软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(10):4781−4793 [doi: 10.13328/j.cnki.jos.006998]

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

基于自适应权值融合的多模态情感分析方法*

罗渊贻, 吴 锐, 刘家锋, 唐降龙

(哈尔滨工业大学 计算机科学与技术学院, 黑龙江 哈尔滨 150006)

通信作者: 吴锐, E-mail: simple@hit.edu.cn

摘 要: 多模态情感分析是利用多种模态的主观信息对情感进行分析的一种多模态任务, 探索模态间的有效交互是多模态分析中的一项重要研究. 在最近的研究中发现, 由于模态的学习速率不平衡, 导致单个模态收敛时, 其余模态仍处于欠拟合的状态, 进而削弱了多模态协同决策的效果. 为了能更有效地将多种模态结合, 学习到更具有表达力的情感特征表示, 提出一种基于自适应权值融合的多模态情感分析方法. 所提方法分为两个阶段: 第 1 个阶段是根据不同模态的学习梯度差异自适应地改变单模态特征表示的融合权值, 实现动态调整模态学习速率的目的, 把该阶段称为 B 融合 (balanced fusion). 第 2 个阶段是为了消除 B 融合的融合权值对任务分析的影响, 提出模态注意力探究模态对任务的贡献, 并根据贡献为各模态分配权重, 把该阶段称为 A 融合 (attention fusion). 用于情感分析的多模态表示由 B 融合和 A 融合的结果共同组成. 实验结果显示, 将 B 融合方法引入现有的多模态情感分析方法中, 能够有效提升现有方法对情感分析任务的分析准确度; 消融实验结果显示, 在 B 融合的基础上增加 A 融合方法能有效减小 B 融合权重对任务的影响, 有利于提升情感分析任务的准确度. 与现有的多模态情感分析模型相比, 所提方法结构更简单、运算时间更少, 且任务准确率优于对比模型, 表明所提方法在多模态情感分析任务中的高效性和优异性能.

关键词: 多模态情感分析; 学习平衡; 多模态融合; 自适应学习

中图法分类号: TP18

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/6998.htm

Multimodal Sentiment Analysis Method Based on Adaptive Weight Fusion

LUO Yuan-Yi, WU Rui, LIU Jia-Feng, TANG Xiang-Long

(College of Computer Science and Technology, Harbin Institute of technology, Harbin 150006, China)

Abstract: Multimodal sentiment analysis is a task that uses subjective information from multiple modalities to analyze sentiment. Exploring how to effectively learn the interaction between modalities has always been an essential task in multimodal analysis. In recent research, it is found that the learning rate of different modalities is unbalanced, leading to the convergence of one modality while the rest of the modalities are under-fitting, which weakens the effect of multimodal collaborative decision-making. In order to combine multiple modalities more effectively and learn the multimodal sentiment features with rich expression, this study proposes a multimodal sentiment analysis method based on adaptive weight fusion. The method of adaptive weight fusion is divided into two phases. The first phase is to adaptively change the fusion weights of unimodal feature representations according to the difference of unimodal learning gradients to dynamically balance the modal learning rate. The study calls this phase balanced fusion (B-fusion). The second phase is to eliminate the impact of the fusion weights of B-fusion on task analysis, propose the modal attention to explore the contributions of modalities to the task, and dynamically allocate the fusion weight to each modality. The study calls this phase attention fusion (A-fusion). The experimental

* 基金项目: 国家自然科学基金 (61672190)
收稿时间: 2022-12-07; 修改时间: 2023-03-06, 2023-05-29; 采用时间: 2023-07-08; jos 在线出版时间: 2023-09-27
CNKI 网络首发时间: 2023-10-07

results show that the introduction of the B-fusion method into existing multimodal sentiment analysis methods can effectively improve the accuracy of sentiment analysis. The ablation experiment results show that adding the A-fusion method to B-fusion can effectively reduce the impact of B-fusion weights on the task, which is conducive to improving the analysis results of sentiment analysis. Compared with the existing multimodal sentiment analysis models, the proposed method has a simpler structure, lower computational consumption, and better task accuracy than these comparison models, which shows that the method has high efficiency and excellent performance in multimodal sentiment analysis tasks.

Key words: multimodal sentiment analysis; balanced learning; multimodal fusion; adaptive learning

在线产品评论的观点判断是自然语言处理研究的一个重要主题, 有着重要而广泛的应用前景. 随着在线视频的普及和视频平台的日益增多, 越来越多的人被吸引通过视频表达自己的观点. 利用机器学习算法了解这些视频中涉及的情感取向有助于我们快速捕捉人们对某些主题的态度, 这使得多模态情感分析成为一个关键的研究领域, 而如何获取模态间的有效交互也成为情感分析中的重要任务[1].

本文的研究对象是以文本、语音、视频这 3 种模态为基础的多模态情感分析任务. 理论上, 现有的多模态情感分析模型接收更多元化的情感信息, 应该优于单模态模型的分析结果. 然而在最近的研究中指出[2], 某些场景下单模态模型的分析结果要优于多模态模型, 并且我们发现此种现象也会出现在多模态情感分析模型中. 上述现象产生的原因是由于不同模态以不同的速率进行拟合和泛化, 而采用单一的优化策略训练所有模态会存在某个模态已拟合时, 其余模态仍处于欠拟合的情况, 导致单模态学习网络无法得到充分学习, 进而影响融合效果. 为解决上述问题, 现有的研究采用额外的单模态损失来辅助多模态模型训练[3], 将单模态之间的训练分离, 得到最优的单模态特征表示; 另一些研究采用单模态之间的特征数值比例作为判断依据对单模态学习速率进行调整[4]. 上述方法在特征层面和任务学习层面来使不同模态的学习速率达到平衡, 没有考虑梯度对模态训练速率的影响.

为了更直观地体现模态间学习不平衡对多模态情感分析任务的影响, 本文以 CMU-MOSI 数据集[5]为基础, 选取 3 种现有的情感分析模型: 基于深度神经网络的后期融合模型 (late fusion based on deep neural network, LF-DNN)、基于长短期记忆的后期融合模型 (late fusion based on long short-term memory, LF-LSTM)、张量融合模型 (tensor fusion network, TFN)[6]进行对比分析, 分析结果如表 1 和图 1 所示. 其中 T, A, V 分别表示文本模态 (text)、音频模态 (audio) 和视频模态 (vision). 表 1 描述了在情感分析任务中, 单模态学习模型和多模态学习模型的二分类测试结果 (Acc-2). 结果表明, 单模态 (文本模态) 的任务分析结果优于多模态. 此外, 我们获取多模态模型中单模态学习网络的最后一层线性变化的梯度幅度进行对比, 结果如图 1 所示. 图 1 中的纵坐标表示单模态学习网络的最后一层线性变换的学习梯度幅度, 即梯度的绝对值, 代表了单个模态的学习速率. 横坐标表示模型训练的轮次. 从图 1 中可以看出, 文本模态的学习梯度幅度在训练过程中远超出其余模态, 而音频模态和视频模态的学习梯度幅度很低, 接近于 0, 由此印证了表 1 中文本模态因更快收敛而表现出更优的情感任务分析结果.

表 1 单模态与多模态情感分析结果对比

模型	单模态训练结果 (T)	多模态训练结果
LF-DNN	0.782 8	0.766 8
LF-LSTM	0.760 9	0.747 8
TFN	0.777 4	0.753 0

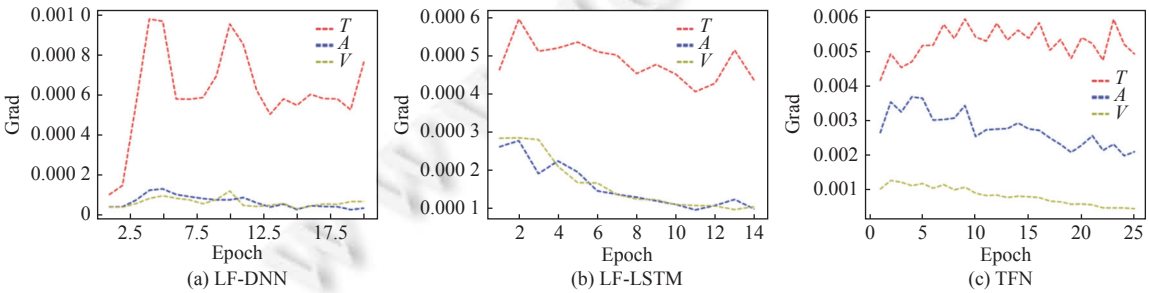


图 1 多模态情感分析模型中单模态学习网络的学习梯度幅度对比

综上分析, 为了增强单模态特征的情感语义表达, 提升模型的情感分析能力, 本文提出一种基于自适应权值融合的多模态情感分析方法. 自适应权值融合的过程分为两个阶段: 第 1 个阶段用于平衡单模态网络的学习速率, 我们称它为 B 融合 (balanced fusion), 第 2 个阶段用于调整单模态对任务分析的贡献, 我们称它为 A 融合 (attention fusion). 在本文模型中, B 融合首先对单模态学习网络输出的特征表示进行加权融合, 并采用各模态学习网络之间的梯度幅度差异作为损失值对 B 融合的融合权值进行优化, 实现在反向传播过程中自适应调整单模态学习梯度的目的. 然而在前向传播过程中, 单模态的融合权值也代表了各模态在对任务分析的贡献度, 较低的权值表示模态包含的任务信息弱, 对任务分析的影响小[7,8]. 因此, 本文基于软注意力机制提出多模态注意力方法对各单模态特征表示进行注意力加权融合, 即 A 融合. 在反向传播过程中, 由于 A 融合权值用于决定单模态对任务的贡献, 而对单模态的学习速率没有影响, 因此 A 融合模块的参数学习过程只与任务损失有关, 与单模态特征学习网络和 B 融合学习无关. 用于任务决策的多模态特征表示由 A 融合的结果与 B 融合的结果共同组成. 综上所述, 本文提出的自适应权值融合方法不仅能有效平衡模态间的学习速率, 也能学习各模态对任务的贡献.

1 相关工作

1.1 多模态情感分析

情感分析被广泛定义为对主观因素的计算研究, 其中包括对人的意见、态度和情绪[7]. 多模态情感分析是利用多种形式的主观表达对情感进行分析判断的一种情感分析方法, 当前的多模态情感分析方法主要针对多模态融合[2,8,9]和多模态表示进行研究[6,10,11].

多模态融合方面, Zadeh 等人[6]于 2017 年提出一种多模态张量融合方法对 3 种模态进行外积融合, 得到的融合结果通过全连接深度网络模型进行最终的情感分析, 该方法不仅能保存单个模态的内部信息, 也能学习模态之间的互补信息; Tsai 等人[12]于 2019 年提出一种多模态转换器结构, 该结构利用交叉模态注意力将文本、音频、图像这 3 种模态进行交叉融合, 使单个模态可以从其他的模态中获取信息. 多模态表示方面, Hazarika 等人[13]于 2020 年提出一种模态共性和特性的表示方法, 利用损失函数学习单模态表示之间的共性和特性, 将这些模态间的共性表示和特性表示进行融合, 减少了模态之间的信息冗余; Rahman 等人[14]于 2020 年提出一种多模态适应门的结构用于多模态模型的微调; Yu 等人[15]设计了一种自监督学习策略的标签生成模块, 并以多任务学习的方式分别学习模态间的一致性表示和差异性表示; Wu 等人[16]利用视频和音频模态的共享语义和私有语义分别对文本模态进行增强和互补, 并提出一种以文本为中心的共享私有框架进行情感分析. 上述的多模态情感分析模型都是采用复杂的融合方式形成用于决策的多模态特征表示, 没有考虑到模态间学习速率不平衡对情感分析任务的影响. 此外, 不同模态的情感语义表达强度存在差异, 进而对任务的贡献度也不同. 本文提出的方法既平衡了模态学习, 也在此基础上利用软注意力机制动态为模态分配注意力权重, 学习不同模态对任务的贡献.

1.2 多模态平衡学习

虽然结合多种模式信息的多模态学习模型在理论上要优于单模态的学习模型, 然而在一些情况下, 依然存在单模态模型结果优于多模态模型的现象. 一些研究将此类现象归结于模态间拟合速率的差异[2,3]. 针对上述问题, Sun 等人[3]在多模态处理任务的基础上增加单模态任务模型, 并根据单模态任务得到的结果计算出可自适应调整的乘积因子, 用于调节梯度变化的速率. Wang 等人[2]通过模态间损失下降的速度差异提出一种梯度混合方法. Chen 等人[17]增加单模态任务损失, 并通过单模态任务损失的权值动态调整单模态的学习速率. Peng 等人[4]发现在反向传播过程中, 单模态之间训练的结果差异会严重影响单模态学习的平衡, 由此对各单模态融合前学习到的最后结果进行打分, 并利用模态间的分数比作为学习率来改善模态间的学习不平衡问题. 上述方法在特征层面和任务学习层面来解决多模态的学习不平衡问题, 没有考虑到模型训练过程中产生的参数学习梯度对模态学习速率的影响.

1.3 注意力机制

在现实生活中, 人们会选择性地关注所有信息的一部分, 同时忽略其他信息, 并以不同的重要程度对获取的信

息进行处理. 注意力模型最初用于机器翻译领域, 目前已成为大多数神经网络结构的重要组成部分, 在多个领域都有着大量的应用[8,17]. Xu 等人[18]于 2015 年提出软注意力机制为输入的信息分配权值, 并进行加权平均处理. 近年来, 注意力机制也多被使用在多模态机器学习的研究领域, Long 等人[19]于 2018 年以多种场景下的视频信息为基础, 对 RGB 图、光流图、音频这 3 种模态进行重要性分析, 结论表明在不同的应用场景下, 各模态表现出来的重要程度不同. 例如在运动场景下, 光流图占比高, 在清洗牙刷等声音识别性较高的场景下, 音频占比高, 而在静止且无声的环境下, RGB 图占比高. Ghosal 等人[7]发现, 在多模态任务的研究中, 不是合并所有可用的模态都能有利于提高模型分析性能, 也不是所有的模态都能在训练中发挥平等的作用, 并针对这些问题提出基于多模态注意力框架的循环神经网络结构. 李群等人[20]利用协同注意力机制能增加模态间交互能力的特点, 实现视觉流和自然语言流之间的有效交互.

2 本文模型

模型的学习过程体现在模型参数的更新, 而参数更新的频率和幅度决定了模型学习的速率大小. 为更有效地结合 3 种模态, 防止因学习速率不平衡导致的性能下降, 本文提出一种基于自适应权值融合的多模态情感分析方法, 根据模态学习的梯度差异自适应地调整模态间的融合权值, 促使单模态融合权值能在反向传播过程中平衡不同模态的参数更新速率. 此外, 本文方法根据不同模态对任务的贡献动态调整单模态在任务分析中的权重, 加强重要模态对任务的作用, 减小次要模态对任务的影响. 本文提出的方法模型如图 2 所示, 其中 $\varphi_i=(X_i;W_i)$ 为单个模态的表示学习网络, $i\in(T,A,V)$. 3 种单模态表示学习网络的学习过程如公式 (1)–公式 (3) 所示. 其中 W_i 是线性变换参数, BERT[21]为文本模态的预训练模型, sLSTM[22]为单向长短期记忆模型, *Batch* 表示单批次训练的样本个数, d_T、d_A、d_V 分别为文本模态、音频模态、视频模态的特征维度, 为了方便计算, 3 种模态表示的特征维度相等, 即 $d_T=d_A=d_V=d$.

$$\varphi_T=W_T^{\mathrm{T}}\mathrm{BERT}(X_T)\in\mathbb{R}^{Batch\times d_T} \tag{1}$$

$$\varphi_A=W_A^{\mathrm{T}}\mathrm{sLSTM}(X_A)\in\mathbb{R}^{Batch\times d_A} \tag{2}$$

$$\varphi_V=W_V^{\mathrm{T}}\mathrm{sLSTM}(X_V)\in\mathbb{R}^{Batch\times d_V} \tag{3}$$

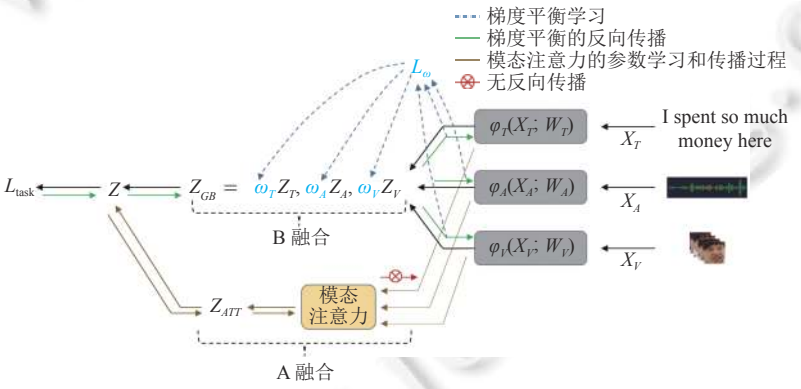


图 2 本文模型

由图 2 可以看出, 模型在前向传播中存在两次加权融合的过程, 目的分别为平衡模态间的学习速率和动态调整模态对任务的贡献. 两次加权融合分别得到的多模态特征表示为 Z_{GB} (B 融合的结果) 和 Z_{ATT} (A 融合的结果), 最终用于任务决策的多模态特征表示为 $Z=(Z_{GB}+Z_{ATT})/2$.

2.1 梯度平衡学习

定义文本模态学习网络、音频模态学习网络、视频模态学习网络输出的特征表示分别为 Z_T, Z_A, Z_V. 其中 $Z_i=\varphi_i(X_i;W_i)$, X_i 为模态初始输入特征, W_i 为模态训练过程中的可学习参数. 用于 B 融合的单模态融合权值分别

为 ω_T, ω_A, ω_V . 为了平衡各模态之间的学习速率, 在学习过程中需要尽可能减小大梯度学习模态的学习速率. 我们在实验中选择 1 作为 B 融合的初始权重. 当模态的 B 融合权重小于 1 时, 能有效减小模态学习速率. 此外, 将 1 作为 B 融合权重的初始值, 能通过 B 融合权重的变化直观看出本文模型是否能有效平衡不同模态的学习速率. B 融合的结果可以表示为:

$$Z_{GB} = [\omega_T Z_T, \omega_A Z_A, \omega_V Z_V] \tag{4}$$

其中, $[\cdot]$ 为特征拼接. 单模态学习过程中的参数更新可以表示为如公式 (5) 所示. 其中 W_i 为单模态训练中的可学习参数, L_{task} 为任务损失.

$$W_i(t+1) = W_i(t) - \frac{\partial L_{\text{task}}}{\partial Z_{GB}} \cdot \frac{\partial Z_{GB}}{\partial Z_i} \cdot \frac{\partial Z_i}{\partial W_i} = W_i(t) - \frac{\partial L_{\text{task}}}{\partial Z_{GB}} \cdot \omega_i \cdot \frac{\partial Z_i}{\partial W_i} \tag{5}$$

从公式 (5) 可以看出, 单模态的参数更新幅度由对应的模态融合权值 ω_i 进行自适应调整. 为使 B 融合的方法能有效平衡模态间的学习速率, 单模态的融合权值 ω_i 应实现如下功能: (1) 通过模态间学习速率的差异而自适应调整; (2) 模态的学习速率越低, 对应的 B 融合权值 ω_i 越大, 模态学习速率越高, 对应的 B 融合权值越小, 目的是降低以高速率学习的模态梯度幅度, 增大以低速率学习的模态梯度幅度. 综上分析, 本文采用单模态学习梯度与阈值梯度之间的距离作为损失函数对 B 融合的融合权值 ω_i 进行学习. 损失函数的表达式如公式 (6) 所示:

$$L_\omega = \sum_{i=(T,A,V)} |g_i(t) - \bar{g}(t) r_i(t)| \tag{6}$$

其中, $g_i(t)$ 为单模态学习过程中所有参数梯度的 1 范数均值. 为减小计算量, 本文采用单模态学习网络中最后一层线性变换的参数梯度进行计算. $g_i(t)$ 的计算可以表示为如公式 (7) 所示. $\bar{g}(t)$ 为 3 种模态的梯度幅度均值, 表示为如公式 (8) 所示. $r_i(t)$ 为针对不同模态设置的阈值梯度调整因子, 目的是提升低速率学习模态的梯度阈值, 增大学习速率, 具体可表示为如公式 (9) 所示.

$$g_i(t) = average\left(\left\| \frac{\partial L_{\text{task}}}{\partial W_i} \right\|_1\right) \tag{7}$$

其中, L_{task} 表示任务损失. W_i 为单模态学习网络中最后一层线性变换的可学习参数, 即公式 (1)–公式 (3) 中的 W . $\|\cdot\|_1$ 为 1 范数计算.

$$\bar{g}(t) = \frac{\sum_{i=(T,A,V)} g_i(t)}{3} \tag{8}$$

$$r_i(t) = \frac{\max_{k=(T,A,V)} g_k(t)}{g_i(t)} \tag{9}$$

从公式 (6)–公式 (9) 中可以看出, 当单个模态 i 的学习速率较小时 (即 $g_i(t)$ 小), 通过 $r_i(t)$ 的尺度变换, 增大该模态的梯度阈值 $\bar{g}(t) r_i(t)$, 则该模态能获得更大的学习梯度, 增大模态的学习速率; 而当模态的学习速率大时 (即 $g_i(t)$ 大), 该模态的梯度阈值 $\bar{g}(t) r_i(t)$ 小于现有梯度 $g_i(t)$, 则该模态的学习梯度变小, 减小模态的学习速率.

本文采用随机梯度下降法 (stochastic gradient descent, SGD) 对 B 融合中的模态融合权值 ω_i 进行优化. 第 t 个时间步的融合权值 $\omega_i(t)$ 可以由公式 (10) 所示的表达式进行更新.

$$\omega_i(t+1) = \omega_i(t) - r_l \cdot \nabla_{\omega_i} L_\omega \tag{10}$$

其中, $\nabla_{\omega_i} L_\omega$ 为 ω_i 的学习梯度, r_l 为超参数 (学习率).

2.2 模态注意力

为了避免 B 融合方法的融合权值在前向传播过程中对任务决策产生较大的影响, 本文以软注意力机制为基础, 设计一种模态注意力方法根据单模态对任务的贡献赋予各单模态特征表示不同的权重, 并进行特征连接 (A 融合). 具体结构如图 3 所示.

图 3 中, Z_T, Z_A, Z_V 分别为文本模态、音频模态、视频模态的特征表示. α_T, α_A, α_V 分别为模型学习到的注意力权重, 即 3 种模态对任务的贡献度. 模态的注意力权值可以表示为:

$$h = ReLU\left((avepool[Z_T, Z_A, Z_V])\, W_{ATT}^{3\times 3}\right) \tag{11}$$

$$[\alpha_T,\alpha_A,\alpha_V]=Sigmoid\left(hW_{ATT}^{3\times3}\right) \tag{12}$$

其中, *Sigmoid* 为 S 型函数 $\frac{1}{1+e^{-x}}$, *avepool* 表示平均池化层.

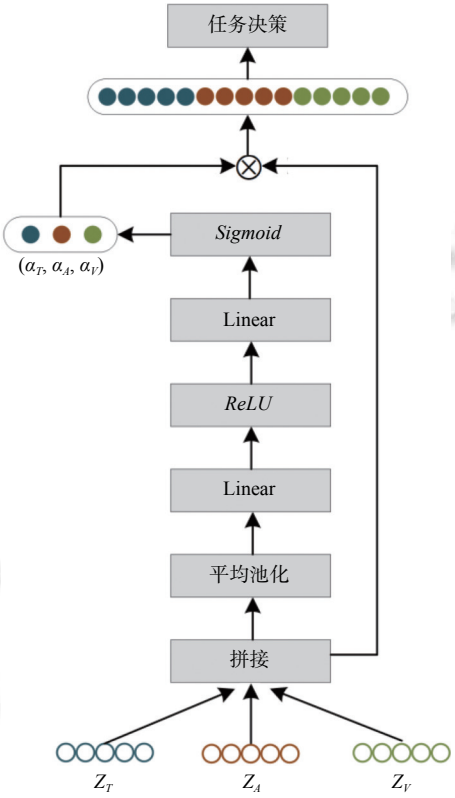


图 3 模态注意力

经过 A 融合后的多模态表示为:

$$Z_{ATT}=[\alpha_T Z_T,\alpha_A Z_A,\alpha_V Z_V] \tag{13}$$

其中, $[\cdot]$ 表示特征拼接.

本文提出的自适应权值融合算法如算法 1 所示.

算法 1. 自适应权值融合算法.

1. 前向传播:
2. 获取单模态特征表示: Z_T,Z_A,Z_V
3. 初始化 B 融合的融合权值 $\omega_i(t=0)=1,\ i=T,A,V$
4. 计算 B 融合的结果 $Z_{GB}=[\omega_T Z_T,\omega_A Z_A,\omega_V Z_V]$
5. 计算 A 融合的融合权值 $\alpha_i,\ i=T,A,V$
6. 计算 A 融合的结果 $Z_{ATT}=[\alpha_T Z_T,\alpha_A Z_A,\alpha_V Z_V]$
7. 计算用于任务决策的多模态表示 $Z=(Z_{GB}+Z_{ATT})/2$
8. 反向传播:
9. 从 t=0 到最终训练结束:
10. 输入单个批次样本计算任务损失 L_{task}

11. 计算 $g_i(t)$
12. 计算 $\bar{g}(t)$
13. 计算用于 B 融合权值学习的损失 L_ω
14. 通过任务损失计算参数梯度 $\nabla_W L_{\text{task}}$ (其中包含 A 融合的参数训练梯度 $\nabla_{W_{ATT}} L_{\text{task}}$)
15. 通过模态的梯度损失 L_ω 计算 B 融合的训练梯度 $\nabla_{\omega_i} L_\omega$
16. 利用 $\nabla_{\omega_i} L_\omega$ 更新 B 融合的融合权值 $\omega_i(t) \to \omega_i(t+1)$
17. 利用 $\nabla_W L_{\text{task}}$ 更新剩余训练参数 $W(t) \to W(t+1)$
18. 结束循环

3 实 验

3.1 数据集

将本文模型与现有的多模态模型在情感分析和情绪分析两种下游任务中进行对比, 测试数据集包括情感分析数据集: CMU-MOSI、CMU-MOSEI、CH-SIMS; 情绪分析数据集: IEMOCAP. 所有数据集的介绍如下所示.

CMU-MOSI[5] & CMU-MOSEI[23]: CMU-MOSI 数据集是由 Zadeh 等人[5]在 2016 年提出的第 1 个通过观点进行标注的情感分析数据集, 它包括了独白、演讲、电影等多种形式的观点数据, 共计 93 个视频, 2 198 个视频片段. 这些视频片段被人为手工标注为 [−3, 3] 之间的情感评分, −3 表示十分消极的情感, 3 表示十分积极的情感. CMU-MOSEI 数据集是 Zadeh 等人[23]在 2018 年针对 CMU-MOSI 进行改进提出的情感分析数据集. CMU-MOSEI 具有更多的样本数量, 表达者和主题也具备了更大的多样性. 该数据集包含 23 453 个视频片段, 分别来自于 5 000 个不同的视频. 本文对 CMU-MOSI 和 CMU-MOSEI 两种数据集的分类结果使用相同的评估标准: 二分类 (将数据标签分为积极和消极两类)、$F1$ 分数 (考虑了召回率和准确率的同等重要性)、平均绝对误差 (mean absolute error, MAE) 和皮尔逊相关 (Pearson correlation, Corr). 对于这些评估指标, 除 MAE 外, 值越高表示该模型对于情感分析任务的性能越好.

IEMOCAP[24]: 该数据集是情绪分析数据集, 其中包含了 10 000 个人类情绪分析的视频. 这是一个多标签数据集, 即每个样本都会针对多种情绪进行打分. 与文献 [12] 相同, 本文选择 4 种情绪 (开心、悲伤、生气、中立) 用于任务分析, 并针对每个情绪的二分类结果和 $F1$ 分数进行测试.

CH-SIMS[25]: 该数据集不仅包含了多模态表标签, 也包含了单模态标签, 利于研究单个模态之间的关系. CH-SIMS 包含了 2 281 个视频片段, 分别来自于 60 部不同种类的电影. 模型针对该数据集的测试结果采用二分类和 $F1$ 分数作为评价标准.

3.2 对比模型

在本文实验中, LF-DNN 模型与 LF-LSTM 模型的实验结果来源于文献 [15] 提供的可获取开源代码. 所有对比实验模型的描述如下.

1) LF-DNN: 基于深度神经网络的后期融合模型.

2) LF-LSTM: 基于长短期记忆的后期融合模型.

3) 张量融合网络 (tensor fusion network, TFN)[6]. 该网络通过创建一个多维的张量, 端到端学习模态内和模态间的动态信息.

4) 低秩多模态融合网络 (low-rank multimodal fusion, LMF)[10]. 该网络能减少由张量运算带来的复杂度, 并获取模态内和模态间的信息.

5) 多模态转换器 (multimodal Transformer, MulT)[12]. 该方法利用交叉模态注意机制获取多模态序列之间的远程交互, 并使单个模态可以从其余模态中获取辅助信息.

6) 基于 Transformer 的低秩融合方法 (low rank fusion based Transformer, LMF-MulT)[26]. 该文献将 LMF 和

MulT 两种模型进行融合, 提出一种新的算法模型, 该模型相比于 LMF 和 MulT 拥有更低的运算时间和更高的精确度.

7) 模态的共性和特性表示学习模型 (modality-invariant and -specific representation, MISA)[13]. 该方法利用损失函数学习各个单模态表示之间的共性特征和特性特征, 利用 Transformer 将学习到的特征进行融合.

8) 多模态自适应门限 (the multimodal adaptation gate for BERT, MAG-BERT)[14]. 文献 [14] 提出了一个基于 BERT 和 XLNet 微调模型的多模态自适应门, 允许在微调的过程中接受多模态非语言数据.

9) 自监督的多任务学习 (self-supervised multi-task learning, Self-MM)[15]. 该方法设计了一种单模态标签自动生成器, 为每个模态提供单独的单模态训练任务, 该模型属于多任务处理模型.

10) 层次图对比学习 (hierarchical graph contrastive learning, HGraph-CL)[27]. 文献 [27] 提出了一种新的层次图对比学习框架用于探索模态内部和模态之间的复杂关系, 增强情感特征提取能力.

3.3 超参数设置

为获取合适的超参数, 本文以验证集的结果为基础, 采用网格搜索的方式对超参数进行选择. 最终实验的超参数设置如表 2 所示.

表 2 对每个数据集进行实验的超参数设置

超参数	MOSI	MOSEI	SIMS	IEMOCAP
batch size	64	64	16	24
learning rate	1E–4	1E–5	1E–5	1E–4
hidden size	256	128	128	256
post dim	128	128	64	128
epoch	50	50	50	50
dropout	0.0	0.4	0.0	0.4
early stop	20	1	8	8

表 2 中, “batch size”是单次训练样本的个数; “learning rate”是神经网络的学习率; “hidden size”是单模态学习的隐藏层特征维度; “post dim”是多模态表示学习的隐藏层特征维度; “epoch”是总的训练次数; “dropout”是随机忽略部分神经元的概率; “early stop”是为防止过拟合而设置的提前结束循环条件.

3.4 B 融合方法的有效性分析

本文将提出的 B 融合方法加入到现有的情感分析模型 LF-DNN 和 LF-LSTM 中, 观察模型中单模态的学习梯度幅度变化与情感分析任务的结果, 判断本文提出方法对平衡多模态学习与提升多模态情感分析模型学习能力的有效性. 实验结果如图 4、表 3 所示. 其中, “(B)”表示加入 B 融合方法后的结果.

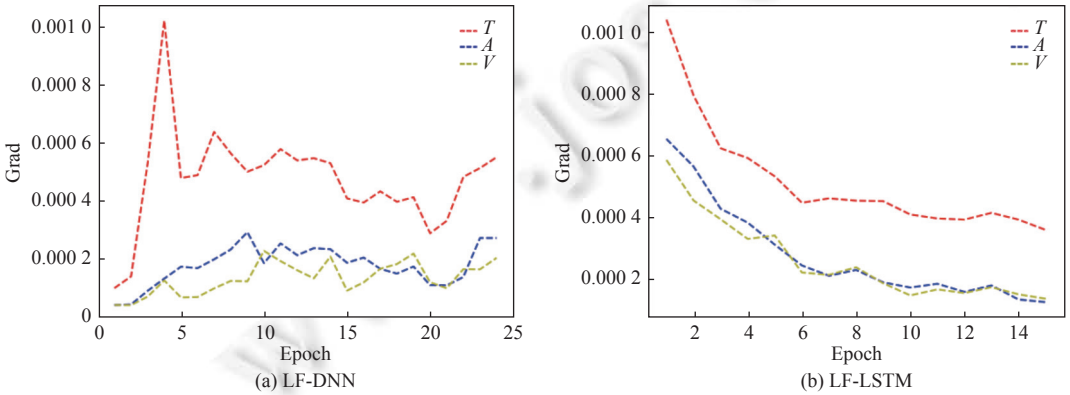


图 4 加入 B 融合方法后的单模态训练梯度幅度对比

表 3 加入 B 融合方法前后的模型测试结果对比

模型	多模态训练结果	多模态训练结果 (B)
LF-DNN	0.766 8	**0.794 2**
LF-LSTM	0.747 8	**0.774 1**
LF-DNN (BERT)	0.780 5	**0.806 1**
LF-LSTM (BERT)	0.779 9	**0.799 8**

图 4 描述了通过 MOSI 数据集验证两种现有情感分析模型 LF-DNN、LF-LSTM 加入 B 融合方法后的单模态梯度幅度变化. 与图 1 进行对比后可以看出, 加入 B 融合方法后, 训练过程中各模态的参数变化幅度更接近, 增大了音频模态和视频模态的学习梯度幅度, 有助于学习到更具表达力的音频特征表示和视频特征表示. 表 3 描述了多模态模型在加入 B 融合方法前后的二分类结果对比. 从表 3 中可以看出, 加入 B 融合方法后, 模型的分类结果有明显提高. 其中, BERT 表示文本模态采用了 BERT 预训练模型进行处理.

3.5 多模态情感分析对比实验

将本文提出的模型与现有的多模态情感分析模型在数据集 CMU-MOSI、CMU-MOSEI、SIMS、IEMOCAP 上进行对比, 验证本文模型在情感分析任务中的有效性. 实验结果如表 4、表 5 所示. 其中, “1”的数据来源于文献 [13], “2”的数据来源于文献 [15], “3”的数据来源于文献 [26], “4”的数据来源于文献 [27], “*”的数据结果是根据可获取的开源代码进行复现得出.

表 4 基于 CMU-MOSEI、CMU-MOSI、SIMS 数据集的模型对比结果 (%)

模型	CMU-MOSI				CMU-MOSEI				SIMS	
	二分类	*F*1	MAE	Corr	二分类	*F*1	MAE	Corr	二分类	*F*1
TFN[1,2]	80.8	80.7	90.1	69.8	82.5	82.1	59.3	70.0	79.86	80.15
LMF[1,2]	82.5	82.4	91.7	69.5	82.0	82.1	62.3	67.7	79.37	78.65
MulT[1]	83.4	82.8	87.1	69.8	82.5	82.3	58.0	70.3	—	—
MISA[1]	83.4	83.6	78.3	76.1	84.23	83.97	56.8	72.4	—	—
MAG-BERT[2]	84.3	84.3	73.1	78.9	85.23	85.08	53.9	75.3	—	—
Self-MM[2]	85.98	85.95	71.3	79.8	85.17	85.30	53.0	76.5	80.74	80.78
HGraph-CL[4]	86.2	86.2	71.7	**79.9**	85.9	85.8	**52.7**	76.9	—	—
MAG-BERT*	83.54	83.57	74.91	78.33	85.16	85.02	54.34	76.02	—	—
Self-MM*	84.03	84.05	72.41	79.09	85.22	85.07	53.89	76.63	79.21	78.87
本文模型	**86.28**	**86.38**	**71.63**	79.1	**86.38**	**86.46**	53.49	**77.02**	**80.96**	**81.28**

表 5 基于 IEMOCAP 数据集的模型对比结果 (%)

模型	Happy		Sad		Angry		Neutral	
	Acc	*F*1	Acc	*F*1	Acc	*F*1	Acc	*F*1
MulT[3]	86.4	82.9	82.3	82.4	85.3	85.8	71.2	70.0
Fusion-Based-CM-Attn-MulT[3]	85.6	83.7	83.6	83.7	84.6	85.0	70.4	69.9
LMF-MulT[3]	85.3	84.1	84.1	83.4	85.7	86.2	71.2	70.8
本文模型	**86.4**	**84.2**	**85.2**	**84.7**	**86.3**	**86.7**	**71.3**	**71.4**

在表 4 体现的情感分析实验结果中, 本文模型在 CMU-MOSI、CMU-MOSEI、SIMS 这 3 个数据集中的结果均优于对比模型. 在 SIMS 数据集的对比实验中, 相比于多任务分析模型 Self-MM 的测试结果, 本文模型采用单任务处理方法获得了更优异的任务分析结果, 说明了本文提出的情感分析方法在减少网络训练分支的情况下, 能获取更具表现力的多模态情感特征表示, 提升情感分析任务的分析结果. 相较于 HGraph-CL 模型, 本文模型在 MOSI 数据集中的皮尔逊相关指标和 MOSEI 数据集中的平均绝对误差指标略低, 原因是图对比学习相比于监督学习更有利于挖掘数据内部和数据之间的关系, 这将会在我们的未来研究中进行探讨.

在表 5 的情绪分析实验结果中, 本文模型在“Happy”和“Sad”两种情绪下的分类结果相比于其他模型的实验结果提升较多, 而在“Neutral”情绪下的分类结果相比于其他模型的实验结果提升较少.

为了直观地体现 B 融合在实验过程中对融合权值的自适应调整过程, 本文采用百分比堆积条形图对 3 种模态在每个轮次训练中的 B 融合权值变化进行展现, 具体如图 5 所示. 由图 5 可以看出, 从第 1 个训练轮次到第 7 个训练轮次, 文本模态的 B 融合权值在不断减小, 而其余两种模态的 B 融合权值在不断增加, 说明文本模态在训练初期的学习速率较高, 而其余模态的训练速率较低, B 融合权值的减小和增大能有效平衡各模态之间的学习速率, 促使单模态特征提取网络学习到更具表达力的特征表示.

为体现本文模型在任务分析中的效率, 将本文模型与现有的多模态情感分析模型进行运算时间的对比. 对比结果如表 6 所示.

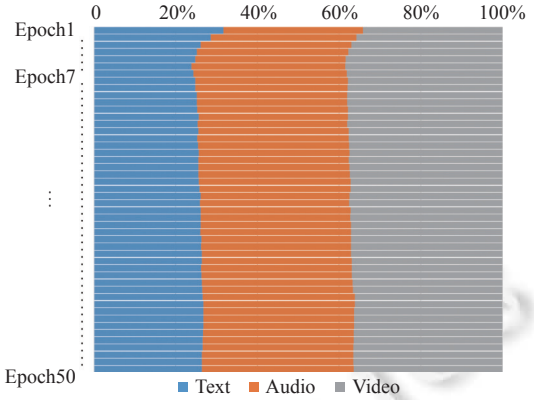


图 5 B 融合权值的变化图

表 6 单个循环时间 (s)

模型	CMU-MOSI	CMU-MOSEI	IEMOCAP
MulT	9.91	221.12	43.33
LMF-MulT	11.01	137.35	23.53
Self-MM	7.45	75.99	—
本文模型	**6.78**	**65.33**	**12.76**

本文提出的模型没有使用复杂的多模态融合方法, 仅采用简单的特征拼接方法进行融合. 从表 6 中可以看出, 本文提出的模型相比于 MulT 等多模态复杂融合模型有更少的训练时间, 证明了本文模型在任务分析中的高效性.

3.6 消融实验分析

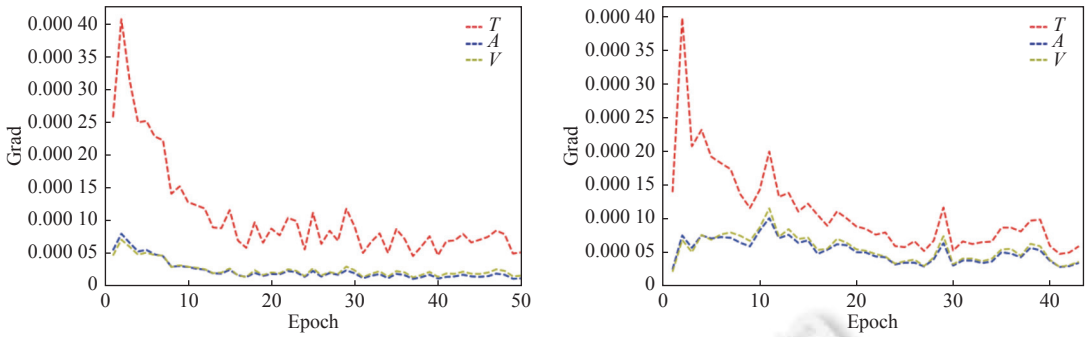
为探究基于不同优化器进行学习的效果, 本文分别利用随机梯度下降优化器 (SGD) 和自适应梯度优化器 (Adam) 的方法进行学习, 判断本文方法是否能基于不同的优化器都获得良好效果. 实验结果如表 7 所示, 其中, (GB) 表示使用了梯度融合方法.

表 7 消融分析结果 (%)

方法	MOSI		MOSEI		SIMS	
	二分类	*F*1	二分类	*F*1	二分类	*F*1
SGD	84.45	84.48	84.84	84.92	79.65	79.48
SGD (GB)	**86.28**	**86.38**	**86.38**	**86.46**	**80.96**	**81.28**
Adam	84.30	84.34	84.56	84.52	79.43	79.21
Adam (GB)	**85.06**	**85.1**	**85.14**	**85.3**	**79.87**	**80.95**

从表 7 中可以看出, 针对不同的优化器, 本文方法都能获得良好的提升效果. 在以随机梯度下降法作为优化器的实验中, 本文方法能获得最优的提升效果, 而在使用自适应梯度优化器对于实验结果的提升较小. 此外, 本文以 MOSI 数据集为基础, 对比 3 种模态在模型使用 B 融合方法前后的学习梯度幅度变化. 对比结果如图 5 所示. 其中, 纵坐标表示模态的学习梯度幅度, 横坐标表示模型的训练轮次.

图 6(a) 描述了使用 B 融合方法之前 3 种模态学习的梯度幅度变化, 图 6(b) 描述了使用 B 融合方法后 3 种模态的梯度幅度变化. 从中可以看出, 未使用 B 融合方法时, 文本模态的学习梯度幅度远高于其余模态, 有较大的学习速率, 而音频模态和视频模态的学习梯度幅度很小, 最大幅度不超过 0.000 1, 在第 10 次训练后接近于 0. 使用 B 融合方法后, 减小了文本模态的学习速率, 并使其余模态能有效学习.



(a) 加入 B 融合方法前 (b) 加入 B 融合方法后

图 6 针对 B 融合方法的消融对比实验

为进一步体现本文方法对单模态和多模态任务结果的影响, 对梯度平衡方法使用前后的单模态和多模态任务结果进行比较, 结果如表 8 所示. 其中, "(–)GB"表示未使用梯度平衡方法. 单模态的实验结果是将单模态特征提取网络的输出用于情感分析得到. 从结果可以看出, 未使用梯度平衡方法时, 多模态任务的结果由文本模态主导, 音频和视频模态的任务准确率低, 而当使用了梯度平衡方法后, 增加了单模态的任务分析准确度, 多模态任务结果也因此得到提升.

我们针对 A 融合方法进行了消融分析, 判断 A 融合方法是否有利于提升情感分析任务的准确度. 结果如表 9 所示. 其中, "(–)A 融合"表示将模态注意力模块 (A 融合) 从模型中消除, 最终用于任务决策的多模态表示仅为 B 融合的输出结果.

表 8 基于 MOSEI 数据集的单模态和多模态任务结果对比 (%)

模态	(–)GB	本文模型
文本	84.76	85.03
音频	65.22	79.50
视频	67.10	77.79
多模态	84.84	86.38

表 9 A 融合方法的消融分析结果 (%)

方法	MOSI		MOSEI	
	二分类	*F*1	二分类	*F*1
(–)A融合	85.37	85.43	85.99	86.11
本文模型	**86.28**	**86.38**	**86.38**	**86.46**

从表 9 的结果可以看出, 在未使用 A 融合的情况下, 实验能获得良好的情感任务分类结果, 但未能达到最优. 在加入 A 融合方法后, 减少了 B 融合对任务决策的影响, 动态地调整单模态对任务分析的贡献, 有利于提升情感任务的分析结果. 此外, 为了体现 A 融合权重对情感分析任务带来的影响, 我们冻结 A 融合参数以外的其余参数, 通过观察 A 融合方法的权重变化, 分析不同的 A 融合权重对情感分析结果的影响. 结果如表 10 所示. 其中, 3 种模态的融合比例为归一化后的结果.

表 10 A 融合权重分析 (%)

A融合比例 (文本:音频:视频)	二分类	*F*1
(0:0:0)	85.99	86.11
(0.38:0.20:0.42)	86.05	86.18
(0.35:0.39:0.26)	85.86	86.00
(0.36:0.34:0.30)	86.38	86.46

从表 10 的实验结果中可以看出, 文本模态因能学习到更多的语义信息而始终被赋予较高的权重, 表明文本模态对任务的贡献较大, 当文本对任务的贡献处于最大时, 任务的分类结果达到最高. 在最优的结果中 (0.36:0.34:0.30), 3 种模态的 A 融合比例差距最小, 证明了本文方法能较好地平衡模态学习, 有效增强模态的情感语义表达能力.

4 结 论

为了能更有效地结合多种模态进行情感分析, 防止因模态间学习速率不平衡而导致模型的性能下降. 本文提

出一种基于自适应权值融合的多模态情感分析方法, 其中的自适应权值融合方法分为两个阶段. 第 1 个阶段为 B 融合, 该阶段利用不同单模态学习网络的梯度幅度差异对模态的 B 融合权值进行优化, 使模型能通过 B 融合权值自适应地调整单模态网络的学习梯度, 实现多模态学习平衡. 第 2 个阶段为 A 融合, 其设计的目的是避免 B 融合权值对情感分析任务的决策产生影响. A 融合方法通过注意力机制为单模态特征表示动态地分配权重, 每个模态的权重代表了该模态对任务分析的贡献. 最终的多模态特征表示为 A、B 两种融合结果的结合. 将 B 融合方法引入现有的多模态情感分析模型中, 结果证明了 B 融合方法能有效提升多模态情感分析模型的学习能力. 将本文模型与现有的多模态情感分析模型在情感分析和情绪分析两种任务下进行对比, 结果显示本文模型在运算时间少的基础上, 任务测试结果均优于对比模型, 验证了本文模型在情感分析任务中的有效性和高效性. 在未来的研究中, 我们将更深入地探究导致模态间学习不平衡的因素.

References:

[1] Wu F, Wang ZQ, Zhou XB, Zhou GD. Joint model for sentiment analysis and review quality detection with user and product representations. Ruan Jian Xue Bao/Journal of Software, 2020, 31(8): 2492−2507 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5895.htm [doi: 10.13328/j.cnki.jos.005895]

[2] Wang WY, Tran D, Feiszli M. What makes training multi-modal classification networks hard? In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12692–12702. [doi: 10.1109/CVPR42600.2020.01271]

[3] Sun Y, Mai S, Hu HF. Learning to balance the learning rates between various modalities via adaptive tracking factor. IEEE Signal Processing Letters, 2021, 28: 1650–1654. [doi: 10.1109/LSP.2021.3101421]

[4] Peng XK, Wei YK, Deng AD, Wang D, Hu D. Balanced multimodal learning via on-the-fly gradient modulation. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 8228–8237. [doi: 10.1109/CVPR52688.2022.00806]

[5] 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.

[6] 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]

[7] Ghosal D, Akhtar MS, Chauhan D, Poria S, Ekbal A, Bhattacharyya P. Contextual inter-modal attention for multi-modal sentiment analysis. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: ACL, 2018. 3454–3466. [doi: 10.18653/v1/D18-1382]

[8] Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Wu TJ, Zhang SH, Martin RR, Cheng MM, Hu SM. Attention mechanisms in computer vision: A survey. Computational Visual Media, 2022, 8(3): 331–368. [doi: 10.1007/s41095-022-0271-y]

[9] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423–443. [doi: 10.1109/TPAMI.2018.2798607]

[10] Liu Z, Shen Y, Lakshminarasimhan VB, Liang PP, Zadeh A, 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: ACL, 2018. 2247–2256. [doi: 10.18653/v1/P18-1209]

[11] 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. 691.

[12] 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]

[13] 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]

[14] 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]

[15] 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. Palo Alto: AAAI Press, 2021. 10790–10797. [doi: 10.1609/aaai.v35i12.17289]

[16] Wu Y, Lin ZJ, Zhao YY, Qin B, Zhu LN. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment

analysis. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. ACL, 2021. 4730–4738. [doi: 10.18653/v1/2021.findings-acl.417]

[17] Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholmsmässan: PMLR, 2018. 794–803.

[18] Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

[19] Long X, Gan C, De Melo G, Liu X, Li YD, Li F, Wen SL. Multimodal keyless attention fusion for video classification. In: Proc. of the 32nd Conf. on Artificial Intelligence. New Orleans: AAAI Press, 2018. 882.

[20] Li Q, Xiao F, Zhang ZY, Zhang F, Li YC. Video summarization based on spacial-temporal transform network. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3195–3209 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6621.htm [doi: 10.13328/j.cnki.jos.006621]

[21] 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]

[22] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[23] Zadeh A, 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]

[24] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42(4): 335–359. [doi: 10.1007/s10579-008-9076-6]

[25] 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]

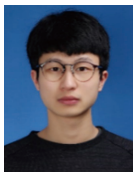
[26] Sahay S, Okur E, Kumar SH, Nachman L. Low rank fusion based transformers for multimodal sequences. In: Proc. of the 2nd Grand-challenge and Workshop on Multimodal Language (Challenge-HML). Seattle: ACL, 2020. 29–34. [doi: 10.18653/v1/2020.challengehml-1.4]

[27] Lin ZJ, Liang B, Long YF, Dang YX, Yang M, Zhang M, Xu RF. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: Int'l Committee on Computational Linguistics, 2022. 7124–7135.

附中文参考文献:

[1] 吴璠, 王中卿, 周夏冰, 周国栋. 基于用户和产品表示的情感分析和评论质量检测联合模型. 软件学报, 2020, 31(8): 2492–2507. http://www.jos.org.cn/1000-9825/5895.htm [doi: 10.13328/j.cnki.jos.005895]

[20] 李群, 肖甫, 张子屹, 张锋, 李延超. 基于空时变换网络的视频摘要生成. 软件学报, 2022, 33(9): 3195–3209. http://www.jos.org.cn/1000-9825/6621.htm [doi: 10.13328/j.cnki.jos.006621]



罗渊贻(1996－), 男, 博士生, 主要研究领域为多模态学习.

刘家锋(1968－), 男, 博士, 副教授, 主要研究领域为模式识别, 机器学习.

吴锐(1976－), 男, 博士, 副教授, 博士生导师, 主要研究领域为模式识别, 多模态学习.

唐降龙(1960－), 男, 博士, 教授, 博士生导师, 主要研究领域为模式识别, 计算机视觉.