基于超图的多模态情绪识别

宗林林[1)] 周佳慧[1)] 谢秋婕[2)] 张宪超[1)] 徐 博[3)]

[1)](大连理工大学软件学院 辽宁 大连 116000)

[2)](复旦大学计算机科学技术学院 上海 200433)

[3)](大连理工大学计算机科学与技术学院 辽宁 大连 116000)

摘 要 近年来多模态情绪识别获得广泛关注,模态间的特征融合决定了情绪识别的效果,现有基于图的情绪特征融合方法多基于二元关系图,在处理三种及以上模态数据时难以实现有效的模态间特征融合,限制了多模态情绪识别的效果.为解决该问题,本文提出基于超图的多模态情绪识别模型(Multi-modal Emotion Recognition Based on Hypergraph,MORAH),引入超图来建立多模态的多元关系,以此替代现有图结构采用的多个二元关系,实现更加充分、高效的多模态特征融合.具体来说,该模型将多模态特征融合分为两个阶段:超边构建阶段和超图学习阶段.在超边构建阶段,通过胶囊网络实现对序列中每个时间步的信息聚合,并建立单模态的图,然后使用图卷积进行第二次信息聚合,并以此作为下一阶段建立超图的基础,得益于图胶囊聚合方法的加入,MORAH可以同时处理对齐数据和未对齐数据,无需手动对齐;在超图学习阶段,模型建立同一样本不同模态节点之间的关联,以及同类样本所有模态之间的关联,同时,在超图卷积过程中,使用分层多级超边来避免过于平滑的节点嵌入,并使用简化的超图卷积方法来融合模型之间的高级特征,以确保所有节点特征仅在必要时更新.在两个基准数据集上的综合实验表明,本文模型利用超图实现了对多模态数据之间多元关系的充分利用.与现有的先进方法相比,在CMU-MOSI数据集的未对齐数据上,MORAH将二分类准确率提高了1.3%,F1得分提高了1.1%.在CMU-MOSEI数据集的未对齐数据上,MORAH将二分类准确率和F1分数分别提高了0.2%.

关键词 情绪识别;多模态学习;超图学习;超边扩展;胶囊网络

中图法分类号 TP311 **DOI号** 10.11897/SP.J.1016.2023.02520

Multi-modal Emotion Recognition Based on Hypergraph

ZONG Lin-Lin[1)] ZHOU Jia-Hui[1)] XIE Qiu-Jie[2)] ZHANG Xian-Chao[1)] XU Bo[3)]

[1)](*Department of Software, Dalian University of Technology, Dalian, Liaoning* 116000)

[2)](*School of Computer Science and Technology, Fudan University, Shanghai* 200433)

[3)](*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116000)

Abstract With the rapid progress of artificial intelligence technology, machines need to recognize users' emotions to provide users with a better human-computer interaction experience. Therefore, emotion recognition has become one of the active fields of artificial intelligence. Traditional emotion recognition is mostly based on text modality. Compared with single modality, multi-modal emotion recognition has the advantages of data complementarity and model robustness. In multi-modal emotion recognition, feature fusion between modalities determines the effect of emotion recognition. Recently, graph-based intra-modality fusion has attracted much attention of related research, which uses graphs of binary relationships between two modalities.

收稿日期:2022-11-07;在线发布日期:2023-07-13.本课题得到国家自然科学基金(No. 62006034)、大连市青年科技之星项目(2021RQ056)资助.**宗林林**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、机器学习.E-mail:llzong@dlut.edu.cn. **周佳慧**,硕士研究生,主要研究领域为情感计算. **谢秋婕**,硕士研究生,主要研究领域为情感计算. **张宪超**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习. **徐 博**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为情感计算、信息检索.E-mail:xubo@dlut.edu.cn.

When processing data of three or more modalities, the graph can hardly effectively establish the feature fusion between all modalities without introducing redundant information, limiting the performance of multi-modal emotion recognition. Therefore, it is necessary to design more effective method to model and fuse multi-modal emotion features. To solve this problem, this paper proposes an emotion recognition model Multi-modal Emotion Recognition Based on Hypergraph (MORAH) which introduces hypergraph to establish multivariate relations among multi-modal data instead of binary relations and achieves efficient multi-modal feature fusion. Specifically, the model divides multi-modal feature fusion into two stages: the hyperedge construction stage and the hypergraph learning stage. In the hyperedge construction stage, we aggregate the information of each time step in the sequence through the capsule network and establish the graph of a single modality. Then, we use graph convolution for the second aggregation, which is used as the basis for establishing hypergraph in the next stage. Benefiting from the graph capsule aggregation method, the model can work with aligned data and unaligned data at the same time, without manual alignment of unaligned data. In the hypergraph learning stage, we not only establish the association between the nodes of different modalities of the same sample but also establish the association between all modalities of the same sample. At the same time, we use hierarchical multi-level hyperedges to avoid too smooth node embedding and the simple hypergraph convolution method to fuse the high-level features between modalities, ensuring that all node features are only updated when necessary in the hypergraph convolution process. Simplified graph convolution can guarantee the effect of emotion recognition and improve the training speed without nonlinear activation and convolution filter matrix. Comprehensive experiments on two benchmark datasets show that the proposed model makes full use of the multiple relations between multi-modal data by using hypergraph. Compared with the existing advanced methods, MORAH improves the binary accuracy by 1.3% and F1-score by 1.1% on the unaligned data of the CMU-MOSI dataset. On the unaligned data of the CMU-MOSEI dataset, MORAH improves the binary accuracy and the F1-score by 0.2%, respectively. To demonstrate the generality of the hypergraph learning stage in various multimodal tasks, we apply the hierarchical multi-level hyperedges to the emotion recognition in conversation (ERC). The experimental results indicate that MORAH can improve the performance of ERC to a certain extent. This suggests that the MORAH model can function as a universal tool to assist downstream natural language processing tasks.

Keywords emotion recognition; multi-modal learning; hypergraph learning; hyperedge extension; capsule network

1 引 言

情绪是人类心理活动的核心要素之一，它在人类日常生活中无处不在，可以影响甚至决定人们的判断和决策. 随着人工智能技术的快速进步，人机交互在人类生活中扮演着越来越重要的作用，为了像人类一样对交互对象的行为做出反应，机器需要对用户当前情绪进行识别，从而为用户提供更好的人机交互体验. 因此，情绪识别成为人工智能热点研究领域之一.

在情绪识别技术的发展过程中，早期的情绪识别通常使用单模态数据，如文本情感分析方法[1]和语音情感识别方法[2]等，但在一些情况下，仅通过单模态数据难以准确区分情绪的类型及程度. 利用多模态数据进行情绪识别具有数据互补性和模型鲁棒性等优势，一定程度上能够缓解单模态数据的局限性[3]，因此近年来多模态情绪识别研究备受关注，相关方法有效提升了情绪识别的效果.

根据模态融合方式的不同，多模态情绪特征融

合可以分为无模型融合和有模型融合两种策略[4].无模型融合以模板和规则等定制化技术为主,旨在利用人工规则实现多模态数据特征的整合,很大程度上限制了方法的可扩展性,因此近年来面向多模态情绪识别的有模型融合方法得到广泛关注,常见的方法有基于核函数的融合方法[5]、基于神经网络的融合方法[6-7]、基于注意力的融合方法[8-15]和基于图的融合方法等[16-18].

基于神经网络的融合方法多以循环神经网络(Recurrent Neural Network,RNN)为主要模块,此类方法训练和测试的耗时较长,且容易出现梯度消失或爆炸的问题,限制其学习长期依赖的能力;基于注意力的方法通过自注意转换器收集时间信息,虽然避免了RNN结构所带来的梯度问题,但其对于不同时间步长之间的信息融合依旧不够充分;近年来基于图融合的情绪识别方法取得了该任务上的最佳效果,当前面向情绪识别的图结构中每条边最多仅能表示两个模态间的关系,在处理三种及以上模态的数据时难以有效建立所有模态间的关系,进而引入冗余信息,降低了情绪识别中多模态特征融合的效果.

为解决该问题,本文提出了基于超图的多模态情绪识别模型(Multi-modal Emotion Recognition Based on Hypergraph,MORAH),引入超图来建立多模态的多元关系以替代原有的多个二元关系,实现更加充分、高效的多模态特征融合.具体来说,本文采用图胶囊聚合方法聚合模态间特征和模态内特征,并将其作为下一阶段构建超图的基础.然后采用改进的超边扩展策略根据超边之间的相似性构建超图,使用超图卷积方法融合模态之间的特征并辅助情绪分类.在两个基准数据集上的实验结果表明,该方法在模态内和模态间的特征融合过程中避免了已有方法在模态融合中的不足,且可以处理对齐数据和未对齐数据,有效提升了情绪识别的效果.本文首次将超边扩展策略和超图卷积网络应用到多模态情绪识别任务,为未来相关研究的开展提供了新的研究思路和优化方向.

本文的主要创新之处总结如下:

(1)提出基于多元情绪关系超图学习的多模态情绪识别方法,该方法用多元情绪关系替代多个二元情绪关系,从而在不引入冗余信息的同时达到对单模态内部和多模态之间多元情绪关系的充分利用,更有效地识别用户情绪.

(2)提出融合超图卷积的情绪超边扩展策略,针对多模态数据场景的情绪识别任务,充分利用各模态细粒度特征建立超边,基于超图卷积方法捕获模态间多元关系,从而打破成对二元关系的限制,提升多模态情绪识别的效果.

(3)基于两个大规模通用多模态情绪识别数据集开展模型评估,实验结果表明,本文方法可以显著提升多模态情绪识别的整体性能,所提出的MORAH模型可以处理对齐和未对齐数据,无需为未对齐数据进行手动对齐等步骤.

2 相关工作

本节将从多模态情绪识别和超图学习两个维度对相关研究工作加以介绍.

2.1 多模态情绪识别

多模态情绪识别的关键在于不同模态特征的融合,按照采用方法的不同,多模态特征融合方法可以划分为基于核函数的融合方法、基于神经网络的融合方法、基于注意力的融合方法和基于图的融合方法,下面分别介绍这四类方法.

2.1.1 基于核融合的情绪识别

基于核函数的融合是基于不同核函数的分类器进行情绪识别,如使用支持向量机(Support Vector Machine,SVM)的方法[5]等.在多模态场景中,通常针对不同的模态可以使用不同的核函数.由于具有核选择的灵活性和损失函数的凸性,多核学习融合在包括多模态情绪识别在内的许多应用中流行起来.然而,在测试过程中,这些融合方法很大程度上依赖训练数据中的支持向量,可能导致较大的内存开销和过拟合.

2.1.2 基于神经网络融合的情绪识别

基于神经网络的融合方法[6-7]采用不同策略融合神经网络对不同模态的特征表示或预测结果,常见方法有RNN及其变体长短期记忆神经网络(Long Short Term Memory,LSTM)和门控循环神经网络(Gated Recurrent Unit,GRU)等.此类方法以RNN为核心模块,主要用于单词对齐的多模态序列.Zadeh等[6]提出了由LSTM和GRU构建的记忆融合网络,利用外部多模态记忆机制存储多模态信息,探索特定视图和跨视图交互.随着注意力机制的广泛应用,部分方法中也加入了注意力作为辅助模块.Zadeh等[7]提出了多注意力循环网络(Multi-attention Recurrent Network,MARN),使用多个注意力系数表示多个跨模态交互,以便对模态内和模态间两种动力学特征进行建模,该模型包含两个关

键组件：长短期混合记忆和多注意块，长短期混合记忆是LSTM的扩展形式，通过重组记忆组件来融合混合信息；多注意块是用于发现不同的跨模态交叉视图动态的组件，该方法在神经网络的基础上加入注意力机制，效果较基于神经网络的方法提升显著. 虽然这些方法使用注意力机制提高了模型性能，但由于它们以RNN为主要模块，仍存在训练和测试时间较长和梯度消失等问题.

2.1.3 基于注意力融合的情绪识别

基于注意力的融合方法[8-15]进一步利用注意机制关注不同模态的不同部分，以避免RNN结构训练时间较长和梯度问题. 早期的工作通常只是采用特征融合策略，将来自不同模态的输入序列连接起来，或者将从每个单独模态学习到的高级信息结合起来. 虽然这些工作获得了比单一模态学习更好的性能，但它们没有明确考虑来自不同模态的序列元素之间的内在依赖关系. Vaswani等[8]首次将仅依赖注意力机制的转换器网络(Transformer)引入神经机器翻译任务，其中编码器和解码器各自利用自注意转换器，这种基于注意力机制的Transformer避免了以RNN为主要模块的模型存在的弊端. 受此启发，Tasi等[9]提出了基于跨模态注意力和自注意力机制的多模态转换器(Multi-modal Transformer，MulT)，他们没有使用传统的编码器与解码器结构，而是建立了多个成对的双向跨模态注意块，通过定向成对跨模态注意关注多模态序列之间的相互作用，用其他模态的特征增强一种模态，有效捕获了跨模态信号，在情绪识别任务中表现出较好的效果.

但这种成对且双向的跨模态注意力会反复地对一种模态的信息进行强化，不可避免地带来冗余信息. Lv等[10]在MulT的基础上做出改进，提出渐进式模态强化(Progressive Modality Reinforcement，PMR)方法，引入消息中心与所有模态进行交互. 消息中心可以向每个模态发送公共信息，从而通过跨模态注意力增强其特征，使用每个模态的增强特征生成改进的公共信息. 和MulT相比，PMR具有两点优势，首先，消息中心促进了跨模态之间的有效信息流，并通过跨模态注意力探索三种模式之间的元素级依赖，而不是定向成对依赖，避免了模态间两两交互带来的冗余信息；其次，渐进式强化策略使源模态的特征可以在强化单元中作为目标模态，为利用源模态的高级特征进行模态强化提供了一种有效的方法. Fu等[11]提出了非齐次融合网络，该方法使用一个带有约束的注意聚合模块实现了对源模态特征的强化，并克服了成对注意复杂性，提高了在模态之间聚合信息的能力. Wang等[12]提出了一种基于变分自编码器的对抗多模态领域迁移算法，并通过多头注意力模块降低了单模态表示之间的距离. Arjmand等[13]提出了一种基于Transformer网络的前缀语言模型，利用传统预训练语言模型实现了跨模态信息的转换. Yang等[14]提出一种面部特征敏感的图像到文本的情感转换方法，从面部表情挖掘视觉情绪语义，匹配目标文本的情感类型，提升了方面级情感分析的性能. Yang等[15]提出一种特征解耦合的多模态情绪识别方法，通过通用编码器和私有编码器的建立，融合各模态特征子空间，提升了多模态情绪识别的效果. 然而，由于大多数基于注意力的方法通过自注意转换器收集时间信息，对于不同时间步长之间的信息融合依旧不够充分.

2.1.4 基于图融合的情绪识别

得益于图卷积网络在处理非欧几里得数据中的强大能力，基于图融合的方法[16-18]近年来得到广泛应用. 在多模态情绪识别任务中，基于图的融合方法为每个模态构造单独的图，将这些图组合成一个融合图，并通过基于图的学习获得融合后的特征，更好地适应数据不全的情况. Yang等[16]提出了模态注意图(Modal-temporal Attention Graph，MTAG)，首先将未对齐的多模态序列转换为具有异构节点和边的图，该图捕获了不同模态之间和不同时间步之间的丰富交互. 然后提出了一种新的图融合方法MTAG融合方法，以及动态剪枝和读取技术，有效地处理了模态时间图并捕获其中的各种交互. MTAG的优势在于可以同时捕获任意数量模态的各种类型的交互，但和之前基于图融合的方法一样，它使用了池化和剪枝来删除一些节点以获得图的最终表示，导致部分信息丢失.

相比之下，图胶囊聚合方法(Graph Capsule Aggregation，GraphCAGE)[17]应用胶囊网络的动态路由从序列构建节点，使模型能够处理更长的序列，具有更强的远程依赖学习能力. 该方法先通过跨模态转换器融合模态间特征，然后在图构造部分通过胶囊网络从低级的特征序列中提取高级特征表示并建立图，最后进行图聚合，使用图卷积和动态路由技术聚合每个模态的特征. 该方法应用动态路由机制替代了池化操作，保留了图的所有信息，可以在不丢失信息的情况下生成高级精细的序列表示. 此外，由于动态路由具有可解释性，更大的路由系数表明

更大的贡献，可以找出模型通过关注哪些信息做出预测．这一方法很好地利用了胶囊网络的聚合能力和图卷积网络的优势，同时避免了RNN的不足，提高了模型效率．然而，GraphCAGE方法并没有组合不同模态的图，依旧使用跨模态注意力机制融合模态间特征，存在模态间特征融合不充分的问题．与GraphCAGE建模多模态数据的方法不同，Hu等[18]使用图卷积融合模态间特征，提出多模态图卷积神经网络，该模型在相同模态不同话语的节点间和不同模态相同话语的节点间建立边，通过图卷积聚合节点信息，取得较好效果．

在最新的研究中，Han等[19]将互信息与多模态情感分析结合，提出了一种用于多模态情绪分析的分层互信息最大化框架（MultiModal InfoMax，MMIM）．Hu等[20]结合了情感和情绪两者的互补信息，提出了一个多模态情感知识共享框架（Unified Multimodal Sentiment Analysis and Emotion Recognition，UniMSE），该框架将多模态情感分析（Multimodal Sentiment Analysis，MSA）和对话情绪识别任务（Emotion Recognition in Conversation，ERC）相结合，从功能、标签和模型等角度优化，在两个任务上均取得了目前最好的效果．另外，相关研究尝试建模人机对话中的多模态情绪，辅助实现更加精准的对话回复生成[21-23]．但是这些基于图的方法主要建立成对节点之间的关系，在处理三种及以上模态的数据时，难以有效地建模所有模态间的多元关系，因此基于图融合的深度学习方法有待进一步改进．

2.2 超图学习方法

超图学习提供了一种有效的方法捕获复杂的高阶关系．与普通图不同，超图的每条边可以包含多个节点，即每条超边不受二元关系的限制，可以同时表示多元关系．随着深度学习的发展，超图学习受到了广泛的关注．在图卷积的基础上，Feng等[24]提出的超图神经网络（Hypergraph Neural Networks，HGNN）和Yadati等[25]提出的超图卷积网络是最早适用于超图的卷积方法．超图学习的目标是最小化超图上连接较强的顶点之间的标签差异．HGNN是基于超图上的谱卷积方法，作者将每层卷积划分为三个阶段：节点特征转换、超边特征收集、节点特征聚合，这样可以利用超图结构更好地细化特征．具体来说，该模型首先通过可学习的滤波器矩阵$\boldsymbol{\Theta}$对初始节点特征$\boldsymbol{X}$进行处理，提取C_2维特征．然后，根据超边收集节点特征，通过与关联矩阵转置的乘法形成维度是$E\times C_2$的超边特征．最后，聚合节点的相关超边特征以得到更新后的节点特征，因此HGNN通过节点到边再到节点的变换有效地提取超图上的高阶相关性．以上过程可以用公式（1）表示，其中$\boldsymbol{D}_v$和$\boldsymbol{D}_e$分别是节点度和超边度的对角矩阵，起归一化作用．

$$Y=D_v^{-1/2}HWD_e^{-1}H^TD_v^{-1/2}X\Theta \tag{1}$$

与现有的图卷积方法相比，HGNN可以自然地建模数据之间的高阶关系，这种关系在形成特征提取中得到了有效的利用和编码．与传统的超图方法相比，该模型不需要超图拉普拉斯行列式的逆运算，因此具有更高的计算效率．值得注意的是，HGNN对多模态特征具有很大的可扩展性，并具有超边生成的灵活性．随后，超图被广泛应用于各种下游任务中，Xia等[26]在会话推荐系统中提出的超图卷积方法将简易图卷积[27]的思想与HGNN结合，从而得到简易的超图卷积，作者通过实验证明，改进后的模型与图卷积网络（Graph Convolutional Network，GCN）相比并不会降低下游应用程序的准确性，并且比快速图卷积网络（Fast Learning with Graph Convolutional Networks，FastGCN）[28]的速度提升了两个数量级．因此，简易图卷积可以在保证模型效果的同时得到更加轻量的模型．Shi等[29]发现多数方法为了应用超图学习模型，需要手动分配超边，比如通过计算节点特征之间的距离来构建超边或用k个最近的节点构造超图等．然而，基于这些方法创建的超边都是单一的，一个单一超边可以被视为一个完整的图，其中无论节点间是否有真实连接都会被视为有连接，这会导致超图中不同的节点嵌入过度平滑的问题．为解决该问题，Sun等[30]提出一种自动学习高质量超边的新方法，通过这种超边扩展策略生成的超边是分层的，遵循幂律分布，显著提高了链路预测性能．在实施超边扩展策略之前，他们先计算了超边扩展的可能性，从而为下一步的扩展提供依据．具体计算如公式（2）和公式（3）所示．

$$\alpha(i,j)=\sigma(Z_e^T\cdot Z_e)_{ij} \tag{2}$$

$$\beta(i,j)=\frac{\sum_{k=1}^{|V|}H(k,i)\cdot H(k,j)}{\sum_{p=1}^{|V|}H(p,i)+\sum_{q=1}^{|V|}H(q,j)} \tag{3}$$

其中，$\boldsymbol{Z}$表示超边的特征，$\boldsymbol{H}$是关联矩阵．公式（2）是通过特征的内积来计算超边i和超边j的相关性；公式（3）是通过Jaccard相似度来计算两条超边的连通性，即用两条超边集合的交集和并集之比表示．最后结合相关性和连通性计算超边扩展的可能性，如公式（4），表示从超边i扩展到超边j的可能性．

$$p(j|i)=\frac{\alpha(i,j)+\beta(i,j)}{\sum_{k\in N_i}(\alpha(i,k)+\beta(i,k))} \tag{4}$$

针对不同的实际情况，有三种超边扩展策略可供选择：深度优先扩展、广度优先扩展和混合扩展．每一种扩展方式都需要对扩展可能性进行排序，按照不同的扩展方式找到即将扩展的超边集合．由于在扩展过程中有部分超边被吞并，也有部分超边保持不变，于是超边数量减少并且获得有层次的多级超边，这种方法可以有效避免节点嵌入过于平滑．鉴于超图学习在多模态特征融合中的效果，本文尝试将超图学习引入多模态情绪识别任务，利用任务特征增强超图学习的效果，达到有效识别多模态情绪的目的．

3　基于超图的多模态情绪识别

3.1　整体结构

为将超图学习应用于多模态情绪识别任务，本文将情绪识别分为了三个阶段：情绪超边构建、情绪超图学习和情绪分类．其中，情绪超边构建阶段旨在从未对齐的多模态情绪数据中学习每个模态内部的长距离依赖，以提取每个模态序列更高级、更精细的表示．进而将聚合后的图表示作为新的节点，建立同一样本不同模态节点间的超边；情绪超图学习阶段旨在融合模态间的特征以实现情绪识别，使用改进的超边扩展策略根据扩展可能性在第一阶段构建的超边之间建立连接，即建立不同样本节点间的边，为接下来的超图卷积提供支持；情绪分类阶段利用超图卷积得到的节点特征对情绪评分进行预测．图1展示了基于超图的多模态情绪识别模型(MORAH)的整体结构．下面详细介绍模型各个模块．

3.2　情绪超边构建

情绪超边构建过程首先获取单模态情绪特征表示，以文本(t)、语音(a)和视觉(v)三种模态的情绪特征融合为例，给定三种模态的数据序列，每种模态的序列表示为$\boldsymbol{X}^{\{t,a,v\}}\in\mathbb{R}^{d^{\{t,a,v\}}\times T^{\{t,a,v\}}}$，其中$d^{\{t,a,v\}}$表示各模态数据的特征维度，$T^{\{t,a,v\}}$是序列长度．本文方法首先将三种模态的特征输入到跨模态转换器(Cross-modal Transformer，CT)中，如下所示．

$$Z^t=\mathrm{CT}^{v\to t}(X^t,X^v)\oplus\mathrm{CT}^{a\to t}(X^t,X^a) \tag{5}$$

$$Z^a=\mathrm{CT}^{t\to a}(X^a,X^t)\oplus\mathrm{CT}^{v\to a}(X^a,X^v) \tag{6}$$

$$Z^v=\mathrm{CT}^{t\to v}(X^v,X^t)\oplus\mathrm{CT}^{a\to v}(X^v,X^a) \tag{7}$$

其中，$\oplus$代表连接操作，输出的各模态特征向量表示为$Z^{\{t,a,v\}}\in\mathbb{R}^{d\times T^{\{t,a,v\}}}$，该向量可以初步融合模态间动力学特征，但其仍需要进一步抽取情绪在时间维度上的长距离依赖．为此，本文尝试利用胶囊网络和图卷积操作进一步萃取模态内特征，充分利用胶囊网络在多个网络层间传输时信息不易丢失和可解释性强等优势，从而更加充分地融合不同时间步的情绪信息．

具体地，首先基于胶囊网络的动态路机制从序列的各时间步中聚合信息，以创建表示单个模态序列特征的情绪图．动态路由机制中路由系数在训练中动态确定，可以使聚合信息时分配的权重更符合聚合目标，同时捕获更有价值的情绪信息．为了便于表示，本文接下来使用m代表模态种类．各模态情绪胶囊的创建过程可以形式化表示如下．

$$Caps^m_{i,\ j}=W^m_{i,j}Z^m_i \tag{8}$$

其中，$\boldsymbol{Z}^m_i$是序列$\boldsymbol{Z}^m$第i个时间步的特征，$\boldsymbol{W}^m_{i,j}\in\mathbb{R}^{d_c\times d}$是可训练的参数矩阵，$d_c$表示胶囊的特征维度，由于

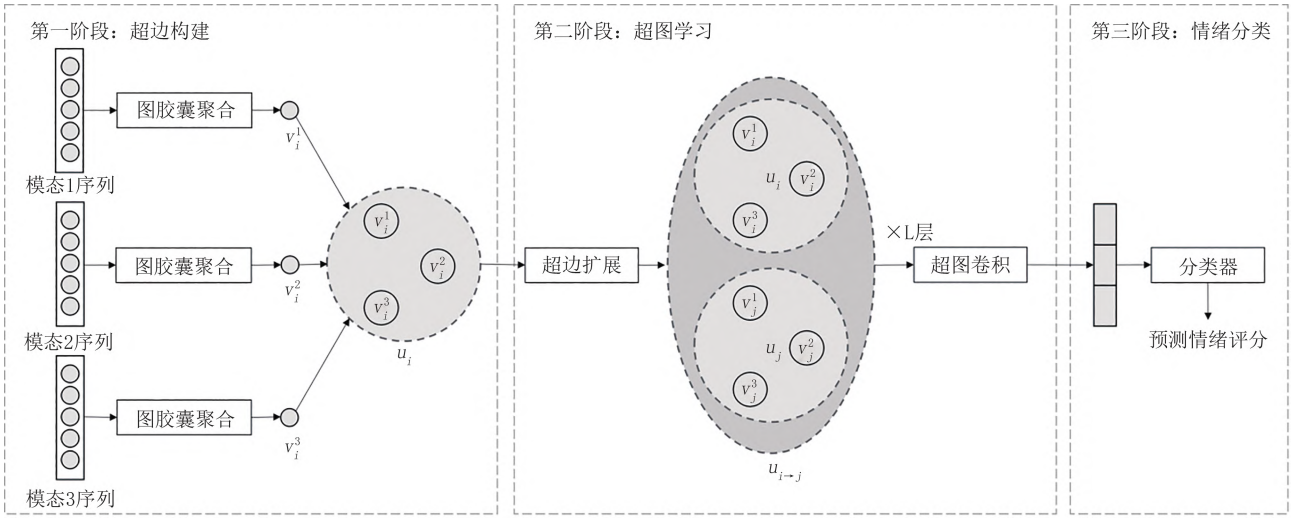

图1　基于超图的多模态情绪识别框架

本文的模型可以推广到n个模态，因此设置$m\in[1,n]$. $\boldsymbol{Caps}_{i,j}^{m}$表示用第$i$个时间步的信息建立的胶囊，该胶囊用于建立第$j$个节点$\boldsymbol{N}_j^m$.

在训练过程中，根据该胶囊与节点的相似性动态更新路由系数$b_{i,j}^{m}$，具体计算方法如下所示.

$$N_j^m=\sum_i Caps_{i,j}^m\times r_{i,j}^m \tag{9}$$

$$r_{i,j}^m=\frac{exp\left(b_{i,j}^m\right)}{\sum_j exp\left(b_{i,j}^m\right)} \tag{10}$$

$$b_{i,j}^m\leftarrow b_{i,j}^m+Caps_{i,j}^m\odot N_j^m \tag{11}$$

其中，$r_{i,j}^m$是$\boldsymbol{Caps}_{i,j}^{m}$的路由系数$b_{i,j}^{m}$标准化的结果.

重复利用公式(9-11)进行多次迭代，可以更新$r_{i,j}^m$得到最终的节点表示$\boldsymbol{N}^m$，从而聚合每个模态序列中所有时间步的信息，通过上述单模态表示学习，可以初步解决长距离情绪依赖的问题，更加充分地实现情绪信息融合.

在单模态表示基础上，需要构建单个模态的图结构，本文采用自注意机制计算单模态图的邻接矩阵，具体计算方法如公式(12)所示.

$$A^m=f\left(\frac{(W_q^m N^m)^T(W_k^m N^m)}{d_c}\right) \tag{12}$$

其中，$\boldsymbol{W}_q^m$和$\boldsymbol{W}_k^m$是可训练的参数矩阵，f是非线性激活函数，d_c表示胶囊的特征维度，该公式用于计算各模态超图的邻接矩阵，从而完成单模态图的构建.

在此基础上，本文迭代利用图卷积和胶囊网络聚合节点信息，即对于每次图卷积都迭代p次动态路由，以获得最终的表示向量，第k次图卷积和胶囊网络的计算如公式(13-15)所示.

$$n^{m,k}=W^k N^{m,k-1}(A^m+I) \tag{13}$$

$$N^{m,k}=f(W_o^k n^{m,k}) \tag{14}$$

$$V^{m,k}=\mathrm{CapsNet}(N^{m,k}) \tag{15}$$

其中，$\boldsymbol{W}^k$和$\boldsymbol{W}_o^k$表示可训练参数矩阵，$CapsNet(\cdot)$表示多次迭代公式(8-11)所获得的胶囊网络的计算过程.

上述两次胶囊网络的区别在于：第一次胶囊网络是从各模态特征序列的时间步聚合信息到单模态图的节点，为单模态图卷积做准备；第二次胶囊网络是从单模态图的各节点聚合信息到每个样本的单模态特征表示. 针对每个模态，通过两阶段的聚合将原始特征序列中的细粒度信息聚合为高表达性的单模态特征. 节点表示$\boldsymbol{N}^m$经过上述计算过程，可以得到高表达性的图表示$\boldsymbol{V}^m$. 此时长距离依赖得到充分学习，模态内特征得到进一步聚合. 为继续融合模态间的特征，本文将该图表示$\boldsymbol{V}^m$视为新的节点，并在同一样本不同模态的节点之间建立超边. 以三种模态的样本为例，$\boldsymbol{v}_i^m$表示第i个样本m模态的节点特征，第i个样本的超边u_i表示如下.

$$u_i=\left\{v_i^t,v_i^a,v_i^v\right\} \tag{16}$$

通过上述过程，MORAH模型完成了情绪超边的构建，在所构建的情绪超边中充分融合模态内的情绪内在特征与模态间的情绪耦合特征，便于后续超图学习模型训练中充分利用多模态情绪信息.

3.3 情绪超图学习

在超图学习阶段，MORAH模型首先在初始超边的基础上改进超边扩展策略，使其适用于多模态情绪识别，进而对扩展生成的多级超边进行简易超图卷积，以获取有效多模态情绪特征表示.

3.3.1 情绪超边扩展

经过情绪超边构建阶段，MORAH模型已提取了样本中每个模态的精细特征并建立了情绪超边. 为有效融合多模态特征，本文拟改进超边扩展策略，基于超图卷积方法捕获模态间的多元关系. 具体地，定义超边表示$\boldsymbol{U}$如下所示.

$$U=B^{-1}\cdot W_e\cdot H^T\cdot V \tag{17}$$

其中，$\boldsymbol{B}$是边的度矩阵，$\boldsymbol{W}_e$是边的权重矩阵，$\boldsymbol{H}$表示模态间的关联矩阵，$\boldsymbol{V}$表示单模态图. 由于每条超边的重要性相同，本文将每条超边的权重设为1. 关联矩阵$\boldsymbol{H}$的定义如公式(18)所示，该公式用于表示节点和超边的关系.

$$H(i,j)=\begin{cases}1 & \text{节点}i\text{在超边}j\text{上}\\ 0 & \text{节点}i\text{不在超边}j\text{上}\end{cases} \tag{18}$$

为避免节点嵌入过于平滑，本文对广度优先的超边扩展策略进行改进，由于本文的初始超边之间没有交集，不存在超边间的连通性，因此用超边表示之间的曼哈顿距离计算从超边$\boldsymbol{U}_{ik}$扩展到超边$\boldsymbol{U}_{jk}$的似然性，计算方法如公式(19)所示.

$$\alpha(i,j)=\sum_{k=1}^{d\_e}\left|U_{ik}-U_{jk}\right| \tag{19}$$

接下来对所有超边并行扩展，对于每条超边，本文选择与它扩展似然性最大的k条超边与其组成新的超边，扩展完成后删除重复超边并更新关联矩阵$\boldsymbol{H}$，该过程如图2所示，该图展示了多级超边扩展的过程，以三个样本$\{u_1,u_2,u_3\}$为例，每个样本包含三种模态的特征$\{\boldsymbol{v}_i^t,\boldsymbol{v}_i^a,\boldsymbol{v}_i^v\}$. 左图是经过第一阶段超边构建形成的三条超边，右图是这三条超边经过超边扩展得到的四条超边，其中$u_{1\to 2}$是超边u_1扩展到u_2生成的新超边. 这里需要说明的是，MORAH模型在扩展生成新的超边之后，没有删除原来的超边，

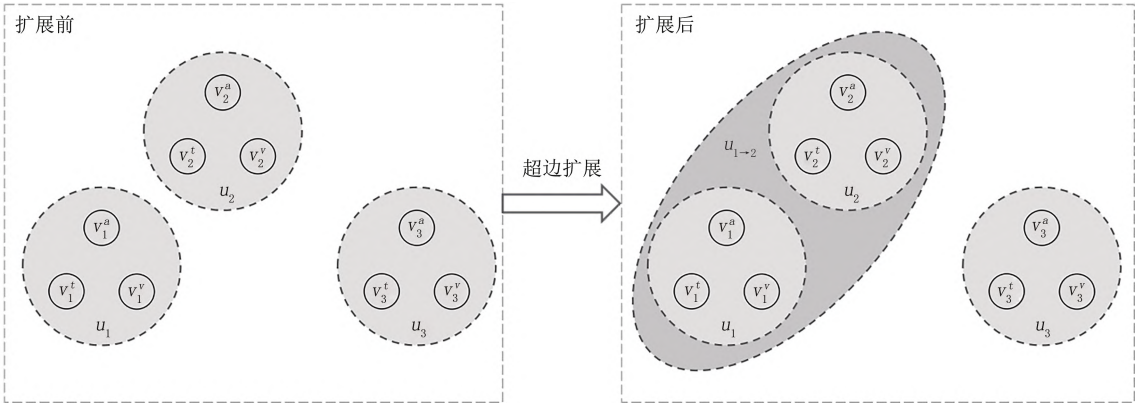

图2　多级超边扩展过程示意图

以此保证同一样本不同模态的节点间仍保持紧密联系．经过超边扩展，可以增加超边数量并得到具有层次的多级超边结构，用于后续超图卷积．

3.3.2　情绪超图卷积

基于多级超边结构，MORAH模型采用超图卷积进一步融合多模态情绪特征．具体地，情绪超图卷积方法如公式(20)所示．

$$V^{(l+1)}=D^{-1}\cdot H\cdot W_e\cdot B^{-1}\cdot H^T\cdot V^{(l)} \tag{20}$$

其中，D表示节点的度矩阵，B表示边的度矩阵，W_e表示边的权重矩阵，H表示模态间的关联矩阵，V表示单模态图．

该方法可以在无需非线性激活和卷积滤波器矩阵的情况下，保证情绪识别效果，降低模型复杂度，以提升训练速度．

多级超边结构使卷积得到的特征同时包含不同层次的信息．情绪超图卷积过程可视为“节点-超边-节点”的聚合顺序，即从节点聚合特征到超边，再从超边聚合特征到节点．与Sun等[30]提出的超边扩展策略不同，MORAH模型在超边扩展后未删除初始超边，如图2所示，节点v_1^a不仅可以聚合来自超边u_1的信息，还可以聚合来自超边u_2的信息．虽然所有超边的默认权重都是1，但同一样本不同模态节点会比不同样本节点聚合的次数多，因此可以保证不同样本的节点特征不会过于平滑．经过情绪超图卷积，每个节点聚合了所在超边其他节点的情绪信息，最终的节点表示如公式(21)所示．

$$P=\frac{1}{L+1}\sum_{l=0}^{L}V^{(l)} \tag{21}$$

经过超图学习阶段，MORAH模型实现了模态内部和模态之间的特征融合．一方面通过动态路由和情绪图卷积两次聚合模态内部特征，充分融合每个模态序列的所有时间步之间的情绪信息．另一方面，通过情绪超边扩展策略以及超图卷积方法实现了对模态间多元情绪关系的充分融合，打破成对关系的限制，更好地利用多元情绪信息．

3.4　情绪分类

在情绪超图卷积的基础上，在MORAH模型的具体实现中采用多层感知器作为分类器，将最终的节点表示输入多层感知器以获得每个样本对应的用户情绪评分，如公式(22)和公式(23)所示．

$$P'=f(W_1P+b_1) \tag{22}$$

$$Y=W_2P'+b_2 \tag{23}$$

其中，f表示非线性激活函数，W_1与b_1、W_2与b_2分别是两个全连接层的参数．

通过上述三个模块，MORAH模型实现了针对未对齐数据的多模态情绪识别，该模型学习过程的整体流程如算法1所示．

算法1. MORAH模型的学习过程

输入：未对齐序列$X^{\{t,a,v\}}\in\mathbb{R}^{d^{\{t,a,v\}}\times T^{\{t,a,v\}}}$；

输出：预测的情绪评分Y．

1. $V\leftarrow GraphCAGE(X^t, X^a, X^v)$
2. $V=\{v_1,v_2,\cdots,v_{3q}\}$，$\mathcal{E}=\{u_1,u_2,\cdots,u_q\}$，$u_j=\{v_j^t,v_j^a,v_j^v\}$
3. $\mathcal{G}=\{V,\mathcal{E}\}$，$H\in\mathbb{R}^{3q\times q}$
4. FOR $u_j\in\mathcal{E}$ DO
5. 　$C_j=\{u_{j1},u_{j2},\cdots,u_{jk}\}$ /*使用公式(19)为u_j选择k个扩展可能性最大的超边*/
6. 　$u_j'=u_j\cup u_{j1}\cup u_{j2}\cup\cdots u_{jk}$
7. 　$\mathcal{E}'=\mathcal{E}\cup u_j'$
8. END FOR
9. 删除$\mathcal{E}'$中的重复超边
10. 更新H
11. FOR $l=1\cdots L$ DO
12. 　$P\leftarrow SimplifyingHGNN(H, V)$

13. END FOR
14. $Y=MLP(P)$
15. RETURN Y

4 实验与结果分析

4.1 数据与评价指标

4.1.1 数据集

本文实验采用两个主流的多模态情绪分析数据集验证所提出模型的性能，这两个数据集分别是CMU-MOSI[31]和CMU-MOSEI[32]. CMU-MOSI包含2199个视频片段，每个片段包含一个句子，内容均为独白，其中的音频采样率为12.5 Hz，视频采样率为15 Hz，文本数据以每个单词分独立单位，以离散的单词嵌入来表示. 每个视频剪辑的情绪强度标注为[－3，＋3]中的一个整数值，反映了情绪的强度，其中＋3表示强烈的积极情绪，其余正数由大到小表示积极情绪由强到弱，－3表示强烈的消极情绪，其余负数由大到小表示消极情绪的由弱到强. 参照MulT[9]的划分，本文使用1284、229和686个视频剪辑分别作为训练集、验证集和测试集.

CMU-MOSEI由来自YouTube的22 856个电影评论视频片段组成，音频采样率为20 Hz，视频采样率为15 Hz. 和CMU-MOSI相同，该数据集使用标注值为[－3，＋3]表示每个视频片段的情绪强度. 同样参照MulT[9]的划分，本文使用16 326、1871和4659个片段分别作为训练集、验证集和测试集.

本文采用与MulT[9]相同的数据处理方法，以便于模型的对比与实验结果的复现. 对于文本模态，使用预先训练的GloVe单词嵌入(glove.840B.300d)[33]获取句子中每个单词的特征向量，对句子中所有单词的特征向量取平均得到句子的特征表示，即得到特征维度为300的文本特征向量. 对于视觉模态，使用Facet来表示35个面部动作单元，记录面部肌肉运动来表示每帧的基本情绪和高级情绪[34-35]. 对于语音模态，使用COVAREP来提取低水平的声学特征[36]，该特征维度为74维. 为实现公平的模型对比，本文使用的各模态特征提取方法均为当前主流特征提取方法，由于所提出模型的创新点主要在于多模态特征的融合，因此我们未在单模态特征提取方面做相应优化，对于各模态特征的进一步优化将作为模型未来优化的重要方向.

4.1.2 评价指标

为实现公平有效的模型性能对比，本文选择与MulT[9]和GraphCAGE[17]等研究工作相同的评价指标，下面简要介绍这些指标.

(1)Acc_7

本文选用Acc_7作为模型在[－3，＋3]七个情绪强度上的准确率评价指标. 首先将模型输出的情绪评分与真实的情绪评分进行四舍五入取整数，然后计算在全部样本中正确预测样本所占的比例.

(2)Acc_2

与Acc_7类似，Acc_2是模型在两个情绪极性上的准确率评价指标，即正数代表积极情绪，非正数代表消极情绪. 首先将模型输出的情绪评分与真实的情绪评分中的正数用1表示，非正数用0表示，然后计算在全部样本中正确预测样本所占的比例.

(3)$F1\ Score$

$F1\ Score$是精确率和召回率的加权平均值，取值范围为[0，1]，在1处取最佳值，0处取最差值. 本文使用二分类的$F1\ Score$，即评分大于0表示积极情绪，评分小于0表示消极情绪，在此基础上，计算情绪二分类的精确率和召回率，进而计算二者的调和平均值. 公式(24)给出了该指标的计算方法，其中p表示精确率，r表示召回率.

$$\mathrm{F1}=2\times\frac{p\times r}{p+r} \tag{24}$$

(4)MAE

平均绝对误差(MAE)是绝对误差的平均值，即真实值y和预测值x之差的绝对值之和的平均值，取值范围为[0，∞]. 该指标不考虑误差方向，仅计算误差的模长. 计算方法如公式(25)所示.

$$MAE=\frac{\sum_{i=1}^{n}|y_i-x_i|}{n} \tag{25}$$

(5)$Corr$

$Corr$是真实值y和预测值x的皮尔逊积矩相关系数，用于反映二者的相关性，取值范围为[－1，＋1]，计算方法为公式(26)所示.

$$Pearsion(x,y)=\frac{\sum xy-\dfrac{\sum x\sum y}{n}}{\sqrt{\left(\sum x^2-\dfrac{(\sum x)^2}{n}\right)\left(\sum y^2-\dfrac{(\sum y)^2}{n}\right)}} \tag{26}$$

4.1.3 对比算法

为充分验证所提出模型的效果，本文与多组先进方法进行了实验性能的对比. 所对比算法可以分为两类：面向非对齐数据的对比模型和面向对齐数据的对比模型. 在面向未对齐数据的对比模型中，

本文选择了早期融合的长短期记忆网络(EF-LSTM)、晚期融合的长短期记忆网络(LF-LSTM)、反复参与的编译嵌入网络(RAVEN)[37]、多模态循环平移网络(MCTN)[38]、基于跨模态注意力和自注意力机制的多模态转换器(MulT)[9]、图胶囊网络(GraphCAGE)[17]、多模态非齐次融合网络(NHFNet)[11]和分层互信息最大化框架(MMIM)[19]为基线模型. 下面对这些方法简要介绍.

(1)EF-LSTM属于早期融合方法,又称特征级融合. 具体做法是先将多模态特征连接,再通过LSTM提取特征并进行预测.

(2)LF-LSTM属于晚期融合方法,又称决策级融合. 具体做法是先应用LSTM提取特征,然后将特征连接起来进行推断预测. 与EF-LSTM类似,由于只是简单的特征连接,该方法无法探索模态内和模态间的动力学特征.

(3)RAVEN[37]考虑了人际交流中会同时使用语言和非语言的行为表达意图,建模这一过程不仅要考虑单词的字面含义,还要考虑这些单词的非语言上下文. 因此首先分析在单词片段中出现的细粒度视觉和语音模态,以此建模非语言子词序列的细粒度结构,并基于非语言线索动态更新单词表征.

(4)MCTN[38]将机器翻译中序列到序列模型(Seq2Seq)的方法借鉴到学习多模态联合表示中,该模型提出从源模态到目标模态的转换会产生一个中间表示,该表示可以捕获两个模态之间的联合信息,用于多模态情绪特征融合.

(5)MulT[9]建立了多个成对的双向跨模态注意块,通过定向成对跨模态注意关注多模态序列之间的相互作用.

(6)GraphCAGE[17]结合了胶囊网络和图卷积,创新了模态内特征提取方法.

(7)NHFNet[11]使用了基于注意力的聚合模块,在MulT的基础上提升了多模态信息的融合能力.

(8)MMIM[19]将互信息引入多模态情绪识别任务,取得目前该任务上的最佳性能.

上述方法中EF-LSTM、RAVEN和MCTN均为对齐数据设计,为了能够使这些模型可以直接应用于未对齐数据,本文参照MulT[9],使用连接时间分类方法(CTC)作为额外的对齐模块[39],并采用了MulT论文中报告的结果. 对于其他对比算法,本文在两个数据集的未对齐数据上均进行了复现,对比算法均按原文给出的参数和代码进行训练,模型结构和参数均与原文相同. 例如,参照原文设置,MULT根据验证集损失值选择模型,GraphCAGE通过测试集准确率选择模型,其他模型的参数设置和结构可参见对应文献,在相应设置下,各种对比算法的性能可以达到最优值. 本文模型通过验证集损失值选择模型,并通过早停法避免过拟合.

4.1.4　实验参数设置

表1展示了在情绪识别任务上,各模型训练过程中主要参数的设置. 对于各个模型所对应的超参数,实验中均进行了网格搜索,以选择最优参数,便于模型性能的公平对比与结果的复现. 对于超边扩展参数k,本文通过二分类的$F1\,Score$进行选择,将区间[1,16]中每个整数值作为k的取值进行模型效果的验证,最终选择$k=8$,模型性能达到最优.

表1　模型中使用的超参数

超参数	CMU-MOSI	CMU-MOSEI
批量大小	16	16
学习率	0.0008	0.001
优化器	Adam	Adam
跨模态注意块	4	4
跨模态注意头	10	8
时序卷积核	1	1
动态路由迭代层数	3	3
超边扩展个数	8	8
训练轮数	38	60

4.2　实验结果

由于所提出MORAH模型没有针对对齐数据做特殊处理,因此MORAH模型主要针对未对齐数据进行多模态情绪识别性能的优化. 本节实验结果主要给出了面向未对齐数据的实验结果,并辅之以对齐数据上的实验结果,在此基础上,通过消融实验验证方法各个功能模型对于性能的贡献程度.

4.2.1　面向未对齐数据的实验结果

面向未对齐数据的实验结果如表2和表3所

表2　CMU-MOSI数据集上的未对齐数据实验

模型	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
EF-LSTM+CTC	31.0	73.6	74.5	1.078	0.542
LF-LSTM	33.7	77.6	77.8	0.988	0.624
RAVEN+CTC	31.7	72.7	73.1	1.076	0.544
MCTN+CTC	32.7	75.9	76.4	0.991	0.613
MulT	34.1	80.3	80.3	0.976	**0.685**
GraphCAGE	35.3	80.3	80.4	0.955	0.659
NHFNet	30.6	76.8	76.9	1.051	0.615
MMIM	24.6	74.2	74.1	1.205	0.529
MORAH	**35.6**	**81.6**	**81.5**	**0.937**	0.679

表3 CMU-MOSEI数据集上的未对齐数据实验

模型	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
EF-LSTM+CTC	46.3	76.1	75.9	0.680	0.585
LF-LSTM	**48.8**	77.5	78.2	0.624	0.656
RAVEN+CTC	45.5	75.4	75.7	0.664	0.599
MCTN+CTC	48.2	79.3	79.7	0.631	0.645
MulT	48.4	80.1	80.6	0.623	0.669
GraphCAGE	48.6	80.8	81.0	0.627	0.653
NHFNet	48.4	81.1	81.3	**0.614**	**0.684**
MMIM	44.3	75.8	76.1	0.715	0.534
MORAH	48.7	**81.3**	**81.5**	0.634	0.675

示，从实验结果可以看出MORAH模型在CMU-MOSI和CMU-MOSEI两个数据集上，尽管个别指标略低于其他基线模型，但大多数评价指标均优于其他对比基线模型，与性能最佳的MulT和GraphCAGE模型相比，经双尾显著性t检验，MORAH模型对多模态情绪识别效果具有显著提升($p<0.05$). 具体地，本文使用二分类的F1值作为主要评估指标，将本文模型分别与MulT和GraphCAGE的二分类F1值进行配对t检验，以0.05的默认显著性水平测试性能改进的显著性，结果表明本文方法在各指标上提升的性能均是显著的.

经过对面向未对齐数据的实验结果进一步分析，可以得出如下发现：

(1)与EF-LSTM、RAVEN和MCTN等未专门针对未对齐数据优化的方法相比，MORAH模型得益于跨模态转换器的加入，可以同时处理对齐和未对齐数据，特别是在未对齐数据上具有显著优势. 因此MORAH模型避免了RNN方法的长距离问题，通过跨模态转换器和自注意转换器关注到模态间和模态内的特征，更好发挥了多模态数据的优势.

(2)与基于注意力机制的MulT模型相比，GraphCAGE和MORAH模型使用胶囊网络和图卷积替代了MulT中用于捕获模态内部特征的自注意模块，使模型融合模态内特征的能力得到提高，在多模态情绪识别任务的多项指标上显著提升了模型效果. 胶囊网络的动态路由技术使得融合过程具有可解释性，揭示了多模态特征融合时模型更加关注的情绪信息.

(3)与基于胶囊网络和图卷积的GraphCAGE模型相比，MORAH模型所提出的多元关系超图和改进的超边扩展策略更加适用于多模态情绪识别场景，在不引入冗余信息的同时实现了对模态内部和模态间特征的充分融合，因此MORAH模型在Acc_2和$F_1\,score$等关键指标上均优于GraphCAGE模型.

(4)与当前最先进的研究方法NHFNet、MMIM相比，我们的方法在两个数据集的主要指标上均表现出色，取得了更好或可比较的结果. MMIM方法在两个数据集上均没有达到最佳效果，这可能是由于互信息的使用对数据要求较高，而本文所用数据缺少相关信息. 在CMU-MOSEI数据集上，MORAH模型性能略低于NHFNet模型，其原因在于MORAH模型对相似超边进行了扩展，使得正确的预测值和真实值更加接近，而错误的预测值和真实值差距变大，因此总体的预测值和真实值的绝对误差增加，二者相关性的降低导致这些指标上略有下降，未来研究可以以此为目标进一步优化MORAH模型.

为进一步验证MORAH模型在两个数据集未对齐数据上的测试集准确率的变化趋势，本文对模型性能的变化趋势作以展示，如图3和图4所示，其中准确率指标采用模型在两个情绪极性上的准确率(Acc_2)表示. 通过图3和图4的实验结果可知，在模

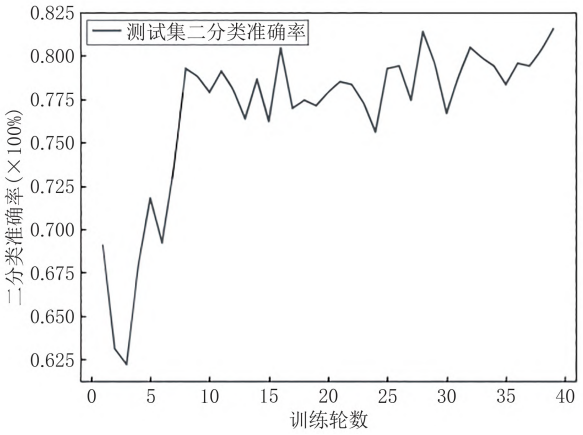

图3 CMU-MOSI数据集未对齐数据的准确率

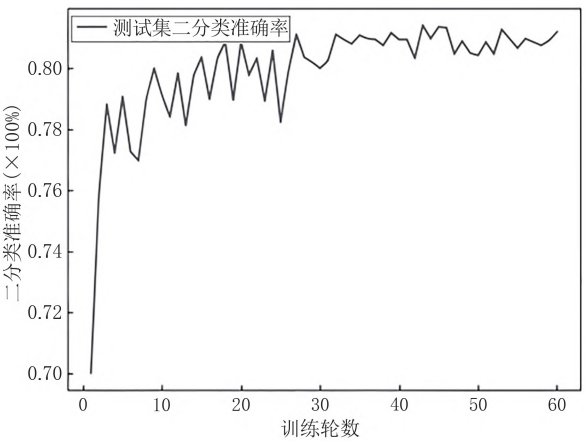

图4 CMU-MOSEI数据集未对齐数据的准确率

型的训练过程中，前期测试集的准确率波动较为剧烈，随着训练轮数增加，模型的整体性能逐渐趋于平缓，并达到收敛.

上述实验结果表明，MORAH模型相比于现有多模态情绪识别方法可以更加准确地识别用户情绪，具有更强的模型性能和模型鲁棒性.

4.2.2　面向对齐数据的实验结果

为进一步验证MORAH模型的性能，本文面向对齐数据进行了实验，实验结果如表4和表5所示，从实验结果可以看出，尽管本文模型并非专门针对对齐数据开展模型优化，MORAH模型仍在Acc_2和F1等指标上相比于最佳的基线模型获得了显著的性能提升.

表4　CMU-MOSI数据集上的对齐数据实验

模型	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
EF-LSTM	33.7	75.3	75.2	1.023	0.608
LF-LSTM	35.3	76.8	76.7	1.015	0.625
RAVEN	33.2	78.0	76.6	0.915	**0.691**
MCTN	**35.6**	79.3	79.1	**0.909**	0.676
MORAH	33.7	**80.0**	**80.2**	1.008	0.626

表5　CMU-MOSEI数据集上的对齐数据实验

模型	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
EF-LSTM	47.4	78.2	77.9	0.642	0.616
LF-LSTM	48.8	80.6	80.6	0.619	0.659
RAVEN	**50.0**	79.1	79.5	0.614	0.662
MCTN	49.6	79.8	80.6	**0.609**	**0.670**
MORAH	46.8	**80.7**	**80.8**	0.646	0.630

然而，由于本文模型没有面向对齐数据进行优化，在其他评价指标如Acc_7上略逊色于专门用于对齐数据的RAVEN模型和MCTN模型，该实验结果的原因在于与未对齐数据相比，对齐数据更容易进行模态内和模态间的特征融合，即在超边构建阶段对齐数据序列已经得到充分融合，接下来的超图融合操作使各样本的节点嵌入过于平滑，所以模型在正负情绪极性分类任务上的准确率高于用于对齐数据的模型，但在更为精细的情绪极性分类任务上的准确率有所下降，同时导致预测值和真实值的平均绝对误差增加，相关系数降低等现象. 如何在多模态情绪超图学习过程中进一步面向对齐数据开展优化，也是未来MORAH模型优化的重要方向.

4.2.3　消融实验

为进一步研究模型中各个模块对多模态情绪识别性能的影响，以验证超图学习在多模态情绪识别中的作用，本文使用未对齐数据集对本文模型进行了消融实验，CMU-MOSI数据集上的消融实验结果如表6所示，CMU-MOSEI数据集的消融实验可以观测到相同的性能变化趋势，由于篇幅限制，在此不再展示.

表6　CMU-MOSI数据集上未对齐数据的消融实验

模型	Acc_7^h	Acc_2^h	$F1^h$	MAE^l	$Corr^h$
MORAH 去掉超边扩展	31.3	79.7	80.0	0.979	0.637
MORAH 去掉超图学习	32.7	80.0	80.3	0.987	0.643
MORAH 去掉多级超边	33.7	80.6	80.6	0.953	0.667
MORAH	**35.6**	**81.6**	**81.5**	**0.937**	**0.679**

从消融实验的结果分析可以看出：

(1)当去掉模型中的超边扩展模块时，即在超边构建阶段完成的初始超边基础上直接进行超图卷积，模型效果大幅下降. 这是由于初始超边建立在同一样本不同模态的节点间，因此每个模态的节点只能聚合同一样本其他模态节点的信息，不能通过超图卷积使相近类别的节点更加接近.

(2)当去掉模型中的超边扩展和超图卷积模块时，即将模型大致简化为了GraphCAGE模型. 但与GraphCAGE模型不同的是，在输入多层感知机之前不是简单连接三种模态的特征，而是用三种模态特征的加权和作为该样本的最终表示. 此时在五个评价指标上均较MORAH模型有一定差距，这是由于图胶囊聚合方法没有充分捕获多模态信息之间的关系，而本文提出的模型通过超边捕获这些多元信息，没有引入多余信息，而且对模态间情绪特征融合得更加充分.

(3)为了论证多级超边在特征融合中发挥的作用，本文在超边扩展完成之后将初始超边删除，在扩展生成的较大超边上进行超图卷积，该模型如图5所示，以四个样本$\{u_1, u_2, u_3, u_4\}$为例，每个样本包含三种模态的特征$\{\boldsymbol{v}_i^t, \boldsymbol{v}_i^a, \boldsymbol{t}_i^v\}$. 左图是经过第一阶段超边构建形成的四个超边，右图是这三个超边经过超边扩展得到的两个新超边$u_{1\to 2}$和$u_{3\to 4}$. 从表6列出的第三行消融实验的结果可以看出，去掉多级超边后模型的所有指标均显著下降. 本文还注意到，与Acc_2指标相比，Acc_7指标下降更加明显，这是由于初始超边的删除使模型中只留下单一的超边，导致同一超边内的节点嵌入过于平滑，因此模型进

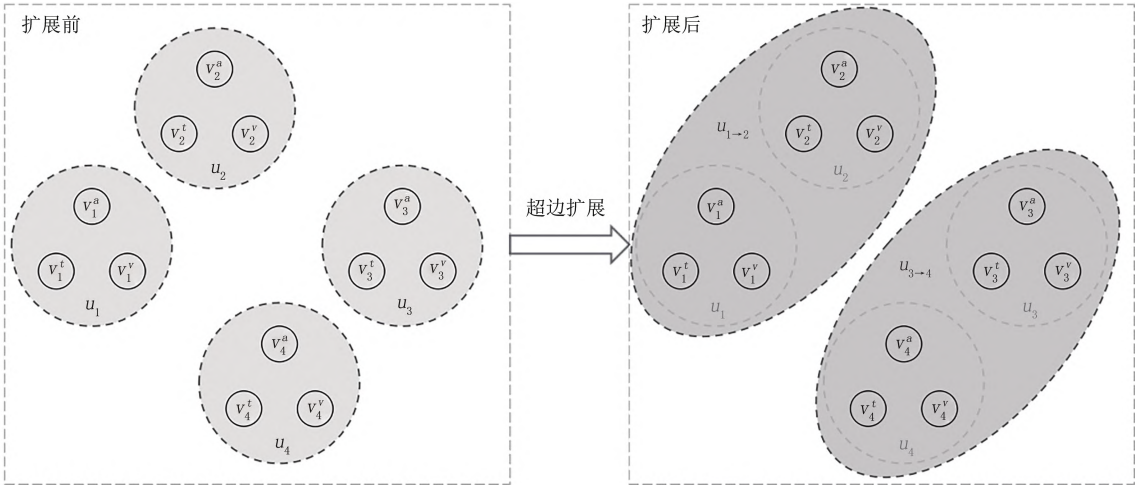

图5 单一超边扩展过程示意图

行精细分类能力减弱得更为显著. 相关研究已证明[9],多模态融合能够带来相比于单模态处理的性能提升,因此本节消融实验主要从模型各模块的性能角度进行验证,未针对单模态处理性能进行验证. 上述消融实验结果进一步证明了所提出方法在多模态融合上的性能提升,该提升来源于融合超图卷积的情绪超边扩展策略,该策略基于超图卷积方法捕获了模态间多元关系,打破了成对二元关系的限制,保证了各模态之间情绪表征信息的互补性.

4.2.4 通用化工具测试

超图学习部分可以抽象化为通用组件,应用于相关的多模态任务上. 本文以MMGCN方法[18]为例,将融合超图卷积的情绪超边扩展策略应用于多模态对话情绪识别任务,实验结果如表7所示.

表7 基于MORAH的对话情绪识别性能($F1^h$)

模型	MELD
MMGCN	57.95
MMGCN+MORAH	58.35

实验结果表明在MELD对话情绪识别数据集上,MORAH模型可以一定程度上提升对话情绪识别的性能,这也表明MORAH模型可以作为通用化工具辅助人机对话等下游自然语言处理任务,增强相关模型对于多模态情绪的识别和感知能力,该研究可以作为未来模型优化和改进的重要方向.

4.3 结 论

为解决多模态情绪识别中特征融合不充分、难以同时捕获三种及以上模态间关系的问题,本文提出了基于超图的多模态情绪识别模型(MORAH). 该模型将多模态特征融合分成两个阶段:情绪超边构建阶段和情绪超图学习阶段. 在情绪超边构建阶段基于图胶囊聚合方法,有效地对各模态低级特征进行了提取和融合,并以此为基础构建多模态情绪超图;在情绪超图学习阶段,本文通过情绪超边扩展策略和情绪超图卷积网络对模态间高级特征进行了有效的融合. 在两个基准数据集上的实验结果表明,MORAH模型在未对齐数据的方法性能方面显著优于最佳的基线方法,获得了更为先进的情绪识别性能,在对齐数据的实验性能上与专门针对对齐数据进行优化的基线方法也具有可比性,消融实验从多个角度验证了促使模型性能提升的原因. 该模型可以推广至n种模态的任务场景,因此下一步的研究工作尝试将用户特征、上下文特征等多元信息加入到情绪超图学习,使模型适用于多人对话等任务场景的情绪识别任务. 同时也可以从各模态精细化特征萃取等角度尝试进一步优化多模态情绪识别的效果.

参 考 文 献

[1] Yadollahi A, Shahraki A G, Zaiane O R. Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 2017, 50(2): 1-33

[2] Semwal N, Kumar A, Narayanan S. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models//Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis. New Delhi, India, 2017: 1-6

[3] Liu Jiming, Zhang Peixiang, Liu Ying, et al. A review of multi-modal emotion analysis technology. Journal of Frontiers of Computer Science and Technology, 2021, 15(07): 1165-1182 (in Chinese)
(刘继明,张培翔,刘颖等. 多模态的情感分析技术综述. 计算机

科学与探索，2021，15(07)：1165-1182)

[4] Zhao S，Jia G，Yang J，et al. Emotion recognition from multiple modalities：fundamentals and methodologies. IEEE Signal Processing Magazine，2021，38(6)：59-73

[5] Zhu Kang，Yan Jingjie，song Yukang，et al. Multi-modal emotion feature recognition method based on multi class kernel canonical correlation analysis，China，2019-05-28(in Chinese)
(朱康，闫静杰，宋宇康等. 基于多类核典型相关分析的多模态情感特征识别方法，中国，2019-05-28)

[6] Zadeh A，Liang P P，Mazumder N，et al. Memory fusion network for multi-view sequential learning//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans，USA，2018：5634-5641

[7] Zadeh A，Liang P P，Poria S，et al. Multi-attention recurrent network for human communication comprehension//Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans，USA，2018：5642-5649

[8] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need//Proceedings of the Annual Conference on Neural Information Processing Systems. Long Beach，USA，2017：5998-6008

[9] Tsai Y H H，Bai S，Liang P P，et al. Multi-modal transformer for unaligned multi-modal language sequences//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence，Italy，2019：6558-6569

[10] Lv F，Chen X，Huang Y，et al. Progressive modality reinforcement for human multi-modal emotion recognition from unaligned multi-modal sequences//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021：2554-2562

[11] Fu Z，Liu F，Xu Q，et al. NHFNET：a non-homogeneous fusion network for multimodal sentiment analysis//2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei，China，2022：1-6.

[12] Wang Y，Wu J，Furumai K，et al. VAE-based adversarial multimodal domain transfer for video-level sentiment analysis. IEEE Access，2022，10：51315-51324.

[13] Arjmand M，Dousti M J，Moradi H. Teasel：a transformer-based speech-prefixed language model. arXiv preprint arXiv：2109.05522，2021.

[14] Yang H，Zhao Y，Qin B. Face-sensitive image-to-emotionaltext cross-modal translation for multimodal aspect-based sentiment analysis//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. AbuDhabi，United Arab Emirates，2022：3324-3335

[15] Yang D，Huang S，Kuang H，et al. Disentangled representation learning for multimodal emotion recognition//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa，Portugal，2022：1642-1651

[16] Yang J，Wang Y，Yi R，et al. MTAG：Modal-temporal attention graph for unaligned human multi-modal language sequences//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2021：1009-1021

[17] Wu J，Mai S，Hu H. Graph capsule aggregation for unaligned multi-modal sequences//Proceedings of the 2021 International Conference on Multi-modal Interaction. Montréal，Canada，2021：521-529

[18] Hu J，Liu Y，Zhao J，et al. MMGCN：Multi-modal fusion via deep graph convolution network for emotion recognition in conversation//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021：5666-5675

[19] Han W，Chen H，Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021：9180-9192

[20] Hu G，Lin T E，Zhao Y，et al. UniMSE：Towards unified multimodal sentiment analysis and emotion recognition//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi，United Arab Emirates，2022：7837-7851

[21] Singh G，Firdaus M，Ekbal A. EmoInt-Trans：A multimodal transformer for identifying emotions and intents in social conversations. IEEE ACM Transactions on Audio Speech Language Processing. 2023，31：290-300

[22] Zhang Y，Wang J，Liu Y，et al. A Multitask learning model for multimodal sarcasm，sentiment and emotion recognition in conversations. Information Fusion，2023，93：282-301

[23] Joshi A，Bhat A，Jain A，et al. COGMEN：COntextualized GNN based Multimodal Emotion recognitioN. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies (NAACL-HLT)，2022：4148-4164

[24] Feng Y，You H，Zhang Z，et al. Hypergraph neural networks//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Honolulu，USA，2019：3558-3565

[25] Yadati N，Nimishakavi M，Yadav P，et al. HyperGCN：A new method for training graph convolutional networks on hypergraphs//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver，Canada，2019：1509-1520

[26] Xia X，Yin H，Yu J，et al. Self-supervised hypergraph convolutional networks for session-based recommendation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. 2021：4503-4511

[27] Wu F，Souza A，Zhang T，et al. Simplifying graph convolutional network//Proceedings of the 36th International Conference on Machine Learning. Long Beach，USA，2019：6861-6871

[28] Chen J，Ma T，Xiao C. Fastgcn：fast learning with graph convolutional networks via importance sampling//Proceedings of the 6th International Conference on Learning Representations. Vancouver，Canada，2018：1-15

[29] Shi H，Zhang Y，Zhang Z，et al. Hypergraph-induced convolutional networks for visual classification. IEEE Transactions on Neural Networks and Learning Systems，2019，30(10)：2963-2972

[30] Sun X, Yin H, Liu B, et al. Multi-level hyperedge distillation for social linking prediction on sparsely observed networks// Proceedings of the Web Conference 2021. Ljubljana, Slovenia, 2021: 2934-2945

[31] Zadeh A, Zellers R, Pincus E, et al. Multi-modal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intelligent Systems, 2016, 31(6): 82-88

[32] Zadeh A, Pu P. Multi-modal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 2236-2246

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

[34] Ekman P, Freisen W V, Ancoli S. Facial signs of emotional experience. Journal of Personality and Social Psychology, 1980, 39(6): 1125-1134

[35] Ekman, Paul. An argument for basic emotions. Cognition & Emotion, 1992, 6(3-4): 169-200

[36] De Gottex G, Kane J, Drugman T, et al. COVAREP: a collaborative voice analysis repository for speech technologies// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy, 2014: 960-964

[37] Wang Y, Shen Y, Liu Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 7216-7223

[38] Pham H, Liang P P, Manzini T, et al. Found in translation: learning robust joint representations by cyclic translations between modalities//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 6892-6899

[39] Graves A, FernándezSantiago, Gomez F. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks//Proceedings of the 23th International Conference on Machine Learning. Pittsburgh, USA, 2006: 369-376

ZONG Lin-Lin, Ph. D., associate professor. Her research interest includes natural language processing, machine learning.

ZHOU Jia-Hui, M. S. candidate. Her main research interests focus on affective computing.

XIE Qiu-Jie, M. S. candidate. Her main research interests focus on affective computing.

ZHANG Xian-Chao, Ph. D., professor. His main research interests focus on machine learning.

XU Bo, Ph. D., associate professor. His main research interest includes affective computing and information retrieval.

Background

With the rapid progress of artificial intelligence, machines need to recognize users' emotions to provide users with a better human-computer interaction experience. Therefore, emotion recognition has become one of the most popular areas of artificial intelligence. Compared with single modal, multi-modal emotion recognition has the advantages of data complementarity and model robustness. In multi-modal emotion recognition, feature fusion between modalities directly affects the effect of emotion recognition. The effective multi-modal fusion methods are mainly based on neural network, attention and graph. The emotion recognition method based on neural network fusion takes Recurrent Neural Network (RNN) as the main module which takes a long time to train and test. The attention-based method collects time information through the self-attention transformer, and the information fusion between different time steps is insufficient. Although the emotion recognition method based on graph fusion avoids the disadvantages of the above methods, most of them use graphs of binary relationships, in which each edge can only represent the relationship between two modalities. When processing data of three or more models, the graph cannot effectively establish the relationship between all modalities without introducing redundant information, so it cannot achieve the ideal feature fusion effect. Therefore, it is necessary to find a better method to model multi-modal data.