DOI:10.19603/j.cnki.1000-1190.2026.02.011 **文章编号**:1000-1190(2026)02-0284-12

融合多子空间及通道注意力的多模态情感分析

米小锋*, 王旭阳, 史浩君

(兰州理工大学计算机与通信学院,兰州 730050)

摘 要:当前多模态情感分析主要依赖复杂的技术对各模态特征进行融合,但由于不同模态特征的分布差异较大,直接融合效果不佳. 为了解决这一问题,本文构建了一种融合多子空间框架及通道注意力的交互学习网络模型. 首先,借助混合神经网络完成各模态特征提取,并利用堆叠的双向长短期记忆网络对话语序列进行语言级表示,将固定大小的话语向量映射到模态不变、模态特定两种不同的表示中,采用时域卷积网络对模态特定表示进行双模态交互;然后,利用通道注意力提取更有意义的信息,并提出跨模态交互双向门控循环神经网络和双模态交互注意力机制,对提取后的模态不变表示向量进行更深层地交互,再经由损失函数完成损失优化;最后,执行基于Transformer的多头注意力机制,获得联合向量,并利用全连接层预测最终结果. 在CMU-MOSI和CMU-MOSEI数据集上进行实验,实验结果表明,该方法能有效消除多模态差异,完成多模态融合.

关键词:多模态情感分析; 混合神经网络; 多模态融合; Transformer; 注意力机制

中图分类号:TP391.41 **文献标识码**:A **开放科学(资源服务)标识码(OSID)**:

早期的情感分析研究主要是基于文本的,然而在当今信息爆炸的时代,通过各种数字平台,人们表达情感的方式日益多样化,从传统的纯文本形式扩展到了包含图像、视频、音频在内的多模态形式. 这种多模态表达不仅丰富了情感交流的层次和深度,也为情感分析带来了前所未有的挑战与机遇. 情感分析,作为自然语言处理(natural language processing, NLP)和多媒体处理交叉领域的一个重要研究方向,旨在自动识别和量化文本、图像、声音等媒介中表达的情感倾向,在理解用户情绪、预测社会趋势、优化用户体验等方面具有广泛的应用价值.

传统的情感分析主要聚焦于单一模态的数据,如纯文本,通过提取情感词汇、句法结构等特征来判断情感极性. 然而,随着社交媒体、在线视频平台等多媒体内容的兴起,用户生成的内容往往融合了多种模态的信息,这些模态之间既相互独立又相互补充,共同构成了完整的情感表达. 因此,如何有效地融合多模态信息,实现更准确、全面的情感分析,成为当前研究的热点和难点.

近年来,研究者在多模态情感分析领域取得了显著的进展,尤其是在深度学习技术的推动下,通过构建复杂的神经网络模型,实现了对图像、文本、音频等模态特征的联合学习和融合. 然而,这些模型往往依赖大量的标注数据,且对于不同模态间的差异性和互补性缺乏深入地理解,导致在实际应用中仍存在性能瓶颈. 此外,多模态情感分析还面临着模态对齐、特征选择、模型泛化能力等方面的挑战.

本研究针对多模态情感分析领域的这些关键问题,提出一种创新的方法,结合多子空间框架和通道注意力机制,更有效地融合多模态数据. 该模型首先利用混合神经网络提取各个模态的特征,然后对话语序列进行语言级表示,将固定大小的向量映射到模态不变和模态特定两种不同的话语表示中. 在模态特定表示中,利用TCN(temporal convolutional network)网络完成双模态交互. 随后,通过通道注意力机制提取关键信息,并提出跨模态交互Bi-GRU(bidirectional gated recurrent unit)和双模态交互注意力机制,对提取后的模态不变表示向量进行更深层的交互. 最后利用基于Transformer的多头注意力机制获得联合向量,并通过全连接层完成任务预测.

1 相关工作

在多模态情感分析中,数据主要来源于社交平

收稿日期: 2025-10-21.
基金项目: 国家自然科学基金项目(62161019).
* 通信联系人. E-mail:1784910394@qq.com.

台的视频，这些视频中的话语按照特定的时间顺序逐一叙述，相互之间具有紧密关联．然而，现有的方法往往将视频中的每一句话视为孤立的个体，忽略了它们之间的上下文联系，这影响了情感分类的准确性．因此，在情感分类过程中充分考虑语境信息可以为分类提供重要的线索．Poria 等[1]提出了一种基于 LSTM（long short-term memory）的模型，该模型利用话语在视频中的邻近环境捕捉上下文关系，以提高分类效果．Chauhan 等[2]提出了一种上下文感知注意力模型，用来提取上下文信息，并根据相邻话语在预测任务中的占比大小计算注意力权重，取得了较好的实验结果．

尽管利用上下文信息建模已取得了显著成果，但这些方法仍然主要关注单一模态中的情感信息，未能充分考虑不同模态之间的相互关联．因此，在多模态情感分析任务中，除了需要捕捉相邻话语的上下文信息，还应重视探索各模态之间的交互关系．缪裕青等[3]设计了一种跨模态门控机制的多模态情感分析模型，成功解决了多模态特征融合不充分的问题．李文雪等[4]提出了一个分层交互融合的多模态情感分析模型，主要使用双向门控循环网络完成各模态的特征提取，然后利用基于门控的注意力机制和自注意力机制完成句子级和篇章级的特征融合，并使用自适应权重判断每个模态对最终情感分析值的贡献．Yang 等[5]提出了一种对比特征融合方法，通过将特征分解为共享特征和模态特异性特征，并将分解后的特征进行融合，形成全面的表示用于情感预测，从而提高了情感分析的性能．

近年来，注意力机制在深度学习领域受到广泛关注，尤其在机器翻译和图像分类等应用中，它与深度学习模型的结合能显著提高性能．研究表明，该机制能集中处理输入数据的关键信息，并忽略无关内容，从而优化模型整体表现．Wang 等[6]使用基于注意力机制的结构，通过结合非文本模态的信息，对文本模态的词向量进行重新优化．Ghosal 等[7]提出了基于 RNN（recurrent neural network）的多模态注意力模型，利用上下文信息及模态间的结合进行情感分类．Xu 等[8]通过结合视觉与文本信息，解决了单模态信息不足的问题，并利用视觉与语义注意力网络，开发了一个双向多层次注意力模型．Xi 等[9]开发了一种创新的方法来提取视频、音频和文本的情感特征，采用多头注意力机制构建了一个多模态情感分析模型，以实现更精确的情感预测．Kumar 等[10]通过门控注意力机制选择性地学习不同模态的交叉特征，并利用自注意力机制来捕捉单词间的长期依赖，从而提升了多模态情感分类的效果．

2017 年，Google 发布了基于注意力机制的 Transformer 模型，该模型专门用于处理长序列信息，并摒弃了传统的卷积神经网络（convolutional neural networks，CNN）结构．利用注意力机制提取模态特征，并支持并行计算，这一设计理念显著提高了处理效率．Delbrouck 等[11]提出了一个基于 Transformer 的联合编码框架用于情感分析，该框架通过模块化的共同注意力和 Glimpse 层对一个或多个模态进行编码．Zadeh 等[12]提出了改进版的分解多模态 Transformer 模型，该模型通过对单模态与多模态交互进行因式分解，并增加了自注意力的数量，提升了情感分类的精度．

2　融合多子空间框架及通道注意力的多模态情感分析模型

本文设计的 MSF-CA 模型总体框架如图 1 所示．该模型主要包含以下步骤：1）完成数据预处理及利用混合神经网络提取各模态的特征；2）将各模态特征进行模态表示学习，分为语言级表示、模态不变表示和模态特定表示．在模态特定表示中，利用 TCN 网络进行双模态交互，然后将模态不变表示和模态特定表示生成的 9 个向量输入通道注意力机制完成关键信息的提取，并设计跨模态交互 Bi-GRU 与双模态交互注意力机制用于对提取后的模态不变表示向量进行深度交互；3）将上述处理后的向量利用基于 Transformer 的多头注意力机制完成模态特征融合，并通过全连接层生成任务预测．

2.1　数据预处理

对视频中的文本特征，使用预训练的中文 BERTbase 模型提取词向量．此模型由 12 个堆叠的 Transformer 层构成，参数量大约 1.1 亿．由于 BERT 的特性，未使用任何分词工具．最终，每个词都被转换为一个具有 768 维的向量表示．

对视频中的视觉，使用 Facet 工具从面部动作单元和面部姿势[13]中提取面部表情特征．对其中的每个采样帧重复这个操作，最终得到特征向量 d_v，MOSI（CMU-multimodal opinion sentiment intensity）的视觉特征维度是 47，MOSEI（CMU-multimodal opinion sentiment emotion intensity）是 35.

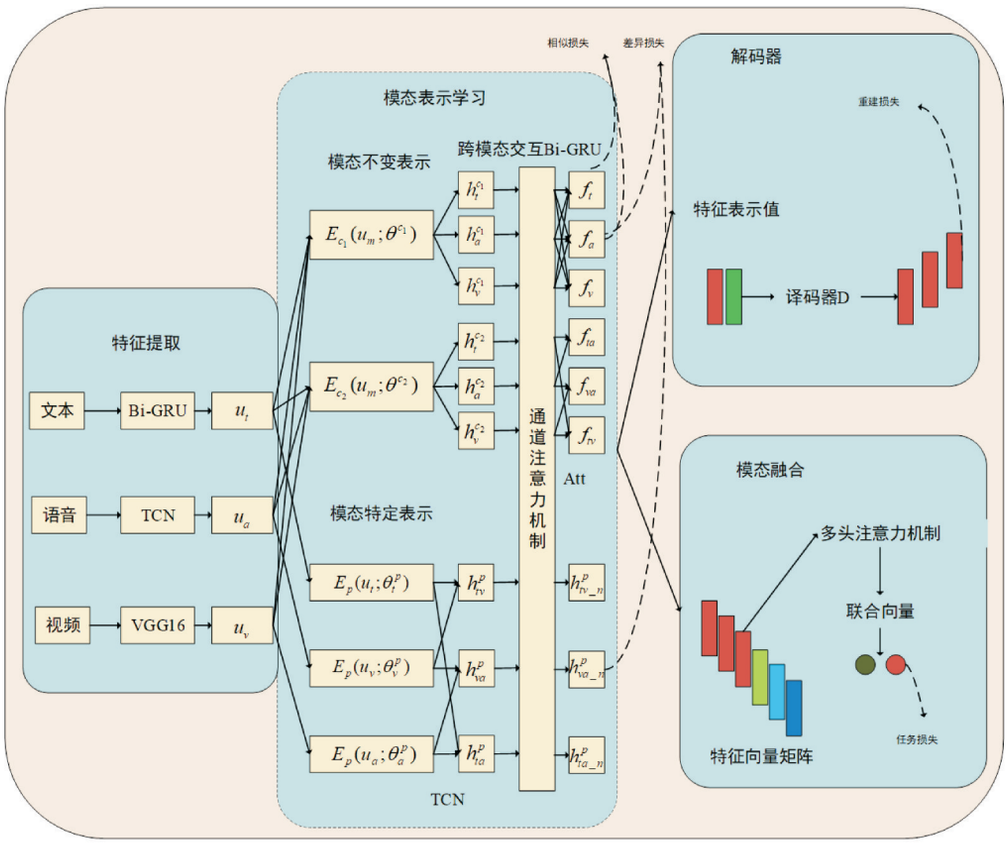

图 1 MSF-CA 模型结构

Fig. 1 Structure of the MSF-CA model

对视频中的语音,本文使用 COVAREP (collaborative voice analysis repository for speech technologies)[14],COVAREP 是一种声学分析框架,它提取的特征有 12 梅尔频率倒谱系数、声门源参数等,得到特征向量 d_a,MOSI 和 MOSEI 数据集的特征维度为 74.

2.2 特征提取及问题定义

在视觉特征提取中,卷积模型(如 CNN)是常用的技术之一. 传统的 CNN 主要借助较深的网络结构和较大的卷积核,这不仅可能导致梯度消失或爆炸现象,还会增加计算负担. 为了应对这些问题,本文引入了 VGG16(visual geometry group 16)模型. 该模型采用三个 3×3 的卷积核等效替代一个 7×7 的卷积核,以及用两个 3×3 的卷积核等效替代一个 5×5 的卷积核. 这样做的目的是在不改变感受野大小的基础上增加网络的层数,通过叠加更多的卷积层来增强模型的非线性映射能力,从而提升网络的整体性能. 语音特征由 TCN 网络提取,文本特征由 Bi-GRU 网络提取.

本文旨在探讨如何通过多模态信息来检测视频中的情感极性. 将视频按照话语单元进行切分,对于较长的视频进行裁剪,其中,话语定义为以呼吸或语言停顿为界限的言语片段. 每个言语片段 u 被视为模型的输入,分为文本、视觉和声学特征. 这些提取的模态特征表示分别是:$u_t \in R^{T_t \times d_t}$、$u_v \in R^{T_v \times d_v}$、$u_a \in R^{T_a \times d_a}$,其中,$T_m$ 表示话语序列的长度,d_m 表示模态各自特征的维度,主要任务是根据这些数据预测相应的情绪类别. 此序列为 $u_{m \in \{t,v,a\}}$,其任务是从预定义的一组 C 类别 $y \in R^C$ 或连续强度变量 $y \in R$ 中预测多模态数据的情感倾向,其中,C 是类别的数量.

2.3 模态对齐

本文使用时间对齐法进行模态对齐,具体步骤为:1) 在编码过程中,应为每个数据块(例如视频帧、音频包、文本段落)精确标记开始和结束时间戳,这可以通过时间戳服务器来实现;2) 确定以国际标准时间作为时间标注的基准,确保所有媒体元素都基于相同的时间基准进行同步;3) 根据每个数据块的时间戳,将文本、音频和图像在时间线上进行对齐. 对齐完成后,进一步进行插值和平滑处理,以确保对齐的准确性和连续性.

2.4 模态表示学习

2.4.1 语言级表示 利用堆叠式双向长短期记忆网络,将各个模态 $m \in \{t,v,a\}$ 对应的话语序列

$u_m \in R^{T_m \times d_m}$ 转化成固定大小的向量 $u_m \in R^{d_h}$. 此网络结构主要包括LSTM层和以ReLU为激活函数的全连接层. 其最终隐藏状态表示与完全连接的密集层相结合，得到 u_m，公式为：

$$u_m = sLSTM(u_m; \theta_m^{lstm}). \tag{1}$$

本文选择使用LSTM而非GRU的原因是LSTM能有效处理长期依赖问题导致的梯度消失和梯度爆炸. 此外，虽然GRU结构较简单、参数较少且更易收敛，但在数据集庞大且复杂的情况下，LSTM的性能通常优于GRU.

2.4.2 模态不变和模态特定表示　每个话语向量 u_m 都被映射到两种不同的表示：1) 模态不变表示，模态不变表示构建两个分布对齐的子空间，分别是共享子空间和辅助共享子空间，辅助共享子空间旨在提升跨模态表示学习中模态间共性的强化效果，补充独立共享子空间在促进效能上的不足；2) 模态特定表示，每个模态一个子空间. 模态不变表示学习共享表示的相似性约束，有助于最小化异质性差距并促进多模态融合；模态特定表示能有效地捕获每个模态的独特特征. 通过这两个表示，提供了有效融合所需的整体视图. 借助编码函数给定模态 m 的话语向量 u_m，学习隐藏模态不变表示 $h_m^{c_1} \in R^{d_h}$、$h_m^{c_2} \in R^{d_h}$ 和模态特定表示 $h_m^p \in R^{d_h}$，具体的计算公式为：

$$\begin{cases} h_m^{c_1} = E_{c_1}(u_m; \theta^{c_1}), \\ h_m^{c_2} = E_{c_2}(u_m; \theta^{c_2}), \\ h_m^p = E_p(u_m; \theta_m^p). \end{cases} \tag{2}$$

其中，E_{c_1}、E_{c_2} 在三种模态中共享参数 θ^{c_1}、θ^{c_2}，而 E_p 为每种模态分配单独的参数 θ_m^p.

2.4.3 模态特定表示向量交互　将上述得到的三个模态特定表示特征向量 h_t^p、h_v^p、h_a^p 进行双模态特征融合，计算公式如式(3)～(5)所示：

$$H_{tv}^p = \mathrm{ReLU}([h_t^p \oplus h_v^p] \cdot W_l^{tv} + b_l^{tv}), \tag{3}$$

$$H_{va}^p = \mathrm{ReLU}([h_v^p \oplus h_a^p] \cdot W_l^{va} + b_l^{va}), \tag{4}$$

$$H_{ta}^p = \mathrm{ReLU}([h_t^p \oplus h_a^p] \cdot W_l^{ta} + b_l^{ta}), \tag{5}$$

其中，$\oplus$ 代表两个矩阵的拼接，W_l^{tv}、W_l^{va}、$W_l^{ta} \in R^{2d_h}$，b_l^{tv}、b_l^{va}、$b_l^{ta} \in R^{d_h}$. 在该网络中，首先对两个单模态特征矩阵执行拼接操作，随后进行特征降维处理. 通过这一过程，最终生成了三个双模态特征矩阵. 将三个双模态特征矩阵输入到TCN网络层中，用以提取双模态序列特征. 相关的计算公式如式(6)～(8)所示.

$$h_{va}^p = N_{\mathrm{TC}}(H_{va}^p), \tag{6}$$

$$h_{tv}^p = N_{\mathrm{TC}}(H_{tv}^p), \tag{7}$$

$$h_{ta}^p = N_{\mathrm{TC}}(H_{ta}^p). \tag{8}$$

2.4.4 通道注意力机制　Woo等[15]改进并实现了一种卷积块注意力机制(convolutional block attention module，CBAM)，这种机制结合了通道注意力模块(channel attention module，CAM)和空间注意力模块(spatial attention module，SAM). 该机制可以在通道和空间上分别计算注意力，节省了参数和计算资源.

本文在构建模型时，利用了CBAM中的通道注意力机制CAM. 该机制首先对输入特征进行基于空间的全局最大池化(global max pooling)和全局平均池化(global average pooling)；然后将处理后的特征输入到一个共享的多层感知机(multi-layer perception，MLP)中学习，该网络使用了ReLU激活函数；最后，将多层感知机输出的特征相加，并通过Sigmoid激活函数进行激活处理，生成最终的通道注意力特征 M_c. CAM模型结构如图2所示，计算公式为：

$$M_c(F) = \mathrm{Sigmoid}(P_{\mathrm{ML}}(\mathrm{AvgPool}(F)) + P_{\mathrm{ML}}(\mathrm{Maxpool}(F))). \tag{9}$$

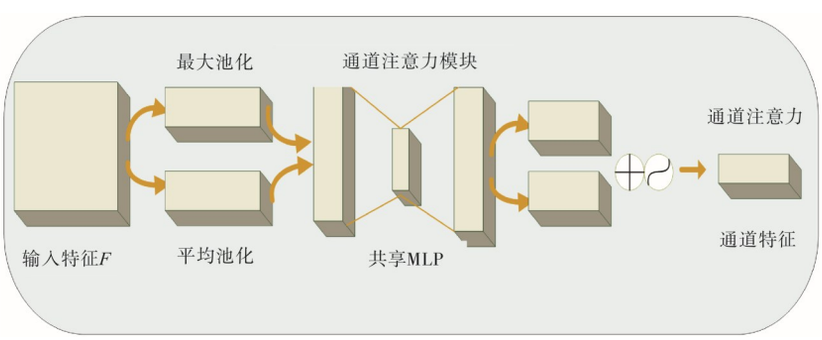

图2　通道注意力结构

Fig. 2　Channel attention structure

利用通道注意力模块，MSF-CA模型能够关注关键信息. 具体过程是将模态特定表示和模态不变表示中的文本、声学、视觉等9个不同的特征视为互异的通道，堆栈为一个通道为9的特征，经

过通道注意力机制处理,得到新特征,这个新特征的维度为9,然后将其拆解为9个经过通道注意力模块不同模态的新特征:{$h_{t\_n}^{c_1}$、$h_{v\_n}^{c_1}$、$h_{a\_n}^{c_1}$、$h_{t\_n}^{c_2}$、$h_{v\_n}^{c_2}$、$h_{a\_n}^{c_2}$、$h_{tv\_n}^{p}$、$h_{va\_n}^{p}$、$h_{ta\_n}^{p}$}.

2.4.5 跨模态交互Bi-GRU 为了深层次减少模态之间的差异,本文在构建模型时,设计了一种跨模态交互Bi-GRU. 旨在对模态不变表示中共享子空间的特征向量经由通道注意力机制处理后进一步融合,实现各模态深层次的交互. 以文本特征为例,首先,分别将文本与语音、文本与视觉特征进行拼接;然后将拼接的两个双模态特征(文本+语音、文本+视觉)与文本特征再一次进行融合,并完成降维处理,输出文本补充特征. 具体融合过程如图3所示,拼接计算过程如式(10)~(11)所示.

$$F_{TV}=\mathrm{ReLU}([h_{t\_n}^{c_1}\oplus h_{v\_n}^{c_1}]\cdot W_t^{VT}+b_t^{VT}),\quad(10)$$

$$F_{TA}=\mathrm{ReLU}[h_{t\_n}^{c_1}\oplus h_{a\_n}^{c_1}]\cdot W_t^{TA}+b_t^{TA}),\quad(11)$$

其中,·代表矩阵乘法,⊕代表矩阵拼接,W_t^{TV}、$W_t^{TA}\in R^{2d_h}$,b_t^{TV}、$b_t^{TA}\in R^{d_h}$,F_{TA}、$F_{TV}\in R^{d_h}$.

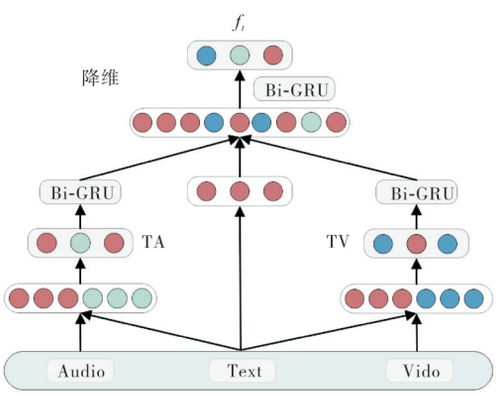

图3 跨模态交互Bi-GRU

Fig. 3 Cross-modal interaction Bi-GRU

利用双模态交互网络提取的特征仅进行了简单级联. 为了更深入地挖掘这些级联特征的关联语义,本文采用了双向门控循环神经网络. 该方法有效融合了两种模态的信息,并将Bi-GRU处理后的双模态数据用作文本特征的补充. 具体计算过程见公式(12)~(13).

$$G_{TV}^{*}=U_{\text{Bi-GR}}(F_{TV}),\quad(12)$$

$$G_{TA}^{*}=U_{\text{Bi-GR}}(F_{TA}),\quad(13)$$

其中,G_{TV}^{*}、$G_{TA}^{*}\in R^{d_h}$是Bi-GRU提取后的双模态情感特征.

接着,将两种多模态文本信息与初始文本信息进行拼接,再传入Bi-GRU网络提取特征信息. 计算公式如下:

$$\begin{cases}F_t=\mathrm{ReLU}([h_{t\_n}^{c_1}\oplus G_{TA}^{*}\oplus G_{VT}^{*}]W_t^{T}+b_t^{T}),\\ f_t=U_{\text{Bi-GR}}(F_t).\end{cases}\quad(14)$$

视觉和语音的融合特征获取方式和文本一致,计算过程如式(15)~(16)所示.

$$\begin{cases}F_a=\mathrm{ReLU}([h_{a\_n}^{c_1}\oplus G_{AT}^{*}\oplus G_{AV}^{*}]W_a^{A}+b_a^{A}),\\ f_a=U_{\text{Bi-GR}}(F_a).\end{cases}\quad(15)$$

$$\begin{cases}F_v=\mathrm{ReLU}([h_{v\_n}^{c_1}\oplus G_{VT}^{*}\oplus G_{VA}^{*}]W_v^{V}+b_v^{V}),\\ f_v=U_{\text{Bi-GR}}(F_v).\end{cases}\quad(16)$$

其中,W_t^{T}、W_a^{A}、$W_v^{V}\in R^{3d_h}$,b_t^{T}、b_a^{A}、$b_v^{V}\in R^{d_h}$,经过多次融合得到f_t、f_a、$f_v\in R^{d_h}$,⊕代表拼接. 将f_t、f_a、f_v作为文本、语音和面部视觉特征参与最终的情感分类.

2.4.6 双模态交互注意力机制 双模态交互注意力机制能够融合不同模态的数据,有效捕捉并利用模态之间的相关性,从而提高模型的整体性能和理解能力. 本文借助双模态交互注意力机制,旨在对模态不变表示中辅助共享子空间的特征向量经由通道注意力机制处理后进一步融合,识别对情感极性判断至关重要的模态信息. 以文本和视觉模态为例,计算公式如下:

$$\begin{cases}M_1=h_{t\_n}^{c_2}\cdot(h_{v\_n}^{c_2})^{\mathrm{T}},\\ M_2=h_{v\_n}^{c_2}\cdot(h_{t\_n}^{c_2})^{\mathrm{T}},\end{cases}\quad(17)$$

其中,$(h_{t\_n}^{c_2})^{\mathrm{T}}$、$(h_{v_n}^{c_2})^{\mathrm{T}}$分别代表相应矩阵的转置,·代表矩阵相乘. 得到两个模态信息的交互矩阵后,借助Softmax函数来运算交互矩阵M_1与M_2中的概率分布分数,计算公式如(18)~(19)所示.

$$N_1(i,j)=\frac{\mathrm{e}^{M_1(i,j)}}{\sum_{k=1}^{u}\mathrm{e}^{M_1(i,k)}},\quad(18)$$

$$N_2(i,j)=\frac{\mathrm{e}^{M_2(i,j)}}{\sum_{k=1}^{u}\mathrm{e}^{M_2(i,k)}},\quad(19)$$

其中,$i,j=1,2,\cdots,u$,$N_1(i,j)$代表相应模态的第i个特征与第j个特征的相关概率分数,其值越大,代表相关性越强.

首先,将上述计算获得的注意力矩阵N_1、N_2分别与文本和视觉特征矩阵进行矩阵乘积,生成注意力矩阵O_1、O_2;随后,将该矩阵O_1、O_2分别与文本和视觉特征矩阵进行Hadamard乘积,以获取交互注意力矩阵$A_1\in R^{d_h}$、$A_2\in R^{d_h}$;最后,拼接这两个交互注意力矩阵,得到融合后的文本和视觉双模态特征. 计算公式如式(20)~(23)所示.

$$\begin{cases}O_1=N_1\cdot h_{t\_n}^{c_2},\\ O_2=N_2\cdot h_{v\_n}^{c_2}.\end{cases}\quad(20)$$

$$\begin{cases} A_1 = O_1 \odot h_{t\_n}^{c_2}, \\ A_2 = O_2 \odot h_{v\_n}^{c_2}. \end{cases} \tag{21}$$

$$F_{tv} = A_1 \oplus A_2, \tag{22}$$

$$f_{tv} = \mathrm{Re}\,lu(F_{tv} W + b), \tag{23}$$

其中，$\cdot$代表矩阵乘法、$\odot$代表Hadamard乘积、$\oplus$代表矩阵拼接，最后得到$f_{tv} \in R^{d_h}$. 文本、语音及视觉、语音的融合方法与上述方法一致. 经过上述操作后，得到一个新特征f_t、f_a、f_v、f_{tv}、f_{va}、f_{ta}、$h_{tv\_n}^p$、$h_{va\_n}^p$、$h_{ta\_n}^p$，在这些特征的基础上再去分析情感.

2.5　模态融合与任务预测

2.5.1 模态融合　完成上述操作后，应用Transformer，执行自注意力的计算，将九个模态向量串联起来. Transformer使用了点积自注意力的方法. 计算公式为：

$$\mathrm{Attention}(Q, K, V) = \mathrm{soft\,max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{24}$$

其中，Q、K、V是查询、键和值矩阵. 这三个矩阵是由一个输出经过三次不同计算得到的. 利用Transformer计算多个并行注意力，其中每个注意力机制的输出被称为一个头(head). head_i计算公式如下：

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^q, KW_i^k, VW_i^v), \tag{25}$$

其中，$W_i^{q/k/v}$表示对查询矩阵、键矩阵和值矩阵的变换矩阵，其主要功能是将线性矩阵投影到局部空间.

具体融合过程如下. 首先，将上述九个模态特征叠加为一个矩阵，如式(26)所示. 然后，对这些模态特征进行多头自注意，使每个向量意识到其他的跨模态(和跨子空间)表示. 这样做可以让每个表征从同伴表征中诱导潜在的信息，这些信息对整体的情感取向是协同的，最终Transformer生成一个新矩阵$\bar{M}$，如式(27)所示. 对于自注意力，设置$Q = K = V = M \in R^{6 \times d_h}$，这里的每个head都是根据公式(25)计算，$\oplus$代表串联，$\theta^{att} = \{W^q, W^k, W^v, W^o\}$，$\bar{M}$的表达式如式(28)所示.

$$M = [f_t, f_a, f_v, f_{tv}, f_{va}, f_{ta}, h_{tv\_n}^p, h_{va\_n}^p, h_{ta\_n}^p] \in R^{9 \times d_h}, \tag{26}$$

$$\bar{M} = [\bar{f}_t, \bar{f}_a, \bar{f}_v, \bar{f}_{tv}, \bar{f}_{va}, \bar{f}_{ta}, \bar{h}_{tv\_n}^p, \bar{h}_{va\_n}^p, \bar{h}_{ta\_n}^p] \in R^{9 \times d_h}, \tag{27}$$

$$\bar{M} = \mathrm{MultiHead}(M; \theta^{att}) = (\mathrm{head}_1 \oplus \cdots \oplus \mathrm{head}_n) W^o. \tag{28}$$

2.5.2 任务预测　通过Transformer串联构造一个联合向量$h^{out} = [\bar{f}_t \oplus \cdots \oplus \bar{h}_{ta_n}^p]$，任务预测由全连接层方程$\hat{y} = G(h^{out}; \theta^{out})$完成.

2.6　损失函数

整个模型的损失函数由四部分组成，分别是相似损失、差异损失、重构损失和任务损失. 计算公式如下：

$$\lambda = \lambda_{\mathrm{task}} + \alpha\lambda_{\mathrm{sim}} + \beta\lambda_{\mathrm{diff}} + \gamma\lambda_{\mathrm{recon}}, \tag{29}$$

其中，α、β、γ代表在总的损失中各项损失所占据的权重比例.

2.6.1 相似损失　为了减小不同模态数据间的差异，MSF-CA模型采用同一编码器来学习多种模态的特征，并使用相似损失函数度量两个特征向量之间的相似性，其目标是在训练过程中最小化该损失. 此处应用了中心距差异(central moment discrepancy，CMD)，CMD是令X和Y为有界随机样本，其概率分布为p和q在区间$[a, b]^N$上，中心距差异正则化器CMD_K被定义为CMD度量的经验估计，$C_k(X) = E(x - E(X))^k$. CMD是一种先进的距离度量方法，通过匹配两种表示的顺序矩差来衡量它们之间分布的差异. 计算公式如下：

$$\mathrm{CMD}_K(X, Y) = \frac{1}{|b-a|}\left\| E(X) - E(Y) \right\|_2 + \sum_{k=2}^{K} \frac{1}{|b-a|^k}\left\| C_k(X) - C_k(Y) \right\|_2, \tag{30}$$

其中，$E(X)$代表样本的经验期望，$E(X) = \frac{1}{|X|}\sum_{x \in X} x$，$C_k(X)$是$X$所有坐标的第$k$个样本中心矩的向量. 对于不同模态的特征，都使用中心距差异计算相似度，最后计算出平均值，计算公式如下：

$$\lambda_{\mathrm{sim}} = \frac{1}{3} \sum_{(m_1, m_2) \in \{(t,a),(t,v),(a,v)\}} \mathrm{CMD}_K(h_{m_1}^{c_1}, h_{m_2}^{c_1}). \tag{31}$$

在相似损失的作用下，模态不变表示子空间的编码器会学习到不同模态特征之间相关联的信息.

首先，本文选择CMD而不是最大平均差异(maximum mean discrepancy，MMD)或KL散度(kullback-leibler divergence)的原因是CMD不仅是一种流行的度量[16]，而且它可以执行高阶矩的显式匹配且无需复杂的距离和核矩阵计算. 其次，尽管对抗损失(adversarial loss)提供了另一种相似性训练方案，但其鉴别器和共享编码器参与极大极小博弈，这会增加额外的参数和复杂度，因此选择运算简单的CMD.

2.6.2 差异损失　差异损失的主要作用是确保两种不同的表示能够高效捕捉输入数据的特征. 在

每一批训练数据中,将矩阵$H_m^{c_1}$和H_m^p转变为零均值矩阵,其行代表每个话语对应模态m的隐藏向量,分别为$h_m^{c_1}$和h_m^p. 接着,计算这些模态向量对的正交性约束$\left\| H_m^{c_1{}^T} H_{\mathrm{m}}^p \right\|_F^2$. 在此式中,$\|\cdot\|_F^2$表示弗罗贝尼乌斯范数的平方. 除了对模态不变向量和模态特定向量之间施加约束,还在模态特定向量之间引入了正交约束. 计算公式如下:

$$\lambda_{\mathrm{diff}} = \sum_{m\in(t,v,a)} \left\| H_m^{c_1{}^T} H_m^p \right\|_F^2 + \sum_{(m_1,m_2)\in\{(t,a),(t,v),(a,v)\}} \left\| H_{m_1}^{p^T} H_{m_2}^p \right\|_F^2. \tag{32}$$

2.6.3 重建损失 此损失是针对差异损失设定的. 由于差异损失的反向传播是强制执行的,为了避免模型最小化差异损失,故让编码器输出不同特征之间琐碎的信息,构造了重建损失. 重建损失使得编码器在输出特征的时候,尽可能捕捉到模态资料的主要信息和更多细节. 为此设计了一个具有θ^d参数的译码器D,当输入同一模态的三个子空间内的向量时,译码器会译码出与该模态初始特征相似的特征:$\hat{u}_m = D(h_m^{c_1} + h_m^{c_2} + h_m^p;\ \theta^d)$. 比较译码器译码出的不同模态的特征$\hat{u}_m$和初始特征$u_m$,并计算它们的均方误差损失,计算公式如下:

$$\lambda_{\mathrm{recon}} = \frac{1}{3}\left(\sum_{m\in\{t,v,a\}} \frac{\left\| u_m - \hat{u}_m \right\|_2^2}{d_h} \right), \tag{33}$$

其中,$\|\cdot\|_2^2$代表L^2范数的平方.

2.6.4 任务损失 本文在情感分析的过程中实现了分类任务和回归任务. 对于分类任务,使用了交叉熵损失进行评估,计算公式如下:

$$\lambda_{\mathrm{task}} = -\frac{1}{n}\sum_{i=0}^{n} y_i \cdot \log \hat{y}_i, \tag{34}$$

其中,n代表一轮训练话语的话语量,y_i代表真实数据集的情感标记,$\hat{y}_i$代表模型预测的情感标记. 对于回归任务,使用均方误差损失,计算公式如下:

$$\lambda_{\mathrm{task}} = -\frac{1}{n}\sum_{i=0}^{n} \left\| y_i - \hat{y}_i \right\|_2^2. \tag{35}$$

通过任务损失函数,使得模型可以得到分类任务和回归任务真实的情感类别和情感值.

3 实验设置与结果分析

3.1 实验数据集

实验采用了卡内基梅隆大学从YouTube视频网站获取的两个数据集:CMU-MOSI和CMU-MOSEI,用于MSF-CA模型的实验评估. MOSI数据集包含93个视频,涵盖2 198个话语视频片段,由89个不同的说话者提供. 作为MOSI的扩展,MOSEI数据集包含3 228个视频和23 453个带注释的视频片段,由1 000多名贡献者提供,涉及250个不同主题. MOSI和MOSEI数据集都是七分类的情感数据集,其区间是[-3,+3],其中,-3表示强烈消极情绪、-2表示较强消极情绪、-1表示普通消极情绪、0表示中性情绪、+1表示普通积极情绪、+2表示较强积极情绪、+3表示强烈积极情绪. 具体的数据集详细统计如表1所示.

表1 实验数据集

Tab. 1 Experimental dataset

数据集	训练集	测试集
MOSI	1 283	686
MOSEI	16 315	4 654

3.2 实验环境配置与参数

实验选择Google colab服务器来训练模型,编程语言为Python 3.7,具体环境如表2所示.

表2 实验环境参数设置

Tab. 2 Experimental environment parameter settings

名称	版本号
GPU	NVIDIA-SMI 460.32.03
显卡	Tesla T4
内存大小	16 G
编程框架	Pytorch框架
编程语言	Python 3.7
操作系统	Windows 11

为了获取模型参数最优的性能组合,所有参数使用Adam优化器进行更新,经过多次实验,依据实验损失率和准确度进行动态调参,确定并选择了提供最佳结果的最佳参数. 具体实验超参数如表3所示.

表3 实验超参数设置

Tab. 3 Experimental hyperparameter settings

名称	MOSI	MOSEI
batch_size	32	64
optimizer	Adam	Adam
learning rate	0.001	0.000 1
dropout	0.5	0.5
epoch	50	70
embeding-size	300	300
hidden-size	128	128

3.3 二分类标签

二分类指标区分为非负/负(non-neg/neg)和正/负(pos/neg)标签，非负/负标签包括0和负数，正/负标签则仅包括正数和负数．实验结果表格中，“－/－”左侧代表非负/负标签，右侧代表正/负标签．

3.4 评价指标

模型评价指标是用来系统评价模型性能的参考，针对不同的任务有着不同的评价指标．在分类任务中，通过利用准确率(accuracy)和$F1$分数($F1$-score)来评估模型的效果，较高的准确率和$F1$分数通常表明模型的性能更佳．对于回归任务，则采用平均绝对误差(mean absolute error, MAE)和皮尔逊相关系数(Pearson correlation coefficient, Corr)作为评估指标．MAE的优点是可以精准反映实际预测误差的大小，主要利用它来评价真实值和拟合值的相似程度．MAE的值M_{AE}越趋向于0，代表模型拟合能力越强．Corr是用于评估两个变量间线性关系强度的统计量，其值的增加表明变量之间的相关性增强．计算公式如式(36)～(39)所示．

$$\mathrm{Accuracy}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}+N_{\mathrm{FP}}+N_{\mathrm{TN}}},\quad(36)$$

$$F1-\mathrm{score}=\frac{2\times N_{\mathrm{TP}}}{2\times N_{\mathrm{TP}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}},\quad(37)$$

上式中下标符号意义分别是真正例(TP)、假正例(FP)、真负例(TN)、假负例(FN)．其中，N_{TP}和N_{TN}分别表示正确分类的正样本和负样本的数量，而N_{FP}和N_{FN}则表示被错误分类的样本数量．

$$M_{\mathrm{AE}}=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i|,\quad(38)$$

$$C_{\mathrm{orr}}=\frac{\mathrm{cov}(X,X_i)}{\sigma_x\sigma_Y},\quad(39)$$

其中，平均绝对误差表达式中，n代表数据点的数量，y_i是第i个数据点的真实值，$\hat{y}_i$是模型对第i个数据点的预测值，$|\cdot|$表示取绝对值．皮尔逊相关系数表达式中，cov代表取协方差，σ表示标准差．

3.5 实验结果与分析

本节首先呈现了在MOSI和MOSEI数据集上的对比实验结果．随后，通过两组实验验证了模型的有效性和参数设置的合理性，这两组实验分别是消融实验及参数性能分析实验．

3.5.1 对比实验　为了对本文所提出的模型进行评估，特与下列模型进行比较，以验证本文所提出模型的性能．

1) ICCN[17]：一种用于多模态语言分析的深度学习模型，通过构建关联学习模块，探索文本、音频和视频数据之间的相关性，从而发掘各个模态之间的关系．

2) BBFN[18]：通过文本和音频双模态数据进行情感分析，并采用相关性控制技术来挖掘两种模态之间的相互关系．

3) CubeMLP[19]：通过建立一种共享的多层感知神经网络模型，探索文本、视觉和音频模态之间的相关性．

4) MCGMF[3]：构建了一种多模态情感分析模型，采用跨模态门控机制，解决了多模态特征融合不足的问题．

5) HFM-AM[4]：设计了一种基于注意力机制的多层次交互融合的多模态情感分析模型，利用自注意机制交互融合策略分别提取属于句子级和篇章级层次的不同特征，有效解决了各模态融合的异构性．

6) MISA[20]：此模型将各个模态分别映射至两个独立的子空间．第一个子空间致力于实现模态不变性，促使不同模态的表征学习其共通特征，并有效缩减模态间的差异；而第二个子空间则专注于模态的特异性，它独特地对应于每个模态，并精准捕捉其独有的特征信息．

7) MSSA[21]：此模型是在MISA模型的基础上添加辅助共享子空间，并加强重建损失的作用．

8) Hycon-B[22]：设计了多种对比学习方法，深入探究模态内与模态间的交互关系，并同时分析样本间以及类别间的关联性．

9) ICDN[23]：整合跨模态Transformer架构与自监督学习策略，以提取单模态情感标签，此方法着重于同步学习信息的一致性特质与差异性属性．

10) PS-Mixer[24]：借助MLP-Mixer框架下的极性向量与强度向量集成模型，旨在促进不同模态数据之间的信息交流与融合．

从表4和表5中呈现的结果可知，本文所提出的融合多子空间框架及通道注意力的算法模型性能优于其他的基线模型．在MOSI和MOSEI数据集上的二分类准确率(Acc-2)、七分类准确率(Acc-7)以及F1值均显著提高．此外，皮尔逊相关系数(Corr)增大，而平均绝对误差(MAE)减小，表明回归任务的性能也得到了改善．特别地，以MOSI数据集为例，相较于传统的ICCN模型，七分

类准确度增长了7%,平均绝对误差减小了0.129,皮尔逊相关系数增大了0.081.而相对于最近的MCGMF模型,七分类准确度增长了0.8%,平均绝对误差减小了0.029,皮尔逊相关系数增大了0.035.这些性能的提升可能源于本文提出的模型对融合前表示学习的重视,以及在模态不变表示和特定表示中应用的模态交互机制.

表4 MOSI基准实验对比结果

Tab. 4 Comparative results of MOSI benchmark experiments

Models	Acc-2	F1-score	Acc-7	M_{AE}	C_{orr}
ICCN	-/83.2	-/82.9	38.8	0.858	0.713
BBFN	-/82.8	-/82.3	42.5	0.787	0.748
CubeMLP	-/85.2	-/84.6	44.1	0.770	0.755
MCGMF	-/83.6	-/83.9	45.1	0.760	0.756
HFM-AM	-/79.6	-/79.3	-	-	-
MISA	78.0/80.9	78.1/80.7	43.1	0.804	0.739
MSSA	80.9/83.1	80.3/82.8	43.9	0.791	0.750
Hycon-B	-/85.0	-/84.9	42.7	0.721	0.795
ICDN	-/81.7	-/81.9	43.4	0.884	0.681
PS-Mixer	80.7/82.4	80.5/82.4	44.0	0.797	0.749
MSF-CA	84.1/85.6	84.4/85.7	45.9	0.731	0.791

表5 MOSEI基准实验对比结果

Tab. 5 Comparison results of MOSEI benchmark experiment

Models	Acc-2	F1-score	Acc-7	M_{AE}	C_{orr}
ICCN	-/84.0	-/84.1	51.9	0.562	0.717
BBFN	-/84.4	-/84.6	53.0	0.548	0.704
CubeMLP	-/84.7	-/84.3	53.2	0.538	0.760
MCGMF	-/85.6	-/85.1	52.8	0.541	0.768
HFM-AM	-/78.7	-/78.2	-	-	-
MISA	81.9/84.6	82.4/84.9	52.7	0.549	0.758
MSSA	83.5/85.9	83.7/85.8	53.6	0.540	0.761
Hycon-B	-/85.6	-/85.8	53.4	0.598	0.779
ICDN	-/81.7	-/81.9	53.0	0.584	0.712
PS-Mixer	83.3/85.8	83.2/85.9	53.7	0.531	0.762
MSF-CA	85.7/88.3	83.9/86.7	54.6	0.521	0.783

3.5.2 消融实验 为了评估模型中各模态对输出数值的贡献度,本文采用了四种模态组合进行测试,包括双模态(T+A、A+V、T+V)和三模态(T+V+A).在双模态组合(T+A、A+V、T+V)测试中,两种特征提取网络的输出会被输入模态不变和模态特定表示中,模态特定表示经由TCN网络融合后,再将每个向量输入通道注意力提取关键信息,然后利用模态融合机制对模态不变表示向量进行双模态融合.随后利用基于Transformer的多头注意力机制完成模态特征融合,最终通过全连接层输出预测结果.对于三模态组合(T+V+A),所有特征均按照上述步骤进行特征提取并用于情感分析,其中,T、A、V分别表示文本、语音和视觉特征.具体的消融实验数据如表6所示.

根据表6的数据,在双模态情感分析中,文本和语音组合表现最佳,视觉和文本组合次之,语音和视觉组合表现最差.这进一步说明了每个模态所包含的情感信息的差异.当引入三模态组合时,模型的性能显著提升,从而验证了本研究在情感分析中采用多模态方法的有效性.

表 6　模态消融实验对比结果

Tab. 6　Modality ablation experiment comparison results

Models	MOSI		MOSEI	
	$M_{AE}(\downarrow)$	$C_{orr}(\uparrow)$	$M_{AE}(\downarrow)$	$C_{orr}(\uparrow)$
V+A	1.448	0.047	0.789	0.098
T+A	0.789	0.767	0.544	0.767
V+T	0.837	0.745	0.576	0.751
T+V+A	0.731	0.791	0.521	0.783

为了进一步评估本文模型中各模块的作用及对模型整体性能的影响，设计了 8 组实验，具体的实验结果如表 7 所示．其中，MSF-CA-α 为消除模型的相似损失；MSF-CA-β 为消除模型中的差异损失；MSF-CA-γ 为消除模型中的重构损失；MSF-CA-CA 为消除模型中的通道注意力机制；MSF-CA-MI 为消除模型中的模态不变表示；MSF-CA-MS 为消除模型中的模态特定表示；MSF-CA-TCN 为消除模态特定表示中的 TCN 融合网络；MSF-CA(Bi-GRU)为使用 Bi-GRU 网络代替混合神经网络提取单模态特征；MSF-CA 为本文所提模型．

表 7　模型消融实验对比结果

Tab. 7　Model ablation experiment comparison results

Models	MOSI		MOSEI	
	$M_{AE}(\downarrow)$	$C_{orr}(\uparrow)$	$M_{AE}(\downarrow)$	$C_{orr}(\uparrow)$
$-\alpha$	0.746	0.771	0.541	0.767
$-\beta$	0.749	0.767	0.548	0.759
$-\gamma$	0.751	0.764	0.544	0.769
−CA	0.764	0.756	0.553	0.771
−MI	0.784	0.751	0.532	0.771
−MS	0.781	0.754	0.537	0.762
−TCN	0.749	0.770	0.541	0.764
Bi-GRU	0.742	0.775	0.533	0.769
MSF-CA	0.731	0.791	0.521	0.783

从表 7 可以看出，模态不变表示、模态特定表示、相似损失、差异损失、重构损失均能影响模型的性能．这说明了融合前表示学习中模态不变表示和模态特定表示的重要性，并表明以最小化损失函数为目标有助于模型学习各子空间表示的重要性．移除通道注意力机制后性能有所下降，表明该机制能有效识别并提取关键信息．此外，去除模态特定表示中的 TCN 融合网络后，模型性能亦见降低，这反映了模态特定表示中不同模态信息差异较大，需要对其进行重要性选择及交互融合．将混合神经网络替换为 Bi-GRU 网络后，模型在捕捉视频序列的情感信息方面能力降低，验证了混合神经网络相对于 Bi-GRU 网络的优越性．

3.5.3 参数性能分析实验　图 4 呈现了特征维度从 260 增加到 310 对模型七分类结果的影响．在维度达到 300 时，模型性能达到最优，并在两个数据集上展示了最佳结果．然而，当维度超过 300，性能开始下降，这可能是由于模型开始对训练数据中的噪声进行过度学习，而忽略了数据的潜在分布，从而导致了过拟合．

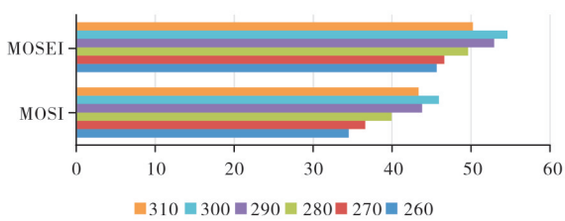

图 4　参数选择实验

Fig. 4　Parameter selection experiment

4　结束语

本文设计的 MSF-CA 多模态情感分析模型，主要完成了在融合前各模态表示学习模型的构建工作．表示学习部分包括模态不变表示和模态特定表示两大部分．在模态不变表示方面，设计了两个共享子空间，以为解码函数提供更多的输入，从而更有效地促进模型学习模态间的共性，并通过重建损失的作用，促使隐藏表征得到优化，进而更加精准地捕获不同模态下的特有细节信息．在模态特定表示中，设计了一种双模态交互网络，能够有效地完成异构模态的特征交互．此外，还利用通道注意力机制提取重要信息，并设计跨模态交互 Bi-GRU 和双模态交互注意力机制对处理后的模态不变表示向量进行更深层次的交互．通过三种不同的交互机制，充分利用了各模态的信息，提高了模型的泛化能力和鲁棒性．最后，使用基于 Transformer 的多头注意力机制完成上述特征融合．在公开的 MOSI 和 MOSEI 数据集上进行大量实验，实验结果显示，本文提出的 MSF-CA 模型能有效完成多模态融合，在回归任务和分类任务中表现出一定的提升．然而，由于实验中直接采用串联方法构造联合向量，导致模型未能充分利用表示学习来优化特征．因此，未来的研究将进一步改进此方法，总结相关经验，并探索其他的模态融合技术．

参考文献:

[1] PORIA S, CAMBRIA E, HAZARIKA D, et al. Context-dependent sentiment analysis in user-generated videos[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Thesiss). Vancouver, Canada:ACL, 2017:873-883.

[2] CHAUHAN D S, AKHTAR M S, EKBAL A, et al. Context-aware interactive attention for multi-modal sentiment and emotion analysis[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: ACL, 2019:5647-5657.

[3] 缪裕青,杨爽,刘同来,等. 基于跨模态门控机制和改进融合方法的多模态情感分析[J]. 计算机应用研究,2023,40(7):2025-2030;2038.
MIAO Y Q, YANG S, LIU T L, et al. Multimodal sentiment analysis based on cross-modality gating mechanism and improved fusion method [J]. Computer Applications in Research, 2023,40(7):2025-2030;2038.

[4] 李文雪, 甘臣权. 基于注意力机制的分层次交互融合多模态情感分析[J]. 重庆邮电大学学报(自然科学版),2023,35(1):176-184.
LI W X, GAN C Q. Hierarchical interactive fusion multimodal sentiment analysis based on attention mechanism [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2023,35(1):176-184.

[5] YANG J, YU Y, NIU D, et al. ConFEDE: contrastive feature decomposition for multimodal sentiment analysis [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Thesis). Toronto, Canada:ACL, 2023:7617-7630.

[6] WANG Y, SHEN Y, LIU Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors [C]//The AAAI Conference on Artificial Intelligence. Hawaii, USA:AAAI, 2019,33(1):7216-7223.

[7] GHOSAL D, AKHTAR M S, CHAUHAN D, et al. Contextual inter-modal attention for multi-modal sentiment analysis [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium:EMNLP, 2018:3454-3466.

[8] XU J, HUANG F, ZHANG X, et al. Visual-textual sentiment classification with bidirectional multi-level attention networks [J]. Knowledge-Based Systems, 2019,178:61-73.

[9] XI C, LU G, YAN J. Multimodal sentiment analysis based on multi-head attention mechanism[C]//Proceedings of the 4th International Conference on Machine Learning and Soft Computing. Hai Phong, Vietnam:ICMLSC, 2020:34-39.

[10] KUMAR A, VEPA J. Gated mechanism for attention based multi modal sentiment analysis [C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York, USA: IEEE, 2020:4477-4481.

[11] DELBROUCK J B, TITS N, BROUSMICHE M, et al. A transformer-based joint-encoding for emotion recogni-tion and sentiment analysis [DB/OL]. (2020-06-29)[2025-10-21]. https://arxiv.org/abs/2006.15955v1.

[12] ZADEH A, MAO C, SHI K, et al. Factorized multimodal transformer for multimodal sequential learning [DB/OL]. (2019-11-22) [2025-10-21]. https://arxiv.org/abs/1911.09826.

[13] WIBOWO H, FIRDAUSI F, SUHARSO W, et al. Facial expression recognition of 3D image using facial action coding system (FACS)[J/OL]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2019, 17(2): 628-636.

[14] VERDE L, MARULLI F, DE FAZIO R, et al. HEAR set: a lightweight acoustic parameters set to assess mental health from voice analysis [J/OL]. Computers in Biology and Medicine, 2024, 182 [2025-10-21]. https://doi.org/10.1016/j.compbiomed.2024.109021.

[15] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module [C]//Proceedings of the European conference on computer vision (ECCV). Munich, Germany: ECCV, 2018:3-19.

[16] MENG Z, CAO W, SUN D, et al. Research on fault diagnosis method of MS-CNN rolling bearing based on local central moment discrepancy[J/OL]. Advanced Engineering Informatics, 2022, 54 [2025-10-21]. https://doi.org/10.1016/j.aei.2022.101797.

[17] SUN Z, SARMA P, SETHARES W, et al. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020, 34(5): 8992-8999.

[18] HAN W, CHEN H, GELBUKH A, et al. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis [C]//Proceedings of the 2021 International Conference on Multimodal Interaction. New York, USA:ACM, 2021:6-15.

[19] SUN H, WANG H, LIU J, et al. CubeMLP: an MLP-based model for multimoda-l sentiment analysis and depression estimation[C]//Proceedings of the 30th ACM International Conference on Multimedia. Los Angeles, USA:ACM, 2022: 3722-3729.

[20] 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, USA:ACM, 2020: 1122-1131.

[21] 胡新荣,陈志恒,刘军平,等. 基于多模态表示学习的情感分析框架[J]. 计算机科学,2022,49(S2):631-636.
HU X R, CHEN Z H, LIU J P, et al. Sentiment analysis framework based on multimodal representation learning[J].

Computer Science, 2022,49(S2): 631-636.(Ch).
[22] MAI S, ZENG Y, ZHENG S, et al. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis [J]. IEEE Transactions on Affective Computing, 2022, 14(3): 2276-2289.
[23] ZHANG Q, SHI L, LIU P, et al. RETRACTED ARTICLE: ICDN: integrating consistency and difference networks by transformer for multimodal sentiment analysis [J]. Applied Intelligence, 2023, 53(12):16332-16345.
[24] LIN H, ZHANG P, LING J, et al. PS-mixer: a polar-vector and strength-vector mixer model for multimodal sentiment analysis [J/OL]. Information Processing & Management, 2023, 60 [2025-10-21]. https://doi. org/10. 1016/j. ipm. 2022. 103229.

Multimodal sentiment analysis fusing multi-subspace and channel attention

MI Xiaofeng, WANG Xuyang, SHI Haojun
(School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: Current multimodal sentiment analysis primarily relies on complex techniques for fusing multimodal features. However, due to the significant distribution differences among various modal features, direct fusion yields poor results. To address this issue, this paper proposes an interactive learning network model that integrates a multi-subspace framework and channel attention. Firstly, a hybrid neural network is utilized to extract features from each modality, and a stacked bidirectional long short-term memory network is employed to represent the utterance sequence at the linguistic level. Fixed-size utterance vectors are mapped into two different representations: modal-invariant and modal-specific, with the latter undergoing bimodal interaction using a temporal convolutional network. Subsequently, channel attention is leveraged to extract more meaningful information, and a cross-modal interactive bidirectional gated recurrent neural network and a bimodal interactive attention mechanism are proposed for deeper interaction among the extracted modal-invariant representation vectors. Loss optimization is then performed using a loss function. Finally, a multi-head attention mechanism based on Transformer is executed to obtain a joint vector, and a fully connected layer is utilized to predict the final result. Experiments conducted on the CMU-MOSI and CMU-MOSEI datasets demonstrate that this method can effectively eliminate multimodal differences and achieve multimodal fusion.

Key words: multimodal sentiment analysis; hybrid neural network; multimodal fusion; Transformer; attention mechanism