计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(5):280−287 [doi: 10.15888/j.cnki.csa.009492]

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

结合模态表征学习的多模态情感分析①

刘若尘[1], 冯　广[2], 罗良语[1], 林浩泽[2]

[1](广东工业大学 计算机学院, 广州 510006)
[2](广东工业大学 自动化学院, 广州 510006)
通信作者: 冯　广, E-mail: von@gdut.edu.cn

摘　要: 在当前视频多模态情感分析研究中, 存在着未充分考虑模态之间的动态独立性和模态融合缺乏信息流控制的问题. 为解决这些问题, 本文提出了一种结合模态表征学习的多模态情感分析模型. 首先, 通过使用 BERT 和 LSTM 分别挖掘文本、音频和视频的内在信息, 其次, 引入模态表征学习, 以获得更具信息丰富性的单模态特征. 在模态融合阶段, 融合了门控机制, 对传统的 Transformer 融合机制进行改进, 以更精确地控制信息流. 在公开数据集 CMU-MOSI 和 CMU-MOSEI 的实验结果表明, 与传统模型相比, 准确性和 *F*1 分数都有所提升, 验证了模型的有效性.

关键词: 多模态情感分析; 表征学习; 特征融合; 门控机制; 多头注意力机制

引用格式: 刘若尘,冯广,罗良语,林浩泽.结合模态表征学习的多模态情感分析.计算机系统应用,2024,33(5):280–287. http://www.c-s-a.org.cn/1003-3254/9492.html

Multimodal Sentiment Analysis Incorporating Modal Representation Learning

LIU Ruo-Chen[1], FENG Guang[2], LUO Liang-Yu[1], LIN Hao-Ze[2]

[1](School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)
[2](School of Automation, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: In the context of current multi-modal emotion analysis in videos, the influence of modality representation learning on modality fusion and final classification results has not been adequately considered. To this end, this study proposes a multi-modal emotion analysis model that integrates cross-modal representation learning. Firstly, the study utilizes Bert and LSTM to extract internal information from text, audio, and visual modalities separately, followed by cross-modal representation learning to obtain more information-rich unimodal features. In the modal fusion stage, the study fuses the gating mechanism and improves the traditional Transformer fusion mechanism to control the information flow more accurately. Experimental results on the publicly available CMU-MOSI and CMU-MOSEI datasets demonstrate that the accuracy and *F*1 score of this model are improved compared with the traditional models, validating the effectiveness of this model.

Key words: multimodal sentiment analysis; representation learning; feature fusion; gating mechanism; multi-head attention mechanism

随着互联网和新媒体技术的蓬勃发展, 社交媒体平台、短视频应用以及各种在线社交互动逐渐成为人们表达情感和观点的主要方式. 用户通过文字、图像、音频和视频等多种模态, 向世界传达他们对事件、产品、甚至生活的情感和看法. 这一现象带来了海量的多模态数据, 其中包含着丰富的情感信息. 多模态情感分析 (multimodal sentiment analysis) 作为一门新兴的研究领域, 旨在利用多种模态信息, 如文本、图

① 基金项目: 国家自然科学基金 (62237001); 广东省哲学社会科学青年项目 (GD23YJY08)
收稿时间: 2023-11-21; 修改时间: 2023-12-22; 采用时间: 2023-12-27; csa 在线出版时间: 2024-03-15
CNKI 网络首发时间: 2024-03-19

像、音频和视频等, 来深入理解和预测用户的情感和情感倾向[1,2].

这些多模态数据不仅包含情感信息的丰富性, 而且具有多样性. 用户的情感表达不再局限于单一模态, 而是在多种数据形式中相互交织. 情感信息可能隐藏在文字评论的背后、图像中人物的面部表情、音频中的语调和情感色彩等. 因此, 传统的情感分析方法, 主要基于单一模态的文本数据, 已不足以满足多模态数据带来的情感分析需求[3]. 多模态情感分析的兴起, 不仅有助于增强我们对用户情感的理解, 还扩展了应用领域的广度. 多模态情感分析对于舆情监测、智能推荐系统、心理健康评估、情感驱动的产品设计等领域具有重要价值. 通过综合考虑来自不同模态的信息, 多模态情感分析为这些应用提供了更加全面、准确和有深度的情感洞察[4,5].

随着深度学习和神经网络技术的不断发展, 多模态情感分析取得了显著进展. 然而, 其中仍然存在一系列挑战.

(1) 尽管多模态情感分析模型在整合不同模态信息方面取得了显著进展, 但现有模型主要集中于多模态数据的融合, 却往往忽视了单模态数据的内部独特性[6]. 每种模态, 例如文本、音频和图像, 都包含着独特的情感线索, 但这些线索往往被现有模型忽略. 这种忽视可能导致对情感分析任务的不足, 因为单模态信息的独特性未能充分挖掘和利用.

(2) 传统的多模态融合技术通常缺乏适应性, 无法有效地处理不同模态的信息流[7]. 这可能导致在融合过程中信息丢失或混淆, 影响情感分析的准确性. 传统方法通常无法有效解决模态不平衡的问题, 使得一些模态的信息被低估, 从而降低了情感分析的性能.

为解决上述问题, 本文提出了一种结合模态表征学习的多模态情感分析模型. 本模型的独特之处在于它不仅致力于实现多模态数据的融合, 还注重学习每个模态内部的情感特征. 提供了更全面的情感分析视角, 有助于克服现有模型忽视的单模态内部信息. 这使得我们的模型能够更准确地预测情感. 此外, 我们引入了一种基于门控机制的创新控制方法, 以增强基于 Transformer 架构的融合过程. 我们引入的门控机制允让模型动态调节信息流, 确保每种模态的贡献根据其与特定上下文的相关性得到适当加权. 这种门控机制不仅增强了融合过程的适应性, 还解决了模态不平衡的问题. 在许多多模态数据集中, 某些模态可能对于特定任务来说信息较少或不太可靠. 控制机制有助于降低或甚至忽略这些信息较少的模态, 从而使预测的结果更加具有准确性和鲁棒性.

1 相关工作

多模态情感分析是一个涉及自然语言处理、计算机视觉、语音识别等多个领域的交叉研究领域. 研究人员的关注点主要集中在情感分析、情感识别和个性特征识别等任务上. 多模态情感分析的研究通常包括多模态数据的表示和多模态数据的融合两个主要方面. Zadeh 等[8]提出了一种创新的方法, 即张量融合网络, 通过利用张量的笛卡尔积将每个模态的特征表示进行整合. 这一方法在综合多模态信息时取得了显著的性能改进. 与此同时, Liu 等[9]采用低阶多模态融合方法对权重张量进行了分解, 从而降低了计算复杂性. 这种方法通过与模态特定的低阶因子进行高效的多模态融合, 实现了对模态特定和跨模态相互作用的学习. 随着注意力机制在多模态融合中的重要性日益凸显, Tsai 等[10]引入了定向成对的跨模态注意机制, 进一步提高了模型的性能. 另一方面, Ghosal 等[11]提出了一种名为 MMMU-BA 的模型, 该模型采用了双注意力机制, 通过充分利用所有模态对内部跨模态的上下文交互信息, 取得了令人瞩目的结果. Pham 等[12]提出了一种基于机器翻译的序列到序列 (sequence to sequence, Seq2Seq) 模型, 该模型利用循环神经网络提取各模态的时序特征. 通过编码-解码过程学习模态之间的关联性, 将编码后的上下文特征作为跨模态的融合特征表示, 为多模态情感分析提供了一种有效的解决方案. 然而, Zhan 等[13]的研究表明, 噪声是跨模态转换中的常见问题, 对模型性能造成了一定影响. 因此, 在对跨模态交互信息进行建模时, 必须考虑去除模态之间的噪声干扰. 在这个方向上的探索中, Han 等[14]提出了分层最大化互信息的框架, 有效地减少有价值的任务相关信息的丢失.

总的来说, 研究者们在多模态情感分析领域取得了显著的进展. 从早期的基于机器学习的方法到近年来的注意力机制和张量融合网络, 各种技术都为提高模型性能和挖掘跨模态关联性提供了丰富的选择. 不同模型在处理多模态信息时呈现出各自的优势, 有的侧重于处理特定类型的噪声, 有的注重模态之间的关联性学习. 然而, 仍然存在一些挑战需要进一步解决.

噪声的处理、模态不平衡、模态融合的高效性等问题仍然是当前研究的热点和难点. 未来的工作可以探索更多创新性的模型结构, 结合深度学习、注意力机制等先进技术, 以更好地实现模态之间的有效融合和情感信息的准确捕捉.

2 结合模态表征学习的多模态情感分析方法

本文提出的结合模态表征学习的多模态情感分析模型框架图如图 1 所示, 模型主要由以下 4 个部分组成, 分别为多模态特征提取、模态表征学习、多模态融合以及最后的情感分类预测.

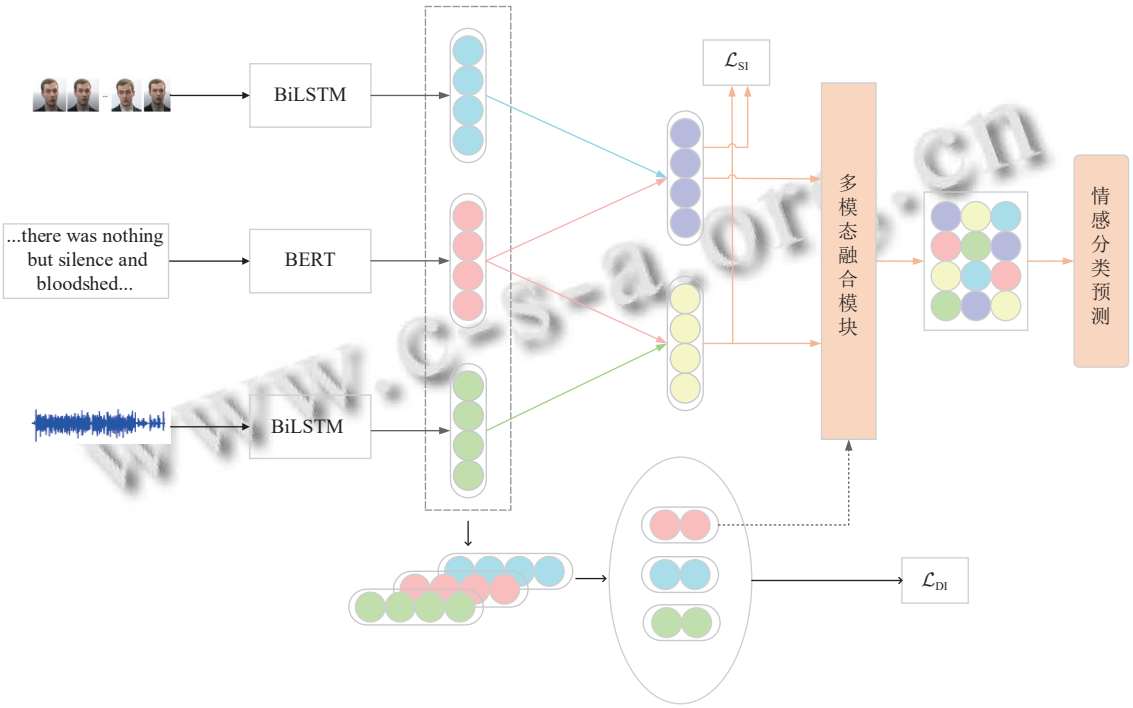

图 1 结合模态表征学习的多模态情感分析方法模型图

(1) 多模态特征提取. 针对文本 (T)、音频 (A) 和视频 (V) 分别采用 BERT 和 LSTM 进行提取, 旨在从原始数据中抽取有代表性的信息.

(2) 模态表征学习. 为了得到更好的模态表示, 模型学习每个模态的表征, 即学习模态之间的动态独立性. 这一步旨在建立每个模态的高层次表征, 以便更好地捕捉模态内部的语义信息.

(3) 多模态融合. 模型通过跨模态交互的方式, 探索和捕捉不同模态之间的关联和交互信息. 这有助于模型更好地理解不同模态之间的语义关系, 提升情感分析性能. 模型通过门控机制及多头注意力机制将不同模态融合起来. 以生成一个更全面的、集成的多模态表示.

(4) 情感分类预测. 得到的多模态融合结果送到前馈神经网络中得到最后的分类结果.

2.1 任务描述

在多模态情感分析 (MSA) 任务中, 模型的输入为从同一视频片段中提取的 3 个单模态原始序列: 文本模态 $X_t \in R^{T_t \times D_t}$, 视觉模态 $X_v \in R^{T_v \times D_v}$, 和声学模态 $X_a \in R^{T_a \times D_a}$, 其中 T_t, D_t 分别为文本序列的长度和表示向量维数, T_v, D_v 为视觉序列的长度和表示向量维数, T_a, D_a 为声学序列的长度和表示向量维数. 设计模型的主要目标是从这些不同模态的输入向量中提取和整合与任务相关的信息, 形成一个统一的表示, 然后利用这一表示对反映情感强度的真实值 y 进行准确预测.

2.2 特征提取

首先, 我们对多模态序列的输入 X_m 进行编码, 得到单位长度的表示 h_m, 其中 $m \in \{t, v, a\}$. 具体来说, 文本部分, 我们使用 BERT 对输入句子进行编码, 并从最后一层的输出中提取头部嵌入作为隐藏层的输出 h_t. 对于视觉和声学, 我们采用了双向 LSTM 来捕捉这些模态的时序特征:

$$h_t = \text{BERT}(X_t; \theta_t^{\text{BERT}}) \tag{1}$$

$$h_v = \text{BiLSTM}(X_v;\theta_v^{\text{BiLSTM}}) \tag{2}$$

$$h_a = \text{BiLSTM}(X_a;\theta_a^{\text{BiLSTM}}) \tag{3}$$

其中, θ_t^{BERT} 表示 BERT 模型的参数, θ_v^{BiLSTM}, θ_a^{BiLSTM} 表示 BiLSTM 模型的参数.

2.3 模态表征学习

模态表征学习是指在多模态数据处理中, 通过学习每个模态 (如文本、图像、声音等) 的有效表示方式, 使得这些表示能够捕捉跨模态动态独立性和相关性.

对于模态提取到的特征, 我们选择使用中心矩差 (CMD) 指标来进一步学习表征信息. CMD 是一种流行的度量, 它可以执行高阶矩的明确匹配, 而无需进行昂贵的距离和核矩阵计算. 同时也是一种先进的距离指标, 通过匹配两个表示的顺序矩差来衡量它们之间的差异. 简而言之, CMD 距离会随着两个分布的相似性而减小.

定义 CMD, 让 X 和 Y 成为有界随机样本分别用 p 和 q 的概率分布表示, 在区间 $[a,b]^N$, 中心矩差异正则化项 CMD_K 被定义为 CMD 指标的经验估计, 由:

$$CMD_K(X,Y) = \frac{1}{|b-a|}\|E(X)-E(Y)\|_2 + \sum_{k=2}^{K}\frac{1}{|b-a|^k}\|C_k(X)-C_k(Y)\|_2 \tag{4}$$

其中, $E(X)$ 为样本经验期望向量, $C_k(X)$ 是所有样本的 k 阶中心矩.

在本模型中, 我们计算每对模态之间的 CMD 损失.

$$\mathcal{L}_{\text{DI}} = \frac{1}{3}\sum_{\substack{(m_1,m_2)\\ \in\{(t,a),\\ (t,v),\\ (v,a)\}}} CMD_K\left(h_{m_1}^r, h_{m2}^r\right) \tag{5}$$

其中, $h_{m_1}^r$, $h_{m_2}^r$ 分别表示两种模态的特征向量.

2.4 多模态融合

多模态融合模块由图 2 所示. 对于来自单个视频剪辑的模态表示对 X, Y, 尽管它们似乎是独立的序列, 但它们之间存在一定的相关性. 我们希望通过模态间交互可以过滤掉与任务无关的模态特定随机噪声, 并尽可能保留跨所有模态的不变内容. 如前所述, 我们选择一个易处理的下界:

$$I(X;Y) \geqslant E_{p(x,y)}[\log q(y|x)] + H(Y) \triangleq I_{\text{SI}} \tag{6}$$

其中, $H(Y)$ 表示 Y 的微分熵. 对于熵 $H(Y)$, 我们使用高斯混合模型 (GMM) 来解决它的计算问题. 我们为每个类别建立两个正态分布, 参数在足够大的采样批次 $D_i \in D_{\text{train}}$ 通过最大似然法进行估计:

$$\hat{\mu}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} h_c^i \tag{7}$$

$$\hat{\sum}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} h_c^i \odot h_c^i - \hat{\mu}_c^{\text{T}}\hat{\mu}_c \tag{8}$$

其中, $\hat{\mu}_c$ 代表类别 c 的均值向量, 是该类别在 GMM 中的中心位置. $\hat{\sum}_c$ 代表类别 c 的协方差矩阵, 是该类别内数据分布的形状和方向. 其中 $c \in \{pos, neg\}$ 表示样本属于的极性类别, N_c 是类别 c 中的样本数, $\odot$ 表示向量逐元素相乘, $\hat{\mu}_c^{\text{T}}$ 为 $\hat{\mu}_c$ 的转置.

将下限作为近似, 我们得到了互信息下限的熵项:

$$H(Y) = \frac{1}{4}\left[\log\left(\det\left(\sum_1\right)\det\left(\sum_2\right)\right)\right] \tag{9}$$

其中, $\det\left(\sum_1\right)$、$\det\left(\sum_2\right)$ 表示两个协方差矩阵的行列式. 在式 (9) 中, 我们隐含地假设两个类别的先验概率相等, 互信息下限的损失函数由以下公式给出:

$$\mathcal{L}_{\text{SI}} = -I_{\text{SI}}^{t,v} - I_{\text{SI}}^{t,a} \tag{10}$$

多模态融合模块由堆叠的两个结合门控机制的 Transformer 组成, 形成对称结构. 跨模态融合过程主要发生在多头注意力操作中, 我们发现由于缺乏信息流的控制, 其性能表现不佳. 为了以一种精细和可控的方式改善它, 我们引入了两个门控: 保留门 g_r, 它决定目标模态的成分中有多少比例保持前进; 融合门 g_f, 它决定有多少比例的目标模态进行融合. 生成门控公式如下:

$$g_r^i = LN(W_r^i(h_1^i \oplus h_2^i)) \tag{11}$$

$$g_f^i = LN(W_f^i(h_1^i \oplus h_2^i)) \tag{12}$$

其中, W_r^i、W_f^i 用于映射输入的权重矩阵, 表示门中不同模态的权重. 然后将这些门应用于多头注意力, 以限制剩余块的信息流, 作为融合的一部分:

$$m^i = MH\text{-}ATT(Q^i, K^i, V^i) \tag{13}$$

$$\tilde{Z}_m^i = LN(g_f^i \odot m^i + g_r^i \odot Z_m^i) \tag{14}$$

其中, $MH\text{-}ATT$ 表示多头注意力, $\odot$ 表示逐分量乘法, LN 表示层归一化. 接下来, 将结果通过前馈网络产生当前层的最终输出:

$$Z_m^i = LN(\tilde{Z}_m^i + FFN(\tilde{Z}_m^i)) \tag{15}$$

最终, 我们将模态交互以及模态表征学习到的 h_t^r 拼接起来, 得到最后的融合结果:

$$Z_z = Z_{TA} \oplus Z_{TV} \oplus h_t^r \tag{16}$$

其中, $\oplus$表示拼接操作.

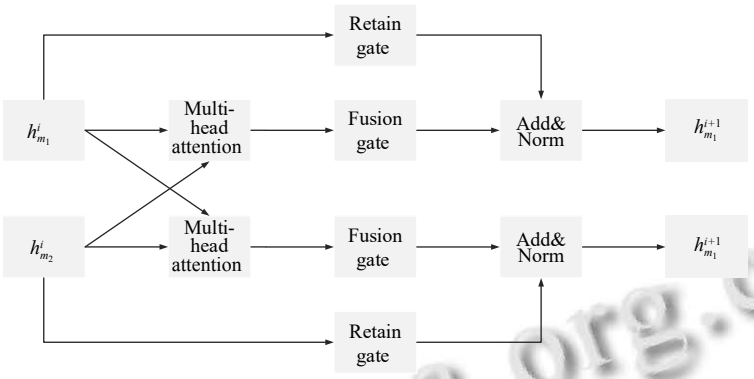

图 2 多模态融合结构示例

2.5 情感预测

训练过程在每次迭代中由两个阶段组成: 在第 1 阶段, 我们将前面式 (5) 中的模态表征学习添加到主损失中进行优化. 在第 2 阶段, 我们通过最小化式 (10) 中的跨模态预测的负对数似然, 用 $q(y|x)$ 近似 $p(y|x)$. 在获得最终预测和真值后, 我们得到了任务损失:

$$\mathcal{L}_{task} = MAE(\hat{y}, y) \tag{17}$$

其中, MAE 代表平均绝对误差损失, 这是回归任务中的常见做法. 最后, 我们计算所有这些损失的加权和以获得该阶段的主要损失:

$$\mathcal{L}_{main} = \mathcal{L}_{task} + \alpha\mathcal{L}_{SI} + \beta\mathcal{L}_{DI} \tag{18}$$

其中, α, β为超参数, 分别表示不同损失的权重.

最后使用全连接层和 $Softmax$ 函数对融合总特征进行情感分类, 分类结果为:

$$y_i = Softmax(w_s(ReLU(w_f X_z + b_f) + b_s)) \tag{19}$$

其中, w_f、b_f 为全连接层的权重和偏置, w_s、b_s 为 $Softmax$ 层的权重和偏置, y_i 为最终的情感分类结果.

3 实验

3.1 实验数据集

本文采用 CMU-MOSI 和 CMU-MOSEI 两个数据集[15,16], 数据集的具体划分如表 1 所示.

CMU-MOSI 的全称是 CMU multimodal opinion-level sentiment intensity, 该数据集是多模态情感分析任务中最为常用的数据集. 数据集中的素材来源于 YouTube 视频网站, 其类型为单人场景下对某个主题的评价和看法. 该数据集包含 93 个视频, 被划分为 2 199 个语音视频片段. 每个片段都手动标注了一个范围从[−3~3]的情绪评分, 来表示负面 (得分低于 0) 或正面 (得分高于 0) 情绪的相对强度.

表 1 实验数据集的划分

划分	训练集	验证集	测试集	总计
CMU-MOSI	1 284	229	686	2 199
CMU-MOSEI	16 326	1 871	4 659	22 856

CMU-MOSEI 全称 CMU multimodal opinion sentiment and emotion intensity, 是规模最大的三模态数据集, 且具有情感和情绪两类标注. CMU-MOSEI 数据集包含超过 65 h 的带标注视频, 视频来自 1 000 多个演讲者, 涉及 250 个主题. 与 CMU-MOSI 相同的是该数据集也具有多标签特性, 即每一个样本对应的情绪可能不止一种, 对应情绪的强弱也不同, 在 [−3~3]之间.

3.2 参数设置及评价指标

本实验使用的 Python 版本为 3.8, 使用的框架为 PyTorch 1.11.0, 显卡为 RTX 4090, 显存 24 GB. 本实验使用的 batch size 大小为 64, 隐藏层大小 128, 学习率为 5E–4, 损失超参α, β分别为 0.1 和 1, 优化器使用 Adam.

本文使用多模态情感分析常用的评价指标. MAE (平均绝对误差), 衡量模型预测值和真实值之间差异的平均绝对值. Corr (相关性) 衡量模型预测值和真实值之间的线性关系. 在情感分析任务中, 通常将情感标签划分为 7 个类别 (非常负面、负面、略微负面、中性、

略微正面、正面、非常正面), Acc-7 衡量模型对所有类别的准确率. Acc-2 将情感标签划分为两个类别, 正面和负面, Acc-2 衡量模型对二分类任务的准确率. *F*1 综合考虑了精确度和召回率, 适用于不平衡的数据集.

3.3 对比实验

为了验证本模型的有效性, 选择以下几种多模态情感分析模型进行比较.

(1) TFN: 能够端到端地学习模态内和模态间的动态, 采用一种新的多模态融合方法 (张量融合) 对模态间动态进行建模, 模态内动态则通过 3 个模态嵌入子网络进行建模.

(2) MFM[17]: 采用了一种路由的方法来动态调节输入模态和输出表示之间的权重以及输出表示.

(3) ICCN[18]: 模型使用特征对的外积以及深度典型相关分析来生成有用的多模态嵌入特征.

(4) CubeMLP[19]: 从特征混合的角度, 接受所有相关模态特征作为输入, 并将它们跨 3 个轴混合, 降低计算成本.

(5) MISA[20]: 学习模态不变量和特定模态的表示, 对多模态数据给出一个全面的、分解的观点, 从而帮助融合预测情感状态.

(6) MAG-BERT[21]: 提出了一个连接到 BERT 的多模式自适应门, MAG 利用以非语言行为为条件的注意力, 允许模型无缝地适应多模态输入.

(7) MMIM[14]: 提出了一种用于多模态情感分析的分层 MI 最大化框架. MI 最大化发生在输入层和融合层, 以减少有价值的任务相关信息的丢失.

3.4 实验结果与分析

在 CMU-MOSI 和 CMU-MOSEI 数据集上的实验结果如表 2 所示. 我们对比了本文提出的模型与其他几个先进的多模态情感分析模型, 包括 TFN、MFM、ICCN、CubeMLP、MISA、MAG-BERT 以及 MMIM*. 其结果从已发论文中得到, 表 2 中*表示相同情况下原论文的复现结果.

表 2 模型在 CMU-MOSI 和 CMU-MOSEI 上的实验结果

模型	CMU-MOSI					CMU-MOSEI				
	Corr	*MAE*	Acc-7 (%)	Acc-2 (%)	*F*1 (%)	Corr	*MAE*	Acc-7 (%)	Acc-2 (%)	*F*1 (%)
TFN	0.698	0.901	34.9	80.8	80.7	0.700	0.593	50.2	82.5	82.1
MFM	0.706	0.877	35.4	81.7	81.6	0.717	0.568	51.3	84.4	84.3
ICCN	0.714	0.862	39.0	83.0	83.0	0.713	0.565	51.6	84.2	84.2
CubeMLP	0.755	0.768	44.5	83.8	83.7	0.759	0.547	**53.3**	84.8	84.8
MISA	0.764	0.804	42.3	82.1	82.03	0.724	0.568	52.2	84.23	83.97
MAG-BERT	0.781	0.727	43.6	84.43	**84.61**	0.755	0.543	52.6	84.82	84.71
MMIM*	0.794	0.719	45.6	84.30	84.31	0.761	0.552	52.1	84.76	84.72
本文模型	**0.796**	**0.698**	**47.1**	**84.76**	**84.61**	**0.767**	**0.535**	**52.9**	**86.13**	**86.03**

从各项指标来看, 本文模型在 CMU-MOSI 上表现出色. 本文模型在 Acc-7 方面达到了 47.1%, 相较最接近的竞争者 MAG-BERT 的 43.6% 有显著提升. 这表明在 7 个情感类别上, 本文模型能够更准确地进行分类. *F*1 分数为 84.76%, 相较 MMIM*的 84.31%, 提高了 0.45 个百分点. 在情感分析任务中, *F*1 分数是平衡了准确率和召回率的重要指标, 显示了本文模型在综合性能上的优势.

在 CMU-MOSEI 上, 本文模型在 Acc-2 方面达到了 86.13%, 而最接近的竞争者是 MMIM*的 84.76%. 这意味着在对情感进行二分类时, 本文模型能够更准确地进行判别, 取得了明显的优势. *F*1 分数为 86.03%, 相比 MMIM*的 84.72%, 提高了 1.31 个百分点. 这再次强调了本文模型在综合性能上的卓越表现.

本文模型综合了模态表征学习和多模态融合模块, 充分发挥了它们的协同效应, 提高了模型在多模态情感分析任务中的性能.

此外, 本模型在 CMU-MOSEI 数据集上的提升比 CMU-MOSI 数据集的提升要高, 原因可能是 CMU-MOSEI 数据集拥有更多的数据样本, 样本数量的增加有助于提高模型的鲁棒性, 使其更好地适应各种输入情况. 拥有足够多的样本可以确保模型不容易受到噪声或特定情况的影响. 过少的样本可能导致模型过拟合, 即在训练数据上表现良好但在新数据上泛化能力差. 相反, 大量样本有助于防止过拟合, 使得模型更好地适应未见过的数据.

3.5 消融实验

为了验证模态表征学习模块和多模态融合模块的有效性, 在 CMU-MOSEI 数据集上设计了多组消融实验. 实验结果如表 3 所示.

表 3 消融实验结果 (%)

模型	Acc-2	*F*1
MMIM*	84.76	84.72
w/o rl	85.09	85.00
w/o mf	85.73	85.72
本文模型	86.31	86.03

表 3 中, w/o rl 表示移除模态表征学习, w/o mf 表示移除模态融合模块. MMIM*代表基准模型, 未进行任何模块的增加. 我们关注了准确率 (Acc-2) 和 *F*1 分数这两个关键性能指标. 在移除了多模态融合模块, 准确率和 *F*1 分数分别下降了 0.58% 和 0.31%. 这表示多模态融合模块能以更加可控的方式控制融合过程, 提高模型的分数. 在移除了模态表征学习模块后, 准确率和 *F*1 分数分别下降了 1.22% 和 1.03%. 这代表模态表征学习能更加充分地学习到模态内部的独特信息, 从而提升模型的准确性. 最终的本文模型在考虑了模态表征学习和多模态融合的情况下表现最佳, 取得了最高的准确率和 *F*1 分数. 这强化了模态表征学习和多模态融合模块在提高性能方面的重要作用, 证明了它们在多模态情感分析中的有效性.

4 结语

本文提出了一种结合模态表征学习的多模态情感分析模型, 克服了传统单模态方法的限制. 通过整合来自文本、视觉和声学等多个模态的信息, 我们能够更全面、准确地捕捉情感的丰富特征. 在多模态训练的过程中, 我们把模态表征学习和结合多头注意力的门控机制巧妙地结合起来, 实现了对模态独立性和模态间关联信息的有力挖掘. 实验证明, 我们的方法在多个数据集上取得了令人满意的性能. 模型不仅提高了情感分类任务的准确性, 还对不同模态的表达进行了有效整合, 提高了模型的鲁棒性. 然而, 尽管取得了一些进展, 多模态情感分析仍然面临一些挑战. 模态之间的异构性、不同模态数据的不平衡性等问题仍然需要进一步研究和解决. 未来的工作可以着重于改进模型的泛化能力, 适应更多领域和场景, 以推动多模态情感分析在实际应用中的广泛应用.

参考文献

1 Soleymani M, Garcia D, Jou B, *et al*. A survey of multimodal sentiment analysis. Image and Vision Computing, 2017, 65: 3–14. [doi: 10.1016/j.imavis.2017.08.003]

2 Majumder N, Hazarika D, Gelbukh A, *et al*. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-based Systems, 2018, 161: 124–133. [doi: 10.1016/j.knosys.2018.07.041]

3 Morency LP, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: Harvesting opinions from the Web. Proceedings of the 13th International Conference on Multimodal Interfaces. Alicante: ACM, 2011. 169–176.

4 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. Information Fusion, 2023, 91: 424–444. [doi: 10.1016/j.inffus.2022.09.025]

5 Das R, Singh TD. Multimodal sentiment analysis: A survey of methods, trends, and challenges. ACM Computing Surveys, 2023, 55(13s): 270.

6 Peng W, Hong XP, Zhao GY. Adaptive modality distillation for separable multimodal sentiment analysis. IEEE Intelligent Systems, 2021, 36(3): 82–89. [doi: 10.1109/MIS.2021.3057757]

7 Wang ZL, Wan ZH, Wan XJ. Transmodality: An end2end fusion method with Transformer for multimodal sentiment analysis. Proceedings of the 2020 Web Conference. Taipei: ACM, 2020. 2514–2520.

8 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.

9 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.

10 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.

11 Ghosal D, Akhtar MS, Chauhan D, *et al*. Contextual inter-modal attention for multi-modal sentiment analysis. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 3454–3466.

12 Pham H, Manzini T, Liang PP, *et al*. Seq2Seq2Sentiment: Multimodal sequence to sequence models for sentiment analysis. Proceedings of the 2018 Grand Challenge and

Workshop on Human Multimodal Language. Melbourne: ACL, 2018. 53–63.

13 Zhan YB, Yu J, Yu Z, *et al*. Comprehensive distance-preserving autoencoders for cross-modal retrieval. Proceedings of the 26th ACM International Conference on Multimedia. Seoul: ACM, 2018. 1137–1145.

14 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.

15 Zadeh A, Zellers R, Pincus E, *et al*. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259, 2016.

16 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 (Volume 1: Long Papers). Melbourne: ACL, 2018. 2236–2246.

17 Li CY, Gan Z, Yang ZY, *et al*. Multimodal foundation models: From specialists to general-purpose assistants. arXiv:2309.10020, 2023.

18 Sun ZK, Sarma PK, Sethares WA, *et al*. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. Proceedings of the 2020 AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 8992–8999.

19 Sun H, Wang HY, Liu JQ, *et al*. CubeMLP: An MLP-based model for multimodal sentiment analysis and Depression estimation. Proceedings of the 30th ACM International Conference on Multimedia. Lisboa: ACM, 2022. 3722–3729.

20 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.

21 Rahman W, Hasan MK, 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.

(校对责编: 孙君艳)