DOI:10.16356/j.2097-6771.2026.03.020

基于特征解耦与跨模态深度交互增强的多模态情感分析

赵智伟[1],张顺香[1,2,3],孙 亮[1],魏可欣[1],陈 梦[1]
(1.安徽理工大学计算机科学与工程学院,淮南 232001;2.淮南师范学院计算机学院,淮南 232038;
3.合肥综合性国家科学中心人工智能研究院,合肥 230026)

摘要: 当前多模态情感分析(Multimodal sentiment analysis, MSA)模型主流方法使用跨模态注意力机制处理不同模态特征信息,但该类方法没有考虑到不同模态特征之间的相似性与差异性,在跨模态交互中容易产生模态相似冗余信息并增加噪声,导致模型性能降低。本文提出一种基于特征解耦与跨模态深度交互增强(Feature decoupling and cross-modal deep interaction enhancement, FD-CMDIE)的多模态情感分析模型。首先在特征提取模块引入NeoBERT提取高质量文本特征,使用堆叠长短期记忆(Long short-term memory, LSTM)网络提取视觉与听觉特征。然后利用共同编码器与独立编码器将3种模态特征解耦成相似性特征与相异性特征,并使用对比学习以文本相似性特征为锚点,在特征空间中拉近相似性特征,同时推远相异性特征。最后设计一种跨模态交互增强网络实现解耦后特征的深度交互与融合,并利用门控注意力池化模块过滤交互产生的噪声信息。在两个基准数据集上进行实验,并与多个当前先进方法比较,在绝大部分指标上都超越了当前先进方法,验证了本文方法的有效性。

关键词: 多模态情感分析;特征解耦;跨模态深度交互增强;对比学习;NeoBERT
中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1005-2615(2026)03-0666-16

Feature Decoupling and Cross-Modal Deep Interaction Enhancement for Multimodal Sentiment Analysis

ZHAO Zhiwei[1], *ZHANG Shunxiang*[1,2,3], *SUN Liang*[1], *WEI Kexin*[1], *CHEN Meng*[1]
(1. School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, China;
2. School of Computer, Huainan Normal University, Huainan 232038, China; 3. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230026, China)

Abstract: Current multimodal sentiment analysis (MSA) models predominantly employ cross-modal attention mechanisms to process feature information from different modalities. However, these approaches often overlook the inherent similarities and dissimilarities among modal features, which can easily lead to the generation of redundancy from modal similarities and an increase in noise during cross-modal interaction, thereby degrading model performance. To address these issues, this paper proposes a novel multimodal sentiment analysis model based on feature decoupling and cross-modal deep interaction enhancement (FD-CMDIE). Firstly, for feature extraction, NeoBERT is utilized to extract high-quality textual features, while stacked long short-term memory (LSTM) networks are employed for visual and acoustic features. Subsequently, common and private encoders are used to decouple the features of the three modalities into

基金项目: 国家自然科学基金面上项目(62476005, 62076006);认知智能全国重点实验室开放课题(COGOS-2023HE02);安徽高校协同创新项目(GXXT-2021-008)。
收稿日期: 2025-08-15;**修订日期:** 2026-10-06
通信作者: 张顺香,男,教授,博士生导师,E-mail:sxzhang@aust.edu.cn。

引用格式: 赵智伟,张顺香,孙亮,等. 基于特征解耦与跨模态深度交互增强的多模态情感分析[J]. 南京航空航天大学学报(自然科学版),2026,58(3):666-681. ZHAO Zhiwei, ZHANG Shunxiang, SUN Liang, et al. Feature decoupling and cross-modal deep interaction enhancement for multimodal sentiment analysis[J]. Journal of Nanjing University of Aeronautics & Astronautics(Natural Science Edition),2026, 58(3):666-681.

similar and dissimilar features. Contrastive learning is then applied, using the textual similar features as anchors, to pull similar features from different modalities closer in the feature space while pushing dissimilar features further apart. Finally, a cross-modal interaction enhancement network is designed for deep interaction and fusion of the decoupled features, and a gated attention pooling module is utilized to filter out noise generated during the interaction. Experiments conducted on two benchmark datasets demonstrate that our proposed method surpasses several state-of-the-art approaches across most metrics, validating its effectiveness.

Key words: multimodal sentiment analysis (MSA); feature decoupling; cross-modal deep interaction enhancement (CMDIE); contrastive learning; NeoBERT

随着互联网的飞速发展,微博、BiliBili、抖音、小红书等网络社交平台用户激增,并且更多用户倾向于通过视频来表达自己的情绪和观点,这使得多模态情感分析(Multimodal sentiment analysis, MSA)相关研究快速发展[1]。多模态情感分析利用文本、视觉和听觉3种常见模态的序列数据,实现情感分析预测[2]。相比较于传统的文本单模态情感分析,MSA在文本模态基础上融合声学信息(声音大小、音调高低)与视觉信息(面部表情、肢体动作)。这些不同模态的信息能够相互补充,有助于解决语义分歧等问题,并提高情感预测的准确性。目前,多模态情感分析已经成为热门研究领域,在舆论分析、智能客服和人机交互[3]等领域有着广泛的应用。

当前多模态分析领域主要关注如何有效利用不同模态的信息以及通过不同的融合方法实现多模态特征有效融合[4]。大多数方法采用注意力机制实现不同模态数据信息的提取与融合,但是由于不同模态数据中情感表达的内容与方式存在差异,并且都一定程度上包含着一些与情感无关的噪声信息,在视觉模态与听觉模态中尤为明显[5],这影响特征提取的质量以及后续多模态特征融合的效果,故MSA任务的关键挑战在于如何获取高质量的单模态特征,以及使用更有效的融合方法来实现模态间情感信息交互与融合,提升最后的情感预测效果[6]。当前广泛使用的注意力机制被认为是跨模态信息交互的有效方法[7],注意力机制通过模拟人类集中注意力观察事物的方式[8],让模型能够重点关注输入序列中的关键部分,从而进行高效学习。但是不同模态之间存在着相似的共有信息与特定的相异性信息,当使用跨模态注意力机制实现不同模态间信息交互时,会产生大量相似性冗余信息与噪声信息,也无法有效利用各模态独有的相异性信息。

针对多模态情感分析领域中如何提高单模态特征的表达能力以及减少模态交互冗余信息的关键问题,本文提出基于特征解耦与跨模态深度交互增强(Feature decoupling and cross-modal deep interaction enhancement, FD-CMDIE)的多模态情感分析模型。在特征提取模块中首先引入使用NeoBERT[9]预训练模型提取高质量的文本特征。它是新一代双向编码器,采用最优深度与宽度比,在保持BERT-base的768宽度的同时增加了深度。使用旋转位置嵌入有效处理长序列,支持4 096个词块的上下文长度。并且在拥有600 B标记数据的RefinedWeb上进行预训练,能够更有效地提取文本序列特征,捕获更丰富的上下文语义信息。此外,使用堆叠长短期记忆网络(stacked long short-term memory, sLSTM)有效捕捉视觉与听觉的时序特征。之后在特征解耦模块中,通过公共的模态相似性编码器与独立的模态相异性编码器实现原始特征解耦,得到每个模态的相似性特征与独有的相异性特征,并利用对比学习,以文本相似性特征为锚点,拉近其他相似特征,推离相异特征,增强模型对模态相似性与差异性的分辨能力,提高特征解耦效果。在多模态特征交互融合阶段,设计跨模态深度交互增强网络(Cross-modal deep interaction enhancement, CMDIE),将3个模态的相似性特征融合为全局一致特征,分别与单模态相异性特征进行多层跨模态深度交互,捕捉彼此重要信息,并减少冗余信息的产生,实现信息交互与增强。同时设计了门控注意力池化模块(Gated attention pooling, GAP),在每层交互后自适应衡量来自不同模态信息的重要性,并调整权重,减少交互产生的噪声信息。最后,连接融合经过多层交互增强后的全局相似性特征与单模态相异性特征,输入到多层感知机中进行预测。使用平均绝对误差作为主要损失函数,对比损失函数作为辅助损失函数进行优化。其中对比损失函数旨在减小相似性特征的差异,而增大相异性特征与相似性特征的差异。在CMU-MOSI与CMU-MOSEI情感分析数据集上进行实验,使用二分类精确度、二分类F1分数、七分类精确度、皮尔逊相关系数和平均绝对误差作为情感分析性能衡量指标。与当前先进方法比较,实

验结果表明本文模型的有效性。

1 相关工作

1.1 情感分析发展

情感分析旨在让计算机通过对文本、音频和视频等不同信息载体中的情感信息进行分析理解,进而识别出其中所包含的具体情绪或情感倾向,例如喜悦、悲伤和愤怒等。早期的情感分析主要聚焦于文本数据,相关研究集中在文本方面级情感分析,如Zhang等[10]构建全局-局部的提取机制,整合全局与局部上下文信息,提高情感的预测准确性。

随着网络社交平台多模态数据的持续增长,融合其他模态数据的多模态情感分析需求也不断增大。Morency等[11]最早提出使用三模态数据进行情感,通过自动识别话语文本中的情感线索,生成文本特征,再与从视频中自动提取的视觉特征、音频特征串联融合,使用三模态分类器进行情感分析。而后续的Poria等[12]在提取3种模态特征后,分别使用基于循环的特征子集选择器和基于主成分分析的特征选择器来减少特征数量,以此提高特征的质量,最后将选择处理后的特征向量串联,利用多核学习算法训练分类器。

1.2 经典模型方法

当情感分析研究进入多模态情感分析阶段,研究者们不断提出新的方法与模型,以提升情感分析任务效果,在这过程中逐渐产生了几类经典模型。

1.2.1 基于张量的融合模型

基于张量的模型主要通过各模态特征表示张量之间积实现信息交互。如Zadeh等[13]提出一种张量融合网络(Tensor fusion network,TFN),通过对三模态张量使用三阶笛卡尔积,实现文本、图像、音频三模态之间的单模态、双模态与三模态交互。之后为了解决张量计算过于复杂的问题,Liu等[14]提出了一种低秩多模态融合(Low-rank multimodal fusion,LMF)方法,在TFN的基础上做出了改进,使用张量分解方法,将权值分解为低秩因子,减少计算参数,输出一个低维向量进行预测。而为了解决缺失模态以及噪声干扰对张量表示的影响,Liang等[15]提出一种基于张量秩最小化正则化的时间张量融合网络(Temporal tensor fusion network, T2FN),在TFN的基础上添加了时间分量,提高张量表示能力,并对张量的秩进行归一化。

1.2.2 基于神经网络的模型

随着深度学习的不断发展,多模态情感分析也结合神经网络提高模型性能,如循环神经网络和长短期记忆网络。Liang等[16]提出递归多阶段融合网络(Recurrent multistage fusion network, RMFN),将融合分解为多个阶段,每个阶段使用门控机制动态调整模态间的信息流,采用LSTM结构,逐步建模多模态数据的时序依赖关系。此外Majumder等[17]提出一种感知上下文的分层融合模型,使用门控循环单元(Gate recurrent unit ,GRU)在层次化融合的基础上引入上下文GRU,解决长时序上下文依赖问题。

1.2.3 基于记忆网络的模型

记忆网络能够实现长期记忆功能,虽然RNN、LSTM和其他变种GRU具有一定的记忆能力,但是在Kumar等[18]看来,这些记忆能力是不足的,他们提出了记忆网络实现长期记忆。之后研究者们使用记忆网络提升MSA模型性能。Zadeh等[19]提出记忆融合网络(Memory fusion network,MFN),通过独立LSTM编码时序特征,每个模态拥有独立记忆模块存储长期上下文信息,通过门控机制动态更新记忆,并实现分层融合。此外,Yu等[20]提出一种基于自监督多任务学习的多模态(Self-supervised multi-task multimodal, Self-MM)情感分析框架,设计一个自监督学习的单模态标签生成模块实现单模态自监督学习,最后联合多模态和单模态任务,分别学习一致性与差异性。

1.3 大语言模型方法

近年来,随着大语言模型(Large language models, LLMs)技术的迅速发展,多模态大模型(Multimodal large language models, MLLMs)逐渐被引入多模态情感分析任务中。与传统方法中依赖复杂的特征融合与交互网络不同,MLLMs利用强大的图像、音频、视频和文本理解与推理能力来完成情感分析预测任务,此类研究主要分为两种技术路径。

1.3.1 基于提示学习的通用模型

通用大模型无需专门的任务微调,依赖其在大规模跨模态数据上训练而学得的强大泛化能力,直接从图像、视频、音频及文本输入中分析情感信息。如GPT-4[21]、Gemini[22]等具备跨模态输入能力的通用模型,通过提示学习实现零样本或少样本的情感预测。然而,这种通用模型在特定情感任务上的稳定性与细粒度情感区分能力仍然有限。

1.3.2 基于大模型微调的专用模型

基于微调大模型的方法,将LLaVA[23]、Video-ChatGPT[24]等开源多模态大模型作为基础,通过引入投影层将视觉、音频等模态特征与语言模型对齐,并在多模态情感分析数据集上进行指令微调,使微调后的模型专精于情感分析任务。如Liu

等[25]提出的EmoLLMs系列模型，构建了首个包含23.4万条样本的多任务情感指令数据集AAID，涵盖3类分类任务与两类回归任务，并建立了统一评测基准。EmoLLMs通过在多任务情感指令数据上对多种LLM进行精调，实现了从情感极性到情感强度的综合分析，在多数情感任务上超越 ChatGPT和GPT-4。

为了充分利用LLMs在语言生成与推理方面的能力，近期的研究也有探索结合大模型推理能力与传统情感模型结构优势的混合框架。如Han等[26]提出了基于大语言模型的多模态情感问答框架DEQA，将视觉与文本模态信息通过“情感描述生成”机制转化为自然语言，再使用大语言模型进行推理，从而实现可解释的多模态情感分析。

多模态大模型虽然拥有强大的跨模态对齐与推理能力，但仍面临计算资源消耗大、任务适应性不足等问题。因此，面向具体情感分析任务的轻量化专用模型仍具有实际应用价值与研究意义。

1.4　最新模型方法

自从注意力机制[8]被提出以来，已经被广泛应用于众多研究领域，在多模态情感分析中也被认为是实现模态间信息交互与融合的有效方法，所以近年来被广泛应用于MSA任务中。Devlin等[27]提出双向上下文编码器BERT（Bidirectional encoder representations from Transformers），通过堆叠多层Transformer实现深度上下文建模，并通过预训练-微调范式提出统一的迁移学习框架，预训练后的BERT能快速适配下游任务，BERT成为MSA任务中最常用的文本特征提取器。Rahman等[28]也利用了注意力机制，使用语言模型BERT和XL-Net在特定任务进行微调时利用文本之外的模态信息，通过分层跨模态注意力实现高效的特征融合。

Tsai等[29]直接使用注意力机制实现多个模态之间特征信息的交互融合，基于跨模态注意力提出一种多模态Transformer（Multimodal Transformer，MulT）模型。每个模态通过独立的Transformer编码器提取特征，再使用双向跨模态注意力建模不同模态间的交互，并通过独立位置编码处理未对齐数据。而Ghosal等[30]提出一种基于上下文感知的跨模态注意力机制，考虑了所有可能的双模态组合，分别计算注意力，构建多模态表示。

为了学习模态不变表示与模态特定表示，Hazarika等[31]将每个模态投影到两个不同的子空间，在模态不变子空间学习共性减少模态差距，在模态专用空间学习特有的特征。而Mai等[32]实现混合三模态对比学习，结合模态内对比学习与跨模态对比学习，增强三模态标准的一致性与判别性。Yang等[33]则提出一个统一的学习框架ConFEDE，结合对比学习与对比特征分解来增强模态特征表示，通过多任务学习提升模型预测效果。Hu等[34]发现情感分析与情绪识别在联合训练中相互促进，因此提出了统一多模态情感分析和离散情绪识别的框架UniMSE。在模态与样本间引入对比学习，捕捉情感和情绪之间的相似性和差异性。

由于跨模态注意力计算复杂，Sun等[35]提出了一个具有双层特征恢复的高效多模态Transformer（Efficient multimodal Transformer，EMT）框架。利用3个单模态特征表示构建全局上下文，与局部单峰进行交互，使用跨模态注意力更新自身信息。此外Sun等[36]在跨模态Transformer中加入卷积层，有效建模局部依赖关系，同时挖掘全局多模态特征与局部特征之间的内在关联。Cheng等[37]同样将注意力与卷积结合，采用膨胀因果卷积与多头注意力机制构建注意力时序卷积网络，并设计多层特征融合（Muti-layer feature fusion，MFF）网络实现特征融合。

这些模型使用各不相同的跨模态注意力机制实现不同模态之间的信息交互与融合，但是在跨模态交互过程中难免会产生模态间相似信息的冗余，并且不加区分地融合会引入多余的噪声信息，影响最终的模型效果。而在使用模态特征分解的方法中，存在特征分解效果不佳，或是没有充分发挥分解特征作用的问题，导致多模态特征融合不充分。

为解决上述方法中存在的不足，本文在这些研究的基础上做出改进。为了提取包含更丰富语义信息的文本特征，引入NeoBERT提取文本特征。结合特征解耦与对比学习得到有效的模态相似性特征与相异性特征，并通过多层跨模态交互增强网络实现特征信息的深度融合，减少相似性冗余信息的产生，使用门控注意力池化模块减少噪声信息。

2　本文方法

本节详细介绍特征提取、对比特征解耦以及跨模态交互增强网络。本文模型架构如图1所示，先将特征提取阶段得到的单模态特征通过相似编码器与相异编码器，实现特征解耦，得到模态相似性特征与模态相异性特征，并通过对比学习提升模型的特征解耦能力，再通过跨模态交互增强网络实现信息深度交互与融合，最终联合预测损失和对比损失进行优化。图1中各变量含义见本文相关章节说明。

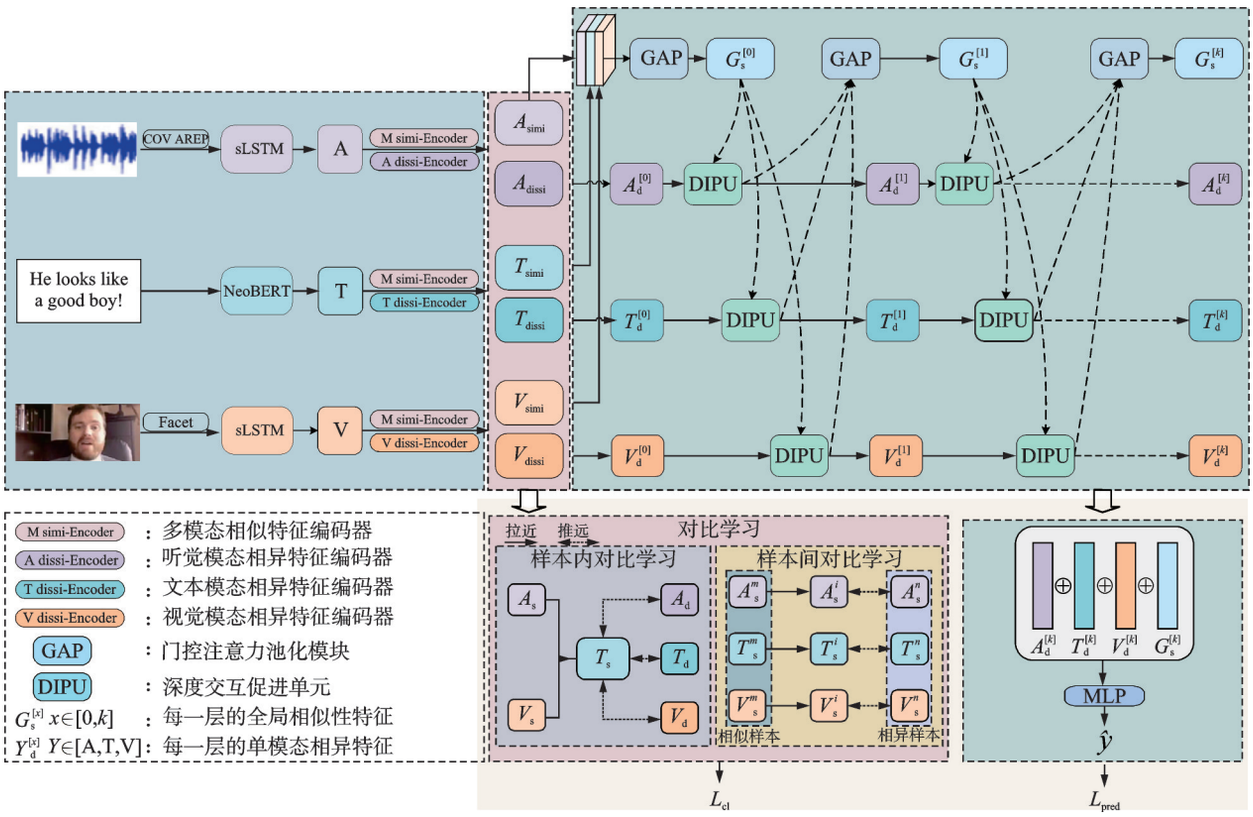


图 1 FD-CMDIE 整体架构

Fig.1 Overall architecture of the FD-CMDIE

2.1 任务定义

多模态情感分析任务通过对文本、视觉和听觉模态数据进行分析，判断所包含的情感倾向。给定多模态序列数据 $\boldsymbol{X}_m \in \mathbf{R}^{l_m \times d_m}$，其中 $m \in \{t, v, a\}$，l_m 与 d_m 分别表示模态 m 的序列长度以及模态特征向量维度。将三模态序列数据输入模型，最终得到模型输出 $\hat{\boldsymbol{y}} \in \mathbf{R}$ 作为情感预测结果，用于下游多模态情感分析任务。

2.2 特征提取

本文使用 NeoBERT 预训练模型提取文本模态的特征。对于每一个文本序列 T，经过分词后在序列首部与尾部分别添加[CLS]与[SEP]两个标签，得到输入 $T=\{[\mathrm{CLS}], w_1, w_2, \cdots, w_n, [\mathrm{SEP}]\}$。$T$ 经过 Token 化之后，得到对应的 Token ID 序列 $\boldsymbol{X}_\mathrm{t}$，作为 NeoBERT 模型的输入，输入模型得到文本模态的特征表示，计算公式为

$$\boldsymbol{F}_\mathrm{t} = \mathrm{NeoBERT}(\boldsymbol{X}_\mathrm{t}; \theta_\mathrm{t}^{\mathrm{NeoBERT}}) \in \mathbf{R}^{l_\mathrm{t} \times d_\mathrm{t}} \tag{1}$$

式中：l_t 表示文本特征序列长度，d_t 表示特征向量维度，$\theta_\mathrm{t}^{\mathrm{NeoBERT}}$ 表示 NeoBERT 模型的参数，并且参数在任务中可以进行微调。

对于听觉与视觉模态数据，分别使用 COVAREP 与 Facet 从原始数据中提取初级特征向量表示 $\boldsymbol{X}_\mathrm{a}$ 与 $\boldsymbol{X}_\mathrm{v}$，由于它们具有较强的时序性，为了捕获特征中的时间依赖信息，本文采用堆叠的长短期记忆网络 sLSTM 获取听觉与视觉序列的时序特征表示，计算公式为

$$\boldsymbol{F}_\mathrm{a} = \mathrm{sLSTM}(\boldsymbol{X}_\mathrm{a}; \theta_\mathrm{a}^{\mathrm{sLSTM}}) \in \mathbf{R}^{l_\mathrm{a} \times d_\mathrm{a}} \tag{2}$$

$$\boldsymbol{F}_\mathrm{v} = \mathrm{sLSTM}(\boldsymbol{X}_\mathrm{v}; \theta_\mathrm{v}^{\mathrm{sLSTM}}) \in \mathbf{R}^{l_\mathrm{v} \times d_\mathrm{v}} \tag{3}$$

式中：l_a 与 l_v 分别表示听觉与视觉模态的时间序列长度；d_a 与 d_v 分别表示听觉与视觉模态的特征维度；$\theta_m^{\mathrm{sLSTM}}(m \in \{\mathrm{a}, \mathrm{v}\})$ 表示 sLSTM 网络的参数；$\boldsymbol{X}_\mathrm{a}$ 与 $\boldsymbol{X}_\mathrm{v}$ 分别表示输入的听觉与视觉特征。

2.3 对比特征解耦

2.3.1 特征解耦

在多模态情感分析中，情感相关信息既存在于不同模态间的共享语义之中，也存在于模态自身的特有表达中。本文将前者称为相似性信息，指在不同模态中均可表达相似情感语义、能够互相预测的部分；将后者称为相异性信息，指仅在某一模态中存在、可补充整体情感理解的部分。例如，在听觉模态中的“音调高”与视觉模态中的“微笑表情”都表达积极情绪，属于相似信息；而听觉模态中的“语速加快”或视觉模态中“眉毛上扬”类似的表情变化则属于特异性信息，它们仅在特定模态中出现，但仍对情感预测有帮助。

为了减少由于不同模态之间信息的结构与分布差异而导致的后续模态融合时产生的噪声信息，并且更有效地利用每个模态的信息，将模态特征解

耦成模态相似性特征与模态相异性特征。本文特征解耦目标是在语义表示层面上实现模态间的区分与互补。通过共同编码器提取模态共享的相似特征，通过独立编码器提取不相似特征，从而为后续的跨模态交互提供语义上可分的特征基础。

通过1个参数共享的模态相似特征编码器与3个参数独立的模态相异特征编码器实现特征解耦。两种编码器的主要结构都是由归一化层、全连接层以及随机丢弃层组成，各个组成部分的详细分析如下。

首先对输入特征进行层归一化。规范化不同模态特征的统计分布，减少原始数据特性带来的尺度影响。稳定分布的输入，不仅帮助后续线性层更快学习，加速模型收敛，提高泛化能力，同时还能够防止梯度消失或爆炸。层归一化公式为

$$\mathrm{LayerNorm}(F_{m\in\{\mathrm{t,v,a}\}})=\frac{F_m-\mu}{\sqrt{\sigma^2+\varepsilon}}\gamma+\beta \quad (4)$$

式中：μ与σ^2分别为输入的均值和方差；ε为一个很小的常数，防止分母为零；γ和β分别为缩放和偏移可学习参数，用于恢复层归一化后特征的表征能力，γ与β都是可学习参数；$F_{m\in\{\mathrm{t,v,a}\}}$为经过特征提取后的3个模态特征。

经过归一化后，特征进入一个全连接线性层，该层通过可学习的权重矩阵将特征从原始编码空间投影到解耦空间。相似性投影的线性层旨在学习一个映射，以提取模态之间的相似表征信息；差异性投影学习的映射捕捉每个模态独有表征信息。线性层公式为

$$\mathrm{Linear}(x)=\mathcal{W}x+b \quad (5)$$

式中：x为经过层归一化后的特征，$\mathcal{W}$为可学习的权重矩阵，b为可学习的偏置向量参数。

随后线性投影层的输出通过双曲正切(Tanh)激活函数，Tanh激活函数为特征变换过程增加了非线性，使模型能够学习从原始编码空间到解耦空间复杂的映射关系，并且Tanh将投影后的特征压缩到$[-1,+1]$之间，稳定特征表示的尺度，计算公式为

$$\mathrm{Tanh}(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (6)$$

最后，为了增强模型的泛化能力并防止过拟合，在Tanh激活函数输出后应用Dropout层。训练阶段以设置的概率随机丢弃部分激活单元的输出，促使模型学习到更加鲁棒的相似性特征与相异性特征。

将原始特征F_{t}、F_{v}和F_{a}输入模态相似特征编码器以及各自模态的相异性特征编码器，分别得到模态相似性特征T_{s}、V_{s}、A_{s}，和模态相异性特征T_{d}、V_{d}、A_{d}这6个解耦后特征。

2.3.2　对比学习

为了提高模型的模态解耦能力以及模型对不同模态间相似特征与相异特征的分辨能力，构建样本内与样本间对比学习，并整合到一个结合NT-Xent[38]对比损失与三元组对比损失的对比学习框架中。

(1) 正负样本对构建

为了扩大对比范围，增加样本对数量，构建样本间对比学习，这需要先构造每个样本的情感相似样本集合与情感相异样本集合。首先计算每个样本对之间的余弦相似度得分，计算公式为

$$\mathrm{sim}(\alpha,\beta)=\frac{\alpha^{\mathrm{T}}\beta}{||\alpha|\cdot|\beta||} \quad (7)$$

$$\mathrm{Score}^{(i,j)}=\mathrm{sim}([F_{\mathrm{t}}^{i};F_{\mathrm{v}}^{i};F_{\mathrm{a}}^{i}],[F_{\mathrm{t}}^{j};F_{\mathrm{v}}^{j};F_{\mathrm{a}}^{j}]) \quad (8)$$

式(7)计算两个向量α与β之间的余弦相似度；式(8)计算每个样本对(i,j)之间的相似度。

再对每一个样本i，选出与其具有相同多模态情感标签的样本j，再按照相似度得分排序，选出得分最高的k_{cl}个样本构成样本i的情感相似样本集合$\mathrm{similar}^{i}$。而对于与样本i情感标签不同的样本，按照相似度得分排序后，选出得分最低的k_{cl}个样本构成样本i的情感相异样本集合$\mathrm{different}^{i}$。通过构建这两个集合筛选出合适的样本，之后构建样本i的正负样本对时，使用这两个集合里的所有样本与之进行组合。

对于样本i，样本内的正样本对$\mathrm{Postive}_{\mathrm{in}}^{i}$由自身的模态相似特征对构成，如$(T_{\mathrm{s}}^{i},V_{\mathrm{s}}^{i})$和$(T_{\mathrm{s}}^{i},A_{\mathrm{s}}^{i})$，并使用来自情感相似样本集合$\mathrm{similar}^{i}$中的相似样本$j$，使用该样本的相似特征对，如$(T_{\mathrm{s}}^{j},V_{\mathrm{s}}^{j})$和$(T_{\mathrm{s}}^{j},A_{\mathrm{s}}^{j})$，扩充样本$i$的样本内正样本对。而样本间正样本对$\mathrm{Postive}_{\mathrm{out}}^{i}$仅由样本$i$与来自$\mathrm{similar}^{i}$的情感相似样本$j$之间相同模态的相似特征组成，如$(T_{\mathrm{s}}^{i},T_{\mathrm{s}}^{j})$、$(V_{\mathrm{s}}^{i},V_{\mathrm{s}}^{j})$和$(A_{\mathrm{s}}^{i},A_{\mathrm{s}}^{j})$，因为不同样本间不同模态的特征差异过大，不宜作为正样本对。

样本内负样本对$\mathrm{Negtive}_{\mathrm{in}}^{i}$由自身的文本相似特征$T_{\mathrm{s}}^{i}$与所有相异性特征组成负样本对，如$(T_{\mathrm{s}}^{i},T_{\mathrm{d}}^{i})$，并从集合$\mathrm{similar}^{i}$中选取情感相似样本$j$，使用其负样本对，如$(T_{\mathrm{s}}^{j},V_{\mathrm{d}}^{j})$扩充样本$i$的负样本对。而样本间负样本对$\mathrm{Negtive}_{\mathrm{out}}^{i}$的构造，与样本间正样本构造对类似，从$\mathrm{different}^{i}$中选取不相似样本，使用不相似样本间的同模态相似特征，如$(T_{\mathrm{s}}^{i},T_{\mathrm{s}}^{j})$。最终的正负样本对分别为

$$P^i = \text{Postive}^i_{\text{in}} \cup \text{Postive}^i_{\text{out}}$$
$$N^i = \text{Negtive}^i_{\text{in}} \cup \text{Negtive}^i_{\text{out}}$$

(2) 对比损失计算

使用NT-Xent损失函数计算二元组对比损失,计算公式为

$$\ell^i_{\text{NT-Xent}} = \sum -\log \frac{\exp\left(\text{sim}(\boldsymbol{\alpha},\boldsymbol{\beta})/\tau\right)}{\sum \exp\left(\text{sim}(\boldsymbol{\alpha},\boldsymbol{\gamma})/\tau\right)} \tag{9}$$

式中:$(\boldsymbol{\alpha},\boldsymbol{\beta})\in P^i$,$(\boldsymbol{\alpha},\boldsymbol{\gamma})\in P^i \cup N^i$且$\boldsymbol{\gamma}\neq\boldsymbol{\beta}$,$\tau$为可调节的温度参数,用于调整相似度的分布。

而三元组对比学习样本对Pair_{Tri}的构建,同样使用文本相似特征$\boldsymbol{T}^i_{\text{s}}$作为锚点,与相似特征构成正对,而与相异特征构成负对。

$$\text{Pair}^i_{\text{Tri}} = \{(a,p,n)\,|\,a=\boldsymbol{T}^i_{\text{s}};\ p=\boldsymbol{V}^i_{\text{s}},\boldsymbol{A}^i_{\text{s}};\ n=\boldsymbol{T}^i_{\text{d}},\boldsymbol{V}^i_{\text{d}},\boldsymbol{A}^i_{\text{d}}\}$$

三元组样本对(a,p,n)中,a作为锚点,p作为正样本,n作为负样本。而三元组对比损失ℓ^i_{Tri}目标是学习一个目标空间,使得相似样本之间距离缩小,而增大不相似样本之间距离,计算公式为

$$\ell^i_{\text{Tri}} = \sum \max(0, d(a,p) - d(a,n) + \text{margin}) \tag{10}$$

式中$(a,p,n)\in\text{Pair}^i_{\text{Tri}}$,$d(x,y)$为距离函数,用来衡量两个嵌入向量之间的不相似度,这里使用余弦不相似度作为距离函数,有

$$d(x,y) = 1 - \text{sim}(x,y) \tag{11}$$

而margin是可调节的正超参数,控制锚点到负样本的距离比锚点到正样本的距离至少大多少。最终的对比损失$\mathcal{L}_{\text{cl}}$计算公式为

$$\mathcal{L}_{\text{cl}} = \frac{1}{n}\sum_{i=1}^{n}\ell^i_{\text{NT-Xent}} + \lambda_{\text{Tri}}\cdot\frac{1}{n}\sum_{i=1}^{n}\ell^i_{\text{Tri}} \tag{12}$$

式中λ_{Tri}为调节三元组对比损失的权重参数。NT-Xent损失作为主要部分,最大化正样本对之间的相似度,拉近相似样本的表示距离,同时推远不相似样本的距离,从而学习有意义的特征表示。而三元组损失直接对距离进行操作,明确正负样本之间在嵌入空间中的间隔要求,更精细地调节特定样本对之间的相对距离。

2.4 跨模态深度交互增强网络

特征解耦使模型能够分别学习跨模态相似的情感信息与模态特有的补充信息。然而,单独使用这两类特征都会导致情感表示的不完整性。仅依赖相似性特征可能损失模态细节,而仅依赖特异性特征又会缺乏跨模态一致性。为了实现不同模态之间深层次的信息交互与互补增强,并且减少在跨模态交互中产生的噪声,本文设计了一个深度交互的跨模态交互增强网络。

核心思想是使用模态相似性特征$\boldsymbol{T}_{\text{s}}$、$\boldsymbol{V}_{\text{s}}$、$\boldsymbol{A}_{\text{s}}$构建一个全局相似特征$\boldsymbol{G}_{\text{s}}$,再使$G_{\text{s}}$分别与模态相异性特征$\boldsymbol{T}_{\text{d}}$、$\boldsymbol{V}_{\text{d}}$、$\boldsymbol{A}_{\text{d}}$通过深度交互促进单元(Deep interaction promotion unit,DIPU)实现跨模态深层交互,经过多层交互,增强各自信息,并且通过门控注意力池化单元GAP调整不同模态信息的重要性,过滤模态交互产生的噪声。

2.4.1 跨模态深度交互促进单元

利用两次跨模态注意力与一次自注意力实现全局相似性特征$\boldsymbol{G}^{[i]}_{\text{s}}$与单模态相异性特征$\boldsymbol{X}^{[i]}_{\text{d}}$,$\boldsymbol{X}\in\{T,V,A\}$的信息交互增强。增强的原理是全局相似性特征可以从每个单模态相异性特征中学习不同模态的私有特征,而单模态相异性特征则从全局相似特征中学习模态共性特征,二者相互促进,实现各自信息互补增强,DIPU的架构如图2所示。

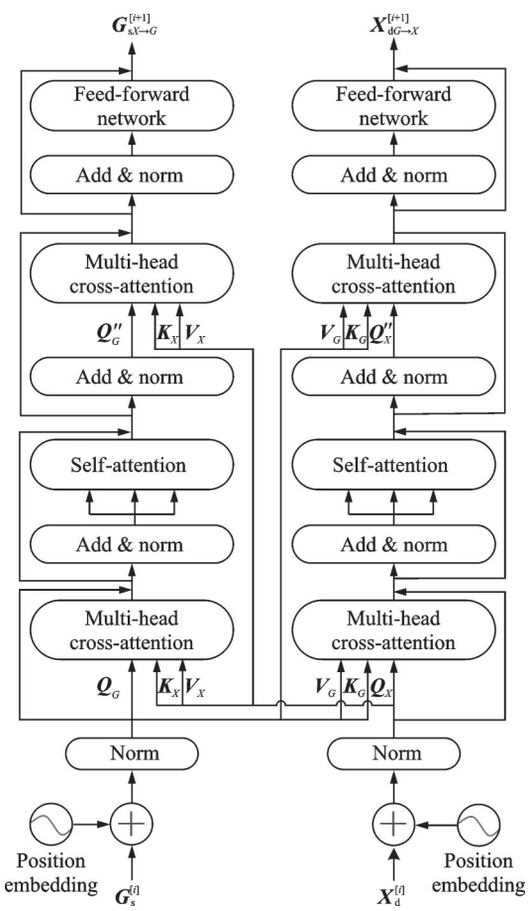


图2 跨模态深度交互促进单元

Fig.2 Cross-modal deep interaction promotion unit

首先,对两个输入进行位置编码,添加位置信息,之后再经过层归一化,计算公式为

$$\boldsymbol{G}_{\text{s}} = \text{LNorm}(\boldsymbol{G}^{[i]}_{\text{s}} + \text{PE}(\boldsymbol{G}^{[i]}_{\text{s}})) \tag{13}$$

$$\boldsymbol{X}_{\text{d}} = \text{LNorm}(\boldsymbol{X}^{[i]}_{\text{d}} + \text{PE}(\boldsymbol{X}^{[i]}_{\text{d}})) \tag{14}$$

式中:$\text{PE}(x)$表示计算位置编码,$\text{LNorm}(x)$表示

层归一化。然后将经过归一化之后的 $\boldsymbol{G}_{\mathrm{s}}\in\mathbf{R}^{T_g\times d}$ 与 $\boldsymbol{X}_{\mathrm{d}}\in\mathbf{R}^{T_x\times d}$ 输入进入第一个多头交叉注意力，输出经过残差连接与归一化，计算公式为

$$\mathrm{CrossAtt}(\boldsymbol{Q}_a,\boldsymbol{K}_b,\boldsymbol{V}_b)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_a\boldsymbol{K}_b^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_b \quad (15)$$

$$\boldsymbol{G}_{\mathrm{s}}'=\mathrm{LNorm}(\mathrm{CrossAtt}(\boldsymbol{Q}_G,\boldsymbol{K}_X,\boldsymbol{V}_X)+\boldsymbol{G}_{\mathrm{s}}) \quad (16)$$

$$\boldsymbol{X}_{\mathrm{d}}'=\mathrm{LNorm}(\mathrm{CrossAtt}(\boldsymbol{Q}_X,\boldsymbol{K}_G,\boldsymbol{V}_G)+\boldsymbol{X}_{\mathrm{d}}) \quad (17)$$

式(15)为多头交叉注意力计算公式；式(16)中，$\boldsymbol{Q}_G=\boldsymbol{G}_{\mathrm{s}}\boldsymbol{\mathcal{W}}_Q$、$\boldsymbol{K}_X=\boldsymbol{X}_{\mathrm{d}}\boldsymbol{\mathcal{W}}_K$、$\boldsymbol{V}_X=\boldsymbol{X}_{\mathrm{d}}\boldsymbol{\mathcal{W}}_V$ 分别作为查询、键与值。其中 $\boldsymbol{\mathcal{W}}_Q$、$\boldsymbol{\mathcal{W}}_K\in\mathbf{R}^{d\times d_k}$，$\boldsymbol{\mathcal{W}}_V\in\mathbf{R}^{d\times d_v}$。式(17)中的查询、键与值分别为 $\boldsymbol{Q}_X=\boldsymbol{X}_{\mathrm{d}}\boldsymbol{\mathcal{W}}_Q$、$\boldsymbol{K}_G=\boldsymbol{G}_{\mathrm{s}}\boldsymbol{\mathcal{W}}_K$、$\boldsymbol{V}_G=\boldsymbol{G}_{\mathrm{s}}\boldsymbol{\mathcal{W}}_V$。

然后对经过初步交叉注意力后的全局相似特征 G_{s}' 与单模态相异特征 $\boldsymbol{X}_d'$ 进行自注意力处理，使得内部整合新吸收到的外部信息，并根据这些新信息调整和精炼自身的内部表示，计算公式为

$$\mathrm{SelfAtt}(\boldsymbol{Q}_a,\boldsymbol{K}_a,\boldsymbol{V}_a)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_a\boldsymbol{K}_a^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_a \quad (18)$$

$$\boldsymbol{G}_{\mathrm{s}}''=\mathrm{LNorm}(\mathrm{SelfAtt}(\boldsymbol{Q}_{G'},\boldsymbol{K}_{G'},\boldsymbol{V}_{G'})+\boldsymbol{G}_{\mathrm{s}}') \quad (19)$$

$$\boldsymbol{X}_{\mathrm{d}}''=\mathrm{LNorm}(\mathrm{SelfAtt}(\boldsymbol{Q}_{X'},\boldsymbol{K}_{X'},\boldsymbol{V}_{X'})+\boldsymbol{X}_{\mathrm{d}}') \quad (20)$$

再将经过初步信息交互与自我信息整合后的 $\boldsymbol{G}_{\mathrm{s}}''$ 与 $\boldsymbol{X}_{\mathrm{d}}''$ 再分别与位置嵌入后的原始输入 $\boldsymbol{X}_{\mathrm{d}}$、$\boldsymbol{G}_{\mathrm{s}}$ 进行第二阶段交叉注意力。重新审视原始数据，更精细地捕捉第一轮交互中可能忽视、或经过内部整合后才显现出重要性的信息，进行更深层次的交互，计算公式为

$$\boldsymbol{G}_{\mathrm{s}}'''=\mathrm{LNorm}(\mathrm{CrossAtt}(\boldsymbol{Q}_G'',\boldsymbol{K}_X,\boldsymbol{V}_X)+\boldsymbol{G}_{\mathrm{s}}'') \quad (21)$$

$$\boldsymbol{X}_{\mathrm{d}}'''=\mathrm{LNorm}(\mathrm{CrossAtt}(\boldsymbol{Q}_X'',\boldsymbol{K}_G,\boldsymbol{V}_G)+\boldsymbol{X}_{\mathrm{d}}'') \quad (22)$$

式中：$\boldsymbol{Q}_G''=\boldsymbol{G}_{\mathrm{s}}''\boldsymbol{\mathcal{W}}_Q$，$\boldsymbol{Q}_X''=\boldsymbol{X}_{\mathrm{d}}''\boldsymbol{\mathcal{W}}_Q$，而 $\boldsymbol{K}_X$、$\boldsymbol{V}_X$ 与 $\boldsymbol{K}_G$、$\boldsymbol{V}_G$ 分别来自于式(16～17)中的原始键与值。

最后将经过二次深度交互增强后的 $\boldsymbol{G}_{\mathrm{s}}'''$ 与 $\boldsymbol{X}_{\mathrm{d}}'''$ 送入前馈神经网络(Feed forward network，FNN)，得到经过深度跨模态信息交互增强后的全局相似特征 $\boldsymbol{G}_{\mathrm{s}X\to G}^{[i+1]}$ 与单模态相异特征 $\boldsymbol{X}_{\mathrm{d}G\to X}^{[i+1]}$，计算公式为

$$\boldsymbol{G}_{\mathrm{s}X\to G}^{[i+1]}=\mathrm{FNN}(\boldsymbol{G}_{\mathrm{s}}''')+\boldsymbol{G}_{\mathrm{s}}''' \quad (23)$$

$$\boldsymbol{X}_{\mathrm{d}G\to X}^{[i+1]}=\mathrm{FNN}(\boldsymbol{X}_{\mathrm{d}}''')+\boldsymbol{X}_{\mathrm{d}}''' \quad (24)$$

式中：$\boldsymbol{X}\in\{\mathrm{T},\mathrm{V},\mathrm{A}\}$，$\boldsymbol{X}_{\mathrm{d}}$ 表示单模态相异性特征，$\boldsymbol{X}\to\boldsymbol{G}$ 表示信息从单模态特有的相异特征流向全局相似特征，从而增强全局相似特征 $\boldsymbol{G}_{\mathrm{s}}$；$\boldsymbol{G}\to\boldsymbol{X}$ 同理。

2.4.2　门控注意力池化

在跨模态交互增强网络的每一层交互中，全局相似特征 $\boldsymbol{G}_{\mathrm{s}}$ 与 3 个单模态相异特征通过跨模态深度交互单元后会产生 3 个被不同模态独有的相异特征增强后的全局相似特征，如 $\boldsymbol{G}_{\mathrm{s}T\to G}$ 是被文本私有的相异性特征增强后的全局相似特征。由于不同模态对于情感分析的重要性不同，往往文本中包含最丰富的情感信息以及比图像与音频更少的噪声，所以在每层交互之后，计算下一层所需的全局相似特征时，使用门控注意力池化，学习来自不同模态增强后特征的重要性，结构如图 3 所示。

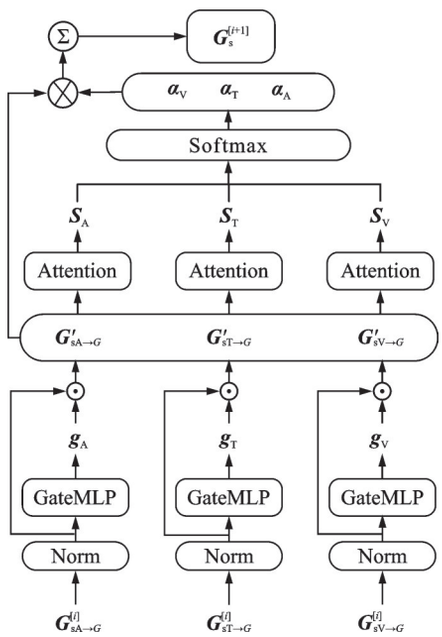


图 3　门控注意力池化

Fig.3　Gated attention pooling

对于第 i 层交互后产生的 3 个全局相似特征 $\boldsymbol{G}_{\mathrm{sT}\to G}^{[i]}$、$\boldsymbol{G}_{\mathrm{sV}\to G}^{[i]}$ 和 $\boldsymbol{G}_{\mathrm{sA}\to G}^{[i]}$，先经过层归一化之后通过独立的门控多层感知机(Gated multilayer perceptron，GateMLP)与 Sigmoid 函数，学习其标量门控值 $g_M\in[0,1]$，$M\in\{\mathrm{T},\mathrm{V},\mathrm{A}\}$。计算公式为

$$g_M=\mathrm{Sigmoid}\Big(\mathrm{GateMLP}(\mathrm{LNorm}(\boldsymbol{G}_{\mathrm{s}M\to G}^{[i]}))\Big) \quad (25)$$

式中门控值表示该全局相似特征的贡献置信度，控制信息通过量。随后再将门控值与归一化后的全局相似特征相乘，公式为

$$\boldsymbol{G}'_{\mathrm{s}M\to G}=\boldsymbol{G}_{\mathrm{s}M\to G}\cdot g_M \quad (26)$$

之后将经过门控处理后的特征分别输入到注意力网络，该网络通过另一个多层感知机计算输入的原始注意力分数 S_M，计算公式为

$$S_{M\in\{\mathrm{T},\mathrm{V},\mathrm{A}\}}=\mathrm{Attention}(\boldsymbol{G}'_{\mathrm{s}M\to G}) \quad (27)$$

再将 3 个注意力分数通过 1 个 Softmax 函数计算最终的注意力权重，最终对经过门控后的特征进行加权求和，有

$$\alpha_{\mathrm{T}},\alpha_{\mathrm{V}},\alpha_{\mathrm{A}}=\mathrm{Softmax}(S_{\mathrm{T}},S_{\mathrm{V}},S_{\mathrm{A}}) \quad (28)$$

$$\boldsymbol{G}_{\mathrm{s}}^{[i+1]}=\sum\nolimits_{M\in\{\mathrm{T},\mathrm{V},\mathrm{A}\}}\boldsymbol{G}'_{\mathrm{s}M\to G}\cdot\alpha_M \quad (29)$$

门控注意力池化首先通过门控独立评估每个信息源的绝对贡献度，然后通过注意力机制学习他

们之间的相对重要性,再加权求和,其能够更灵活地整合来自不同交互路径的特征,并且突出最重要的部分,抑制噪声信息。

2.5 损失计算

对于一个包含n个样本的批次中第i个样本,将经过k层跨模态深层交互后实现信息增强的全局相似特征$\boldsymbol{G}_{\mathrm{s}}^{i[k]}$与单模态相异特征$\boldsymbol{T}_{\mathrm{d}}^{i[k]}$、$\boldsymbol{V}_{\mathrm{d}}^{i[k]}$、$\boldsymbol{A}_{\mathrm{d}}^{i[k]}$拼接融合后送入最终用于预测的多层感知机中,得到样本i的情感预测结果,计算公式为

$$\hat{y}^i=\mathrm{MLP}\left(\mathrm{Concat}\left(\boldsymbol{G}_{\mathrm{s}}^{i[k]},\boldsymbol{T}_{\mathrm{d}}^{i[k]},\boldsymbol{V}_{\mathrm{d}}^{i[k]},\boldsymbol{A}_{\mathrm{d}}^{i[k]}\right)\right) \tag{30}$$

使用平均绝对误差作为预测损失,平均损失计算公式为

$$\mathcal{L}_{\mathrm{pred}}=\frac{1}{n}\sum_{i=1}^{n}|y^i-\hat{y}^i| \tag{31}$$

式中y^i表示样本i的真实情感标签。而最终联合的训练损失为

$$\mathcal{L}=\mathcal{L}_{\mathrm{pred}}+\lambda_{\mathrm{cl}}\mathcal{L}_{\mathrm{cl}} \tag{32}$$

以预测损失为主要部分,式中λ_{cl}参数用于控制对比学习损失的贡献量。

3 实验结果与分析

3.1 数据集

在多模态情感分析主流数据集CMU-MOSI[39]、CMU-MOSEI[40]上训练、测试和验证本文提出的模型,数据集划分如表1所示。

表1 数据集划分

Table 1 Data set split

Dataset	Train	Valid	Test	All
CMU-MOSI	1 284	229	686	2 199
CMU-MOSEI	16 326	1 871	4 659	22 856

CMU-MOSI数据集是一个被广泛使用的英语多模态情感分析数据集,包含文本、视觉和听觉模态信息。该数据集由来自于YouTube平台93个视频组成,视频内容涵盖各个主题,经分割后得到2 199个情感强烈的视频片段,每个视频片段都手动标注了情感强度,范围为−3到+3,将情感划分为7种类别:强消极、消极、弱消极、中性、弱积极、积极和强积极。

CMU-MOSEI数据集是更大规模的情感分析和情感预测数据集,由来自1 000位不同的演讲者的3 228个YouTube视频经过剪辑标注后的23 453个视频片段组成,视频内容涵盖250个不同的主题。同样包含3个模态数据,情感标签在−3到+3之间。

3.2 实验设置

实验环境:Ubuntu 20.04.3 LTS 64位操作系统,Python 3.9.21,Pytorch 2.8.0,CUDA 12.8。硬件环境:NVIDIA RTX 5090。相关参数设置如表2所示。

表2 参数设置

Table 2 Parameter settings

Parameter	MOSI	MOSEI	说明
Batch size	32	16	批次大小
Learning rate	1e−4	1e−4	学习率
k_{cl}	10	50	样本集合构建参数
Temperature τ	0.5	0.5	NT-Xent温度参数
Margin	0.5	0.5	ℓ_{Tri}边界参数
Layers k	3	3	CMDIE交互层数
λ_{Tri}	0.1	0.1	$\mathcal{L}_{\mathrm{cl}}$中$\ell_{\mathrm{Tri}}$权重
λ_{cl}	0.1	0.1	对比损失权重
Optimizer	AdamW	AdamW	优化器选择
Epoch	50	12	迭代次数

3.3 基线模型与评价指标

3.3.1 基线模型

MFN[19]提出一种记忆融合网络,利用LSTM学习模态间特定的交互,再用注意力机制学习跨模态交互,最后使用多模态门控记忆网络融合信息。

TFN[13]使用张量融合网络架构,应用三重笛卡尔积来聚合单模态、双模态及三模态的相互作用。

LMF[14]同样使用张量实现融合,将高维输入张量分解为多个低秩因子矩阵,再利用分解后的矩阵实现融合,提高计算效率。

MulT[29]使用独立的Transformer编码器进行初步特征提取,再通过跨模态注意力机制实现信息交互,能够有效处理异步、非对齐的多模态输入。

MISA[31]将不同模态信号投射到两个不同的特征空间,学习共享与私有表示,实现有效融合。

HyCon[32]基于对比学习方法,实现3个模态混合对比学习框架,关注不同模态之间的交互。

Self-MM[20]使用自监督机制生成单模态标签,并通过多任务学习指导多模态特征融合。

EMT-DLFR[35]使用全局上下文与单模态特征进行交互,提高多模态融合效率,并通过多种重建损失学习缺失的语义信息。

CET-M[36]在跨模态Transformer中加入卷积层,有效建模局部依赖,同时挖掘全局多模态上下文与局部特征之间的内在联系。

ConFEDE[33]提出一个统一的学习框架,利用

不同的编码器将每个模态特征投影到模态不变特征空间与模态特定特征空间，联合对比损失、单模态损失与任务预测损失。

DLF[41]提出了一个以解纠缠语言为重点的多模态表示学习框架，通过解纠缠模块分离共享特征与独立特征信息，并通过语言引导的交叉注意力机制，利用互补的特定模态信息加强语言表征。

TCHFN[42]提出一个以文本为中心的分层融合网络，实现以文本为核心的跨模态细微融合，并通过多模态融合输出门减轻融合表示内的冗余信息。

UniMSE[34]提出一个统一的情感知识贡献框架，利用对比学习与情感识别进行统一分析。该模型是目前的SOTA模型。

3.3.2　评价指标

实验从分类与回归两个角度的5个指标评价模型表现。分类任务评价指标包括二分类精度(Acc-2)、七分类精度(Acc-7)和F_1分数，回归任务评价指标包括皮尔逊相关系数(Corr)和平均绝对误差(Mean absolute error, MAE)。Acc-2与F_1值均包括负/非负两种，对于MAE外的所有指标，数值越高表示性能越好。

3.4　实验结果与分析

本文模型在MOSI、MOSEI数据集上的实验结果与其他基线模型对比如表3所示。表中“/”左边是含0二分类，“/”右边是不含0二分类，粗体表示最佳结果。

表3　实验结果对比

Table 3　Experimental results comparison

Model	MOSI					MOSEI				
	Acc-2	F_1	Acc-7	MAE	Corr	Acc-2	F_1	Acc-7	MAE	Corr
MFN	77.4/—	77.3/	34.1	0.965	0.632	78.94/82.86	79.55/82.85	51.34	0.573	0.718
LMF	—/82.5	—/82.4	33.2	0.917	0.695	80.54/83.48	80.94/83.36	51.59	0.576	0.717
TFN	—/80.8	—/80.7	34.9	0.901	0.698	78.50/81.89	78.96/81.74	51.60	0.573	0.714
MulT	—/83.0	—/82.8	40.0	0.871	0.698	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA	81.8/83.4	81.7/83.6	42.3	0.783	0.776	83.6/85.5	83.8/85.3	52.2	0.555	0.756
HyCon	—/85.2	—/85.1	46.6	0.713	0.790	—/85.4	—/85.6	52.8	0.601	0.776
Self-MM	83.44/85.46	83.36/85.43	46.67	0.708	0.796	83.76/85.15	83.82/84.90	53.87	0.531	0.765
EMT	83.3/85.0	83.2/85.0	47.4	0.705	0.798	83.4/86.0	83.7/86.0	54.5	0.527	0.774
CET-M	84.0/86.0	83.8/85.9	47.7	0.696	0.805	83.4/86.2	83.6/86.1	**54.9**	0.523	0.773
ConFEDE	84.17/85.52	84.13/85.52	42.27	0.742	0.784	81.65/85.82	82.17/85.83	54.86	0.522	0.780
DLF	—/85.06	—/85.04	47.08	0.731	0.781	—/85.42	—/85.27	53.90	0.536	0.764
TCHFN	85.57/86.13	85.41/86.31	44.75	0.748	0.780	84.01/86.27	84.14/86.48	53.19	0.538	0.770
UniMSE	85.85/86.90	85.83/86.42	**48.68**	0.691	0.809	**85.86**/87.50	**85.79**/87.46	54.39	0.523	0.773
FD-CMDIE	**86.75/88.60**	**86.71/88.59**	46.42	**0.671**	**0.817**	85.06/**87.59**	85.28/**87.49**	54.48	**0.501**	**0.801**

当FD-CMDIE与最先进的SOTA模型UniMSE对比。在MOSI数据集上，Acc-2、F_1指标分别了提升0.9%/1.7%与0.88%/2.17%，而在Acc-7指标上下降了2.26%。在Corr指标上提升了0.008，并且MAE降低了0.02。在MOSEI数据集上与SOTA模型相比，不含0二分类任务的Acc-2、F_1指标分别提升了0.09%与0.03%，而且Corr提高了0.028，MAE下降了0.022。

与经典的多模态情感分析模型相比，如MFN、TFN、LMFT MulT，FD-CMDIE在所有指标上都显著优于经典方法。而在与近年来的先进模型对比中，如MISA、HyCon、Self-MM、EMT、CET-M和ConFEDE以及近期发表的DLF和TCHFN，本文模型在几乎所有指标上都有明显提高。

首先与使用相似模态交互方法的EMT和CET-M模型对比，FD-CMDIE在两个数据集上的Acc-2与F_1等指标上显著提升。这两个模型直接使用提取后的特征进行交互，为后续融合引入了噪声，而本文使用经过特征解耦后的相似特征构建全局上下文特征与单模态相异特征交互，通过深度交互网络实现信息增强，并设计门控注意力池化调整不同模态重要性，有效降低了交互过程中产生的噪声。

再与同样关注不同模态相似信息与互补信息的Self-MM和ConFEDE模型对比。FD-CMDIE在MOSI数据集上的Acc-2与F_1值均提升近3个百分点。在MOSEI数据集上的Acc-2与F_1值比Self-MM分别提升了1.3%/2.44%和1.46%/2.59%；比ConFEDE分别提高了3.41%/1.77%和3.11%/1.66%。这两个模型在特征融合部分未能

充分挖掘不同模态之间的关联。而本文通过高质量的特征提取以及特征解耦,提升后续特征交互的质量,多层跨模态交互充分挖掘模态间有效信息,为情感分析提供了有力支撑。

最后,综合考虑所有指标,本文模型基本达到SOTA模型水平。本文模型有效解耦并利用了相似特征与相异特征,实现多模态信息深度交互融合,从而获得了更好的性能。

3.5 超参数分析

在对比学习的样本集合构建中,计算当前样本i与其他样本之间的相似度得分后由超参数k_{cl}决定构建相似样本集合similar^i与不相似样本集合different^i的大小。在此分析该超参数对于实验结果的影响,由于MOSEI数据集规模远大于MOSI数据集,所以在两个数据集上k_{cl}采用不同的递增策略,实验具体结果如表4,5所示。

表4 MOSI上k_{cl}性能分析

Table 4 k_{cl} performance analysis on MOSI dataset

k_{cl}	MOSI	
	Acc-2	F_1
2	79.01/81.10	78.83/81.01
3	81.49/82.93	81.46/82.96
4	83.53/84.91	83.51/84.93
5	83.48/85.01	83.50/85.10
6	84.38/85.37	84.32/85.27
7	85.27/85.93	85.32/85.92
8	85.62/86.52	85.35/86.81
9	86.21/87.13	86.27/87.31
10	**86.75/88.60**	**86.71/88.59**
11	86.44/88.26	86.41/88.27
12	86.34/88.17	86.39/88.25

表5 MOSEI上k_{cl}性能分析

Table 5 k_{cl} performance analysis on MOSEI dataset

k_{cl}	MOSI	
	Acc-2	F_1
10	79.07/85.06	79.93/85.20
20	79.59/85.72	80.41/85.85
30	83.38/85.52	83.22/85.43
40	83.26/86.63	83.92/86.69
50	**85.06/87.59**	**85.28/87.49**
60	83.61/86.71	83.23/86.77
70	83.90/87.12	84.23/87.05
80	82.36/87.01	82.89/87.03

在数据规模较小的MOSI数据集上,随着对比学习样本集合构建参数k_{cl}的递增,模型的Acc-2与F_1指标也逐渐提升,当k_{cl}达到10后基本不再增加。而在数据规模更大的MOSEI数据集上,当k_{cl}在50左右达到最佳效果。在构建样本集合时需要选择数量合适的样本,样本过少则无法有效学习,而样本过多则会被区分度不高的样本影响学习效果。k_{cl}在两个数据集上取值变化分别如图4和图5所示。

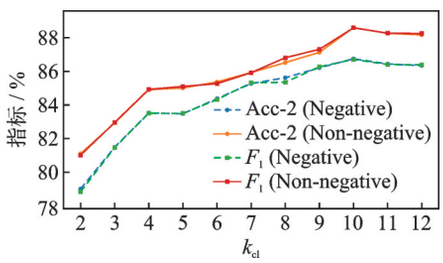


图4 MOSI上k_{cl}性能分析

Fig.4 k_{cl} performance analysis on MOSI

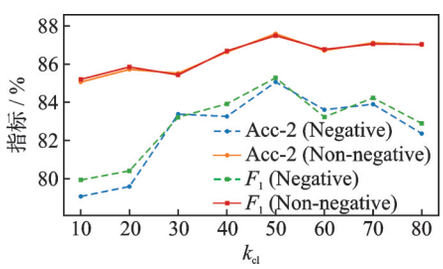


图5 MOSEI上k_{cl}性能分析

Fig.5 k_{cl} performance analysis on MOSEI

此外,CMDIE模块中深度交互网络的层数Layers k的不同取值,也会对模型效果产生影响。具体实验结果如表6所示,表中数据为不含0二分类任务的精确度Acc-2与F_1值。

表6 CMDIE网络层数与性能分析

Table 6 CMDIE network layers and performance analysis

Layers k	MOSI		MOSEI	
	Acc-2	F_1	Acc-2	F_1
1	83.69	83.58	85.53	85.66
2	85.37	85.81	87.11	87.13
3	**88.60**	**88.59**	**87.59**	**87.49**
4	87.04	86.98	86.79	86.73
5	85.37	85.37	85.17	85.35

从实验结果可以看出,不断增加网络深度并不一定能获得模型性能的提升。当网络层数较低时,不能完全地实现模态间信息交互,而当网络过深时会导致模型过于复杂,不能有效学习。

3.6 可视化分析

在多模态情感分析中,不同模态的贡献有所不同,在两个数据集上各个模态的注意力权重可视化分别如图6、7所示。从图中可以看出文本模态在多模态情感分析中始终占据主导地位,说明文本中包含着最丰富的情感相关信息,而视觉与听觉模态

信息作为文本的补充，提供与情感相关的辅助信息，比如视觉模态中说话人的面部表情以及听觉模态中语速和语调等信息，因此视觉与听觉模态在情感分析任务中也都有着一定的贡献。

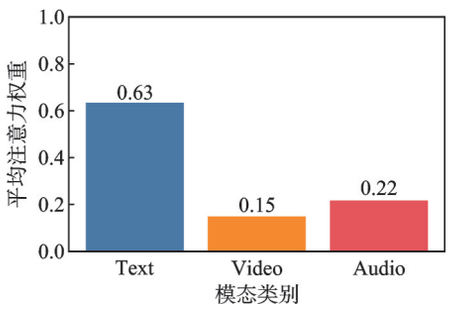


图 6　MOSI上模态平均贡献

Fig.6　Average contribution of modalities on MOSI

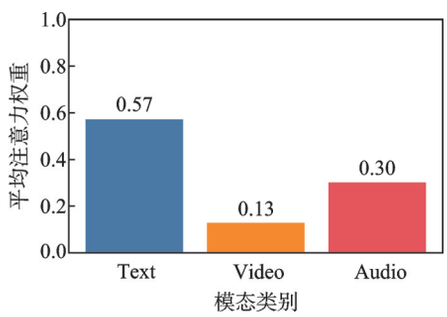


图 7　MOSEI上模态平均贡献

Fig.7　Average contribution of modalities on MOSEI

关于对比学习对于不同模态特征解耦效果影响的可视化，在MOSI数据集上进行实验，结果如图8和图9所示。从有无对比学习时的特征投影可以看出，图8中不使用对比学习时，特征解耦效果不佳，在表示空间中相似特征与相异特征区分不够明显。而在图9中，添加对比学习后3种模态相似性特征更加靠近，而相异性特征远离文本相似特征锚点，并且特征在表示空间中聚集程度更高，说明模型在表示空间中成功区分了模态共享的相似情感信息与各模态私有的相异情感信息。

3.7　消融实验

在CMU-MOSI与CMU-MOSEI两个数据集上对模型进行消融实验，进一步评估模型的各个组成部分对情感分析的作用，消融实验结果如表7所示。

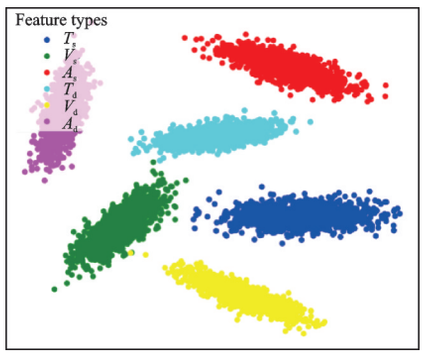


图 8　无对比学习时的特征

Fig.8　Features without contrastive learning

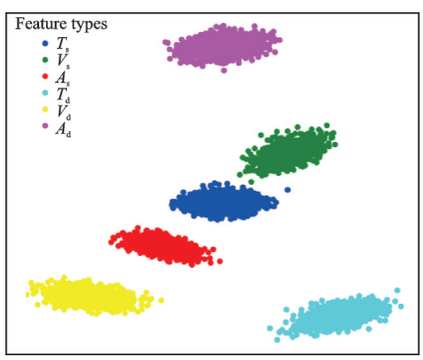


图 9　有对比学习时的特征

Fig.9　Features with contrastive learning

具体操作与结果分析如下：

(1)w/o NeoBERT：改用BERT提取文本特征。在两个数据集上都出现了Acc-2与F_1值的下降，说明NeoBERT能够提取包含更丰富情感语义的高质量文本特征。

(2)w/o CL：不使用对比学习，此时模型在两个数据集上的指标均有下降，说明由于模型不能有效解耦模态特征，从而影响后续多层跨模态交互，最终导致情感预测效果不佳。

(3)w/o CMDIE：不使用多层深度交互增强网络，特征解耦之后直接拼接融合。此时模型的所有指标均出现了较大幅度的下降，这表明CMDIE网络对于模型的重要性，能够通过解耦后的特征实现模态间的深度交互并减少噪声信息，捕捉不同模态内在情感联系，最终实现有效融合。

表 7　消融实验结果

Table 7　Ablation experiment results

Model	MOSI		MOSEI	
	Acc-2	F_1	Acc-2	F_1
w/o NeoBERT	84.12/85.68	84.84/85.77	82.15/85.05	82.76/85.61
w/o CL	83.88/85.72	83.68/85.45	81.64/85.42	82.14/85.32
w/o CMDIE	81.73/83.96	81.94/83.80	79.91/82.68	80.27/82.43
w/o GAP	85.47/86.73	85.52/86.84	82.28/85.77	82.39/85.90
w/o CrossAtt-2	83.36/84.45	83.12/84.29	82.77/83.29	82.35/83.21
FD-CMDIE	**86.75/88.60**	**86.71/88.59**	**85.06/87.59**	**85.28/87.49**

(4)w/o GAP:使用CMDIE,但是不使用门控注意力池化单元,改用平均池化。与完整模型的性能仍有一定差距,表明加入门控注意力池化模块能够提升跨模态交互的效果。原因在于GAP能够通过门控机制与注意力机制自适应调整与重要模态交互后的全局相似特征,降低噪声信息干扰,提升特征融合质量。

(5)w/o CrossAtt-2:消融掉DIPU模块中的第二层交叉注意力网络。结果显示模型性能有明显下降,说明该部分的有效性,原因在于第二层交叉注意力网络能够实现对原始输入信息的再次分析,关注原始输入在第一层交叉注意力中被遗漏的重要信息。

(6)FD-CMDIE:完整模型的情感分析实验结果。消融实验结果的二分类精确度Acc-2与F_1分数可视化如图10和图11所示,从中可直观地看出模型每一部分消融后,与完整模型性能的差距。

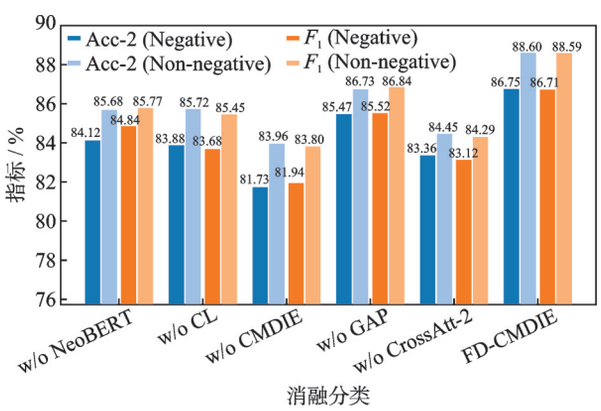


图10 MOSI上消融实验结果

Fig.10 Ablation experiment results on MOSI

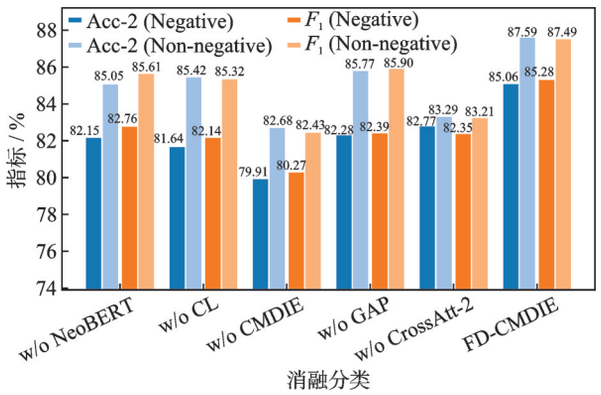


图11 MOSEI上消融实验结果

Fig.11 Ablation experiment results on MOSEI

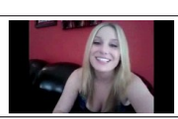
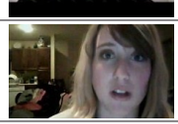
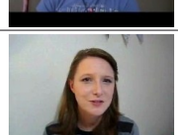
从上述消融实验分析以及图8~10和图11可以得出结论,模型的每个部分均有作用。FD-CMDIE通过NeoBERT提取高质量文本特征;提取高质量文本特征;对比学习提高模型对于模态相似特征与相异特征的分辨能力,提高特征解耦的效果;多层跨模态深度交互增强网络利用解耦之后的特征实现模态间信息深度交互,同时门控注意力池化模块自适应调整来自不同模态信息的重要性,有效地减少噪声,提高了模态融合质量。

3.8 案例分析

为了进一步验证本文模型在情感分析任务中的有效性,从CMU-MOSI数据集中选取样例进行案例分析,结果如表8所示。表中依次展示了每个样例的3个模态数据,包括文本、视频关键帧和音频描述,对比了每个样例的真实情感标签值与模型预测值以及对应的情感极性。

表8 案例分析

Table 8 Case analysis

序号	案例		真实值	预测值	情感极性	
					真实	预测
1	文本:Just the great great movie 图像:露齿微笑 音频:语调高昂		3.0	2.85	积极	积极
2	文本:I do not understand this movie 图像:眉毛皱起,表情严肃 音频:语调下降,逐渐低沉		−2.0	−1.79	消极	消极
3	文本:Anyhow it was really good 图像:神色平静 音频:语调轻微起伏		2.4	2.25	积极	积极
4	文本:But mostly its just Kate Hudson being very annoying 图像:深色平淡,面无表情 音频:语调平缓,语速较慢		−1.6	−1.46	消极	消极
5	文本:And you knew you shouldn't be laughing 图像:眉头上挑,嘴角上扬 音频:语调起伏,语速较快		−1.0	−0.82	消极	消极

案例 1 与案例 2 均是情感倾向极为明显，样例 1（露齿大笑）表达积极情感，而案例 2（眉毛皱起、表情严肃、语调低沉）表达了消极情感，本文模型能够捕获这些明显的情感表达信息并做出准确预测。

对于案例 3 与案例 4 这类平淡地表达积极或消极情感的例子，此时图像与音频不仅无法提供多少有效信息帮助情感分析，还有可能引入噪声。而模型能够通过调整文本模态的权重，以文本信息为主导实现准确预测。

而对于案例 5 这种包含反讽意味的案例，文本表达与其他模态不一致，演讲者表情似乎表达积极情感，而文本表达的情感却是消极的。此时图像和音频模态数据反而可能会干扰真实的情感预测。而在这种情况下本文模型依然能够准确预测真实情感极性，再次证明了模型的有效性。

4　结　　论

本文提出了一种基于特征解耦与跨模态深度交互增强的多模态情感分析模型。引入新一代双向编码器 NeoBERT 提取包含丰富语义的文本特征，使用堆叠 LSTM 捕捉视觉与听觉模态时序特征，增强单模态特征的表达能力。再通过对比学习与表示学习实现特征解耦，获取模态相似特征与模态相异特征，有效减少后续跨模态交互时存在的相似性冗余。构建全局相似特征分别与 3 个模态相异特征在跨模态交互增强网络实现深度信息交互，同时通过门控注意力池化模块自适应调整来自各模态信息的权重，减少引入的噪声信息。最后联合预测损失与对比损失进行优化。本文模型在 CMU-MOSI 与 CMU-MOSEI 数据集上与多个基线模型对比，证明了模型的良好性能。

多模态情感分析中，视觉模态的图像数据以及听觉模态的音频数据中容易存在噪声信息，影响提取特征的表示能力。此外，虽然不同模态私有的相异性特征能够相互补充，但是也存在部分模态信息表达的情感倾向是相反的，例如包含讽刺情感的数据。此时就需要综合考虑所有模态以及上下文语境分析真实情感。因此，在未来的工作中，将关注如何在特征提取阶段减少噪声信息，重点研究不同模态特征信息之间更细粒度的相互作用，以提高多模态情感分析模型的性能。

参考文献：

[1] GANDHI A, ADHVARYU K, PORIA S, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information Fusion, 2023, 91: 424-444.

[2] HAN W, CHEN H, PORIA S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 9180-9192.

[3] PORIA S, CAMBRIA E, BAJPAI R, et al. A review of affective computing: From unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125.

[4] 郭续，买日旦·吾守尔，古兰拜尔·吐尔洪. 基于多模态融合的情感分析算法研究综述[J]. 计算机工程与应用, 2024, 60(2): 1-18.
GUO Xu, GUSHUER B, TULHUN G. A review of research on sentiment analysis algorithms based on multimodal fusion[J]. Computer Engineering and Applications, 2024, 60(2): 1-18.

[5] FU Z, LIU F, XU Q, et al. NhfNet: A non-homogeneous fusion network for multimodal sentiment analysis[C]//Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). [S.l.]: IEEE, 2022: 1-6.

[6] ZHU L, ZHU Z, ZHANG C, et al. Multimodal sentiment analysis based on fusion methods: A survey[J]. Information Fusion, 2023, 95: 306-325.

[7] GKOUMAS D, LI Q, LIOMA C, et al. What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis[J]. Information Fusion, 2021, 66: 184-197.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[9] LE BRETON L, FOURNIER Q, EL MEZOUAR M, et al. NeoBERT: A next-generation BERT[EB/OL]. (2025-02-26)[2026-05-24]. https://arxiv.org/abs/2502.19587.

[10] ZHANG Z, LI X, BAI H, et al. GAL: A global aspect local extraction mechanism for aspect-based sentiment classification[J]. Information Sciences, 2025, 717: 122299.

[11] 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. New York: Association for Computing Machinery (ACM), 2011: 169-176.

[12] PORIA S, CAMBRIA E, GELBUKH A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal senti-

ment analysis[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 2539-2544.

[13] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 1103-1114.

[14] LIU Z, SHEN Y. Efficient low-rank multimodal fusion with modality-specific factors[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018.

[15] LIANG P P, LIU Z, TSAI Y H H, et al. Learning Representations from imperfect time series data via tensor rank regularization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 1569-1576.

[16] LIANG P P, LIU Z, ZADEH A A B, et al. Multimodal language analysis with recurrent multistage fusion[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 150-161.

[17] MAJUMDER N, HAZARIKA D, GELBUKH A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-based systems, 2018, 161: 124-133.

[18] KUMAR A, IRSOY O, ONDRUSKA P, et al. Ask me anything: Dynamic memory networks for natural language processing[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2016: 1378-1387.

[19] ZADEH A, LIANG P P, MAZUMDER N, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI Conference On Artificial Intelligence. Palo Alto, California, USA: AAAI Press, 2018: 5634-5641.

[20] YU W, XU H, YUAN Z, 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. Palo Alto, California, USA: AAAI Press, 2021: 10790-10797.

[21] HURST A, LERER A, GOUCHER A P, et al. GPT-4o system card[EB/OL]. (2023-12-19)[2026-05-24]. https://arxiv.org/abs/2312.11805.

[22] TEAM G, ANIL R, BORGEAUD S, et al. Gemini: A family of highly capable multimodal models[J]. arXiv preprint arXiv:2312.11805, 2023.

[23] LIU H, LI C, WU Q, et al. Visual instruction tuning [J]. Advances in Neural Information Processing Systems, 2023, 36: 34892-34916.

[24] MAAZ M, RASHEED H, KHAN S, et al. Video-chatgpt: Towards detailed video understanding via large vision and language models[EB/OL]. (2023-06-08)[2026-05-24]. https://arxiv.org/abs/2306.05424.

[25] LIU Z, YANG K, XIE Q, et al. EmoLLMs: A series of emotional large language models and annotation tools for comprehensive affective analysis[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.[S.l.]: ACM, 2024: 5487-5496.

[26] HAN Z, HU M, BAI Y, et al. DEQA: Descriptions enhanced question-answering framework for multimodal aspect-based sentiment analysis[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2025, 39(22): 23987-23995.

[27] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[28] RAHMAN W, HASAN M K, LEE S, et al. Integrating multimodal information in large pretrained transformers[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 2359-2369.

[29] TSAI Y H H, BAI S, LIANG P P, et al. Multimodal Transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 6558-6569.

[30] 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: Association for Computational Linguistics, 2018: 3454-3466.

[31] 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 Multime-

dia. [S.l.]: ACM, 2020: 1122-1131.

[32] 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.

[33] 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 Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 7617-7630.

[34] HU G, LIN T E, ZHAO Y, et al. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 7837-7851.

[35] SUN L, LIAN Z, LIU B, et al. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2023, 15(1): 309-325.

[36] SUN B, JIA L, CUI Y, et al. Conv-enhanced transformer and robust optimization network for robust multimodal sentiment analysis[J]. Neurocomputing, 2025, 634: 129842.

[37] CHENG H, YANG Z, ZHANG X, et al. Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion [J]. IEEE Transactions on Affective Computing, 2023, 14(4): 3149-3163.

[38] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of International conference on machine learning. [S.l.]: PMLR, 2020: 1597-1607.

[39] ZADEH A, ZELLERS R, PINCUS E, et al. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages[J]. IEEE Intelligent Systems, 2016, 31(6): 82-88.

[40] ZADEH A A B, LIANG P P, PORIA S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018: 2236-2246.

[41] WANG P, ZHOU Q, WU Y, et al. DLF: Disentangled-language-focused multimodal sentiment analysis [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2025: 21180-21188.

[42] HOU J, OMAR N, TIUN S, et al. TCHFN: Multimodal sentiment analysis based on text-centric hierarchical fusion network[J]. Knowledge-Based Systems, 2024, 300: 112220.

（编辑：刘彦东）