软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007628] [CSTR: 32375.14.jos.007628]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

融合对抗学习与多视图网络的鲁棒多模态情感分析*

蔡林沁, 刘道宏, 刘岚睿, 牟仁云

(重庆邮电大学 自动化学院/工业互联网学院, 重庆 400065)

通信作者: 蔡林沁, E-mail: cailq@cqupt.edu.cn

摘　要: 提高模型对特征噪声的鲁棒性已经成为多模态情感分析 (multimodal sentiment analysis, MSA) 领域最具挑战性的任务之一, 最近的研究已经提出了考虑缺失模态的高效 MSA 模型. 然而, 现有方法通常只关注特定类型的缺陷, 导致在同时存在多种类型噪声的实际场景中模型鲁棒性和泛化性弱, 而且现有研究忽略了不同模态之间的深层交互, 难以全面理解多模态情感信息. 针对这些问题, 提出一种融合对抗学习与多视图网络的鲁棒多模态情感分析网络 (robust multimodal sentiment analysis integrating adversarial learning and multi-view network, ALMV). 具体而言, 首先使用时间模态特征缺失作为噪声数据, 与完整序列构成噪声-原始实例对以构建增强数据. 其次, 通过时间卷积和 Transformer 编码器提取噪声-原始实例对中每个模态序列的局部信息和全局信息, 并构建权重调控的多视图网络来学习噪声-原始实例对之间的多模态联合表示. 然后, 提出一种具有语义重建监督的多重对抗训练策略, 在模态级和话语级两个层面学习噪声和完整数据之间的统一联合表征. 最后, 在公开数据集上进行了大量实验, 对多种类型噪声场景下 ALMV 的性能进行测试. 实验结果证明, ALMV 方法有效地提升了多模态情感分析在多种异构数据缺陷场景下的性能和鲁棒性.

关键词: 鲁棒多模态情感分析; 对抗学习; 语义重建; 多视图网络

中图法分类号: TP18

中文引用格式: 蔡林沁, 刘道宏, 刘岚睿, 牟仁云. 融合对抗学习与多视图网络的鲁棒多模态情感分析. 软件学报. http://www.jos.org.cn/1000-9825/7628.htm

英文引用格式: Cai LQ, Liu DH, Liu LR, Mou RY. Robust Multimodal Sentiment Analysis Integrating Adversarial Learning and Multi-view Network. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7628.htm

Robust Multimodal Sentiment Analysis Integrating Adversarial Learning and Multi-view Network

CAI Lin-Qin, LIU Dao-Hong, LIU Lan-Rui, MOU Ren-Yun

(School of Automation/School of Industrial Internet, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Improving robustness to feature noise is one of the most challenging problems in multimodal sentiment analysis (MSA). Recent studies have proposed efficient MSA models that consider missing modalities, but they typically focus on specific types of defects, which leads to limited robustness and generalization in real-world scenarios where multiple types of noise coexist. In addition, deep interactions across different modalities are often inadequately modeled, making it difficult to fully capture multimodal emotional semantics. To address these issues, this study proposes a robust MSA method that integrates adversarial learning with a multi-view network, namely ALMV. Specifically, temporal modality feature masking is adopted to simulate noisy data, and noise-original instance pairs are constructed with intact sequences for data augmentation. Secondly, temporal convolutional networks and Transformer encoders are used to extract local and global information from every modality sequence, and a weight-controlled multi-view network is constructed to learn joint multimodal representations from the noise-original instance pairs. Additionally, a novel multi-level adversarial training strategy with semantic reconstruction supervision is introduced to learn unified representations between noisy and complete data at both the modality level and the

* 基金项目: 国家自然科学基金 (62277008); 重庆市自然科学基金创新发展联合基金 (重点) 项目 (CSTB2024NSCQ-LZX0133)
收稿时间: 2025-07-07; 修改时间: 2025-10-19; 采用时间: 2025-12-29; jos 在线出版时间: 2026-07-08

utterance level. Extensive experiments on public datasets are conducted to verify the performance of ALMV under various heterogeneous noise scenarios. Experimental results demonstrate that ALMV substantially improves the robustness and performance of multimodal sentiment analysis in the presence of diverse data defects.

Key words: robust multimodal sentiment analysis (MSA); adversarial learning; semantic reconstruction; multi-view network (MVN)

多模态情感分析旨在从各种模态中提取有意义的情感特征, 并分析人类的情感倾向和态度. 由于其适用于广泛的领域, 包括智慧教育[1,2]、智能人机交互[3]、医疗保健[4]、社交媒体分析[5]、群体决策系统[6]和营销管理[7,8], 因此吸引了广泛的研究兴趣. 例如, 在智慧教育的领域中, 多模态情感分析在互动式课堂, 在线网络教育等方面有着广泛的应用场景. 通过多模态情感分析, 教师根据学生的情绪波动和参与度调整教学策略, 鼓励学生更多地参与讨论[1,2]. 在虚拟学习环境人机交互领域[3,9], 通过融合面部表情、语音和手势信号来实时感知学习者的情绪变化, 从而动态调整教学内容或交互方式, 提升沉浸感与学习效果. MSA 主要涉及 3 种常见模态的时序数据, 即: 音频 (声学行为)、视觉 (面部表情) 和语言 (口头语言). 这些不同类型的数据为我们深入了解说话人的情感提供了丰富的信息.

最近的研究通过整合来自不同模态的信息以改善模态内和模态间的相互作用, 在 MSA 方面取得了重要进展, 其中如何有效地捕获情感信息以学习有意义的表示是一个关键的方面. 现有方法主要通过采用复杂的交互和融合机制来整合不同模态的信息[10,11], 包括基于注意力的方法[12,13]、基于张量的方法[14,15]和基于机器翻译的方法[16]. 另有一些方法探索模态的权重调控机制, 如罗渊贻等人[17]通过不同模态的学习梯度改变模态的权值动态融合. 虽然这些方法已经取得了较好的性能, 但它们无法捕获各种交互状态下的多视图线索. 为此, Tang 等人[18]提出了多视图交互式表征模型 MVIR (multi-view interactive representations), 专注于学习在不同的交互状态下的多视图交互式表征, 从不同角度全面理解情感信息, 增强 MSA 整体分析过程和性能. 但是, MVIR 主要探索单视图和双视图的交互, 对三视图的探索不足, 这可能会忽视多模态的内在关联和交互信息, 难以分析说话者微妙而有价值的情感暗示.

同时, 这些方法都假定在训练和推理阶段所有模态都可用. 但在实际场景中, 由于不可避免的因素, 多模态数据往往会出现模态不确定性缺失的情况. 例如, 语音可能会因背景噪声或传感器故障而暂时中断; 说话人的面部表情可能会因遮挡或移动而无法捕获; 自动语音识别错误可能会导致某些口语单词不可用[19,20]. 在基于完整模态训练的 MSA 任务中, 每种模态都传达了不同的信息, 不同模态之间存在着互补关系. 当某些模态缺失时, 系统只能依赖剩余模态的信息, 这可能导致 MSA 准确度下降[21], 严重影响 MSA 模型的有效性. 如图 1 所示, 在完整模态下, 根据话语信息, 可以预测出“Positive”的正确结果. 而由于一些潜在的原因导致视觉、语言和音频模态中的一些包含与情感相关联的关键信息的帧级特征缺失, 这会让模型倾向于预测“Neutral”的错误情感结果.

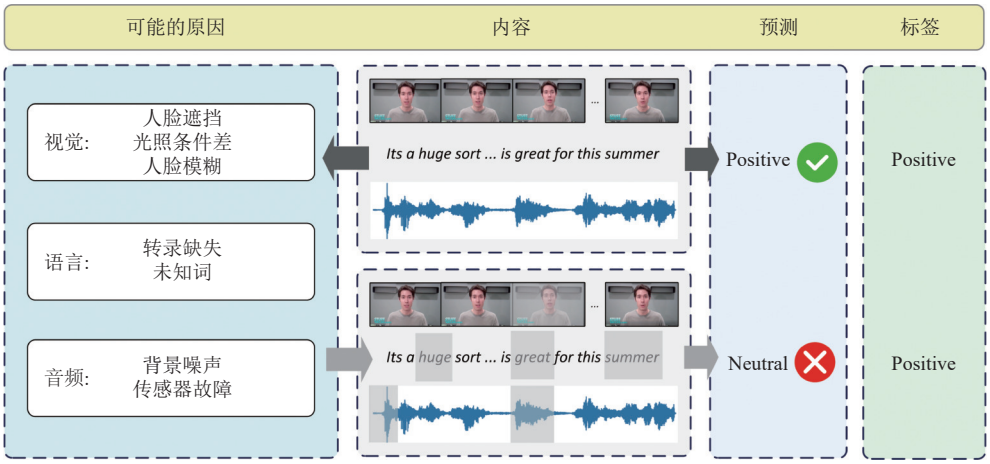


图 1 不完整模态预测情感示例

解决模态缺失问题的一个简单方法是在模型推理过程中执行零、平均或最近邻填充[22]. 然而, 在基准实况和估算实况之间存在很大的域差距. 因此, 使用这些朴素插补时, 模型性能通常会严重下降. 为解决噪声干扰, Meta-

NA[23]从元学习的角度提出了一种元噪声自适应方法. 在元训练时段期间获取用于所有潜在类型的特征噪声的共享知识, 在测试时使用具有目标噪声模式的实例来进一步细化. 但元学习的方法针对小样本数据, 在大范围的数据集上作用有限. TFRNet[24]提出了基于 Transformer 编码器-解码器结构的特征重建网络来解决随机模态特征缺失, 期望通过重建缺失的部分从不完整的多模态序列中隐式地学习语义特征. 虽然取得了有希望的结果, 但 TFRNet 仅在模态级重构单模态特征, 有恢复情感无关信息的风险, 同时它仅针对随机缺失这一种缺陷. NITA[25]提出了基于噪声感知的对抗训练和话语级语义重建, 以缩小原始数据和噪声数据之间的表示差距, 但这种方法仅在话语级恢复多模态联合表示, 忽视了模态特定信号的补充作用.

为克服上述问题, 本文提出一种融合噪声对抗学习与多视图网络的鲁棒多模态情感分析网络 ALMV (robust multimodal sentiment analysis integrating adversarial learning and multi-view network), 对模态缺失场景下的噪声具有较强的鲁棒性. 具体而言, 首先使用时间模态特征缺失作为噪声数据, 与完整模态数据构成噪声-原始实例对以构建增强数据. 然后, 我们通过时间卷积和 Transformer 编码器提取噪声-原始实例对的每个模态序列的局部信息和全局信息, 并使用一种权重调控的多视图网络学习噪声-原始实例对之间的多模态联合表示, 从而充分探索多视图特征的情感线索. 在此基础上, 基于成对的中间表示, 提出一种新的具有语义重建监督的多重对抗训练策略, 在模态级和话语级两个层面学习噪声和完整数据之间的统一联合表示, 以恢复完整的模态特定信号和话语信号.

本文的主要贡献可以概括为 3 点.

(1) 提出一种基于噪声注入的对抗学习训练框架, 该框架集成了噪声注入的对抗训练和语义重建, 在模态级和话语级两个维度缩小原始数据对和噪声数据对之间的表示差距, 并进一步促进鲁棒的表示学习.

(2) 提出一种权重调控的多视图网络, 显式地建模单视图、双视图和三视图特征, 并通过视图特征顶点的权重动态地调控多视图特征的重要性, 从而充分探索多模态特征的情感线索.

(3) 在两个流行的基准 MSA 数据集上的大量实验表明, 所提出的 ALMV 方法在所有 4 种形式的异构数据缺陷的鲁棒性测试中, 都取得了最先进的性能.

1 相关工作

1.1 鲁棒多模态情感分析

在理想的情况下, MSA 所有的模态都是完整的, 但是现实场景中, 可能存在模态缺失或者噪声混入使得多模态数据呈现出不完整性特征, 致使基于完整数据设计的传统 MSA 方法对不完整数据较为敏感. 近年来针对潜在数据不完整性的鲁棒多模态情感分析引起了越来越广泛的关注[26]. 本节总结了在推理期间针对典型类型的数据不完整性的最新工作, 包括模态缺失和模态特征缺失两大类.

模态缺失源于传统多模态分析中的模态消融研究, 是 MSA 最受关注的一种缺陷类型. 一个简单而有效的方法是模态缺失填补. Tsai 等人[27]使用表征分裂并基于学习的多模态判别和模态特定生成因子来估算缺失模态. Ma 等人[28]通过添加具有学习的高斯分布权重的聚类中心向量来估算缺失模态. 最近, Han 等人[29]通过对齐矩阵进一步增强了表征插补. 另一种范式涉及模态转换[16,30], 其利用源和目标模态之间的中间表示作为联合多模态特征. Liang 等人[31]提出了一种张量秩正则化方法, 从对齐的不完美时间序列数据中学习多模态表征. 王楠等人[32]提出一种基于知识蒸馏的方法, 通过教师模型对学生模型施加约束. Pham 等人[33]利用顺序循环平移来学习鲁棒的联合表达, 这可能不需要在推理时将所有模态作为输入. 此外, Zhao 等人[34]采用基于级联残差自动编码器的数据增强和循环平移来解决不确定的缺失模态问题. Guo 等人[35]为缺失模态引入生成提示来促进模型的鲁棒性. Reza 等人[36]提出了一种简单且参数有效的预训练多模态网络自适应方法, 利用调制的中间特征弥补缺失的模态. 然而, 基于缺失模态估算的方法和基于模态估计的方法都需要知道哪些条目或模态在推断期间是不完整的, 并且在更复杂的情况, 如不完整的信号存在于帧粒级而不是模态级[19,37]的情况下容易遭遇失败.

模态特征缺失是 MSA 的一种更普遍的缺陷类型, 其理想地将细粒度的多模态扰动公式化为模态序列中缺失的特征. 最早文献 [31] 基于对干净数据的低秩性质的观察, 提出了一种低秩正则化方法. Li 等人[38]利用跨时间域的数据动态来改进模型, 从而降低内存资源的消耗. 近年来, 已经出现了与辅助特征重建相关的方法. Yuan 等人[24]

提出了一种基于 Transformer 的特征重构网络. Sun 等人[39]提出了一种双层特征恢复方法, 即对低维特征重建, 对高级特征吸引以增强鲁棒性能. Meta-NA[23]提出了元噪声自适应学习策略, 在元训练期间获取用于所有潜在类型的特征噪声的共享知识, 并且在元测试期间使用具有目标噪声模式的实例来进一步细化.

1.2 对抗表征学习

对抗表征学习起源于生成对抗网络[40], 是一种新兴的技术, 可以将分布显式匹配到任意先验分布[41,42]. 最近, 对抗性学习策略进一步扩展到多模态领域, 如文本到图像合成[43,44]. 在多模态情感分析领域, 研究人员利用对抗性训练来学习每种模态的区分和生成表示, 从而提高模型性能[27]. Mai 等人[45]引入了对抗训练为各种模态开发一个有区别的联合嵌入空间. Yuan 等人[25]提出了一种基于噪声模仿的对抗训练框架 NIAT, 将结构时间特征擦除策略引入噪声增强, 并在话语级多模态融合特征上进行对抗训练和语义重构以指导对噪声数据的鲁棒表征学习. 然而, NIAT 仅在话语级重构融合后的多模态完整信号, 实际上对噪声信号的恢复不充分, 这可能会带来次优的结果.

1.3 多视图学习

多视图学习 (multi-view learning, MVL) 旨在利用同一对象的多种视图 (view) 或表征信息来提升模型的性能与鲁棒性, 它利用来自多个视角或模态的信息来增强预测性能并更深入地了解复杂数据[46]. 在许多现实世界的场景中, 例如图像识别, 生物信息学和社交网络分析, 仅仅依靠数据的单个视图可能无法提供足够的信息进行准确的分析和预测[47,48]. MVL 的核心概念是利用不同视图中的互补信息[49], 与传统的单视图方法相比, 可以提高学习效果. 通过整合多个视图, 我们可以捕捉不同信息, 发现隐藏的模式, 并克服单个视图的局限性[46]. 在 MVL 领域, 采用了各种策略, 包括将多个视图合并为单个表示的早期融合技术, 以及将来自各个视图的联合预测组合在一起的后期融合方法[50]. 在多模态情感分析领域, Tang 等人[18]提出了一个新的框架, 专注于学习在不同交互状态下的多视图交互式表征 (MVIR). 该框架允许从单视图和双视图的各个角度全面理解情感信息, 从而增强整体分析过程. 然而该方法对三视图的探索不足, 这可能会忽视多模态的内在关联和交互信息.

2 方 法

本文融合对抗学习的多视图网络 (ALMV) 整体架构如图 2 所示. 这是一种基于噪声注入的多重对抗学习框架的多视图融合网络模型, 旨在解决存在不完整数据的情况下使用语言, 音频和视觉模态预测情绪. 为了增强鲁棒性, 该模型采用了对抗性训练机制. 重建网络负责在噪声条件下在模态级和话语级两个层面恢复模态特定语义信息和多模态融合特征语义信息, 而多个判别网络通过对抗学习提取噪声不变特征. 多视图网络用于融合模态特定特征, 从而获得更具鲁棒性的融合表征, 确保噪声和干净数据之间在话语级别的情感一致性.

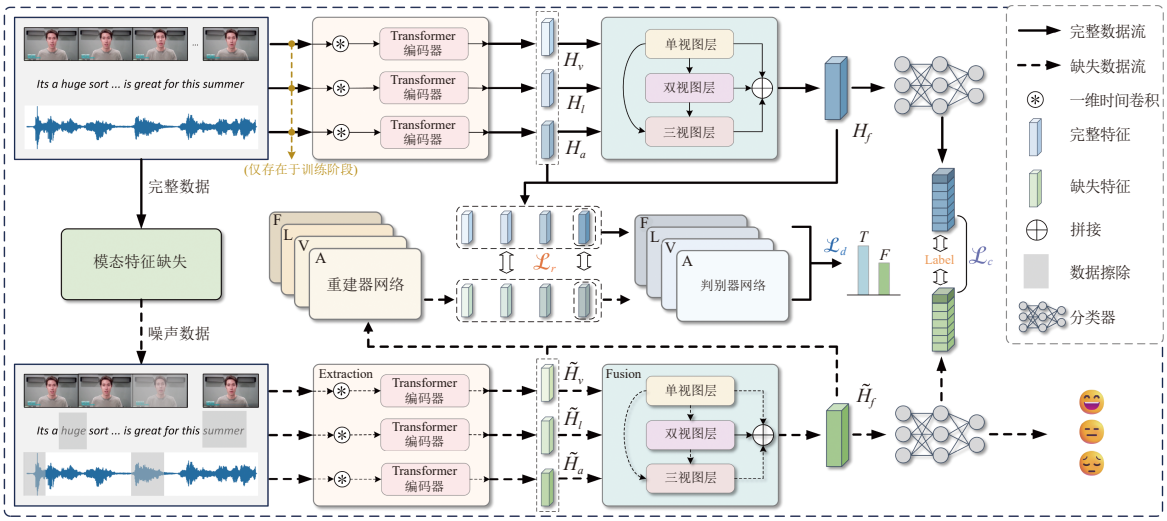


图 2 ALMV 总体框架图

2.1 任务设置

多模态情感分析是利用同一视频片段的音频 (a), 视觉 (v) 和语言 (l) 这 3 种模态来预测情感, 其表达式为 $I_m \in \mathbb{R}^{T_m \times d_m}$, 其中 T_m 表示序列长度, d_m 表示每种模态的维数, $m \in \{a,v,l\}$. 鲁棒多模态情感分析是在现实生活中可能遇到的潜在的模态数据缺失的场景下, 利用缺失的多模态信号 $\tilde{I}_m$ 来预测说话人的情感. 本文旨在开发鲁棒的 MSA 模型, 可以有效地整合所有可用的多模态信息, 在不完整的模态设置中准确预测情感强度得分 $\tilde{y} \in \mathbb{R}$. 在训练期间, 由于手工数据收集, 我们假设来自语言、音频和视觉模态的完整模态序列被提供有对应的情感注释 y. 在测试阶段, 利用未知扰动的实例对训练好的模型进行评估, 以验证模型对数据缺陷的有效性和鲁棒性.

2.2 基于噪声注入的数据增强

为了增加模型对缺失模态数据的鲁棒性, 使用时间特征擦除用于模仿推断同一时段中的潜在缺陷, 具体地, 给定具有原始模态序列 $I_m = [U_l; U_a; U_v]$ 的完整数据序列, 随机选择同一帧的时间步长, 并且同时擦除在这些时间步长的所有模态特征.

$$\tilde{I}_m = F(I_m, g_m) \in \mathbb{R}^{T_m \times d_m} \tag{1}$$

其中, $\tilde{I}_m = [\tilde{U}_l; \tilde{X}_a; \tilde{X}_v]$ 表示对应的缺失模态的噪声序列, $F(\cdot)$ 是掩码函数, $g_m \in \{0,1\}^{T_m}$ 是要遮掩的帧的随机时间掩码, T_m 表示序列长度, d_m 表示每种模态的维数. 具体来说, 对于语言模态, 将缺失的特征设为 BERT 词汇表中的 [UNK] 以模拟潜在的翻译错误. 对于视觉和听觉模态, 缺失特征被设置为零填充向量. 时间特征擦除以对齐的方式进行以引导模型, 而不是直接利用来自其他模态的同步信息, 而是集中于噪声实例中剩余的承载情感的部分.

然后, 使用预训练的 BERT 模型[51]对语言序列进行处理:

$$X_l = BERT(U_l),\ \tilde{X}_l = BERT(\tilde{U}_l) \in \mathbb{R}^{T_m \times d_m} \tag{2}$$

将完整序列 I_m=[X_l; X_a; X_v] 和噪声序列 $\tilde{I}_m = [\tilde{X}_l; \tilde{X}_a; \tilde{X}_v]$ 输入 ALMV 模型进行对抗训练.

2.3 特征提取网络

为了提取模态特征的局部信息和全局信息, 采用由一维时间卷积和 Transformer 两部分组成的特征提取网络, 如图 2. 首先使用模态特定的一维卷积 (1D Conv) 模块对原始数据和噪声数据进行卷积操作. 该步骤通过滑动大小为 k_m 的卷积核, 在时序维度上提取局部特征, 能够捕捉相邻帧/词之间的短期模式并实现初步的维度压缩, 从而去除冗余信息并突出局部判别信号.

$$h_m = Conv1d(I_m, k_m) \in \mathbb{R}^{T \times d_m} \tag{3}$$

$$\tilde{h}_m = Conv1d(\tilde{I}_m, k_m) \in \mathbb{R}^{T \times d_m} \tag{4}$$

其中, $m \in \{a,v,l\}$, k_m 指的是内核大小, 而 d_m 是模态 m 的隐藏维数. 然后在时间维度上将卷积序列连接以获得隐藏序列表示. 更进一步, 将一维卷积的输出送入 *Transformer*, 采用多头自注意力机制, 在全序列范围内建立各时刻之间的关联, 学习长距离依赖. 对于输入序列 $h_m \in \mathbb{R}^{T \times d_m}$, 我们定义查询为 $Q_m = h_m W_Q$, 键为 $K_m = h_m W_K$, 值为 $V_m = h_m W_V$, 其中 $W_Q \in \mathbb{R}^{d_m \times d_q}$, $W_K \in \mathbb{R}^{d_m \times d_k}$, $W_V \in \mathbb{R}^{d_m \times d_v}$, 从而自注意力可以被公式化为:

$$Attention(h_m) = Softmax\left(\frac{Q_m K_m^{\mathrm{T}}}{\sqrt{d_k}}\right) V_m = Softmax\left(\frac{h_m W_Q W_K^{\mathrm{T}} h_m^{\mathrm{T}}}{\sqrt{d_k}}\right) h_m W_V \tag{5}$$

我们使用含自注意力机制的 Transformer 编码器来从噪声-原始数据的不同语义空间中学习重要的全局语义信息, 得到特征提取后的特征:

$$H_m = Transformer(h_m) \in \mathbb{R}^{T \times d_m} \tag{6}$$

$$\tilde{H}_m = Transformer(\tilde{h}_m) \in \mathbb{R}^{T \times d_m} \tag{7}$$

从而, 特征提取网络既保留了局部的细粒度特征, 又通过全局注意力获得了对序列整体上下文的深层理解.

2.4 多视图网络

假设多模态融合处于多个阶段, 并考虑到需要保留所有视角模态的相互作用, 我们引入了一种多视图网络 (multi-view network, MVN), 通过逐级建模单视图、双视图和三视图特征的相互作用来捕获跨模态交互, 如图 3 所

示. MVN 将每个视图编码记为顶点 m, 并动态计算顶点间的相似性 $S_{m_i m_j}$ 和重要性 $\alpha_{m_i m_j}$, 从而使模型结构具有高度的可解释性和灵活性.

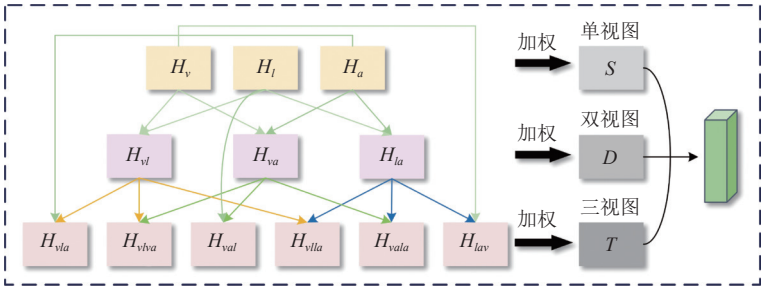


图 3 多视图网络结构

● 权重调控机制: 多视图权重调控的核心逻辑源于信息论中的互信息 (mutual information, MI). 设模态 m_1 和 m_2 的特征分布为 $P(H_{m_1})$ 和 $P(H_{m_2})$, 则两模态的互信息为:

$$MI(H_{m_1};H_{m_2})=\mathbb{E}_{P(H_{m_1},H_{m_2})}\left[\log\frac{P(H_{m_1},H_{m_2})}{P(H_{m_1})\cdot P(H_{m_2})}\right] \tag{8}$$

其中, $P(H_{m_1},H_{m_2})$ 是联合概率, $P(H_{m_1})$ 和 $P(H_{m_2})$ 为边缘概率, $\mathbb{E}_P[\cdot]$ 为联合概率分布 $P(H_{m_1},H_{m_2})$ 的数学期望. 互信息 $MI(H_{m_1};H_{m_2})$ 越大, 说明两模态特征的信息重叠度越高, 此时双视图交互仅能获取少量新增信息, 信息增益有限; 互信息越小, 说明两模态特征的差异度越大, 双视图交互能整合不同维度的有效信息, 信息增益显著. 因此, 双视图交互的权重应与互信息呈反比, 即通过互信息量化模态冗余性, 实现权重的自适应调控.

多视图网络的权重调控机制通过“顶点权重”与“相似度”的协同计算, 实现对模态间互补性的自适应建模. 其中顶点权重确保重要模态在交互中占据主导地位, 相似度 (通过特征内积近似互信息) 则抑制冗余模态的无效交互.

● 单视图层: 由 3 个模态的单模态顶点组成, 3 个模态的信息向量被表示为 H_l、H_a 和 H_v. 为了使网络能够自适应地关注每个样本中最具信息量的单一视图, 首先对 3 类视图分别计算注意力分数, 并确定这些单模态特征相互作用的重要性, 因为并非所有模态都有同等的贡献. 然后, 我们通过计算来自所有单视图顶点的信息的加权平均值来获得最终的单视图特征, 如下所示:

$$\alpha_m=Softmax(\tanh(W_{\text{att}}H_m)) \tag{9}$$

$$S=\frac{1}{3}\sum\nolimits_m \alpha_m\cdot H_m \tag{10}$$

其中, S 是最终的单视图向量, α_m 是模态 m 的权重, $m\in\{a,v,l\}$, $W_{\text{att}}\in\mathbb{R}^{d\times 1}$ 为可学习的线性映射矩阵, tanh 为激活函数, *Softmax* 为归一化函数.

● 双视图层: 使用视图对映射子网络 $F_{\text{map}}\in\mathbb{R}^{2d}\to\mathbb{R}^{d}$, 融合每两路视图, 提取其潜在的协同表征以获得每个双视图顶点:

$$F_{\text{map}}(z)=\tanh(W_2(LeakyReLU(W_1 z))),\ z\in\mathbb{R}^{2d} \tag{11}$$

$$H_{m_1m_2}=LeakyReLU(\beta_{m_1m_2}\cdot F_{\text{map}}(m_1\oplus m_2)) \tag{12}$$

其中, W_1 和 W_2 为可学习参数矩阵, $\oplus$ 表示特征拼接, $m_1,m_2\in\{a,v,l\}$ 且 $m_1\neq m_2$, $H_{m_1m_2}$ 是 m_1 和 m_2 顶点的双视图向量, *LeakyReLU* 和 tanh 均为激活函数.

对于连接第 1 层和第 2 层的边的权重, 我们首先利用内积估计第 1 层每两个均匀的单模信息向量的相似度. 由信息论可知, 两个信息向量越相似, 它们的双视图相互作用就越不重要, 并且它们的信息在单视图层中已经被很好地探索. 两个信息向量的相似度定义为:

$$S_{m_1m_2}=\hat{H}_{m_1}^{\mathrm{T}}\hat{H}_{m_2} \tag{13}$$

其中, $\hat{H}_{m_1}$ 表示 H_{m_1} 的 *Softmax* 归一化向量 (归一化确保计算的相似度在 0–1 之间), T 表示转置操作, $S_{m_1m_2}$ 是表示

m_1 和 m_2 顶点的相似度的标量. 将连接第 1 层 m_1 顶点和第 2 层 m_1m_2 顶点的边的权值定义为 $\alpha_{m_1}/(S_{m_1m_2}+\epsilon)$, 它与 α_{m_1} 成正比, 与 $S_{m_1m_2}$ 成反比, 然后, m_1m_2 顶点的权重定义为:

$$\breve{\alpha}_{m_1m_2}=\frac{\alpha_{m_1}+\alpha_{m_2}}{S_{m_1m_2}+\epsilon} \tag{14}$$

$$\alpha_{m_1m_2}=\frac{\mathrm{e}^{\breve{\alpha}_{m_1m_2}}}{\sum_{m_{i\neq j}}\mathrm{e}^{\breve{\alpha}_{m_im_j}}} \tag{15}$$

其中, $\alpha_{m_1m_2}$ 是表示 m_1m_2 顶点的权重的标量 (分母上的值 ϵ 被应用于控制相似性和顶点权重对 $\alpha_{m_1m_2}$ 的相对重要性), 公式 (15) 是第 2 层中权向量的软最大一致化运算. 双视图层最终的双视图特征表示 D 可以通过加权获得:

$$D=\sum_{m_1,m_2\in\{a,v,l\},m_1\neq m_2}\alpha_{m_1m_2}\cdot H_{m_1m_2} \tag{16}$$

● 三视图层: 三视图层中每两个双视图顶点使用与生成双视图相同的视图对映射子网络 F_{map} (但参数不共享) 进一步融合, 以获得第 3 层中的三视图特征. 此外, 每个特定的双视图顶点与在先前融合阶段中对该双视图顶点的形成没有贡献的单视图顶点融合, 从而产生 3 个其他三视图顶点. 因此, 共有 6 个三视图特征顶点, 如图 3 所示. 我们对连接到第 3 层的边应用相同的权重计算方法, 对三视图顶点应用与第 2 层相同的重要性度量方法. 然后将每个三视图顶点的加权信息相加, 得到最终的三视图特征表示.

当完成单视图表征 S、双视图表征 D 以及三视图高阶表征 T 后, 将它们在通道维度拼接获得最终的多视图融合特征:

$$f=Concat(S,D,T) \tag{17}$$

其中, f 表示多视图融合特征, $Concat$ 表示拼接操作.

将特征提取网络得到的特征输送到多视图网络中, 得到干净数据和噪声数据各自的融合特征:

$$H_f=MVN(H_m,\theta_{MVN}) \tag{18}$$

$$\tilde{H}_f=MVN(\tilde{H}_m,\theta_{MVN}) \tag{19}$$

其中, H_f 是完整模态的融合特征, $\tilde{H}_f$ 是缺失模态的融合特征, θ_{MVN} 表示多视图网络的可训练参数.

2.5 分类器网络

分类器使用两层全连接前馈网络来实现, 以根据原始噪声数据对的提取的表示 H_f 和 $\tilde{H}_f$ 来预测情感. 我们将单层前馈网络定义为:

$$FFN(x)=\sigma(x\cdot W_f+b_f) \tag{20}$$

其中, σ 表示可选的激活函数, W_f 和 b_f 是可学习的模型参数.

从而, 我们可以将分类器网络公式化为:

$$C_{\theta_C}(H_f)=FFN(FFN(BN(h)))\in\mathbb{R} \tag{21}$$

其中, BN 是 BatchNorm, θ_C 是分类器网络中的可学习参数, h 表示分类器的输入.

对于情感预测监督, 分类损失 $\mathcal{L}_c$ 由完整数据 H_f 和噪声数据 $\tilde{H}_f$ 的数据对的加权和计算:

$$\mathcal{L}_c=\frac{1}{1+\alpha}\cdot L1(y,C_{\theta_C}(H_f))+\frac{\alpha}{1+\alpha}\cdot L1(y,C_{\theta_C}(\tilde{H}_f)) \tag{22}$$

其中, $L1(\cdot)$ 是指 $L1$ 损失, y 是情感标签, 超参数 α 是控制来自完整数据流与噪声数据流的相对权重. 降低 α 使训练更多依赖干净样本, 从而使模型倾向于降低估计偏差, 同时相对削弱由噪声流引入的不变性约束. 相反, 增大 α 有利于通过噪声注入的鲁棒学习增强对扰动的不变表征, 但过高则可能削弱判别能力.

2.6 重建器网络

重建器网络的目的是指导特征提取网络和融合网络在噪声情况下再生丢失的语义, 通过执行模态级重建 (即针对单模态噪声数据 $\tilde{H}_m$, 其中 $m\in\{a,v,l\}$) 和话语级语义重建 (即对噪声多视图融合特征 $\tilde{H}_f$). 从噪声数据流中

接收模态特定表示 $\tilde{H}_m$ 和多模态融合表示 $\tilde{H}_f$, 重建模块由 3 层前馈网络实现:

$$R_{\theta_r}(H_n) = FFN(FFN(FFN(BN(\tilde{H}_n)))) \in \mathbb{R}^d \tag{23}$$

其中, BN 是 BatchNorm, θ_r 是重建模块的可学习参数, $n \in \{a, v, l, f\}$. 重建损失 $\mathcal{L}_r$ 被设计为来自缺失数据的重建特征与原始特征之间的 $L1$ 损失:

$$\mathcal{L}_r^m = L1(H_m, R_{\theta_r^m}(\tilde{H}_m)) \tag{24}$$

$$\mathcal{L}_r^f = L1(H_f, R_{\theta_r^f}(\tilde{H}_f)) \tag{25}$$

$$\mathcal{L}_r = \mathcal{L}_r^f + \sum\nolimits_{m\in\{a,v,l\}} \lambda_1 \mathcal{L}_r^m \tag{26}$$

其中, $\mathcal{L}_r^m$ 是模态级重建损失, $\mathcal{L}_r^f$ 是话语级重建损失, λ_1 是模态级和话语级语义重建的重建损失的平衡系数, $\mathcal{L}_r$ 是整个重建器族的总损失, θ_r 是重建器的参数.

2.7 判别器网络

在对抗训练框架中, 噪声对抗训练策略通过匹配原始-噪声数据对的单模态的模态特定表示和融合级的多视图融合表示分布来学习统一表示. 通过一个二人博弈过程, 使模型学到噪声不敏感的特征. 在该博弈中, 提出的判别器模块 D_{θ_d} 经过训练, 以区分原始干净数据和缺少特征的有噪数据; 而特征提取网络和多视图网络作为对手, 目标是经过训练调整自身参数, 使得带噪样本经模态特定提取后得到的特征 $\tilde{H}_m$ 和多模态融合后得到的特征 $\tilde{H}_f$ 在判别器看来难以区分, 即学习表示以混淆判别器 D_{θ_d}. 在 ALMV 框架中, D_{θ_d} 是一个基本的二元分类器:

$$D_{\theta_d}(h) = \sigma(FFN(FFN(FFN(BN(h))))) \in \mathbb{R}^d \tag{27}$$

其中, σ 是 Sigmoid 函数, BN 是 BatchNorm, θ_d 是可学习参数.

最后, 辅助噪声感知对抗训练可以通过以下最小-最大博弈来描述:

$$\min_{\theta_T^m}\max_{\theta_d^m}\mathcal{L}_d^m = -\frac{1}{N}\sum_{i=1}^{N}\log D_{\theta_d^m}\left(F_{\theta_T^m}(H_n)\right) - \frac{1}{N}\sum_{i=1}^{N}\log\left(1 - D_{\theta_d^m}\left(F_{\theta_T^m}(\tilde{H}_n)\right)\right) \tag{28}$$

$$\min_{\theta_f}\max_{\theta_d^f}\mathcal{L}_d^f = -\frac{1}{N}\sum_{i=1}^{N}\log D_{\theta_d^f}\left(F_{\theta_f}(H_f)\right) - \frac{1}{N}\sum_{i=1}^{N}\log\left(1 - D_{\theta_d^f}\left(F_{\theta_f}(\tilde{H}_f)\right)\right) \tag{29}$$

$$\mathcal{L}_d = \mathcal{L}_d^f + \sum\nolimits_{m\in\{a,v,l\}} \lambda_2 \mathcal{L}_d^m \tag{30}$$

其中, $\mathcal{L}_d^m$ 是模态级对抗损失, $\mathcal{L}_d^f$ 是话语级对抗损失, λ_2 是模态级和话语级的对抗损失的平衡系数, $\mathcal{L}_d$ 是整个判别器族的总损失. θ_d 是判别器的参数, θ_T 是 Transformer 编码器的参数, θ_f 是多视图网络的参数.

2.8 对抗学习优化目标

在提出的 ALMV 框架中, 表示学习过程由 3 种不同的监督指导. 所有原始噪声数据对的平均分类损失 $\mathcal{L}_c' = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_c$ 是多模态情感预测的基本监督. 而所有原始噪声数据对上的用于促使表示对注入噪声不变的平均判别损失 $\mathcal{L}_{\mathrm{d}}$ 和用于保持语义保真性的平均重建损失 $\mathcal{L}_r' = -\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_r$ 被视为鲁棒多模态表示学习的辅助损失.

综合所有目标, 最终的优化目标如下:

$$\min_{\theta_T,\theta_f,\theta_c,\theta_r}\max_{\theta_d}\mathcal{L} = ((1-\beta)\cdot\mathcal{L}_r' + \beta\cdot\mathcal{L}_d) + \mathcal{L}_c' \tag{31}$$

其中, β 是平衡重建损失与判别损失之间博弈强度的超参数. 较大的 β 强化了判别约束, 提升噪声鲁棒性, 但过大的对抗权重会压制细粒度语义信息, 导致部分判别信号丢失; 相反, 较小的 β 使重建器更关注样本级语义重构, 但在分布层面抗噪性能减弱.

3 实验与分析

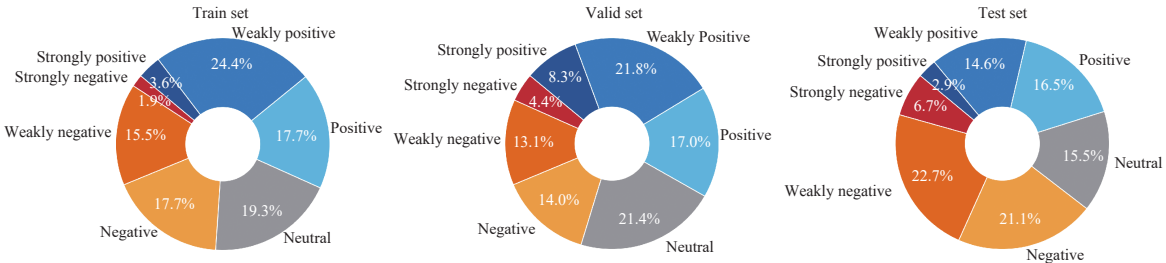
为了验证 ALMV 的性能, 我们在两个流行的数据集上进行了多种类型的实验.

3.1 数据集

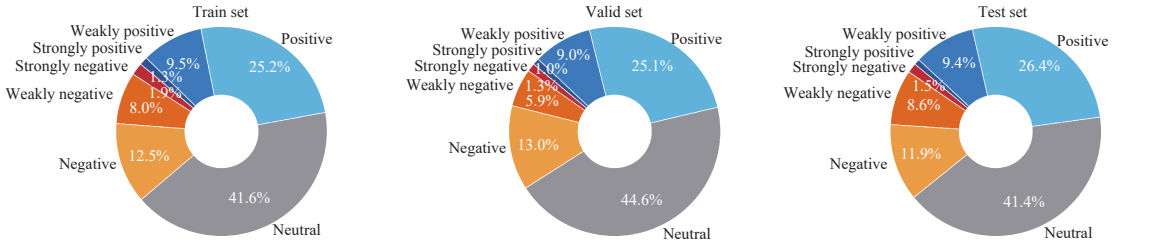
本文在两个基准 MSA 数据集 CMU-MOSI 和 CMU-MOSEI 上进行了实验.

CMU-MOSI[52]: 这个数据集是用英语开发的, 包括音频、文本和视频形式, 共 2 199 个多模态样本, 这些形式是从 YouTube 独白电影评论中收集的 2 199 个带注释的视频片段汇编而成的. 它提供了一个集中的方法来研究电影评论背景下的情感检测. 具体而言, 训练集由 1 284 个样本组成, 验证集包含 229 个样本, 测试集包含 686 个样本. 每个单独的样本被分配一个范围从–3 (指示强烈否定) 到 3 (指示强烈肯定) 的情感分数.

CMU-MOSEI[53]: 它是 CMU-MOSI 的扩展, 结合了来自 YouTube 视频的相同形式的音频、文本和视频, 但它具有更广泛的范围, 涵盖更广泛的主题, 并且规模更大. 该数据集包含从 YouTube 收集的 22 856 个视频片段, 具有多种因素 (例如: 自发的表情、头部姿势、遮挡、照明). 该数据集分为 16 326 个训练实例、1 871 个验证实例和 4 659 个测试实例. 每个实例都被精心标记了一个从–3 到 3 的情感分数. 而情绪得分从–3 到 3 表示最消极到最积极. 两个数据集的详细划分如图 4 所示.



(a) CMU-MOSI 数据集的详细数据划分



(b) CMU-MOSEI 数据集的详细数据划分

图 4 所用数据集的情感划分

3.2 多模态数据预处理

我们利用 BERT[51]、COVAREP[54]和 Facet 作为情感分析中的特征提取工具, 以准确捕捉语言、音频和视觉模态对情感表达的独特贡献. 这些特征在实现情感识别的总体目标中相互补充, 促进更深刻的多模态情感融合.

对于音频: COVAREP 是一个功能强大的音频特征提取工具, 专为语音情感分析和韵律特征提取而设计. 该工具能够高效地从音频信号中提取多种声学特征, 包括基频 (F0)、音量、共振峰、声道特征以及时域特征. 所有特征都经过零均值和方差归一化处理, 对于音频中损坏的帧数不计入. 输入音频数据后, COVAREP 计算出所有的音频特征, 将这些特征堆叠在一起, 最终形成预处理后的音频特征数据.

对于视频: 面部表情和微表情是情感的重要视觉表征. Facet 工具包为每个视觉帧提取面部特征, 包括面部标志, 7 种基本情绪和动作单元, 形成 35 维特征向量. 这些特征精确地表征了面部表情的变化. 它们有助于识别突出的情绪, 揭示情绪强度的细微变化和微表情的微妙变化.

3.3 基线方法

对于模态细粒度特征缺失: 我们选取部分经典的 MSA 方法和部分专门针对缺失模态特征而设计的鲁棒 MSA 方法作为基线方法. 传统的 MSA 基线是第 1 种基线方法. 具体而言, 我们选择了多模态 Transformer (MulT)[11]、

模态不变和特定表示 (MISA)[55]、BERT 的多模态适应门 (MAG-BERT)[12]和自监督多任务多模态情感分析网络 (Self-MM)[56]. 这些经典的 MSA 方法通过复杂的融合策略在完美的测试数据上实现了令人印象深刻的性能. 模态特征缺失的基线是第 2 种基线方法, 包括时态张量融合网络 (T2FN)[31]、时间积融合网络 (TPFN)[38]、基于 Transformer 的特征重建网络 (TFRNet)[24]、基于噪声模仿的对抗训练方法 (NIAT)[25]和基于元学习的噪声自适应方法 (Meta-NA)[23].

对于整个模态缺失的基线方法, 在前面基础上, 我们增加文本增强型 Transformer (TETFN)[57]、耦合翻译融合网络 (CTFN)[16]、多模态分解模型 (MFM)[27]、SMIL[28]、基于模态转换的方法 (Modal-Trans)[30]、MM-Align[29]、基于提示学习的方法 (MPLMM)[35]、自适应训练方法 (Adapted)[36]和文本引导的重建网络 (TgRN)[58].

3.4 验证指标

为了定量评估模型对不同缺失率的鲁棒性, 采用了文献 [24] 中提出的指标折线图下面积 *AUILC* (area under indicators line chart) 度量. 通过考虑对应的模型性能 $X=\{x_0, x_1,\ldots, x_t\}$在递增缺失率序列$\{r_0, r_1,\ldots, r_t\}$下的表现, 并计算折线图上每对相邻点之间的面积之和:

$$AUILC_X = \sum_{i=0}^{t-1} \frac{(x_i + x_{i+1})}{2} \cdot (r_{i+1} - r_i) \tag{32}$$

对于模态特征缺失和其他异质缺陷中的每个预设缺失率, 情感分析预测被公式化为以平均绝对误差 (MAE) 作为度量的回归问题. 此外, 根据以前的工作[11,55], Acc-2 和 *F*1-score 指标被用作负/非负分类标准. 对于所有上述指标, 除 MAE 是指标越低性能越高之外, 其余的指标都是较高的值表示较好的模型性能.

3.5 实验细节

我们在具有 24 GB 内存的 NVIDIA GeForce RTX 4090 GPU 上, 使用 PyTorch 1.11.0 和 CUDA 11.3 环境实现本文模型. CMU-MOSI 和 CMU-MOSEI 上的详细配置如表 1 所示.

表 1　两个 MSA 基准数据集的详细参数配置

超参数	CMU-MOSI	CMU-MOSEI
批大小	128	128
学习率	0.002	0.002
优化器	Adam	Adam
提前停止 (训练轮数)	8	8
卷积核大小 (l, a, v)	(3, 3, 9)	(3, 5, 3)
注意力头数	8	8
Transformer层数	3	3
损失函数权重 ($\alpha, \beta, \lambda_1, \lambda_2$)	(1.0, 0.4, 0.6, 0.5)	(1.0, 0.6, 0.5, 0.5)
融合网络隐藏层维度	50	50
重构器隐藏层维度	(80, 96)	(128, 64)
判别器隐藏层维度	(128, 64)	(80, 32)
融合网络Dropout比例	0.1	0.1
重构器Dropout比例	0.4	0.2
分类器Dropout比例	0.3	0.2
随机种子	[1111, 1112, 1113]	[1111, 1112, 1113]

我们使用对齐版本的 2 个数据集进行实验以实现对同一时间步的模态缺失, 用 Adam 优化器训练. Batchsize 的调整范围为{64, 128, 256}, 模态嵌入层将序列长度统一设定为 50, 将维度统一设置为 128 以方便接下来的处理. ALMV 的卷积层的内核大小 (k_v, k_l, k_a), 对于 CMU-MOSI 被设置为 (9, 3, 3), 并且对于 CMU-MOSEI 被设置为 (3, 3, 5). Transformer 编码器的层数设置为 3 层, 在验证集上执行网格搜索以选择隐藏维度和丢弃率. 对于不同损失的权重, 超参数 α, β 和 λ_1, λ_2 以 0.1 的步长从 0 调整到 1, 平衡每个模块的贡献. 此外, 当验证集上的最佳 MAE 连续

8 个周期未更新时, 采用提前停止策略停止模型训练, 以防止模型过拟合. 本文的所有模型都使用 3 种不同的随机种子进行训练然后取平均值, 以进行公平的比较.

3.6 细粒度模态特征缺失

对于细粒度模态特征缺失, 我们在干净数据、随机模态缺失、时间模态缺失和结构时间模态特征缺失等 4 种场景下做了大量的实验. 对于每个缺失场景, 在缺失率区间{0.0, 0.1,…, 1.0}上验证总体性能. 表 2 和表 3 分别给出了 CMU-MOSI 和 CMU-MOSEI 数据集上的实验结果. 显而易见的是, 干净数据和所有缺失场景之间的明显性能差距, 这表明扰动对现实世界的情感检测是一个不可避免的威胁. 此外, 在 3 种缺失场景中, 所有模型在随机模态缺失和时间模态特征缺失的情况下表现相似 (大多数模型性能变化在 1% 以内). 相比之下, 在结构时间模态缺失的场景下, 所有的方法的表现相较于前两个场景都明显较差. 结果表明, 结构性时间模态缺失更具挑战性, 因为连续序列缺失会阻止模型从附近的模态信号中恢复缺失的语义.

表 2 CMU-MOSI 的细粒度模态特征缺失实验结果

Methods	干净数据	随机模态缺失			时间模态缺失			结构时间模态缺失		
	MAE	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)
MulT[11]	0.881	1.209	65.59	64.39	1.210	65.71	64.45	1.322	60.21	57.58
MISA[55]	0.809	1.233	67.20	64.51	1.234	67.13	65.20	1.426	60.75	56.82
MAG-BERT[12]	0.802	1.316	65.07	64.39	1.319	65.10	63.37	1.528	58.82	54.59
Self-MM[56]	0.790	1.295	67.43	65.11	1.295	67.65	65.41	1.615	60.80	56.37
T2FN[31]	0.890	1.211	65.60	64.76	1.211	64.45	64.63	1.303	61.51	60.75
TPFN[38]	0.896	1.195	65.23	62.67	1.196	65.23	62.67	1.267	61.41	58.58
TFRNet[24]	0.980	1.204	65.83	63.25	1.201	65.99	63.55	1.265	62.34	59.03
NIAT[25]	0.758	1.131	68.02	66.13	1.130	67.95	66.06	1.261	61.99	58.70
Meta-NA[23]	—	1.193	67.14	67.06	—	—	—	—	—	—
ALMV (ours)	**0.749**	**1.122**	**68.87**	**68.16**	**1.123**	**68.78**	**68.05**	**1.224**	**64.08**	**62.63**

注: 加粗数据表示最佳结果

表 3 CMU-MOSEI 的细粒度模态特征缺失实验结果

Methods	干净数据	随机模态缺失			时间模态缺失			结构时间模态缺失		
	MAE	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)
MulT[11]	0.559	0.715	68.67	68.89	0.715	68.52	68.70	0.763	61.35	61.70
MISA[55]	0.571	0.721	73.75	73.09	0.720	73.77	73.09	0.766	71.71	69.69
MAG-BERT[12]	0.536	0.697	74.33	74.11	0.698	73.48	73.55	0.723	70.17	69.97
Self-MM[56]	0.574	0.722	70.39	70.30	0.723	70.40	70.35	0.762	65.43	65.35
T2FN[31]	0.580	0.723	73.27	71.63	0.722	73.31	71.66	0.760	67.72	66.24
TPFN[38]	0.590	0.725	73.78	72.84	0.724	73.71	72.78	0.758	69.73	68.98
TFRNet[24]	0.593	0.725	73.39	71.44	0.724	73.40	71.44	0.756	71.28	67.74
NIAT[25]	0.554	0.690	77.81	75.24	0.690	77.79	75.22	0.735	75.29	71.26
ALMV (ours)	**0.536**	**0.678**	**78.12**	**75.56**	**0.678**	**78.13**	**75.57**	**0.734**	**75.44**	**71.56**

注: 加粗数据表示最佳结果

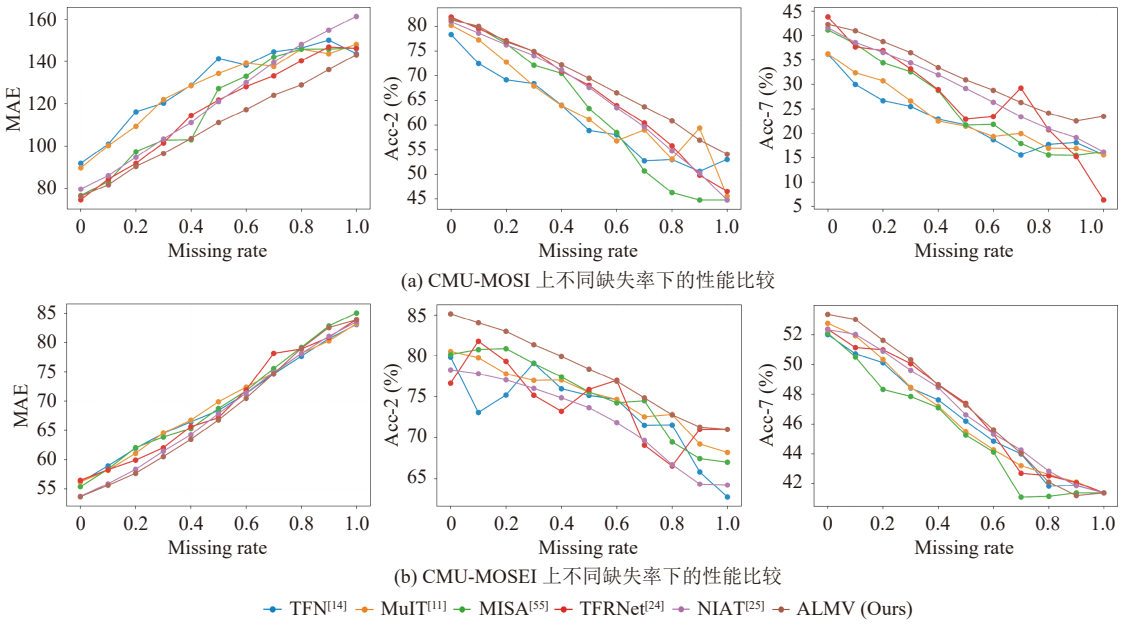
显而易见, ALMV 模型在各种测试场景中都表现出色. 虽然我们的方法是为测试噪声场景的数据而设计的, 但在干净数据测试中, 本方法在两个数据集上都领先其他包括传统的设计用于干净数据的 MSA 方法. 而在所有的 3 种缺失场景下, 本方法在两个数据集的所有指标上都处于领先地位.

具体而言, ALMV 在 CMU-MOSI 表现突出, *F*1 分数方面在随机缺失和时间缺失场景分别大幅领先次优的方法 NIAT 超过约 2%. 究其原因, ALMV 相比于 NIAT[25]在模态级层面和话语级层面都实现了对抗训练, 不仅是恢复最终的多模态融合特征, 同时重建每个单模态的模态特定特征, 这对恢复多模态特征的完整性更有利. 此外值得

一提的是, ALMV 在 CMU-MOSI 的结构时间缺失的场景下, 准确率和 $F1$ 分数分别超过了 64% 和 62%, 显著超过其他方法超过 2% 以上的性能, 这证明了 ALMV 在处理复杂模态缺失和结构缺陷方面的优势. 而在 CMU-MOSEI 上尽管我们的领先优势没有那么大, 但所有的指标也全面领先其他的方法, 特别是 MAE 指标, 我们的方法在干净, 随机和时间缺失上领先超过 1%–2%, 这突出了 ALMV 的泛化性.

3.7 定性分析

接下来, 我们专注于评估 ALMV 在各种缺失率下的鲁棒性. 为了生成不完整的视图, 在训练和测试阶段对 3 种模态应用相同的缺失率. 选择最流行的随机缺失下的噪声场景作为比较, 独立地生成每个模态的随机时间掩模. 在 CMU-MOSI 和 CMU-MOSEI 上的性能比较分别如图 5(a) 和图 5(b) 所示.



(a) CMU-MOSI 上不同缺失率下的性能比较

(b) CMU-MOSEI 上不同缺失率下的性能比较

图 5 不同模型在不同缺失率下的性能表现

为了简单且有说服力, 我们选用 3 个代表性度量 (即 MAE、Acc-2 和 Acc-7). 请注意, CMU-MOSI 和 CMU-MOSEI 上的 Acc-2 是指根据阴性/非阴性样本计算的二进制准确度. 从图 5 中可得到以下观察: 总体上, 随着缺失率的增加, 所有方法的性能都呈准线性下降. 它表明, 多模态序列的随机模态特征丢失将适度降低模型的性能, 这个问题需要得到仔细解决以便在现实世界中部署 MSA 模型. 更重要的是, 在大多数情况下, ALMV 优于所有其他比较方法. 具体而言, 在 CMU-MOSI 数据集, 我们的方法在缺失率区间{0.0, 0.1,…, 1.0}下, 几乎所有的指标都好于其他方法. 而在 CMU-MOSEI 数据集, ALMV 的相对于其他方法的优势主要体现在{0.0, 0.1,…, 0.6}的低缺失率的区间上, 而这也是现实中大部分模态缺失的情况, 我们归因于 CMU-MOSEI 的数据量和场景复杂度相比 CMU-MOSI 要高很多, 在高缺失率下对缺失模态特征的重建会更加困难. 此外, 由图 5 所示, 其中的 MulT[11]和 TFRNet[24]等方法在缺失率变化时, 性能起伏波动很大, 而 ALMV 随着缺失率的上升, 几乎呈现线性的变化, 这体现出 ALMV 模型的鲁棒性.

3.8 整体模态缺失实验比较

在此设置中, 语言、音频或视觉模态序列之一在推理过程中被完全删除. 表 4 和表 5 显示了 ALMV 模型在 CMU-MOSI 和 CMU-MOSEI 数据集上与基线方法的比较. 总体来看, 所提 ALMV 方法在大部分指标下, 尤其是在音频模态缺失和视觉模态缺失等噪声场景下都实现了最佳的性能. 具体来说, 在 CMU-MOSI 数据集上, ALMV 方法在音频缺失场景下达到了最佳性能, 在视觉缺失下达到了与最佳基线相当的性能. 而在 CMU-MOSEI 数据集上,

ALMV 方法在几乎所有的缺失场景都能实现最佳的指标. 具体来说, 在语言缺失情况, ALMV 在准确率方面达到了最佳性能. 值得一提的是, 在音频缺失和视觉缺失的 MAE 指标分别达到了 0.537 和 0.539, 高于排名第 2 的方法 NIAT[25]约 2%. 这得益于 ALMV 在模态级和话语级两个层面上对单模态特征和多模态融合特征的语义重建作用, 尤其是在音频或视觉模态序列之一缺失时, 重建器网络可以帮助多视图网络在话语级层面上从语义信息更高的语言特征中学习到情感知识, 从而恢复多模态融合特征中的音频或视觉信号.

表 4 CMU-MOSI 的整体模态缺失性能比较

Methods	语言模态缺失			音频模态缺失			视觉模态缺失		
	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)
MISA[a][55]	1.472	54.72	54.33	1.007	78.00	78.11	0.999	77.29	77.35
MulT[a][11]	1.384	58.89	58.80	0.923	79.22	79.18	0.953	78.51	78.5
TETFN[a][57]	1.453	52.18	43.04	0.916	80.34	80.33	0.941	79.78	79.80
TPFN[a][38]	1.426	54.07	51.50	0.923	77.54	77.55	0.934	78.05	78.10
T2FN[a][31]	1.393	57.47	57.35	0.906	80.34	80.27	0.900	79.47	79.45
CTFN[a][16]	1.417	54.53	45.56	0.940	79.46	79.45	0.926	80.21	80.19
TFRNet[a][24]	1.410	54.11	52.72	0.946	78.10	78.16	0.905	79.47	79.52
Meta-NA[a][23]	**1.357**	59.50	59.63	0.886	81.71	81.53	0.874	81.00	80.78
MPLMM[35]	—	**65.02**	**65.41**	—	81.12	81.19	—	80.76	81.09
Adapted[36]	—	55.49	53.96	—	—	—	—	—	—
TgRN[58]	1.426	51.85	57.99	0.756	81.44	81.50	0.762	**82.17**	**82.23**
ALMV (ours)	1.399	58.89	59.95	**0.751**	**81.92**	**81.79**	**0.756**	81.78	81.69

注: 标记a的数据来自文献[23], 其余的数据来自原始论文; 加粗数据表示最佳结果

表 5 CMU-MOSEI 的整体模态缺失性能比较

Methods	语言模态缺失			音频模态缺失			视觉模态缺失		
	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)
T2FN[b][31]	0.851	51.93	47.85	0.583	82.25	82.30	0.599	82.11	82.13
TPFN[b][38]	0.828	62.12	59.93	0.614	79.61	79.95	0.593	80.66	80.84
TFRNet[b][24]	0.867	64.91	58.99	0.589	81.34	81.28	0.626	79.91	80.02
MFM[b][27]	0.821	62.00	—	0.658	79.10	—	0.658	79.20	—
SMIL[b][28]	0.820	63.10	—	0.684	78.50	—	0.680	78.30	—
Modal-Trans[b][30]	0.817	65.10	—	0.643	79.90	—	0.645	79.60	—
MM-Align[b][29]	**0.811**	66.20	—	0.635	81.00	—	0.637	80.80	—
NIAT[b][25]	0.836	70.19	58.99	0.556	84.42	84.23	0.562	83.99	83.96
MPLMM[35]	—	68.21	**69.91**	—	80.11	80.13	—	80.45	80.43
Adapted[36]	—	63.32	60.69	—	—	—	—	—	—
ALMV (ours)	0.838	**70.61**	59.47	**0.537**	**84.71**	**84.59**	**0.539**	**84.83**	**84.63**

注: 标记b的数据来自文献[25], 其余的数据来自原始论文; 加粗数据表示最佳结果

同样我们观察到, 语言模态缺失情况下, Meta-NA[23]和 MPLMM[35]取得了最好的效果, 在视觉模态缺失的情况下也取得了和 ALMV 相当的性能. 我们分析原因是因为 Meta-NA[23]在元训练期间获取潜在类型特征噪声的共享知识, 在元测试期间进一步微调模型以获得当前噪声模式的最优解, 这种基于元学习的方法对 CMU-MOSI 这样的小样本数据效果显著, 故在 3 种模态缺失下都展现出富有竞争力的结果. 而 MPLMM[35]为缺失模态引入生成性提示、遗漏信号提示和遗漏类型提示, 这种来自 NLP 领域的通过设计提示指令引导模型完成特定任务的方法对恢复自然语言信号更加有利, 所以在缺失语言模态情况下效果更加突出. 文本引导的重建网络 TgRN[58]在特征提取过程中使用文本增强视听模态特征, 进而通过重建模块恢复缺失特征, 这种方法对于音频缺失和视觉缺失的场景

起到了显著的效果, 但语言模态缺失会极大地影响该方法的性能.

而我们提出的 ALMV 方法通过时间模态特征缺失作数据增强, 使用 Transformer 和多视图网络作为骨干网络, 辅以语义重建监督的多重对抗训练策略. 在不同模态缺失场景下都取得了最先进或是相当的性能, 对多种类型噪声的实际场景中到了新的平衡性和泛化性. 尽管 ALMV 的结果非常突出, 但在语言模态缺失的情况下, ALMV 的表现比非言语模态的表现下降得更多, 这种现象可能是由于语言模态缺失与其他常见噪声模态之间存在显著差异, 这突出了语言模态在 MSA 任务中的主导地位.

3.9 消融实验

表 6 列出了 CMU-MOSI 和 CMU-MOSEI 数据集上随机模态特征缺失和结构时间模态缺失的消融研究结果, 以充分研究 ALMV 每个组件的有效性.

表 6 ALMV 不同组件的消融实验

Model	CMU-MOSI						CMU-MOSEI					
	随机模态缺失			结构时间模态缺失			随机模态缺失			结构时间模态缺失		
	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)	MAE	Acc-2 (%)	*F*1 (%)
w/o Rec	1.177	67.99	67.52	1.291	62.47	61.21	0.726	74.42	72.83	0.751	73.16	71.13
w/o Dis	1.214	68.10	67.67	1.328	61.06	57.96	0.691	73.89	72.1	**0.729**	69.91	67.69
w/o Rec & Dis	1.215	66.52	64.49	1.334	60.37	56.33	0.687	74.06	73.27	0.733	68.98	68.02
w/o Adver-M	1.193	66.36	64.26	1.290	61.25	57.41	0.684	73.86	73.55	0.739	68.29	67.05
w/o Adver-F	1.182	67.14	64.85	1.312	61.47	56.48	0.684	72.38	71.68	0.737	63.96	62.42
w/o TransEnc	1.177	67.51	66.39	1.308	62.73	60.36	0.680	68.99	67.97	0.734	60.18	57.17
w/o Fusion	1.181	66.69	65.58	1.265	62.25	59.73	0.687	72.90	72.54	0.741	63.19	63.64
w/o NoisyTrain	1.159	65.43	60.09	1.231	61.87	56.47	0.731	75.44	71.08	0.754	73.98	69.29
ALMV (ours)	**1.122**	**68.87**	**68.16**	**1.224**	**64.08**	**62.63**	**0.678**	**78.12**	**75.56**	0.734	**75.44**	**71.56**

注: 加粗数据表示最佳结果

首先, 我们研究对抗学习的组件对鲁棒 MSA 的影响. 对于去除重建器和判别器, 可以观察到近乎所有的指标都出现了不同程度的下降, 除了 CMU-MOSEI 的结构时间缺失下的 MAE, 尽管其性能出现小范围好转 (我们认为这是处于误差范围内), 但准确率和 *F*1 分数实际上出现了超过 5% 的大幅下降. 可以观察到, 对于同时删除重建器和判别器比单独删除其中某个组件在两个数据集的大部分指标性能下降更多. 这强调了对抗训练和语义重建的有效性, 表明判别和重建模块是互补的, 因为当另一个组件保留时, 可以减轻其中一个组件的去除, 且可以相互增强.

其次, 我们研究模态级和话语级的对抗学习对性能的有效性. 对于前者, 本文删除了针对每个单模态的重建器和判别器, 对于后者本文删除了融合后特征的重建器和判别器. 可以观察到, 这两者性能指标均出现下降, 特别是在 CMU-MOSEI 上, 后者性能下降幅度更甚. 我们认为, 由于话语级对抗学习直接作用于最终的情感分析, 相比于模态级对抗学习效果更明显和直接. 相比而言, 模态级对抗学习则对恢复模态特定性更加有利, 能提供更多的补充信息.

然后, 我们研究特征提取网络和融合网络对性能的影响, 对于前者, 删除了 Transformer 编码器, 对于后者, 删除多视图网络, 将特征提取后的特征直接在通道上拼接以代替融合网络. 可以观察到删除这两个骨干网络, 性能指标在两个数据集均出现下滑, 而且在 CMU-MOSEI 数据集上出现了高达 10% 以上的性能下降. 由于 CMU-MOSEI 的样本量更大且主题更丰富, 平均话语长度和持续时间相较于 CMU-MOSI 也分别增加了 53% 和 74%, 这导致其对骨干网络的要求更高. 删除 Transformer 编码器, 则无法充分提取单模态特征; 删除多视图网络, 则融合特征的表示能力不足.

最后, 我们直接删除了整个噪声注入训练的模组, 即仅保留骨干网络. 也即是说, 仅使用骨干网络在干净的数据上训练, 然后在不同的缺失条件下做噪声测试. 由表 6 可知, 在 CMU-MOSEI 上出现了 2%–4% 的性能下降, 而在 CMU-MOSI 下降最多的情况下则超过了 8%, 远远高于前者的幅度. 这并不是偶然, 在第 3.6 节细粒度模态特征缺失实验中, ALMV 在小样本数据集 CMU-MOSI 上相比于基线方法的领先程度明显高于在大规模数据集 CMU-

MOSEI 上的实验结果. 此外, 在第 3.7 节中, 在小样本数据集 CMU-MOSI 上几乎在所有缺失率下的所有指标都好于其他方法, 而在 CMU-MOSEI 数据集, ALMV 的优势更多限于低缺失率下. 这些实验结果表明, ALMV 在小样本/低资源场景 (如 CMU-MOSI) 中带来的相对增益明显高于在大规模数据集 (如 CMU-MOSEI) 上的增益. 原因可归结为: 噪声注入与对抗约束在数据有限时充当结构化正则化并提高样本多样性, 从而显著提升鲁棒性; 而在样本充足的场景中, 模型可从大量真实样本自然学习到多样性, 人工注入噪声的边际收益相对减小. 因此, ALMV 更适用于中小规模, 缺陷敏感或噪声多变的应用场景; 对于大规模的数据和场景, ALMV 模型的骨干网络将会更具决定性.

3.10 其他异质性缺陷泛化实验

虽然本文使用时间特征擦除训练 ALMV 模型, 但我们认为 ALMV 模型也对其他异质缺陷测试具有良好的泛化性. 在本节, 进一步使用随机特征缺失来训练模型用以测试非对齐模态的数据集. 此外, 还测试了 ALMV 防御转录文本攻击的能力, 我们直接在构造的噪声测试集上记录模型的评估指标, 而不使用 *AUILC* 值.

(1) 非对齐模态数据实验

本文 ALMV 方法采用基于噪声注入的数据增强, 我们使用时间特征擦除用于模仿推断同一时段中的潜在缺陷, 即随机选择相同帧的时间步长, 并且同时擦除在这些时间步长的 3 个模态特征. 在这个要求下, 需要在模态对齐的数据集上才能实现数据增强. 为探索 ALMV 在时间步非对齐情况下的性能, 我们选择当下最流行的数据集之一的 CH-SIMS[59]来进行实验, 它是一个模态未对齐数据集. 具体来说, 遵循文献 [24,60], 在缺失率区间{0.0, 0.1,…, 0.5}上验证总体性能, 使用随机模态特征擦除来替代数据增强方法, 同时探索 ALMV 在随机模态缺失场景下的性能. 实验结果如表 7 所示. 实验结果表明, ALMV 仅在 Corr 指标上略微逊色于最先进的方法 CTRN[60], 而在其他所有指标中均取得了领先的结果. CTRN[60]引入了动量模型来帮助模型聚焦于特定于情感分析任务的特征. ALMV 采用随机模态特征擦除的数据增强, 对随机噪声的模仿训练, 极大地提高了模型的鲁棒性. 更重要的是 ALMV 在模态级和话语级两个层面做双重对抗训练对模态特定信号和话语信号起到了充分的恢复作用.

表 7 CH-SIMS 的随机模态特征缺失实验结果

Model	Acc-2	Acc-5	MAE	Corr
TFN[c][14]	0.373	0.181	0.233	0.259
MulT[c][11]	0.37	0.173	0.244	0.227
MISA[e][55]	0.347	0.106	0.294	0.038
Self-MM[d][56]	0.374	0.183	0.231	0.258
TFRNet[c][24]	0.377	0.18	0.237	0.253
NIAT[e][25]	0.36	0.18	0.24	0.253
EMT-DLFR[d][39]	0.371	0.194	0.232	0.269
CTRN[d][60]	0.382	0.21	0.223	**0.276**
ALMV	**0.385**	**0.22**	**0.225**	0.272

注: 标记c的数据来自文献[24], 标记d的数据来自文献[60], 标记e的数据来自通过源代码实现; 加粗数据表示最佳结果

(2) 转录文本攻击实验

在现实生活中, 除了常见的细粒度模态特征缺失和整体模态缺失之外, 对转录文本的攻击同样是 MSA 系统的一个巨大威胁. 我们探索了几种常见类型的转录文本攻击的实验[61,62], 即情感词语删除, 情感词语替换和 ASR 转录错误. 具体而言, 对情感词删除, 即将转录文本中的情感词替换成 [UNK] 标记. 而对情感词替换, 即将提供的转录文本中的情感词替换为其在 CMU-MOSI 数据集上的反义词. 对于 ASR 错误, 则使用 ASR 转录系统得到的转录文本, 替换 CMU-MOSI 测试集上提供的文本, 经统计, 替换后在测试集上 ASR 系统的误码率约为 35%.

如表 8 所示, 在完美数据下, ALMV 领先最先进的模型的性能均超过 1% 以上. 此外, ALMV 在情感词删除和 ASR 转录错误方面的性能优于所有的基线方法. 在情感词替换方面 ALMV 在 Acc-2 上达到了最佳的性能, 但在 MAE 上略微落后于 SWRM 方法[62]. SWRM 为情感词感知多模态精炼模型, 它是专门针对文本转录错误所设计的

方法. 它能够检测情感词在文本中的位置, 并利用多模态情感线索动态地精炼错误的情感词. 这些结果表明, 在CMU-MOSI 数据集上针对情态特征缺失而设计的 ALMV 模型, 用于防御文本转录攻击依然是有效的. ALMV 在转录文本攻击实验中, 更多是通过捕获情感相关的非语言线索即视频模态特征和音频模态特征来防御情感词攻击. 尽管本文提出的 ALMV 方法对情感词替换, 情感词删除和 ASR 转录错误等多个方面都取得了优异的效果, 但考虑到语言是 MSA 的主要模态, 在未来仍然需要针对自动语音识别错误, 情感词攻击等转录文本攻击场景做更进一步的优化和提升.

表 8　CMU-MOSI 数据集转录文本攻击实验结果

Model	Acc-2 (%)				MAE			
	完美数据	情感词删除	情感词替换	ASR转录错误	完美数据	情感词删除	情感词替换	ASR转录错误
MulT[11]	80.80	71.82	62.24	74.15	0.881	1.114	1.344	1.013
MISA[55]	81.80	72.50	62.05	75.63	0.809	1.062	1.326	0.946
Self-MM[56]	80.81	72.25	62.73	76.15	0.790	1.067	1.327	0.923
T2FN[31]	79.16	70.94	62.34	72.69	0.890	1.111	1.320	1.022
TPFN[38]	79.30	69.44	62.05	71.82	0.896	1.146	1.352	1.038
TFRNet[24]	78.77	70.02	62.39	72.55	0.980	1.136	1.326	1.084
SWRM[62]	80.14	71.07	61.13	76.45	0.945	1.122	**1.294**	0.894
NIAT[25]	81.82	73.08	62.49	77.21	0.758	1.026	1.368	0.887
ALMV (ours)	**83.24**	**73.81**	**62.80**	**77.64**	**0.747**	**1.019**	1.309	**0.881**

注: 加粗数据表示最佳结果

3.11　对抗学习的超参数研究

ALMV 使用 3 种不同的损失来监督训练过程中的表征学习. 在训练期间, 由于不同损失的预设权重显著影响模型性能, 因此平衡不同的损失成为关键问题. 图 6 是 CMU-MOSI 数据集上的超参数研究, 用深色表示最好性能.

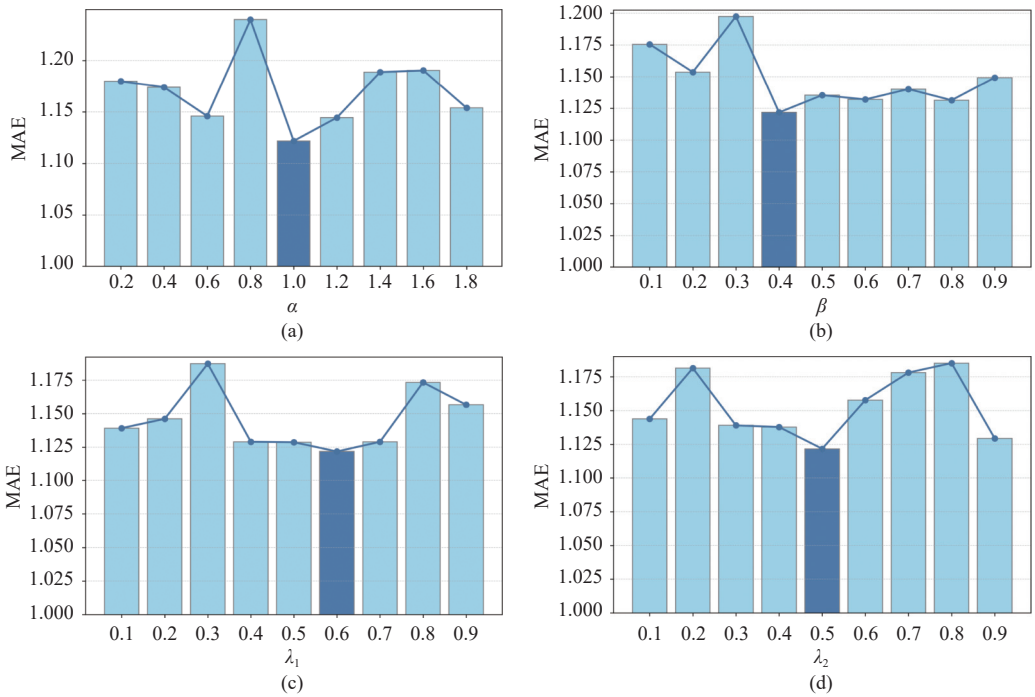


图 6　CMU-MOSI 数据集上不同超参数的性能

原始数据和噪声数据之间的平衡: 如图 6 所示, 超参数 α 确定了来自完整数据流和噪声数据流的分类损失之间的权重. 我们分析图 6 中给出的实验结果表明, 当 α=1.0 时, 该模型达到最佳性能, 这表明来自完美数据流和噪声数据流的分类损失对所提出的 ALMV 框架的贡献相等. 通过超参数分析, 强调了基于噪声注入增强数据的有效性.

判别器和重建器之间的平衡: 超参数 β 决定了判别损失和重建损失之间的权衡. 根据公式 (30), 较大的 β 对应于分配给辨别损失的较高权重. 判别项推动重建特征向干净特征的分布靠拢, 而重构项则保证样本级语义保真. 基于图 6 中所示的调整实验结果, 其中 β 的范围为 0.1–0.9, 我们发现 β=0.4 时, MAE 最低, 产生了最佳的模型性能. 事实上, 该参数也和数据样本量有关, 在 CMU-MOSI 这样的小样本情形下, 适度偏重重建器有利于保证语义保真和稳定表示. 而对于 CMU-MOSEI, 基于其复杂样本的特性, 则更需要增大判别权重以提高对复杂噪声的鲁棒性. 实验结果表明, 判别损失对于有效学习损坏数据的表示至关重要.

模态级对抗学习和话语级对抗学习的平衡: λ_1 和 λ_2 分别是模态级和话语级的语义重建的平衡系数和对抗损失的平衡系数. 话语级的对抗学习旨在指导融合网络通过话语级的语义重构, 在噪声环境下重新生成话语缺失的语义. 模态级的对抗学习旨在指导特征提取网络恢复单模态的模态特定信号, 为多模态融合提升补充信息. 所以, 事实上话语级对抗学习由于直接作用于情感分析, 其权重占比会更高, 起到更加主导的作用. 而模态级对抗学习作为补充, 其权重更低. 由图 6 可知, 当 λ_1=0.6 和 λ_2=0.5 时达到了最佳的平衡. 其模态级的重建系数略高于判别系数, 这也再次印证了在小样本下, 模型需要适度地偏重重建器网络以实现语义保真. 结果表明, 区分模态级的对抗学习和话语级的对抗学习的权重对于有效学习对抗表征学习至关重要.

由此我们可以获得在 CMU-MOSI 上模型的最佳参数为 $(\alpha, \beta, \lambda_1, \lambda_2) = (1.0, 0.4, 0.6, 0.5)$. 类似的, 可以得到在 CMU-MOSEI 上的最佳参数为 $(\alpha, \beta, \lambda_1, \lambda_2) = (1.0, 0.6, 0.5, 0.5)$.

3.12 多视图网络的权重可视化

本节探究多视图网络的融合过程, 以证明多视图网络的可解释性. 我们在 CMU-MOSI 数据集做实验, 输出同一个训练好的模型在低缺失率和高缺失率的 4 种缺失率下的每个视图的权重可视化. 横坐标为 12 个视图, 纵坐标为 CMU-MOSI 中样本的索引 (我们将所有样本的权重拼接同时输出), 如图 7 所示.

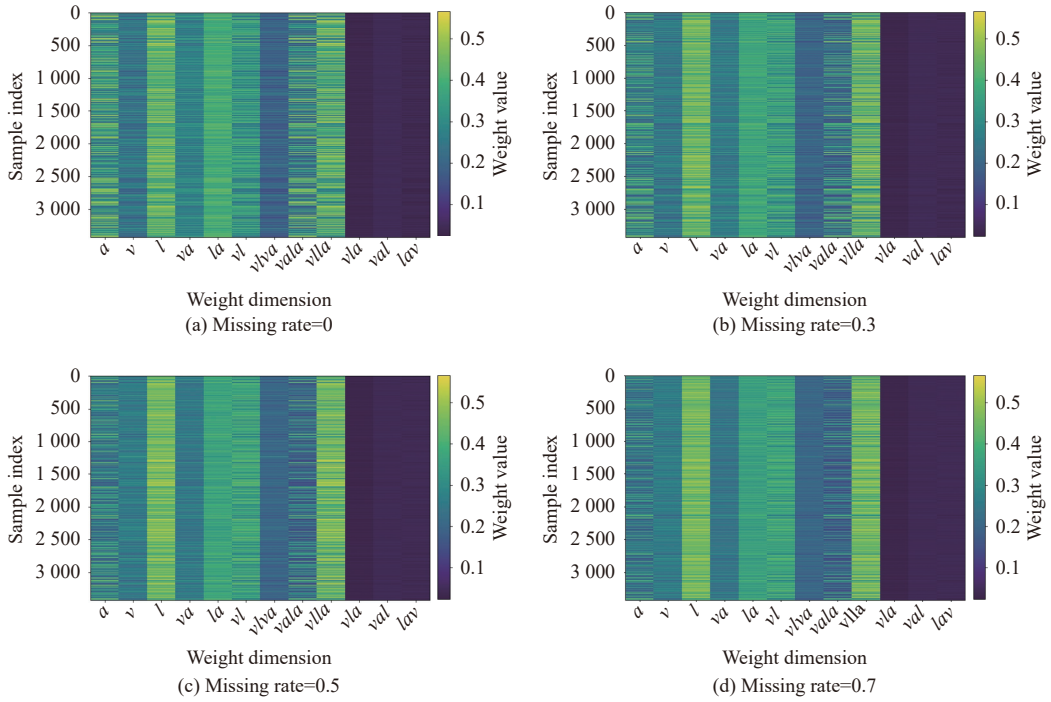


图 7 不同缺失率下多视图特征的权重可视化

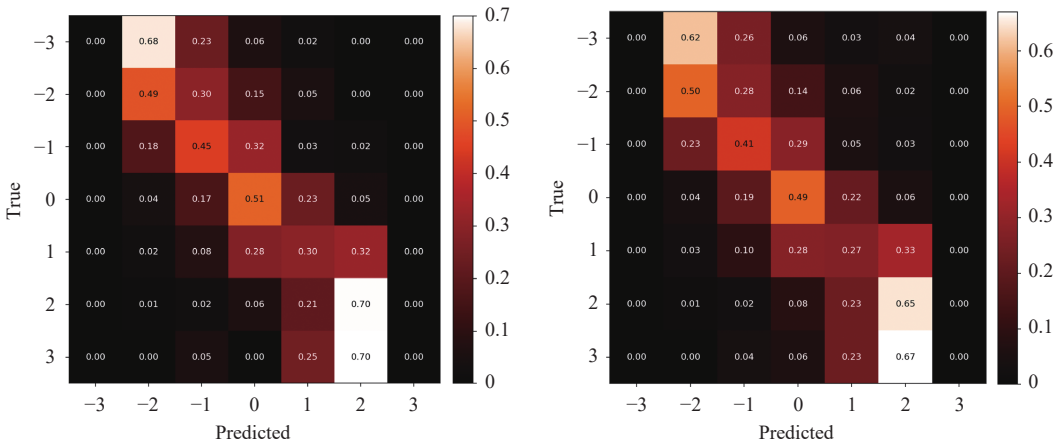
对于单视图特征 (图 7 前 3 列), 显然语言模态对于大多数样本来说是最具预测性的, 因为语言是情感分析最重要的线索, 但它的重要性在样本之间差异很大. 这表明当语言模态对于特定样本来说微不足道时, 以其他模态的多模态形式分析情感的优势可以发挥主导作用. 对于双视图特征 (图 7 第 4–6 列), *la* 和 *vl* 视图的权重非常接近, 其次是 *va*, 这是因为语言模态在双视图中起着巨大的作用, 不包含语言模态的视图权重在多数样本中会更低.

而对于三视图层, 可以发现对于融合互补的双视图和单视图特征得到的三视图特征对于情感分析的贡献不高, 几乎仅起到了补充作用. 而通过融合两个双视图特征得到的三视图特征主导了三视图层的情感分析贡献. 分析原因, 这和我们设计网络的信息增益的层级差异有关, 多视图特征的本质和视图协同带来的信息增益有关. 双视图特征相互融合, 是高阶特征融合. 以三视图特征 *vlla* 为例, 其来自 *vl* 视觉-语言双视图和 *la* 语言-音频双视图融合得到, 高阶协同带来叠加信息增益, 语言模态的语义锚点放大了两个双视图融合的信息增益, 使模型在高阶融合中捕获到单模态或双模态无法表达的复杂跨模态关系. 而双视图与单视图特征的融合是低阶融合, 两者的信息在之前的单视图层和双视图层已充分表达, 两者融合带来的信息增益有限, 在三视图层中更多起到补充作用.

同时可以观察到其中 *vlla* 的三视图特征相比于单视图特征和双视图特征, 对情感预测的贡献十分显著, 这证明了探索每两个双视图特征之间的相互作用的必要性. 此外, 随着缺失率的逐渐上升, 各个样本的不同的视图对于情感预测的贡献也在动态的调整, 如在高缺失率为 0.7 的情况下, 三视图特征 *vlla* 的权重相比于低缺失率的时候更高.

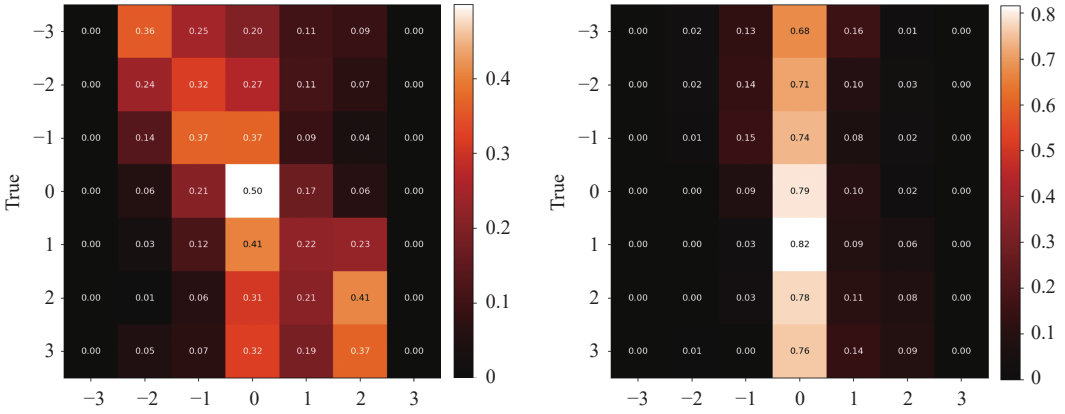
3.13 混淆矩阵可视化

我们输出了在随机缺失的噪声条件下, CMU-MOSI 数据集上不同缺失率下模型输出标签的混淆矩阵图, 由于 MSA 的标签为 (−3, 3) 之间离散的情感分数, 我们将其归为 7 类以方便观察. 如图 8 所示, 强烈积极和强烈消极几乎都无法准确预测, 这是因为在图 4 的数据集标签分布中我们可以知道, 该数据集中此类情感样本非常少, 导致模型无法获得足够的训练, 从而把强烈积极和强烈消极都归类在了积极和消极这一类中. 而除此以外, 对于其他情感没有那么强烈的类别, 模型都能做出最好的判断. 而随着缺失率的上升, 可以观察到, 模型预测的情感逐渐向中性情感靠拢. 例如, 当缺失率为 0.1 时, 模型有 0.5 的概率能将消极这一类别预测正确, 而只有 0.14 的概率会将其预测为中性; 但当缺失率为 0.5 时, 模型只有 0.24 的概率能正确预测消极, 而有高达 0.27 预测出中性的结果. 而当缺失率为 0.9 时, 几乎所有类别预测的情感均为中性. 显而易见的情况是, 缺失率越高, 数据的有用性就越少, 模型只能从大量零向量特征中预测出中性的情感.



(a) Missing rate=0　　(b) Missing rate=0.1

图 8　不同缺失率下混淆矩阵可视化



(c) Missing rate=0.5 (d) Missing rate=0.9

图 8 不同缺失率下混淆矩阵可视化 (续)

4 结 论

本文针对真实世界场景下多模态数据常见的缺失与噪声干扰问题, 提出了一种融合对抗学习与多视图网络的鲁棒多模态情感分析框架 ALMV. 与传统方法仅依赖完整模态不同, 本方法通过人为模拟时间模态特征缺失并与原始数据配对, 生成噪声-原始实例对, 形成多样化的增强样本; 随后利用时间卷积与 Transformer 编码器分别挖掘每种模态的局部和全局信息, 并引入一种权重调控的多视图融合网络学习原始实例对的多模态联合表示, 从多个视角探索细腻的情感线索. 针对噪声与完整样本的中间表示, 我们进一步设计了带有语义重建监督的多重对抗训练策略, 分别在模态级和话语级统一对齐噪声与真实分布, 既恢复了各模态的特定信号, 又强化了融合特征的表达能力. 在多种不同模态缺失场景的实验中, ALMV 相较于现有基线方法在准确率, 鲁棒性及泛化能力上均表现出显著提升, 充分验证了所提方法在处理非理想多模态数据时的有效性. 在未来的工作中, 我们将进一步扩展该方法到更多现实世界中的其他类型的异构数据缺陷如引入图像遮挡或传感器漂移等多种缺失模式.

References

[1] Chauhan GS, Agrawal P, Meena YK. Aspect-based sentiment analysis of students' feedback to improve teaching-learning process. In: Satapathy SC, Joshi A, eds. Information and Communication Technology for Intelligent Systems. Singapore: Springer, 2019. 259–266. [doi: 10.1007/978-981-13-1747-7_25]

[2] Zhou J, Ye JM. Sentiment analysis in education research: A review of journal publications. Interactive Learning Environments, 2023, 31(3): 1252–1264. [doi: 10.1080/10494820.2020.1826985]

[3] Narayanan V, Manoghar BM, Dorbala VS, Manocha D, Bera A. ProxEmo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 8200–8207. [doi: 10.1109/IROS45743.2020.9340710]

[4] Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 2023, 91: 424–444. [doi: 10.1016/j.inffus.2022.09.025]

[5] Sun YH, Liu ZZ, Sheng QZ, Chu DH, Yu J, Sun HX. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 2024, 110: 102454. [doi: 10.1016/j.inffus.2024.102454]

[6] Trillo JR, Herrera-Viedma E, Morente-Molinera JA, Cabrerizo FJ. A large scale group decision making system based on sentiment analysis cluster. Information Fusion, 2023, 91: 633–643. [doi: 10.1016/j.inffus.2022.11.009]

[7] Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 2017, 37: 98–125. [doi: 10.1016/j.inffus.2017.02.003]

[8] Punetha N, Jain G. Game theory and MCDM-based unsupervised sentiment analysis of restaurant reviews. Applied Intelligence, 2023,

53(17): 20152–20173. [doi: 10.1007/s10489-023-04471-1]

[9] Dirin A. User experience of mobile virtual reality: Experiment on changes in students' attitudes. The Turkish Online Journal of Educational Technology, 2020, 19(3): 80–93.

[10] Ma LY, Yao Y, Liang T, Liu TL. Multi-scale cooperative multimodal Transformers for multimodal sentiment analysis in videos. In: Proc. of the 37th Australasian Joint Conf. on Artificial Intelligence. Melbourne: Springer, 2025. 281–297. [doi: 10.1007/978-981-96-0351-0_21]

[11] Tsai YHH, Bai SJ, Liang PP, Kolter JZ, Morency LP, Salakhutdinov R. Multimodal Transformer for unaligned multimodal language sequences. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6558–6569. [doi: 10.18653/v1/P19-1656]

[12] Rahman W, Hasan MK, Lee S, Zadeh AAB, Mao CF, Morency LP, Hoque E. Integrating multimodal information in large pretrained Transformers. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 2359–2369. [doi: 10.18653/v1/2020.acl-main.214]

[13] Ma YK, Ma B. Multimodal sentiment analysis on unaligned sequences via holographic embedding. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP 2022). Singapore: IEEE, 2022. 8547–8551. [doi: 10.1109/ICASSP43922.2022.9747646]

[14] Zadeh A, Chen MH, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1103–1114. [doi: 10.18653/v1/D17-1115]

[15] Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh AB, Morency LP. Efficient low-rank multimodal fusion with modality-specific factors. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne: ACL, 2018. 2247–2256. [doi: 10.18653/v1/P18-1209]

[16] Tang JJ, Li K, Jin XY, Cichocki A, Zhao QB, Kong WZ. CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). ACL, 2021. 5301–5311. [doi: 10.18653/v1/2021.acl-long.412]

[17] Luo YY, Wu R, Liu JF, Tang XL. Multimodal sentiment analysis method based on adaptive weight fusion. Ruan Jian Xue Bao/Journal of Software, 2024, 35(10): 4781–4793 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6998.htm [doi: 10.13328/j.cnki.jos.006998]

[18] Tang ZM, Xiao Q, Qin YC, Zhou X, Zhou JT, Li KL. Multi-view interactive representations for multimodal sentiment analysis. IEEE Trans. on Consumer Electronics, 2024, 70(1): 4095–4107. [doi: 10.1109/TCE.2024.3357480]

[19] Zeng JD, Zhou JT, Liu TY. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. IEEE Trans. on Multimedia, 2023, 25: 6301–6314. [doi: 10.1109/TMM.2022.3207572]

[20] Zeng JD, Liu TY, Zhou JT. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Madrid: ACM, 2022. 1545–1554. [doi: 10.1145/3477495.3532064]

[21] Zeng JD, Zhou JT, Liu TY. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 2924–2934. [doi: 10.18653/v1/2022.emnlp-main.189]

[22] Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423–443. [doi: 10.1109/TPAMI.2018.2798607]

[23] Yuan ZQ, Zhang BZ, Xu H, Gao K. Meta noise adaption framework for multimodal sentiment analysis with feature noise. IEEE Trans. on Multimedia, 2024, 26: 7265–7277. [doi: 10.1109/TMM.2024.3362600]

[24] Yuan ZQ, Li W, Xu H, Yu WM. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 4400–4407. [doi: 10.1145/3474085.3475585]

[25] Yuan ZQ, Liu YH, Xu H, Gao K. Noise imitation based adversarial training for robust multimodal sentiment analysis. IEEE Trans. on Multimedia, 2024, 26: 529–539. [doi: 10.1109/TMM.2023.3267882]

[26] Ma MM, Ren J, Zhao L, Testuggine D, Peng X. Are multimodal Transformers robust to missing modality? In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 18156–18165. [doi: 10.1109/CVPR52688.2022.01764]

[27] Tsai YHH, Liang PP, Zadeh A, Morency LP, Salakhutdinov R. Learning factorized multimodal representations. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–20.

[28] Ma MM, Ren J, Zhao L, Tulyakov S, Wu C, Peng X. SMIL: Multimodal learning with severely missing modality. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 2302–2310. [doi: 10.1609/aaai.v35i3.16330]

[29] Han W, Chen H, Kan MY, Poria S. MM-Align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: ACL, 2022. 10498–10511. [doi: 10.18653/v1/2022.emnlp-main.717]

[30] Wang ZL, Wan ZH, Wan XJ. TransModality: An End2End fusion method with Transformer for multimodal sentiment analysis. In: Proc. of the Web Conf. 2020. Taipei, China: ACM, 2020. 2514–2520. [doi: 10.1145/3366423.3380000]

[31] Liang PP, Liu Z, Tsai YHH, Zhao QB, Salakhutdinov R, Morency LP. Learning representations from imperfect time series data via tensor rank regularization. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 1569–1576. [doi: 10.18653/v1/P19-1152]

[32] Wang N, Wang Q, Ouyang DT. Multimodal sentiment analysis model based on knowledge distillation and dynamic adjustment mechanism. Chinese Journal of Computers, 2025, 48(8): 1923–1942 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2025.01923]

[33] Pham H, Liang PP, Manzini T, Morency LP, Póczos B. Found in translation: Learning robust joint representations by cyclic translations between modalities. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 6892–6899. [doi: 10.1609/aaai.v33i01.33016892]

[34] Zhao JM, Li RC, Jin Q. Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (Vol. 1: Long Papers). ACL, 2021. 2608–2618. [doi: 10.18653/v1/2021.acl-long.203]

[35] Guo ZR, Jin T, Zhao Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Bangkok: ACL, 2024. 1726–1736. [doi: 10.18653/v1/2024.acl-long.94]

[36] Reza MK, Prater-Bennette A, Asif MS. Robust multimodal learning with missing modalities via parameter-efficient adaptation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2025, 47(2): 742–754. [doi: 10.1109/TPAMI.2024.3476487]

[37] Liang PP, Zadeh A, Morency LP. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. ACM Computing Surveys, 2024, 56(10): 264. [doi: 10.1145/3656580]

[38] Li BH, Li C, Duan F, Zheng N, Zhao QB. TPFN: Applying outer product along time to multimodal sentiment analysis fusion on incomplete data. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 431–447. [doi: 10.1007/978-3-030-58586-0_26]

[39] Sun LC, Lian Z, Liu B, Tao JH. Efficient multimodal Transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Trans. on Affective Computing, 2024, 15(1): 309–325. [doi: 10.1109/TAFFC.2023.3274829]

[40] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems—Vol. 2. Montreal: MIT Press, 2014. 2672–2680.

[41] Pang TY, Yang X, Dong YP, Su H, Zhu J. Bag of tricks for adversarial training. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–21.

[42] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 2018, 35(1): 53–65. [doi: 10.1109/MSP.2017.2765202]

[43] Reed S, Akata Z, Yan XC, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: Proc. of the 33rd Int'l Conf. on Machine Learning—Vol. 48. New York: JMLR.org, 2016. 1060–1069.

[44] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas D. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 5908–5916. [doi: 10.1109/ICCV.2017.629]

[45] Mai SJ, Hu HF, Xing SL. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 164–172. [doi: 10.1609/aaai.v34i01.5347]

[46] Yan XQ, Hu SZ, Mao YQ, Ye YD, Yu H. Deep multi-view learning methods: A review. Neurocomputing, 2021, 448: 106–129. [doi: 10.1016/j.neucom.2021.03.090]

[47] Han ZB, Zhang CQ, Fu HZ, Zhou JT. Trusted multi-view classification with dynamic evidential fusion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2551–2566. [doi: 10.1109/TPAMI.2022.3171983]

[48] Wen GQ, Cao P, Bao HW, Yang WJ, Zheng T, Zaiane O. MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis. Computers in Biology and Medicine, 2022, 142: 105239. [doi: 10.1016/j.compbiomed.2022.105239]

[49] Huang HN, Zhou GX, Zhao QB, He LF, Xie SL. Comprehensive multiview representation learning via deep autoencoder-like

nonnegative matrix factorization. IEEE Trans. on Neural Networks and Learning Systems, 2024, 35(5): 5953–5967. [doi: 10.1109/TNNLS.2023.3304626]

[50] Zheng QH, Zhu JH, Li ZY, Tian ZQ, Li C. Comprehensive multi-view representation learning. Information Fusion, 2023, 89: 198–209. [doi: 10.1016/j.inffus.2022.08.014]

[51] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1 Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[52] Zadeh A, Zellers R, Pincus E, Morency LP. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 2016, 31(6): 82–88. [doi: 10.1109/MIS.2016.94]

[53] Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne: ACL, 2018. 2236–2246. [doi: 10.18653/v1/P18-1208]

[54] Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP—A collaborative voice analysis repository for speech technologies. In: Proc. of the 2014 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP 2014). Florence: IEEE, 2014. 960–964. [doi: 10.1109/ICASSP.2014.6853739]

[55] Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 1122–1131. [doi: 10.1145/3394171.3413678]

[56] Yu WM, Xu H, Yuan ZQ, Wu JL. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI Press, 2021. 10790–10797. [doi: 10.1609/aaai.v35i12.17289]

[57] Wang D, Guo XT, Tian YM, Liu JH, He LH, Luo XM. TETFN: A text enhanced Transformer fusion network for multimodal sentiment analysis. Pattern Recognition, 2023, 136: 109259. [doi: 10.1016/j.patcog.2022.109259]

[58] Yang ZH, He Q, Yu MH, Du NS, Lu YJ. TCTR: Text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. Information Fusion, 2026, 126: 103571. [doi: 10.1016/j.inffus.2025.103571]

[59] Yu WM, Xu H, Meng FY, Zhu YL, Ma YX, Wu JL, Zou JY, Yang KC. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 3718–3727. [doi: 10.18653/v1/2020.acl-main.343]

[60] Sun B, Jia L, Cui YM, Wang N, Jiang T. Conv-enhanced Transformer and robust optimization network for robust multimodal sentiment analysis. Neurocomputing, 2025, 634: 129842. [doi: 10.1016/j.neucom.2025.129842]

[61] Balakrishnan S, Fang YH, Zhu XD. Exploring robustness of prefix tuning in noisy data: A case study in financial sentiment analysis. In: Proc. of the 4th Workshop on Financial Technology and Natural Language Processing (FinNLP). Abu Dhabi: ACL, 2022. 78–88. [doi: 10.18653/v1/2022.finnlp-1.9]

[62] Wu Y, Zhao YY, Yang H, Chen S, Qin B, Cao XH, Zhao WT. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. In: Findings of the Association for Computational Linguistics: ACL 2022. Dublin: ACL, 2022. 1397–1406. [doi: 10.18653/v1/2022.findings-acl.109]

附中文参考文献

[17] 罗渊贻, 吴锐, 刘家锋, 唐降龙. 基于自适应权值融合的多模态情感分析方法. 软件学报, 2024, 35(10): 4781–4793. http://www.jos.org.cn/1000-9825/6998.htm [doi: 10.13328/j.cnki.jos.006998]

[32] 王楠, 王淇, 欧阳丹彤. 基于知识蒸馏与动态调整机制的多模态情感分析模型. 计算机学报, 2025, 48(8): 1923–1942. [doi: 10.11897/SP.J.1016.2025.01923]

作者简介

蔡林沁, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为人工智能, 模式识别, 人机交互.

刘道宏, 硕士, 主要研究领域为多模态情感分析, 计算机视觉.

刘岚睿, 硕士, 主要研究领域为多模态, 视觉问答, 计算机视觉.

牟仁云, 硕士生, 主要研究领域为多模态, 知识图谱, 自然语言处理.