

混合监督对比学习的多模态情感分析特征对齐方法

金程浩,朱翠涛*

(中南民族大学 智能无线通信湖北省重点实验室,湖北 武汉 430074)

摘要 针对多模态情感分析中特征融合前模态对齐不足的问题,提出了一种由情感分数差异引导的,作用于模态内和跨模态的混合监督对比学习.该方法利用样本标签之间的差异,使差异小的样本之间构成正样本对,差异大的样本之间构成负样本对,并基于此开展监督对比学习,通过学习模态内和跨模态的情感语义关联,促使不同模态特征在语义空间中更一致地对齐.同时,引入动态注意力加权融合机制,通过自适应评估不同模态对当前样本的贡献,优化特征融合效果.在不同数据集上的对比实验表明:所提方法具有良好的性能.相较于基线模型,在CMU-MOSI数据集上,五分类准确率最高提升19.98%,皮尔逊相关系数最高提升0.167;在CMU-MOSEI数据集上,七分类准确率最高提升3.07%;在CH-SIMS数据集上,二分类准确率和三分类准确率最高分别提升7.66%和9.19%,F1分数最高提升7.71.

关键词 多模态情感分析;模态对齐;监督对比学习

中图分类号 TP391 文献标志码 A 文章编号 1672-4321(2026)06-0823-11

doi:10.20056/j.cnki.ZNMDZK.20260611

Feature alignment method for multimodal sentiment analysis based on hybrid supervised contrastive learning

Jin Chenghao, Zhu Cuitao*

(Hubei Key Laboratory of Intelligent Wireless Communication, South-Central Minzu University, Wuhan 430074, Hubei China)

Abstract To address the problem of insufficient modal alignment before feature fusion in multimodal sentiment analysis, an emotion score difference-guided hybrid supervised contrastive learning method that operates on both intra-modal and cross-modal relationships is proposed. Specifically, the method uses differences between sample labels to construct positive and negative sample pairs, where samples with smaller differences are regarded as positive pairs and those with larger differences are regarded as negative pairs. Based on this, supervised contrastive learning is performed to capture affective semantic relationships within and across modalities, thereby encouraging features from different modalities to align more consistently in the semantic space. In addition, a dynamic attention-weighted fusion mechanism is introduced to optimize feature fusion by adaptively estimating the contribution of each modality for the current sample. Comparative experiments on different datasets demonstrate the effectiveness of the proposed method. Compared with baseline models, the proposed method achieves a maximum improvement of 19.98% in five-class accuracy and 0.167 in Pearson correlation on the CMU-MOSI dataset; a maximum improvement of 3.07% in seven-class accuracy on the CMU-MOSEI dataset; and maximum improvements of 7.66% and 9.19% in binary and three-class accuracy, respectively, as well as 7.71 in F1 score, on the CH-SIMS dataset.

Keywords multimodal sentiment analysis; modal alignment; supervised contrastive learning

多模态情感分析融合文本、语音和视觉的互补信息,能同时把握语言、韵律和表情信息,弥补单模

收稿日期 2026-02-14

*通信作者 朱翠涛(1967-),男,教授,博士,研究方向:无线通信及最优化理论,E-mail:cuitaozhu@mail.scuec.edu.cn

基金项目 国家自然科学基金资助项目(61671483);湖北省自然科学基金资助项目(2016CFA089)

态在讽刺、反语和强度刻画上的不足,并且因其更加贴合真实世界的情感表达场景而备受关注^[1]. 在客服质检、直播风控、车载交互和教育测评等场景具有重要应用价值^[2-3]. 然而,不同模态之间存在异质性,在时间粒度方面,不同模态数据的时序信息并不同步,同一时间戳的特征可能对应不同的情感片段^[4];在语义信息方面,模态之间的情感语义与真实的语义之间存在差异^[5],因此,多模态特征往往呈现“不对齐”的特点^[6].

早期研究通常在模型训练前对多模态序列进行预对齐,以缓解时序不一致问题^[7]. 然而,此类方法依赖于独立的预处理步骤,使模型不能在训练过程中动态适应真实场景中的模态错位;同时,他们主要关注在时间维度上的对齐,忽略了不同模态在情感语义表达上的差异^[8]. 为克服上述不足,部分研究在模型内部显式建模模态间的对齐关系. 例如,基于跨模态 Transformer 的方法通过注意力机制建模模态间的交互关系^[9]. 然而,此类方法的计算复杂度高,对情感差异和语义对齐的建模能力不足^[10-11]. 统计对齐的方法通过最大化模态间的相关性实现特征对齐,具有直观、可解释性优势,但其多依赖线性相关性假设,难以刻画复杂的非线性交互和高维语义关系^[12]. 此外,图神经网络方法通过构建图结构以建模模态间的复杂关系,虽具表达能力,但对图结构与关系定义高度依赖,设计成本较高^[13-14].

近年来,对比学习逐渐成为多模态学习的重要工具,其通过构建正负样本对并优化对比损失增强表征一致^[15-17]. 然而,现有多模态情感分析中的对比学习方法大多仅依据标签是否一致来构建正负样本对,侧重整体情感一致性的约束,难以充分刻画情感强度变化和表达差异,从而限制了其在复杂情感场景中的对齐效果.

针对上述方法的局限性,本文提出了一种由情感分数差异引导的,作用于模态内和跨模态的混合监督对比学习,用于多模态特征对齐. 与跨模态 Transformer 方法和图神经网络方法相比,本文方法更关注语义信息层面的对齐,不需要建立复杂的序列级交互,也不需要额外构建图结构并人为定义节点间关系,结构简洁,计算复杂度低. 与统计对齐方法相比,本文方法并非简单使不同模态特征在整体统计分布上保持一致,而是利用对比学习约束样本级特征关系,融合更深层次语义信息;与传统对比学习方法相比,本文方法利用连续情感分数差异反映样本间的情感接近程度,从而实现更细粒度的正

负样本划分.

具体而言,本方法利用连续情感分数差异构造细粒度正负样本对,使模型从样本“是否同类”的粗粒度约束转换为“接近何种程度”的细粒度约束,从而增强对情感强度变化的刻画能力;同时,本文构建作用于模态内和跨模态的混合监督对比学习(HSCL),通过使前者保持模态内部情感结构稳定、后者增强不同模态之间的语义一致性,实现多模态特征对齐;最后,通过动态注意力加权融合模块(DAWF)根据不同样本中各模态的贡献自适应分配权重,从而突出有效信息,提升融合表示的质量和预测性能.

1 模型设计

本文模型的整体结构如图 1 所示. 首先,模型对原始多模态输入进行特征提取,获得各模态的低级特征. 其中,视觉特征包含面部动作单元、姿态变化等动态信息,语音特征包含韵律参数、音色等声学信息,文本特征包含词汇语义和语法结构等语言信息. 随后,各模态特征分别输入对应的编码器获得高级表示,并通过线性映射投影到统一特征空间. 为缓解不同模态特征在融合前的语义不一致问题,模型在统一特征空间中引入情感分数差异引导的混合监督对比学习,从跨模态语义对齐和模态内结构保持两个层面对特征施加对比约束,实现多模态特征对齐. 最后,将对齐后的各模态特征输入动态注意力加权融合模块,得到融合特征,并由多层感知机输出连续的情感强度预测分数.

1.1 混合监督对比学习

为了在统一特征空间内实现多模态特征对齐,本文设计了混合监督对比学习机制,其整体过程如图 2 所示. 以情感分数标签为 0.8 的样本为例,首先根据情感分数差异与阈值之间的关系,将样本对划分为正样本对和负样本对;随后,通过权重矩阵给不同的样本对分配不同的监督强度,以增强细粒度约束. 在此基础上,分别计算跨模态和模态内相似度矩阵,并结合权重矩阵,得到加权后的相似度矩阵. 接着,对加权相似度矩阵中的正样本项以及全部正负样本项分别进行聚合,得到正样本相似度之和与正负样本相似度总和,并代入到 InfoNCE 损失式中,构造单项对比损失. 最后,通过对两个方向的单向损失进行聚合,得到模态对之间的双向对比损

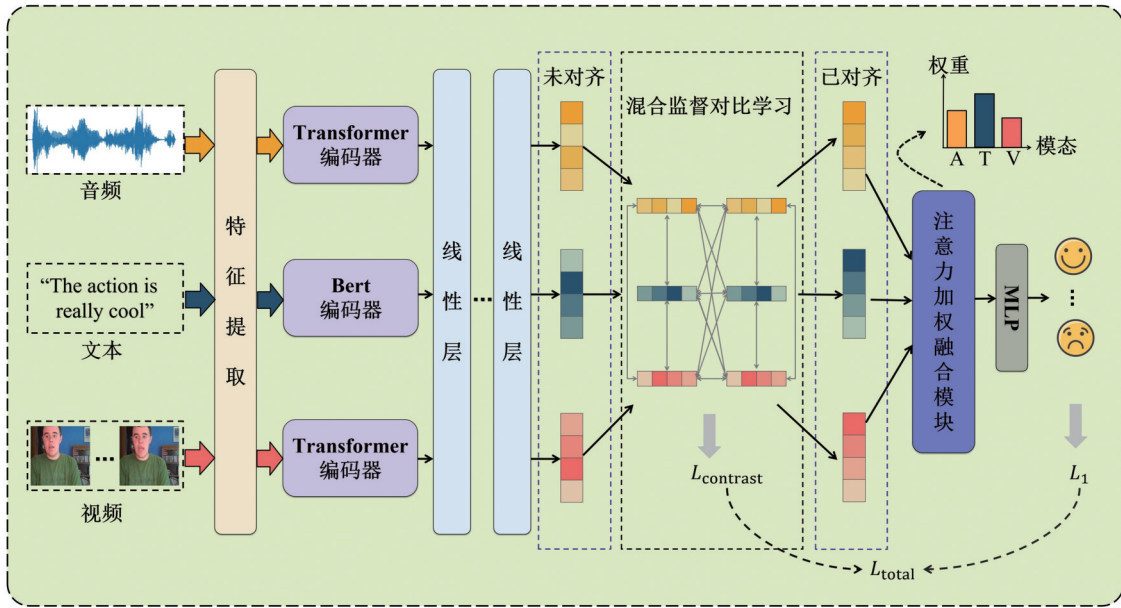


图1 模型整体结构

Fig. 1 Overall structure of model

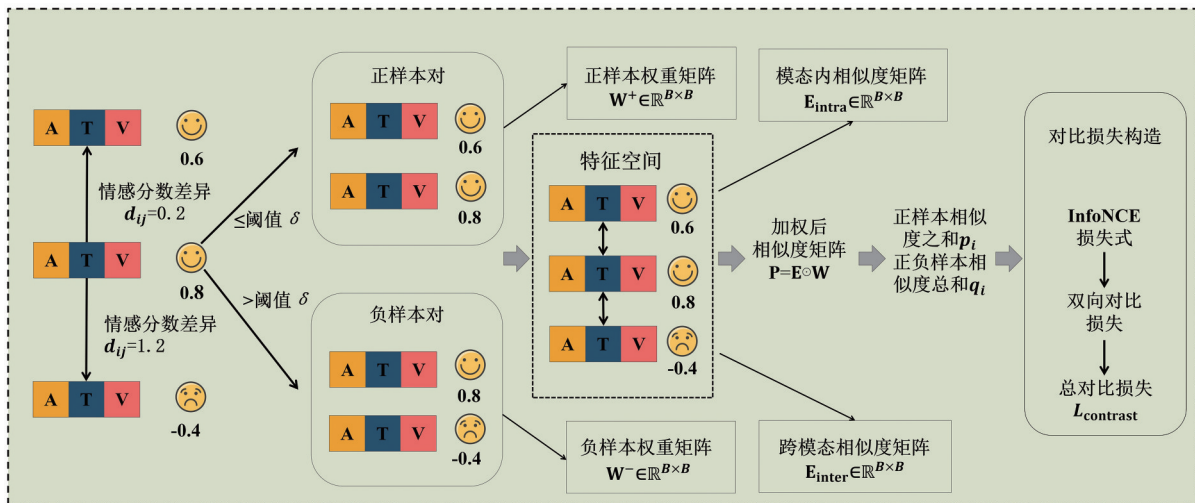


图2 混合监督对比学习技术图

Fig. 2 Technical diagram of hybrid supervised contrastive learning

失,并进一步对所有模态对的双向损失求和,得到总对比损失。

1.1.1 阈值确定与样本对构造

由于不同数据集的情感分数标签范围有所差异,并且阈值 δ 的设定会对正负样本构造产生影响,因此,在基于情感分数差异构造正负样本对之前,首先需要统一不同数据集下情感标签的度量尺度,并确定样本对划分阈值 δ 。为了统一标签尺度,本文将情感分数标签 y_i 线性映射到区间 $[-1, 1]$,其形式可表示为:

$$\tilde{y}_i = 2 \cdot \frac{y_i - y_{\min}}{y_{\max} - y_{\min}} - 1, \quad (1)$$

其中, $y_{\min}$ 和 $y_{\max}$ 表示原始标签的最小值和最大值, $\tilde{y}_i$

为映射后的标签。对于任意样本对 (i, j) ,定义其情感差异为:

$$d_{ij} = |\tilde{y}_i - \tilde{y}_j|, \quad (2)$$

由于 $\tilde{y} \in [-1, 1]$,因此样本对的情感差异满足 $d_{ij} \in [0, 2]$ 。避免阈值设定完全依赖经验,本文借鉴最大类间方差阈值分割思想,在训练集样本对的情感差异分布上确定理论参考阈值。具体地,记 t 为候选阈值, $T \subset [0, 2]$ 为情感差异区间 $[0, 2]$ 离散化得到的候选阈值集合。对于任意候选阈值 t ,先依据样本对的情感差异将训练集中的样本对划分为正负两类:

$$P^+(t) = \{(i, j) | d_{ij} \leq t\}, \quad P^-(t) = \{(i, j) | d_{ij} > t\}. \quad (3)$$

在此基础上,定义类间分离度函数 $J(t)$,用于衡量候选阈值 t 下正负样本对的可分离程度:

$$J(t) = \omega_+(t)\omega_-(t)(\mu_+(t) - \mu_-(t))^2, t \in T, \quad (4)$$

其中, $\omega_+(t)$ 与 $\omega_-(t)$ 分别表示候选阈值 t 下正负样本对在训练集中的比例. 设训练集中全部样本对总数为 M ,则有:

$$\omega_+(t) = \frac{|P^+(t)|}{M}, \omega_-(t) = \frac{|P^-(t)|}{M}. \quad (5)$$

$\mu_+(t)$ 和 $\mu_-(t)$ 分别表示两类样本对的平均情感差异,表示为:

$$\mu_+(t) = \frac{1}{|P^+(t)|} \sum_{(i,j) \in P^+(t)} d_{ij}, \mu_-(t) = \frac{1}{|P^-(t)|} \sum_{(i,j) \in P^-(t)} d_{ij}. \quad (6)$$

当 $J(t)$ 取值最大时,说明在当前阈值下正负样本对之间的差异越明显,类间分离效果越好. 因此,本文将类间分离度最大的阈值作为理论参考值,即:

$$\delta = \arg \max_{t \in T} J(t). \quad (7)$$

确定阈值 δ 后,本文在训练过程中根据批内样本间的情感分数差异构造正负样本对. 对于同一批次内任意样本对 (i,j) ,若其分数差异满足 $d_{ij} \leq \delta$,认为二者在情感语义上足够接近,将其划分为正样本对;若 $d_{ij} > \delta$,认为二者在情感语义上具有明显差异,将其划分为负样本对. 由此,正负样本对集合可表示为:

$$P^+ = \{(i,j) | d_{ij} \leq \delta\}, P^- = \{(i,j) | d_{ij} > \delta\}. \quad (8)$$

为了进一步说明正负样本对的构造过程,将其总结为算法 1.

算法 1 正负样本对的构造流程

输入: 标签 y , 阈值 δ

输出: 正样本集合 P^+ , 负样本集合 P^-

步骤:

- 1 按式(1)将标签 y 映射到区间 $[-1, 1]$, 得到 $\tilde{y}$
- 2 初始化正样本集合 $P^+ = \emptyset$, 负样本集 $P^- = \emptyset$
- 3 对任意样本对 (i,j) 按式(2)计算情感分数差异 d_{ij}
- 4 若 $d_{ij} \leq \delta$, 按式(8)将样本对加入正样本集合 P^+
- 5 若 $d_{ij} > \delta$, 按式(8)将样本对加入负样本集合 P^-
- 6 输出 P^+, P^-

1.1.2 权重矩阵构造

完成正负样本对的划分后,若对所有正样本和负样本赋予统一权重,则会忽略样本对之间仍然存在的细粒度情感差异. 对于正样本,情感分数越接近的样本对在语义上越一致,应施加更强的聚合约

束;对于负样本,不同样本对与阈值边界的相对位置也存在差异,情感差异越大的样本对应施加更强的分离约束. 因此,本文进一步构造权重矩阵,以增强混合监督对比学习的针对性.

设批次内样本对的情感差异矩阵为 $\mathbf{D} = [d_{ij}] \in \mathbb{R}^{B \times B}$,其中 B 表示批次大小. 基于情感差异矩阵 $\mathbf{D}$,构造正样本权重矩阵 $\mathbf{W}^+ \in \mathbb{R}^{B \times B}$ 和负样本权重矩阵 $\mathbf{W}^- \in \mathbb{R}^{B \times B}$.

对于正样本,采用 $\tanh(\cdot)$ 非线性函数构造连续权重,并设定权重上限为 g ,得到正样本权重矩阵 $\mathbf{W}^+$:

$$W_{ij}^+ = \begin{cases} -g \tanh(d_{ij} - 2\delta), & d_{ij} \leq \delta, \\ 0, & \text{otherwise}, \end{cases} \quad (9)$$

对于负样本,为了避免所有负样本施加统一惩罚,构造负样本权重矩阵 $\mathbf{W}^-$:

$$W_{ij}^- = \begin{cases} \beta \cdot (d_{ij}) \cdot \tanh(d_{ij} - 2\delta), & d_{ij} > \delta, \\ 0, & \text{otherwise}, \end{cases} \quad (10)$$

其中, $\beta(d_{ij})$ 定义为:

$$\beta(d_{ij}) = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot \text{Sigmoid}(k(d_{ij} - \delta)), \quad (11)$$

这里 $\beta_{\min}$ 和 $\beta_{\max}$ 分别表示负样本惩罚系数的下界和上界, k 为Sigmoid函数的斜率参数. 此设计使靠近阈值边界的负样本受到较弱惩罚,远离阈值边界的负样本受到更强惩罚.

1.1.3 构造相似度矩阵

权重矩阵仅表示不同样本对的相对重要性,不能直接反映样本之间的相近程度. 为此,本文分别构造跨模态相似度矩阵和模态内相似度矩阵,并与正负权重矩阵相结合.

设经过编码和线性映射后的音频、文本、视频在特征空间里的特征矩阵分别为 $\mathbf{Z}_a, \mathbf{Z}_t, \mathbf{Z}_v \in \mathbb{R}^{B \times D}$,其中 B 为批次大小, D 为特征维度. 不同样本的特征向量幅值存在差异,为了减小幅值差异对相似度计算的影响,对各模态特征逐行进行L2归一化,得到:

$$\hat{\mathbf{Z}}_m = \frac{\mathbf{Z}_m}{\|\mathbf{Z}_m\|_2}, m \in \{a, t, v\}. \quad (12)$$

对任意两种模态 $m, n \in \{a, t, v\}$ 且 $m \neq n$,跨模态相似度矩阵定义为:

$$\mathbf{S}_{\text{inter}}^{(m,n)} = \hat{\mathbf{Z}}_m \hat{\mathbf{Z}}_n^T, \quad (13)$$

对应地,模态 m 的内部相似度矩阵定义为:

$$\mathbf{S}_{\text{intra}}^{(m,m)} = \hat{\mathbf{Z}}_m \hat{\mathbf{Z}}_m^T. \quad (14)$$

其中, $\mathbf{S}_{\text{inter}}^{(m,n)} \in \mathbb{R}^{B \times B}$ 反映不同模态特征的语义相似度; $\mathbf{S}_{\text{intra}}^{(m,m)} \in \mathbb{R}^{B \times B}$ 反映模态内部样本之间的语义关联. 为了调节不同样本对相似度差异的区分程度,

并将相似度转化为非负形式,引入温度系数 τ 进行缩放并做指数化处理:

$$\mathbf{E}_{\text{inter}}^{(m,n)} = \exp\left(\frac{\mathbf{S}_{\text{inter}}^{(m,n)}}{\tau}\right), \quad (15)$$

$$\mathbf{E}_{\text{intra}}^{(m,m)} = \exp\left(\frac{\mathbf{S}_{\text{intra}}^{(m,m)}}{\tau}\right). \quad (16)$$

随后,将相似度矩阵与对应的权重矩阵联合计算,得到加权后的正负相似度矩阵:

$$\mathbf{P}_{\text{inter}}^+(m,n) = \mathbf{E}_{\text{inter}}^{(m,n)} \odot \mathbf{W}^+, \quad (17)$$

$$\mathbf{P}_{\text{inter}}^-(m,n) = \mathbf{E}_{\text{inter}}^{(m,n)} \odot \mathbf{W}^-, \quad (18)$$

$$\mathbf{P}_{\text{intra}}^+(m,m) = \mathbf{E}_{\text{intra}}^{(m,m)} \odot \mathbf{W}^+, \quad (19)$$

$$\mathbf{P}_{\text{intra}}^-(m,m) = \mathbf{E}_{\text{intra}}^{(m,m)} \odot \mathbf{W}^-, \quad (20)$$

其中, $\odot$ 表示逐元素乘法.

1.1.4 双向对比损失

为了将相似度信息转化为可以直接优化的训练目标,本文构造双向对比损失.在该损失约束下,表达相似情感的跨模态样本表示在统一特征空间中相互靠近,增强模态间语义对齐;同时,模态内情感接近的样本表示相互靠拢,增强模态内部语义结构的一致性.相反,表达不同情感的样本表示相互远离.

对任意两个模态 m,n ,计算 m 到 n 的单方向对比损失:

$$L_{m \rightarrow n} = -\frac{1}{B} \sum_{i=1}^B \log \frac{p_i^{m \rightarrow n}}{q_i^{m \rightarrow n}}, \quad (21)$$

其中, $p_i^{(m \rightarrow n)}$ 表示第 i 个样本在跨模态和模态内正样本的加权相似度之和, $q_i^{(m \rightarrow n)}$ 表示第 i 个样本在跨模态和模态内全部正负样本上的加权相似度总和,具体写为:

$$p_i^{m \rightarrow n} = \sum_{j=1}^B P_{ij,\text{inter}}^+(m,n) + \sum_{j=1}^B P_{ij,\text{intra}}^+(m,m), \quad (22)$$

$$q_i^{m \rightarrow n} = p_i^{m \rightarrow n} + \sum_{j=1}^B P_{ij,\text{inter}}^-(m,n) + \sum_{j=1}^B P_{ij,\text{intra}}^-(m,m). \quad (23)$$

为了保证模态对齐的双向性,取双向损失的均值:

$$L_{m \leftrightarrow n} = \frac{1}{2} (L_{m \rightarrow n} + L_{n \rightarrow m}). \quad (24)$$

最终,总对比损失定义为所有模态的双向损失之和:

$$L_{\text{contrast}} = L_{v \leftrightarrow a} + L_{v \leftrightarrow t} + L_{t \leftrightarrow a}. \quad (25)$$

通过最小化 L_{contrast} ,模型能在统一特征空间内强化跨模态语义对齐,同时保持模态内情感结构一致.

1.2 动态注意力加权融合模块

在获得对齐后的文本、语音与视频三种模态特征后,本文进一步设计了一种动态注意力加权融合模块.该模块针对每个样本自适应地估计文本、语音和视频模态在当前语义表达中的相对重要性,并据此完成加权融合.如图3所示,首先,将对齐后的3种模态特征在特征维度上进行拼接形成联合表示;随后,将联合表示输入由线性层、Dropout层和Relu激活函数组成的注意力感知网络,生成3个模态对应的未归一化注意力得分;在此基础上,通过缩放和softmax归一化得到样本级模态权重;最后,利用该权重对3种模态特征进行加权求和,形成统一维度的融合表示,供后续分类预测.

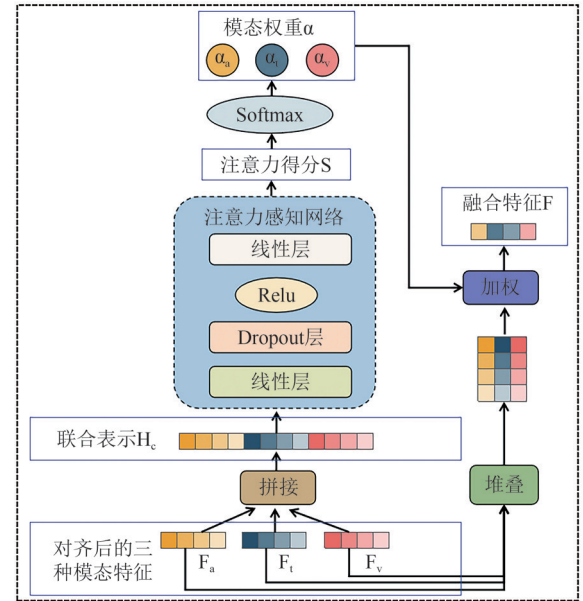


图3 动态注意力加权融合模块

Fig. 3 Dynamic attention-weighted fusion module

设经过对齐后的3种模态特征分别为 $\mathbf{F}_a, \mathbf{F}_t, \mathbf{F}_v \in \mathbb{R}^{B \times D}$,其中 B 表示批次大小, D 表示特征维度.为了生成模态权重,首先将三种模态特征在特征维度上进行拼接,形成联合表示:

$$\mathbf{H}_e = \text{Concat}(\mathbf{F}_a, \mathbf{F}_t, \mathbf{F}_v), \in \mathbb{R}^{B \times 3D}. \quad (26)$$

随后,将联合表示 $\mathbf{H}_e$ 输入注意力感知网络,得到3个模态对应的注意力得分 $\mathbf{S} \in \mathbb{R}^{B \times 3}$.为了控制注意力得分的数值尺度,本文进一步引入缩放因子 d_s ,并将其大小设置为 $\sqrt{D}$,再通过softmax函数对3个模态的得分进行归一化,得到样本级模态权重:

$$\boldsymbol{\alpha} = \text{softmax}\left(\frac{\mathbf{S}}{d_s}\right), \boldsymbol{\alpha} \in \mathbb{R}^{B \times 3}. \quad (27)$$

获得权重后,本文将3种模态特征沿新维度堆叠形成三模态特征张量,再与注意力权重进行加权

求和,得到融合表示.其结果等价表示为:

$$\mathbf{F} = \alpha_a \odot \mathbf{F}_a + \alpha_t \odot \mathbf{F}_t + \alpha_v \odot \mathbf{F}_v, \mathbf{F} \in \mathbb{R}^{B \times D}, \quad (28)$$

其中, $\alpha_a, \alpha_t, \alpha_v \in \mathbb{R}^{B \times 1}$ 分别表示音频、文本和视频模态的样本级权重, $\odot$ 表示逐元素乘法. 通过上述方式,模型能够根据不同样本的具体内容,自适应调整各模态在融合过程中的贡献,从而提升融合表示的有效性与判别性.

1.3 目标函数

本文采用由回归损失与对比损失加权组成的联合目标函数. 其中,回归损失用于计算预测情感分数与真实分数之间的误差;对比损失用于约束多模态表示空间中的特征对齐(具体形式见 1.2 节).

设批次大小为 B ,第 i 个样本的模型预测值与真实标签分别为 $\hat{y}_i$ 与 y_i . 回归损失采用 L_1 损失:

$$L_1 = \frac{1}{B} \sum_{i=1}^B |y_i - \hat{y}_i|, \quad (29)$$

对比损失 L_{contrast} 按第 1.1 节所述定义,最终训练目标为两者的加权和:

$$L_{\text{total}} = L_1 + \lambda L_{\text{contrast}}, \quad (30)$$

其中 λ 是对比损失的权重系数,用于平衡回归学习与表示对齐之间约束. 所提方法的整体流程如算法 2 所示.

算法 2 混合监督对比学习的多模态情感分析特征对齐方法

输入: 单模态特征 $\mathbf{X}_m, m \in \{a, t, v\}$, 标签 y_i , 模型参数

输出: 模型预测标签 $\hat{y}_i$

步骤:

- 1 按公式(1)-(7)确定阈值 δ
- 2 for each sample in batch do
- 3 按公式(1)、(2)、(8)划分正负样本对
- 4 按式(9)-(11)计算正负样本权重矩阵
- 5 按式(12)-(14)计算跨模态/模态内相似度矩阵
- 6 按照式(15)、(16)将相似度矩阵指数化
- 7 按照式(17)-(20)将相似度矩阵加权
- 8 按式(21)-(25)计算总对比损失 L_{contrast}
- 9 按式(26)-(28)将特征融合并预测 $\hat{y}_i$
- 10 按式(29)、(30)得联合损失 L_{total}
- 11 反向传播更新网络参数
- 12 end

2 实验

2.1 数据集和评估指标

本文在 3 个公开且广泛使用的数据集上进行实验:英文数据集 CMU-MOSI^[18]、CMU-MOSEI^[19] 和中文数据集 CH-SIMS^[20]. CMU-MOSI 数据集包含 2199

个独白视频样本,按训练集、验证集和测试集划分为 1284、229 和 686. CMU-MOSEI 含有 22856 个样本,按训练集、验证集和测试集划分为 16326、1871 和 4659. CH-SIMS 由中文互联网视频片段构成,共 2281 个样本,训练集、验证集和测试集数量分别为 1368、456 和 457.

评估指标遵循多模态情感分析常用设置,包含七分类准确率 Acc_7、五分类准确率 Acc_5、三分类准确率 Acc_3、二分类准确率 Acc_2、F1 分数、平均绝对误差 MAE 和皮尔逊相关系数 Corr. 其中,在 CMU-MOSI 与 CMU-MOSEI 数据集上,Acc_2 和 F1 采用两种测评设置同时报告,并以“/”分割:左侧为包含零标签样本的结果,右侧为去除零标签样本的结果.

2.2 实验设置

实验基于 PyTorch 2.4.1 与 Python 3.8.0 实现,运行环境为 NVIDIA GeForce RTX4060 Laptop GPU, CUDA 版本为 11.8. 对于主流多模态情感分析基线模型,本文采用 MMSA^[20] 统一框架进行复现;其余 SOTA 方法参照原论文的实验配置并结合公开代码复现. 训练过程中采用早停策略:若验证集损失指标连续 8 个 epoch 未提升时,提前终止训练并在测试集上进行测试. 对于 CMU-MOSI 数据集,设置阈值 δ 为 0.64,对于 CMU-MOSEI 数据集,设置 δ 为 0.5,对于 CH-SIMS 数据集,设置 δ 为 0.8.

为了说明模型的训练与收敛过程,本文以 CMU-MOSEI 数据集为例,如图 4 所示,模型在训练初期验证集损失快速下降、相关系数迅速提升,说明模型能较快学习到有效的多模态表征. 在第 6 个 epoch 之后,验证损失与相关性波动较小,模型逐渐收敛,并且在第 14 个 epoch 上取得验证最小损失. 因此,本文在早停策略下选择第 14 个 epoch 的模型作为最终测试模型.

本文采用的基线模型如下:

TFN^[7]:张量融合网络,用于解决特征融合不充分的问题.

LMF^[21]:低秩张量融合网络,用于降低模型的数量与计算复杂度.

Graph_MFN^[19]:一种引入图结构网络来进行多模态交互的模型.

BERT-MAG^[22]:通过在微调阶段用门控机制注入其余模态信息的模型.

Mult^[9]:通过成对的跨模态 Transformer 进行对齐的模型.

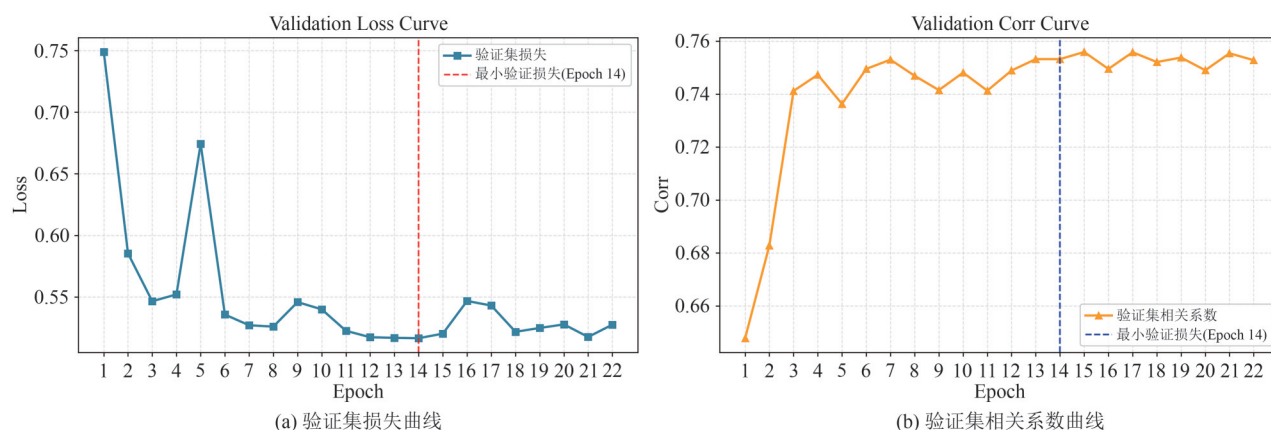


图4 在CMU-MOSEI上验证集损失与相关系数随训练轮次的变化曲线

Fig. 4 Curves of validation loss and correlation coefficient over training epochs on CMU-MOSEI

MISA^[8]:通过特征解耦来缓解模态间分布差异的模型。

MMIM^[23]:通过分层互信息最大化策略来抑制模态间冗余问题的模型。

ALMT^[24]:用文本模态进行指导,消除多模态特征的冗余与冲突的模型。

TETFN^[25]:以文本特征为中心,对非语言模态进行文本增强捕获时序依赖的模型。

SELF_MM^[26]:自监督多任务学习模型,在标签在缺乏单模态标注时让模型仍能学习到更具区分的特有表征。

Prisa^[27]:优先融合与近距离感知对比学习,改善跨模态对齐不足问题的框架。

2.3 性能对比

在未对齐的CMU-MOSI数据集上,将所提出模型与多个具有代表性的多模态情感分析模型进行对比实验,结果如表1所示。表中Acc₂和F1的双数值结果分别对应包含零标签样本与去除零标签样本两种测评设置:前者表示在全部测试样本上进行二分类评估,后者表示将情感标签为零的中性样本剔除后,仅对正负样本进行二分类评估。

相较于基线模型,本文模型在整体指标上表现良好。具体而言,二分类准确率Acc₂(含/不含零标签样本)分别提升0.88%~8.17%与0.92%~9%,五分类准确率Acc₅提升0.59%~19.98%,七分类准确率Acc₇最高提升13.56%;同时F1分数(含/不含零标签样本)分别提升0.83~8.08与0.93~8.89,相关系数Corr提升0.007~0.167,平均绝对误差MAE降低0.012~0.309。上述结果表明,情感分数差异引导的混合监督对比学习能够促进跨模态特征对齐,动态注意力加权融合模块可按样本自适应分配模态权

重,从而提升预测的稳定性与精度。需要指出的是,本文模型在Acc₇指标上略低于BERT-MAG和SELF_MM。这是因为Acc₇对细粒度离散类别边界更为敏感,而BERT-MAG通过门控方式将音频和视觉信息注入预训练语言表示,更有利于增强对细粒度类别的判别能力;SELF_MM则通过自监督多任务学习强化模态特异表示学习,因此在七分类任务上表现出一定优势。

在CMU-MOSEI数据集上,进一步将所提模型与多个主流多模态情感分析模型进行对比实验,结果如表2所示。与表1相同,表2中的Acc₂和F1的双数值结果分别对应包含零标签样本和去除零标签样本的两种测评设置。总体来看,本文模型在多数评价指标上均表现领先:二分类准确率Acc₂(含/不含零标签样本)最高提升11.62%与4.65%,五分类准确率Acc₅提升0.07%~3.31%,七分类准确率Acc₇提升0.28%~3.07%;同时,F1分数(含/不含零标签样本)最高分别提升10.23与4.31,相关系数Corr提升0.009~0.066,平均绝对误差MAE降低0.007~0.05。上述实验表明:本文模型在更大规模、标签更丰富的场景下仍具备稳定优势和良好的泛化能力。然而,在去除零标签样本的Acc₂和F1指标上,ALMT的表现略优于本文模型。这是因为ALMT的方法更强调在语言引导下抑制辅助模态中的冗余和冲突信息,因此更适用于去除中性样本后的极性判别任务。

在中文数据集CH-SIMS上实验结果如表3所示。本文模型在各项指标上有明显提升:二分类准确率Acc₂提升2.41%~7.66%,三分类准确率Acc₃提升1.97%~9.19%,五分类准确率Acc₅提升0.87%~6.78%,F1分数提升2.01~7.71,相关系数Corr最高提

表1 CMU-MOSI测试集上的性能比较
Tab. 1 Performance comparison on the CMU-MOSI test set

Model	MOSI					
	Acc_2/%	F1	Acc_5/%	Acc_7/%	MAE	Corr
TFN	76.82/77.74	76.72/77.70	37.76	33.67	0.961	0.633
LMF	75.36/76.37	75.43/76.51	32.94	31.34	1.010	0.656
Graph_MFN	77.55/78.66	77.40/78.56	41.11	36.88	0.942	0.642
BERT_MAG	82.65/84.45	82.68/84.41	52.04	45.77	0.713	0.793
Mult	80.61/82.01	80.40/81.85	41.69	36.01	0.893	0.689
MISA	80.32/82.01	80.35/82.10	47.81	42.42	0.780	0.774
MMIM	82.36/84.45	82.32/84.47	46.36	41.11	0.779	0.767
ALMT	81.20/82.62	81.16/82.64	48.40	43.73	0.767	0.774
TETFN	80.90/82.32	80.91/82.38	49.42	43.29	0.736	0.790
SELF_MM	80.32/82.47	80.28/82.50	52.33	45.48	0.750	0.772
Prisa	81.60/83.50	81.60/83.60	49.90	43.00	0.768	0.772
本文模型	83.53/85.37	83.51/85.40	52.92	44.90	0.701	0.800

表2 CMU-MOSEI测试集上的性能比较
Tab. 2 Performance comparison on the CMU-MOSEI test set

Model	MOSEI					
	Acc_2/%	F1	Acc_5/%	Acc_7/%	MAE	Corr
TFN	75.75/82.03	76.75/82.19	53.19	51.64	0.574	0.713
LMF	81.50/85.17	82.01/85.17	54.43	52.89	0.553	0.745
Graph_MFN	73.36/80.88	74.59/81.17	52.56	51.21	0.573	0.723
BERT_MAG	80.58/84.87	81.23/84.94	54.48	52.61	0.544	0.762
Mult	79.24/84.59	79.97/84.63	54.52	52.87	0.558	0.740
MISA	82.72/85.36	83.02/85.25	53.87	52.35	0.547	0.763
MMIM	83.24/83.27	83.00/82.76	53.81	52.01	0.575	0.724
ALMT	81.61/85.75	82.16/85.77	54.97	53.40	0.535	0.770
TETFN	81.07/85.11	81.58/85.08	55.76	54.00	0.539	0.763
SELF_MM	79.70/84.18	80.31/84.17	55.25	53.53	0.532	0.763
Prisa	79.20/84.40	79.90/84.40	55.80	54.00	0.540	0.758
本文模型	84.98/85.53	84.82/85.48	55.87	54.28	0.525	0.779

升0.114,平均绝对误差最高降低0.055. 这表明模型在跨语言数据上具有可迁移性;另一方面,由于本文模型偏向于提升跨模态对齐效果和特征融合,在中文数据集上对连续情感强度的绝对误差控制和趋势一致性方面建模相对不足,相关指标未达到最优.

2.4 消融实验

2.4.1 模态消融实验

为了验证不同模态对模型性能的贡献,本文在 MOSI 数据集上进行了模态消融实验,结果如表 4 所示. 实验中 w/o 分别表示剔除文本、音频和视频后的模型. 结果表明,移除任一模态后,模型性能均明显下降;其中,二分类准确率 Acc_2(含/不含零标签样

本)最高下降 22.01%与 23.63%,F1 分数(含/不含零标签样本)最高下降 21.94 与 23.5,五分类准确率 Acc_5 下降 5.4%~34.55%,七分类准确率 Acc_7 下降 4.96%~26.53%,平均绝对误差 MAE 提升 0.134~0.636,相关系数 Corr 下降了 0.026~0.559. 说明文本、音频和视频模态对情感分析均有重要影响. 其中,移除文本模态后的性能退化最为显著,表明文本模态是情感表达的核心信息来源;移除音频或移除视频模态后,模型性能也有较大幅度下降,说明二者能为情感分析提供有效辅助. 相比之下,完整模型在 Acc_2、F1、Acc_5、Acc_7、MAE 和 Corr 等指标上均取得最佳结果,验证了多模态融合策略的有效性.

表3 CH-SIMS测试集性能比较
Tab. 3 Performance comparison on the CH-SIMS test set

Model	SIMS					
	Acc_2/%	Acc_3/%	Acc_5/%	F1	MAE	Corr
TFN	74.84	62.80	38.29	75.21	0.444	0.546
LMF	77.46	67.18	42.01	76.97	0.428	0.588
Graph_MFN	77.02	67.18	43.33	77.18	0.433	0.584
BERT_MAG	72.65	59.96	39.17	72.56	0.471	0.473
Mult	77.24	65.65	38.73	77.15	0.443	0.578
MISA	77.24	64.33	41.14	77.24	0.440	0.546
MMIM	73.74	63.02	37.42	73.46	0.478	0.482
ALMT	76.81	65.65	40.70	77.10	0.424	0.583
TETFN	77.90	64.33	42.45	78.26	0.410	0.582
SELF_MM	77.90	64.11	41.58	77.64	0.417	0.582
Prisa	76.37	65.65	38.73	76.22	0.424	0.610
本文模型	80.31	69.15	44.20	80.27	0.423	0.587

表4 模型在CMU-MOSI数据集上模态消融实验
Tab. 4 Modal ablation experiments of the model on the CMU-MOSI dataset

Model	Acc_2/%	F1	Acc_5/%	Acc_7/%	MAE	Corr
w/o 文本	61.52/61.74	61.57/61.90	18.37	18.37	1.337	0.241
w/o 音频	81.05/82.32	81.10/82.42	45.19	38.63	0.835	0.762
w/o 视频	80.32/81.10	80.37/81.21	47.52	39.94	0.853	0.774
完整模型	83.53/85.37	83.51/85.40	52.92	44.90	0.701	0.800

2.4.2 模块消融实验

为了验证混合监督对比学习机制(HSCL)和动态注意力加权融合模块(DAWF)的有效性,本文在MOSI数据集上进行了模块消融实验.具体而言,在完整模型的基础上,分别移除HSCL机制(w/o HSCL)、移除DAWF模块(w/o DAWF),以及同时移除两个模块(w/o ALL),并与完整模型进行比较,结果如表5所示,模型出现了不同程度的退化.移除HSCL后,Acc_2(含/不含零标签样本)下降1.17%和1.53%,F1分数分别下降1.24和1.48,Acc_5和Acc_7分别下降4.52%和3.79%,MAE上升0.06,Corr下

降0.018,说明HSCL通过增强跨模态语义对齐和保持模态内情感结构的一致性,有效提升了特征表示质量;移除DAWF后,Acc_2(含/不含零标签样本)下降1.02%和1.22%,F1分数分别下降1.11和1.29,Acc_5和Acc_7分别下降1.32%和0.15%,MAE上升0.047,Corr下降0.009;同时移除两个模块后,模型性能退化最明显,Acc_2、F1、Acc_5、Acc_7、MAE和Corr均达到最差.与之相比,完整模型在各项指标上均为最佳结果,进一步说明了HSCL和DAWF联合设计的有效性.

表5 模型在CMU-MOSI数据集上模块消融实验
Tab. 5 Module ablation experiments of the model on the CMU-MOSI dataset

Model	Acc_2/%	F1	Acc_5/%	Acc_7/%	MAE	Corr
w/o HSCL	82.36/83.84	82.27/83.92	48.40	41.11	0.761	0.782
w/o DAWF	82.51/84.15	82.40/84.11	51.60	44.75	0.748	0.791
w/o ALL	81.78/83.08	81.80/83.16	46.36	40.23	0.794	0.774
完整模型	83.53/85.37	83.51/85.40	52.92	44.90	0.701	0.800

为了更直观地感受4种设置对情感预测前融合特征表示的影响,进一步对融合特征进行可视化分析,结果如图5所示.图5(a)中,同时移除HSCL和DAWF后,融合特征分布最为涣散,类别间交叠明显;图5(b)中,移除HSCL后,特征聚合程度有所提升,但

整体仍较松散;图5(c)中,移除DAWF后,特征分布进一步收敛,但局部仍存在一定混叠;相比之下,图5(d)中的完整模型特征分布最为紧凑,类别边界更加清晰.该结果与表5中的定量分析一致,说明HSCL和DAWF均有助于提升融合表示的可分性与模型性能.

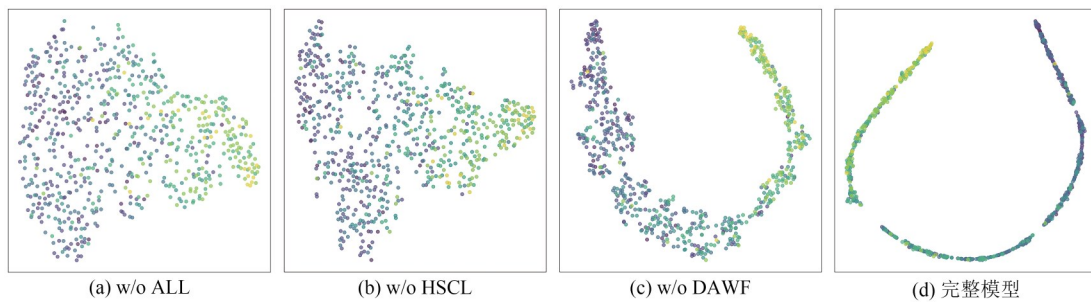


图 5 特征分布可视化

Fig. 5 Feature distribution visualization

3 结论

针对多模态情感分析中模态异质性导致的特征对齐不足问题,本文提出了一种由情感分数差异引导的混合监督对比学习方法,通过把特征映射到统一特征空间,构建跨模态和跨样本间的对比学习来增强特征对齐;同时,设计了动态注意力加权融合模块,优化特征融合的效果.在主流的数据集上进行的大量实验表明,所提出的方法能够有效缓解模态对齐不足的问题,并提升模型的预测能力.此外,消融实验进一步验证了对齐机制和融合机制的有效性.尽管本文方法达到了较好的整体性能,但仍存在不足:一方面,本文方法更侧重跨模态对齐和融合表示优化,因此在部分细粒度分类指标和连续回归指标上未达到最优,说明其在类别边界建模和情感强度拟合方面仍有提升空间;另一方面,本文主要针对完整模态场景展开研究,未充分考虑模态缺失、噪声干扰等复杂条件下的鲁棒性问题.未来将进一步研究细粒度情感类别边界建模与连续情感强度拟合方法,并面向模态缺失、噪声干扰等复杂场景,探索更鲁棒的特征对齐与自适应融合机制.

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] Nguyen T H, Shirai K. Topic modeling based sentiment analysis on social media for stock market prediction [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing: Association for Computational Linguistics, 2015: 1354-1364.
- [3] Zeng Y, Mai S J, Hu H, et al. Multimodal sentiment analysis in the wild: A survey [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(10): 4795-4814.
- [4] Huan Ruohong, Zhong Guowei, Chen Peng, et al. UniMF: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences [J]. IEEE Transactions on Multimedia, 2024, 26: 5753-5768.
- [5] Pan Y, Shen Y, Liang P P, et al. Cross-modal conflict resolution for multimodal sentiment analysis [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023: 1895-1906.
- [6] Chen Bingzhi, Cao Qi, Hou Mixiao, et al. Multimodal emotion recognition with temporal and semantic consistency [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3592-3603.
- [7] Zadeh A, Chen Minghai, 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: Association for Computational Linguistics, 2017: 1103-1114.
- [8] 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: ACM, 2020: 1122-1131.
- [9] Tsai Y H, Bai Shaojie, 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: Association for Computational Linguistics, 2019: 6558-6569.
- [10] Sun Licai, Lian Zheng, Liu Bin, et al. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis [J]. IEEE Transactions on Affective Computing, 2024, 15(1):

- 309-325.
- [11] Wu Daiqing, Yang Dongbao, Shen Huawen, et al. Resolving sentiment discrepancy for multimodal sentiment detection *via* semantics completion and decomposition[J]. Pattern Recognition, 2026, 172: 112719.
 - [12] He Ping, Qi Huaying, Wang Shiyi, et al. Cross-modal sentiment analysis of text and video based on Bi-GRU cyclic network and correlation enhancement[J]. Applied Sciences, 2023, 13(13): 7489.
 - [13] Yang Jianing, Wang Yongxin, Yi Ruitao, et al. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences [C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Virtual Event: Association for Computational Linguistics, 2021: 1009-1021.
 - [14] SijieMai, Xing Songlong, He Jiaxuan, et al. Multimodal graph for unaligned multimodal sequence analysis *via* graph convolution and graph pooling[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(2): 1-24.
 - [15] Sharma Vishal, Rishu Null, Kukreja Vinay, et al. A novel integration of simple framework for contrastive learning of visual representations and data-efficient image transformer for diabetic macular edema diagnosis [J]. Iran Journal of Computer Science, 2025, 8(4): 1615-1645.
 - [16] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//Proceedings of the 38th International Conference on Machine Learning. Cambridge: PMLR, 2021: 8748-8763.
 - [17] Li Junnan, Selvaraju R, Gotmare A, et al. Advances in neural information processing systems [C]//Advances in Neural Information Processing Systems. New York: Curran Associates, Inc., 2021, 34: 9694-9705.
 - [18] 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.
 - [19] Bagher Zadeh A, 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: Association for Computational Linguistics, 2018: 2236-2246.
 - [20] Yu Wenmeng, Xu Hua, Meng Fanyang, et al. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event: Association for Computational Linguistics, 2020: 3718-3727.
 - [21] Liu Zhun, Shen Ying, Lakshminarasimhan V B, et al. Efficient low-rank multimodal fusion with modality-specific factors [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 2018: 2247-2256.
 - [22] 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. Virtual Event: Association for Computational Linguistics, 2020: 2359-2369.
 - [23] Han Wei, Chen Hui, 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. Virtual Event: Association for Computational Linguistics, 2021: 9180-9192.
 - [24] Zhang Haoyu, Wang Yu, Yin Guanghao, et al. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 756-767.
 - [25] Wang Di, Guo Xutong, Tian Yumin, et al. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis [J]. Pattern Recognition, 2023, 136: 109259.
 - [26] Yu Wenmeng, Xu Hua, Yuan Ziqi, 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. Washington, DC: AAAI Press, 2021: 10790-10797.
 - [27] Ma Feipeng, Zhang Yueyi, Sun Xiaoyan. Multimodal sentiment analysis with preferential fusion and distance-aware contrastive learning [C]//2023 IEEE International Conference on Multimedia and Expo (ICME). Brisbane: IEEE, 2023: 1367-1372.