Computer Science and Application 计算机科学与应用, 2025, 15(5), 615-623
Published Online May 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.155134

Hans汉斯

基于分层距离感知对比学习的多模态情绪分析

吕欣阳*，金媛媛，韩　旭，杨　明

沈阳城市建设学院信息与控制工程学院，辽宁 沈阳

收稿日期：2025年4月14日；录用日期：2025年5月15日；发布日期：2025年5月23日

摘　要

多模态情感分析(multimodal sentiment analysis, MSA)利用视觉、文本和音频等模态数据来提升情感分析的准确性。尽管多模态信息能够提供更丰富的语境，但如何有效地处理异构模态数据之间的交互与融合仍然是一个重要挑战。为了解决这一问题，本文提出了一种基于分层距离感知对比学习(hierarchical distance-aware contrastive learning, HDACL)的多模态情感分析方法。具体而言，HDACL通过引入跨模态注意力机制，实现了不同模态数据之间的充分交互。与此同时，我们设计了一种基于情感强度距离差异引导的对比学习策略，进一步增强了多模态数据的一致性对齐。在CMU-MOSI多模态情感分析数据集上进行验证，实验结果表明，HDACL方法在Acc-2和Acc-7指标上分别取得了0.7%和0.8%的性能提升。

关键词

多模态情感分析，跨模态注意力机制，对比学习

Multimodal Sentiment Analysis Based on Hierarchical Distance-Aware Contrastive Learning

Xinyang Lyu*, Yuanyuan Jin, Xu Han, Ming Yang

Department of Information and Control Engineering, Shenyang Urban Construction University, Shenyang Liaoning

Received: Apr. 14th, 2025; accepted: May 15th, 2025; published: May 23rd, 2025

Abstract

Multimodal sentiment analysis (MSA) utilizes visual, textual, and audio data to improve the accuracy of sentiment analysis. Although multimodal information can provide richer context, how to

*通讯作者。

文章引用: 吕欣阳, 金媛媛, 韩旭, 杨明. 基于分层距离感知对比学习的多模态情绪分析[J]. 计算机科学与应用, 2025, 15(5): 615-623. DOI: 10.12677/csa.2025.155134

effectively handle the interaction and fusion across heterogeneous multimodal data remains an important challenge. To this end, this paper proposes a multimodal sentiment analysis method based on hierarchical distance-aware contrastive learning (HDACL). Specifically, HDACL achieves full interaction across different modal data by introducing a cross-modal attention mechanism. Meanwhile we design a contrastive learning strategy guided by the difference in sentiment intensity distance to further enhance the consistency alignment of multimodal data. The method was validated on the CMU-MOSI multimodal sentiment analysis dataset. Experimental results show that the HDACL method achieved 0.7% and 0.8% performance improvements on the Acc-2 and Acc-7 indicators, respectively.

Keywords

Multimodal Sentiment Analysis, Cross-Modal Attention Mechanism, Contrastive Learning

1. 引言

多模态情感分析(multimodal sentiment analysis, MSA)旨在通过整合来自不同模态的信息(如视觉、文本和音频)来全面分析人类情感。与基于单模态的情绪分析不同，MSA 不仅利用单模态数据，还探索模态间的相互关系，其提供了更丰富，更准确的情感识别[1]。现如今，MSA 已在市场决策[2]、人机交互[3]以及社交媒体[4]等领域广为涉猎。

早期的研究主要集中在单模态情感分析上，通常依赖于单一的数据源，如视觉、文本和音频，以提取情感信息[5]。然而，单模态方法在应对复杂情感表达时无法全面捕捉和理解情感的细微变化。尽管多模态数据的组合提高了情感预测的准确性[6]，但模态间的异质性问题增加了多模态融合的难度。为了解决上述挑战，前人的共工作通常使用两种策略：1) 基于多模态注意力机制的 MSA 方法[7] [8]通过动态建模模态间各元素的相关性，实现模态间的有效交互，从而缩小非对齐多模态数据之间的鸿沟。2) 基于多模态一致性学习的 MSA 方法[9] [10]能够有效地弥合不同模态之间的差异，促进模态间信息的协同融合。在模态融合过程中，仅依靠跨模态交互难以充分实现模态间的信息融合。跨模态一致性学习可以有效实现跨模态对齐，减少多模态融合难度。然而，现存的跨模态一致性学习策略缺乏情绪信息的有效约束，导致多模态融合特征中的语义信息有所下降。

鉴于上述挑战，我们提出了一种基于分层距离感知对比学习(hierarchical distance-aware contrastive learning, HDACL)的多模态情感分析方法，旨在解决异质多模态数据融合过程中面临的模态非对齐问题。本文的主要贡献如下。

首先，我们提出了一种多模态交互注意力(cross-modal interaction attention, CIA)机制，以实现多模态数据间的交互和融合。CIA 以文本模态为中心，充分利用文本信息的主导作用，确保各模态在情感分析中的协同工作。

其次，我们进一步设计了一种情绪距离感知对比学习(sentiment distance-aware contrastive learning, SDACL)方法，其可以在情感强度差异指导下实现多模态正负样本对的选择。这为对比学习引入了细粒度的语义信息，从而增强了对齐后特征的情感表达能力和语义一致性。

最后，我们在多模态情感分析数据集 CMU-MOSI 上进行实验。结果表明，相较于前人的方法，HDACL

在 Acc-2、F1 分数、Acc-7 以及平均绝对误差(mean absolute error, MAE)指标上取得了显著提升。

2. 相关工作

2.1. 单模态情感分析

早期的研究主要聚焦于单一模态的情感分析，其中构建情感词典是文本情感分析的常用方法。与图像和音频模态相比，文本模态提供了更丰富的语义层次信息。例如，Taboada 等[11]构建了具有极性和强度的单词词典，名为语义取向计算器(semantic orientation calculator, SO-CAL)，以捕捉文本对情绪信息的关系。吴杰胜等[12]通过添加程度副词、否定词等丰富了情感词典，提高了中文微博情感分析的准确率。随着机器学习的发展，支持向量机[13]、循环神经网络[14]以及 Transformer [15]等方法通过有监督学习的策略进一步改善了基于文本情感分析的性能。基于视觉的情感分析通常通过提取面部动作单元或微表情特征，计算出能够识别出个体的情感状态[16]。例如，Li 等[17]利用卷积神经网络(Convolutional Neural Network, CNN)来捕获面部表情特征，以实现情感分析。另一方面，基于音频的情感分析则通过分析语音的声学特征，如基频、语音强度、语速以及音调等，来识别情感状态。特征提取方法如梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)常用于捕捉音频信号中的情感信息[18]。随着深度神经网络的引入，特别是 CNN 和 Transformer，这些模型能够更好地处理音频中的时序信息和复杂的声学特征，进而提升情感识别的效果[19]。由于情感表达方式复杂且多样，因此单一模态往往难以充分捕捉情感信息。

2.2. 多模态情感分析

近年来，随着多模态情感分析的研究进展，学者们开始探索通过结合多种模态(如文本、音频和视觉)来提高情感识别的准确性。其中，多模态融合是提升情感识别精度的核心所在。常见的融合方法包括张量融合[20]和注意力机制融合[21] [22]。然而，由于不同模态之间存在异质性导致这些方法难以弥合模态之间的差距，进而导致多模态融合不充分。近期的研究[7] [8]也尝试引入多模态 Transformer 来实现跨模态交互并减少模态间的异质性问题。例如，Tsai 等[7]提出了 MulT，其利用定向成对跨模态注意力机制来关注不同模态之间的密集相互作用。尽管如此，这些方法缺乏跨模态一致性建模，导致信息融合不完善。此外，一些基于跨模态一致性学习的方法[9] [10]，通过提高模态间的对齐程度，实现了增强了多模态的融合效果。例如，Han 等[9]通过最大化模态间的互信息，来提升跨模态一致性程度，进而提升多模态联合表示的质量。然而，这些方法往往缺乏细粒度的情感信息引导，尤其是在复杂情感场景下，单纯的跨模态一致性可能无法有效捕捉情感的多维特征，导致情感分析的鲁棒性不足。因此，本文提出了基于 HDACL 的 MSA 方法，其通过具有分层跨模态一致性学习的多模态注意力机制来实现多模态信息充分交互。在多模态交互过程中，HDACL 可以在情绪距离感知对比学习约束下实现多模态信息的细粒度对齐。

3. 方法

3.1. 总体框架

所提出的 HDACL 框架的结构图如图 1 所示。针对不同的输入模态，采用了不同的编码方式：文本模态使用 BERT [23]作为文本编码器以提取编码特征 $Z_t \in \mathbb{R}^{T_t \times D_t}$；而视觉和音频使用预训练工具[22]进行初步特征提取并使用 Transformer 以分别获得其对应的编码特征 $Z_a \in \mathbb{R}^{T_a \times D_a}$ 和 $Z_v \in \mathbb{R}^{T_v \times D_v}$。随后，音频编码特征和视觉编码特征作为非文本特征，与文本特征通过距离感知对比学习进行融合，以实现低级特征在情感信息约束下的多模态对齐。进一步地，跨模态 Transformer 被应用于以文本模态为中心的跨模态交互，生成两种高级跨模态特征 $h_{a\to t}$ 和 $h_{v\to t}$。接着，通过距离感知对比学习以实现高级特征的对齐。最

后，进行多模态特征融合并通过多层感知机进行情感分析。

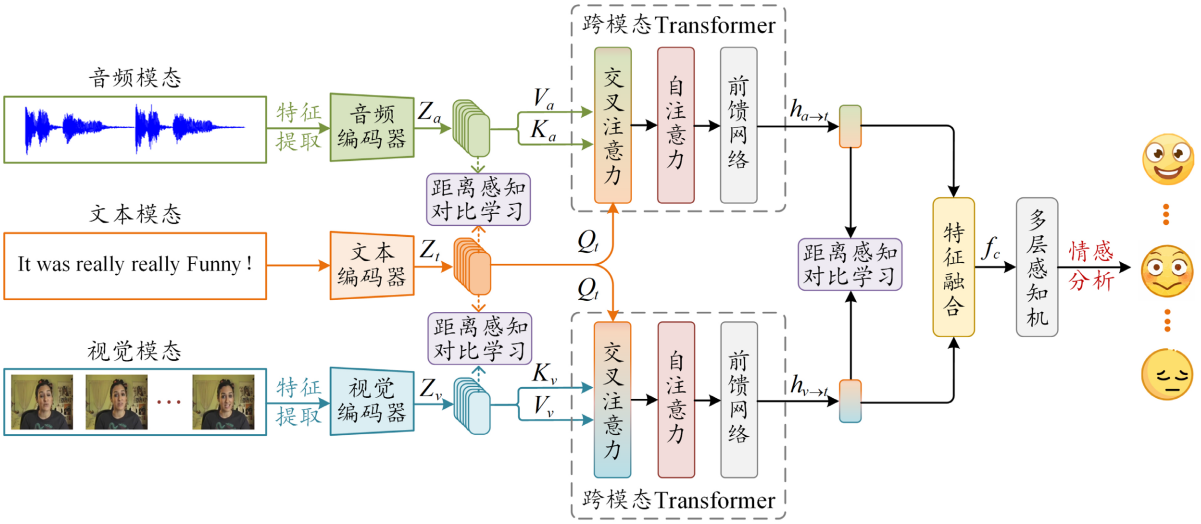

Figure 1. Diagram of the proposed HDACL structure
图 1. 所提出的 HDACL 结构图

3.2. 跨模态 Transformer

如图 1 所示，所提出的跨模态 Transformer 主要包括三个子网络：交叉注意力机制、自注意力机制以及前馈网络。其中，每个子网络还包括恒等映射和层归一化，其可以表示为 Add&Norm$(\bullet)$。我们利用主模态，即文本模态，隐式地指导模态间学习。具体来说，我们使用交叉注意力机制执行文本模态特征和非文本模态特征间的交互与融合。值得一提的是，文本模态在此配置中充当查询 Q_t，将其定位为目标模态，而非文本模态信息被视为键 $K_{\bar{m}}$ 和值 $V_{\bar{m}}$，$\bar{m}\in\{a,v\}$。因此，文本模态与非文本模态的跨模态注意力机制操作可以公式化为：

$$\begin{aligned} H_{\bar{m}\to t} &= \text{Add\&Norm}\left(\text{Atten}\left(Q_t,K_{\bar{m}},V_{\bar{m}}\right)\right) \\ &= \text{Add\&Norm}\left(\text{Softmax}\left(\frac{Q_t K_{\bar{m}}^T}{\sqrt{d_{\bar{m}}}}\right)V_{\bar{m}}\right), \end{aligned} \tag{1}$$

其中，$H_{\bar{m}\to t}$ 表示跨模态特征，$d_{\bar{m}}$ 为非文本模态特征的维度。随后，我们通过自注意力机制来对跨模态特征进一步学习，以实现序列内的全局依赖性建模，其可以表示为：

$$\begin{aligned} H'_{\bar{m}\to t} &= \text{Add\&Norm}\left(\text{Atten}(Q_{\bar{m}\to t},K_{\bar{m}\to t},V_{\bar{m}\to t})\right) \\ &= \text{Add\&Norm}\left(\text{Softmax}\left(\frac{Q_{\bar{m}\to t}K_{\bar{m}\to t}^T}{\sqrt{d_{\bar{m}\to t}}}\right)V_{\bar{m}\to t}\right), \end{aligned} \tag{2}$$

其中，$H'_{\bar{m}\to t}$ 表示全局建模后的跨模态特征，$d_{\bar{m}\to t}$ 为跨模态特征的维度。最后，通过前馈网络进行非线性变换，以捕获更复杂的表示，进而获得最终的跨模态特征 $H''_{\bar{m}\to t}$，可以表示为：

$$H''_{\bar{m}\to t} = \text{Add\&Norm}\left(\text{Relu}\left(W_2\left(W_1 H'_{\bar{m}\to t}+b_1\right)+b_2\right)\right), \tag{3}$$

其中，W_1，W_2，b_1，以及 b_2 表示前馈网络的权重和偏执。至此，跨模态 Transformer 构建完毕，其可以有效地在可变的时间步长内捕获多模态序列之间的相互作用。

3.3. 情绪距离感知对比学习

为了在对齐过程中有效利用情感强度连续变化的细粒度信息，我们设计了情绪距离感知对比学习，其包括低级层次对比和高级层次对比。首先，我们通过欧氏距离来计算不同样本间的情绪强度差异，以样本 i 和 j 为例，可以公式化为：

$$D(i,j)=\left|y_i-y_j\right|, j\neq i, \tag{4}$$

其中，y 表示情绪强度，也是 MSA 任务的真实样本标签。然后，我们设置了阈值 λ =0.5 确定积极样本对和消极样本对。当 $D(i,j)\leq\lambda$ 时，样本 i 和 j 为积极样本对；当 $D(i,j)>\lambda$ 时，样本 i 和 j 为消极样本对；之后，我们根据情感强度差异来为不同的样本对赋予不同的权重。具体来说，我们利用使用非线性激活函数设计了权重函数，如下：

$$w_{i,j}=\begin{cases}\mu\left|\mathrm{Tanh}\left(D(i,j)-2\lambda\right)\right|, D(i,j)\leq\lambda\\ \mu\left|\mathrm{Tanh}\left(D(i,j)\right)\right|, D(i,j)>\lambda\end{cases} \tag{5}$$

其中，$w_{i,j}$ 表示样本 i 和 j 的权重，$\mu=1.5$ 为幅值，τ 为温度系数。公式(5)的权重分布图如图 2 所示。当 $D(i,j)\leq\lambda$ 时，随着距离的增大，$w_{i,j}$ 逐渐下降。此时，情感强度差异越大，积极对拉近的强度应该越小。当 $D(i,j)>\lambda$ 时，随着距离的增大，$w_{i,j}$ 逐渐上升。此时，情感强度差异越大，消极对推远的强度应该越大。至此，情绪距离感知对比学习可以表示为：

$$l_{\mathrm{cl}}=-\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\sum_{j=1}^{2B}1_{D(i,j)\leq\lambda}*e^{w_{i,j}*\frac{\mathrm{sim}(i,j)}{\tau}}}{\sum_{j=1}^{2B}1_{D(i,j)>\lambda}*e^{w_{i,j}*\frac{\mathrm{sim}(i,j)}{\tau}}}, i\neq j, \tag{6}$$

其中，B 表示批次数量，$\mathrm{sim}(i,j)$ 表示样本 i 和 j 的余弦相似度，1_x 表示指示器(如果满足 x，则为 1，否则为 0。)对于低级层次对比，i 或 j 用 z_m （Z_m 的[CLS]特征），$m\in\{t,a,v\}$ 表示；对于高级层次对比，i 或 j 用 h_n （$H^{''}_{\overline{m}\to t}$ 的[CLS]特征），$n\in\{a\to t, v\to t\}$ 表示。

3.4. 融合层和预测端

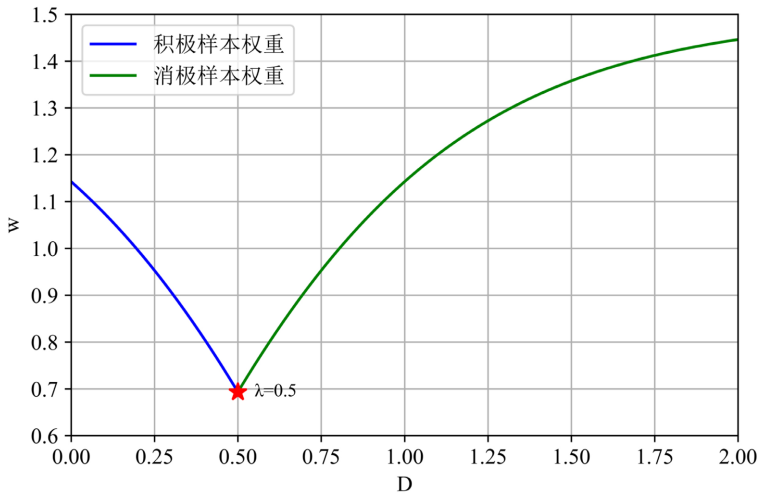

Figure 2. Diagram of the weight change rule
图 2. 权重变化规律图

在经历过跨模态交互和一致性学习后，我们通过拼接操作来融合[CLS]特征 $h_{a\to t}$ 和 $h_{v\to t}$。然后通过全连接层来实现最终情感强度分数的预测，其目标函数使用 MAE，如下：

$$l_{\mathrm{MAE}} = \left| y - \hat{y} \right| \tag{7}$$

式中，y 为真实情感强度标签，$\hat{y}$ 为全连接层预测出来的情感标签。结合距离感知对比学习，我们的总体目标函数可以表示如下：

$$l = l_{\mathrm{MAE}} + \alpha\left(l_{\mathrm{cl}}^{1} + l_{\mathrm{cl}}^{2} + l_{\mathrm{cl}}^{3}\right) \tag{8}$$

式中，α 表示损失函数的权重系数，l_{cl}^{1} 和 l_{cl}^{2} 表示低级层次对比学习，l_{cl}^{3} 表示高级层次对比学习。

4. 实验结果及分析

4.1. 实验配置

所有实验在 python 语言 3.8 版本和 pytorch 深度学习框架 1.10 版本下执行。此外，所有模型均在一个 NVIDIA RTX 3090 GPU 上进行。在训练过程中，我们使用 Adamw 优化器，学习率为 0.001。为了避免过拟合，我们实现了一个早期停止策略，当连续 8 个连续 epoch 的 MAE 不下降即认为训练完毕。

4.2. 数据集和评价指标

我们使用了 CMU-MOSI [24]数据集来评估 HDACL 的性能。CMU-MOSI 数据集是评估 MSA 性能的最常用基准之一，它是从 YouTube 上的视频博客中收集而来的，包含从 93 个视频中分割出来的 2199 个视频片段。CMU-MOSI 数据集的每个片段都是在[−3, 3]的范围内以正/负分别区分积极情绪和消极情绪。在实验中，我们使用了二类精度(Acc-2)、F1-Score、七类精度(Acc-7)以及 MAE 作为性能评价指标。

4.3. 对比实验结果分析

为了验证 HDACL 的有效性，我们将其与 MSA 任务中以前的模型，即 MULT [7]、MMIM [9]、TFN [20]以及 MISA [22]进行了比较。从表 1 中我们观察到所提出的 HDACL 的 Acc-2、F1-Score、Acc-7 以及 MAE 分别达到了 85.8%、85.7%、46.8%以及 0.71，在各个指标上的性能均优于前人的方法。这归因于以文本模态为中心的跨模态信息交互和分层情绪距离感知对比学习对异构模态数据的有效融合，能够更好地捕捉情感信息中的细微差异，并促进多模态情感分析任务中的信息传递和表达，从而提高模型的准确性。

Table 1. Comparative experimental results
表 1. 对比实验结果

模型	Acc-2	F1-Score	Acc-7	MAE
MULT	83.6	83.6	45.5	0.83
MMIM	84.8	84.8	46.0	0.77
TFN	84.3	84.4	45.2	0.81
MISA	85.1	84.9	44.8	0.78
HDACL	**85.8**	**85.7**	**46.8**	**0.71**

4.4. 特征可视化分析

如图 3 所示，我们通过对输入到预测端前的特征进行 T-SNE 可视化来分析情绪距离感知对比学习对模型性能的影响。从图 3(a)中可以看出，在未应用情绪距离感知对比学习的情况下，所提取的特征在不

同类别之间的判别性较差，且同一类别的特征分布较为分散，显示出较低的类内紧密度和较高的类间重叠。相比之下，图 3(b)展示了在使用对比学习后，模型所提取的特征显著改善了判别性，不同类别间的特征具有明显的差异性，且各类别的特征分布更加集中。这表明情绪距离感知对比学习有效地提升了特征的判别能力，减少了类别之间的混淆，从而优化了模型的性能。

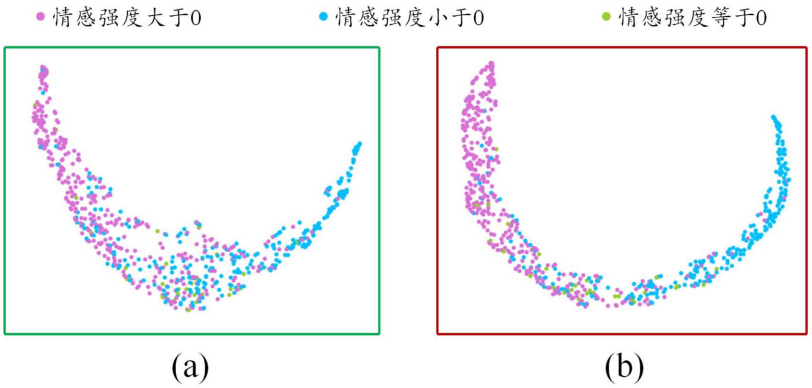

Figure 3. Feature visualization results, (a) is the method without contrastive learning, (b) is the HDACL
图 3. 特征可视化结果，(a) 是未使用对比学习的方法，(b) 是所提出的 HDACL

4.5. 重要超参数分析

如图 4 所示，我们进一步对两个重要超参数，阈值 λ 和权重系数 α 的敏感性进行了分析。具体而言，我们在{0.1, 0.2, ⋯, 0.9}的范围内探索了这两个超参数的影响。从图 4(a)中可以看出，当阈值 λ 设定为 0.5 时，模型的 MAE 最低，性能表现最佳。若阈值过小，模型难以有效拉近积极样本对之间的距离，而过大的阈值则可能导致忽视消极样本对的推远效果，进而影响模型的训练效果和性能。图 4(b)显示了权重系数对模型性能的影响。我们观察到，当权重系数 α 设为 0.3 时，模型的表现最佳。若权重系数过小，无法充分实现跨模态对齐，从而影响模型的整体性能；而过大的权重系数则可能引入干扰，导致主任务预测损失的优化不充分。因此，合理选择这两个超参数对于模型性能的提升至关重要。

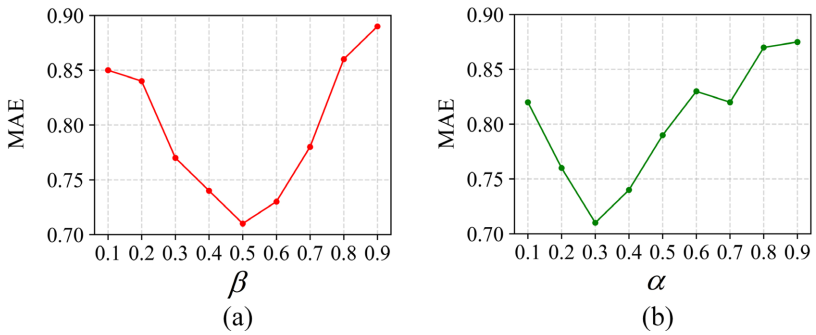

Figure 4. Hyperparameter sensitivity analysis results, (a) is the threshold λ, (b) is the weight coefficient α
图 4. 超参数敏感性分析结果，(a) 是阈值 λ，(b) 是权重系数 α

5. 结论

本文提出了一种基于分层距离感知对比学习(HDACL)的 MSA 方法，旨在解决异构模态数据之间的交互与融合问题。通过引入跨模态注意力机制(CIA)和情绪距离感知对比学习(SDACL)策略，HDACL 能

够有效地增强多模态数据的一致性对齐，提升情感分析的准确性和语义一致性。实验结果表明，HDACL在 CMU-MOSI 数据集上的 Acc-2、F1-Score、Acc-7 以及 MAE 分别达到了 85.8%、85.7%、46.8%以及0.71，相比传统方法取得了显著的性能提升。

基金项目

沈阳市科技创新智库决策咨询课题《数字经济下推进沈阳制造业工业互联网平台建设与发展对策研究》(SYZK2022ZX087)，沈阳城市建设学院校级科研基金项目《基于机器视觉技术的机器人分拣系统的研究》(XKJ202306)。

参考文献

[1] Islam, M.S., Kabir, M.N., Ghani, N.A., Zamli, K.Z., Zulkifli, N.S.A., Rahman, M.M., *et al.* (2024) Challenges and Future in Deep Learning for Sentiment Analysis: A Comprehensive Review and a Proposed Novel Hybrid Approach. *Artificial Intelligence Review*, **57**, Article No. 62. https://doi.org/10.1007/s10462-023-10651-9

[2] Poria, S., Hazarika, D., Majumder, N. and Mihalcea, R. (2023) Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, **14**, 108-132. https://doi.org/10.1109/taffc.2020.3038167

[3] Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion*, **37**, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

[4] Somandepalli, K., Guha, T., Martinez, V.R., Kumar, N., Adam, H. and Narayanan, S. (2021) Computational Media Intelligence: Human-Centered Machine Analysis of Media. *Proceedings of the IEEE*, **109**, 891-910. https://doi.org/10.1109/jproc.2020.3047978

[5] 彭李湘松, 张著洪. 基于三角形特征融合与感知注意力的方面级情感分析[J]. 计算机工程, 2025: 1-10. https://doi.org/10.19678/j.issn.1000-3428.0070397, 2025-03-25.

[6] Fan, C., Zhu, K., Tao, J., Yi, G., Xue, J. and Lv, Z. (2025) Multi-Level Contrastive Learning: Hierarchical Alleviation of Heterogeneity in Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*, **16**, 207-222. https://doi.org/10.1109/taffc.2024.3423671

[7] Tsai, Y.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L. and Salakhutdinov, R. (2019) Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 6558-6569. https://doi.org/10.18653/v1/p19-1656

[8] Li, Y., Wang, Y. and Cui, Z. (2023) Decoupled Multimodal Distilling for Emotion Recognition. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 6631-6640. https://doi.org/10.1109/cvpr52729.2023.00641

[9] Han, W., Chen, H. and Poria, S. (2021) Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, 7-11 November 2021, 9180-9192. https://doi.org/10.18653/v1/2021.emnlp-main.723

[10] Wang, D., Liu, S., Wang, Q., Tian, Y., He, L. and Gao, X. (2023) Cross-Modal Enhancement Network for Multimodal Sentiment Analysis. *IEEE Transactions on Multimedia*, **25**, 4909-4921. https://doi.org/10.1109/tmm.2022.3183830

[11] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, **37**, 267-307. https://doi.org/10.1162/coli_a_00049

[12] 吴杰胜, 陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件, 2019, 36(9): 93-99.

[13] Chang, C. and Lin, C. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 1-27. https://doi.org/10.1145/1961189.1961199

[14] 王彬, 蒋鸿玲, 吴槟. 基于 Attention-Bi-LSTM 的微博评论情感分析研究[J]. 计算机科学与应用, 2020, 10(12): 2380-2387.

[15] Naseem, U., Razzak, I., Musial, K. and Imran, M. (2020) Transformer Based Deep Intelligent Contextual Embedding for Twitter Sentiment Analysis. *Future Generation Computer Systems*, **113**, 58-69. https://doi.org/10.1016/j.future.2020.06.050

[16] Fasel, B. and Luettin, J. (2003) Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, **36**, 259-275. https://doi.org/10.1016/s0031-3203(02)00052-3

[17] Li, J., Zhang, D., Zhang, J., Zhang, J., Li, T., Xia, Y., *et al.* (2017) Facial Expression Recognition with Faster R-CNN.

Procedia Computer Science, **107**, 135-140. https://doi.org/10.1016/j.procs.2017.03.069

[18] Nancy, A.M., Kumar, G.S., Doshi, P. and Shaw, S. (2018) Audio Based Emotion Recognition Using Mel Frequency Cepstral Coefficient and Support Vector Machine. *Journal of Computational and Theoretical Nanoscience*, **15**, 2255-2258. https://doi.org/10.1166/jctn.2018.7447

[19] Koolagudi, S.G. and Rao, K.S. (2012) Emotion Recognition from Speech: A Review. *International Journal of Speech Technology*, **15**, 99-117. https://doi.org/10.1007/s10772-011-9125-1

[20] Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L. (2017) Tensor Fusion Network for Multimodal Sentiment Analysis. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 9-11 September 2017, 1103-1114. https://doi.org/10.18653/v1/d17-1115

[21] He, L., Wang, Z., Wang, L. and Li, F. (2023) Multimodal Mutual Attention-Based Sentiment Analysis Framework Adapted to Complicated Contexts. *IEEE Transactions on Circuits and Systems for Video Technology*, **33**, 7131-7143. https://doi.org/10.1109/tcsvt.2023.3276075

[22] Hazarika, D., Zimmermann, R. and Poria, S. (2020) MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 1122-1131. https://doi.org/10.1145/3394171.3413678

[23] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., *et al.* (2020) Transformers: State-Of-The-Art Natural Language Processing. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing*: *System Demonstrations*, 16-20 November 2020, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[24] Zadeh, A., Zellers, R., Pincus, E. and Morency, L.P. (2016) MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv: 1606. 06259.