计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(4):13−25 [doi: 10.15888/j.cnki.csa.009461]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于双编码器表示学习的多模态情感分析①

冼广铭, 阳先平, 招志锋

(华南师范大学 软件学院, 佛山 528225)
通信作者: 冼广铭, E-mail: xgm20011@163.com

摘　要: 多模态情感分析旨在通过用户上传在社交平台上的视频来判断用户的情感. 目前的多模态情感分析研究主要是设计复杂的多模态融合网络来学习模态之间的一致性信息, 在一定程度上能够提升模型的性能, 但它们大部分都忽略了模态之间的差异性信息所起到的互补作用, 从而导致情感分析出现偏差. 本文提出了一个基于双编码器表示学习的多模态情感分析模型 DERL (dual encoder representation learning), 该模型通过双编码器结构学习模态不变表征和模态特定表征. 具体来说, 我们利用基于层级注意力机制的跨模态交互编码器学习所有模态的模态不变表征, 获取一致性信息; 利用基于自注意力机制的模态内编码器学习模态私有的模态特定表征, 获取差异性信息. 此外, 我们设计两个门控网络单元对编码后的特征进行增强和过滤, 以更好地结合模态不变和模态特定表征, 最后在融合时通过缩小不同多模态表示之间的 L2 距离以捕获它们之间潜在的相似情感用于情感预测. 在两个公开的数据集 CMU-MOSI 和 CMU-MOSEI 上的实验结果表明该模型优于一系列基线模型.

关键词: 多模态情感分析; 双编码器; 层级注意力; 门控网络单元; 相似情感

引用格式: 冼广铭,阳先平,招志锋.基于双编码器表示学习的多模态情感分析.计算机系统应用,2024,33(4):13–25. http://www.c-s-a.org.cn/1003-3254/9461.html

Multimodal Sentiment Analysis Based on Dual Encoder Representation Learning

XIAN Guang-Ming, YANG Xian-Ping, ZHAO Zhi-Feng

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: Multimodal sentiment analysis aims to assess users' sentiment by analyzing the videos they upload on social platforms. The current research on multimodal sentiment analysis primarily focuses on designing complex multimodal fusion networks to learn the consistency information among modalities, which enhances the model's performance to some extent. However, most of the research overlooks the complementary role played by the difference information among modalities, resulting in sentiment analysis biases. This study proposes a multimodal sentiment analysis model called DERL (dual encoder representation learning) based on dual encoder representation learning. This model learns modality-invariant representations and modality-specific representations by a dual encoder structure. Specifically, a cross-modal interaction encoder based on a hierarchical attention mechanism is employed to learn the modality-invariant representations of all modalities to obtain consistency information. Additionally, an intra-modal encoder based on a self-attention mechanism is adopted to learn the modality-specific representations within each modality and thus capture difference information. Furthermore, two gate network units are designed to enhance and filter the encoded features and enable a better combination of modality-invariant and modality-specific representations. Finally, during fusion, potential similar sentiment between different multimodal representations is captured for sentiment prediction by reducing the L2 distance among them. Experimental results on two publicly available datasets CMU-MOSI and CMU-MOSEI show that this model

① 基金项目: 国家自然科学基金 (61070015)
收稿时间: 2023-10-15; 修改时间: 2023-11-15; 采用时间: 2023-11-24; csa 在线出版时间: 2024-01-30
CNKI 网络首发时间: 2024-02-01

outperforms a range of baselines.

Key words: multimodal sentiment analysis; dual encoder; hierarchical attention; gate network unit; similar sentiment

1 引言

情感分析是一种系统性过程, 旨在从不同数据来源确定人们的观点和态度. 在早期的研究中通常使用单一类型的数据, 例如文本、音频或图像来分析人类的情感[1]. 然而, 随着社交媒体的普及, 人们越来越倾向于通过上传视频来分享对日常生活中各种事物的看法. 这一趋势催生了多模态情感分析的发展, 特别是在单模态数据由于其内容限制和情感歧义而显得不足以全面理解情感表达的背景下[2]. 多模态情感分析的目标是将基于单模态的分析方法扩展到包含文本、音频和视觉信息的视频中, 以充分理解人们所表达的情感. 例如, 在分析讽刺话语时, 结合说话人的面部表情和语音语调能够更准确地识别出其中的情感含义.由于在处理复杂的多模态数据方面显示出的显著效果, 多模态情感分析已经吸引了越来越多的研究和关注[3].

多模态情感分析任务并非简单地组合多个模态的数据, 将具有异质性的不同模态数据进行有效融合一直是该领域的难点之一, 过去的研究也大多集中在这一问题上. Tsai 等人[4]提出使用成对的跨模态 Transformer[5]来执行两两模态之间的对齐. 这一方法使不同模态特征通过跨模态注意力机制探索潜在的跨模态情感映射, 并在一定程度上解决了该问题. 但这样的对齐方式存在两个缺陷, 一方面模态之间的交互并不充分, 只能从另一个模态中直接学习情感信息而无法同时与其他所有模态进行信息共享, 从而使学习到的模态不变表征可能会存在偏差. 另一方面, 其计算复杂度为$O(C_n^2)$ (n 为模态数), 随着模态的增加计算复杂度呈指数级增长, 因此对计算资源的要求非常高. Lv 等人[6]和 Sun 等人[7]提出的方法同时在所有模态之间共享信息, 能够很好地学习到全局情感信息, 但它们都增加了额外的开销来存储全局共享信息, 并且没有考虑到每个模态私有的特定于模态的信息.

一个更有效的多模态表示应同时包含一致性信息和差异性信息. 学习模态不变表征意味着在不同模态之间学习一致性, 从相关模态中提取共同的语义. 而学习模态特定表征则涉及理解不同模态对同一内容的不同表达, 重点关注模态之间的差异性. Hazarika 等人[8]和 Yu 等人[9]提出的方法同时将模态特征的一致性信息和差异性信息考虑在内, 使用不同的任务损失或设计一个单模态标签生成模块来学习表征的多样性. 设计复杂的多模态融合网络能够使模型的预测能力得到极大的提升, 但忽略不同模态的差异性, 模型的预测效果也会受到限制. 如图 1 所示, 根据文本内容“sick”很容易判断它所表达的是消极的情感, 但视觉信息“smile”展示的是积极的情感, 最后结合 3 个模态的互补信息才能够准确判断它所表达的是积极的情感.

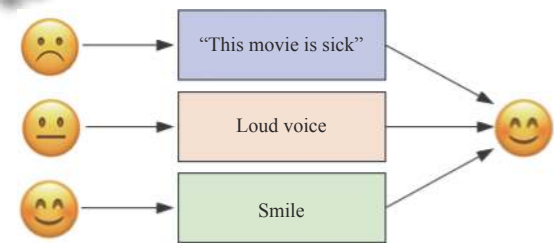

图 1 不同模态之间的差异性表达示例

为了学习到同时具有一致性和差异性信息的多模态表示, 本文提出了一个基于双编码器表示学习的多模态情感分析模型 DERL. DERL 的主要组成部分是特征编码层, 通过所有模态的深层交互来获取高质量的多模态表示. 特征编码层包含两种编码器, 其中一种是基于层级注意力机制的跨模态交互编码器. 考虑到使用成对的跨模态注意力交互存在情感信息流通受限于极少模态之间的问题, 本文采用了层级注意力的方法. 它通过层级递进地在多个模态之间计算多头跨模态注意力, 以学习模态不变表征. 与之前的方法相比, 该方法并没有添加附加的装置来存储多模态信息, 而是直接通过信息的不断前向流通来建立各个模态之间的复杂映射, 学习它们之间的一致性情感线索. 另一种是基于自注意力机制的模态内编码器, 它将提取到的目标模态特征经过多层 Transformer encoder 进行编码, 以学习高水平的语义特征, 然后与模态不变表征相互补充. 此外, 在层与层之间设计了两个门控网络单元, 分别用于增强模态不变表征和过滤模态特定表征内部的噪声. 通过这两个门控单元的不断优化, 最终学习到包含丰富而全面的情感信息的多模态表示.

本文的主要贡献有以下几点.

(1) 本文提出了一个基于双编码器表示学习的多模态情感分析模型 DERL. 该模型通过基于层级注意力机制的跨模态交互编码器捕获模态的一致性信息, 通过基于自注意力机制的模态内编码器学习模态的差异性信息, 并有效地结合了这两种信息.

(2) 本文设计了两个门控网络单元, 分别用于动态增强模态不变表征和过滤模态特定表征内部的噪声. 在融合阶段, 通过减少多模态表示之间的 L2 距离, 捕获不同表示之间的相似信息.

(3) 在 CMU-MOSI[10]和 CMU-MOSEI[11]两个公开多模态情感分析数据集上评估本文所提出的模型 DERL, 实验结果表明, DERL 优于一系列的基线模型.

2 相关工作

2.1 多模态情感分析

多模态情感分析是一个热门的研究方向, 旨在从多模态数据中挖掘情感信息[2]. 它在应对社会突发事件舆论监控和新闻谣言检测等实际应用方面有着重要研究意义[12]. 随着大型多模态数据集的出现, 研究者们提出了许多先进的多模态情感分析模型. Zadeh 等人[13]提出张量融合网络 (tensor fusion network, TFN), 将多模态情感分析问题视为模态内和多模态动力学建模, 通过计算三重笛卡尔积来显式建立单模态、双模态和三模态之间的联系. Liu 等人[14]在 TFN 的基础上应用张量分解方法来降低三重笛卡尔积的计算复杂度. 张量融合网络能够有效地建立模态内和模态间的相互作用, 这是其显著优势. 然而, 它的主要挑战在于其计算复杂度随着特征维数的增加呈指数级增长. 这意味着, 它可能会超出某些计算资源的处理能力. 此外, 张量融合网络在建立模态之间的关联时缺乏细粒度的词级交互, 无法精确地捕捉跨模态语义关联. 为解决此问题, Zadeh 等人[15]将长短期记忆网络 (long short-term memory, LSTM) 应用于每个模态视图, 学习特定于视图的交互作用, 重构 LSTM 记忆网络, 并保留多模态交互信息. Liang 等人[16]提出递归多阶段融合网络 (RMFN), 将融合问题分解为多个阶段, 每个阶段关注多模态信号的一个子集进行有效地融合. 此外, 受机器翻译的启发, 一些工作采用类似结构来建立模态之间的联系. Chauhan 等人[17]采用基于循环神经网络的多模态情感和情绪分析方法 (context-aware interactive attention, CIA), 它通过模态内编码器机制将一种模态编码成另一个模态, 以学习模态之间的相互作用. Yang 等人[18]通过将视觉和音频特征转换为双向编码器 BERT[19]提取的文本特征来提高视觉和音频特征的质量. 吕学强等人[20]提出了一种以文本模态为中心的模态融合策略, 通过带有注意力机制的编解码器网络区分不同模态之间的共有语义和私有语义, 并最终实现情感预测. 由于每个模态数据的采样频率不同, 不同模态的同一序列之间对应关系并不明确, 仅将一种模态编码成另一种模态, 编码后的序列与原序列并不一定对齐, 这将影响最后的情感预测.

2.2 基于注意力机制的多模态情感分析

相较于上述模型, 基于注意力机制的模型在长序列表示分析方面表现出卓越的性能. 它能够充分地提取上下文语义, 并捕获不同模态样本在不同时间步之间的隐式对齐关系. Ou 等人[21]提出了一种多模态局部-全局注意力网络来整合不同模态的表征, 从而产生一种区分性的情感表征. 杨青等人[22]提出具有注意力更新门的 BiGRU 网络, 使用注意力机制对模型进行优化. Wang 等人[23]应用注意门控机制来学习一种视觉和音频模态的非线性组合, 从而产生了非语言的移位向量. 类似地, Rahman 等人[24]引入注意门控记忆, 将文本和非语言线索整合到另一个向量中, 随后将其添加到文本模态. 这些研究致力于将注意力机制与神经网络有机结合, 成功解决了其他模型难以捕捉序列中不同位置依赖关系的问题, 从而在一定程度上提升了模型的预测能力. 为了更充分发挥注意力机制在理解和处理上下文信息方面的潜力, 研究人员创新性地将跨模态注意力机制引入 Transformer 结构. Tsai 等人[4]提出的 MulT 模型首次使用定向成对的跨模态 Transformer, 以跨模态注意力机制捕获多模态序列间不同时间步长上的语义联系, 并潜在地将序列从一个模态调整到另一个模态. Liang 等人[25]指出直接使用成对的跨模态 Transformer 可能导致不同模态之间的分布不匹配问题, 因此提出了模态不变跨模态注意力方法, 旨在学习模态不变空间上的跨模态相互作用. 宋云峰等人[26]提出了一种基于注意力的多层次混合融合的多任务多模态情感分析模型 (multi-level attention and multi-task, MAM), 将注意力机制与多任务学习相结合来学习更泛化的模态特征表示. 包广斌等人[27]通过研究分析相邻话语之间的依赖关系和文本、语音和视频模态之间的交互作用, 建立一种融合上下文和双模态交互注意力的多模态情感分析模型. Sun 等人[7]利用从每个模态

中提取的特征来构造注意力张量. 该方法假设其中一个模态为目标模态, 而其余模态为源模态, 通过计算源模态到目标模态注意力来生成相应的交互特征. 目前, 采用基于跨模态注意力机制的 Transformer 网络架构进行多模态情感分析的研究已经成为主流趋势. 尽管该方法的效果超越了大多数其他模型, 但其关注点在于寻找不同模态之间的相关性, 以增强特征表达能力. 然而, 这可能使最终得到的特征包含大量的冗余信息, 并且无法捕捉不同模态之间的差异性.

2.3 基于表示学习的多模态情感分析

除了采用复杂的多模态融合网络构建模态的映射关系外, 许多研究工作的重点在于如何提取高质量的多模态情感表示. Hazarika 等人[8]提出一种新的架构学习多模态情感分析的模态不变和模态特定表征, 以提供多模态数据的全局视图, 从而帮助预测情感状态. Yu 等人[9]考虑到单模态标签的稀少, 设计了一个基于自监督策略的单模态标签生成模块来辅助多模态任务进行情感预测. 相较于使用跨模态注意力机制融合不同模态间一致性信息的方法, 上述两个研究侧重于学习模态特定的差异性信息, 并将其与模态不变表征相互补充, 这一策略极大地提高了情感预测的准确率. Han 等人[28]提出 MMIM 框架分层地最大化了单模态输入对和多模态融合结果与单模态输入之间的互信息, 以此来优化模态的表示. Mai 等人[29]创造性地将信息瓶颈原理与多模态表示学习相结合以过滤掉噪声信息, 减少冗余来学习最小充分单模态和多模态表示. 程子晨等人[30]根据信息瓶颈理论, 设计了包含两个互信息估计器的互信息估计模块, 寻找一种简洁的、具有较好预测能力的多模态表示向量. 考虑到数据集中提取的原始特征存在大量噪声, 这些研究的主要关注点在于如何在融合前有效改善特征质量并减少噪声. 随着对比学习的广泛应用, 多模态情感分析任务也开始尝试使用对比学习来优化表示学习. Li 等人[31]提出一种基于 Transformer encoder 的多层融合模块, 并使用了两种基于标签和数据的对比学习任务来学习多模态数据中与情感相关的共同特征. Mai 等人[32]提出了一种新的基于对比学习的三模态表示学习框架 HyCon, 它探索了在现有工作中被忽略的样本间和类间关系, 以获得更具区别性的联合嵌入. 对比学习在大规模的多模态预训练模型中表现出强大的表示学习能力, 能够实现文本与图像之间的语义对齐. 在近两年的多模态情感分析研究中, 对比学习被用于实现不同模态间同一序列的语义对齐, 并被用于探索同一类中不同样本之间的关系. 然而, 如何将对比学习与复杂的多模态融合模型相结合, 仍然需要进一步创新.

受上述工作的启发, 我们的工作结合了基于跨模态注意力机制的跨模态交互和多模态表示学习, 旨在深入研究多模态情感分析任务. 传统的研究方法通常将跨模态注意力机制融入 Transformer 编码器, 将目标模态作为 Query, 源模态作为 Key 和 Value, 计算定向的跨模态注意力分数. 不同于其他方法, 我们的研究采用层级注意力机制, 在跨模态交互编码器中通过两次计算跨模态注意力分数, 以充分共享各模态之间的情感信息, 从而学习模态不变表征. 同时, 我们并行使用模态内编码器捕捉不同模态间的差异性信息. 尤为重要的是, 在编码器的每一层中间设计了两个门控网络单元, 以增强模态不变表征并过滤目标模态特定表征中的噪声. 增强单元通过评估目标模态特征对多模态聚合特征的贡献度, 提取目标模态内有效的情感信息, 从而增强模态不变表征. 而过滤单元则利用模态不变表征中的一致性情感信息, 引导目标模态的更新, 以过滤模态内的噪声. 正是通过这两种不同的编码器结构, DERL 逐层学习了模态不变与模态特定表征, 并最终获得了既包含一致性信息又包含差异性信息的多模态表示, 从而更准确地预测情感强度.

3 方法

本文所提出的基于双编码器表示学习的多模态情感分析模型 DERL 结构如图 2 所示, 其主要由特征提取层、特征编码层和多模态融合层 3 部分组成. 特征提取层对原始文本、视觉和音频特征进行上下文语义提取, 得到具有较高质量的模态特征. 特征编码层由两个分支编码器和门控单元构成, 每一个模态有两个不同的编码器: 跨模态交互编码器 CIE (cross-modal interaction encoder) 和模态内编码器 IE (intra-modal encoder). 编码器层之间有增强单元 EU (enhancement unit) 和过滤单元 FU (filter unit) 分别对编码后的特征进行增强和过滤. 多模态融合层首先计算由特征编码层生成的 3 个多模态表示之间的 L2 距离, 并通过减小它们在同一特征空间中的距离, 来捕获一致的情感表达. 随后, 将这 3 个多模态表示拼接, 并通过一个多层感知机 (MLP) 进行融合, 最后输出情感强度值$\hat{y}$作为预测结果.

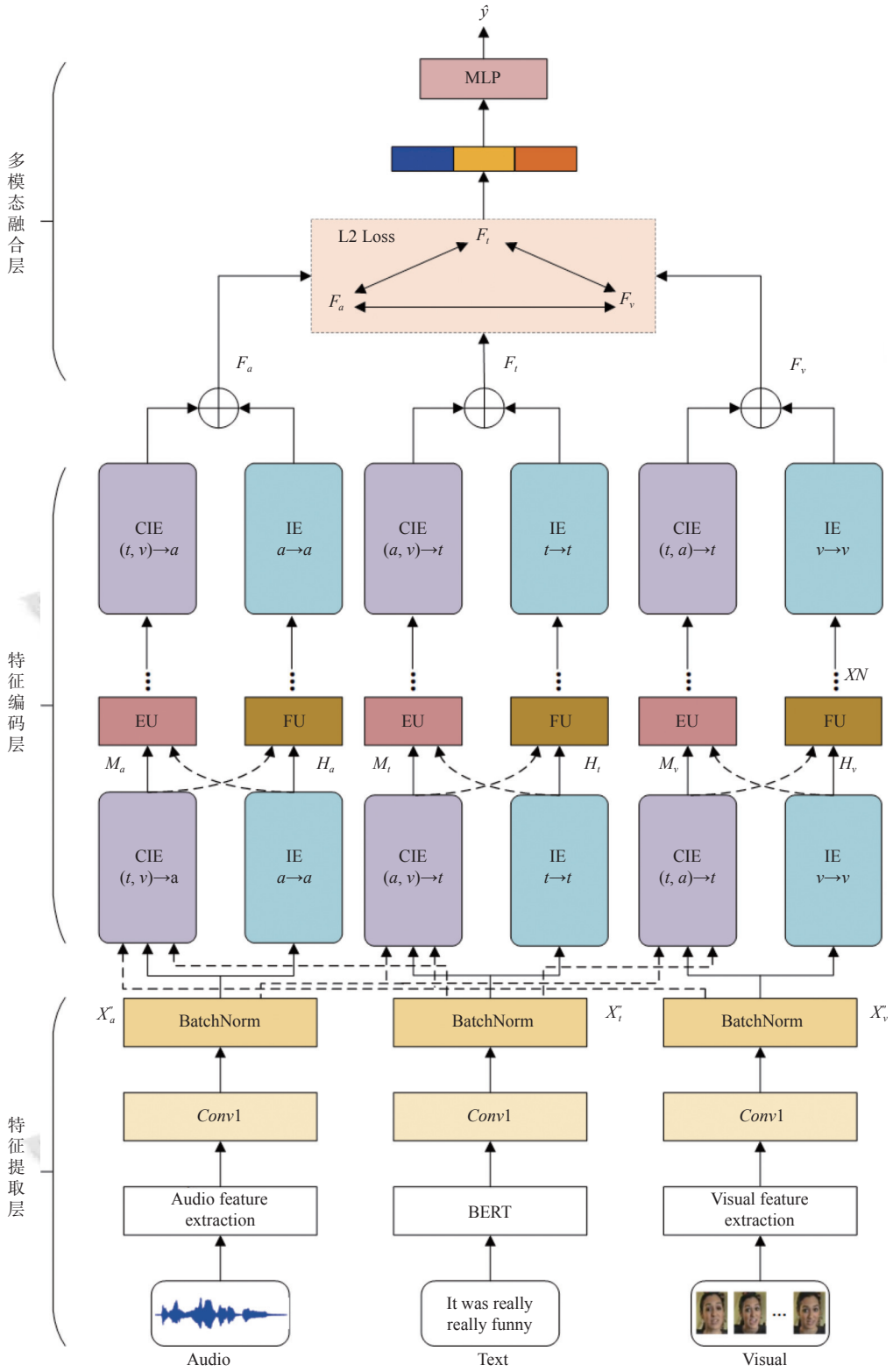

图 2 基于双编码器表示学习的多模态情感分析模型 DERL 整体结构图

3.1 任务定义

多模态情感分析是利用多模态信号来检测视频片段中的所表达的情感. 一个视频片段X由文本 (t)、音频 (a) 和视觉 (v) 序列 3 部分组成, 分别表示为X_m =

$\{x_m^1, x_m^2, \cdots, x_m^{L_m}\} \in R^{L_m \times d_m}$, $m \in \{t,a,v\}$, 其中L_m表示样本的序列长度, d_m表示特征维度, x_t^i、x_a^i、x_v^i分别表示第i个词、音频帧和视觉帧. 目标是给定一个视频序列X预测其情感强度y, 其中$y \in [-3,3]$是一个连续的情感强度变量, $y>0$表示积极情感, $y<0$表示消极情感, $y=0$表示中立情感.

3.2 特征提取层

3.2.1 文本特征提取

模型利用 BERT 预训练模型来提取词嵌入, 每个序列首先与两个特殊的标记“[CLS]”和“[SEP]”拼接, 并分别拼接在每个序列的头部和尾部. 然后, 这些句子被标记化并输入到预训练模型 BERT 中进行特征提取, 最后获得维度为 768 的特征序列X_t.

$$X_t = \text{BERT}(Text) \in R^{L_t \times 768} \tag{1}$$

3.2.2 音频特征提取

对于音频模态, 采用 COVAREP[33]提取 74 维音频特征X_a, 它包括梅尔频谱倒谱系数、音高、声门源参数和其他与情感相关的特征.

3.2.3 视觉特征提取

Facet 工具包用于基于面部动作编码系统 (FACS) 为 CMU-MOSI 和 CMU-MOSEI 数据集提取视觉特征X_v. 视觉特征包括面部动作单位和面部姿势, 它的维度在 CMU-MOSI 中为 47, 在 CMU-MOSEI 中为 35.

由于不同模态序列是在不同的采样率下收集, 为了确保输入序列中每个元素对其局部邻域元素有足够的感知, 我们将提取到的原始特征序列通过一维卷积层来提取局部语义信息, 同时将它们通过批归一化层来稳定特征分布. 最后, 所有模态特征的维度保持统一:

$$X'_m = Conv1D(X_m, K_m) \in R^{L_m \times d} \tag{2}$$

$$X''_m = BatchNorm(X'_m) \in R^{d \times L_m} \tag{3}$$

其中, K_m为各模态对应的卷积核大小, $m \in \{t,a,v\}$, d为统一后的特征维度.

3.3 特征编码层

特征编码层由N层基于 Transformer 的编码器组成, 为了使特征携带位置信息, 首先将 3 个模态的特征经过一个位置嵌入层:

$$Z_m = X''_m + PE(L_m, d) \in R^{L_m \times d}, m \in \{t,a,v\} \tag{4}$$

其中, $PE(\cdot)$表示对序列中的每一个位置索引计算位置嵌入, Z_m为各个模态经过位置编码后的特征, 接下来将这些特征作为输入通过不同的编码器中.

3.3.1 跨模态交互编码器 CIE (cross-modal interaction encoder)

每一个模态特征的编码器层有两个分支, 分别是跨模态交互编码器和模态内编码器. 跨模态交互编码器采用层级注意力机制以获取 3 个模态的模态不变表征. 它的输入为位置编码后 3 个模态的特征, 其中一个模态作为目标模态, 另外两个模态作为源模态, 目标模态通过执行多头跨模态注意力从源模态中学习情感线索, 最后得到包含一致性信息的模态不变表征. 因为 3 个模态分别作为目标模态时工作流程一致, 接下来以文本模态作为目标模态为例, 介绍其具体工作流程, 3 个模态的跨模态交互编码器内部结构如图 3 所示. 在第 1 层的输入中, 得到经过位置编码后的特征Z_t、Z_a与Z_v. 首先, 为了获取计算注意力时所需要的Q、K和V, 分别将文本模态Z_t进行线性映射得到Q, 音频模态Z_a进行线性映射得到K和V, 然后计算它们之间的多头注意力分数$attn_{ta}$, 再经过残差连接和层归一化得到最后的多模态特征M_{ta}, 它包含了文本和音频模态之间的共享情感信息. 随后将M_{ta}作为Q, 视觉模态特征Z_v作为K和V, 再计算一次多头注意力分数$attn_{tav}$, 经过残差连接和层归一化得到 3 个模态交互的多模态特征M_{tav}, 最后通过一个前馈层输出模态不变特征M_t. 它建立了 3 个模态之间的情感映射关系, 使不同模态特征之间的情感线索得到共享:

$$attn_{ta}^i = Softmax\left(\frac{\left(W_t^{Q_i} Z_t\right)\left(W_a^{K_i} Z_a\right)^{\mathrm{T}}}{\sqrt{d}}\right)\left(W_a^{V_i} Z_a\right) \tag{5}$$

$$attn_{ta} = concat\left[attn_{ta}^1, attn_{ta}^2, \cdots, attn_{ta}^h\right] W^{ta} \tag{6}$$

$$M_{ta} = LayerNorm(attn_{ta} + Z_t) \tag{7}$$

$$attn_{tav}^i = Softmax\left(\frac{\left(W_{ta}^{Q_i} M_{ta}\right)\left(W_v^{K_i} Z_v\right)^{\mathrm{T}}}{\sqrt{d}}\right)\left(W_v^{V_i} Z_v\right) \tag{8}$$

$$attn_{tav} = concat\left[attn_{tav}^1, attn_{tav}^2, \cdots, attn_{tav}^h\right] W_{tav} \tag{9}$$

$$M_{tav} = LayerNorm(attn_{tav} + M_{ta}) \tag{10}$$

$$M_t = LayerNorm(Feedforward(M_{tav}) + M_{tav}) \tag{11}$$

其中, $W_t^{Q_i}$、$W_a^{K_i}$、$W_a^{V_i}$、$W_{ta}^{Q_i}$、$W_v^{K_i}$、$W_v^{V_i}$、W_{ta}、

W_{tav} 为可学习的权重矩阵, d 为特征的维度, $attn_{ta}^{i}$、$attn_{tav}^{i}$ 分别表示文本与音频特征第 i 个头的注意力分数, 文本和音频的多模态特征 M_{ta} 与视觉特征第 i 个头的注意力分数, $attn_{ta}$ 和 $attn_{tav}$ 为将所有头拼接后的多头注意力分数, *concat* 表示拼接操作, *Feedforward* 为前馈层, 包括两个全连接层和一个 *ReLU* 激活函数.

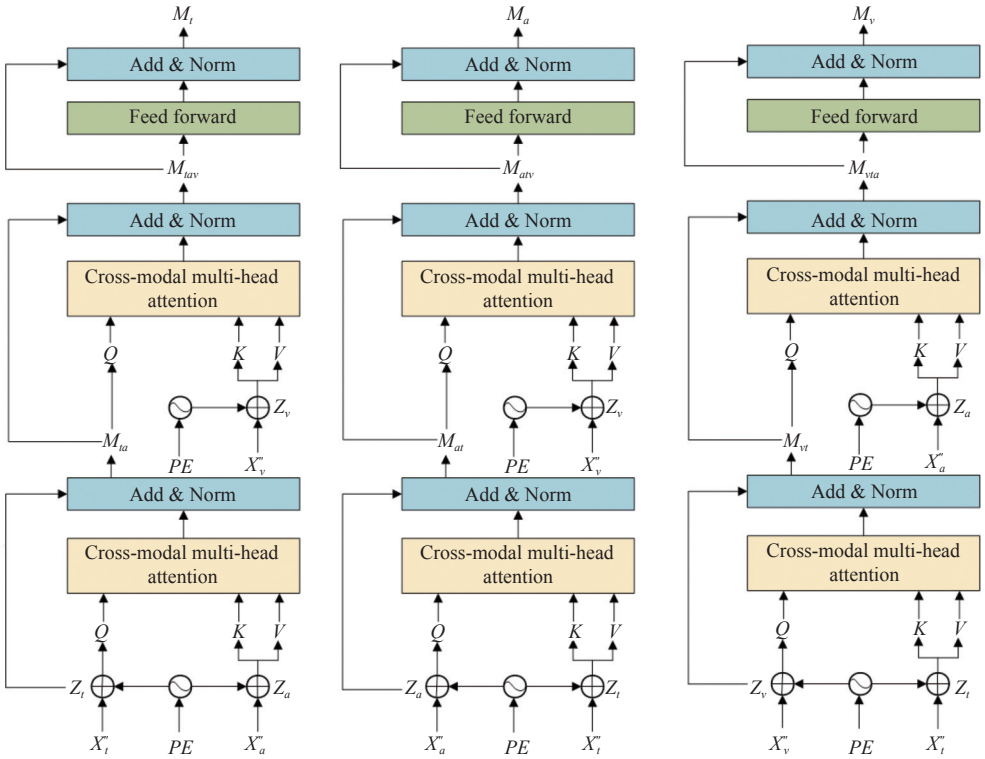

图 3 跨模态交互编码器 CIE 内部结构图

3.3.2 模态内编码器 IE (intra-modal encoder)

模态内编码器为目标模态的模态特定表征编码器. 我们考虑到跨模态交互编码器在执行的过程中, 目标模态会逐渐地向另外两个模态靠拢, 导致自身的语义信息逐层减弱, 最后可能会丧失自身特有的情感语义的问题. 因此, 设计一个与跨模态交互编码器并行的模态内编码器来保留自身的语义信息并过滤低级别特征中包含的噪声. 同样 3 个模态的编码过程一致, 这里主要分析文本模态的编码过程. 首先获取经过位置编码后的特征 Z_t, 将其经过线性映射得到 Q、K 和 V 并计算多头自注意力分数, 然后经过残差连接和归一化层, 最后通过前馈层输出模态特定表征 H_t, 其结构如图 4 所示.

$$attn_t^i = Softmax\left(\frac{\left(W_t^{Q_i'}Z_t\right)\left(W_t^{K_i'}Z_t\right)^{\mathrm{T}}}{\sqrt{d}}\right)\left(W_t^{V_i'}Z_t\right) \tag{12}$$

$$attn_t = concat\left[attn_t^1, attn_t^2, \cdots, attn_t^h\right]W_t \tag{13}$$

$$attn_t' = LayerNorm\left(attn_t + Z_t\right) \tag{14}$$

$$H_t = LayerNorm\left(Feedforward\left(attn_t'\right) + attn_t'\right) \tag{15}$$

其中, $W_t^{Q_i'}$、$W_t^{K_i'}$、$W_t^{V_i'}$、W_t 为可学习的权重矩阵, $attn_t^i$ 为第 i 个头的注意力分数, *concat* 表示拼接操作, *Feedforward* 为前馈层, 包括两个全连接层和一个 *ReLU* 激活函数.

3.3.3 增强单元 EU (enhancement unit)

跨模态交互编码器和模态内编码器所得到的编码特征要么仅具有各模态的相同情感信息, 要么仅包含自身的语义信息. 然而, 更为有效的情感特征向量应该是这两者的有机结合. 一种简单的方法是在层与层之间对它们进行加权求和得到多模态特征, 然后将其输入到下一层的跨模态交互编码器中. 然而, 这种方式并不能获得最优的增强多模态特征. 因此, 我们设计了一个增强单元 EU, 根据单模态特征对聚合情感特征的贡献程度来增强模态不变表征. 具体如下, 先将经过编码得到的模态不变特征 M_m 和模态特定特征 H_m 线性映射到同一特征空间得到多模态特征 M_m' 和单模态特征 H_m', 然后计算两个特征向量的哈达玛积得到聚合情感特征

F. 此时F在这一特征空间中聚合了两个特征中的情感信息. 最后, 通过计算单模态特征H'_m对聚合向量F的贡献程度, 以保留最后增强到模态不变表征中的单模态特征部分:

$$M'_m = M_m W_M \in R^{L_m \times d} \tag{16}$$

$$H'_m = H_m W_H \in R^{L_m \times d} \tag{17}$$

$$F = M'_m \odot H'_m \in R^{L_m \times d} \tag{18}$$

$$WF = Softmax\left(H'_m \frac{F^{\mathrm{T}}}{\|F\|_2}\right) \in R^{L_m \times L_F} \tag{19}$$

$$M''_m = dropout(WF) H'_m + M_m \in R^{L_m \times d} \tag{20}$$

其中, W_M、W_H为可学习的权重矩阵, $m \in \{t,a,v\}$, $\odot$表示哈达玛积, $\|\cdot\|_2$为 L2 范式.

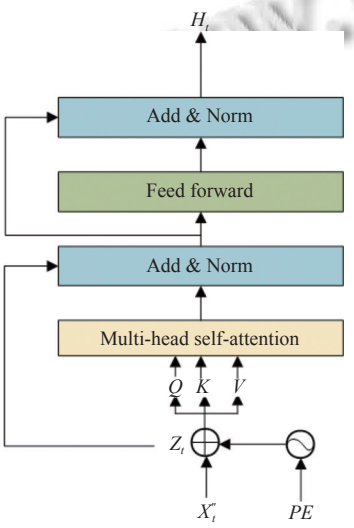

图 4 模态内编码器 IE 内部结构图

3.3.4 过滤单元 FU (filter unit)

单模态原始特征内含有大量的噪声信息, 例如文本中与情感无关的词语, 图像中的背景噪声以及音频中的停顿等, 这些都会影响最后的情感判断. 此外, 随着编码器层数的增加, 单模态特征的语义可能会产生偏移. 因此, 我们设计一个过滤门控单元 FU, 与模态不变特征M_m相结合来过滤模态特定特征H_m内的噪声, 同时使单模态的语义编码过程更加偏向于情感语义编码而忽略掉其他信息, 具体过程如下:

$$G = Sigmoid\left(M_m W'_M + H_m W'_H + b\right) \tag{21}$$

$$H''_m = G \odot H_m + (I - G) \odot M_m \in R^{L_m \times d} \tag{22}$$

其中, W'_M、W'_H为可学习的权重矩阵, b为偏置, $\odot$表示按元素逐位相乘, I为全为 1 的张量.

3.4 多模态融合层

经过 N 层特征编码层后得到 6 个特征向量, 分别是 3 个模态不变特征向量M_m和 3 个模态特定特征向量H_m, $m \in \{t,a,v\}$. 为了使它们对最后的情感预测具有一致的表达, 在同一特征空间中减少它们之间的 L2 距离:

$$F_m = M_m + H_m \tag{23}$$

$$L_{dis} = \frac{1}{3}(dis(F_t,F_a) + dis(F_t,F_v) + dis(F_a,F_v)) \tag{24}$$

其中, $dis(\cdot)$表示 L2 距离. 随后将优化后的F_t, F_a和F_v拼接并通过一层 MLP 进行融合, 最后输出情感强度预测值$\hat{y}$.

$$F = concat[F_t,F_a,F_v] \in R^{L \times 3d} \tag{25}$$

$$F' = FC(ReLU(FC(F))) \in R^{L \times 3d} \tag{26}$$

$$\hat{y} = outlayer(F') \tag{27}$$

其中, *concat* 表示拼接操作, *outlayer* 为一层输出维度为 1 的全连接层.

3.5 训练损失函数

实验过程使用均方误差 *MSE* 作为多模态情感分析任务的损失函数来指导学习过程, 可以表示为:

$$L_{\text{task}} = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 \tag{28}$$

最后模型总的训练目标为:

$$L = L_{\text{task}} + \beta L_{\text{dis}} \tag{29}$$

其中, β为权重系数, 是一个超参数. N为总样本数. 通过不断优化损失L来训练模型.

4 实验

4.1 数据集

在两个公开多模态情感分析数据集 CMU-MOSI 和 CMU-MOSEI 上评估所提出的模型 DERL.

CMU-MOSI: CMU-MOSI[10]是一个从 YouTube 上收集的广泛应用于多模态情感分析的数据集, 演讲者通过独白来表达他们对某些主题的意见. 其中训练集有 1 281 个话语, 验证集有 229 个话语, 测试集有 685 个话语, 每个话语的情感强度都在[–3, 3]范围内, 其中 +3 和–3 分别代表强积极和强消极的情感.

CMU-MOSEI: CMU-MOSEI[11]是对 CMU-MOSI 的改进, 具有更多的话语数量, 样本、演讲者和主题.

具体来说, 训练集有 16 265 个话语, 验证集有 1 869 个话语, 测试集有 4 643 个话语.

4.2 评价指标与实验细节

多模态情感分析任务通常被视为回归任务, 为了对基于双编码器表示学习的多模态情感分析模型 DERL 进行全面的评价, 采用了平均绝对误差 (*MAE*) 和皮尔逊相关系数 (*Corr*) 等各种标准回归任务评价方法. *MAE* 与 *Corr* 的具体计算公式如下:

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i| \tag{30}$$

$$Corr = \frac{cov(y,\hat{y})}{\sigma y \sigma \hat{y}} = \frac{E[(y-\mu y)(\hat{y}-\mu\hat{y})]}{\sigma y \sigma \hat{y}} \tag{31}$$

其中, N 表示总样本数, y 表示真实标签值, $\hat{y}$ 表示预测值, σ 是标准差, μ 是期望, $cov(y,\hat{y})$ 是 y 和 $\hat{y}$ 之间的协方差.

此外, 为了与其他任务保持一致, 二分类准确率 (Acc-2)、七分类准确率 (Acc-7) 和 $F1$ 分数 ($F1$) 也用于评估模型的性能. 上述评价指标中 *MAE* 越低表示模型的性能越好, 其他指标则越高表示模型性能越好.

模型代码基于 PyTorch 框架构建, 在 24 GB 显存的 NVIDIA A30 GPU 计算资源上进行训练和测试. 使用 Adam 优化器来训练模型, 在两个数据集上的批处理大小均为 48, BERT 微调时学习率为 1E–5, 模型学习率在 CMU-MOSI 上为 5E–5, CMU-MOSEI 上为 2E–5. 经过实验选取最优特征编码层数分别为 2 和 5, 注意力头数均为 10, 一维卷积的核大小在所有模态上都为 3. 具体实验设置如表 1 所示.

表 1 实验设置

Setting	CMU-MOSI	CMU-MOSEI
Optimizer	Adam	Adam
Batch size	48	48
Learning rate	5E–5	2E–5
BERT learning rate	1E–5	1E–5
Feature size	50	50
Attention head	10	10
Transformer layer N	2	5
Kernel size (t/v/a)	3/3/3	3/3/3
β	0.005	0.5

4.3 基线方法

为了验证基于双编码器表示学习的多模态情感分析模型 DERL 的有效性, 将实验结果与一系列基线方法进行了比较.

TFN[13]: 该模型通过计算模态向量的外积得到一个多维张量, 获取单模态、双模态和三模态之间的相互作用.

LMF[14]: 该模型是对 TFN 的一种改进, 它采用低秩多模态张量融合技术来提高效率.

MFN[15]: 该模型分别通过 LSTM 和特殊注意力机制学习视图特定交互和跨视图交互, 然后通过多视图门控记忆进行时间总结.

RAVEN[23]: 该模型学习非语言子词序列的细粒度结构, 并基于非语言线索动态转移单词表示, 有效地解决了词汇语义表示不足的问题.

MCTN[34]: 该模型通过从源模态到目标模态的循环翻译学习联合表示.

MulT[4]: 该模型利用跨模态注意力机制提供的潜在跨模态自适应, 通过直接关注其他模态中的低层次特征而不需要对齐来融合多模态信息.

MFM[35]: 该模型由生成网络和判别网络组成, 通过同时优化两个网络得到多模态表示.

ICCN[36]: 该模型使用深度典型相关分析 (DCCA) 来分析文本、音频和视觉之间的隐藏关系.

MAG-BERT[24]: 该模型将非语言行为映射到具有轨迹和大小的向量, 随后用于在 BERT 中转移词汇表示.

MISA[8]: 该模型结合了包括分布相似性损失、正交损失、重建损失和任务预测损失来学习模态不变和模态特定表征.

Self-MM[9]: 该模型设计了一个基于自监督学习策略的标签生成模块, 以生成特定的单模态标签. 然后进行多模态任务和单模态任务的联合训练, 以探索模态之间的一致性和差异性.

4.4 实验结果

表 2 和表 3 分别展示了 DERL 与基线模型在 CMU-MOSEI 和 CMU-MOSI 两个数据集上的对比实验结果, 其中“—”表示原论文中没有报道该项数据, 加粗内容则表示该项指标表现最好的模型. 我们可以看到 DERL 在两个公开数据集上都表现出非常好的效果, 并且在所有评价指标上高于基线模型或与基线模型持平. 具体来说, 在 CMU-MOSEI 数据集上平均绝对误差 *MAE* 为 0.530, 相关系数 *Corr* 为 0.770, 二分类准确率 Acc-2 为 86.5%, 七分类准确率 Acc-7 为 54.1%, $F1$ 分数为 86.5%. 其中 Acc-2 和 Acc-7 比最好的基线模型 MISA 分别高出 1% 和 1.9%, 其他评估指标也有显著的提升. CMU-MOSI 数据集的规模大约是 CMU-MOSEI

的 1/10, DERL 同样能够很好地适用于该数据集. 在 CMU-MOSI 数据集上平均绝对误差 *MAE* 为 0.693, 相关系数 *Corr* 为 0.798, 二分类准确率 Acc-2 为 86.6%, 七分类准确率 Acc-7 为 48.3%, *F*1 分数为 86.5%. 同样, 在 Acc-2 和 Acc-7 上比最好的基线模型 MAG-BERT 和 MISA 分别高出 0.5% 和 6.0%.

表 2 DERL 与基线模型在 CMU-MOSEI 数据集上的对比结果

Model	*MAE*	*Corr*	Acc-2 (%)	*F*1 (%)	Acc-7 (%)
MFN	—	—	76.0	76.0	—
RAVEN	0.614	0.662	79.1	79.5	50.0
MCTN	0.609	0.670	79.8	80.6	49.6
MulT	0.580	0.703	82.5	82.3	51.8
TFN	0.593	0.700	82.5	82.1	50.2
LMF	0.623	0.677	82.0	82.1	48.0
MFM	0.568	0.717	84.4	84.3	51.3
ICCN	0.565	0.713	84.2	84.2	51.6
MISA	0.555	0.756	85.5	85.3	52.2
Self-MM	0.530	0.765	85.2	85.3	—
DERL	**0.530**	**0.770**	**86.5**	**86.5**	**54.1**

表 3 DERL 与基线模型在 CMU-MOSI 数据集上的对比结果

Model	*MAE*	*Corr*	Acc-2 (%)	*F*1 (%)	Acc-7 (%)
MFN	0.965	0.632	77.4	77.3	34.1
RAVEN	0.915	0.691	78.0	76.6	33.2
MCTN	0.909	0.676	79.3	79.1	35.6
MulT	0.871	0.698	83.0	82.8	40.0
TFN	0.901	0.698	80.8	80.7	34.9
LMF	0.917	0.695	82.5	82.4	33.2
MFM	0.951	0.662	78.1	78.1	36.2
ICCN	0.860	0.710	83.0	83.0	39.0
MAG-BERT	0.784	0.782	84.3	84.3	—
MISA	0.783	0.761	83.4	83.6	42.3
Self-MM	0.713	0.798	86.0	86.0	—
DERL	**0.693**	**0.798**	**86.6**	**86.5**	**48.3**

MFN、RAVEN、MCTN、TFN 以及 LMF 等早期工作使用传统的融合方式来建立模态间的交互, 而 DERL 使用 Transformer 架构学习模态间的一致性情感信, 因此, 其性能远超这些模型. MulT 和 DERL 在设计上有相似之处, 均运用了基于跨模态注意力机制的方法来实现模态间的交互. 然而, MulT 采用的是定向成对的跨模态 Transformer, 这产生了大量的冗余信息, 并且未能考虑到不同模态之间的差异性信息. 因此, 其在预测准确率方面远不及 DERL. MAG-BERT 将非语言模态语义集成到文本模态上, 然后使用 BERT 学习非语言的位移向量, 它在一定程度上降低了冗余信息, 因此其预测准确率超越了 MulT. 然而, 它同样忽略了模态间的差异性信息. MISA 和 Self-MM 基于模态不变与模态特定表示学习理念, 通过在最终预测任务的损失函数中加入不同子任务学习的正则化项, 以获得高质量的多模态表示. 这些方法的实验结果优于大部分仅专注于融合一致性情感以增强多模态表示的方法, 表明结合一致性和差异性信息的多模态表示在提升情感预测准确率方面更为有效. 尽管如此, 这些方法在模态融合上的策略相对简单, 例如 Self-MM 只是通过拼接不同模态的特征得到多模态联合特征向量, 再将它通过线性层建立模态间的交互. 相比之下, DERL 采用了层级注意力机制来捕捉各个模态之间的情感线索, 因此, 它在性能上超越了这两个模型.

与所有基线模型相比, DERL 在情感强度的预测方面表现得非常出色. 实验结果表明, 基于双编码器表示学习的多模态情感分析模型 DERL 在大数据集 CMU-MOSEI 和较小的数据集 CMU-MOSI 上都取得了出色的效果, 这表明 DERL 将基于层级注意力的跨模态交互编码和目标模态编码相结合是有效的, 并适用于不同的数据场景. 此外, 设计的两个门控网络单元也取得了预期的效果, 增强单元 EU 通过利用高水平的目标模态特征向量增强模态不变表征, 很好地将多模态信息与单模态特定语义相结合, 进行互补. 过滤单元 FU 也能够保留单模态表征的有用信息, 过滤掉内部无关的噪声. 利用 L2 距离减少多模态表示之间的语义距离, 也能够使多模态表示对最后的情感表达具有一致性. 综上所述, 这些网络结构对多模态情感分析任务具有一定程度的贡献.

4.5 消融实验

为了验证 DERL 模型各部分的有效性, 在较大的数据集 CMU-MOSEI 上进行了一系列的消融实验.

4.5.1 不同编码器的组合

基于双编码器表示学习的多模态情感分析模型的主要框架由两个分支编码器组成, 因此, 我们探究了不同编码器组合的有效性, 实验结果如表 4 所示. 表 4 前 4 行表示仅使用模态内编码器, 其中 *t*, *a*, *v* 分别表示每次仅使用 1 个模态的数据, 经过比较可以发现文本模态经过编码后对情感的预测比音频和视觉两个模态都要准确. 因为文本模态经过微调过的 BERT 预训练模型, 本身它就具有良好的语义信息, 而原始的音频和视觉特征则是包含大量噪声的低水平特征, 因此对情感的预测效果较差. 当将 3 个模态的模态特定表征进行拼接然后预测时, 即表中的 *t*+*a*+*v*, 可以发现 Acc-2 比

单模态的音频和视觉特征高出近 20%. 然而, 与单模态的文本特征相比, 提升并不显著. 这可能是因为其他两个模态特征引入了噪声, 从而影响了预测结果. 采用这种简单的晚期融合方式并不能有效消除噪声的影响, 但总体效果仍优于单模态的文本编码器. 因此, 可以得出多模态数据相较于单模态数据更有利于情感分析任务.

表 4 CMU-MOSEI 数据集上不同编码器组合的消融实验结果

Model	Modalities	*MAE*	*Corr*	Acc-2 (%)	*F*1 (%)	Acc-7 (%)
仅模态内编码器 (IE)	*t*	0.551	0.757	84.6	84.7	51.3
	a	0.837	0.254	65.6	61.8	38.2
	v	0.875	0.281	64.6	63.5	37.1
	t+a+v	0.547	0.764	84.7	84.7	51.9
仅跨模态交互编码器 (CIE)	*tav*	0.550	0.757	85.2	85.2	52.4
	atv	0.540	0.759	85.3	85.3	52.9
	vta	0.547	0.756	85.0	85.0	52.3
	tav+atv+vta	0.538	0.765	85.4	85.4	53.2
完整模型	**DERL**	**0.530**	**0.770**	**86.5**	**86.5**	**54.1**

表 4 中间 4 行表示仅使用基于层级注意力的跨模态交互编码器对 3 个模态的特征进行跨模态交互. *tav*、*atv*、*vta* 分别表示以文本、音频和视觉特征作为目标模态, 其他两个模态作为源模态. 相较于简单的晚期融合方式, 层级注意力机制更有利于提高分类准确率. 各项准确率都有显著提升, 其中 Acc-7 高出 1%, 验证了层级注意力机制的有效性. 此外, 将 3 个模态的模态不变表征拼接并预测情感强度值, 各项指标相较于单独使用一个模态不变表征进行预测都有一定程度的提升.

表格最后一行是完整的 DERL 的实验结果, 在各类指标都优于以上不同的模型结构. Acc-2 比仅使用音频和视觉单模态编码器高出 20% 以上, 比最佳的 3 个跨模态交互编码器组合高出 1.1%, Acc-7 也分别高出近 16% 和 1.1%. 通过实验结果的对比, 充分体现了将跨模态交互编码和模态内编码器相结合的想法是正确的. 它既能够学习到不同模态特征之间的共同情感线索, 又能够有效地将私有于目标模态的模态特定情感语义整合, 从而对话语的情感做出全面的预测.

4.5.2 不同模块的组合

为了深入分析模型的不同模块对整个模型性能的影响, 我们通过删除不同模块进行实验来评估各部分的有效性, 实验结果如表 5 所示. 从表中可以看出, 当删除单模态过滤单元 FU 时, 模型的性能有所下降, 表明学习到的模态不变表征包含一定的噪声, 而过滤单元的设计能够在一定程度上去除噪声. 同样, 当删除增强单元 EU 时, 模型效果也有所下降, 这表明设计的增强单元能够将单模态的特定语义有效地增强至模态不变表征中, 使其更加全面. 当同时删除 FU 和 EU 时, 这一结论表现得更加明显, Acc-2 下降了 1.1%. 实验结果突显出针对两个分支编码器所设计的中间层网络单元对多模态情感表示学习的重要性. 最后验证了融合部分 L2 损失的有效性, 缺少 L2 损失时模型的性能下降, 表明在融合前从相关的多模态表示中学习相似情感表达, 确保这些特征在一个共同的投影空间中更接近有助于提升预测的准确性.

表 5 CMU-MOSEI 数据集上不同模块的消融实验结果

Model	*MAE*	*Corr*	Acc-2 (%)	*F*1 (%)	Acc-7 (%)
w/o FU	0.532	0.765	85.8	85.7	53.9
w/o EU	0.534	0.767	85.8	85.8	53.7
w/o FU & EU	0.538	0.760	85.4	85.4	53.4
w/o L2 Loss	0.533	0.768	86.0	86.0	53.8
DERL	**0.530**	**0.770**	**86.5**	**86.5**	**54.1**

综上所述, DERL 中所设计的网络模块对于最终的模型性能都有相当程度的贡献.

4.6 编码器层数的对比

为了找出最适合的编码器层数, 我们在 CMU-MOSEI 和 CMU-MOSI 两个数据集上进行了对比实验, 结果如图 5 所示. 实验中, 编码器从 1 层增加至 6 层. 在图 5 中, CMU-MOSEI 数据集的结果以蓝色折线表示, 实验结果表明在编码器层数为 5 时, Acc-2 达到最高, 为 86.5%. 而 CMU-MOSI 数据集的结果则以黄色折线表示, 当编码器层数为 2 时, Acc-2 达到最高点, 为 86.6%. 折线图的趋势表明, 编码器层数过少不足以充分学习多模态表示, 从而使模型的预测准确率无法达到最高. 同时, 层数过多则容易造成模型过拟合, 使预测准确率出现抖动. 两者均会影响模型的最终预测效果. 此外, CMU-MOSI 数据集在 Acc-2 上比 CMU-MOSEI 更早达到最高点, 可能的原因是 CMU-MOSI 的样本数量远少于 CMU-MOSEI, 所以浅层模型已足以拟合所有样本. 因此, DERL 针对不同数据集选择适当的编码器层数.

5 结论与展望

本文提出了一个基于双编码器表示学习的多模态情感分析模型 DERL, 该模型通过将模态不变和模态特定表征相结合来学习高质量的多模态情感表示. 在跨模态交互编码器中, 采用层级注意力机制, 以建立各个模态之间的复杂映射关系, 使最终的模态不变表征

能够捕捉到所有模态中的共享信息. 基于双编码器结构分别设计了两个网络单元, 特征增强单元 EU 用于将目标模态特征中特定情感语义增强到模态不变表征中, 过滤单元 FU 用于过滤模态特定表征中的噪声和对于情感预测无关的部分. 最后, 融合前通过在同一特征空间中缩小多模态表示之间的 L2 距离, 从而捕获不同表示之间的潜在相似性. DERL 模型在 CMU-MOSEI 和 CMU-MOSI 两个流行数据集上的实验结果超越一系列基线模型, 并通过不同的消融实验, 验证了模型中各部分的有效性. 未来, 我们将使用更加优秀的策略来结合模态不变与模态特定表征, 可以引入数学知识设计损失函数来聚合两个特征之间的关联部分. 同时, 我们考虑将对比学习应用至同序列或同类的两个特征之间, 探索它们之间的潜在联系.

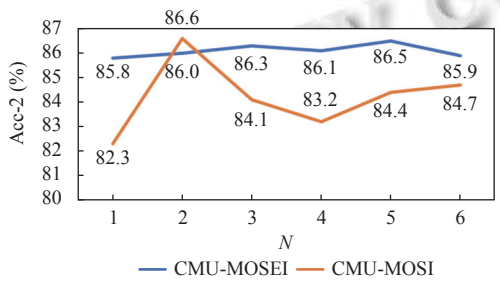

图 5 不同编码器层数的二分类准确率 Acc-2

参考文献

1 Mehendale N. Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2020, 2(3): 446. [doi: 10.1007/s42452-020-2234-1]

2 刘青文, 买日旦·吾守尔, 古兰拜尔·吐尔洪. 双元双模态下二次门控融合的多模态情感分析. 计算机工程与应用. http://kns.cnki.net/kcms/detail/11.2127.TP.20230613.0928.004.html. (在线出版)(2023-06-13).

3 Li ZH, Guo QB, Pan YS, *et al*. Multi-level correlation mining framework with self-supervised label generation for multimodal sentiment analysis. Information Fusion, 2023, 99: 101891. [doi: 10.1016/j.inffus.2023.101891]

4 Tsai YHH, Bai SJ, Liang PP, *et al*. Multimodal transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 6558–6569.

5 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

6 Lv FM, Chen X, Huang YY, *et al*. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2554–2562.

7 Sun H, Chen YW, Lin LF. TensorFormer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection. IEEE Transactions on Affective Computing, 2023, 14(4): 2776–2786. [doi: 10.1109/TAFFC.2022.3233070]

8 Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020. 1122–1131. [doi: 10.1145/3394171.3413678]

9 Yu WM, Xu H, Yuan ZQ, *et al*. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 10790–10797. [doi: 10.1609/aaai.v35i12.17289]

10 Zadeh A, Zellers R, Pincus E, *et al*. 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]

11 Zadeh AAB, Liang PP, Poria S, *et al*. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 2236–2246. [doi: 10.18653/v1/P18-1208]

12 陈志毅, 隋杰. 基于 DeepFM 和卷积神经网络的集成式多模态谣言检测方法. 计算机科学, 2022, 49(1): 101–107. [doi: 10.11896/jsjkx.201200007]

13 Zadeh A, Chen MH, Poria S, *et al*. Tensor fusion network for multimodal sentiment analysis. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017. 1103–1114. [doi: 10.18653/v1/D17-1115]

14 Liu Z, Shen Y, Lakshminarasimhan VB, *et al*. Efficient low-rank multimodal fusion with modality-specific factors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 2247–2256. [doi: 10.18653/v1/P18-1209]

15 Zadeh A, Liang PP, Mazumder N, *et al*. Memory fusion network for multi-view sequential learning. Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018. 5634–5641. [doi: 10.1609/aaai.v32i1.12021]

16 Liang PP, Liu ZY, Zadeh AAB, *et al*. Multimodal language analysis with recurrent multistage fusion. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 150–161. [doi: 10.18653/v1/D18-1014]

17 Chauhan DS, Akhtar S, Ekbal A, *et al*. Context-aware interactive attention for multi-modal sentiment and emotion analysis. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019. 5647–5657. [doi: 10.18653/v1/D19-1566]

18 Yang B, Shao B, Wu LJ, *et al*. Multimodal sentiment analysis with unidirectional modality translation. Neurocomputing, 2022, 467: 130–137. [doi: 10.1016/j.neucom.2021.09.041]

19 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

20 吕学强, 田驰, 张乐, 等. 融合多特征和注意力机制的多模态情感分析模型. 数据分析与知识发现. http://kns.cnki.net/kcms/detail/10.1478.G2.20230508.1710.008.html. (在线出版)(2023-05-10).

21 Ou YJ, Chen ZZ, Wu F. Multimodal local-global attention network for affective video content analysis. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(5): 1901–1914. [doi: 10.1109/TCSVT.2020.3014889]

22 杨青, 张亚文, 朱丽, 等. 基于注意力机制和 BiGRU 融合的文本情感分析. 计算机科学, 2021, 48(11): 307–311. [doi: 10.11896/jsjkx.201000075]

23 Wang YS, Shen Y, Liu Z, *et al*. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 7216–7223. [doi: 10.1609/aaai.v33i01.33017216]

24 Rahman W, Hasan K, Lee S, *et al*. Integrating multimodal information in large pretrained Transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 2359–2369.

25 Liang T, Lin GS, Feng L, *et al*. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 8148–8156. [doi: 10.1109/ICCV48922.2021.00804]

26 宋云峰, 任鸽, 杨勇, 等. 基于注意力的多层次混合融合的多任务多模态情感分析. 计算机应用研究, 2022, 39(3): 716–720. [doi: 10.19734/j.issn.1001-3695.2021.08.0357]

27 包广斌, 李港乐, 王国雄. 面向多模态情感分析的双模态交互注意力. 计算机科学与探索, 2022, 16(4): 909–916. [doi: 10.3778/j.issn.1673-9418.2105071]

28 Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 9180–9192.

29 Mai SJ, Zeng Y, Hu HF. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. IEEE Transactions on Multimedia, 2023, 25: 4121–4134. [doi: 10.1109/TMM.2022.3171679]

30 程子晨, 李彦, 葛江炜, 等. 利用信息瓶颈的多模态情感分析. 计算机工程与应用. 2024, 60(2): 137–146.

31 Li Z, Xu B, Zhu CH, *et al*. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. Proceedings of the 2022 Findings of the Association for Computational Linguistics. Seattle: ACL, 2022. 2282–2294.

32 Mai SJ, Zeng Y, Zheng SJ, *et al*. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 2023, 14(3): 2276–2289. [doi: 10.1109/TAFFC.2022.3172360]

33 Degottex G, Kane J, Drugman T, *et al*. COVAREP-A collaborative voice analysis repository for speech technologies. Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence: IEEE, 2014. 960–964. [doi: 10.1109/ICASSP.2014.6853739]

34 Pham H, Liang PP, Manzini T, *et al*. Found in translation: Learning robust joint representations by cyclic translations between modalities. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 6892–6899. [doi: 10.1609/aaai.v33i01.33016892]

35 Tsai YHH, Liang PP, Zadeh A, *et al*. Learning factorized multimodal representations. Proceedings of the 7th International Conference on Learning Representations. New Orleans: OpenReview.net, 2019.

36 Sun ZK, Sarma P, Sethares W, *et al*. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 8992–8999. [doi: 10.1609/aaai.v34i05.6431]

(校对责编: 牛欣悦)