软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(2):537−553 [doi: 10.13328/j.cnki.jos.007144] [CSTR: 32375.14.jos.007144]

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

多模态可信度感知的情感计算*

罗佳敏, 王晶晶, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

通信作者: 王晶晶, E-mail: djingwang@suda.edu.cn

摘　要: 多模态情感计算是情感计算领域一个基础且重要的研究任务, 旨在利用多模态信号对用户生成的视频进行情感理解. 尽管已有的多模态情感计算方法在基准数据集上取得了不错的性能, 但这些方法无论是设计复杂的融合策略还是学习模态表示, 普遍忽视了多模态情感计算任务中存在的模态可信度偏差问题. 认为相较于文本, 语音和视觉模态往往能更真实的表达情感, 因而在情感计算任务中, 语音和视觉是高可信度的, 文本是低可信度的. 然而, 已有的针对不同模态特征抽取工具的学习能力不同, 导致文本模态表示能力往往强于语音和视觉模态 (例如: GPT3 与 ResNet), 这进一步加重了模态可信度偏差问题, 不利于高精度的情感判断. 为缓解模态可信度偏差, 提出一种模型无关的基于累积学习的多模态可信度感知的情感计算方法, 通过为低可信度的文本模态设计单独的文本模态分支捕捉偏差, 让模型在学习过程中从关注于低可信度文本模态的情感逐步关注到高可信度语音和视觉模态的情感, 从而有效缓解低可信度文本模态导致的情感预测不准确. 在多个基准数据集上进行实验, 多组对比实验的结果表明, 所提出的方法能够有效地突出高可信度语音和视觉模态的重要性, 缓解低可信度文本模态的偏差; 并且, 该模型无关的方法显著提升了多模态情感计算方法的性能, 这表明所提方法在多模态情感计算任务中的有效性和通用性.

关键词: 多模态可信度感知; 多模态情感计算; 可信度偏差; 累积学习

中图法分类号: TP18

中文引用格式: 罗佳敏, 王晶晶, 周国栋. 多模态可信度感知的情感计算. 软件学报, 2025, 36(2): 537–553. http://www.jos.org.cn/1000-9825/7144.htm

英文引用格式: Luo JM, Wang JJ, Zhou GD. Multi-modal Reliability-aware Affective Computing. Ruan Jian Xue Bao/Journal of Software, 2025, 36(2): 537–553 (in Chinese). http://www.jos.org.cn/1000-9825/7144.htm

Multi-modal Reliability-aware Affective Computing

LUO Jia-Min, WANG Jing-Jing, ZHOU Guo-Dong

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: Multi-modal affective computing is a fundamental and important research task in the field of affective computing, using multi-modal signals to understand the sentiment of user-generated video. Although existing multi-modal affective computing approaches have achieved good performance on benchmark datasets, they generally ignore the problem of modal reliability bias in multi-modal affective computing tasks, whether in designing complex fusion strategies or learning modal representations. This study believes that compared to text, acoustic and visual modalities often express sentiment more realistically. Therefore, voice and vision have high reliability, while text has low reliability in affective computing tasks. However, existing learning abilities of different modality feature extraction tools are different, resulting in a stronger ability to represent textual modality than acoustic and visual modalities (e.g., GPT3 and ResNet). This further exacerbates the problem of modal reliability bias, which is unfavorable for high-precision sentiment judgment. To mitigate the bias caused by modal reliability, this study proposes a model-agnostic multi-modal reliability-aware affective computing approach (MRA) based

* 基金项目: 国家自然科学基金 (62006166, 62076175, 62076176); 江苏高校优势学科建设工程
收稿时间: 2023-04-03; 修改时间: 2023-07-06, 2023-10-08; 采用时间: 2023-12-30; jos 在线出版时间: 2024-05-08
CNKI 网络首发时间: 2024-05-11

on cumulative learning. MRA captures the modal reliability bias by designing a single textual-modality branch and gradually shifting the focus from sentiments expressed in low-reliability textual modality to high-reliability acoustic and visual modalities during the model learning process. Thus, MRA effectively alleviates inaccurate sentiment predictions caused by low-reliability textual modality. Multiple comparative experiments conducted on multiple benchmark datasets demonstrate that the proposed approach MRA can effectively highlight the importance of high-reliability acoustic and visual modalities and mitigate the bias of low-reliability textual modality. Additionally, the model-agnostic approach significantly improves the performance of multi-modal affective computing, indicating its effectiveness and generality in multi-modal affective computing tasks.

Key words: multi-modal reliability-aware; multi-modal affective computing; reliability bias; cumulative learning

人类的情感复杂、抽象且丰富, 通常可以通过多种模态的信息来体现, 例如文字、图像、音频等. 图 1(a) 给出了来源于某电影片段的视频例子, 其中“你不是很上镜”是文本 (体现了负面情感的语言), 视频帧是图像 (体现了正面情感的微笑表情、肢体动作等), 音频频谱是语音 (体现了正面情感的愉悦语气、音调音色等). 基于此, 在自然语言处理领域, 多模态情感计算 (multi-modal affective computing, MAC) 通常被定义为结合两种或两种以上模态信息对情感进行预测的任务. 相比于单模态情感计算任务, 多模态情感计算可以同时考虑多种模态信息, 因而往往具有更高的预测精度和鲁棒性. 近几年随着通用深度学习技术的发展, 以及多模态情感计算在社会计算、多媒体推荐以及心理健康咨询等领域的广泛应用前景, 融合多个模态信息来进行情感预测渐渐成为研究热点, 吸引了越来越多的研究者们的关注. 现有的研究工作 (例如: 杨杨等人[1]) 主要围绕多模态情感计算任务的 5 大核心问题 (对齐, 翻译, 表示, 融合以及联合学习) 展开. 尽管已有的工作取得了一定的研究进展, 但是这些工作无论是设计多模态融合策略[2–8]还是学习多模态表示[9–14], 都忽略了多模态情感计算任务中存在的模态可信度 (可信度: 对人或者事物可以信赖的程度, 是根据经验对一个事物或者一件事情为真的相信程度) 偏差问题.

本文认为, 在现实世界中, 相较于文本, 语音和图像更能够真实反映人类的情感, 因为文本常具有欺骗性. 如图 1(a) 中的例子所示 (其中, T, V, A 分别表示文本, 视觉, 语音模态; NG, WN, WP, PS 分别表示消极, 弱消极, 弱积极, 积极情感标签), 语音 (欢乐的音调) 和视觉模态 (笑脸) 的情感是正面的, 而文本模态 (不上镜) 的情感是负面的, 然而正确的情感标签应该是正面的. 这表明文本模态所反映的情感并不一定可信. 相关实例在日常生活中还有很多, 例如抑郁症患者往往会在语言中掩盖自己的真实情感状态, 而在语音和视觉信息中很难掩饰自己的真实情感[15]. 因此, 本文认为不同模态信息在情感表达上存在可信度偏差问题, 并且认为语音和视觉模态是高可信度的, 文本模态是低可信度的. 然而, 已有的针对不同模态的特征抽取工具学习能力存在差异, 导致不同模态的表示能力之间存在强弱之分, 例如: 文本模态的表示能力往往强于语音和视觉模态 (参考 GPT3 与 ResNet). 再具体一点, 从图 1(b) 和 (c) 的折线图实验结果中可以看出, 单文本模态的情感预测性能强于单语音或者单视觉模态的性能. 这主要是因为文本模态语义表示能力本身就强于语音和视觉模态, 容易误导多模态情感计算方法的情感预测结果高度偏向文本模态的预测结果. 同时, 通过图 1(a) 中的例子也可以看出, 文本模态往往会误导多模态情感计算方法给出错误的情感预测标签, 这进一步加重了模态可信度偏差问题.

为了缓解上述模态可信度偏差问题, 本文受累积学习[16]思想启发设计了一种新的处理多模态情感计算任务的方法. 该方法采用了累积学习的渐进式学习策略, 让模型在训练过程中首先关注低可信度的文本模态, 再随着训练过程的推进, 逐步关注到高可信度的语音和视觉模态. 具体而言, 本文提出了一种模型无关的基于累积学习的多模态可信度感知的情感计算方法 MRA (multi-modal reliability-aware affective computing approach). 该 MRA 方法主要包含两个部分: 基础多模态情感计算模型和累积学习模块 (cumulative learning module), 其中累积学习模块由文本模态分支 (textual-modality branch) 和多模态主模型 (multi-modal main block) 组成. MRA 的工作流程主要分为 3 个步骤: 首先将文本模态输入到文本模态分支 (包含文本特征编码器, 神经网络以及任务输出函数) 中学习低可信度的文本模态的偏差, 得到文本模态输出; 其次, 基础多模态情感计算模型将输入的 3 种模态进行表示学习并融合得到多模态输出; 最后, 多模态主模型利用累积学习衡量文本模态输出与多模态输出之间的偏差, 然后随着训练过程的推进去除该偏差. 相比于传统的多模态情感计算方法, MRA 更关注于模态可信度偏差问题, 能够有效突出高可信度的语音和视觉模态在情感预测中的重要性, 减少低可信度的文本模态对模型的误导.

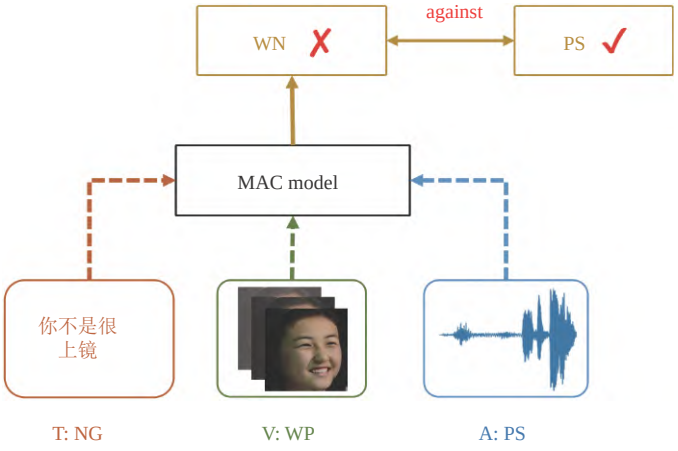
综上所述, 本文的主要贡献如下.

(1) 本文考虑多模态情感计算任务中的模态可信度偏差问题, 并且认为语音和视觉模态相较于文本在情感预测上是高可信度的.

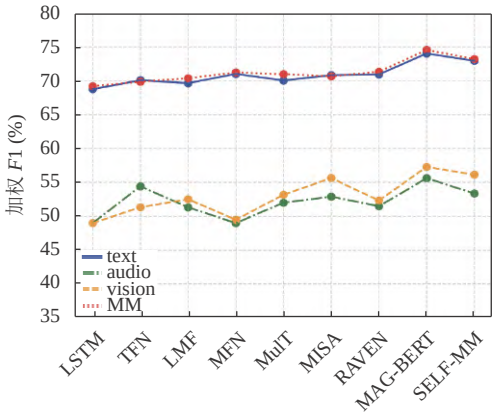
(2) 本文通过初步实验分析发现, 由于不同模态的特征抽取工具的学习能力不同, 导致文本语义表示往往强于语音和视觉模态, 因而单文本模态的性能往往强于单语音与单视觉, 这进一步加剧了模态可信度偏差问题.

(3) 本文提出了一种模型无关的多模态可信度感知的情感计算方法 MRA; 并且由于该方法是模型无关的, 因此可以轻松地扩展到已有的多模态情感计算方法上.

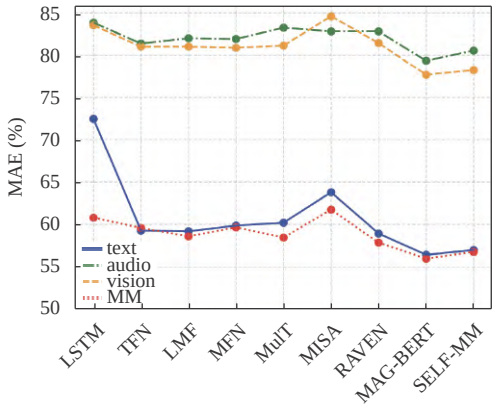
(4) 本文提出的方法在多个基准数据集上相较于传统的性能优异的多模态情感计算方法能获得显著的性能提升. 这充分说明该方法能够有效突出高可信度语音和视觉模态表示的重要性, 缓解低可信度文本模态表示的偏差; 并且验证了本文的方法在多模态情感计算任务中的有效性和通用性.



(a) 可信度偏差样本, 选自 CH-SIMS 数据集



(b) 不同多模态情感计算模型在 CMU-MOSEl 数据集上的加权 *F*1 结果折线图



(c) 不同多模态情感计算模型在 CMU-MOSEI 数据集上的 MAE 结果折线图

图 1 多模态情感计算任务中存在可信度偏差的样本及实验现象

1 相关工作

1.1 多模态情感计算

多模态情感计算是 NLP 领域常见的任务之一, 其处理了来自于多源信号 (如文本描述、语音音频、视频图片) 的信息来理解丰富多彩的人类情感. 由于不同模态的信息在数据形式和处理方式上有很大的差别, 在模型中多增加一种模态信息虽然可以带来潜在的建模效果提升, 但同时也增加了建模的复杂度和难度. 目前, 多模态情感计

算模型广义上可以分为两种形式: 一种是如何设计多模态融合策略, 另一种是如何学习更好的模态表示.

对于多模态融合的方法, EF-LSTM[2]首先将单模态输入进行简单的拼接融合, 并使用了一个基于 RNN 的模型来捕捉多模态输入中的时序依赖关系. TFN[3]为了建模模态内部和模态之间的动态, 提出了张量融合网络来获得张量表示, 端到端的学习这两种动态. MFN[4]是一种用于多视图序列的新颖神经网络, 称为记忆融合网络, 明确的解释了不同视图之间存在的特定视图交互和交叉视图交互形式. 为了缓解转换张量过程中遭受的维度指数增长和多模态融合过程中计算复杂度高的问题, Liu 等人[5]提出了低秩多模态融合方法, 利用低秩张量进行多模态融合以提高效率. 在建模多模态人类语言时间序列数据中存在两大挑战: 模态序列不对齐和跨模态元素之间的长距离依赖, Tsai 等人[6]引入了 Transformer[17], 以端到端的方式解决两大挑战, 并且无需模态数据对齐. Multi-modal routing (多模态路由)[7]是一种动态调整不同模态权重的路由机制, 它针对每个输入样本动态调整输入模态和输入表示之间的权重, 进而识别单个模态和跨模态特征的相对重要性. Han 等人[8]提出了一种层次化的互信息 (MI) 最大化模型, 在单模态输入内以及多模态融合结果与单模态输入之间最大化互信息, 以便通过多模态融合保持与任务相关的信息. Fu 等人[18]认为模态之间存在信息密度差异, 即视觉和音频具有低水平的信号特征, 而文本具有高水平的语义特征, 提出使用非齐次融合网络来实现多模态交互.

对于模态表示学习的方法, RAVEN[9]考虑了文本单词出现的非语言语境, 通过分析单词段中发生的细粒度视觉和听觉模态来模拟表达的非语言表示, 并通过改变非语言行为的词语表示来捕捉非语言意图的动态本质. MCTN[10]提出了一种通过转换模态来学习鲁棒联合表示的方法, 当使用成对的多模态数据进行训练时, 只需要在测试时使用源模态的数据就可以进行最终的情感预测, 确保模型在其他模态中保持鲁棒性, 不受扰动或信息缺失的影响. ICCN[11]假设可以通过学习从文本和基于文本的音频以及类似的基于文本的视频中提取的特征之间的 (隐藏) 相关性来改进多模态情感计算, 基于深度典型相关分析学习输入所有模态之间的相关性. MISA[12]为了学习有效的模态表示, 将每个模态映射到两个不同的子空间: 模态不变和特定于模态的子空间, 前者学习跨模态表示以学习它们的共性, 减小模态之间的差距; 后者是每个模态私有的, 用以捕获它们的特定特征. MAG-BERT[13]发现近期基于预训练模型的上下文表示在 NLP 的多个领域展示了最先进的性能, 但对这些预训练模型进行微调在文本模态上是很简单的, 对于多模态语言来说却并不容易, 因为当前预训练模型不具备接受视觉和听觉两种额外模态的必要组件. 因此, 提出了一个连接 BERT 的多模态适应门, 允许 BERT 在微调期间接受多模态非语言数据. SELF-MM[14]设计了一个基于自监督学习策略的标签生成模块, 为每种模态生成额外的单模态标签, 然后对多模态任务和单模态任务进行联合训练, 分别学习一致性和差异性.

然而, 上述研究工作无论是融合策略还是模态表示学习, 都忽略了多模态情感计算任务中的关键问题, 即模态的可信度偏差问题, 导致多模态的结果会更偏向于低可信度的文本模态所预测的情感. 基于此, 本文首次将累积学习引入多模态情感计算任务中, 提出了一种新的多模态可信度感知的情感计算方法, 可以有效地缓解低可信度的文本模态带来的偏差, 同时突出高可信度的语音和视觉模态在多模态情感计算任务上的重要性.

1.2 累积学习

累积学习是一种渐进式学习策略, 是人类或者人工智能系统积累知识和能力的认知过程, 这些知识和能力是后续认知发展的基础[16]. 累积学习的优点在于可以整合通过经验获得的知识, 并在学习的过程中通过类比知识转移来促进进一步的学习, 例如学生可以在不同的情境中传递知识的能力. 简言之, 累积学习是知识和能力的逐步发展, 并且随着时间的推移进一步提升.

目前基于累积学习的研究工作相对较少, 大部分的工作集中在计算机视觉中的长尾问题上. 累积学习通过这种渐进式的学习策略, 使得模型在训练过程中先关注非尾部类的学习, 再随着模型的训练, 逐渐关注到尾部类的学习中去. CLEAR[19]是在 OSOC (one-shot one-class) 图像识别任务中采用累积学习这种类似人类学习的方式, 通过积累迄今为止学习的经验, 利用过去训练过程中获得的所有知识. Zhou 等人[20]利用累积学习的策略设计双边分支网络先学习通用的表示, 然后逐渐学习到少量类别的数据, 这样既能同时处理表示学习和分类器学习, 每个分支都有其单独的职责. CAE[21]提出基于累积学习的思想解决强化学习中的多目标实现算法泛化的问题, 其衡量了在指定范围内从给定状态到目标的可达性.

受到已有工作的启发[20,22], 本文将累积学习引入到多模态情感计算方法中, 提出了一种新的多模态可信度感知的情感计算方法. 该方法通过设计的累积学习模块先学习低可信度的文本模态, 在模型训练过程中衡量低可信度模态与高可信度模态之间的偏差, 并随着训练过程的推进去除该偏差. 这可以有效地缓解低可信度的文本模态导致的情感预测错误. 此外, 该方法是模型无关的, 可以轻松地扩展到已有的多模态情感计算方法上, 并显著提升情感预测的精度.

2 基础多模态情感计算方法

在介绍本文提出的多模态可信度感知的情感计算方法前, 本文先概述了基础多模态情感计算方法. 在本节中, 首先简要介绍多模态情感计算任务的定义 (第 2.1 节); 然后描述不同模态的特征抽取 (第 2.2 节); 最后概括了多模态情感计算方法的工作流程, 并阐述了多模态情感计算任务的损失函数 (第 2.3 节).

2.1 任务定义

在多模态情感计算任务中, 模型的输入是多模态的序列, 记为 $U_m \in R^{l_m \times d_m}$, 其中 l_m 代表序列长度, d_m 代表模态 m 的向量维度. 具体而言, 在本文中, $m \in \{t,v,a\}$, 其中 t,v,a 分别代表 3 种不同的模态, 即文本、视觉和语音. 已有的多模态情感计算模型从这些输入模态表示中提取和整合任务相关的信息, 以形成统一的多模态表示, 然后利用它们对反映情感类别或强度的标签值 y 进行准确的预测.

2.2 特征抽取

(1) 文本特征

传统的文本特征是通过 GloVe Embedding[23]对 utterance 中的每个 token 进行表示. 然而, 随着预训练语言模型在 NLP 领域的兴起, 本文使用预训练的 BERT[24]模型作为文本的特征抽取器. 相较于使用 GloVe 获得的 300 维 token 嵌入, 本文利用 BERT 预训练模型 (CMU-MOSI 和 CMU-MOSEI 使用 uncased, CH-SIMS 使用 Chinese) 来获得文本词向量表示, 得到的文本特征向量的维度 d_t 为 768.

(2) 语音特征

CMU-MOSI 和 CMU-MOSEI 数据集的语音特征包含了从 COVAREP[25]中提取的各种低阶统计音频函数, 这是一种语音分析的框架. 该语音特征包含了 12 梅尔频率倒谱系数 (MFCCs)、音色跟踪、浊音/清音分割特征、声门声源参数、峰值斜率参数和最大离散熵. CH-SIMS 使用了默认参数的 LibROSA[26]演讲工具包来提取 22 050 Hz 的语音特征, 提取出了 33 维的帧级语音特征, 包含了 1 维的对数基频 (logF0), 20 维的梅尔频率倒谱系数 (MFCCs) 和 12 维的康斯坦特色谱图 (CQT). 因此, 本文在 CMU-MOSI、CMU-MOSEI 和 CH-SIMS 这 3 个数据集上分别得到了 5、74 和 33 的语音特征维度 d_a.

(3) 视频特征

CMU-MOSI 和 CMU-MOSEI 数据集都使用 Facet 工具来提取脸部表情特征, 其包含了基于脸部动作编码系统 (FACS)[27]所获得的脸部动作单元和脸部姿势特征, 并对 utterance 视频序列中的每个采样帧重复该过程. CH-SIMS 数据集首先以 30 Hz 的频率从视频片段中提取帧, 然后使用 MTCNN[28]人脸检测算法来提取对齐的人脸. 同时, CH-SIMS 使用 MultiComp OpenFace 2.0[29]工具包来抽取人脸的 68 个基准点集合、17 个人脸动作单元、头部姿势、头部方向和眼神. 因此, 本文在 CMU-MOSI、CMU-MOSEI 和 CH-SIMS 这 3 个数据集上分别得到了 20、35 和 709 的视觉特征维度 d_v.

根据不同模态的特征抽取工具, 可以明显看出文本模态的表示能力远强于语音和视觉模态. 这进一步加重了模态可信度偏差问题, 导致目前的多模态情感计算模型难以达到高精度的情感预测. 因此, 本文提出了一种新颖的多模态可信度感知的情感计算方法, 借助累积学习降低低可信度的文本模态带来的偏差.

2.3 基础多模态情感计算方法

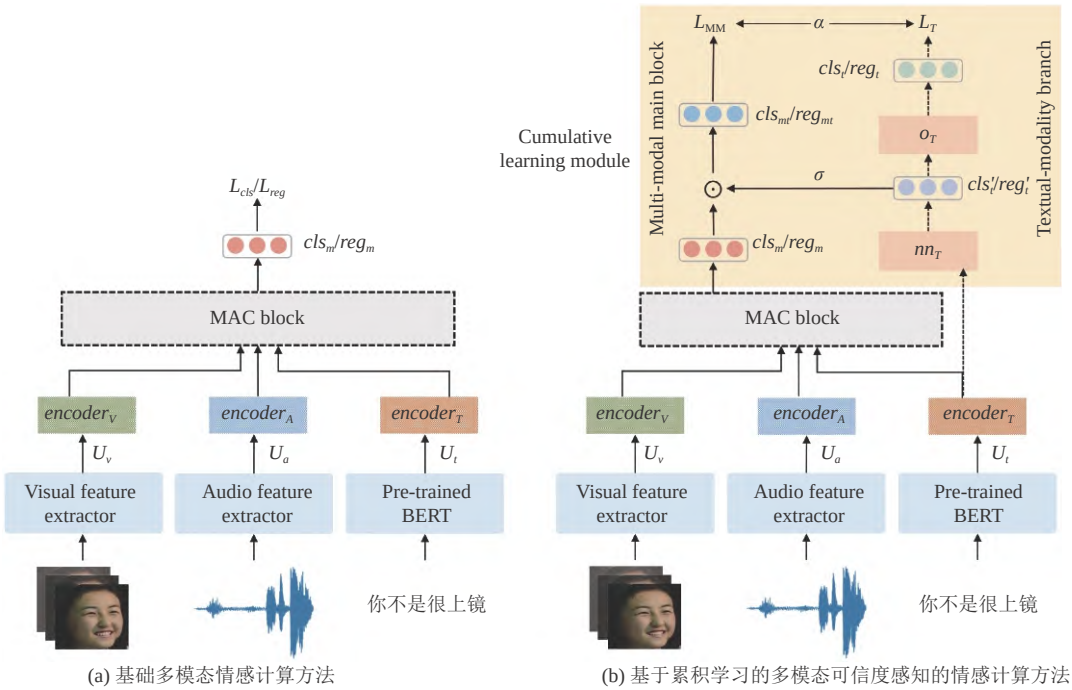
基础多模态情感计算的方法如图 2(a) 所示, 其主要是最小化规模大小为 N 的数据集上的标准交叉熵 (情感分

类任务) 或均方误差损失 (情感回归任务). 根据第 2.1 节多模态情感计算的任务定义, 模型的输入是多模态序列, 记为 $Input=(U_t,U_a,U_v)$, 其中 $U_t=\{U_{t_1},U_{t_2},\ldots,U_{t_N}\}$, $U_a=\{U_{a_1},U_{a_2},\ldots,U_{a_N}\}$, $U_v=\{U_{v_1},U_{v_2},\ldots,U_{v_N}\}$. 接着, 模型分别通过文本模态表示编码器 $encoder_T$、语音模态表示编码器 $encoder_A$、视觉模态表示编码器 $encoder_V$ 对 3 种模态输入进行编码, 得到 3 种模态表示 h_t,h_a,h_v, 如公式 (1)–公式 (3) 所示:

$$h_t=encoder_T(U_t) \tag{1}$$

$$h_a=encoder_A(U_a) \tag{2}$$

$$h_v=encoder_V(U_v) \tag{3}$$



(a) 基础多模态情感计算方法　　(b) 基于累积学习的多模态可信度感知的情感计算方法

图 2　基础多模态情感计算方法和基于累积学习的多模态可信度感知的情感计算方法

MAC 模块主要包含模态交互模块 MIM (modal-interaction module) 和模态融合模块 MFM (modal-fusion module). 这是本文根据已有的多模态情感计算模型总结的通用结构, 但是每个模型有其新颖之处, 这里不具体概述细节, 实验部分的第 4.2 节将会对使用到的每个多模态情感计算方法做较为详细的描述.

MIM 主要用于学习模态之间的交互信息 (部分模型不含该模块, 如早期融合方法), 得到交互后 3 种模态表示 h_t',h_a',h_v', 如公式 (4) 所示:

$$h_t',h_a',h_v'=MIM(h_t,h_a,h_v) \tag{4}$$

经过 MIM 后, 模型将交互后的 3 种模态表示 h_t',h_a',h_v' 通过 MFM 进行多模态融合, 得到融合后的多模态表示 f_m; 再将其通过任务输出函数 (如分类器 $classifier$ 或回归器 $regression$), 得到最终的多模态输出 $Output=cls_m/reg_m$; 最后使用交叉熵损失 (cross-entropy loss) 或均方误差损失 (mean square error loss) 反向传播进行优化, 如公式 (5)–公式 (9) 所示:

$$f_m=MFM(h_t',h_a',h_v') \tag{5}$$

$$cls_m=classifier(f_m) \tag{6}$$

$$reg_m=regression(f_m) \tag{7}$$

$$L_{cls} = \frac{1}{N}\sum_i -[y_i \log(cls_{m_i}) + (1-y_i)\log(1-cls_{m_i})] \tag{8}$$

$$L_{reg} = \frac{1}{N}\sum_i (reg_{m_i} - y_i)^2 \tag{9}$$

3 多模态可信度感知的情感计算方法

本节将会详细描述基于累积学习的多模态可信度感知的情感计算方法 (MRA). 首先将 MRA 与第 2.3 节的基础多模态情感计算方法进行对比, 并详细描述 MRA 方法的流程 (第 3.1 节); 此外, 本文提出了 MRA 方法在训练学习中的损失函数, 用于训练和优化模型 (第 3.2 节).

3.1 基于累积学习的多模态可信度感知方法 (MRA)

从图 1(b) 和图 1(c) 中的折线图实验结果可以明显看出, 基础多模态情感计算方法倾向于从数据集中学习到文本模态的偏差. 换言之, 它们依靠文本模态的统计规律来提供准确的预测, 而不需要考虑语音和视觉模态. 然而, 相较于文本模态, 语音和视觉模态在情感表达上是高可信度的, 更能够反映真实的情感状态 (如图 1(a)). 虽然已有的多模态计算模型取得了一定的进展, 但是它们普遍都忽略了模态可信度偏差问题. 基于此, 本文提出了一种模型无关的基于累积学习的多模态可信度感知方法 (MRA), 用以缓解低可信度的文本模态所带来的情感预测偏差. MRA 主要由基础多模态情感计算模型和累积学习模块 (cumulative learning module) 组成. 累积学习模块包含为低可信度的文本模态单独设计的文本模态分支 (textual-modality branch) 和多模态主模型 (multi-modal main block), 前者用于在训练过程中学习文本模态偏差, 后者用于衡量并去除文本模态输出与多模态输出之间的偏差, 得到最终的预测结果.

本文方法 MRA 的核心思想如图 2(b) 所示. 本文采用了累积学习的渐进式学习策略, 让模型在训练过程中首先关注低可信度的文本模态, 再随着训练的推进, 逐步关注到高可信度的语音和视觉模态. 为了测量低可信度的文本模态的偏差, 本文单独设计了一个文本模态分支 (即图 2(b) 中的 textual-modality branch). 该分支主要由文本特征编码器 $encoder_T$, 神经网络 nn_T 和任务输出函数 o_T (分类或回归) 组成, 并同时与基础多模态情感计算模型训练学习. 在训练过程中, 文本模态分支学习低可信度的文本模态偏差, 得到文本模态输出; 基础多模态情感计算模型得到多模态输出. 在之后的训练过程中, 多模态主模型 (即图 2(b) 中的 multi-modal main block) 利用累积学习衡量多模态输出和文本模态输出之间的偏差, 随着训练过程的推进消除该偏差. 通过这种累积学习的方式, 基础多模态情感计算模型能够更关注于高可信度的语音和视觉模态的正确情感信息, 减少低可信度的文本模态的误导, 从而提升预测的准确性. 文本模态分支得到文本模态输出 cls_t/reg_t, 如公式 (10) 所示:

$$cls_t/reg_t = o_T(nn_T(encoder_T(U_t))) \tag{10}$$

在训练过程中, 单独的文本模态分支能够防止任何基础多模态情感计算模型学习到偏差. 在验证和测试过程中, 删除文本模态分支只使用基础多模态情感计算模型的预测结果. 在将基础多模态情感计算模型的多模态输出 cls_m/reg_m 传递给定义的损失函数之前, 本文将它们与长度为 a (a 为 batch 规模的大小) 的 mask 合并. 这个 mask 是将神经网络 nn_T 的输出经过 *Sigmoid* 函数获得的, 记为 cls'_t/reg'_t.

$$cls'_t/reg'_t = Sigmoid(nn_T(encoder_T(U_t))) \tag{11}$$

本文引入 mask 的目的在于通过修改基础多模态情感计算模型的多模态输出来动态改变损失. 因此, 本文通过计算 mask 和原始多模态输出 cls_m/reg_m 之间的元素乘积得到新的多模态输出 cls_{mt}/reg_{mt}, 如公式 (12) 所示:

$$\begin{cases} cls_{mt} = cls_m \odot cls'_t, & \text{if classifier task} \\ reg_{mt} = reg_m \odot reg'_t, & \text{if regression task} \end{cases} \tag{12}$$

MRA 以这种特定的方式修改基础多模态情感计算模型的多模态输出, 以防止模型从文本模态分支中学习到偏差. 此外, 本文分析了两种场景帮助理解 MRA 方法对基础多模态情感计算模型的影响. 第一, MRA 方法提升了有偏差样本的重要性, 并充分利用无偏差样本 (即可以在不使用另外两种模态的情况下正确进行分类的样本). 具体而言, MRA 通过文本模态分支输出一个 mask, 用以提高正确情感标签的概率, 同时降低其他标签的概率, 使得

有偏差样本的损失要比无偏差样本的损失高. 通过这种方式, 基础多模态情感计算模型反向传播的梯度较大, 从而提升了有偏差样本在学习过程中的重要性. 第二, MRA 增加了使用多模态 (同时使用文本, 语音和视觉这 3 种模态) 才能得到正确情感标签的样本的重要性. 对于这些样本, MRA 通过文本模态分支输出的 mask 增加错误情感标签的分数, 因此损失会高很多; 通过这种方式, 基础多模态情感计算模型鼓励从这些样本中进行学习, 从而提升了同时使用 3 种模态就能得到正确情感标签的样本的重要性.

3.2 基于累积学习的优化目标

本文使用从两个损失计算的梯度, 联合优化基础多模态情感计算模型及其文本模态分支的参数. 本文中, 多模态情感计算任务主要是情感分类和情感回归, 对应的损失函数分别是交叉熵损失和均方误差损失.

主损失函数是从多模态主模型预测相关的交叉熵损失或者均方误差损失, 这里本文记为 $L_{\mathrm{MM}} = L_{cls}/L_{reg}$. θ_{MM} 为反向传播以优化导致该损失的所有参数, 即基础多模态情感计算模型 (包括 3 个模态表示编码器 $encoder_{T/A/V}$, 模态交互模块 MIM, 模态融合模块 MFM, 分类器 $classifier$ 或回归器 $regression$ 等模块) 的参数联合. 在本文的设置中, 基础多模态情感计算模型和文本模态分支之间共享文本模态表示编码器 $encoder_T$ 的参数.

文本模态分支的损失函数 L_T 是从文本模态输出 cls_t/reg_t 中学习到的与预测相关的交叉熵损失或者均方误差损失. 本文使用这个损失来优化文本模态分支的参数, 记为 θ_T , 其中 θ_T 是神经网络 nn_T 和分类器 c_T 的参数联合. 这进一步提高了文本模态分支捕获偏差的能力. 需要注意的是, 模型不会将这个损失反向传播到文本模态表示编码器 $encoder_T$, 这是为了防止基础多模态情感计算模型学习到文本模态的偏差.

本文通过将两个损失相加, 利用累积学习超参数 α 来衡量文本模态表示所损失的权重, 得到了最终损失 L_{MRA} , 如公式 (13) 所示:

$$L_{\mathrm{MRA}}(\theta_{\mathrm{MM}}, \theta_T) = L_{\mathrm{MM}}(\theta_{\mathrm{MM}}) + \alpha \cdot L_T(\theta_T) \tag{13}$$

4 实 验

本节描述了实验的细节, 包括实验所使用的数据集 (第 4.1 节), Baselines 方法 (第 4.2 节), 实验设置 (第 4.3 节), 实验结果对比 (第 4.4 节) 以及实验分析 (第 4.5 节). 本文提出的方法是模型无关的, 因此对比实验的目标是比较基础多模态情感计算方法在增加本文方法 MRA 前后的性能差异. 为了进行公平的比较, 对于每种不同的多模态情感计算方法, 本文在 3 个基准数据集上重新复现得到其情感计算任务的结果, 并且每种方法取 5 次结果的平均作为汇报的实验结果.

4.1 数据集

本文在 3 个多模态情感计算基准数据集上对本文提出的方法进行了评估, 3 个数据集的数据分布统计情况如表 1 所示.

表 1 数据集统计

数据集	训练集	验证集	测试集	总数量
CMU-MOSI	1 284	229	686	2 199
CMU-MOSEI	16 326	1 871	4 659	22 856
CH-SIMS	1 368	456	457	2 281

(1) CMU-MOSI[30]是一个由 2 199 段简短的独白视频片段组成 (每个视频片段都是一个句子的持续时间) 的人类多模态情感计算数据集, 其来源于 YouTube. 视频的内容是说话者表达他们对电影等主题的观点. CMU-MOSI 原始数据总共有 93 个完整的视频, 分别由 89 位不同的说话者演绎, 并将其切割为 2 199 个句子的视频段. 每个视频段由人工标注 [−3, +3] 之间的连续观点分数, 其中−3/+3 分别表示强烈的消极/积极情感.

(2) CMU-MOSEI[31]是对 CMU-MOSI 数据集的一个扩展, 它拥有更大数量的句子, 样本、说话者和主题更加具有多样性. CMU-MOSEI 拥有 22 856 条人工标注的视频段, 其来源于 5 000 个视频, 由 1 000 不同的说话者演绎

以及拥有 250 个不同的主题.

(3) CH-SIMS[32]是一个中文的既标注了单模态标签也标注了多模态标签的情感计算数据集, 其拥有 2 281 个视频片段. CH-SIMS 收集了不同的电影、电视连续剧和综艺节目, 其具有自发的表情、各种头部姿势等. 人工标注时为每个样本标记一个介于 [−1, +1] 之间的情感分数, 其中−1/+1 分别表示强烈的消极/积极情感.

4.2 Baselines

本文选择了传统的性能优异的多模态情感计算模型来验证 MRA 的有效性, 如下所述.

(1) EF-LSTM (early fusion LSTM)[2]首先将 3 种模态的输入特征进行简单的拼接融合, 得到多模态表示, 然后再使用 LSTM 捕获序列中的长距离依赖关系.

(2) TFN (tensor fusion network)[3]通过创建多维张量, 捕捉 3 种模态的单模态、双模态以及三模态之间的相互作用, 显式地建模特定视图和跨视图的动态. TFN 主要包含 3 个部分, 模态嵌入子网络将单模态特征作为输入, 输出丰富的模态表示; 张量融合层使用 3 倍笛卡尔积显式地对单模态、双模态和三模态之间的交互进行建模; 情感推理子网络将张量融合层的输出为条件进行情感推理.

(3) MFN (memory fusion network)[4]考虑了特定视图和跨视图两种交互形式. MFN 主要包含 3 层, 第 1 层是长短期记忆系统 (LSTM), 独立对每个视图进行编码并建模该特定视图中的动态; 第 2 层是 Delta 记忆注意力网络, 用来发现 LSTM 中的跨视图交互; 第 3 层是在多视图门控记忆模块中随时间存储跨视图信息. 需要注意的是, MFN 对于 3 种模态需要词层次的对齐, 但是 CH-SIMS 数据集是一个非对齐的数据集. 因此, 本文采用了一个简单的对齐策略使得 3 个模态在词层级对齐以使用 MFN.

(4) LMF (low-rank multi-modal fusion)[5]解决了多模态张量融合中遭受的维度指数增长和计算复杂度增加的问题, 提出使用低秩张量进行有效的多模态融合, 并学习具体模态和跨模态的交互. LMF 首先通过将 3 个单模态输入分别传入到 3 个子嵌入网络得到 3 个单模态表示, 再通过模态特异性因子进行低秩多模态融合得到多模态输出表示, 然后将其用于预测任务.

(5) RAVEN (recurrent attended variation embedding network)[9]对非语言子词序列的细粒度结构进行建模, 根据非语言行为动态转换词表示. RAVEN 主要由 3 个部分组成: 非语言行为子网络利用两个独立的循环神经网络在一个长词段内编码一系列视觉和听觉模态并输出非语言行为特征; 门控模态混合网络以原词特征、视觉特征和听觉特征作为输入, 利用注意力门控机制得到非语言语境下词的语义变化幅度和方向的非语言位移向量; 多模态偏移通过将非语言位移向量与原词特征相结合, 计算多模态偏移词表示.

(6) MulT (multi-modal Transformer)[6]针对模态序列不对齐和跨模态元素之间的长距离依赖问题. MulT 的核心是跨模态注意力块, 关注整个话语长度上的跨模态交互, 通过重复强化一个模态与其他模态的表示, 将一个模态流转换到另一个模态流, 从而不需要考虑模态序列是否对齐.

(7) MISA (modality-invariant and modality-specific)[12]将每个模态映射到两个不同的子空间, 用以学习有效的模态表示. 第 1 个子空间是模态不变的, 将一个话语的所有模态都映射到一个共享的子空间与分布对齐, 帮助捕获潜在的共性和相关表示, 缩小模态之间的差距; 第 2 个空间是模态特定的, 学习每个模态私有的表示, 补充在模态不变子空间中捕捉到的共同潜在表示. 为了学习这两个子空间, 使用了分布相似性损失 (用于模态不变子空间)、正交损失 (用于模态特定子空间)、重构损失和任务预测损失.

(8) MAG-BERT (multi-modal adaptation gate-BERT)[13]为了弥补预训练模型在视觉和听觉两种模态在微调上的差距, 使用 MAG (多模态适应门) 连接 BERT, 允许 BERT 在微调期间接受多模态非语言数据. MAG 通过对非语言行为的关注, 将视觉和听觉信息映射到一个具有轨迹和量级的向量上, 在微调过程中该适应向量修改 BERT 的内部状态, 允许模型无缝适应多模态输入. 特别地, MAG-BERT 模型在非对齐的 CH-SIMS 数据集上并不适用.

(9) SELF-MM (self-supervised multi-task learning)[14]是一个基于自监督学习策略的标签生成模块, 通过一个多模态任务和 3 个单模态子任务的联合学习获得信息丰富的单模态表示, 其中单模态子任务的标签是在自监督方法中自动生成的, 采用硬共享策略来共享底层表示学习网络. 在训练阶段, 设计了一个权重调整策略来平衡不同子任务之间的学习进度.

4.3 实验设置

(1) 实现细节

为了验证本文方法 MRA 的有效性, 本文采用了多个性能优异的多模态情感计算模型, 并且采用了模型对应论文公开的源代码超参数设置. 例如, MulT 模型使用 Adam 优化器[33]进行梯度更新, 初始化的学习率为 1E–3, L2 正则系数为 1E–3. Transformer 的头数设置为 10, 隐藏层的数目设置为 30. 针对不同的数据集设置不同的批量大小: CMU-MOSI, CMU-MOSEI 和 CH-SIMS 分别为 32, 128, 64. MulT 模型的总训练 epoch 数为 40. 为了防止过拟合, 采用丢弃率为 0.2 的 Dropout 策略, 并在训练过程中采取了提前停止策略, 若在 8 个 epoch 内验证集的准确率没有提升则停止训练. 本文的实验均在验证集上微调参数, 保留在验证集上性能最好的模型在测试集上进行测试, 并报告测试集的结果. 不同的模型超参数设置有所不同, 具体详见模型对应的论文源代码. 为了进行公平的比较, 本文采用了随机的种子初始化模型, 在 3 个数据集上均取 5 次结果的平均作为最终的结果进行汇报.

(2) 评价指标

依据之前的工作, 本文汇报了两种形式的实验结果: 分类和回归. 对于分类, 本文汇报了二分类 (积极/消极) 的准确率 (Acc2) 和加权 $F1$. 对于回归, 本文汇报了 Acc-k (k 取 5 和 7, CH-SIMS 为 Acc5, CMU-MOSI 和 CMU-MOSEI 为 Acc7), 平均绝对误差 (MAE) 以及 Pearson 相关系数 (Corr). 除了 MAE, 其余指标都是越高代表性能越好.

4.4 实验结果对比

表 2–表 4 分别展示了本文的方法 MRA 和多模态情感计算模型在 3 个多模态情感计算基准数据集上的实验结果对比 (Δ 表示基础多模态情感计算方法在增加本文方法 MRA 前后的性能差值, 其中“+”表示提升, “–”表示下降).

表 2 CMU-MOSI 数据集上方法性能比较 (%)

Approach	Acc2	加权$F1$	Acc7	MAE	Corr
EF-LSTM	76.76	76.85	35.42	96.87	64.29
+MRA	78.75	78.75	36.50	92.22	66.62
Δ	+1.99	+1.90	+1.08	−4.65	+2.33
TFN	77.29	77.34	31.43	101.62	63.27
+MRA	79.79	79.62	35.77	92.42	65.93
Δ	+2.50	+2.28	+4.34	−9.20	+2.66
MFN	77.24	77.55	34.34	97.35	65.00
+MRA	79.91	79.75	36.30	91.58	67.01
Δ	+2.67	+2.20	+1.96	−5.77	+2.01
LMF	78.90	78.87	34.84	94.56	66.30
+MRA	80.18	80.10	36.27	92.75	66.91
Δ	+1.28	+1.23	+1.43	−1.81	+0.61
RAVEN	78.93	78.87	35.33	94.53	66.01
+MRA	80.19	80.11	37.61	91.75	67.52
Δ	+1.26	+1.24	+2.28	−2.78	+1.51
MulT	80.21	80.22	36.44	91.39	68.91
+MRA	81.95	81.95	37.82	88.79	69.78
Δ	+1.74	+1.73	+1.38	−2.60	+0.87
MISA	79.79	79.64	34.29	93.39	66.82
+MRA	81.01	80.78	37.00	90.63	67.98
Δ	+1.22	+1.14	+2.71	−2.76	+1.16
MAG-BERT	81.14	81.08	40.59	79.99	75.92
+MRA	82.45	82.39	42.86	76.16	77.24
Δ	+1.31	+1.31	+2.27	−3.83	+1.32
SELF-MM	79.73	79.62	36.24	92.93	66.08
+MRA	80.67	80.45	37.93	90.22	67.40
Δ	+0.94	+0.83	+1.69	−2.71	+1.32

表 3 CMU-MOSEI 数据集上方法性能比较 (%)

Approach	Acc2	加权$F1$	Acc7	MAE	Corr
EF-LSTM	78.39	77.21	49.66	60.30	68.01
+MRA	80.06	79.62	51.23	58.74	69.46
Δ	+1.67	+2.41	+1.57	−1.56	+1.45
TFN	80.36	81.31	50.32	59.00	70.29
+MRA	82.33	82.36	51.61	57.83	71.42
Δ	+1.97	+1.05	+1.29	−1.17	+1.13
MFN	81.00	81.17	51.22	57.30	71.98
+MRA	82.86	82.78	52.10	56.28	73.25
Δ	+1.86	+1.61	+0.88	−1.02	+1.27
LMF	82.45	82.61	50.90	57.92	72.06
+MRA	83.97	83.91	52.01	56.57	73.13
Δ	+1.52	+1.30	+1.11	−1.35	+1.07
RAVEN	82.29	82.32	50.82	57.78	71.82
+MRA	83.79	83.66	51.51	56.65	72.48
Δ	+1.50	+1.34	+0.69	−1.13	+0.66
MulT	83.10	83.15	51.80	57.11	72.61
+MRA	84.04	83.98	52.06	56.73	72.96
Δ	+0.94	+0.83	+0.26	−0.38	+0.35
MISA	81.80	81.81	51.91	57.46	71.82
+MRA	83.62	83.54	52.49	56.89	72.44
Δ	+1.82	+1.73	+0.58	−0.57	+0.62
MAG-BERT	79.64	80.31	50.01	58.73	73.94
+MRA	81.63	82.06	52.32	56.09	75.02
Δ	+1.99	+1.75	+2.31	−2.64	+1.08
SELF-MM	81.79	81.69	49.39	60.84	68.55
+MRA	82.75	82.53	49.82	59.56	69.57
Δ	+0.96	+0.84	+0.43	−1.28	+1.02

表 4 CH-SIMS 数据集上方法性能比较 (%)

Approach	Acc2	加权$F1$	Acc5	MAE	Corr
EF-LSTM	69.37	56.82	21.02	59.34	−4.39
+MRA	69.37	56.82	23.23	57.62	5.48
Δ	+0.0	+0.0	+2.21	−1.72	+9.87
TFN	79.30	79.51	39.47	40.52	64.78
+MRA	80.65	80.95	46.65	37.81	67.31
Δ	+1.35	+1.44	+7.18	−2.71	+2.53
MFN	77.55	77.26	38.64	44.45	56.56
+MRA	79.25	79.10	40.74	42.86	59.64
Δ	+1.70	+1.84	+2.10	−1.59	+3.08
LMF	78.29	77.83	38.29	43.84	58.64
+MRA	80.04	79.56	41.84	42.12	60.75
Δ	+1.75	+1.73	+3.55	−1.72	+2.11
RAVEN	77.46	77.43	40.96	44.50	55.91
+MRA	79.34	78.83	43.89	43.44	58.62
Δ	+1.88	+1.40	+2.93	−1.06	+2.71
MulT	77.68	77.20	39.17	44.79	56.57
+MRA	79.08	78.63	42.49	43.59	58.17
Δ	+1.40	+1.43	+3.32	−1.20	+1.60
MISA	77.99	77.83	38.34	44.92	57.69
+MRA	79.21	78.69	42.06	43.87	58.41
Δ	+1.22	+0.86	+3.72	−1.05	+0.72
SELF-MM	78.07	77.89	40.87	42.80	58.83
+MRA	79.61	79.18	42.06	42.25	59.98
Δ	+1.54	+1.29	+1.19	−0.55	+1.15

从表 2–表 4 的结果中可以明显看出:

(1) 本文的 MRA 方法显著提升了多模态情感计算任务 (分类和回归) 的性能. 这表明了已有的多模态情感计算模型忽视了模态可信度偏差问题, 当低可信度的文本模态与高可信度的语音和视觉模态的情感不同时, 多模态情感计算模型偏向于低可信度的文本模态的情感; 然而 MRA 可以缓解模态可信度偏差问题, 有效突出高可信度的语音和视觉模态表示的重要性, 缓解低可信度的文本模态表示的偏差.

(2) 本文的 MRA 方法在 CMU-MOSI 和 CH-SIMS 这两个规模相对较小的数据集上性能提升显著 (特别是 CMU-MOSI 中的 MAE 指标, p-value<0.05), 本文认为原因在于这两个数据集中模态可信度偏差的样本比例相较于 CMU-MOSEI 更高, 使得模型在累积学习的过程中能够更好地提升对有偏差样本的关注.

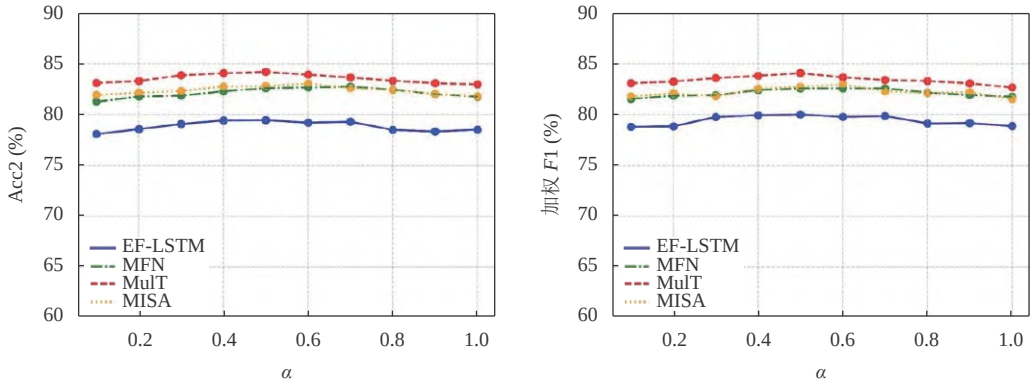
(3) 本文的 MRA 方法在大部分多模态情感计算模型上的性能提升显著 (如 TFN 模型, 其在 CMU-MOSI 数据集上的 Acc7 提升了 4.34% (p-value<0.05), MAE 下降了 9.02% (p-value<0.01)), 而在个别模型 (如 SELF-MM) 上的性能提升较小. 本文认为原因在于不同的多模态情感计算模型的结构复杂度不同, 越复杂的模型往往由很多的模块以及模态间/模态内表示学习层堆叠组合而成, 其更具稳定性和鲁棒性, 因而性能提升幅度也较小.

4.5 实验分析

(1) 累积学习超参数 α 的影响

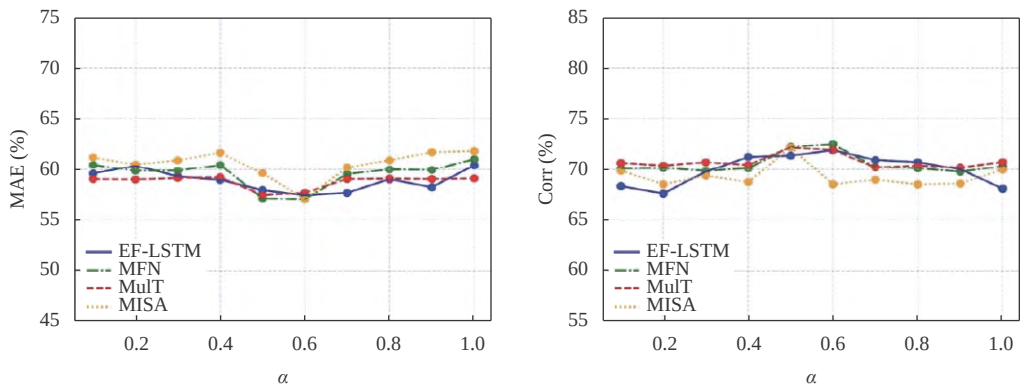
本文分析了累积学习超参数 α 变化 (变化区间为 [0.0, 1.0], 步长为 0.1) 对多模态情感计算任务性能的影响, 选取了 EF-LSTM, MFN, MulT 和 MISA 这 4 个多模态情感计算模型, 探索了这些模型增加本文的 MRA 方法后在 CMU-MOSEI 数据集上情感分类 (二分类 Acc2 和加权 $F1$) 和情感回归 (MAE 和 Corr) 两个任务的结果变化. 图 3 展示了累积学习超参数 α 对情感分类任务的指标 Acc2 (图 3(a)) 和加权 $F1$ (图 3(b)) 的结果, 图 4 展示了累积学习超参数 α 对情感回归任务的指标 MAE (图 4(a)) 和 Corr (图 4(b)) 的结果. 从这 4 个折线图中可以明显看出, 不同的多模态情感计算模型的最优累积学习超参数 α 是不同的, 并且不同任务的最优累积学习超参数 α 也不同, 甚至不

同指标的最优累积学习超参数 α 也不同, 但都集中在 [0.4, 0.6] 的区间内. 本文认为原因可能是由于采用了不同的随机种子, 使得模型的初始化参数不同, 学习过程中也存在不同最优超参数的问题. 在本文的实验中, 选取的 α 也在 [0.4, 0.6] 区间中, 大多数模型选取 $\alpha = 0.5$ 作为最优累积学习超参数 α.



(a) 二分类 Acc2　　(b) 二分类 $F1$

图 3　累积学习超参数 α 变化对分类任务指标 Acc2, 加权 $F1$ 的影响



(a) 回归 MAE　　(b) 回归 Corr

图 4　累积学习超参数 α 变化对回归任务指标 MAE, Corr 的影响

(2) 累积学习过程的可视化

本文从 CH-SIMS 数据集中随机选取了无偏差 (图 5(a)) 和有偏差样本 (图 5(b)) 进行累积学习的可视化分析, 直观地展示 MRA 方法如何缓解可信度偏差问题. 如图 5 所示, mask 是文本模态分支输出的情感预测概率, MRA 借助 mask 来动态修改损失达到渐进式学习目的. 具体而言, 对于无偏差样本, 文本模态分支输出的 mask 增加了正确情感预测的概率, 因此损失相对于有偏差样本会低. 对于有偏差样本, 文本模态分支输出的 mask 增加了错误情感预测的概率, 因此损失相对于无偏差样本会高, 使得基础多模态情感计算模型在反向传播时更关注有偏差样本的学习. MRA 方法通过这种方式提升了模型对有偏差样本的学习能力. 此外, 本文绘制了累积学习过程中不同损失的变化折线图. 如图 6 所示, L_T 的变化趋势与整体 L_{MRA} 的变化趋势接近, 中间存在波动即为学习过程中区分了存在可信度偏差的样本, 而 L_{MM} 的变化趋势则是基础多模态情感计算模型正常变化趋势.

(3) 可信度偏差样本的比例

本文统计了 3 个多模态情感计算数据集上存在可信度偏差样本的比例, 如图 7 所示. 由于 CMU-MOSI 和 CMU-MOSEI 数据集并没有标注单模态标签, 本文随机选取 MulT 和 SELF-MM 两个基础多模态情感计算模型,

并分别对 3 个数据集的测试集预测文本、语音和视觉这 3 个单模态的标签, 用以统计可信度偏差样本的比例. 特别地, CH-SIMS 数据集含有人工标注的单模态标签, 也人工统计了 CH-SIMS 测试集上真实的可信度偏差样本的比例 (19.47%). 从图 7 中可以看出, CMU-MOSI (13.37%), CMU-MOSEI (10.65%) 和 CH-SIMS (19.47%) 这 3 个数据集中可信度偏差样本比例较高, 存在明显的可信度偏差问题.

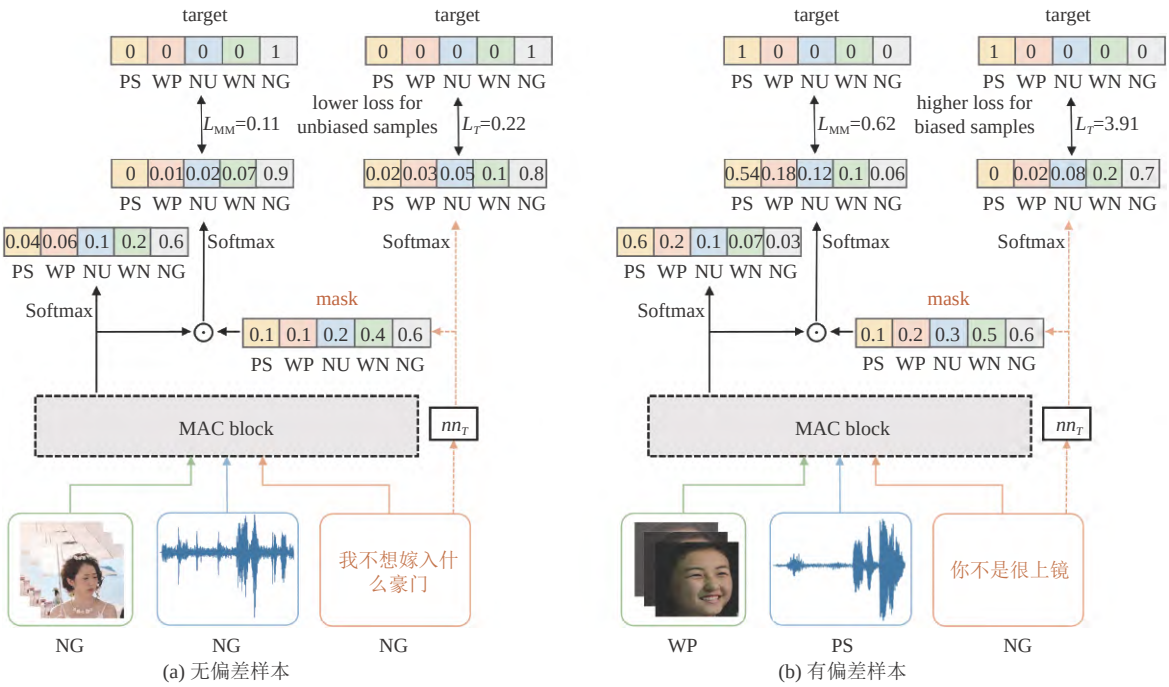


图 5 累积学习过程的可视化

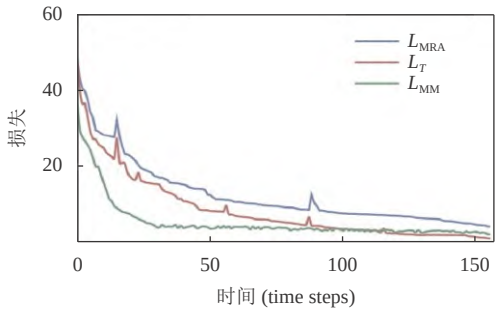


图 6 累积学习过程中不同损失的变化

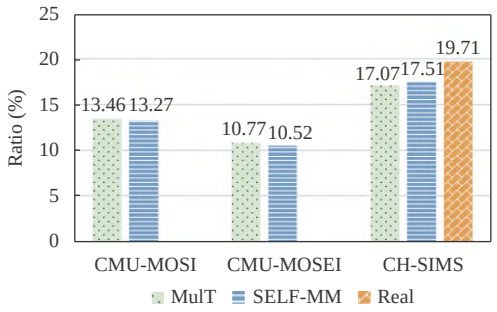


图 7 可信度偏差样本的比例

(4) 可信度偏差样本的分析

在统计可信度偏差样本比例的基础上, 本文在筛选出的存在可信度偏差样本上进一步开展了实验. 由于 CMU-MOSI 和 CH-SIMS 测试集中筛选出的样本数量过少, 实验结果极为不稳定, 因此本文汇报了 CMU-MOSEI 测试集上的结果, 如表 5 所示. 本文随机选取了 4 个基础多模态情感计算模型 (TFN, MulT, MISA 和 SELF-MM), 验证它们在加入 MRA 前后对有偏差样本的情感预测结果. 可以明显看出, 本文的 MRA 方法在有偏差样本上的提升明显高于表 3 (如 MRA 在 SELF-MM 上分别提升了 10.58% (Acc2), 7.68% (加权 $F1$), 3.60% (Acc7), 7.13% (MAE), p-value<0.01), 验证了 MRA 方法在有偏差样本上的有效性.

表 5 CMU-MOSEI 数据集 (有偏差样本) 上方法性能比较 (%)

Approach	Acc2	加权$F1$	Acc7	MAE
TFN	56.59	66.37	35.62	89.25
+MRA	61.48	70.36	37.79	85.74
Δ	+4.89	+3.99	+2.17	−3.51
MulT	58.39	67.99	36.36	84.90
+MRA	65.20	73.10	38.25	78.70
Δ	+6.81	+5.11	+1.89	−6.20
MISA	58.48	68.05	37.10	86.46
+MRA	66.36	73.98	38.94	80.42
Δ	+7.88	+5.93	+1.84	−6.04
SELF-MM	59.03	68.50	34.88	84.30
+MRA	69.61	76.18	38.48	77.17
Δ	+10.58	+7.68	+3.60	−7.13

(5) 不同模态的可信度差异分析

为了衡量不同模态的可信度, 本文统计了 CH-SIMS 数据集中文本, 语音和视觉这 3 个模态与真实多模态标签一致的比例, 如图 8 所示. 具体而言, 本文首先选择 CH-SIMS 数据集进行统计, 因为其包含人工标注的文本, 语音和视觉单模态标签和多模态标签. 其次, 本文根据 CH-SIMS 数据集划分的 5 个类别 (消极, 弱消极, 中性, 弱积极, 积极) 分别判断 3 个单模态与多模态标签是否一致, 并统计每种模态标签一致的样本数目和比例. 从图 8 中可以看出, 文本模态与多模态情感表达一致的比例 (47.52%) 低于语音 (58.44%) 和视觉 (60.42%) 的比例, 从而说明相较于文本模态, 语音和视觉模态在情感表达上更为可信.

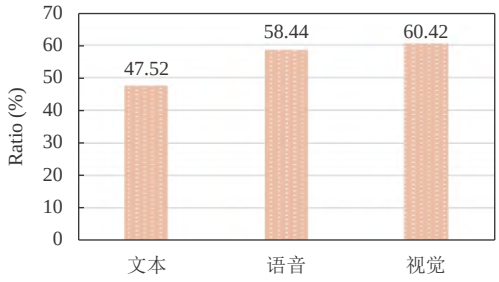


图 8 不同模态的可信度差异

(6) 有效性分析

为了更加直观地说明多模态情感计算任务中存在的可信度偏差问题, 本文从 CH-SMIS 数据集的测试集中选取了部分样本进行了有效性分析, 如图 9 所示. 图中分别展示了 3 个样本的文本, 视频帧以及语音频谱, 并标注了每个模态的情感极性 (PS, WP, NU, WN, NG 分别代表积极, 弱积极, 中性, 弱消极, 消极). 本文选取了 MulT 模型对抽取的 3 个样本进行情感预测, 每个样本的右侧给出了 MulT 和 MulT 模型加上 MRA 方法的情感标签概率扇形图. 可以明显看出, MulT 模型在 3 个样本上都预测错误, 且都倾向于低可信度的文本模态的情感. 然而, 样本真实的多模态情感却与低可信度的文本模态的情感相悖, 与高可信度的语音和视觉模态一致. 由此观之, 目前的多模态情感计算模型忽视了可信度偏差问题, 导致其无法达到高精度的情感预测. 在 MulT 模型的基础上引入本文的方法 MRA 之后, 可以明显地看出模型在 3 个样本上给出了正确的情感预测, 不会再倾向低可信度的文本模态的情感. 这表明了本文的方法 MRA 确实能够有效缓解多模态情感计算模型中的可信度偏差问题, 突出高可信度的语音和视觉模态在情感预测上的重要性, 减少低可信度的文本模态的误导, 从而提升多模态情感计算模型的性能.

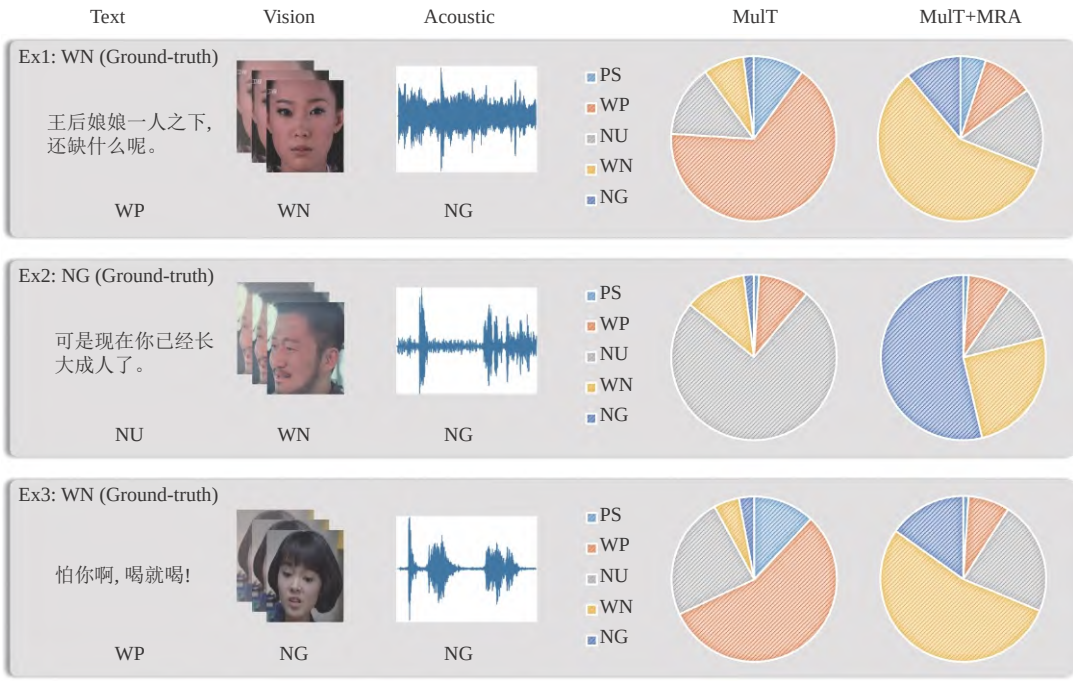


图 9 CH-SIMS 数据集中的样本及情感预测概率扇形图

5 总 结

本文针对多模态情感计算任务中的可信度偏差问题进行了研究. 已有的多模态情感计算模型, 无论是设计复杂的融合策略抑或是学习丰富的模态表示, 都忽视了模态可信度偏差问题. 本文结合初步的实验分析认为, 语音和视觉模态相较于文本模态更能真实反映情感, 是高可信度的. 然而, 目前针对不同模态的特征抽取工具的学习能力不同, 导致不同模态表示之间存在强弱之分, 因而多模态情感计算模型偏向于低可信度的文本模态, 难以达到高精度的情感预测. 为了缓解模态可信度偏差问题, 本文提出了一种模型无关的基于累积学习的多模态可信度感知的情感计算方法 MRA, 其通过单独训练的文本模态分支学习低可信度的文本模态偏差, 再利用累积学习策略使得多模态情感计算模型在训练过程中缓解并消除文本模态偏差. 本文在多个基准数据集上利用不同的多模态情感计算模型进行了实验, 验证了 MRA 的有效性和通用性. 在下一步的工作中, 我们考虑引入可信学习 (trusted learning)[34] 来度量不同模态的可信度, 进一步提升多模态情感计算任务的性能. 此外, 我们将进一步探索多模态情感计算任务中的其他问题, 尝试将 MRA 方法扩展到其他存在模态可信度偏差的多模态任务中, 如多模态抑郁症检测, 视觉问答 (visual question answering, VQA) 等.

References:

[1] Yang Y, Zhan DC, Jiang Y, Xiong H. Reliable multi-modal learning: A survey. Ruan Jian Xue Bao/Journal of Software, 2021, 32(4): 1067–1081 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6167.htm [doi: 10.13328/j.cnki.jos.006167]

[2] Wöllmer M, Weninger F, Knaup T, Schuller W, Sun CK, Sagae K, Morency LP. YouTube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems, 2013, 28(3): 46–53. [doi: 10.1109/MIS.2013.34]

[3] 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: Association for Computational Linguistics, 2017. 1103–1114. [doi: 10.18653/v1/D17-1115]

[4] Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP. Memory fusion network for multi-view sequential learning. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 5634–5641. [doi: 10.1609/aaai.v32i1.12021]

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

[6] 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: Association for Computational Linguistics, 2019. 6558–6569. [doi: 10.18653/v1/P19-1656]

[7] Tsai YHH, Ma M, Yang MQ, Salakhutdinov R, Morency LP. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 1823–1833. [doi: 10.18653/v1/2020.emnlp-main.143]

[8] Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 9180–9192. [doi: 10.18653/v1/2021.emnlp-main.723]

[9] Wang YS, Shen Y, Liu Z, Liang PP, Zadeh A, Morency LP. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 7216–7223. [doi: 10.1609/aaai.v33i01.33017216]

[10] 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]

[11] Sun ZK, Sarma P, Sethares W, Liang YY. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 8992–8999. [doi: 10.1609/aaai.v34i05.6431]

[12] 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]

[13] Rahman W, Hasan MK, Lee S, Zadeh AB, 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. Association for Computational Linguistics, 2020. 2359–2369. [doi: 10.18653/v1/2020.acl-main.214]

[14] 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]

[15] Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. EPJ Data Science, 2017, 6: 15. [doi: 10.1140/epjds/s13688-017-0110-z]

[16] Thórisson KR, Bieger J, Li X, Wang P. Cumulative learning. In: Proc. of the 12th Int'l Conf. on Artificial General Intelligence. Shenzhen: Springer, 2019. 198–208. [doi: 10.1007/978-3-030-27005-6_20]

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

[18] Fu ZW, Liu F, Xu Q, Qi YJ, Fu XL, Zhou AM, Li ZB. NHFNET: A non-homogeneous fusion network for multimodal sentiment analysis. In: Proc. of the 2022 IEEE Int'l Conf. on Multimedia and Expo. Taipei: IEEE, 2022. 1–6. [doi: 10.1109/ICME52920.2022.9859836]

[19] Kozerawski J, Turk M. CLEAR: Cumulative learning for one-shot one-class image recognition. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3446–3455. [doi: 10.1109/CVPR.2018.00363]

[20] Zhou BY, Cui Q, Wei XS, Chen ZM. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9716–9725. [doi: 10.1109/CVPR42600.2020.00974]

[21] Naderian P, Loaiza-Ganem G, Braviner HJ, Caterini AL, Cresswell JC, Li T, Garg A. C-learning: Horizon-aware cumulative accessibility estimation. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[22] Cadene R, Dancette C, Ben-Younes H, Cord M, Parikh D. RUBi: Reducing unimodal biases for visual question answering. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 76. [doi: 10.5555/3454287.3454363]

[23] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014. 1532–1543. [doi: 10.3115/v1/D14-1162]

[24] 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. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[25] 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. Florence: IEEE, 2014. 960–964. [doi: 10.1109/ICASSP.2014.6853739]

[26] McFee B, Raffel C, Liang DW, Ellis DPW, McVicar M, Battenberg E, Nieto O. LibROSA: Audio and music signal analysis in Python. In: Proc. of the 14th Python in Science Conf. Austin: scipy.org, 2015. 18–24.

[27] Ekman P, Rosenberg EL. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS). Oxford: Oxford University Press, 1997.

[28] Zhang KP, Zhang ZP, Li ZF, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016, 23(10): 1499–1503. [doi: 10.1109/LSP.2016.2603342]

[29] Baltrusaitis T, Zadeh A, Lim YC, Morency LP. OpenFace 2.0: Facial behavior analysis toolkit. In: Proc. of the 13th IEEE Int'l Conf. on Automatic Face & Gesture Recognition. Xi'an: IEEE, 2018. 59–66. [doi: 10.1109/FG.2018.00019]

[30] Zadeh A, Zellers R, Pincus E, Morency LP. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259, 2016.

[31] 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. Melbourne: Association for Computational Linguistics, 2018. 2236–2246. [doi: 10.18653/v1/P18-1208]

[32] 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. Association for Computational Linguistics, 2020. 3718–3727. [doi: 10.18653/v1/2020.acl-main.343]

[33] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2015.

[34] 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]

附中文参考文献:

[1] 杨杨, 詹德川, 姜远, 熊辉. 可靠多模态学习综述. 软件学报, 2021, 32(4): 1067–1081. http://www.jos.org.cn/1000-9825/6167.htm [doi: 10.13328/j.cnki.jos.006167]

罗佳敏(1997－), 女, 博士生, CCF 学生会员, 主要研究领域为自然语言处理.

周国栋(1967－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理.

王晶晶(1990－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理.