DOI:10.19651/j.cnki.emt.2416242

整合类内差异与类间关联的隐喻感情预测*

杨亚萍[1] 张敬源[2]
(1.石家庄铁道大学语言文化学院 石家庄 050043; 2.北京科技大学外国语学院 北京 100083)

摘 要: 隐喻感情预测有助于改进社交媒体内容的用户体验,同时在心理健康监测和虚拟心理治疗方面也具有潜在价值。此外,它还可以更精确地识别目标受众的感情需求,优化广告策略,提升商业效益。为了进一步提升情绪识别与情感检测的有效性,提出了一种整合类内差异与类间关联的多模式隐喻感情识别架构。首先引出三种单模式模型,包括视觉语义模型,文本语义模型和音频语义模型,从文本、视觉和音频三种不同的数据源中分别提取每个模式的个性化差异特征;随后引出一种深度层次多模式模型,通过中间层融合的方式对多模式之间的关联性进行学习,更好地利用双模式与三模式之间提供的互补信息;最后,基于决策层融合的方式将上述四种模型进行融合,在一种端到端的框架中,实现多模式隐喻感情的预测。在开源的数据集中进行的大量消融实验与对比实验证明了方法的有效性。

关键词: 隐喻感情预测;多模式模型;图像;文本;语音;模式识别
中图分类号: TN-9 **文献标识码:** A **国家标准学科分类代码:** 510.1050

Metaphorical affective prediction integrating intra-class differences and interclass associations

Yang Yaping[1] Zhang Jingyuan[2]
(1. College of Language and Culture, Shijiazhuang Tiedao University, Shijiazhuang 050043, China;
2. School of Foreign Languages, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: Metaphorical affective prediction can help improve the user experience of social media content, while also having potential value in mental health monitoring and virtual psychotherapy. In addition, it can more accurately identify the affective needs of the target audience, optimize advertising strategies, and improve business efficiency. In order to further enhance the effectiveness of metaphorical affective prediction, architecture on multi-mode metaphorical affective prediction method that consolidating intra-class difference and inter-class coherence is proposed. Firstly, three single-mode models are introduced, including image semantic model, text semantic model, and voice semantic model, to extract personalized differential features from three data sources, respectively. Then, a deep layering multi-mode model is introduced to learn the coherences between multiple modes through intermediate layer fusion, better utilizing the complementary information provided by bi-modal and tri-modal data. Finally, the four aforementioned models are fused using a decision-making layer fusion approach to predict multi-modal metaphorical feelings in an end-to-end architecture. Extensive ablation experiments and comparative studies conducted on open-source datasets have demonstrated the effectiveness of proposed approach.

Keywords: metaphorical affective prediction;multi-mode model;image;ext;voice;pattern recognition

0 引 言

隐喻感情预测主要包括情感检测和情绪识别两种任务,它们是与人类感知相关但有差异性的两种表现[1]。情感通常指的是对事物、事件或主题的一种主观态度或评价。它可以用来描述人类感情的正面或负面倾向,比如喜欢、厌恶、满意、不满等,因此更偏向于一般化的状态,更加稳定和持久。而情绪则是更具体的体验,它是一种强烈的且短期的反应,通常由特定的触发因素引起,比如恐惧、愤怒、快乐、悲伤等。因此,情绪更加情境化,更容易受到外界刺激

收稿日期:2024-06-16
* 基金项目:国家社会科学基金一般项目(19BYY226)、2023 年河北省高等学校英语教学改革(课程思政数字资源库)研究与实践项目(2023YYSZ026)资助

的影响,并且具有更强烈的生理和心理反应。

隐喻感情预测一般通过计算机技术来分析和识别文本、语音、图像等数据中的情绪和情感状态。旨在理解和推断人类情感和情绪,并从中提取有关感情倾向、感情强度、感情类别等方面的信息。由于与人类的切身体验和感受相关联,因此近年来,隐喻感情预测在各个领域都受到了广泛的关注。一般来说,它的应用主要包括但不限于[2]:

1)用户体验分析:企业可以使用相关模型来分析用户在社交媒体或者相关应用程序中通过文本、图像和语音等多种形式表达的感情,以帮助他们全面了解用户对其产品和服务的体验。通过从用户的不同情感和情绪反馈中提取关键点,企业可以改进产品,提高用户满意度,并根据用户实际需求量身定制个性化产品。

2)公众舆论分析:政府机关可以利用隐喻感情预测模型来分析公众对相关政策、公共服务或社会问题的态度。通过进一步的分析,政府可以获取公众舆论对相关政策的满意度调查,评估其决策的影响。这对提高公民参与度,促进我国治理体系和治理能力的现代化,保持我国健康平稳发展意义重大。

早期的隐喻感情预测方法主要集中在对书面文本的分析上,然而,这些单模式方法通常无法充分捕捉人类感情的复杂性与丰富性。近些年,随着社交媒体的兴起,显著增加了视频内容的渗透率,人们更愿意通过视频表达自己的个性和观点,其中的内容形式就包括了视频作者提供的图像和语音模式,以及用户在浏览时发表的个人见解与评论这种文本模式。相比之下,整合文本、图像和语音的多模式感情分析,显然可以提高预测系统的准确性和鲁棒性:因为每种模式往往传达特定与互补的信息,因此通过合成不同模式的信息,可以提供更全面、更细致的接近人类的感情理解。

多模式感情预测的结构可以分为两个主要组成部分:模式内分析和模式间分析[3-4]。前者着重于分析每个模式内部的交互作用。例如,对于文本模式来说,便涉及到分析句子中单词和短语之间的联系。后者处理不同模式之间的交互作用,并根据对齐情况进一步分类为同步交互或异步交互。同步的模式间交互发生在模式对齐的情况下,意味着不同模式在时间戳内同时出现。例如,当演讲者表达其观点时,伴随的文本可以精确同步以匹配讲述的内容。相反,异步的模式间交互指的是没有模式对齐的情况,即多种模式不会同时出现。在多模式感情预测中,主要的难点在于准确表示模式内信息,并有效捕捉模式间的关联性。因此,获得准确的模式内表达并找到最佳的融合方法来提取模式间相关且互补的特征至关重要。

隐喻感情预测强烈依赖于提取到的模式特征。例如,文本模式中不同单词和图像模式中不同区域往往反映着不同程度的感情水平。因此,从单模式数据中提取判别性的信息相当重要。注意力机制一般可用来突出时序数据中最重要的特征,因此引入该机制使得感情分析着重于情感和情绪相关的词语(对文本和语音数据而言)和区域(对图像数据而言)。此外,双模式和三模式数据之间存在着很强的相关性。比如,若只有单模式数据,此时看完电影后表述的"这部电影还行"这句话可能有歧义,但是当存在双模式数据时,伴随用户"微笑"的表情,可以视做正向反馈,即积极的感情。相反,若配上一个"皱眉"的表情会被认为是负面感情。在三模式的例子中,结合文本和视觉模式与语音模式,如"声音较大",则会加剧感情的表现。这3种类型的特征相互补充,可用于全面的感情分析。因此,在多模式感情分析中,将模式表达的特定信息和共享的模式信息结合起来相当重要。基于此,本文引入注意力机制并结合深度神经网络的优势,提出了一种整合类内差异与类间关联的多模式隐喻感情识别架构(architecture on multi-mode metaphorical affective prediction method that consolidating INTra-class difference and INTer-class coherence,INT2-M3AP)。具体来说,INT2-M3AP模型首先采用3个独立的单模式,从文本、图像和音频中提取具有区分性的模式内特征,同时预测单模式表达的特定感情。其中的文本语义模型捕捉显著性的感情词语,视觉语义模型聚焦于与感情相关的图像区域,音频语义模型动态提取和学习反映感情相关的音频特征。随后利用基于中间层融合的深度层次多模态模型(deep hierarchical multimodal model)捕捉3种不同模式之间的互补信息,分层次学习双模式和三模式特征的模式间关联。最后基于决策层融合的方式将所有4个模型组合成一个全面的框架,实现最终的感情分类。INT2-M3AP模型采用了混合融合策略(hybrid fusion strategy)[5],结合了特征层融合和决策层融合技术的优势。具体来说,中期的特征层融合可以在不同多模式的特征之间引入关联性,而后期的决策层融合则克服了来自不同分类器的非相关误差,从而比单分类器产生更好的结果。此外,由于层次模型对输入数据变化和噪声更加鲁棒,能够提取更加抽象和高级的特征,因此基于深度层次的多模式框架INT2-M3AP,可用于结构化地学习复杂表征。综上,从统计学习的理论角度来看,本文方法可以进行全面而细致的针对感情预测的模式识别任务。

1 相关方法

目前的相关研究大都基于文本进行感情预测[6-7],与之相比,基于音频的感情预测在近些年来才崭露头角[8],而基于视觉的感情预测相对最具挑战性,因为其涉及的主观性和抽象程度要高得多[9]。而由于其相较单模式感情预测的有效性,在过去的几年中多模式感情预测已经收到了越来越多的关注[10-13]。以下部分将详细介绍基于单模式和基于多模式(包括双模式与三模式)的感情预测方法。

1.1 单模式方法

1)基于文本信息的感情预测

基于文本信息的感情预测方法是自然语言处理

(natural language processing,NLP)领域中的热点。它可以分为两大类:基于字典的方法[14]和基于机器学习的方法[15]。基于字典的方法通过使用单词中表达的情感极性(sentiment polarities)来确定文本表达的感情。例如,语义取向计算器(semantic orientation calculator,SO-CAL)[14]构建了一种用于文本感情预测的字典,其中每个单词都有一种语义取向标签。文献[15]结合了监督和无监督技术,提出了一种可学习感情信息的向量空间模型。除了这些方法之外,深度学习方法也可用于文本感情的有效分类。文献[16]提出使用深度递归神经网络进行语句层面的感情主观性检测。文献[17]引入了一种基于 n-gram graph-cut 和长短期记忆网络(long short-term memory,LSTM)的混合方法。文献[18]提出了 RoBERTa-LSTM 模型,该模型结合了 Transformer 和序列模型的优势,用于感情预测。

2)基于音频信息的感情预测

基于音频信息的感情预测研究集中在提取情绪和情感相关的声学特征,如语音强度、音高、持续时间和带宽[19]。文献[20]利用高斯混合模型,基于梅尔倒谱系数(mel frequency cepstrum coefficient, MFCC)、音调和韵律等声学特征得到了 98%的识别准确率,还表明了加入音频特征的方法优于不加入音频特征的方法。文献[21]提出了一种用于语音情感识别的深度学习模型。他们结合了双向长短期记忆(bidirectional long short-term memory,Bi-LSTM)、卷积神经网络(convolutional neural network,CNN)和注意力机制,利用从语音数据中提取的对数-梅尔谱图来捕捉情感表征。

3)基于视觉信息的感情预测

与上述两种单模式方法对比,基于视觉的感情预测仍在曲折中快速发展。此研究的主要目的是通过检测和分析面部表情和身体特征传达出的情绪和情感,通常分为三类。

基于低级特征的方法。如文献[22]中,将图像的低级特征与上下文信息融合,提出一种基于主题的情绪识别方法。

基于中级特征的方法。如文献[23]采用传统的机器学习方法检测 1 200 个情感相关的形容词-名词对(adjective-noun pairs,ANP),并使用视觉情感本体(visual sentiment ontology,VSO)和 SentiBank 引入了一种新颖的中级特征表示。

基于高级特征的方法。如文献[24]利用预训练的 AlexNet 作为视觉情感分析的特征提取器。此外,文献[25]也采用 AlexNet 架构,并使用渐进式方法对其进行微调,以减轻视觉情感分析中噪声的影响。文献[26]采用了一种注意力机制来自动确定可用于视觉情感分类的相关图像区域。文献[27]结合了 3 种分类模型,包括定制化 CNN, InceptionV3 和 ResNet50,提出了一种基于集成学习的面部情感识别方法。与单独的分类器相比,在识别人脸表现出的中性和积极的情绪方面显示出了良好的性能。

1.2　多模式方法

多模式方法相对于单模式方法具有更高的性能,因此在学术界和工业界引起了广泛的关注。多模式数据的存在使得可以从视觉、音频和文本等多个不同角度全面理解感情预测。本节介绍了关于多模式感情预测的相关工作,包括双模式(音频-视觉、音频-文本、视觉-文本)和三模式的研究。

1)基于音频-视觉信息的感情预测

受到深度学习在不同模式识别任务中的成功启发,文献[28]提出了一种端到端的音频-视觉情感检测框架,使用经过预训练的 ResNet-50 网络作为特征提取器分别处理语音和视觉模式。将提取的特征进行简单的串联,并作为输入传入两层 LSTM 网络进行端到端训练。为了有效地组合音频和视觉特征向量,文献[29]引入了一种新颖的特征层融合方法,基于线性池化理论(linear pooling theory),动态捕捉模式之间的关联。为了提高多模式情感检测的性能,文献[30]引入了两种包括特征层和模型层的融合策略。他们使用 OpenSmile 工具包、CNN-BLSTM、CNN 和 LBP-TOP 模型,对从音频和视觉通道提取的信息进行融合。文献[31]提出了一种创新的多模式深度学习方法,用于实时视频流情感分类。该方法结合了音频-视觉输入流的特征,并使用 4 个紧凑的深度神经网络模型同时分析视觉和听觉数据。文献[32]提出了一个多模式情感分析系统,利用面部表情和语音识别来提高情感识别的准确性。该研究介绍了 3 种情感检测模型,包括 CNN-SVM、RNN-SVM 和 CRNN-SVM,其中 CRNN-SVM 模型在情感分类中达到了最高的识别率。

2)基于音频-文本信息的感情预测

近年来,注意力机制由于可用于在模式之间建立互补连接而受到了广泛关注。文献[33]提出了一种基于自注意力机制的特征层融合技术,用于对文本和音频模式进行情感分类。随后,文献[26]提出了一种音频-文本情感检测技术。使用了深度卷积神经网络(deep convolutional neural network,DCNN)进行声学特征提取。对于文本特征,考虑了双向 LSTM 和 DCNN 两个并行的分支,并进行了中间层融合。文献[34]提出了一种基于注意力机制方法,以建立语音帧和文本单词之间的对齐,旨在提取更可靠的多模式特征表示。文献[35]引入了一种针对多模式情感分析的改进 BERT 模型,其中优化了 BERT 的内部结构并融入了非语言信息。该模型包括 3 个模块:门控通道、分层多头自注意力和基于自注意力过程的张量融合,它在 CMU-MOSI 数据集上的表现优于原始的 BERT 模型和传统模型,包括 Transformer 和 LSTM。

3)基于视觉-文本信息的感情预测

在情感和情绪预测方面,对视觉和文本模式的融合也进行了大量研究工作。文献[4]提出了一种混合的深度注意融合方法,结合了视觉-文本模式之间的判别性特征和相

关性特征,用于感情预测。文献[36]提出了一种探索文本和图像模式之间连接的多模式自适应情感检测方法。文献[37]首先利用预训练的 CNN 和词嵌入来提取图像和文本特征,随后提出了一种基于注意力的框架,包括视觉-语义注意力和语义自注意力模块,用于跨图像和文本模式学习相关特征进行情感检测。对于在线新闻的情感识别,文献[38]设计了一种基于布局(layout-driven)的多模式注意力方法。该方法利用在线新闻的布局将图像与相应的文本对齐,并结合了图像和文本模式的多模式特征。近些年,文献[39]提出了一种基于迁移学习和集成学习的加权 CNN 混合模型,用于多模式感情预测。该方法利用预训练的 VGG16 和 Mask-RCNN 模型分别提取视觉特征并检测物体,同时利用 BERT 提取文本特征。与其他方法相比,这种混合模型取得了更优异的性能。类似地,文献[40]引入了一种图像-文本交互网络(image-text interaction network,ITIN),用于多模式感情预测。ITIN 模型旨在探索文本和情感相关的图像区域之间的关系,其中包括一个跨模式对齐模块用于检测区域-词语对应关系,一个跨模式门控模块用于处理不匹配的区域-词语对,以及一个上下文特征表示来提高预测准确性。

4)基于音频-视觉-文本信息的感情预测

与双模式感情预测相比,涉及文本、音频和视觉信息的多模式感情预测研究相对有限。文献[41]提出了一种从多个数据源中提取情感的方法,其中采用了决策层融合策略,将从各种模式提取的与情感相关的信息进行整合。文献[42]提出了一种利用 3 个分别的单模式神经网络从文本、视觉和音频模式提取特征的方法,随后通过张量融合将这些特征组合起来。文献[43]提出使用多核学习进行多模式情感分析。该框架将文本、视觉和音频模式整合在一起,从不同的数据集中提取情绪特征。类似地,文献[44]引入了一种分层特征融合技术,首先将两种模式组合起来,然后再加入第 3 种模式。此外还结合了上下文建模技术。文献[45]提出了一种基于深度学习的特征层融合框架,用于多模式感情预测,它们利用 CNN 提取文本和视觉特征,而 openSmile 工具包用于提取音频特征。多模式感情预测主要的挑战之一是如何在融合不同模式之前挖掘上下文信息和提取重要的模式间特征。文献[46]针对此问题,提出了一种多层次提取上下文特征,并利用 BiRNN 注意力融合方法捕捉模式间关键特征的方法。

2 本文方法

本文提出的 INT2-M3AP 架构如图 1 所示。

该框架由 4 个不同的组件组成,其中包括 3 个单模式模型,即视觉语义模型、文本语义模型、音频语义模型,以及一个深度层次多模式模型。视觉和文本语义模型专注于提取具有判别性的图像区域特征和关键的文本词语,而音频语义模型动态学习与情感相关的音频特征。深度层次多模式网络以分层的方式利用所有 3 个模式之间的内部关联。它首先通过将单模式融合为双模式组合来学习双模式间的关联。然后,通过将所有 3 个双模式连接为三模式来获取三模式间的关联。最后,使用决策层融合将所有 4 个模型的输出进行整合。

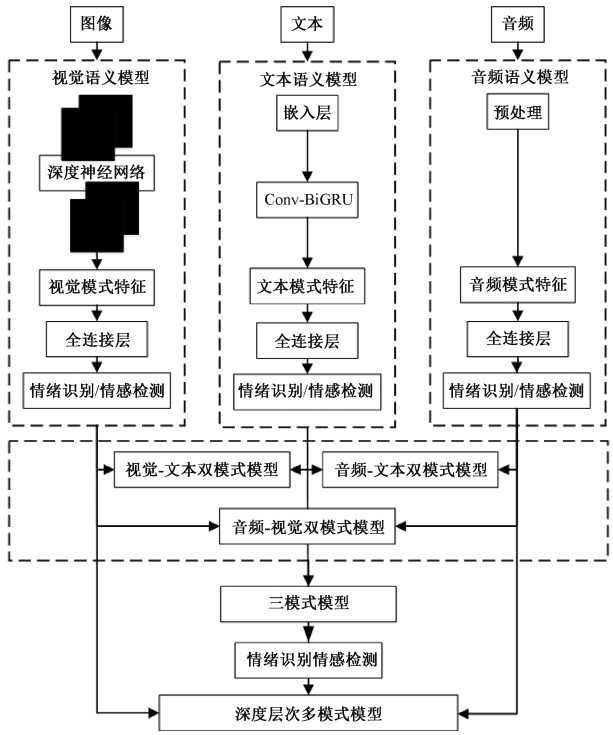

图 1 INT2-M3AP 总体架构

Fig. 1 The overall architecture of INT2-M3AP

2.1 视觉语义模型

通常情况下,某些视觉区域特征与感情的关联性更强。如果能够获取到这些关键视觉特征,就能更成功地进行感情预测任务。本文提出的视觉语义模型,能以端到端的方式自动关注显著的视觉区域特征,并分析视频模式中的情感和情绪。大体上,在一个视频或其片段中出现的相同面部表情和手势,代表着特定的情感或情绪,提取关键帧以代表整个视频或特定片段的情绪是相当有必要的。

本文假设基于面部表情的可变性,对一系列视频进行总结,用一个或几张特定的关键帧来表示整个视频,可以有效地进行感情预测。每个视频根据其对应的情绪或情感被划分为多个范围,根据其起始和结束范围,获取并保留可代表特定情绪或情感的图像帧。要处理获得的帧序列(共计 756 000 帧)需要大量软硬件资源,并可能引发与其他现有模式的对齐问题,因此本文对每个片段保留唯一的关键帧进行感情表征。

总结来说,本文收集了判别性的图像帧,即每个视频片段对应的一个关键帧,最终将视觉信息与其他数据模式对齐。这样降低了模型的复杂性,得到了计算高效的范式。

给定一组 k 个图像,$V=\{V_1,V_2,\cdots,V_i,\cdots,V_k\}$。对

于每个图像 V_i，使用深度预训练模型提取判别性的图像区域特征 $R_i=\{r_i^1,r_i^2,\cdots,r_i^3,\cdots r_i^M\}\in \mathrm{R}^{\mathrm{M}\times\mathrm{N}}$，如式(1)：

$$R^i=f_v(V_i;\theta_v),R_i\in \mathrm{R}^{\mathrm{M}\times\mathrm{N}} \tag{1}$$

其中，θ_v 表示深度模型参数，M 是图像区域的数量，N 是每个图像区域的特征维度。

使用预训练模型并通过微调(fine-tuning)来提取有效的特征，而并不从头设计和训练整个模型，因此计算过程更高效[47]。具体来讲，模型根据目标类别和数据集进行微调。对于情感检测任务，有 3 个类别，即积极、消极和中性。同时，对于情绪识别任务，有 4 个类别，即愤怒、快乐、悲伤和中性。最终，提出的模型可以对视觉感情分类，如式(2)：

$$Y_i^{(v)}=f_r(R_i;\theta_r^{(v)}),Y_i\in R^C \tag{2}$$

其中，$\theta_r^{(v)}$ 表示全连接层(fully connected layer，FC layer)的参数，C 表示感情类别的数量。

2.2　文本语义模型

类似于图像区域特征，文本模式中的某些词与情绪的相关性更强。近年来，注意力机制被越来越多地应用于对关键语义词的关注方面，并且在各种与 NLP 相关的任务中证明了其价值，如文本情感预测[48]、机器翻译[49]等。与以前的工作不同，本文的文本语义模型以端到端的方式突显文本模式中与情绪或情感高度相关的词语。

为了获得更好的感情预测结果，首先对文本数据进行预处理以消除噪声。令 $T=\{T_1,T_2,\cdots,T_i,\cdots T_k\}$ 表示通过嵌入层传递的经过预处理的 k 个文本序列的集合，其中 T_i 表示实值向量表征。对于每个文本 T_i，使用预训练的词嵌入将词序列转换为密集向量空间 $S_i=\{S_i^1,S_i^2,\cdots,S_i^j,\cdots,S_i^L\}\in R^{L\times N}$，以获取丰富的语义特征，如式(3)：

$$S_i=T_i\boldsymbol{W}_t,S_i\in R^{L\times N} \tag{3}$$

其中，$\boldsymbol{W}_t$ 表示参数矩阵，L 表示文本长度，N 是嵌入维度。在上述向量空间中，S_i^j 表示第 i 个文本文档中的第 j 个词。若单独获取每个词的语义信息不足以表示完整的句子，仍然存在上下文词语之间缺乏关联的问题。因此，在每个步骤 j 中，将 S_i^j 输入至 Conv-BiGRU 模型[26](即 CNN 和双向 GRU 的组合)，以挖掘更抽象的语义特征 $X_i=\{x_i^1,x_i^2,\cdots,x_i^j,\cdots,x_i^L\}$，如式(4)：

$$X_i=f_s(S_i,\theta_s),X_i\in R^{L\times R} \tag{4}$$

其中，θ_s 表示 Conv-BiGRU 模型的参数，R 表示 Conv-BiGRU 层的维度。

并非所有特征的贡献程度皆相同，所以将特征向量 X_i 进一步输入至注意力层，以突出最重要的感情语义特征，从而得到经过降维的一组模式内注意力特征。注意力机制根据单词 X_i^j 对感情预测任务的重要性为每个词分配一个注意力强度 a_i^j。a_i^j 是一个非标准化的注意力权重，它可以评估单词 $\{X_i^j\}$ 在文本中与感情相关的程度，并使用 softmax 激活函数计算，如式(5)：

$$\alpha_i^j=\frac{\exp(e_i^j)}{\sum_{j=1}^{L}\exp(e_i^j)} \tag{5}$$

$$e_i^j=\sigma(Wx_i^j+b) \tag{6}$$

其中，W 和 b 是权重和偏置参数。$\sigma(.)$ 表示非线性 tanh 激活函数，有助于增强特征之间的非线性关系。给定 α_i，可以规范化注意力强度对所有词序列 $\{x_i^j\}_{1\leqslant j\leqslant L}$ 的影响。因此，可以通过对单词特征进行加权平均来计算显著性语义特征，如式(7)：

$$z_i^j=\sum\nolimits_{1\leqslant j\leqslant L}\alpha_i^jx_i^j,Z_i\in R^R \tag{7}$$

式(8)表示获取文本注意力特征的完整过程：

$$Z_i=f_b(X_i;\theta_b(t)),Z_i\in R^R \tag{8}$$

其中，$\theta_b(t)$ 是式(6)中的参数集。注意力语义特征映射 Z_i 是与感情相关的文本特征的全面表征，进一步引入多个全连接层以构建感情分类器，如式(9)：

$$Y_i(t)=f_c(Z_i;\theta_c(t)),Y_i\in R^C \tag{9}$$

其中，$\theta_c(t)$ 表示 FC 层的参数，C 表示感情类别的数量。

2.3　音频语义模型

音频语义模型旨在学习在音频模式中最有助于预测感情的特征。在提取模式内特征之前，通过从音频文件的每个注释片段构建音频向量来对数据进行预处理。

令 $A=\{A_1,A_2,\cdots,A_i,\cdots,A_n\}$ 表示 n 个音频集合，它作为浮点时间序列以 44.1 kHz 采样率进行加载，其中每个 A_i 为独立的音频文件。每个 A_i 根据片段的起始和结束时间戳截断，并以音频向量的形式保存为 $\boldsymbol{V}_i=(v_i^1,v_i^2,\cdots,v_i^j)$。每个音频向量 $\boldsymbol{V}_i$ 包括截断的音频信号，进一步处理以提取感情相关的音频特征 $X_i=(x_i^1,x_i^2,\cdots,x_i^j,\cdots,x_i^L)$，其中 L 表示特征总数。

为了能更好地表征各种情绪和情感，本文计算了多种语音信号特征，包括均值、标准差、MFCC、谱质心(spectral centroid)、过零率(zero-crossing rate，ZCR)、谱带宽(spectral band width)、自相关特征、色度以及音高等其他特征：

1)均值。均值是最常用的语音信号特征之一，它计算信号平均值，如式(10)：

$$\mu=\frac{1}{N}\sum\nolimits_{i=0}^{N-1}x_i \tag{10}$$

其中，μ 表示均值，x_i 表示信号处于位置 i 的实值，N 表示音频信号的总长度。

2)标准差。标准差是衡量语音信号与其均值之间波动的特征，如式(11)：

$$\sigma=\sqrt{\frac{1}{N-1}\sum\nolimits_{i=0}^{N-1}(x_i-\mu)^2} \tag{11}$$

其中，σ 表示标准差，x_i 表示信号处于位置 i 的值，N 表示音频信号的总长度。

3)MFCC。MFCC 将标准频率转换为梅尔刻度，将特征缩放为人类可以感知的范围，如式(12)：

$$MFCC_k=\sum_{\theta=1}^{M}X_\theta\cos\left(k(\theta-1)\frac{\pi}{M}\right),k=1,2,\cdots,N \tag{12}$$

其中,N 表示倒谱系数的个数,用于将实际信号从滤波器中分离出来。X_θ 是第 θ 个滤波器的对数能量输出,其中 θ 的范围从 1 到 M,表示三角带通滤波器的序号。

4)谱质心。谱质心是一种用于定位频谱重心的度量,其利用了短时傅里叶变换的幅度和频率信息,如式(13):

$$SpectralCentroid_i = \frac{\sum_{i=0}^{k} kX_i[k]}{\sum_{i=0}^{k} X_i[k]} \tag{13}$$

其中,$X_i[k]$ 表示第 i 帧在频率 bin$k$ 处的傅里叶变换幅度。谱质心通常与声音的亮度相关,较高的值表示能量更集中在更高的频率中。

5)过零率(zero-crossing rate,ZCR)。过零率通过计算信号在特定音频片段内的零交叉次数来衡量声音的平滑程度。换句话说,它是信号从正向到负向或从负向到正向的符号变化率。在平滑信号中,其值较低,而在嘈杂信号中其值较高,如式(14):

$$ZCR = \frac{1}{M-1}\sum_{k=0}^{M-1} II\{s_k s_{k-1} < 0\} \tag{14}$$

其中,s_k 表示长度为 k 的信号,II 为指示函数。

6)谱带宽。谱带宽是指光带在最大峰值的一半处的宽度。它计算了 n 阶谱带宽,如式(15):

$$SpectralBandwidth = (\sum_t S(t)(f(t)-f_c)^n)^{1/n} \tag{15}$$

其中,$S(t)$ 和 $f(t)$ 分别表示频率 bin$t$ 处的谱幅和频率,f_c 表示谱质心。当 $n=2$ 时,它类似于加权标准差。

7)自相关。自相关定义了信号与其滞后版本之间的相似程度。它还可用于识别信号中的重复模式。信号 s 的自相关 α 表示如下:

$$a(t) = \sum_k s(k)s(k-t) \tag{16}$$

其中,t 表示滞后参数,$\alpha(t)$ 在 $t=0$ 时有最大值。

8)色度。色度特征是一个包含十二个元素的特征向量,每个元素对应音高在信号中的能量。

9)音高。音高可以通过利用语音信号的频率尺度来确定。较高的频率表示高音调,较低的频率表示低音调。

如上提取的特征可以充分表征情绪和情感相关的音频特征,进一步输入到多个 FC 层中,以构建音频模式的感情分类器,表示为:

$$Y_i(\alpha) = f_k(X_i;\theta_k(\alpha)), Y_i \in \mathrm{R}^{\mathrm{C}} \tag{17}$$

其中,$\theta_k(\alpha)$ 为 FC 层的参数集合,C 表示情绪类别的数量。

2.4 深度层次多模式模型

深度层次多模式模型旨在以统一的方式融合 3 种不同的模型,以获得 3 种模式之间的内部相关性。传统合并单独模型的方法之一是对每个模型的预测结果仅应用决策层融合,但这种方式并不能有效地学习异构模式之间的内部相关性。本文采用中间层融合方案来集成上述的 3 种单模式语义模型,以层次化方式充分利用所有 3 种模式之间的内部关联性。首先,对于每个双模式组合,整合其内部模式特征,然后将所有双模式特征融合为三模式特征。通过这种方式,可以学习更准确的 3 种模式之间的双模式和三模式相关性,以获得更准确的感情预测结果。

1)双模式融合。上述单模式的特征向量首先被组合成 3 种不同的双模式,包括音频-视觉、音频-文本和视觉-文本。

对于音频-视觉双模式,学习到的模式内声学语义特征 X_i 和视觉语义特征 R_i 融合如下:

$$B_i^{av} = (X_i \oplus R_i), B_i^{av} \in R^D \tag{18}$$

然后将其作为输入传递给多个 FC 层,以挖掘音频和视觉模式特征之间的相关性,进行双模式感情预测:

$$Y_i^{(av)} = f_b(B_i^{av};\theta_b^{(av)}), Y_i^{(av)} \in R^C \tag{19}$$

其中,$\theta_b^{(av)}$ 表示 FC 层的参数,C 表示输出类别的数量。

类似地,对于音频-文本双模式,学习到的模式内声学语义特征 X_i 和文本语义特征 Z_i 融合如下:

$$B_i^{at} = (X_i \oplus Z_i), B_i^{at} \in R^D \tag{20}$$

然后将其作为输入传递给多个 FC 层,以挖掘音频和文本模式特征之间的相关性,进行双模式感情预测:

$$Y_i^{(at)} = f_b(B_i^{at};\theta_b^{(at)}), Y_i^{(at)} \in R^C \tag{21}$$

其中,θ_b^{at} 为 FC 层的参数,C 表示类别数量。

类似地,对于视觉-文本双模式,学习到的模式内图像特征 R_i 和文本特征 Z_i 融合如下:

$$B_i^{vt} = (R_i \oplus Z_i), B_i^{vt} \in R^D \tag{22}$$

然后将其作为输入传递给多个 FC 层,以挖掘视觉和文本模式特征之间的相关性,进行双模式感情预测:

$$Y_i^{(vt)} = f_b(B_i^{vt};\theta_b^{vt}), Y_i^{(vt)} \in R^C \tag{23}$$

其中,$\theta_b^{(vt)}$ 为 FC 层的参数,C 表示类别数量。

自此,每个双模式都获得了一个细粒度连接,从而提供了一组更抽象的跨模式特征向量,进一步帮助获得三模式之间的相关性。

2)三模式融合。为了更好地挖掘所有 3 种模式之间的相互关系,进一步融合了上述得到的情绪和情感相关的双模式特征向量 B_i^{av}、B_i^{at} 和 B_i^{vt}:

$$T_i^{avt} = (B_i^{av} \oplus B_i^{at} \oplus B_i^{vt}), T_i^{avt} \in R^D \tag{24}$$

然后将其作为输入传递给多个 FC 层,以获取 3 种双模式之间的内部相关性,改进多模式感情预测:

$$Y_i^{(avt)} = f_t(T_i^{avt};\theta_t^{(avt)}), Y_i^{(avt)} \in R^C \tag{25}$$

其中,$\theta_t^{(avt)}$ 为 FC 层的参数,C 表示类别数量,融合后的向量 $Y_i^{(avt)}$ 表示与情绪和情感相关的三模式特征。整个模型以深度层次的方式进行训练,分析所有模式之间的相互关系,构建更准确的多模式感情预测系统。

这种单纯中间层的融合方案由于信息重复或者不同模式的特征可能在表征的尺度、分布等方面存在不一致性差异。而机器学习中对于信息冗余的做法一般是特征选择

(如卡方检验、互信息等)或者降维(主成分分析、线性判别分析),本质上是一种信息冗余和信息损失之间的权衡。本文为了应对这种可能出现的信息冗余,在决策层通过对视觉语义模型、文本语义模型、音频语义模型和深度层次多模式模型的预测结果进行多数投票,并在没有达成一致性意见时根据置信度最大原则进行预测。

2.5　基于决策层融合的感情预测

本节将上述 4 个模型(包括文本语义模型、视觉语义模型、音频语义模型和深度分层多模式模型)通过决策层融合成一个综合框架,以得出最终的预测结果。决策层融合是一种将不同性质的特征集成的策略,如下所示:

$$Y_i = MajorityVoting\{Y_i^{(v)}input:(V_i)$$
$$Y_i^{(t)}input:(T_i)$$
$$Y_i^{(a)}input:(A_i)$$
$$Y_i^{(avt)}input:(A_i,V_i,T_i)\}$$

即最终的预测结果 Y_i 是通过多数投票方案从视觉语义模型(V_i)、文本语义模型(T_i)、音频语义模型(A_i)和所有 3 个模型的组合(A_i,V_i,T_i)中计算得出的。如果大多数模型对于某个类别没有达成一致意见,则根据置信度最大的模型选择最终的预测结果。

3　实验结果及分析

3.1　数据集

为了验证 INT2-M3AP 模型的有效性,基于开源准数据集 IEMOCAP[50] 进行多模式感情预测的验证。IEMOCAP 数据集是一个交互式二元多模式语料库,其中包含 20 个月收集到的音频、视频、转录和运动捕捉数据。它源于 10 名演员,男女比例平均,由 5 次即兴表演和脚本表演构成,并由 3 名标注者对情绪类别(包括悲伤、快乐、愤怒、中性、沮丧、兴奋、惊讶、厌恶、恐惧和其他)进行注释,由两个标注者对情绪价值(消极与积极)、支配程度(弱与强)和活跃程度(平静与兴奋)进行注释。对于情绪识别任务,便于与其他方法公平比较,考虑愤怒、快乐、悲伤、和中性类别的多数一致(即 3 个标注者中至少有两个认可同一标签)的话语(utterance)。另一方面,对于情感检测任务,考虑 IEMOCAP 的情感效价,利用 likert5 级评分法进行评估,其中的分数范围从 1(消极)到 5(积极)。简单起见,在每个话语上提供了平均情感效价。由于情感效价标签中存在严重的类不平衡,数据被划分为更有效的消极、中性和积极情感类别,范围分别为 1～2.5、2.5～3.5 和 3.5～5。如表 1 和 2 所示,分别显示了情绪和情感类别的统计信息。

表 1　每个情绪类别中的话语数量

Table 1　Number of utterances in each emotion category

愤怒	快乐	悲伤	中性	总数
1 103	1 636	1 084	1 708	5 531

表 2　每个情感类别中的话语数量

Table 2　Number of utterances in each sentiment category

消极	中性	积极	总数
1 633	2 010	1 888	5 531

与 CMU-MOSI 和 CMU-MOSEI 等其他数据集相比,IEMOCAP 数据集中的感情环境更复杂,侧重于自发情感和情绪表达,为感情预测模型提供了更微妙且接近人类感知表达的数据来源。此外,IEMOCAP 数据集中包含来自不同性别、文化和年龄的演讲者,因此更加具有多样性。该数据集中使用的注释方法更加复杂,涉及多个人类标注者,可以减少标注噪声以得到更精确的注释结果。综上,IEMOCAP 数据集为目前多模式感情预测研究的基准数据集。

3.2　实验设置

本文方法包括 4 个不同的模型:3 个模式内个性化模型和一个双模式和三模式协同模型。在训练多模式模型之前,每个单模式都经过了单独的训练,以获取特定模式内的特征和预测结果。为了处理数据中的噪声,采用了各种预处理技术,包括数据清理、归一化和特征缩放,以减少可能降低模型准确性的冗余信息来提高数据质量。

在视觉语义模型中,首先将提取的关键帧缩放为 224×224 的大小,保留 RGB 通道,然后进行相应预处理。随后,将这些图像输入给在 ImageNet 上预训练的 VGG19 网络,进行视觉语义特征提取。最后一层的卷积层特征表征了图像区域,维度为 196×512。这样一来,视觉语义模型中的每个图像可以表示为维度为 512 的 196 个区域特征。

在文本语义模型中,单词序列的最大长度设定为 100。对于长度较短的序列,采用零填充策略,而对于超过 100 个单词的序列,则进行截断处理。对于文本特征,应用预训练的 GloVe 嵌入将序列中的每个单词转换为 300 维的向量。BiGRU 的隐层状态分别设置为 360 和 230 个神经元,而卷积层的特征图个数为 12。

在音频语义模型中,构建音频向量时,对原始波形以 44.1 kHz 的频率进行采样。使用 Librosa 音频处理库计算每个音频段的 17 个不同声学特征。全连接层的结构设置为 96～64,激活函数为 ReLU。

深度层次多模式模型将所有单模式特征组合起来,包括视觉语义模型产生的 25 088 维视觉语义特征,文本语义模型产生的 12 维文本语义特征,以及音频语义模型产生的 64 维音频语义特征。在融合后,形成一个 25 100 维的视觉-文本特征向量,一个 25 152 维的音频-视觉特征向量和一个 76 维的音频-文本特征向量。之后,3 种双模式特征向量合并为一个 50 328 维的三模式特征向量。

软硬件环境方面,本文模型基于 Python 语言,以及 Keras 和 Tensorflow 框架实现。所有实验都在一台搭载

64 GB 主内存和两个 NVIDIA GeForce RTX 2080 Ti(每个 GPU 的内存为 22 GB)的 HP 28 G4 工作站上进行。在训练之前,所有情绪和情感类别都转换为二进制数组。因此,每个模型的输出层使用 sigmoid 激活函数实现。将数据集按 8∶1∶1 的比例划分为训练集、测试集和验证集。使用 Adam 和 RMSprop 作为优化器,学习率为 0.001。应用参数为 0.000 1 的 L2 正则化以避免过拟合。每个模型的训练 batch 为 32 并使用二元交叉熵作为损失函数。在训练时,通过监测验证集损失,应用 patience 为 30 的早停策略,防止模型过拟合。

3.3 实验结果及其分析

1)情绪识别结果对比与分析

INT2-M3AP 框架在 IEMOCAP 数据集上进行的情绪识别任务的结果如表 3 所示。

表 3 单模式、多模式与本文方法的情绪识别结果对比

Table 3 Comparison of emotion recognition results among unimodal, multimodal, and the proposed method

方法	模型	准确率/%	召回率/%	精准率/%	F1 分数
单模式	视觉语义模型	96.04	96.30	95.55	95.92
	文本语义模型	84.43	80.72	84.75	82.30
	音频语义模型	77.79	73.99	77.92	75.90
多模式	深度层次多模式模型	91.20	92.21	90.00	91.09
单模式+多模式	INT2-M3AP	93.0	94.47	92.16	93.07

结果表明,文本语义模型的准确率为 84.43%,优于音频语义模型的 77.79% 准确率。而视觉语义模型以 96.04%的准确率优于其他两个单模式模型。深度层次多模式模型优于除了视觉语义模型之外的两种单模式模型,准确率为 91.20%,这突显了将多模式信息用于情绪分类的有效性。此外,提出的 INT2-M3AP 框架同样优于除了视觉语义模型之外的所有其他模型,准确率达到 93%。此外,可以看到,召回率,精准率和 F1 分数的结果与准确率类似。

如表 4 所示,呈现了不同类别的情绪识别准确率,表明"愤怒"情绪的分类效果优于其他情绪。

表 4 不同类别的情绪识别结果对比

Table 4 Comparison of emotion recognition results across different categories

情绪类别	准确率/%
愤怒	96.80
开心	90.79
悲伤	91.56
中性	85.64

还比较了在不同预训练模型上训练的视觉语义模型的结果,如表 5 所示。公平对比期间,皆采用了基于 ImageNet 数据集的预训练版本。

与其他情绪识别的基准方法结果对比,如表 6 所示,可以看到提出的方法比基于单模式和多模式方法的结果都要优秀。

2)情感检测结果对比与分析

情感检测结果同样基于 IEMOCAP 数据集进行评估,如表 7 所示。

表 5 不同视觉语义模型的情绪识别结果对比

Table 5 Comparison of emotion recognition results among different visual semantic models

模型	准确率/%	召回率/%	精准率/%	F1 分数
VGG16	96.04	98.11	93.86	95.93
VGG19	96.11	98.48	93.62	96.00
ResNet18	92.47	96.15	88.84	92.35
ResNet34	94.85	97.33	92.23	94.71
ResNet50	95.62	97.72	93.39	95.50
InceptionV3	91.02	94.23	87.6	90.80

结果表明,所提出的 INT2-M3AP 模型与其他除了视觉语义模型的消融模型相比取得了更好的结果,准确率达到了 85%。74.27% 准确率的文本语义模型比准确率为 68.47%的音频语义模型表现更好,而视觉语义模型的表现优于其他两种单模式模型,准确率达到 90.47%。此外,可以看出,除了视觉语义模型之外,所有这些单模式模型的性能都低于深层次多模式模型,后者分类情感时的准确率达到了 83.59%。此外,可以看到,召回率,精准率和 F1 分数的结果与准确率类似,同样证明了整合多模式数据的重要性。

如表 8 所示,呈现了不同类别的情感检测准确率,表明"积极"情感的分类效果优于其他情感。

如表 9 所示,对使用不同预训练模型的视觉模型情感检测结果进行了比较。公平对比期间,皆采用了基于 ImageNet 数据集的预训练版本。

如表 10 所示,将所提出的方法与最先进的方法进行了比较。结果确认了除了音频模式外,所提出的模型在单

表 6　与其他情绪识别基准方法比较

Table 6　Comparison with other emotion recognition benchmark methods

方法	划分策略	模态						
		T	V	A	T+V	T+A	A+V	A+V+T
文献[44]	训练集-测试集	—	—	—	75.9	76.1	69.5	76.5
文献[46]	训练集-测试集	—	—	—	78.32	78.85	67.75	80.87
文献[26]	交叉验证	70.81	—	69.89	—	79.22	—	—
文献[34]	训练集-测试集	54.8	—	57.4	—	69.5	—	—
文献[51]	交叉验证	65.90	—	55.60	—	72.82	—	—
文献[52]	训练集-测试集	66.09	—	—	—	—	—	—
文献[12]	交叉验证	—	—	—	—	75.3	—	—
文献[53]	训练集-测试集	65.13	41.47	32.84	70.61	—	55.70	71.84
文献[54]	交叉验证	80.12	90.86	71.22	87.90	79.22	88.90	92.4
文献[55]	训练集-测试集	82.54	94.87	75.46	94.55	82.11	92.01	92.88
INT2-M3AP	训练集-测试集	84.43	96.04	77.79	97.74	83.98	94.86	93.0

表 7　单模式、多模式与本文方法的情感检测结果对比

Table 7　Comparison of sentiment detection results among unimodal, multimodal, and the proposed method

方法	模型	准确率/%	召回率/%	精准率/%	F1 分数
单模式	视觉语义模型	90.47	88.89	91.39	90.12
	文本语义模型	74.27	69.23	74.23	71.64
	音频语义模型	68.47	61.54	68.26	64.73
多模式	深度层次多模式模型	83.59	84.62	81.33	82.94
单模式＋多模式	INT2-M3AP	85.0	88.46	81.27	84.71

表 8　不同类别的情感检测结果对比

Table 8　Comparison of sentiment detection results across different categories

情感类别	准确率/%
消极	87.01
中性	70.62
积极	93.14

表 9　不同视觉语义模型的情感检测结果对比

Table 9　Comparison ofsentiment detection results among different visual semantic models

模型	准确率/%	召回率/%	精准率/%	F1 分数
VGG16	90.47	94.12	86.42	90.09
VGG19	93.91	96.15	91.31	93.67
ResNet18	87.11	88.00	84.16	86.03
ResNet34	90.32	92.00	87.32	89.59
ResNet50	92.84	96.00	89.15	92.44
InceptionV3	82.71	80.77	82.22	81.49

模式和多模式方面都取得了显著优势。

此外，还比较了 INT2-M3AP 在不同关键超参数下(包括不同优化器和 batch 大小)下的情绪识别和情感检测结果，以检验其对模型性能的影响，结果如表 11 和表 12 所示。

表 10　与其他情感检测基准方法比较

Table 10　Comparison with other sentiment detection benchmark methods

方法	划分策略	模态						
		T	V	A	T+V	T+A	A+V	A+V+T
文献[56]	交叉验证	—	60.98±4.96	49.99±3.63	—	—	64.67±6.48	—
文献[57]	交叉验证	—	55.53	47.08	—	—	63.92	—
文献[58]	交叉验证	—	—	61.9	—	—	—	—
文献[59]	交叉验证	—	—	49.80	—	—	—	—
文献[60]	交叉验证	64.8	—	59.0	—	69.4	—	—

续表 10

Continuation table 10

方法	划分策略	模态						
		T	V	A	T+V	T+A	A+V	A+V+T
文献[10]	训练集-测试集	—	—	—	—	—	—	64.3
文献[61]	训练集-测试集	—	—	74.99	—	—	—	—
文献[53]	训练集-测试集	65.17	46.05	43.83	65.33	—	48.70	68.28
文献[54]	交叉验证	68.95	74.29	59.66	84.43	70.55	78.89	70.0
文献[55]	训练集-测试集	72.85	85.03	65.04	86.23	71.15	86.96	79.07
INT2-M3AP	训练集-测试集	74.27	90.47	68.47	91.66	73.77	91.13	85.0

表 11 不同超参数下的情感检测结果对比

Table 11 Comparison of sentiment detection results under different hyperparameters

优化器	Batch 大小	准确率/%
Rmsprop	16	83.0
	32	85.0
	64	85.0
Adam	16	84.0
	32	85.0
	64	84.0

表 12 不同超参数下的情绪识别结果对比

Table 12 Comparison of emotion recognition results under different hyperparameters

优化器	Batch 大小	准确率/%
Rmsprop	16	91.0
	32	93.0
	64	92.0
Adam	16	92.0
	32	93.0
	64	92.0

可以看出,对于情感检测任务,使用 Rmsprop 和 Adam 优化器,当 batch 大小为 32,准确率最高,达到 85.0%。同样,对于情绪识别任务,Rmsprop 和 Adam 优化器,当 batch 大小为 32,相对于其他超参数组合的结果更好,准确率达到 93.0%。

4 结 论

本文提出了一种用于多模式感情预测的 INT2-M3AP 框架。它首先在不同层次上探索了多种模式之间的内在联系,并有效地识别了文本、图像和音频模式之间的判别性特征。其次它可以直观地将关注重点集中于情绪和情感相关的单词、区域和声学特征上。这种方法使模型能够利用互补信息,从而更准确地预测情绪和情绪。最后,为了协同多种模式,INT2-M3AP 将中层融合和决策层融合技术结合在一个统一的框架内。这种融合方法确保了多模式在深度端到端模型中可以相互补充。在开源的数据集上评估证明了 INT2-M3AP 框架对于情感检测和情绪识别的有效性。本文的方法表明,在精确地感知情感和情绪方面,多模式方法比单模式方法的确更好,因为在学习过程中,单一模式的局限性可以通过其他模式来弥补。

参考文献

[1] 贾林锋,吴黎明,温腾腾,等. 多尺度卷积的时频域语音分离方法研究[J]. 电子测量与仪器学报,2022,36(11):134-140.
JIA L F, WU L M, WEN T T, et al. Multi-scale convolution timefrequency domain speech separation method [J]. Chinese Journal of Electronic Measurement and Instrumentation, 2022, 36(11): 134-140.

[2] 张小恒,李勇明,王品. 双阶段帕金森病语音聚类包络卷积稀疏迁移学习算法[J]. 仪器仪表学报,2022,43(11):151-161.
ZHANG X H, LI Y M, WANG P. A clusteringenvelope convolution sparse transfer l earning algorithm for two-stage Parkinson′s disease speech[J]. Chinese Journal of Scientific Instrument, 2022,43(11):151-161.

[3] MAN A, PU Y, XU D, et al. Multi-feature fusion for multimodal attentive sentiment analysis[C]. MMAsia, 2019, 43:1-6.

[4] JIN P, LI J, MU L, et al. Effective sentiment analysis for multimodal review data on the web[C]. Algorithms and Architectures for Parallel Processing: 20th International Conference, 2020: 623-638.

[5] 闫超,贾振堂. 基于 Transformer 与增强信息融合的双源情感识别[J]. 国外电子测量技术,2023,42(4):187-193.
YAN CH, JIA ZH T. Dual-source emotion recognition based on transformer and enhanced information fusion [J]. Foreign Electronic

Measurement Technology,2023,42(4):187-193.

[6] 周佳鑫,焦亚萌,王彦斌,等.融合注意力和辅助分类器的膨胀残差网络语音情感识别研究[J].国外电子测量技术,2023,42(8):19-25.
ZHOU J X, JIAO Y M, WANG Y B, et al. Research on Inflatable Residual Network speech Emotion Recognition based on Fusion of attention and auxiliary classifier [J]. Foreign Electronic Measurement Technology,2023,42(8):19-25.

[7] 焦亚萌,周成智,李文萍,等.融合多头注意力的VGGNet语音情感识别研究[J].国外电子测量技术,2022,41(1):63-69.
JIAO Y M, ZHOU CH ZH, LI W P, et al. VGGNet Speech Emotion Recognition with multi-head attention[J]. Foreign Electronic Measurement Technology, 202, 41(1):63-69.

[8] 闫舒羽,李小光,顾天昊,等.基于多通道帧级筛选的LSTM网络脑电情感识别[J].国外电子测量技术,2023,42(12):94-101.
YAN SH Y, LI X G, GU T H, et al. Eeg emotion recognition using LSTM network based on multi-channel frame level screening[J]. Foreign Electronic Measurement Technology,2023,42(12):94-101.

[9] LI H, HUANG J, HUANG J, et al. Deep multimodal learning and fusion based intelligent fault diagnosisapproach[J]. Journal of Beijing Institute of Technology,2021, 30(2): 172-185.

[10] ZHANG Z, RINGEVAL F, DONG B, et al. Enhanced semi-supervised learning for multimodal emotion recognition [C]. 2016 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). IEEE, 2016: 5185-5189.

[11] CHEN W, XING X, XU X, et al. Key-sparsetransformer with cascaded cross-attention block for multimodal speech emotion recognition[J]. ArXiv preprint arXiv:2106.11532, 2021.

[12] HU P, PENG D, WANG X, et al. Multimodal adversarial network for cross-modal retrieval [J]. Knowledge-Based Systems, 2019, 180: 38-50.

[13] KOTELNIKOVA A, PASCHENKO D, BOCHENINA K, et al. Lexicon-based methods vs. BERT for text sentiment analysis [C]. International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer International Publishing, 2021: 71-83.

[14] VIELMA C, VERMA A, BEIN D. Single and multibranch CNN-bidirectional LSTM for IMDb sentiment analysis[C]. 17th International Conference on Information Technology-New Generations (ITNG 2020), 2020: 401-406.

[15] IRSOY O, CARDIE C. Opinion mining with deep recurrent neural networks[C]. Proceedings of the 2014 conference on empirical methods in natural language processing(EMNLP), 2014: 720-728.

[16] NANDI B P, JAIN A, TAYAL D K. Aspect based sentiment analysis using long-shortterm memory and weighted N-gram graph-cut [J]. Cognitive Computation, 2023, 15(3): 822-837.

[17] RAMASWAMY S L, CHINNAPPAN J. RecogNet-LSTM + CNN: Ahybrid network with attention mechanism for aspectcategorization and sentiment classification [J]. Journal of Intelligent Information Systems, 2022, 58(2): 379-404.

[18] OHALA J J. Ethological theory and the expression of emotion in the voice [C]. Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96. IEEE, 1996, 3: 1812-1815.

[19] NAVAS E, HERNÁEZ I, LUENGO I. An objective and subjective study of the role ofsemantics and prosodic features in building corpora for emotional TTS[J]. IEEEtransactions on audio, speech, and language processing, 2006, 14(4): 1117-1127.

[20] HU D, CHEN C, ZHANG P, et al. A two-stage attention based modality fusion framework for multi-modal speech emotion recognition [J]. IEICE TRANSACTIONS on Information and Systems, 2021, 104(8): 1391-1394.

[21] YANG Y, JIA J, ZHANG S, et al. How do your friendson social media disclose your emotions? [C]. Proceedings of the AAAIConference on Artificial Intelligence. 2014, 28 (1), DOI: 10.13140/2.1.1484.9920.

[22] LI Z, SUN Q, GUO Q, et al. Visual sentiment analysis based on image caption and adjective-noun-pair description[J]. Soft Computing, 2021(2): 1-13.

[23] CAMPOS V, JOU B, GIRO X. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction[J]. Imageand Vision Computing, 2017, 65: 15-22.

[24] YOU Q, LUO J, JIN H, et al. Robust image sentimentanalysis using progressivelytrained and domain transferred deep networks[C]. Proceedings of the AAAI conference on Artificial Intelligence, 2015, 29(1), DOI:10.48550/arXiv.1509.06041.

[25] NEDIYANCHATH A, PARAMASIVAM P, YENIGALLA P. Multi-head attention for speech

emotion recognition withauxiliary learning of gender recognition[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2020: 7179-7183.

[26] MINAEI B B, ASADI M, PARVIN H. Anensemble based approach for feature selection[C]. International Conference on Engineering Applications of Neural Networks. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 240-246.

[27] BAKHSHI A, WONG A S W, CHALUP S. End-to-end speech emotion recognition based on time and frequency information using deep neural networks[C]. European Conference on Artificial Intelligence, 2020: 969-975.

[28] HOU Y, YU H, ZHOU D, et al. Local-awarespatio-temporal attention network with multi-stage feature fusion for human action recognition [J]. Neural Computing and Applications, 2021, 33: 16439-16450.

[29] CAI J, MENG Z, KHAN A S, et al. Feature-level and model-level audiovisual fusion for emotion recognition in the wild[C]. 2019 IEEE Conference on MultimediaInformation Processing and Retrieval (MIPR). IEEE, 2019: 443-448.

[30] CRUZ F, PARISI G I, TWIEFEL J, et al. Multi-modal integration of dynamic audiovisual patterns for aninteractive reinforcement learning scenario[C]. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2016: 759-766.

[31] NAKAYAMA S, TAKADA T, KIMURA R, et al. Temperatureand humidity deficit model for greenhouses based on system identification method with two-variable input: quick setup and model evaluation duringspring and fall [J]. Agricultural Information Research, 2021, DOI: 10.3173/AIR.30.1.

[32] CHEN J, WU Z, YANG Z, et al. Multimodalfusion network with latent topic memoryfor rumor detection[C]. 2021 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2021: 1-6.

[33] SINGH P, SRIVASTAVA R, RANA K P S, et al. A multimodal hierarchical approach to speech emotion recognition from audioand text[J]. Knowledge-Based Systems, 2021, 229: 107316.

[34] WEI H, ZHANG Q, GU Y. Remaining useful life prediction of bearings based on self-attention mechanism,multi-scale dilated causal convolution, and temporal convolution network [J]. Measurement Science and Technology, 2023, 34(4): 045107.

[35] SHIRZAD A, ZARE H, TEIMOURI M. Deep Learning approach for text, image, and GIFmultimodal sentiment analysis [C]. 2020 10th International Conference onComputer and Knowledge Engineering(ICCKE). IEEE, 2020: 419-424.

[36] HUANG F, WEI K, WENG J, et al. Attention-based modality-gated networks for image-text sentiment analysis [J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2020,16(3): 1-19.

[37] GUO W, ZHANG Y, CAI X, et al. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition[J]. IEEE Transactions on Multimedia, 2020, 23: 1785-1798.

[38] ZAHOOR M M, QURESHI S A, BIBI S, et al. A new deephybrid boosted and ensemble learning-based brain tumor analysis using MRI[J]. Sensors, 2022, 22(7): 2726.

[39] XIAO L, WU X, WU W, et al. Multi-channel attentivegraph convolutional network with sentiment fusion for multimodal sentiment analysis[C]. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 4578-4582.

[40] LING Y, YU J, XIA R. Vision-language pre-trainingfor multimodal aspect-based sentiment analysis[J]. ArXiv preprint arXiv:2204.07955, 2022.

[41] MAN A, PU Y, XU D, et al. Multi-FeatureFusion forMultimodal Attentive Sentiment Analysis [C]. MMAsia. 2019: 43:1-43:6.

[42] NGUYEN T L, KAVURI S, LEE M. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips[J]. Neural Networks, 2019, 118: 208-219.

[43] CAI G, LYU G, LIN Y, et al. Multi-level deep correlative networks for multi-modal sentiment analysis[J]. Chinese Journal of Electronics, 2020, 29(6):1025-1038.

[44] CAMBRIA E, HAZARIKA D, PORIA S, et al. Benchmarking multimodal sentiment analysis [C]. Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17-23, 2017, Revised Selected Papers, Part II 18. Springer International Publishing, 2018: 166-179.

[45] HUDDAR M G, SANNAKKI S S, RAJPUROHIT V S. Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal

sentiment analysisand emotion classification [J]. International Journal of Multimedia Information Retrieval, 2020, 9(2): 103-112.

[46] CHAN J Y L, BEA K T, LEOW S M H, et al. State of the art: A review of sentimentanalysis based on sequential transfer learning[J]. Artificial Intelligence Review, 2023, 56(1): 749-780.

[47] WANG W, PAN S J, DAHLMEIER D, et al. Recursive neural conditional random fields for aspect-based sentiment analysis[J]. ArXiv preprint arXiv: 1603.06679, 2016.

[48] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning toalign and translate[J]. ArXiv preprint arXiv:1409.0473, 2014.

[49] TÜRKER B B, MARZBAN S, SEZGIN M T, et al. Affect burst detection using multi-modal cues[C]. 2015 23nd Signal Processing and Communications Applications Conference(SIU). IEEE, 2015: 1006-1009.

[50] EVTODIENKO L. Multimodal end-to-end group emotion recognition using cross-modal attention[J]. ArXiv preprint arXiv:2111.05890, 2021.

[51] LI W, SHAO W, JI S, et al. BiERU: Bidirectional emotional recurrent unit for conversational sentiment analysis[J]. Neurocomputing, 2022, 467: 73-82.

[52] WÖLLMER M, METALLINOU A, EYBEN F, et al. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling[C]. Annual conference of the International Speech Communication Association,2010.

[53] LI B, FEI H, LIAO L, et al. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition[C]. Proceedings of the 31st ACM International Conference on Multimedia, 2023: 5923-5934.

[54] KHALANE A, MAKWANA R, SHAIKH T, et al. Evaluating significant features in context-aware multimodal emotion recognition with XAI methods[J]. Expert Systems, 2023,1(6): 1-25.

[55] LIU X, XU Z, HUANG K. Multimodal emotion recognition based on cascaded multichannel and hierarchical fusion[J]. Computational Intelligence and Neuroscience, 2023, 1: 5611-5628.

[56] HE Y, SENG K P, ANG L M. Multimodal sensor-input architecture with deep learning for audio-visual speech recognition in wild[J]. Sensors, 2023, 23(4):1834.

[57] ZHANG Z, GU J. Facial affect recognition in the wild using multi-task learningconvolutional network[J]. ArXiv preprint arXiv:2002.00606, 2020.

[58] BEGUŠ G. Local and non-local dependencylearning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks[J]. Computer speech & language, 2022, 71: 101244.

[59] ALDENEH Z, KHORRAM S, DIMITRIADIS D, et al. Pooling acoustic and lexical features for the predictionof valence[C]. Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017:68-72.

[60] ATMAJA B T, AKAGI M. Dimensional speechemotion recognition from speech features and word embeddingsby using multitasklearning [J]. APSIPA Transactions on Signal and Information Processing, 2020, 9:e17.

[61] MA Y, DI G, WANG W. Advances Research in Speech Emotion Recognition Based on Multi-task Learning[C]. 2022 4th International Workshop on Artificial Intelligence and Education(WAIE). IEEE, 2022: 27-31.

作者简介

杨亚萍(通信作者),讲师,硕士,主要研究方向为概念隐喻,多模式,英文演讲。

E-mail:15831930760@163.com

张敬源,教授,博士,主要研究方向为功能语言学,话语分析。