计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250200022

基于主体注意力与多空间域信息协同的多模态情感分析

冯 广[1] 林忆宝[2] 钟 婷[1] 郑润庭[2] 黄俊辉[2] 刘天翔[2] 杨燕茹[2]
1 广东工业大学自动化学院 广州 510006
2 广东工业大学计算机学院 广州 510006

摘 要 多模态情感分析在智慧教育中具有重要应用价值,例如通过分析学生的语言、表情和语调等多模态信息,来评估课堂参与度和情感状态,从而辅助教师实时调整教学策略。然而,现有多模态情感分析领域中,跨模态注意力机制对于异构模态间的关联捕捉不够充分,并且对共享空间与私有空间的信息协同并未进行深入探索,存在跨模态融合学习受限且多空间域信息协同不充分的问题。针对这些问题,文中提出了基于主体注意力融合多空间域异构模态的多模态情感分析模型,该模型通过主体注意力机制,对两个空间域中的异构模态分别进行充分融合,以解决跨模态融合学习受限的问题。然后,利用门控机制补充共享空间域异构模态融合向量的模态独立性,以实现私有空间与共享空间信息的协同,有效解决多空间域信息协同不充分的问题。实验结果表明,该模型在公共数据集 MOSI 和 MOSEI 上的得分整体都有提高,说明该方法可以充分捕捉多模态异构信息间的潜在关系并有效协同不同空间域的异构融合信息。

关键词:多模态情感分析;主体注意力;多空间域;门控机制;智慧教育
中图分类号 TP391.1

Multimodal Sentiment Analysis Based on Dominant Attention and Multi-space Domain Information Collaboration

FENG Guang[1],LIN Yibao[2],ZHONG Ting[1],ZHENG Runting[2],HUANG Junhui[2],LIU Tianxiang[2] and YANG Yanru[2]
1 School of Automation,Guangdong University of Technology,Guangzhou 510006,China
2 School of Computer Science,Guangdong University of Technology,Guangzhou 510006,China

Abstract Multimodal sentiment analysis has significant applications in smart education,such as assessing students' engagement and emotional states through speech,facial expressions,and tone to help teachers adjust teaching strategies in real time. However,existing cross-modal attention mechanisms struggle to capture associations between heterogeneous modalities effectively,and the collaboration between shared and private spaces remains underexplored,limiting multimodal fusion learning. To address these issues,this paper proposes a multimodal sentiment analysis model that integrates heterogeneous modalities across multiple space domains using dominant attention. This mechanism enables effective fusion of heterogeneous modalities in both domains,enhancing cross-modal learning. Additionally,a gating mechanism preserves the modality independence of shared-space fusion vectors,ensuring complementary interactions between private and shared spaces. Experimental results on the MOSI and MOSEI datasets demonstrate that the proposed model achieves overall performance improvements,validating its ability to capture and integrate heterogeneous multimodal information effectively.

Keywords Multimodal sentiment analysis,Dominant attention,Multi-space domain,Gating mechanism,Smart education

在信息爆炸的当代社会,互联网和社交媒体已深度渗透到人们的日常生活中。从文字博客到短视频分享,从语音聊天到虚拟现实体验,用户通过各种形式的数据表达个人观点和情感[1]。这种信息呈现的多样性不仅丰富了数据的来源,也给传统情感分析方法带来了新的挑战。例如,在市场研究中,情感分析能够揭示消费者的需求与偏好;在舆情监控中,政府可通过情感分析追踪公众对热点事件的情绪态度[2]。因此,如何利用社交媒体上的海量数据分析人们的情感,已成为当前研究中的一个热门方向。

早期的情感分析方法主要依赖于文本数据,通常只分析单一模态的信息,这种方法虽然在一定程度上能够识别情感,但也存在显著的局限性。单一模态的情感分析仅依赖于文字本身,而忽略了情感表达中其他关键信息的影响,例如语音语调、面部表情和身体语言等。这样的方法无法准确捕捉到情感的复杂性和多样性,特别是在一些隐晦或具有讽刺性的情境中,可能会导致情感判定出现误差。例如,如果某人说:"真是谢谢你",单纯从文字上看,这句话似乎表达了感激之情,忽略语音的语调和面部表情,就无法察觉到这句话背后可能蕴含的讽刺意味。实际上,这个人可能语气冷淡或者不耐烦,同时面部表情可能带着不屑或轻蔑。在这种情况下,单纯的文

基金项目:国家自然科学基金重点项目(62237001);广东省哲学社会科学青年项目(GD23YJY08)
This work was supported by the Key Program of the National Natural Science Foundation of China(62237001) and Youth Project of Guangdong Provincial Philosophy and Social Sciences(GD23YJY08).
通信作者:冯广(von@gdut.edu.cn)

本情感分析会错误地将其判定为正面情感，而实际上这是一种负面情绪的表达。这样的情感分析结果显然是被误导的。因此，单一模态的情感分析无法充分把握人类情感的复杂性，尤其在面对讽刺、反讽、幽默等复杂情绪表达时，传统的文本分析方法显得力不从心。这也促使研究者逐渐转向多模态情感分析，结合文本、语音、面部表情等多个数据源，以更全面、更准确地理解和识别人类情感。

多模态情感分析的核心在于如何从不同模态中提取有效的特征，并将这些特征进行融合[3]。在目前的多模态分析中，各种模态（例如文本、音频和视觉）采用的是不同的特征提取技术，且这些特征的维度和表示方式也往往各不相同。这使得如何将来自不同模态的信息进行有效融合，成为了一个难题。简单的特征融合方法往往会导致信息丢失或不一致，从而影响情感分析的准确性。在早期的工作中，Tsai 等[4]采用了一种基于跨模态注意力机制的 Transformer 模型来解决 3 种模态中的融合问题。Hazarika 等[5]提出了 MISA 的多模态框架，将多模态数据投影到多个子空间域，以提供多模态数据全面且清晰的视图。这些方法在一定程度上解决了多模态异构数据的融合问题，但仍然存在着两个挑战：1）跨模态注意力机制对多模态异构数据的融合并不充分，多模态协同学习受限的问题仍然存在；2）在多模态子空间信息的深度交互和模态间复杂融合方面仍然不够充分，导致情感分析的强度不足。

为了解决上述问题，本文提出了基于主体注意力与多空间域信息协同的多模态情感分析模型（DCMDA-CDG）。与传统的跨模态注意力机制以及多空间域信息的融合不同，本研究设计了一种动态选择主体的跨模态融合注意力策略，旨在实现多模态的深度融合，并引入多空间域信息互补机制，动态地调节共享空间域与私有空间域的信息流，从而丰富子空间域的信息内容，促进多空间域信息协同。具体来说，本文的创新点如下。

1）动态主体注意力机制。传统的 Cross-Attention 往往缺乏主模态的引导性，各模态特征对等交互，容易导致辅助模态的噪声或冗余信息干扰主体模态特征的提取；本研究提出基于模态信息量的动态主体注意力机制。首先，采用注意力机制对多模态特征进行信息量计算，动态选择当前任务下的主体模态作为信息聚合中心；其次，在注意力计算中实施双重 Softmax 注意力策略，实现主体模态与辅助模态信息的聚合与广播。这种机制可以动态选择主体模态以引导跨模态融合，并有效挖掘辅助模态对情感表达至关重要的细节特征，进一步强化多模态信息的协同表达。

2）跨空间域信息互补。通过学习模态的共享特征和私有特征表示，为不同空间域提供了相互独立的视图。然而，这种独立性导致空间域之间缺乏有效的交互与信息补充。针对这一问题，本文设计了一种门控机制，通过动态调节不同空间域的信息流，实现了共享空间与私有空间的信息互补，从而丰富了子空间域中的特征表达。

3）多模态情感分析性能提升。本文设计的模型在两个公开数据集（CMU-MOSI 和 CMU-MOSEI）上进行了训练测试。实验结果表明，DCMDA-CDG 模型能够有效地提高多模态情感分析的准确性，尤其在模态融合优化和空间域信息交互方面表现出了显著优势。

1 相关研究

多模态情感分析（Multimodal Sentiment Analysis，MSA）近年来成为情感分析研究的一个重要方向[6]，旨在通过融合来自多个模态的信息（如文本、音频和视频）来更准确地理解和预测情感。传统的情感分析方法主要依赖于单一模态（如文本数据），但单一模态往往无法全面捕捉情感表达的多样性，特别是在面对复杂的情感变化时。

随着多模态数据的广泛可用，研究者开始探索如何将不同模态的信息进行有效融合，以提升情感分析的表现。最初的多模态情感分析方法主要集中在早期融合和后期融合。Poria 等[7]提出了一种并行化的决策级数据融合方法，利用多核学习（MKL）算法训练提取的三模态特征向量进行情感分类。Zadeh 等[8]提出的张量融合网络（Tensor Fusion Network）通过张量融合方法对模态间的动态进行建模，以有效地捕捉不同模态之间的交互信息。随后，Liu 等[9]提出的低秩多模态融合方法（Low-Rank Multimodal Fusion Network）作为 TFN 的改进模型，通过低秩分解减少了参数数量，但当特征维度过长时，参数数量仍然容易爆炸。

跨模态注意力机制的引入进一步推动了多模态情感分析的发展，尤其是在处理模态间异质性方面表现出色。注意力机制能够通过动态调整模态间的权重，增强关键模态信息的影响力，从而提升模型的性能。Wang 等[10]提出了 RAVEN（Relational Attention-based Visual-Textual Emotion Network）模型，通过关系注意力机制增强文本与视觉信息之间的交互，自适应地捕捉两种模态中情感表达的关联性。Putra 等[11]则提出了 MAG-BERT（Mutilmodal Adaption Gate），将 RAVEN 集成到了 BERT 中，有效地将非言语信息融入到模型中获得了更优的融合表示。Tasi 等[4]设计了一种多模态 Transformer（Mutilmodal Transformer，MULT），通过引入多模态 Transformer 架构捕捉多模态数据中的长距离依赖关系。Yu 等[12]设计了一种自监督的多任务多模态（Self-MM）情感分析网络，包含 1 个自监督学习的标签生成模块和 3 个独立的单模态任务，进行多目标任务联合训练。尽管这些任务很好地融合了复杂的多模态，但忽略了模态情感信息的私有表征和共享表征。Hazarika 等[5]提出了 MISA（Multimodal Interaction with Self-Attention）模型，将多模态模态投影到多空间域，并采用模块化融合策略保持模态特征的独立性与互补性。然而，这些跨模态注意力方法对多模态异构数据的融合并不充分，多模态融合共现特征的学习受限。

通过相关研究和分析发现，尽管跨模态注意力机制为异构模态融合提供了新的可能性，但现有方法在模态间的融合仍不充分，且缺乏主导模态，未能充分捕捉模态间复杂的关系。并且现有方法大多忽视了私有空间与共享空间之间的信息流动，未能充分利用并协同不同空间域的信息。为了解决上述问题，本研究借鉴了双轨策略，将每个模态投影到两个不同的子空间（模态不变的子空间和模态特定的子空间），同时关注在所有模态中普遍存在的情感特征和每个模态独有的特征，利用设计的主体注意力动态地选择主体模态作为主导，进行异构模态信息的融合，并通过门控机制协同共享空间域与私有空间域的信息流。

2　DCMSA-CDG 模型

本文模型如图 1 所示，主要由以下 4 部分组成：模态特征编码层，特征空间划分层，模态特征融合层及预测层。该模型首先将各模态数据输入到模态特征编码层中提取出各模态的特征，并通过 DMS 模块选择出主体模态，然后将各模态特征输入到特征空间划分层，将各模态划分到共享空间域和私有空间域以探索模态间的区别与联系，然后在模态特征融合层中对两个空间域的模态特征进行融合，通过主体注意力机制，以主体模态为引导得到融合表征，并通过 CDG 生成的门控将共享空间的融合表征与私有空间的信息进行互补协同，最后在预测层对融合向量进行自注意力和多层 MLP 得到预测值。

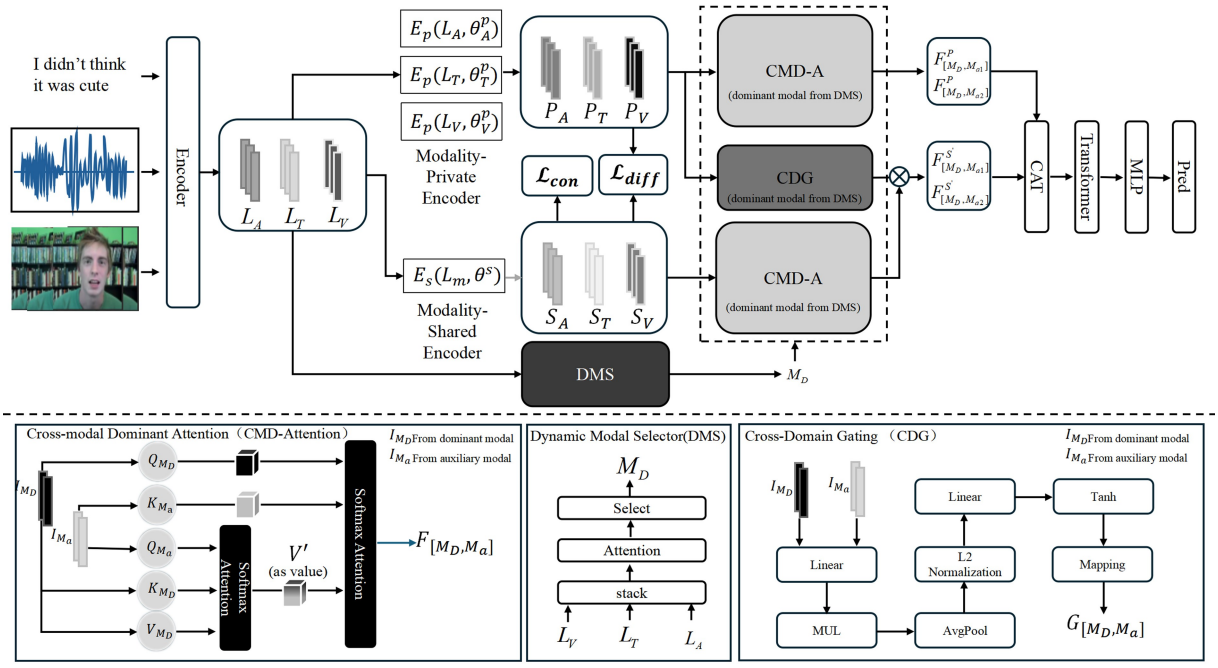

图 1　DCMSA-CDG 模型的实现细节

Fig. 1　Implementation details of DCMSA-CDG model

2.1　模态特征编码层

从每种异构子信息（文本、音频、视频）中提取基本特征以构建对各个模态的初步理解。特征提取方法如下。

1）文本特征提取：对于文本模态，选用了 RoBERTa[13] 模型来提取初步特征。通过利用 RoBERTa 的 tokenizer 对文本进行编码，将其转换为模型可处理的标记序列，并保留了词汇的位置和段落信息。随后，输入经过预处理的文本至 RoBERTa 模型，以获取文本的结构化特征，包括段落和句子级别的表征。这些特征捕捉了文本中的复杂依赖关系，并为后续的情感分析任务提供了丰富的语言和结构信息。

2）音频特征提取：采用了 OpenSMile[14] 和 COVAREP 这两个工具来提取和分析语音信号的特征。首先，利用 OpenSMile[15] 提取了一系列关键的语音特征，包括梅尔频率倒谱系数（MFCC）、音高、能量等。OpenSMile 是一个强大的音频信号处理工具，它能够高效地处理音频数据，并提取出多种声学特征。在初步特征提取的基础上，进一步使用 COVAREP 进行更深入的语音分析。COVAREP 提供了多种语音分析算法。通过 COVAREP，能够进行声门源分析，获取更细致的语音特征，如基频和声门关闭与开放的时序信息，这些特征对于情感状态具有重要意义。

3）视频特征提取：使用 Facet[16] 和 OpenFace[17] 来提取面部表情特征。Facet 是一种面部分析工具，能够从视频中提取多种面部特征，面部标志、面部动作单元（AU）、头部姿势、视线轨迹和 HOG 特征等，这些特征有助于捕捉细微的面部表情变化和情感状态。OpenFace 则提供包括面部标志、头部姿势、眼睛注视和面部动作单元等信息，通过分析每一帧中的人脸特征，提取出动态的面部动作和静态的面部标志，为后续的情感分析提供了全面的面部表情数据。

对于视频和音频的初级特征，使用 Transformer Encoders 来进行特征提取，捕获远程依赖关系。对于文本信息，将输入文本输入 RoBERTa 以增强文本表示。每种异构子信息的输出表示为 L_i，其中 $i\in\{A,V,T\}$。特征提取的计算式如下：

$$L_T = RoBERTa(X_T,\Phi_T) \tag{1}$$

$$L_A = TransformerEncoder(X_A,\Phi_A) \tag{2}$$

$$L_V = TransformerEncoder(X_V,\Phi_V) \tag{3}$$

2.2　特征空间划分层

特征空间划分层将特征投影到两个独立的空间，将每个模态向量分别投影到两个相对独立的表示空间，分别为表示不同模态之间的公共特性的模态共享空间，以及表示各模态特定特性的模态私有空间，通过两个表示空间来提供有效融合所需的全局视图。首先，通过共享编码器对模态的特征进行统一建模，提取出异构模态间的共享表示，从而有效减小异构模态间的异质性差距。其次，为了更精细地捕捉各异构模态的独特特性，引入了 3 种独立的私有编码器，分别针对不同异构模态进行私有表示的学习。通过将各异构模态特征嵌入到其私有子空间，保留模态的独特性，同时避免共享表示对模态私有表征的干扰[18]，从而为后续的模态融合提供更丰富的语义表征。

共享表示 S_M 和私有表示 P_m 的计算式如下：

$$S_M = E_c(L_M;\theta^S) \tag{4}$$

$$P_M = E_p(L_M;\theta_M^p) \tag{5}$$

其中，共享编码器共享参数 θ^S，而私有编码器为每种模态分配各自的参数 θ_M^P，M 是 3 种模态{A,V,T}。

2.3 模态特征融合层

模态特征融合层包含 4 个关键模块：主体注意力机制模块、模态共享表征交互模块、模态私有表征交互模块和跨空间域门控交互模块。主体注意力机制模块通过显式建模文本模态与其他模态之间的相互影响，使得文本模态能够从辅助模态中汲取信息，同时保持其独特的语义特征，从而提升整个多模态系统的协同效应和性能。模态共享表征交互模块中，各异构模态的共享表征已经经过统一建模，具备全局一致性。通过主体注意力机制，进一步捕获模态间的显式交互关系，强化模态间的信息协作，从而增强共享表征之间的互补性和一致性，提升跨模态信息的整合效果。模态私有表征交互模块中，各模态的表征反映了其独特性，通过注意力机制探索模态私有信息之间的互补关系，增强每个模态的表现力，同时保留模态特有的语义信息。这一过程有效提高了私有表征在多模态融合中的贡献。跨空间域门控交互模块则为模态的私有空间和共享空间提供了一种交互机制，通过在不同空间域之间实现信息流动，补充共享表征中的模态独立信息，丰富共享表征的内容，进一步提升多模态融合的表现力。通过门控机制的调节，能够灵活控制不同空间域之间的信息交互和融合。

2.3.1 主体注意力机制

文本作为信息载体，尤其在情感分析和语义理解任务中，通常携带着丰富的情感信息。因此，在多模态情感分析中，文本往往被作为主导模态，以充分利用其语义和情感特征。然而，在某些特定场景下，音频和视频模态可能承载着更具决定性的情感信息。例如，在情感表达更依赖语调、语速或面部表情的场景中，音频或视频特征可能比文本更能反映情感状态。因此，设计了自适应主体注意力机制，通过自适应选择最具信息量的模态作为主体模态，使其与辅助模态进行跨模态交互。该机制不仅能够确保文本在大多数情境下作为主导模态，还能够在音视频信息更具主导性时动态调整，从而提升多模态情感分析的准确性和鲁棒性。

在将模态特征划分到私有空间和共享空间之前，使用 DMS 模块来评估每个模态的信息质量，将 3 种模态特征输入评估模块，通过注意力机制计算每个模态质量分数，让模型学习不同模态的信息量大小，然后根据注意力分数找到信息质量最好的模态，视为后续的主体模态。其具体过程如下：

$$AW_M=\mathrm{Softmax}\left(\frac{Q_M\cdot K_M^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{6}$$

$$s_M=\sigma(f(AW_M\cdot V_M)) \tag{7}$$

$$M_D=\mathrm{argmax}(s_T,s_A,s_V) \tag{8}$$

其中，$f(\cdot)$ 是 MLP 层，$argmax(\cdot)$ 表示取最大值函数，$\sigma(\cdot)$ 是 Sigmoid 归一化，s_M 是 3 种模态的信息量分数，M 属于 3 种模态，M_D 是 3 种模态中选出的主体模态。

在主体注意力机制中，将 DMS 模块中得到的模态视为主体模态输入，与其他两个模态输入进行信息聚合，辅助模态特征作为查询向量参与到注意力计算中，与主体特征（作为键向量和值向量）进行交互，计算得到跨模态特征的聚合表示。通过这种方式，模型能够从所有的主体模态特征中聚合出与辅助模态特征相关的全局信息。在主体注意力的框架下，主体模态特征可以通过辅助模态的交互，逐层感知到辅助模态特征中细粒度的局部信息[19]。然后将聚合后的跨模态特征作为值向量，辅助模态特征作为键向量，主体模态特征作为查询向量，进行第二轮注意力计算，将视频中的信息广播到所有主体模态特征中。通过广播机制，将辅助模态特征的信息传递给每个主体模态特征 token，从而确保每个文本特征都能够获得来自辅助模态特征的上下文信息。这有助于模型整合多模态特征，建立更加完整的情感理解。其具体过程如下：

$$AW_{M_a\to M_D}=\mathrm{Softmax}\left(\frac{Q_{M_a}\cdot K_{M_D}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{9}$$

$$V'=AW_{M_a\to M_D}\cdot V_{M_D} \tag{10}$$

$$AW_{M_D\to M_a}=\mathrm{Softmax}\left(\frac{Q_{M_D}\cdot K_{M_a}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{11}$$

$$F_{[M_D,M_a]}=AW_{M_D\to M_a}\cdot V' \tag{12}$$

其中，d_k 是输入向量的维度，防止点积运算结果过大；AW 是注意力权重；$F_{[M_D,M_a]}$ 是以 I_{M_D} 为主体，I_{M_a} 为辅助模态的融合输出；M_D 是 3 种模态中选出的主体模态；M_a 是两种辅助模态。

2.3.2 跨空间域门控

共享空间的融合倾向于将模态之间的共性特征提取出来，但无法充分保留每个模态的独立信息，尤其是每个模态在表达情感时的独特特征。情感分析不仅依赖于模态间的相似性，还需要各模态提供的独立信息，尤其是在情感推理上，每个模态的独特信息对情感的细粒度预测至关重要。

为了解决共享空间融合后可能丧失模态间独立信息的问题，引入了以因子双线性池化（Factorized Bilinear Pooling，FBP）[20] 为基础的跨空间域门控（Cross-domain Gating，CDG），如图 1 的右下部分所示。在私有空间中，文本特征和辅助模态特征被 CDG 单独处理，从而避免了它们在共享空间中过度融合，避免了信息混淆或模态间依赖关系的丧失。对处理后的融合表示进行门控映射，最终生成一个私有空间融合向量门控值，用于对共享空间融合表征进行补充。

将两个模态输入 I_{M_D} 和 I_{M_a}，分别通过将原始特征映射为查询向量和键向量。然后通过点积计算得到一个更高维的联合表征矩阵 $\boldsymbol{G}=Linear(F_{\mathrm{norm}})\in R^{N\times 1}$，表示了两个模态之间的所有交互作用。接下来，对 $\boldsymbol{G}=Linear(F_{\mathrm{norm}})\in R^{N\times 1}$ 应用平均池化操作，通过在每个窗口内求和来减少特征维度，同时保留重要的特征信息。为了进一步处理特征，进行归一化操作生成最终的特征表示 $\boldsymbol{G}=Linear(F_{\mathrm{norm}})\in R^{N\times 1}$。这里，归一化使得特征向量的幅度不会过大，有助于模型的稳定性和收敛。

$$I'_{M_D}=Linear(I_{M_D})\in R^{N\times(f_{\mathrm{hid}}\cdot k)} \tag{13}$$

$$I'_{M_a}=Linear(I_{M_a})\in R^{N\times(f_{\mathrm{hid}}\cdot k)} \tag{14}$$

$$F_{\mathrm{mul}}=I'_{M_D}\cdot I'_{M_a} \tag{15}$$

$$F'_{\mathrm{mul}}=AvgPool(F_{\mathrm{mul}},k) \tag{16}$$

$$F_{\mathrm{norm}}=\frac{F'_{\mathrm{mul}}}{\|F'_{\mathrm{mul}}\|_2} \tag{17}$$

其中，N 是序列长度，f_{hid} 是隐藏层的维度，k 是池化操作的核大小，$AvgPool(\cdot)$ 是平均池化操作，M_D 是 3 种模态中选出的主体模态，M_a 是两种辅助模态。

将 FBP 层生成的最终的特征表示为 F_{norm}，采用线性层，将 FBP 层的输出 F_{norm} 映射到一个新的维度。这个线性层的输出是一个 1 维的向量 $\boldsymbol{G}$，用于表示门控信号的初值。然后

使用 tanh 激活函数，将线性层的输出映射到(−1,1)的范围内。最后将门控信号的值转换到[0,1]区间，这样正值和负值都会被映射到 0 和 1 之间，正值保持不变，负值被转换为其正值。门控信号接近 1 时，表示该模态特征在融合中占主导地位；接近 0 时，表示该模态特征的影响被抑制。

$$G=Linear(F_{\text{norm}})\in R^{N\times 1} \tag{18}$$

$$G'=\tanh(G) \tag{19}$$

$$G_{[M_D,M_a]}=\frac{1}{2}(G+1) \tag{20}$$

2.3.3　模态共享表征融合模块

模态共享空间的意义在于捕获异构模态之间的公共特性，增强模态间的协同与交互，从而提升对多模态信息的整体理解。然而，由于模态共享空间缺乏各模态的独特特征，可能在后续融合层中对整体融合引入信息冗余，从而导致性能下降。因此，利用模态私有空间的信息进行补充，能够弥补共享空间的不足，实现更加全面的特征表达。在模态共享空间中，使用主体注意力机制分别对文本、视频、音频 3 种异构模态的输入向量进行交互，生成相应的跨模态融合向量。并且将私有空间的模态向量输入 CDG 模块生成对应的跨空间域门控值。最后采用了加权平均的方式，将 CDG 模块生成的门控值与对应的跨模态融合向量相乘，然后与原始融合向量进行加权平均。这里的权重是可学习的参数，模型将在训练过程中自动优化这些权重，以实现共享空间和私有空间融合向量之间的最佳互补，生成最终的跨模态融合向量。

$$F^{S}_{[M_D,M_{a1}]}=CMD\text{-}Attention(S_{M_D},S_{M_{a1}}) \tag{21}$$

$$F^{S}_{[M_D,M_{a2}]}=CMD\text{-}Attention(S_{M_D},S_{M_{a2}}) \tag{22}$$

$$F^{S'}_{[M_D,M_{a1}]}=\lambda\cdot G_{[M_D,M_{a1}]}F^{S}_{[M_D,M_{a1}]}+(1-\lambda)\cdot F^{S}_{[M_D,M_{a1}]} \tag{23}$$

$$F^{S'}_{[M_D,M_{a2}]}=\lambda\cdot G_{[M_D,M_{a2}]}F^{S}_{[M_D,M_{a2}]}+(1-\lambda)\cdot F^{S}_{[M_D,M_{a2}]} \tag{24}$$

其中，*CMD-Attention*(·)为主体注意力机制，$G_{[M_D,M_{a1}]}$和$G_{[M_D,M_{a2}]}$为 CDG 模块生成的门控值，M_D 是 3 种模态中选出的主体模态，M_{a1}和 M_{a2}是另外两种辅助模态，λ 是可学习的权重参数。

2.3.4　模态私有表征融合模块

由于不同模态的数据形式和特性各不相同，这些特性可能无法通过模态共享空间体现。模态私有空间能够单独提取这些模态相关的细粒度特征，避免模态间信息的干扰和混淆，并保留各模态特有的独立信息，从而捕获每种模态的独特特性。在模态私有空间中，分别对文本、视频、音频 3 种异构模态的输入向量使用主体注意力机制进行充分的交互，生成相应的跨模态融合向量。

$$F^{P}_{[M_D,M_{a1}]}=CMD\text{-}Attention(P_{M_D},P_{M_{a1}}) \tag{25}$$

$$F^{P}_{[M_D,M_{a2}]}=CMD\text{-}Attention(P_{M_D},P_{M_{a2}}) \tag{26}$$

其中，*CMD-Attention*(·)为主体注意力机制，M_D 是 3 种模态中选出的主体模态，M_{a1}和 M_{a2}是另外两种辅助模态。

2.4　预测层

预测层中，首先将私有空间和共享空间中最终的跨模态融合数据进行拼接，随后用多头自注意力机制处理拼接后的特征。多头自注意力能够从多个角度捕捉跨模态和跨子空间的关系，确保模型能够在不同模态之间建立有效的关联，并保留每个模态的关键信息[21]。它通过并行工作的多个注意力头，有效地捕捉复杂的模态交互，同时增强对跨模态信息的全局理解，为情感预测提供更为精准的特征表示。然后。经过自注意力处理后的特征会经过求和层，对每个时间步进行平均，得到批量样本的均值。最后，批量样本的均值会传递到多层感知器(MLP)层，进一步整合和映射为最终的情感预测值[22]。MLP 能够处理特征间的非线性关系，将高维特征转化为更加适合情感分类的表示，并为情感预测提供更精确的结果[23]。通过这一预测层，本文模型能够有效整合多模态信息，提升情感分析的准确性和鲁棒性。具体过程如下：

$$h_1=Concat(F^{P}_{[M_D,M_{a1}]};F^{P}_{[M_D,M_{a2}]};F^{S'}_{[M_D,M_{a1}]};F^{S'}_{[M_D,M_{a2}]}) \tag{27}$$

$$h_2=TransformerEncode(h_1) \tag{28}$$

$$h_3=\frac{1}{N}\sum_{n=1}^{N}h_2[n,:,:] \tag{29}$$

$$O=FC(RELU(FC(h_3)))\in R^{N\times 1} \tag{30}$$

其中，*concat*(·)为拼接处理；N 为样本总数；*RELU*(·)为激活函数；*FC*(·)为全连接层；O 是最终输出，维度为 1。

2.5　损失函数

为了实现不同模态表示在不变子空间中的对齐，以约束各模态在共享子空间中的表征保持一致，从而学习到不同模态间的共性特征表示[24]。本研究在解耦模态表征学习过程中引入了一致性损失函数，具体来说，采用中心矩差异[25](Central Moment Discrepancy，CMD)来量化不同模态之间的差异，确保模态特征在共享空间中的一致性，从而有效对齐模态间的表示。该一致性损失的计算式如式(31)所示：

$$L_{\text{con}}=\frac{1}{3}\sum_{(m_1,m_2)\in\{(T,A),(T,V),(A,V)\}}CMD(\boldsymbol{I}_{m_1},\boldsymbol{I}_{m_2}) \tag{31}$$

其中，$\boldsymbol{I}_{m_1}$，$\boldsymbol{I}_{m_2}$分别表示共享子空间的不同模态向量；*CMD*(·)是中心矩差异。

与此同时，引入了差异性损失函数来避免信息冗余，并确保模态共享表征和私有表征能够捕获输入数据的不同方面。这一设计使得不同模态之间的表征能够保持非冗余性，从而在保留个性化信息的同时实现更加有效的多模态融合。该差异性损失的计算式如式(32)所示：

$$L_{\text{diff}}=\sum_{m\in\{T,V,A\}}\|S_m^TP_m\|_{\text{F}}^2+\sum_{(m_1,m_2)\in\{(T,A),(T,V),(A,V)\}}\|P_{m_1}^TP_{m_2}\|_{\text{F}}^2 \tag{32}$$

其中，$\{T,V,A\}$是 3 种模态，分别表示文本、视频和音频；S_m是共享子空间的模态向量；P_m 是私有子空间的模态向量。

对于目标损失函数，采用了均方误差损失函数 MSE，通过最小化情感强度损失，帮助情感强度接近真实值[26]，提高模型预测的准确性。该损失的计算式如式(33)所示：

$$L_{\text{task}}=\frac{1}{N}\sum_{i=0}^{N_b}\|y_i-\hat{y}_i\|_2^2 \tag{33}$$

其中，y_i 是真实值，$\hat{y}_i$ 是模型的预测值，N 为样本数量。

最终的总体损失函数如式(34)所示：

$$L=L_{\text{task}}+\alpha L_{\text{con}}+\beta L_{\text{diff}} \tag{34}$$

其中，α，β是相互作用权重，决定每个正则化分量对总体损失的贡献。

3　实验过程与结果分析

本章介绍了本文选取的数据集、基准方法、实验参数设置，同时对实验结果进行了展示与分析。

3.1　数据集分析

对本文方法在两个广泛应用的多模态情感分析数据集上进行了测试，即 CMU-MOSI 和 CMU-MOSEI。CMU-MOSI

数据集包含了2199个带有情感分数的意见视频片段，这些情感分数用来衡量从负面到正面的情感倾向，分数范围设定为[−3,3]。而CMU-MOSEI数据集则包含了来自1000个不同说话者的23453个视频剪辑，每个视频剪辑同样标注有从−3到3的情感分数。在实验过程中，使用了CMU多模态软件开发工具包(CMU-Multimodal SDK)所提供的分割方法来处理这些视频数据，这种方法能更细致地分析和理解视频中的情感表达。实验中将数据分为训练、验证和测试集，以完成分类任务，具体的切分情况如表1所列。

表1 数据集划分

Table 1 Division of datasets

类型	CMU-MOSI	CMU-MOSEI
训练集	1284	16326
验证集	229	1871
测试集	686	4659
总和	2199	22856

3.2 基准方法

为了验证本文模型在情感分析任务上的有效性，本文选取了多模态情感分析领域部分最先进的方法作为对比。

TFN[8]：通过张量分解来融合多模态特征，以降低模型复杂度并提高参数效率。它通过学习不同模态间的相关性，增强了模型对情感状态的理解。

LMF[9]采用了低秩融合策略，将多模态信息融合在一个低维空间中。这种方法有助于减少计算量，同时保留了关键的多模态交互信息。

MFN[27]：引入了记忆机制来存储和更新模态间的交互信息。这种设计允许模型在处理序列数据时，更有效地利用长短期依赖关系。

MULT[4]：利用统一的Transformer架构来处理多种类型的输入数据。它通过共享参数和注意力机制，实现了模态间的有效融合。

MISA[5]：提出了模态不变和模态特有的表示，通过学习跨模态的共同特征和模态特有的特征，提高了多模态情感分析的准确性。

MAG-BERT[11]：结合了BERT的强大语言表示能力和多模态注意力机制，以更好地理解和融合来自不同模态的信息。

SELF-MM[12]：利用自监督的单模态标签生成方法，联合多任务学习方法，挖掘多模态表征的一致性和差异性。

MCGMF[26]：利用跨模态门控机制去噪并提取互补信息，结合权重和相似约束关注模态情感贡献差异性与表达一致性。

TETFN[28]：通过学习面向文本的跨模态映射，利用注意力机制和预测机制将文本与非文本表示结合，创建统一的多模态表示。

TMBL[29]：提出双模态绑定机制处理模态特定特征，三模态绑定机制处理模态无关特征，从而促进跨模态交互。

3.3 实验细节

3.3.1 实验配置

本实验基于PyTorch框架构建，运行于Ubuntu22.04操作系统，配置为NVIDIA GeForce RTX4090。在标准计算环境下，使用高性能GPU进行加速，确保训练过程的高效性。这一实验环境可广泛应用于类似的情感分析研究。

3.3.2 实验参数和评估指标

CMU-MOSI和CMU-MOSEI数据集的情感标注预测可以看作回归问题，因此本文引入了绝对平均误差MAE、皮尔逊相关性Corr最为主要的评价指标。对于分类任务，用加权的F1-score和二进制分类准确度Acc2作为评估参数，来测量本文模型分类结果的精度和准确度。使用这些综合评价指标，可以有效地评估模型在不同任务中的性能和鲁棒性，确保全面分析其在多模态情感分析中的能力。

3.4 实验结果与对比分析

3.4.1 与基准方法的对比实验

表2列出了在多模态情感分析研究的数据集CMU-MOSI和CMU-MOSEI上进行的实验结果。为保证实验的公平性，本文在相同实验环境下复现了所有对比模型，表中加粗的内容表示各项指标表现最佳的数据。结果显示，本文模型在除MAE以外的指标落后于较新的模型以外，其他各项指标均优于所有基线模型。通过定量分析，在两个数据集上，所提方法不仅展现了最先进的性能，还有强大的泛化能力。

表2 不同模型在数据集CMU-MOSI和CMU-MOSEI上的情感预测结果

Table 2 Sentiment prediction results of different model on the CMU-MOSI and CMU-MOSEI datasets

Model	CMU-MOSI					CMU-MOSEI				
	Acc-7	*Acc-2*	*F1*	*MAE*	*Corr*	*Acc-7*	*Acc-2*	*F1*	*MAE*	*Corr*
TFN	32.1	80.2	80.1	0.925	0.662	50.2	82.6	82.3	0.570	0.716
LMF	32.8	80.1	80.0	0.931	0.670	48.0	83.7	83.8	0.568	0.727
MFN	34.2	80.0	80.0	0.951	0.665	51.1	84.0	83.9	0.575	0.720
MULT	40.0	80.5	80.5	0.918	0.685	52.1	84.0	83.9	0.564	0.732
MISA	42.3	83.5	83.5	0.752	0.784	52.2	84.3	84.3	0.550	0.758
MAG-BERT	42.9	84.6	84.6	0.730	0.789	52.8	85.1	85.1	0.558	0.761
SELF-MM	45.8	84.9	84.8	0.731	0.785	53.0	85.2	85.2	**0.540**	0.763
MCGMF	45.6	83.5	83.5	0.763	0.756	52.8	85.6	85.6	0.541	0.768
TETFN	—	86.10	86.07	**0.717**	0.800	—	85.18	85.27	0.551	0.748
TMBL	36.3	83.84	84.29	0.867	0.762	52.4	85.84	85.92	0.545	0.766
Ours	**46.1**	**86.89**	**86.83**	0.728	**0.802**	**53.2**	**86.95**	**86.94**	0.541	**0.779**

3.4.2 主体注意力的主体模态选择实验

为了验证模型中对自适应选择主体模态的有效性，设计了一系列实验，通过在MOSEI数据集上让不同的模态作为主体进行交互，探讨主体模态的选择对多模态预测任务的影响。具体来说，在固定文本为主体模态和动态选择主体模态两种情况下进行了模型训练和评估。

文本为主体模态：在此设置下，文本模态固定为主导信息源，音频和视频作为辅助模态进行融合。我们期望文本的语义信息能够提供强大的上下文支持，帮助音频和视频模态更好地进行跨模态交互。

动态选择主体模态：在此动态融合框架中，模态主导权由模型根据注意力机制计算得到。我们期望在不同的样本下模型能够选择合适的主体模态，其他模态作为辅助进行跨模态交互，使主体注意力模块摆脱固定文本模态的选择限制。

实验结果如表 3 所列，可以观察到，与固定文本为主体模态的情况相比，动态选择主体模态时，模型的性能在 MOSEI 数据集上有一定程度的提升。MOSEI 数据集提供的模态信息更丰富，使得动态选择可以充分学习哪些情况下视频或音频比文本更重要，避免固定文本带来的信息瓶颈并提升预测能力。

表 3　主体模态选择实验

Table 3　Subject-modal selection experiments

Lead-Modal	*MAE*	*Corr*	*Acc*/%	*F1*/%
Text(MOSEI)	0.545	0.771	86.58	86.36
DMS(MOSEI)	0.541	0.779	86.95	86.93

3.4.3　消融实验

1）不同模态对模型的影响

考虑到不同模态在情感交流中所携带的信息量不同，为了探究不同的模态组合对模型性能的贡献，在 CMU-MOSI 上进行了一系列消融实验，结果如表 4 所列。

表 4　模态消融实验

Table 4　Modal ablation experiments

Modal	*MAE*	*Corr*	*Acc*/%	*F1*/%
T	0.802	0.760	82.58	82.61
A	1.454	0.008	44.71	61.83
V	1.454	0.005	43.93	61.28
T,A	0.766	0.792	86.12	86.20
T,V	0.801	0.788	85.09	85.13
A,V	1.192	0.011	63.21	63.21
T,A,V	0.728	0.802	86.89	86.83

实验结果表明，当删除任意模态时，模型的所有指标均出现了一定程度的下降。在 3 种单一的模态中，文本模态在各项指标上均展现出显著优势。由此可见，文本模态在多模态情感分析任务中蕴含较为丰富的情感信息，更适合于情感识别任务。进一步对比单模态与双模态组合的效果，可以发现多模态的组合显著优于单一模态，这初步验证了不同模态间情感信息存在互补性。特别地，音频模态和视觉模态的组合效果表现较差，而包含文本模态的组合则均表现出较高的性能提升，进一步验证了文本模态在多模态情感分析中的关键作用。而同时使用 3 种模态时，模型性能显著提升，各项指标均达到了最佳水平。不同模态间确实能实现有效的信息互补，凸显了多模态情感分析相较于单一模态分析的优越性。

2）不同模块对模型的影响

本文模型主要由 4 个重要模块组成，分别是特征提取模块（FE）、跨模态主体注意力模块（CMD-A）、跨空间域门控信号（CDG）模块、多模态融合模块。为了验证模型中各个模块的有效性，在 CMU-MOSI 数据集上进行了以下消融实验，结果对比如表 5 所列。

移除主体注意力机制：去除原模型中的主体注意力机制，直接采用 Cross-modal Attention 的方式将文本信息与视频、音频进行融合，拼接 4 种融合向量并进行预测。

移除跨空间域门控：去除原模型中的跨空间域门控信号，对模态共享空间和模态私有空间的数据进行主体注意力机制的融合，然后拼接 4 种融合向量并进行预测。

移除特征提取模块：去除原模型中对多模态特征的进一步提取，采用经典模型 MISA 中的特征提取方法，将提取的特征向量输入融合模块中，拼接 4 种融合向量并进行预测。

首先，移除主体注意力模块，性能出现了显著的下降，Acc-2 下降了 5.69 个百分点，F1 下降了 5.55 个百分点，Corr 降低 0.037，MAE 上升 0.064，这是因为主体注意力机制中，主体模态能够在任务中占主导地位，而辅助模态则通过加强主体模态信息的语境或情感补充来提升模型的整体性能。这种方式优化了传统 crossmodal-attention 中对所有模态的均等关注，避免了无差别的信息融合，并通过跨模态的信息聚合与广播，使主体模态更加充分地感知到辅助模态特征中细粒度的局部信息。

其次，移除跨空间域门控后，性能有稍微下降，Acc-2 下降 2.82 个百分点，F1 下降 2.77 个百分点，Corr 降低 0.020，MAE 上升 0.033，这是因为跨空间域门控确保了私有空间信息能够有效补充共享空间融合信息的不足，从而提升了模型的表现。移除该门控后，模型失去了利用私有空间信息来补充共享空间融合向量的能力，导致了不同空间域信息交互不充分，进而影响了模型的性能。

之后，移除特征提取模块后，模型性能出现了一定程度的降低，Acc-2 下降 0.55 个百分点，F1 下降 0.49 个百分点，Corr 降低 0.008，MAE 上升 0.014，这是退化现象，原因是采用统一特征提取方法时，输入特征质量出现了一定程度的降低，而本文的特征提取方法能更细致地捕捉特征中的细粒度情感线索，如 COVAREP 的声门特征、Facet 的微表情时序模式，并且在相同的特征条件下，模型相较于基线模型最低仍保持 0.44 个百分点的 Acc-2 分数优势和 0.43 个百分点的 F1 分数优势，证明了模型的融合机制具有独立于特征提取的理论增益。

表 5　模块消融实验对比

Table 5　Comparison of module ablation experiments

Model	*MAE*	*Corr*	*Acc*-2/%	*F1*/%
Ours	0.728	0.802	86.89	86.83
w/o(CMD-A)	0.792	0.765	81.20	81.28
w/o(CDG)	0.761	0.782	84.07	84.06
w/o(FE)	0.742	0.794	86.34	86.34

结束语　本文提出了一种基于主体注意力与多空间域信息协同的多模态情感分析模型，旨在通过动态选择主导模态进行多模态交互及多空间域信息交互，显著提升多模态情感分析的性能。该模型通过实现动态主体注意力机制，促进了模态的深层次融合，并打破了固定文本带来的信息瓶颈；同时通过跨空间域门控有效实现了私有空间与共享空间信息的充分交互，显著增强了模型在情感分析任务中的表现。在实验部分，在 CMU-MOSI 和 CMU-MOSEI 两个广泛使用的情感分析数据集上进行了验证，实验结果表明，模型在这两个数据

集上都取得了优异的性能，验证了其在实际情感分析任务中的有效性。

尽管本文模型在 CMU-MOSI 和 MOSEI 数据集上展现了较强的性能，但仍存在一定的局限性。这一局限性源于两个主要因素。首先，文本作为信息载体在情感分析中含有丰富的情感和语义信息，因此本研究选择动态确定主体模态，相比于固定文本模态为主体模态，在小数据集 MOSI 的预测效果上有稍微的退步，这可能是因为数据量不足以支撑可靠的模态选择机制，在情感分析中的贡献未能得到充分体现。其次，尽管音频和视频模态携带重要的情感信息，但提取这些模态中的抽象特征仍面临较大挑战。音频特征需要处理背景噪声和多种音调变化，而视频特征的提取又需要跨时空维度进行有效的信息抽象，在情感信息贡献不均的情况下，音频和视频模态的潜力可能会被低估。因此，未来的研究将致力于探索更加高效的音频和视频特征提取方法，并优化模态间的协同融合策略，旨在更好地发掘各模态之间的关联性，突出音频和视频模态在特定场景的作用，从而提升情感分析的整体性能。

此外，尽管模型在现有数据集上表现出色，但其泛化能力仍有提升空间。为进一步增强模型的适应性和鲁棒性，未来的研究将引入跨领域的数据集和多样化的应用场景。这将有助于推动模型在更广泛实际应用中的部署，为情感分析技术在实际应用中的广泛推广奠定坚实的技术基础。

参 考 文 献

[1] MIAO Y Q, YANG S, LIU T L, et al. Multimodal Sentiment Analysis based on Cross-Modal Gating Mechanism and Improved Fusion method [J]. Computer Applications and Research, 2023, 40(7): 2025-2030, 2038.

[2] MORENCY L P, MIHALCEA R, DOSHI P. Towards multimodal sentiment analysis: Harvesting opinions from the web [C]// Proceedings of the 13th International Conference on Multimodal Interfaces. Alicante, Spain: ACM, 2011: 169-176.

[3] LUO Y Y, WU R, LIU J E, et al. Multimodal Sentiment Analysis Method Based on Adaptive Weight Fusion [J]. Journal of Software, 2024, 35(10): 4781-4793.

[4] TSAI Y H, BAI J, LIANG P P, et al. Multimodal transformer for unaligned multimodal language sequences [C]// Proceedings of the Conference. Association for Computational Linguistics. Meeting. Florence, Italy: Association for Computational Linguistics, 2019.

[5] 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. Seattle, USA: ACM, 2020: 1122-1131.

[6] ZHONG T, FENG G, LIN J Z, et al. Sentiment Analysis Aimed at Multi-Source Information Clustering and Private Feature Learning [J]. Computer Engineering and Applications, Early Access, 2015: 1-12.

[7] PORIA S, CHATURVEDI I, CAMBRIA E, et al. MKL Based Multimodal Emotion Recognition and Sentiment Analysis [C]// IEEE 16th International Conference on Data Mining (ICDM 2016). San Diego, USA: IEEE, 2016: 439-448.

[8] ZADEH A, CHEN M, PORIA S, et al. Tensor Fusion Network for Multimodal Sentiment Analysis [J]. arXiv: 1707. 07250, 2017.

[9] LIU Z, SHEN Y, LAKSHMINARASIMHAN V B, et al. Efficient low-rank multimodal fusion with modality-specific factors [J]. arXiv: 1806. 00064, 2018.

[10] WANG Y S, SHEN Y, LIU Z, et al. Words can shift: Dynamically adjusting word representations using nonverbal behaviors [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 7216-7223.

[11] PUTRA B, AZIZAH K, MAWALIMC O, et al. MAG-BERT-ARL for Fair Automated Video Interview Assessment [C]// IEEE Access. 2024.

[12] YU W M, XU H, YUAN Z Q, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 10790-10797.

[13] LIU Y H. Roberta: A robustly optimized bert pretraining approach [J]. arXiv: 1907. 11692, 2019, 364.

[14] EYBEN F, WÖLLMER M, SCHULLER B, et al. Opensmile: the Munich versatile and fast open-source audio feature extractor [C]// Proceedings of the 18th ACM International Conference on Multimedia. 2010: 1459-1462.

[15] DEGOTTEX G, KANE J, DRUGMAN T, et al. COVAREP—A Collaborative Voice Analysis Repository for Speech Technologies [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014). Florence, Italy: IEEE, 2014: 960-964.

[16] GUSTAFSON, L, ROLLAND, C, RAVI, N, et al. FACET: Fairness in Computer Vision Evaluation Benchmark [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE, 2023: 20370-20382.

[17] AMOS B, LUDWICZUK B, SATYANARAYANAN M. OpenFace: A General-Purpose Face Recognition Library with Mobile Applications [J]. CMU School of Computer Science, 2016, 6(2): 20.

[18] YONG L, WANG Y Z, CUI Z. Decoupled Multimodal Distilling for Emotion Recognition [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 6631-6640.

[19] HAN D C, YE T Z, HAN Y Z, et al. Agent Attention: On the Integration of Softmax and Linear Attention [C]// European Conference on Computer Vision. Cham, Switzerland: Springer, 2025: 124-140.

[20] ZHOU Y, YU J, FAN J P, et al. Multi-Modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1821-1830.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems (NeurIPS). Red Hook, NY, USA: Curran Associates, Inc. , 2017: 5998-6008.

[22] HAO S, WANG H Y, LIU J Q, et al. CubeMLP: An MLP-Based

Model for Multimodal Sentiment Analysis and Depression Estimation [C]// Proceedings of the 30th ACM International Conference on Multimedia. Lisbon, Portugal: ACM, 2022: 3722-3729.

[23] CHEN W Z, HOU Y. A Temporal Multimodal Sentiment Analysis Model Fusing Multi-Level Attention and Sentiment Scale Vectors [J]. Data Analysis and Knowledge Discovery, 2025, 9(3): 1-18.

[24] LUO Y Y, WU R, LIU J F, et al. Multimodal Sentiment Analysis Method for Emotion-Semantic Inconsistency [J]. Computer Research and Development, 2025, 62(2): 374-382.

[25] ZELLINGER W, GRUBINGER T, LUGHOFER E, et al. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning [J]. arXiv: 1702.08811, 2017.

[26] ZMIAO Y Q, YANG S, LIU T L, et al. Multimodal Sentiment Analysis Based on Cross-modal Gating Mechanism and Improved Fusion Method[J]. Application Research of Computers, 2023, 40(7): 2025-2030, 2038.

[27] ZADEH A, LIANG P P, PORIA S, et al. Memory Fusion Network for Multiview Sequential Learning [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 5634-5641.

[28] DI W, GUO X T, TIAN Y M, et al. TETFN: A Text Enhanced Transformer Fusion Network for Multimodal Sentiment Analysis [J]. Pattern Recognition, 2023, 136: 109259.

[29] HUANG J H, ZHOU J, TANG Z C, et al. TMBL: Transformer-Based Multimodal Binding Learning Model for Multimodal Sentiment Analysis [J]. Knowledge-Based Systems, 2024, 285: 111346.

FENG Guang, born in 1973, Ph.D, professor-level senior experimenter, master's supervisor. His main research interests include classroom streaming, big data and artificial intelligence.