

基于双重梯度调制的多模态情感识别

黄昕如,程 丽,陈日清,兰林毅

(福建农林大学 计算机与信息学院,福建 福州 350002)

摘要:针对多模态情感识别下模态数据固有的异质性限制模型全面感知情感信息,导致模态贡献度不平衡的问题,提出一种基于双重梯度幅度调制的多模态情感识别方法。该方法利用预定义的模态贡献比率动态监测各个模态的学习贡献,根据贡献差异适时调整梯度幅度,优化反向传播过程,减轻强势模态对弱势模态的抑制,促进模型全面捕捉各个模态中的关键情感信息。该方法在CREMA-D、CMU-MOSI和RAVDESS 3个公开情感识别数据集上的识别准确率分别达到74.05%、60.06%和46.31%,相较于表现最优的对照模型分别提升了4.01%、4.23%和3.17%,能够更好地完成多模态情感识别任务。

关键词:多模态;梯度调制;模态不平衡;情感识别;模态融合

DOI:10.11907/rjdk.241856

中图分类号:TP391

文献标识码:A

文章编号:1672-7800(2026)002-0052-06

扫描二维码阅读全文:



Multi-modal Emotion Recognition Based on Dual Gradient Modulation

HUANG Xinru, CHENG Li, CHEN Riqing, LAN Linyi

(College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China)

Abstract: A multimodal emotion recognition method based on dual gradient amplitude modulation is proposed to address the inherent heterogeneity of modal data in multimodal emotion recognition, which limits the model's ability to fully perceive emotional information and leads to an imbalance in modal contribution. This method utilizes predefined modal contribution ratios to dynamically monitor the learning contributions of each modality, adjusts the gradient amplitude in a timely manner based on the contribution differences, optimizes the backpropagation process, reduces the suppression of dominant modalities on weaker modalities, and promotes the model to comprehensively capture key emotional information in each modality. The recognition accuracy of this method on three publicly available emotion recognition datasets, CREMA-D, CMU-MOSI, and RAVDESS, reached 74.05%, 60.06%, and 46.31%, respectively. Compared with the best performing control model, it improved by 4.01%, 4.23%, and 3.17%, respectively, and can better complete multimodal emotion recognition tasks.

Key Words: multi-modal; gradient modulation; modal imbalance; emotion recognition; model fusion

0 引言

情感在人类交流中起着重要作用。多模态情感识别作为人机交互和智能分析的关键技术,受到了广泛关注^[1-3]。其旨在通过分析和融合来自不同模态(如文本、音频和视觉)的数据,全面理解和识别人类情感。这一技术在社交媒体、情感分析以及机器人等领域具有重要的应用

价值^[4-6]。然而,多模态数据的异质性导致模型在情感信息融合过程中面临挑战,特别是不同模态间的贡献度不平衡问题,可能会限制模型对情感复杂性的全面捕捉,影响识别的准确性。

在多模态情感识别领域,研究人员致力于通过整合来自不同模态(如文本、音频和视觉信息)的数据,实现对人类情感的全面识别与理解。早期研究中,Zadeh等^[7]提出张量融合网络,采用向量化策略融合多模态特征,有效处

收稿日期:2024-12-31

基金项目:国家自然科学基金项目(61972093);福建省高校产学研合作项目(2022H6006)

作者简介:黄昕如(1999-),女,CCF会员,福建农林大学计算机与信息学院硕士研究生,研究方向为计算机视觉与模式识别;程丽(1968-),女,硕士,福建农林大学计算机与信息学院副教授、硕士生导师,研究方向为计算机视觉与模式识别;陈日清(1979-),男,博士,福建农林大学计算机与信息学院教授、博士生导师,研究方向为多媒体安全和计算机视觉;兰林毅(2000-),男,福建农林大学计算机与信息学院硕士研究生,研究方向为计算机视觉与模式识别。本文通讯作者:程丽。

理了模态间的复杂交互。Tsai等^[8]提出多模态转换器,通过扩展Transformer模型的自注意力机制来强化跨模态学习,增强了单一模态的特征表示。然而,这些研究主要关注模态间的相关性,未能充分考虑模型训练过程中模态贡献度的不平衡问题。由于数据特性、信息量及模型表征能力的差异,模型可能过度依赖某一模态,导致在信息融合时忽视了其他模态中的关键情感信息,从而引发模态贡献度不平衡现象^[9]。为解决这一问题,Lv等^[10]在MULT(Multimodal Transformer)框架下提出一种渐进式强化学习策略,旨在通过逐步学习模态间的潜在特征表示来提升弱势模态的表现。Du等^[11]通过引入额外的单模态分类器或预训练模型来辅助多模态模型的训练,以增强其对弱势模态特征的学习。知识蒸馏技术也被用于提升弱势模态的性能,其通过将强势模态的知识迁移到弱势模态来提升模型的识别能力^[12]。还有研究引入原型聚类方法,通过将数据集中的样本划分为多个簇,以更好地捕捉弱势模态中的潜在结构^[13]。尽管现有方法通过加强弱势模态的学习速率来提升模型对情感信息的捕捉能力,但是通常需要更多的训练步骤来构建和优化神经网络的额外模块,可能会引入噪声干扰,进而影响模型的整体性能。

为此,本文提出一种基于双重梯度调制(Dual Gradient Modulation, DualGM)的多模态情感识别方法。该方法从优化过程的角度深入分析多模态情感识别中的模态贡献度不平衡问题,发现强势模态因对损失函数下降的显著贡献而获得较大梯度更新,而弱势模态则因模型依赖性强

而梯度更新受限。除了提升弱势模态的表现,还需要抑制强势模态而达到双重提升弱势模态的效果。为此,本文提出DualGM方法,根据预定义的模态贡献比率动态调整各模态的梯度,当强势模态的表征能力显著优于弱势模态时,通过降低强势模态的梯度幅度并提高弱势模态的梯度幅度,避免了强势模态因优化过快而导致弱势模态学习停滞,实现了多模态间的均衡优化。实验结果表明,DualGM在多模态情感识别数据集上显著提升了模型性能,有效缓解了模态贡献度不平衡现象。所提方法为解决模态贡献度不平衡问题提供了新的视角和解决方案,推动了多模态情感识别技术的发展与应用。

1 方法构建

本文首先对不平衡问题进行分析。通过实验发现,多模态情感识别中的一个模态可以抑制其他模态的梯度幅度,导致其模态表现较差。基于此,本文提出兼顾抑制强势模态与增强弱势模态的DualGM。DualGM主要由贡献监测模块(Monitoring Module)和贡献对齐模块(Alignment Module)两个模块组成。首先通过音频和图像编码器分别得到各自模态的特征,将特征转换为概率分布作为贡献监测模块的输入。贡献监测模块预定义了一个贡献比 ρ_i^u 来反映音频与图像数据之间的贡献度差距;贡献对齐模块预定义 λ_i^u 介入动态调整梯度,以增强弱势模态的梯度幅度,减弱强势模态的梯度幅度。在训练过程中,通过反向梯度传播调整网络的权重。DualGM的整体框架如图1所示。

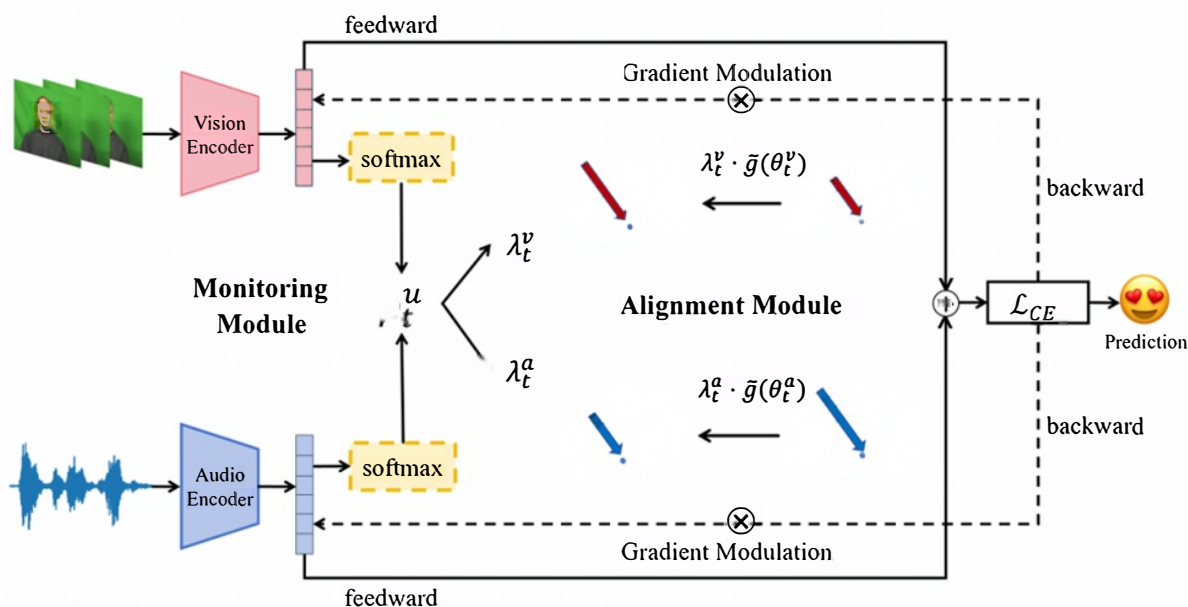


Fig. 1 Overall framework of DualGM

图1 DualGM整体框架

1.1 不平衡分析

以具有音频和图像两种模态的情感识别数据为例对不平衡现象进行分析。首先,将训练数据集表示为 $D =$

$\{x_i, y_i\}_{i=1,2,\dots,N}$, $i=1,2,\dots,N$, x_i 为不同模态的输入,即 $x_i = (x_i^a, x_i^v)$,其中 a 和 v 分别表示音频和图像模态。 $y_i = 1, 2, \dots, M$, M 为情感类别的个数。模型使用音频编码器

$\varphi_a(\theta_a, x_i^a)$ 和图像编码器 $\varphi_v(\theta_v, x_i^v)$ 提取特征, 分别记作 φ_a 和 φ_v , θ_a 和 θ_v 均为编码器的参数。为方便观察每个模态的优化过程, 本文使用最常见的拼接求和来获得多模态模型的输出。计算公式为:

$$f(x_i) = W^a \cdot \varphi^a(\theta^a, x_i^a) + W^v \cdot \varphi^v(\theta^v, x_i^v) + b \quad (1)$$

式中: $W^a \in \mathbb{R}^{M \times d_a}$, $W^v \in \mathbb{R}^{M \times d_v}$, $b \in \mathbb{R}^M$, 均为分类器的参数。

以 CREMA-D 数据集为例, 采用两个模态得到的准确率为 66.13%, 音频模态的准确率达到 57.27%, 图像模态的准确率仅达到 28.76%。音频模态的表现明显优于图像模态, 这意味着模型偏好音频模态的情感变化, 图像模态的情感信息可能被忽视。本文将 m 类的真值输出表示为 $f(x_i)_m$, 定义交叉熵损失函数为:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{f(x_i)_{y_i}}}{\sum_{k=1}^M e^{f(x_i)_k}} \quad (2)$$

利用梯度下降法 (Gradient Descent, GD) 找到最优参数, 以音频模态为例。计算过程为:

$$\begin{aligned} W_{t+1}^a &= W_t^a - \eta \nabla_{W^a} \mathcal{L}_{CE}(W_t^a) \\ &= W_t^a - \eta \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)} \varphi^a(\theta^a, x_i^a) \end{aligned} \quad (3)$$

$$\begin{aligned} \theta_{t+1}^a &= \theta_t^a - \eta \nabla_{\theta^a} \mathcal{L}_{CE}(\theta_t^a) \\ &= \theta_t^a - \eta \frac{1}{N} \sum_{i=1}^N \frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)} \frac{\partial (W_t^a \cdot \varphi_t^a(\theta_t^a, x_i^a))}{\partial \theta_t^a} \end{aligned} \quad (4)$$

式中: η 为学习率。采用同样的方法可以更新得到 W^v 和 θ^v 的公式。由式 (3) 和式 (4) 可知, 除了与训练损失 $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ 相关的项以外, W^a 和 θ^a 的优化不受其他模态优化的影响。各个模态编码器的参数更新通常是独立进行的, 这意味着它们在优化过程中缺乏跨模态的信息交流。根据式 (2), 本文将属于 m 类别的样本输出 $f(x_i)_m$ 的梯度 $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)}$ 重写为:

$$\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)_{y_i}} = \frac{e^{(W^a \cdot \varphi^a + b^a + W^v \cdot \varphi^v + b^v)_{y_i}}}{\sum_{k=1}^M e^{(W^a \cdot \varphi^a + b^a + W^v \cdot \varphi^v + b^v)_k}} - 1 \quad (5)$$

由该式可以推断, 对于音频模态来说, 当其表现出较好的性能时, 能通过 $W^a \cdot \varphi^a$ 对 $\frac{\partial \mathcal{L}_{CE}}{\partial f(x_i)_{y_i}}$ 提供较多的贡献, 从而使整体损失变小。因此, 相对于其他模态, 音频模态对模型的贡献最大, 占据多模态训练过程的主导地位。在反向传播过程中, 音频模态会获得较大的梯度更新, 同时限制了图像模态的表现。当多模态模型快速从音频模态中学习情感信息而趋于收敛时, 图像模态由于梯度更新不足而未能得到充分优化。因此, 如何加快弱势模态、减缓强势模态的学习速度是本文需要解决的关键问题。

1.2 梯度调制

基于以上章节对于模态贡献度不平衡的分析, 可以采取减缓强势模态收敛速度的策略, 使模型在训练过程中也能从弱势模态中获取可能重要的情感信息。DualGM 方法可根据实际需要抑制强势模态的梯度幅度, 亦可增强弱势模态的梯度幅度, 从而充分学习不同模态中的情感特征。以下详细阐述。

1.2.1 贡献监测模块

在使用 GD 方法时将编码器 φ^u 的参数 θ^u 更新为:

$$\theta_{t+1}^u = \theta_t^u - \eta \nabla_{\theta^u} \mathcal{L}_{CE}(\theta_t^u) \quad (6)$$

采用随机梯度下降优化方法 (Stochastic Gradient Descent, SGD), 将编码器 φ^u 的参数 θ^u 进一步更新为:

$$\theta_{t+1}^u = \theta_t^u - \eta \tilde{g}(\theta_t^u) \quad (7)$$

式中: $\tilde{g}(\theta_t^u) = \frac{1}{m} \sum_{x \in B_t} \nabla_{\theta^u} \mathcal{L}(x; \theta_t^u)$, 是对全梯度 $\nabla_{\theta^u} \mathcal{L}(\theta_t^u)$ 的无偏估计; B_t 为选择 t 步中大小为 m 的随机小批。本文定义贡献比 ρ_t^u , 用于监测不同模态对学习目标的贡献差异。表示为:

$$\rho_t^v = \frac{\sum_{i \in B_t} s_i^v}{\sum_{i \in B_t} s_i^a} \quad (8)$$

$$s_i^a = \sum_{k=1}^M \left(W_t^a \cdot \varphi_t^a(\theta_t^a, x_i^a) + \frac{b}{2} \right)_k \quad (9)$$

$$s_i^v = \sum_{k=1}^M \left(W_t^v \cdot \varphi_t^v(\theta_t^v, x_i^v) + \frac{b}{2} \right)_k \quad (10)$$

式中: s_i^v 和 s_i^a 分别表示图像模态和音频模态特征经过 softmax 层之后的概率输出。采用模态的近似预测 $\left(W_t^v \cdot \varphi_t^v(\theta_t^v, x_i^v) + \frac{b}{2} \right)$ 来估计模型中不同模态的单峰性能。当 $\rho_t^v > 1$ 时, 表示图像模态的贡献较大, 图像模态占主导地位; 反之, 当 $\rho_t^v < 1$ 时, 图像模态则是被抑制的一方。此外, ρ_t^a 为 ρ_t^v 的倒数。因此, 可以通过贡献比来动态监测视听模态之间的贡献差异。

1.2.2 贡献对齐模块

贡献比反映不同模态的贡献差异时, 贡献对齐模块可同步介入, 对梯度幅度进行适度调制。本文不仅需要提高弱势模态的学习速率, 而且还要减缓强势模态的学习速率。因此, 定义 λ_t^u 来动态调整贡献比。公式为:

$$\lambda_t^u = \begin{cases} 1 - \tanh(\alpha \cdot \rho_t^u), & \text{if } \rho_t^u > 1 \\ 1 + \text{sigmoid}\left(\frac{\alpha}{\rho_t^u}\right), & \text{other} \end{cases} \quad (11)$$

式中: α 为控制调制程度的超参数。将 λ_t^u 应用于式 (7) 中, 则迭代过程中对 θ_t^u 的更新为:

$$\theta_{t+1}^u = \theta_t^u - \eta \cdot \lambda_t^u \tilde{g}(\theta_t^u) \quad (12)$$

利用 λ_t^u 既可以避免模型对强势模态的过度依赖, 又可以深入挖掘并利用弱势模态中的情感信号, 确保所有模态

在训练过程中得到均衡的重视,从而使模型能够全面捕捉和理解情感信息。

2 实验方法与结果

2.1 实验环境与参数设置

2.1.1 实验环境

硬件环境为 Intel (R) Core (TM) i7-8700K CPU @ 3.70GHz,软件环境为 Ubuntu 22.04.2 操作系统,采用 NVIDIA GeForce RTX 2080Ti GPU,内存为 16 GB。实验均在 PyTorch1.12.1 的深度学习框架下进行。

2.1.2 参数设置

对于音频模态,输入的音频数据经过转换得到 257×1 003 的频谱图。视觉模态则从视频片段中每间隔 1 s 抽取一帧图像作为输入数据。视觉模态和音频模态均使用 ResNet18 作为主干网络,将输入数据映射为 512 维的特征向量^[14]。实验中所有批次大小 batch size 为 16,均使用 SGD 优化器训练多模态模型,模型训练轮次 epoch 设置为 100^[15]。为了平衡模型的学习速度、防止过拟合现象,本文通过大量实验,将动量设置为 0.90,权重衰减为 1e-4,学习率初始化为 1e-3,并根据具体的数据集确定超参数。

2.2 数据集

为验证本文方法的有效性,在常用的多模态情感数据集 CREMA-D、CMU-MOSI 和 RAVDESS 上进行实验。其中,CREMA-D 数据集包含音频和图像模态,具有面部和声音的情感表达,训练集由 6 698 个样本组成,测试集由 744 个样本组成,包含快乐、悲伤、愤怒、恐惧、厌恶和中性 6 个标签^[16]。CMU-MOSI 数据集包含 2 199 个独白短视频片段,以 12.5 Hz 和 15 Hz 的采样率提取音频和图像特征,用于训练和测试的样本分别为 1 284 个和 686 个,包含消极、中性、积极 3 个标签^[17]。RAVDESS 数据集包含纯音频、音频—视频和纯视频 3 种形式的 7 356 个文件,共有平静、快乐、悲伤、恐惧、愤怒、惊讶、厌恶 7 种情感标签,本文使用音频—视频数据^[18]。

2.3 评价指标

为验证模型的有效性,利用测试数据集进行评估,使用准确率 (Accuracy) 作为评价指标。其是指所有预测正确的正例和负例占情感样本总量的比例。计算公式为:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (13)$$

式中:TP、TN 分别为正类样本和负类样本预测正确的数量;FP、FN 分别为正类样本和负类样本预测错误的数量。

2.4 实验结果与分析

2.4.1 与其他缓解不平衡方法的性能比较

将本文方法与 OGM-GE^[19]、AGM^[20] 等梯度调节方法进行比较,结果见表 1。基线模型是对两个模态的简单拼

接,不采用任何调制策略。可以看出,本文方法在 CREMA-D、CMU-MOSI 和 RAVDESS 数据集上均得到最优结果,分别为 74.05%、60.06% 和 46.31%,相较基线方法分别提升了 8.92%、4.81% 和 5.28%,相较表现最佳的对照方法分别提升了 4.01%、4.23% 和 3.17%。

Table 1 Performance comparison with other methods for mitigating imbalance

表 1 与其他缓解不平衡方法的性能比较 (%)			
方法	CREMA-D	CMU-MOSI	RAVDESS
基线模型	66.13	55.25	41.03
OGM-GE	67.88	55.83	41.51
AGM	70.04	55.75	43.14
本文方法	74.05	60.06	46.31

2.4.2 不平衡的缓解效果

以 RAVDESS 数据集为例,对其训练得到的音频和图像模态性能进行比较,结果如表 2 所示。相较于其他缓解不平衡的情感识别方法,本文方法有效减小了模态之间的差距,弱势模态的性能得到了提升,有助于缓解模态贡献度不平衡的问题。

Table 2 The alleviation effect of various methods on the imbalance of modal contribution

表 2 各方法对模态贡献度不平衡的缓解效果 (%)				
方法	融合后的准确率	音频单模态准确率	图像单模态准确率	音频单模态与图像单模态准确率差值
基线模型	41.03	40.71	18.11	22.60
OGM-GE	41.51	39.42	17.47	21.95
AGM	43.14	39.04	16.70	22.34
本文方法	46.31	39.90	23.24	16.66

2.4.3 与基于 CNN 与 Transformer 的情感识别方法的比较

在 CREMA-D 数据集上,将本文方法与一些代表性情感识别方法进行比较,包括基于 CNN 的 PSP^[21]、AVEL^[22]、CMRAN^[23]、MM-Pyramid^[24] 和基于 Transformer 的 MBT^[25]、LAVISH^[26]、MA-AVT^[27]。结果见表 3。现有基于 CNN 和 Transformer 的情感识别方法的性能依赖于对大规模音频和图像数据的预训练和庞大的网络架构,而本文方法采用轻量级 ResNet18 作为编码器,参数量仅为 22.35 M,比对照方法减少 5~10 倍,不仅大大减少了对计算资源的需求,而且保持了较高的准确率。

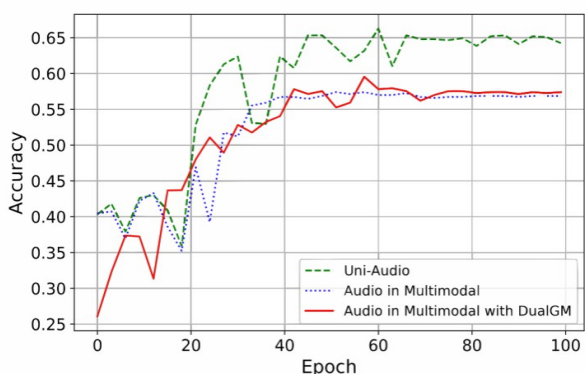
Table 3 Comparison with emotion recognition methods based on CNN and Transformer

表 3 与基于 CNN 与 Transformer 的情感识别方法的比较				
方法	图像编码器	音频编码器	整体参数量/M	准确率/%
PSP	VGG-19	VGGish	231.5	62.5
AVEL	ResNet-152	VGGish	136.0	59.7
CMRAN	ResNet-152	VGGish	148.2	61.8
MM-Pyramid	ResNet-152	VGGish	176.3	63.1
MBT	ViT-B-16	ViT-B-16	206.4	70.7
LAVISH	ViT-B-16	ViT-B-16	110.4	69.7
MA-AVT	ViT-B-16	ViT-B-16	110.2	72.3
本文方法	ResNet18	ResNet18	22.4	74.1

2.5 消融实验

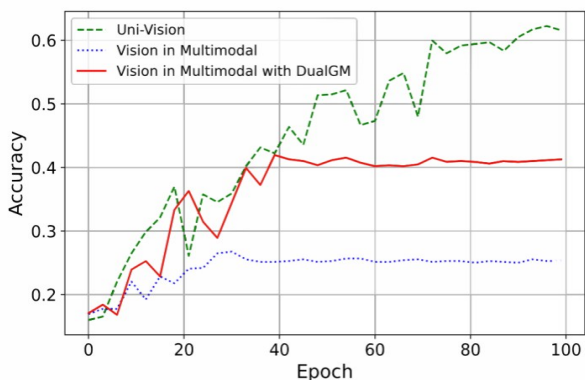
2.5.1 单模态性能比较

模态之间的相互作用会导致每个模态对模型有着不同的贡献,并且每种模态都不能充分利用自身的特性。因此,本文比较使用和不使用 DualGM 时多模态模型的单峰和单峰分支表现。如图 2 所示,在没有 DualGM 的多模态模型中,音频分支的学习效果优于图像分支,图像分支与单一图像模态相比有明显差距。在应用 DualGM 后,虽然模型中有抑制音频分支的措施,但音频分支的性能仍然优于基线模型,没有受到显著影响,表明在促进弱势模态作出贡献的同时相对有效地保护了强势模态,充分发挥了不同模态在情感识别中的作用。



(a) Performance of uni-audio, single-modal branch representation with or without DualGM

(a) 音频模态的单峰模型、有无 DualGM 的单模态分支表现



(b) Performance of uni-vision, single-modal branch representation with or without DualGM

(b) 图像模态的单峰模型、有无 DualGM 的单模态分支表现

Fig. 2 Performance of unimodal models and unimodal branches

图 2 单模态模型和单峰分支性能表现

2.5.2 与传统融合方法的比较

常见的多模态融合策略包括拼接和相加^[28-29]。本文将 DualGM 方法集成到上述融合技术中,实验结果见表 4,符号†表示使用 DualGM 方法。可以看出,与 DualGM 结合的方法在不同数据集上均表现出更好的性能,证明了 DualGM 在各种场景下的有效性和适用性。

Table 4 Comparison with traditional fusion methods

方法	表 4 传统融合方法的比较 (%)	
	CREMA-D 准确率	CMU-MOSI 准确率
拼接融合	66.12	55.25
拼接融合†	74.05	58.46
相加融合	66.53	59.18
相加融合†	71.37	60.06

3 结语

本文提出一种创新性的 DualGM 方法来解决多模态情感识别中的模态贡献不平衡问题,通过预定义的模态贡献比率动态调整梯度幅度,成功提升了弱势模态的学习效率,同时避免了强势模态的过度影响。实验结果表明, DualGM 在多个数据集上均取得显著的性能提升。然而,该方法在超参数调整和计算复杂性方面仍存在挑战。未来计划进一步优化算法,减少对超参数的依赖,并探索其在实时系统中的应用潜力。此外,还计划研究 DualGM 在更多样化数据集上的泛化能力,以期在智能客服、健康监测等领域实现更广泛的应用。

参考文献:

- [1] LIU F Y, YU W Q. A text sentiment analysis model combining BERT and attention [J]. Software Guide, 2023, 22(8): 66-71.
刘斐瑜,俞卫琴. 融合 BERT 与注意力的文本情感分析模型[J]. 软件导刊, 2023, 22(8): 66-71.
- [2] SHI A W, CAI R. Multi-modal emotion recognition method combining multiple attention mechanisms [J]. Software Guide, 2023, 22(10): 105-109.
史爱武,蔡润. 结合多种注意力机制的多模态情感识别方法[J]. 软件导刊, 2023, 22(10): 105-109.
- [3] HUO Y. Research on multi-task expression recognition in multi-dimensional continuous space[J]. Software Guide, 2024, 23(5): 17-23.
霍奕. 多维连续空间的多任务表情识别研究[J]. 软件导刊, 2024, 23(5): 17-23.
- [4] SHEN Z Q. A cross-modal social media sentiment analysis method based on image-text fusion[J]. Software Guide, 2019, 18(1): 9-13, 16.
申自强. 一种基于图文融合的跨模态社交媒体情感分析方法[J]. 软件导刊, 2019, 18(1): 9-13, 16.
- [5] LI H L, LIANG Z Q, HUANG X J, et al. A multimodal psychological evaluation algorithm for the power grid industry [J]. Software Guide, 2024, 23(9): 90-98.
李华亮,梁泽权,黄星杰,等. 面向电网行业的多模态心理评测算算法[J]. 软件导刊, 2024, 23(9): 90-98.
- [6] LIU Z, WU M, CAO W, et al. A facial expression emotion recognition based human-robot interaction system[J]. IEEE/CAA Journal of Automatica Sinica, 2017, 4(4): 668-676.
- [7] ZADE A, CHEN M, PORI S, et al. Tensor fusion network for multimodal sentiment analysis[DB/OL]. https://arxiv.org/abs/1707.07250.
- [8] TSAI Y H H, BAI S, LIA P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the 57th Annual Meet-

- ing of the Association for Computational Linguistics, 2019: 6558–6569.
- [9] WANG W, TRAN D, FEISZLI M. What makes training multi-modal classification networks hard? [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12695–12705.
- [10] LV F, CHEN X, HUA Y, et al. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2554–2562.
- [11] DU C, LI T, LIU Y. Improving multi-modal learning with uni-modal teachers [DB/OL]. <https://arxiv.org/abs/2106.11059>.
- [12] LI Y, WANG Y, CUI Z. Decoupled multimodal distilling for emotion recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 6631–6640.
- [13] FAN Y, XU W, WANG H, et al. Pmr: prototypical modal rebalance for multimodal learning [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 20029–20038.
- [14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
- [15] ROBB H, MONR S. A stochastic approximation method [J]. The Annals of Mathematical Statistics, 1951, 22(3): 400–407.
- [16] CAO H, COOP D G, KEUT M K, et al. Crema-d: crowd-sourced emotional multimodal actors dataset [J]. IEEE Transactions on Affective Computing, 2014, 5(4): 377–390.
- [17] ZADE A, ZELL R, PINC E, et al. Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages [J]. IEEE Intelligent Systems, 2016, 31(6): 82–88.
- [18] LIVING S R, RUSS F A. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English [J]. PloS One, 2018, 13(5): e0196391.
- [19] PENG X, WEI Y, DENG A, et al. Balanced multimodal learning via on-the-fly gradient modulation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8238–8247.
- [20] LI H, LI X, HU P, et al. Boosting multi-modal model performance with adaptive gradient modulation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 22214–22224.
- [21] ZHOU J, ZHEN L, ZHONG Y, et al. Positive sample propagation along the audio-visual event line [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 8436–8444.
- [22] TIAN Y, SHI J, LIB B, et al. Audio-visual event localization in unconstrained videos [C]//Proceedings of the European Conference on Computer Vision, 2018: 247–263.
- [23] XU H, ZENG R, WU Q, et al. Cross-modal relation-aware networks for audio-visual event localization [C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 3893–3901.
- [24] YU J, CHENG Y, ZHAO R W, et al. Mm-pyramid: multimodal pyramid attentional network for audio-visual event localization and video parsing [C]//Proceedings of the 30th ACM International Conference on Multimedia, 2022: 6241–6249.
- [25] NAGR A, YANG S, ARNAB A, et al. Attention bottlenecks for multi-modal fusion [J]. Advances in Neural Information Processing Systems, 2021, 34: 14200–14213.
- [26] LIN Y B, SUNG Y L, LEI J, et al. Vision transformers are parameter-efficient audio-visual learners [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 2299–2309.
- [27] MAHMUD T, MO S, TIAN Y, et al. MA-AVT: modality alignment for parameter-efficient audio-visual Transformers [DB/OL]. <https://arxiv.org/abs/2406.04930>.
- [28] OWEN A, EFRO A A. Audio-visual scene analysis with self-supervised multisensory features [C]//Proceedings of the European Conference on Computer Vision, 2018: 631–648.
- [29] KIEL D, GRAVE E, JOUL A, et al. Efficient large-scale multi-modal classification [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 5198–5204.

(责任编辑:尹晨茹)