENG C Z, et al. Short text classification of Chinese with label information assisting[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(4): 1-19.

[19] 张换香, 彭俊杰. 基于方面级情感分析的深度语义挖掘模型[J]. 电子学报, 2024, 52(7): 2307-2319.
ZHANG H X, PENG J J. A deep semantic mining model based on aspect-level sentiment analysis[J]. Acta Electronica Sinica, 2024, 52(7): 2307-2319. (in Chinese)

[20] 廉筱峪, 夏楠, 戴高乐, 等. 复杂噪声环境下基于轻量化模型的车内交互语音增强和识别方法[J]. 电子学报, 2024, 52(4): 1282-1287.
LIAN X Y, XIA N, DAI G L, et al. An in-vehicle interaction speech enhancement and recognition method based on lightweight models in complex environment[J]. Acta Electronica Sinica, 2024, 52(4): 1282-1287. (in Chinese)

[21] PENG J J, WU T, ZHANG W Q, et al. A fine-grained modal label-based multi-stage network for multimodal sentiment analysis[J]. Expert Systems with Applications, 2023, 221: 119721.

[22] WANG L, PENG J J, ZHENG C Z, et al. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning[J]. Information Processing & Management, 2024, 61(3): 103675.

[23] LI Z Y, PENG J J, LIN X C, et al. Multimodal intent recognition based on text-guided cross-modal attention[J]. Applied Intelligence, 2025, 55(10): 690.

[24] ZHANG H L, XU H, WANG X, et al. A clustering framework for unsupervised and semi-supervised new intent

discovery[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 5468-5481.

[25] 苏建华, 池云仙, 许云峰, 等. 基于注意力模态融合的多模态意图识别[EB/OL]. (2024-01-10)[2024-11-18]. http://kns.cnki.net/KCMS/detail/detail.aspx?filename=JSJC20241114005&dbname=CJFD&dbcode=CJFQ.
SU J H, CHI Y X, XU Y F, et al. Multimodal intention recognition based on attention modal fusion[EB/OL]. (2024-01-10)[2024-11-18]. http://kns.cnki.net/KCMS/detail/detail. aspx?filename=JSJC20241114005&dbname=CJFD&dbcode=CJFQ. (in Chinese)

[26] HUANG S J, QIN L B, WANG B B, et al. SDIF-DA: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection[C]//ICASSP 2024 -2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2024: 10206-10210.

[27] ZADEH A, CHEN M H, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural-Language Processing. Stroudsburg: ACL, 2017: 1103-1114.

[28] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 1-8.

[29] MAI S J, HU H F, XU J, et al. Multi-fusion residual memory network for multimodal human sentiment comprehension[J]. IEEE Transactions on Affective Computing, 2022, 13(1): 320-334.

[30] TSAI Y H, BAI S J, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences[J]. Proceedings of the Conference. Association for Computational Linguistics. 2019, 1: 6558-6569.

[31] RAHMAN W, HASAN M K, LEE S W, et al. Integrating multimodal information in large pretrained transformers[J]. Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2020, 2020: 2359-2369.

[32] HAZARIKA D, ZIMMERMANN R, PORIA S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 1122-1131.

[33] WANG D, LIU S, WANG Q, et al. Cross-modal enhancement network for multimodal sentiment analysis[J]. IEEE Transactions on Multimedia, 2022, 25: 4909-4921.

[34] 樊琳, 龚勋, 郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. 电子学报, 2024, 52(7): 2341-2355.
FAN L, GONG X, ZHENG C Y. A multi-modal medical image analysis algorithm based on text guidance[J]. ACTA Electronica Sinica, 2024, 52(7): 2341-2355. (in Chinese)

[35] WANG D, GUO X T, TIAN Y M, et al. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis[J]. Pattern Recognition, 2023, 136: 109259.

[36] CHEN Y Z, ZHU W H, YU W L, et al. Prompt learning for multimodal intent recognition with modal alignment perception[J]. Cognitive Computation, 2024, 16(6): 3417-3428.

[37] WU S X, DAI D M, QIN Z W, et al. Denoising bottleneck with mutual information maximization for video multimodal fusion[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2023: 2231-2243.

[38] KENTON J D M W C, TOUTANOVA L K. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie. 2019, 1: 4171-4186.

[39] BAEVSKI A, ZHOU H, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[EB/OL]. (2020-10-22)[2024-11-18]. https://arxiv.org/abs/2006.11477v3.

[40] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[41] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning-ICML'06. New York: ACM, 2006: 369-376.

[42] ZHANG H Y, WANG Y, YIN G H, et al. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2023: 756-767.

[43] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Oakland: NIPS, 2017: 5998-6008.

[44] ZADEH A, ZELLERS R, PINCUS E, et al. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos[EB/OL]. (2016-08-12)[2024-11-18]. https://arxiv.org/abs/1606.06259v2.

作者简介

彭俊杰　男,1977年1月出生于湖北省黄冈市. 现为上海大学计算机工程与科学学院教授、博士生导师. 主要研究方向为数据挖掘与分析、自然语言处理. 在国内外发表学术论文130余篇.
E-mail: jjie.peng@shu.edu.cn

张换香　女,1979年9月出生于内蒙古包头市. 现为上海大学计算机工程与科学学院博士研究生. 主要研究方向为自然语言处理、情感分析.
E-mail: zhanghuanxiang@imust.edu.cn

李铮一　男,1999年9月出生于湖南省邵阳市. 现为上海大学计算机工程与科学学院硕士研究生. 主要研究方向为自然语言处理、多模态意图识别.
E-mail: lizhengyishu@gmail.com

王　兰　女,1997年出生于安徽省安庆市. 现为上海大学计算机学院博士研究生. 主要研究方向为自然语言处理、多模态情感分析.
E-mail: wanglan1997@shu.edu.cn