智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于自适应梯度调制的音视频多模态平衡学习方法

王忠美, 敖文秀, 刘建华, 贾林, 张昌凡, 彭深奥, 刘金平

引用本文:
王忠美, 敖文秀, 刘建华, 等. 基于自适应梯度调制的音视频多模态平衡学习方法[J]. 智能系统学报, 2025, 20(5): 1217-1226.
WANG Zhongmei, AO Wenxiu, LIU Jianhua, et al. An audio-visual multimodal balanced learning method based on adaptive gradient modulation[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1217-1226.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202412009

DOI: 10.11992/tis.202412009
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250806.1453.006

基于自适应梯度调制的音视频多模态平衡学习方法

王忠美[1], 敖文秀[1], 刘建华[1], 贾林[1], 张昌凡[1], 彭深奥[1], 刘金平[2]
(1. 湖南工业大学 轨道交通学院, 湖南 株洲 412007; 2. 湖南师范大学 信息科学与工程学院, 湖南 长沙 410081)

摘 要: 针对音视频多模态学习中因异质学习速率导致单一模态主导模型学习过程, 抑制其他模态学习, 进而削弱多模态协同决策效果的问题, 提出一种基于自适应梯度调制的多模态平衡学习方法 (adaptive gradient modulation based compensation and regularization, AGM-CR)。首先, 根据模态间的学习梯度差异引入调制系数来自适应调整各模态的学习速率; 然后, 通过梯度均衡化策略, 将单个模态的梯度损失作为正则项融入总损失来约束模态间梯度差异, 进一步平衡各模态的学习过程; 最后, 实验结果表明在 CREMA-D 和 RAVDESS 数据集上, AGM-CR 将分类准确率分别提高了 2.5 和 3.3 百分点, 并在多次迭代中减小模型的梯度波动, 表现出更高的训练稳定性和收敛速度。与现有的平衡方法相比, AGM-CR 可即插即用, 更具灵活性和通用性。

关键词: 平衡学习; 多模态学习; 梯度调制; 自适应学习; 梯度均衡化; 学习速率; 音视频模态; 协同决策
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2025)05−1217−10

中文引用格式: 王忠美, 敖文秀, 刘建华, 等. 基于自适应梯度调制的音视频多模态平衡学习方法 [J]. 智能系统学报, 2025, 20(5): 1217–1226.
英文引用格式: WANG Zhongmei, AO Wenxiu, LIU Jianhua, et al. An audio-visual multimodal balanced learning method based on adaptive gradient modulation[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1217–1226.

An audio-visual multimodal balanced learning method based on adaptive gradient modulation

WANG Zhongmei[1], AO Wenxiu[1], LIU Jianhua[1], JIA Lin[1], ZHANG Changfan[1], PENG Shen'ao[1], LIU Jinping[2]
(1. School of Railway Transportation, Hunan University of Technology, Zhuzhou 412007, China; 2. College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China)

Abstract: To address the challenge in audio-visual multimodal learning, where differing learning rates across modalities cause one to dominate and suppress others, thereby weakening the multimodal collaborative decision-making process, a novel multimodal balanced learning method based on adaptive gradient modulation (AGM-CR) is proposed. This method employs modulation coefficients that dynamically adjust the learning rates of individual modalities according to their gradient variations. Additionally, it incorporates a gradient balancing strategy that integrates modality-specific gradient losses into the total loss as a regularization term. Together, these mechanisms reduce gradient disparities, fostering a more balanced and effective learning process. Experimental evaluation on the CREMA-D and RAVDESS datasets demonstrates that AGM-CR improves classification accuracy by 2.5 and 3.3 percentage points, respectively. Furthermore, AGM-CR stabilizes training by minimizing gradient fluctuations across iterations, which accelerates convergence. Importantly, AGM-CR functions as a plug-and-play approach, enhancing flexibility and generalizability compared with existing balancing approaches.

Keywords: balanced learning; multimodal learning; gradient modulation; adaptive learning; multimodal gradient balancing; learning rate; audio-visual multimodal; collaborative decision-making

多模态学习通过关联和整合不同模态的数据, 提供更丰富、全面的信息理解和处理能力, 提升模型的整体性能, 在多模态情感分析[1-2]、自动驾驶[3-4]、医学图像分析等多个领域得到广泛应用。

收稿日期: 2024−12−11. **网络出版日期**: 2025−08−06.
基金项目: 国家重点研发计划项目 (2021YFF0501101); 国家自然科学基金项目 (52272347); 国家自然科学基金青年基金项目 (62106074).
通信作者: 王忠美. E-mail: wangzhongmei@hut.edu.cn.

理论上，多模态接收到更多的信息，其性能应优于单模态[5]。然而最近的研究发现[6-8]，基于同一优化目标进行联合训练的多模态模型往往不如最优的单模态模型。这是因为不同模态具有异质学习速率，使用统一的学习速率和单一的优化策略训练所有的模态，会导致学习速率较快的模态先行拟合并过早主导学习过程，从而抑制其他模态的学习。此时，学习速率较慢的模态仍处于欠拟合状态，导致对应的单模态学习网络无法得到充分学习，最终造成多模态网络整体性能退化。

现有的解决多模态不平衡问题的方法可归纳为基于学习目标、基于优化策略和基于模型架构的平衡方法[9]。基于学习目标的平衡方法主要体现在对目标函数优化设计上，通过在常规联合学习损失函数之外引入额外的单模态损失，利用附加损失来打破模态间不平衡[10-12]，可有针对性地对某个特定模态进行优化，尤其适用于在联合学习中某个模态信息较为稀疏或噪声较大的情况，但由于额外损失的加入，可能需要手动调整权重，难以适应不同任务的变化。不同于直接引入单模态损失的基于学习目标的方法，基于优化策略的平衡方法采用模态特征间的关系作为判断依据对单模态学习速率进行调整[13-15]。其优势在于可以自适应地调整学习率，减少了人为干预，适用于多模态数据特征规模差异较大的情况，但如何精确度量和调整模态间的特征关系，确保优化策略的有效和稳定成为该方法面临的挑战。基于模型架构的平衡方法从模型体系结构出发，考虑各学习过程的差异，通过提出新的网络框架来改善联合学习的效果[16-18]，具有灵活性和可拓展性。通过对模型架构的创新，为特定的多模态任务量身定制合适的架构，适用于具有高结构化需求或特定任务要求的场景，但也因此通常伴随着较高计算开销和实现复杂度。上述方法从不同层面考虑了如何平衡不同模态的学习速率，但是没有充分考虑到学习过程中梯度对模态学习速率的影响。

为了更好地平衡模态间的学习速率，提升模型的整体性能，本文提出一种基于自适应梯度调制的平衡学习方法，以动态控制每个模态的学习过程。首先获取每个模态的梯度来计算调制系数，动态地平衡单模态网络的学习速率。其次在反向传播过程中，将单个模态的梯度损失融入总损失中，逐渐减小学习梯度的幅度差异，从而提高模型稳定性和泛化能力。

本文的主要贡献如下：

1) 提出一种自适应调制各模态学习速率的方法，解决了多模态联合训练中异质学习速率导致的某些模态无法充分学习的问题。

2) 提出模态梯度均衡化策略，将单个模态的梯度损失作为正则项融入总损失，平衡模态间梯度幅度差异，增强了模型的泛化能力和稳定性。

1　相关工作

1.1　多模态学习

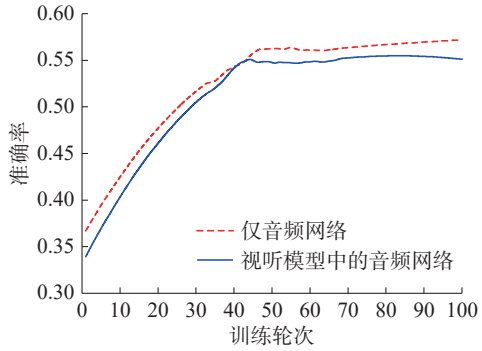
多模态学习是指通过联合利用来自不同模态的信息进行学习和预测，也可理解为对多源异构数据的挖掘分析，旨在实现对多模态信息的融合，弥补单一模态信息获取的局限性，从而提高模型的预测准确性和鲁棒性，在多个领域得到应用。如情感分析[19]中通过整合视觉、音频和文本信息来理解人类情感表达；在自动驾驶[20]中通过整合摄像头提供的 RGB 图像和激光雷达提供的深度信息来提高自动驾驶系统的感知和决策能力；在医学图像分析[21]中联合磁共振图像和正电子发射断层扫描图像，确保融合图像的高对比度和清晰度，有助于做出更全面的诊断。通过在这些下游任务中有效融合多模态信息，模型能够生成更为丰富的特征表示，显著提升在各类复杂任务中的表现。然而，现有的多模态学习方法仍面临着如何有效平衡不同模态间的学习速率、减小模态间异质性带来的影响等挑战，这也成为本文要解决的核心问题。

1.2　多模态平衡学习

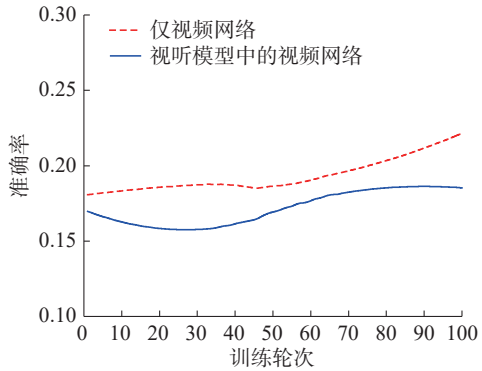
理论上多模态模型接受更多的信息，其性能应优于单模态学习模型。然而近期研究发现，模态间学习速率的差异会导致多模态模型性能的退化，甚至表现不如最优的单模态模型。

为了直观地展现多模态网络性能的下降情况，本文基于 CREMA-D 多模态情感分析数据集，分别训练单模态网络 (仅使用音频或仅使用视频) 和多模态联合训练网络，并将多模态网络中的单模态分支与独立训练的单模态网络进行准确率对比。图 1 给出了音频模态 (图 1(a)) 和视频模态 (图 1(b)) 的准确率对比结果。可以观察到：单模态模型对任务的分析准确率高于从多模态模型中分离出的相应模态；无论是在单模态模型还是多模态模型中，音频模态的准确率均显著高于视频模态，表明模型在情感分析任务中对音频信息偏好更强；在多模态联合训练中，视频模态准确率在初期出现下降，反映出音频模态在多模态学习中占主导作用，抑制了视频模态的学习。综上

分析表明, 在多模态学习中, 存在某些模态主导学习过程, 抑制其他模态学习的现象, 最终削弱了多模态协同决策的效果。



(a) 音频模态准确率对比



(b) 视频模态准确率对比

图 1　单模态模型和多模态模型中单模态分支的准确率对比

Fig. 1　Comparison among the unimodal branching accuracies of unimodal and multimodal models

针对上述问题, Wang 等[5] 提出梯度混合的方法, 将单模态模型的目标损失融合至多模态损失, 根据模态间损失下降的速度差异来寻找最优的权重组合。Peng 等[13] 在训练过程中动态监测不同模态对学习目标的贡献差异, 并基于该差异自适应调整学习速率, 为欠优化模态提供更多学习机会。Sun 等[22] 采用额外的单模态损失来辅助多模态模型训练, 并通过将较低的学习率分配给更接近收敛的模态来获得各模态特定的优化算法。Xiao 等[23] 提出随机丢弃收敛速度较快的模态网络通道作为一种正则化技术, 从而减缓相应模态的训练速度, 保证与其他模态的学习动态更兼容。尽管上述方法从模型架构、学习任务和优化策略等不同层面来解决多模态学习速率不平衡问题, 但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。

2　多模态梯度调制自适应平衡学习

2.1　多模态学习不平衡问题分析

现有解决不平衡问题的方法通常未考虑模型训练过程中参数梯度对模态学习速率的影响。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共同决定了学习速率的大小[24]。由深度学习模型的贪婪特性[25] 和对不平衡问题分析发现, 快速学习的模态参数更新频率更高, 往往会主导模型的决策, 而慢速学习的模态信息可能被忽略或利用不足, 从而影响最终的预测性能。

具体来说, 给定训练数据集 $\boldsymbol{D}=\{\boldsymbol{x}_i,y_i\}_{i=1,2,\cdots,N}$, $\boldsymbol{x}_i$ 和 y_i 分别是训练样本和标签。第 i 个训练样本 $\boldsymbol{x}_i=(\boldsymbol{x}_i^{\mathrm{a}},\boldsymbol{x}_i^{\mathrm{v}})$, 其中 $\boldsymbol{x}_i^{\mathrm{a}}$ 和 $\boldsymbol{x}_i^{\mathrm{v}}$ 分别为音频和视频模态的特征向量, $y_i\in\{1,2,\cdots,M\}$, M 为类别数。针对训练样本 $(\boldsymbol{x}_i,y_i)$, 首先, 使用 Resnet18 的骨干网络 $\boldsymbol{\varphi}^{\mathrm{a}}(\boldsymbol{\theta}^{\mathrm{a}},\cdot)$ 和 $\boldsymbol{\varphi}^{\mathrm{v}}(\boldsymbol{\theta}^{\mathrm{v}},\cdot)$ 分别提取音频和视频特征, $\boldsymbol{\theta}^{\mathrm{a}}$ 和 $\boldsymbol{\theta}^{\mathrm{v}}$ 分别为音频编码器和视频编码器可更新参数。接着, 通过融合音频特征和视频特征, 得到多模态模型的表达式:

$$f(\boldsymbol{x}_i)=\boldsymbol{W}^{\mathrm{a}}\cdot\boldsymbol{\varphi}^{\mathrm{a}}(\boldsymbol{\theta}^{\mathrm{a}},\boldsymbol{x}_i^{\mathrm{a}})+\boldsymbol{W}^{\mathrm{v}}\cdot\boldsymbol{\varphi}^{\mathrm{v}}(\boldsymbol{\theta}^{\mathrm{v}},\boldsymbol{x}_i^{\mathrm{v}})+\boldsymbol{b}$$

式中: $f(\boldsymbol{x}_i)$ 是对样本 $\boldsymbol{x}_i$ 的最终预测输出, $\boldsymbol{W}^{\mathrm{a}}$ 和 $\boldsymbol{W}^{\mathrm{v}}$ 分别为音频和视频特征的权重矩阵, $\boldsymbol{b}$ 为偏置项。分类任务中的交叉熵损失可表示为

$$L=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f(\boldsymbol{x}_i)_{y_i})}{\sum\limits_{k=1}^{M}\exp(f(\boldsymbol{x}_i)_k)}$$

式中: N 为样本数量, $f(\boldsymbol{x}_i)_{y_i}$ 表示模型对第 i 个样本的真实类别 y_i 的预测得分, $f(\boldsymbol{x}_i)_k$ 表示模型对第 k 个类别的预测得分。通过对交叉熵损失求导可以得到

$$\frac{\partial L}{\partial f(\boldsymbol{x}_i)_c}=\frac{\exp(\boldsymbol{W}^{\mathrm{a}}\cdot\boldsymbol{\varphi}^{\mathrm{a}}+\boldsymbol{W}^{\mathrm{v}}\cdot\boldsymbol{\varphi}^{\mathrm{v}}+\boldsymbol{b})_c}{\sum\limits_{k=1}^{M}\exp(\boldsymbol{W}^{\mathrm{a}}\cdot\boldsymbol{\varphi}^{\mathrm{a}}+\boldsymbol{W}^{\mathrm{v}}\cdot\boldsymbol{\varphi}^{\mathrm{v}}+\boldsymbol{b})_k}-1_{c=y_i}$$

式中: 分式是模型在类别 c 上的预测概率, 分子为类别 c 的指数得分, 分母为所有类别的预测得分的指数和, $1_{c=y_i}$ 表示当类别 c 是真实类别 y_i 时取值为 1, 否则为 0。由于 $1_{c=y_i}$ 的取值为 1 或 0 的特性, 模型在真实类别上的输出为预测概率减 1, 而在其他类别上的输出为模型的预测概率。由上式可知当某一模态表现足够好时, 其特征提取和权重在预测中将占据主导地位, 梯度更新将主要由该模态的输出决定, 使得模型会忽视其他模态的优化学习。因此, 为了平衡模态速率, 需充分考虑模型训练过程中不同模态的梯度贡献, 以避免某一模态主导梯度更新, 导致其他模态优化学习不足。

2.2　多模态学习速率自适应调制

为解决联合训练中模态异质学习速率导致的网络性能退化问题, 提出基于自适应梯度调制的

平衡方法 (adaptive gradient modulation based compensation and regularization, AGM-CR), 整体框架如图 2 所示。首先, 对原始音频和视频数据进行预处理, 分别输入到各自的编码器中提取特征, 并进行交互融合得到交叉熵损失; 然后, 在反向传播过程计算音频和视频模态的梯度, 并根据各模态梯度的比值计算调制系数, 利用调制系数动态调整各模态的学习速率; 最后, 将各模态梯度与平均梯度之间的距离作差得到的模态梯度损失作为正则项与交叉熵损失加权, 构建新的损失函数, 利用该重构的损失进行反向传播, 以此循环迭代, 逐步优化模型性能。

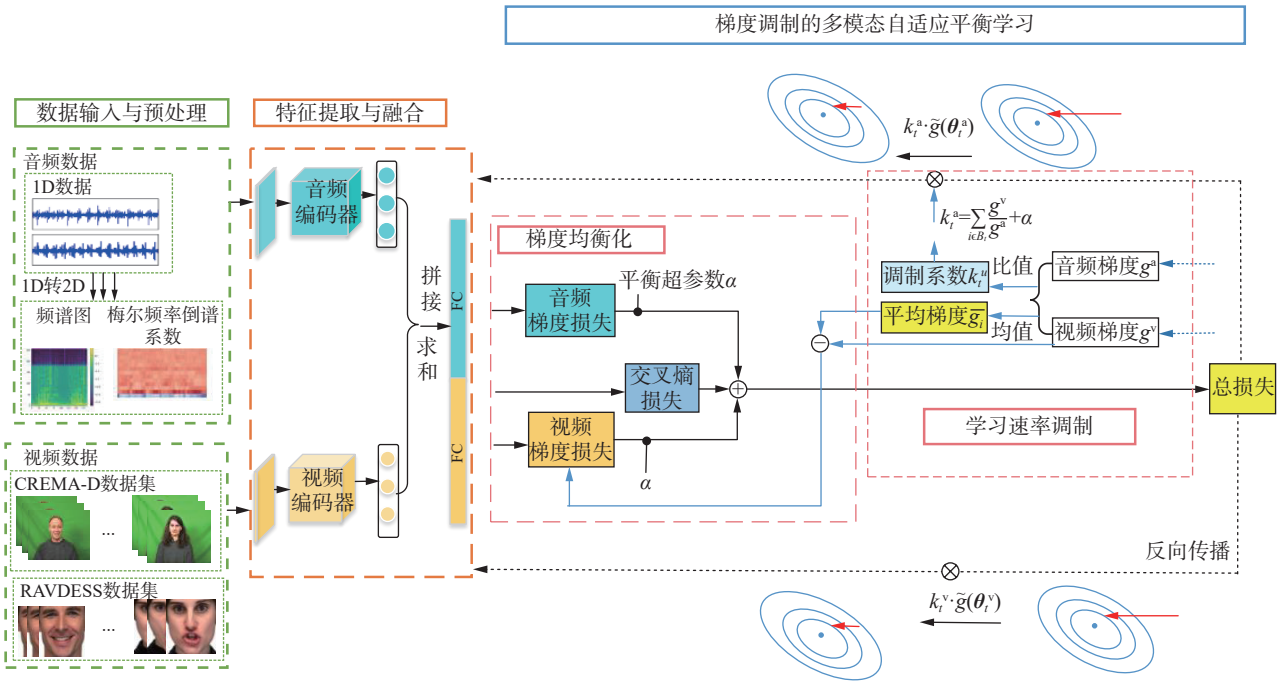


图 2　基于自适应梯度调制的平衡学习框架

Fig. 2　Balanced learning framework based on adaptive gradient modulation

如 2.1 小节所述, 采用随机梯度下降优化算法对编码器参数的更新可表示为

$$\boldsymbol{\theta}_{t+1}^u = \boldsymbol{\theta}_t^u - \eta \nabla_{\boldsymbol{\theta}^u} L(\boldsymbol{\theta}_t^u) = \boldsymbol{\theta}_t^u - \eta \tilde{g}(\boldsymbol{\theta}_t^u) \quad u = \{\mathrm{a}, \mathrm{v}\}$$

式中: η为学习率, $\nabla_{\boldsymbol{\theta}^u} L(\boldsymbol{\theta}_t^u)$和$\tilde{g}(\boldsymbol{\theta}_t^u)$为损失函数 L 对参数$\boldsymbol{\theta}_t^u$的梯度。为解决学习速率不平衡问题, 对参数更新作出修改:

$$\boldsymbol{\theta}_{t+1}^u = \boldsymbol{\theta}_t^u - \eta k_t^u \tilde{g}(\boldsymbol{\theta}_t^u) \tag{1}$$

式中, k_t^u为引入的调制系数, 根据前面分析, k_t^u应该满足以下几点: 通过模态间学习速率的差异自适应调整; 模态学习速率越低, 对应的值应该更大, 模态的学习速率越高, 对应的值应该更小, 目的是增大以低速率学习的模态梯度幅度, 降低以高速率学习的模态梯度幅度。综上所述, k_t^u可表示为

$$k_t^u = \begin{cases} \sum\limits_{i \in B_t} \dfrac{g^{\mathrm{a}}}{g^{\mathrm{v}}} + \alpha, & u = \mathrm{v} \\ \sum\limits_{i \in B_t} \dfrac{g^{\mathrm{v}}}{g^{\mathrm{a}}} + \alpha, & u = \mathrm{a} \end{cases} \tag{2}$$

式中: α为平滑项, 防止梯度比值过大或过小导致的不稳定; g^{a}和g^{v}分别为每个小批量 B_t 样本中的音频和视频模态梯度幅度。当音频模态的梯度较大时, 说明该模态学习速率较快, 此时 k_t^u的值小于 1, 从而降低其学习速率。相反, 视频模态的 k_t^u值大于 1, 提高其学习速率。由于网络最后一层直接影响模型的输出, 为节约计算成本, 所求的梯度为多模态模型中单模态学习网络的最后一层线性变换的梯度 2-范数。

综上所述, 通过调制系数的引入, 分别调控每个模态的学习速率, 确保较慢学习的模态获得更多的训练关注, 而较快学习的模态则会适当减缓学习速率, 避免某一模态主导模型的学习过程, 从而提升了模型的整体性能。

2.3　多模态梯度均衡化

梯度表示模型参数在当前优化方向上的变化率, 其大小决定了参数更新的幅度。在多模态学习中, 不同模态间的梯度差异可能导致模型更新的不均衡性, 即当某一模态的梯度显著大于其他模态, 模型会过度依赖该模态信息, 而忽略其他模态的特征, 这种不平衡不仅会削弱模型对多源输入信息的充分利用能力, 还可能导致模型的泛化性能下降, 从而影响其在不同任务和场景下的表现。为确保各模态在训练过程中对模型参数更新的贡献更加均衡, 提高模型的稳定性和泛化性

能, 本文重新设计了损失函数。在传统的交叉熵损失基础上, 引入了梯度损失作为正则化项, 以约束模态间梯度的差异。重构后的损失函数为

$$L=\sum_{i}^{B_t}L_r+r\times(L_{\mathrm{a}}+L_{\mathrm{v}}) \tag{3}$$

式中: L_r 为交叉熵损失, r 为平衡超参数, 目的是调节梯度损失在总体损失中的占比; L_{a} 和 L_{v} 分别为音频和视频模态的梯度损失, 表示为各模态偏离平均梯度的距离:

$$L_u=\ |g_u-\overline{g_i}|,\quad u=\{\mathrm{a},\mathrm{v}\} \tag{4}$$

$\overline{g_i}$为音频和视频模态的梯度幅度均值, 即平均梯度:

$$\overline{g_i}=\frac{\sum_{u=(a,v)}g_u}{2} \tag{5}$$

从式 (3)~(5) 可知, 正则项通过约束各模态梯度与平均梯度的偏离程度, 促使模型在训练过程中更均衡地利用各模态信息。具体而言, 梯度均衡化在一定程度上减缓梯度较大模态的更新, 同时加强梯度较小模态的更新幅度, 从而防止某一模态在训练中占据主导地位, 削弱其他模态的学习能力。该平衡方法不仅确保了各模态在训练中的均衡学习, 还有效减少了因某一模态过度拟合或欠拟合导致的模型不稳定性, 从而提高了模型的整体稳定性和准确性。

本文提出的梯度调制自适应平衡算法如算法 1 所示。

算法 1 梯度调制自适应平衡算法

输入 训练数据集 $\boldsymbol{D}=\{\boldsymbol{x}_i,y_i\}_{i=1,2,\cdots,N}$, 初始化模态参数 $\boldsymbol{\theta}^u,u=\{\mathrm{a},\mathrm{v}\}$

1) 前向传播:

2) 编码器 $\boldsymbol{\varphi}^{\mathrm{a}}$、$\boldsymbol{\varphi}^{\mathrm{v}}$ 提取音频和视频模态的特征;

3) 特征融合。

4) 反向传播:

5) 从 t=0 到最终训练结束:

6) 计算在单个批次样本 B_t 里多模态模型中单模态子网络最后一层线性变换的梯度幅度 (g^{a} 和 g^{v});

7) 由式 (2) 计算调制系数;

8) 获取单个批次样本任务交叉熵损失 L_r;

9) 由式 (4)、(5) 计算平均梯度和音频、视频模态梯度损失 L_{a}、L_{v};

10) 由式 (3) 将损失重新加权;

11) 由式 (1) 更新参数 $\boldsymbol{\theta}_t$;

12) 结束循环。

3 实验结果与分析

3.1 数据集

本文使用了 2 个数据集: CREMA-D(crowd-sourced emotional multimodal actors dataset)[26] 和 RAVDESS (Ryerson audio-visual database of emotional speech and song)[27]。

CREMA-D 是一个用于语音情感识别的视听数据集, 包含 91 名来自不同种族和民族的演员朗读几个简短词语的 7 442 个视频片段, 每个片段持续 2~3 s。该数据集涵盖 6 种最常见情绪: 愤怒、快乐、悲伤、中性、厌恶和恐惧。情绪标签通过众包方式从 2 443 名评分者的评估中获得。整个数据集按 9∶1 的比例随机分为 6 698 个样本的训练集和 744 个样本的测试集。

RAVDESS 是一个多模态情感识别数据集, 包括 24 名专业演员 (12 男、12 女) 朗读 2 个词法匹配的陈述, 数据总共包括 7 356 个文件, 涵盖纯音频、音频−视频和纯视频形式, 情绪类型为 8 种: 中性、平静、快乐、悲伤、愤怒、恐惧、厌恶和惊讶。本文仅使用其中的 1 440 个短演讲视频片段进行训练。

3.2 实验设置

预处理 在实验中, 均使用 Resnet18[28] 网络作为骨干网来提取特征。在 CREMA-D 数据集中, 从每个视频中随机提取一帧作为视觉编码器的输入, 将音频数据处理成大小为 257×188 的频谱图并将 Resnet18 的输入通道由 3 改为 1, 其余部分保持不变作为音频编码器的输入。在 RAVDESS 数据集中, 从每个视频中提取 30 帧连续图像, 并使用数据集中提供的 2D 人脸标记对每帧进行裁剪, 将人脸区域调整为 224×224 大小作为视觉编码器的输入, 将视频裁剪掉不包含声音的前 0.5 s, 根据文献 [29], 提取前 13 个梅尔频率倒谱系数 (Mel frequency cepstral coefficients, MFCC) 特征作为音频编码器的输入。

由于 RAVDESS 数据集没有预先划分训练集与测试集, 根据文献 [30] 建议, 采用五折交叉验证。具体而言, 将 24 个演员按 5∶1 的比例分成训练集和测试集。由于演员的性别由 ID 奇偶性区分, 因此通过选取 4 位连续 ID 的演员作为每折交叉验证的测试集, 确保性别均匀分布。

训练 所有实验均基于 PyTorch 框架, 在 NVIDIA Tesla T4 GPU 上进行训练。

3.3 超参数设置

为获取合适的超参数, 本文采用网格搜索的方式进行超参数选择。因平衡超参数 r 控制着分类交叉熵函数 L_r 和梯度损失 L_{a}、L_{v} 之间的权重分配, 为确保交叉熵损失仍在训练过程中占据主要地位, 同时让梯度损失也能有效参与但不超过交

叉熵损失的影响，将 r 的取值范围设定为有限集：$r \in \{0.5, 1, 1.5, 2, 2.5\}$。将随机抽取的超参数代入网络中进行训练和测试，以测试准确率最高为指标保存搜索得到的最优模型后，返回对应超参数，如表 1 所示。

表 1　不同数据集超参数设置

Table 1　Hyperparameter settings employed for different datasets

超参数	CREMA-D	RAVDESS
批量大小	64	16
学习率	10^{-3}	10^{-3}
学习率衰减步长	70	70
学习率衰减率	0.1	0.1
平衡超参数r	1	1
平滑项α	10^{-5}	10^{-5}

在迭代训练中，训练轮次 (epoch) 为 100，初始学习率为 0.001，并在每 70 个 epoch 衰减为原来的 1/10，使用随机梯度下降 (stochastic gradient descent, SGD) 优化器训练网络。

3.4　对比实验

为进一步验证 AGM-CR 的有效性和泛化能力，设计了 2 组对比实验。首先在 RAVDESS 数据集上将 AGM-CR 与不同的特征融合策略结合进行对比分析。除数据预处理部分有不同外，其余处理步骤与之前在 CREMA-D 数据集上的实验保持一致，采用五折交叉验证方法来确保结果的可靠性，具体实验结果如表 2 所示。

表 2　基于 RAVDESS 数据集的对比实验结果

Table 2　Results of comparative experiments conducted on the RAVDESS dataset

方法	准确率/%
Resnet18(Summation+FC)	65.4
Resnet18(Concat+FC)	65.7
AGM-CR(Summation+FC✦✦)	68.7
AGM-CR(Concat+FC✦✦)	66.9

注：“✦✦”表示应用AGM-CR。

从表 2 可观察到，使用 Resnet18 提取特征并采用特征求和 (Summation) 和全连接层 (FC) 进行分类时，模型的准确率为 65.4%，采用特征连接 (Concat) 与全连接层的组合，准确率为 65.7%。将 AGM-CR 与这两种特征融合策略结合后，准确率分别提升至 68.7% 和 66.9%，分别提高了 3.3 和 1.2 百分点。这些结果表明，AGM-CR 应用于传统特征融合策略，可以提高模型在验证集上的准确率，验证了 AGM-CR 的有效性和通用性。

接着在 CREMA-D 数据集上将 AGM-CR 与当前最新的平衡方法 OGM[13]、OGM-GE[13] 和 PMR[15] 进行定量对比。OGM 基于不同模态对学习目标的贡献差异动态调整模态学习速率，OGM-GE 在 OGM 基础上额外添加高斯噪声以增强模型泛化能力。PMR 通过引入原型来促进学习较慢的模态，并利用原型熵正则化抑制主导模态的过度影响。OGM 和 OGM-GE 代码公开可用，因此在相同的设置上复现了该算法，但复现结果和原论文中报告的结果存在一定差距，这可能是由于不同的数据预处理和运行环境带来的影响。为保证公平性，本实验采用相同的设置。PMR 数据来源于原始论文。实验结果如表 3 所示。

表 3　在 CREMA-D 数据集上与最新进平衡方法对比

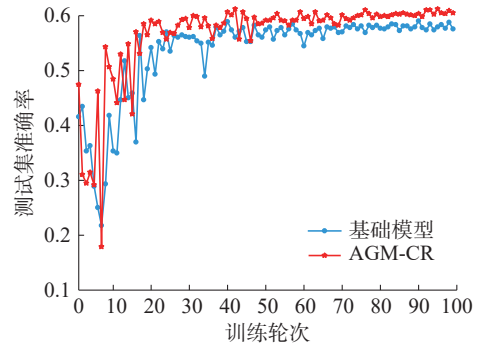
Table 3　Comparison with the latest balancing method on the CREMA-D dataset

方法	准确率/%
基础网络(Resnet18)	58.9
OGM[13] *	59.0
OGM-GE[13]*	59.9
PMR[15]	61.1
AGM-CR	61.4

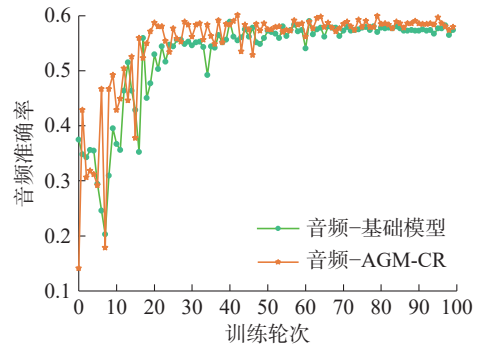
注：“*”表示复现结果。

在未采用任何平衡方法的情况下，基础模型在 CREMA-D 数据集的分类准确率为 58.9%，复现的 OGM 和 OGM-GE 方法分别达到了 59.0% 和 59.9%，原始 PMR 方法为 61.1%。本文提出的 AGM-CR 方法的准确率为 61.4%，相比 OGM-GE 提高了 1.5 百分点，相比 PMR 提高了 0.3 百分点。OGM 和 OGM-GE 通过减缓快速学习模态的学习速率，在一定程度上平衡模态差异，因此准确率较基础模型有所提升。PMR 在此基础上，同时增强了弱势模态的学习速率，进一步提高分类性能。而 AGM-CR 直接利用梯度差异平衡模态学习，取得了最高准确率。值得注意的是，这 4 种方法均仅采用最基本的 Resnet18 特征提取模型，在增加有限计算量和模型参数的基础上，带来了性能的提升。一方面说明了异质学习速率确实对模型性能产生了负面影响，另一方面证明了本文提出的 AGM-CR 方法在解决不平衡问题上的有效性。

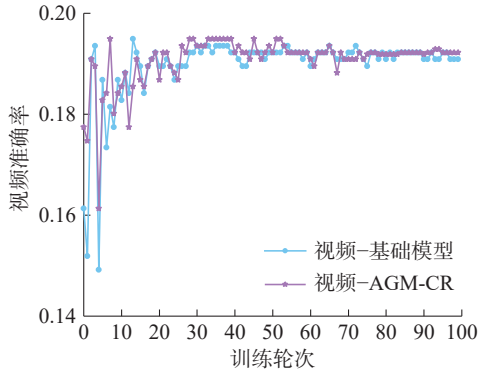
为了更直观地观察 AGM-CR 对模型性能的影响，在 CREMA-D 数据集上对比了基础模型和引入 AGM-CR 方法后的模型随迭代次数的准确率变化情况，图 3(a) ~ (c) 分别给出了基础模型和使用 AGM-CR 的多模态模型在测试集、多模态模型中的单模态音频子网络、多模态模型中的单模态视频子网络的准确率变化情况。



(a) 测试集



(b) 音频数据



(c) 视频数据

图 3　基础模型和采用 AGM-CR 的模型准确率随迭代次数的变化

Fig. 3　Accuracy changes exhibited by the baseline model and the model with AGM-CR over numerous iterations

在测试集上, 训练初期 (前 20 个训练轮次左右), 两种模型的准确率均快速上升, 其中引入 AGM-CR 的模型上升速度更快, 并且在后续训练中其准确率持续高于基础模型。趋于稳定后, 使用 AGM-CR 的模型准确率稳定在 60% 左右, 而未使用的基础模型则在 55% 左右, 且 AGM-CR 模型的波动较小, 稳定性更高。在多模态模型中的单模态子网络上, 引入 AGM-CR 的模型的准确率同样显示出比基础模型更高的准确率和更好的稳定性, 这与整体测试集中的趋势一致。总体而言, 图中结果表明, AGM-CR 在整体测试集、音频和视频数据上的准确率均有提升, 并且使模型更加稳定。

为了更清楚地了解 AGM-CR 对模型的调制作用, 在 CREMA-D 数据集上提取并对比了不同模态下加入 AGM-CR 与未加入时的梯度变化情况, 结果如图 4 所示。

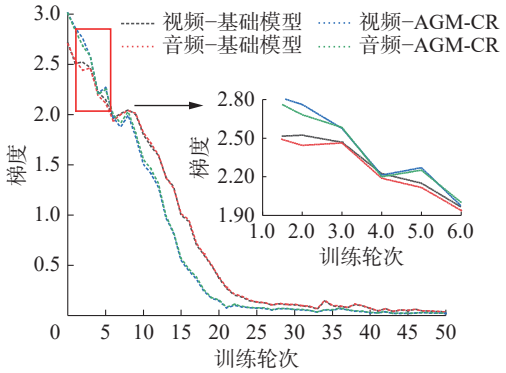
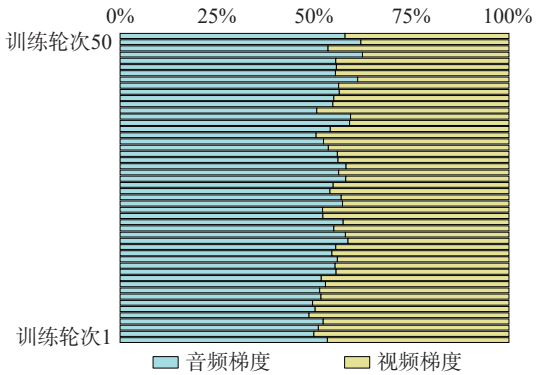


图 4　基准模型和加入 AGM-CR 的模型中模态梯度随迭代次数的变化

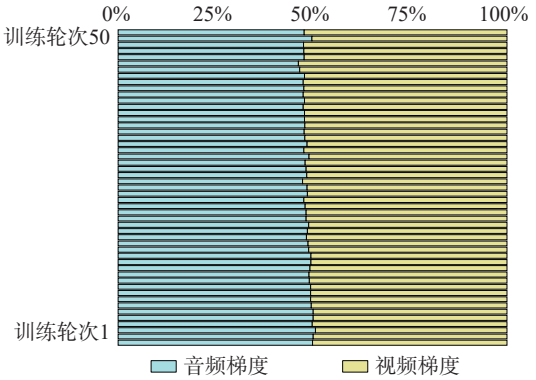
Fig. 4　Comparison between the visual and audio gradients of the normal and AGM-CR models across numerous epochs

从主图可以观察到, 使用 AGM-CR 的模型 (蓝色和绿色虚线) 在各模态的梯度值下降更快, 在训练后期 (约第 35 个训练轮次), 所有梯度趋于平稳, 接近于 0, 表示模型趋于收敛。在这一阶段, 使用 AGM-CR 的模型在视频和音频数据上的梯度波动更小, 显示出更好的训练稳定性。子图展示了训练初期的梯度变化细节, 可以明显看到, 使用 AGM-CR 的模型 (蓝色和绿色虚线) 在各模态之间的梯度差距小于基础模型 (黑色和红色虚线) 之间的梯度差距。这意味着 AGM-CR 减小了两个模态之间的学习速率差异, 显示出更好的训练效果。这些结果验证了 AGM-CR 通过调制两个模态间的梯度差异, 降低了学习速率的不平衡性, 提升了训练过程的稳定性和收敛速度。

基于上述提取的梯度数据查看 AGM-CR 在实验过程中各模态梯度调整过程, 采用百分比堆积条形图展示音频和视频模态在每个轮次训练过程中梯度的变化情况, 结果如图 5 所示。



(a) 基础模型



(b) AGM-CR

图 5　模态梯度占比变化

Fig. 5　Modality gradient proportion

从图 5 可看出，基础模型在整个训练过程中表现出更明显的梯度比例波动和随机性，音频模态的梯度占比始终处于较高水平，视频模态的梯度占比相对较低。相比之下，结合 AGM-CR 模型对不同模态数据的梯度比例进行了动态调整。音频模态的梯度幅度最终减小，对应视觉模态的梯度幅度增加，这表明模型在训练过程中逐渐将更多的关注度分配给了慢速学习的视频模态，避免了初期占优势的音频模态对整体训练过程的主导作用，最终实现了模态间的梯度达到相对平衡。

3.5　消融实验

为验证 AGM-CR 各部分的有效性，设计了一项消融实验以评估 AGM-CR 中各部分对分类准确率的影响。实验比较了 3 种方法的性能：基础骨干网络 (Resnet18)、在基础网络上加入梯度调制 (AGM) 以及在 AGM 基础上进一步加入正则化项 (AGM-CR)。各方法的具体结果如表 4 所示。

表 4　基于 CREMA-D 数据集的消融实验

Table 4　Ablation study results obtained on the CREMA-D dataset

方法	准确率/%
Resnet18 (Summation+FC)	58.9
AGM (Summation+FC✦)	59.2
AGM-CR(Summation+FC✦✦)	61.4

注：“✦”表示不加正则项，“✦✦”表示加入正则项。

表 4 显示，在 CREMA-D 数据集上，使用 Resnet18 提取特征，通过特征求和后经全连接层进行分类的准确率为 58.9%；在此基础上，仅加入梯度调制 (AGM) 时准确率提升了 0.3 百分点；进一步在 AGM 基础上加入正则项 (CR) 时准确率又提升了 2.2 百分点。这表明 AGM 和 CR 对于模型性能的提升均是有效的。最终，引入 AGM-CR 的准确率达到 61.4%，相比于基础网络的精度提升了 2.5 百分点。

为探究基于不同优化器的学习效果，分别采用 SGD 和自适应梯度优化器 (adaptive moment estimation, Adam) 进行实验，以评估所提方法是否能在不同的优化器下取得良好的性能。实验结果如表 5 所示。

表 5　在 CREMA-D 数据集上与不同优化器结合

Table 5　Experiments with SGD and Adam optimizers on the CREMA-D dataset

优化器	准确率/%
SGD	58.9
Adam	59.1
SGD✦✦	61.4
Adam✦✦	59.6

注：“✦✦”表示应用了AGM-CR。

从实验结果可看出，在 CREMA-D 数据集上，SGD 和 Adam 的基础准确率分别为 58.9% 和 59.1%。结合 AGM-CR 后，SGD 的准确率提高至 61.4%，Adam 提升至 59.6%，表明 AGM-CR 在不同优化器上的增强效果。值得注意的是，在相同实验设置下，SGD 与 AGM-CR 结合的提升幅度更大，相比 Adam 取得了更优的性能。这一现象可能是由于 Adam 在训练过程中会自适应调整学习率，从而降低了对 AGM-CR 进行梯度调制的敏感性，限制了进一步性能提升的空间。总体而言，AGM-CR 具有良好的适应性，能够在不同优化器的作用下实现稳定的性能改进。

4　结束语

本文为解决异质学习速率导致的音视频多模态网络性能退化问题，提出了一种基于自适应梯度调制 (AGM-CR) 的多模态平衡学习方法。该方法利用模态间学习梯度的差异计算调制系数，自适应地平衡单模态网络的学习速率，并利用梯度损失作为约束项，有效减小模态间的学习速率差异，平衡各模态对模型训练的影响。对比实验表明，AGM-CR 在 CREMA-D 和 RAVDESS 数据集上相较于基准模型准确率提升了 2.5 和 3.3 百分点，并且减少了模型的波动，表现出更高的稳定性和收敛速度。消融实验进一步验证了各部分的有效性。

然而，当前算法仍存在一些不足。尽管 AGM-CR 提升了模型性能，但由于骨干网络 Resnet18 结构较为简单，未能针对不同模态特点选择最优的特定编码器，导致最终准确性仍低于现有处理同

样数据集的最佳模型。

在未来的工作中, 将更关注以下问题: 将本文方法应用于更多模态的数据集, 验证其在其他多模态任务中的效果; 结合更先进的融合策略和网络架构, 进一步提升模型的性能和鲁棒性。

参考文献:

[1] 黄学坚, 马廷淮, 王根生. 基于样本内外协同表示和自适应融合的多模态学习方法[J]. 计算机研究与发展, 2024, 61(5): 1310−1324.
HUANG Xuejian, MA Tinghuai, WANG Gensheng. Multimodal learning method based on intra-and inter-sample cooperative representation and adaptive fusion[J]. Journal of computer research and development, 2024, 61(5): 1310−1324.

[2] 潘家辉, 何志鹏, 李自娜, 等. 多模态情绪识别研究综述[J]. 智能系统学报, 2020, 15(4): 633−645.
PAN Jiahui, HE Zhipeng, LI Zina, et al. A review of multimodal emotion recognition[J]. CAAI transactions on intelligent systems, 2020, 15(4): 633−645.

[3] CHANG Yicong, XUE Feng, SHENG Fei, et al. Fast road segmentation via uncertainty-aware symmetric network [C]//2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022: 11124−11130.

[4] CUI Can, MA Yunsheng, CAO Xu, et al. A survey on multimodal large language models for autonomous driving [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2024: 958−979.

[5] WANG Weiyao, TRAN D, FEISZLI M. What makes training multi-modal classification networks hard? [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 12695−12705.

[6] HUANG Yu, LIN Junyang, ZHOU Chang, et al. Modality competition: what makes joint training of multi-modal network fail in deep learning?(provably)[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 9226−9259.

[7] XU Ruize, FENG Ruoxuan, ZHANG Shixiong, et al. MMCosine: multi-modal cosine loss towards balanced audio-visual fine-grained learning[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1−5.

[8] LI Hong, LI Xingyu, HU Pengbo, et al. Boosting multi-modal model performance with adaptive gradient modulation[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 22157−22167.

[9] WEI Yake, HU Di, DU Henghui, et al. On-the-fly modulation for balanced multimodal learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2025, 47(1): 469−485.

[10] YANG Liu, WU Zhenjie, HONG Junkun, et al. MCL: a contrastive learning method for multimodal data fusion in violence detection[J]. IEEE signal processing letters, 2022, 30: 408−412.

[11] DU Chenzhuang, TENG Jiaye, LI Tingle, et al. On unimodal feature learning in supervised multimodal learning[C]//International Conference on Machine Learning. Honolulu: PMLR, 2023: 8632−8656.

[12] LIU Shilei, LI Lin, SONG Jun, et al. Multimodal pretraining with self-distillation for product understanding in E-commerce[C]//Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. Singapore: ACM, 2023: 1039−1047.

[13] PENG Xiaokang, WEI Yake, DENG Andong, et al. Balanced multimodal learning via on-the-fly gradient modulation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans: IEEE, 2022: 8238−8247.

[14] 刘成广, 王善敏, 刘青山. 类别平衡调制的人脸表情识别[J]. 计算机科学与探索, 2023, 17(12): 3029−3038.
LIU Chengguang, WANG Shanmin, LIU Qingshan. Class-balanced modulation for facial expression recognition[J]. Journal of frontiers of computer science and technology, 2023, 17(12): 3029−3038.

[15] FAN Yunfeng, XU Wenchao, WANG Haozhao, et al. PMR: prototypical modal rebalance for multimodal learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 20029−20038.

[16] LIN Xun, WANG Shuai, CAI Rizhao, et al. Suppress and rebalance: towards generalized multi-modal face anti-spoofing[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 211−221.

[17] 刘佳, 宋泓, 陈大鹏, 等. 非语言信息增强和对比学习的多模态情感分析模型[J]. 电子与信息学报, 2024, 46(8): 3372−3381.
LIU Jia, SONG Hong, CHEN Dapeng, et al. A multimodal sentiment analysis model enhanced with non-verbal information and contrastive learning[J]. Journal of electronics & information technology, 2024, 46(8): 3372−3381.

[18] ZHOU Yipin, LIM S N. Joint audio-visual deepfake detection[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 14780−14789.

[19] YU Wenmeng, XU Hua, YUAN Ziqi, et al. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis[J].

Proceedings of the AAAI conference on artificial intelligence, 2021, 35(12): 10790−10797.

[20] XIAO Yi, CODEVILLA F, GURRAM A, et al. Multimodal end-to-end autonomous driving[J]. IEEE Transactions on intelligent transportation systems, 2020, 23(1): 537−547.

[21] 刘慧, 朱积成, 王欣雨, 等. 面向医学图像融合的多尺度特征频域分解滤波[J]. 软件学报, 2024, 35(12): 5687−5709.
LIU Hui, ZHU Jicheng, WANG Xinyu, et al. Multi-scale feature frequency domain decomposition filtering for medical image fusion[J]. Journal of software, 2024, 35(12): 5687−5709.

[22] SUN Ya, MAI Sijie, HU Haifeng. Learning to balance the learning rates between various modalities via adaptive tracking factor[J]. IEEE signal processing letters, 2021, 28: 1650−1654.

[23] XIAO Fanyi, LEE Y J, GRAUMAN K, et al. Audiovisual slowfast networks for video recognition[EB/OL]. (2020−01−23)[2024−12−11]. https://arxiv.org/abs/2001.08740.

[24] 罗渊贻, 吴锐, 刘家锋, 等. 基于自适应权值融合的多模态情感分析方法[J]. 软件学报, 2024, 35(10): 4781−4793.
LUO Yuanyi, WU Rui, LIU Jiafeng, et al. Multimodal sentiment analysis method based on adaptive weight fusion[J]. Journal of software, 2024, 35(10): 4781−4793.

[25] WU Nan, JASTRZEBSKI S, Cho K, et al. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks[C]//International Conference on Machine Learning. Baltimore: PMLR, 2022: 24043−24055.

[26] CAO Houwei, COOPER D G, KEUTMANN M K, et al. CREMA-D: crowd-sourced emotional multimodal actors dataset[J]. IEEE transactions on affective computing, 2014, 5(4): 377−390.

[27] LIVINGSTONE S R, RUSSO F. 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.

[28] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[29] JIN Qin, LI Chengxin, CHEN Shizhe, et al. Speech emotion recognition with acoustic and lexical features[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing. South Brisbane: IEEE, 2015: 4749−4753.

[30] TANG Guichen, XIE Yue, LI Ke, et al. Multimodal emotion recognition from facial expression and speech based on feature fusion[J]. Multimedia tools and applications, 2023, 82(11): 16359−16373.

作者简介：

王忠美，讲师，电气与电子工程师协会 (IEEE) 会员，主要研究方向为人工智能、计算机视觉和遥感信息处理。E-mail：wangzhongmei@hut.edu.cn。

敖文秀，硕士研究生，主要研究方向为模态融合、多模态平衡学习。E-mail：m23081100020@stu.hut.edu.cn。

刘建华，教授，博士生导师，主要研究方向为轨道交通电传动控制与智能运维。主持国家自然科学基金项目 2 项、国家重点研发计划课题 1 项。E-mail：jhliu@hut.edu.cn。