The 34th International Joint Conference on Artificial Intelligence (IJCAI)
Paper ID: 754

Towards Equilibrium: An Instantaneous Probe–and–Rebalance Multimodal Learning Approach

Yang Yang, Xixian Wu, Qing–Yuan Jiang*
Nanjing University of Science and Technology, Nanjing, China

Introduction

Modality Imbalance in Multimodal Learning

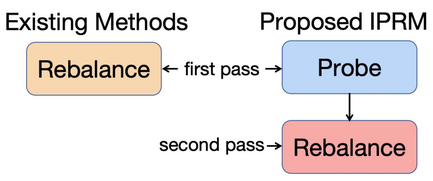
Modality imbalance arises when certain modalities contribute disproportionately during training. Stronger modalities tend to dominate the learning process, leading to insufficient utilization of weaker modalities and ultimately degrading overall model performance.

Limitations in Existing Rebalancing MML Methods

- **Rebalance only after imbalance occurs:** Existing methods use deferred rebalancing, intervening only after imbalance emerges, limiting their ability to prevent it proactively.
- **Learning under biased modality states:** These methods train the model under modality imbalance, causing it to optimize based on biased representations and thereby affecting overall performance.

Our Contributions

- A novel fusion representation strategy.
- A novel two-pass forward strategy.
- A novel integrated two-pass training method.



Methodology

Multimodal Fusion with GMM

- Extract unimodal representation:

$$\forall o \in \{a, v\}, Z_i^o = g_o(X_i^o; \Theta_o)$$

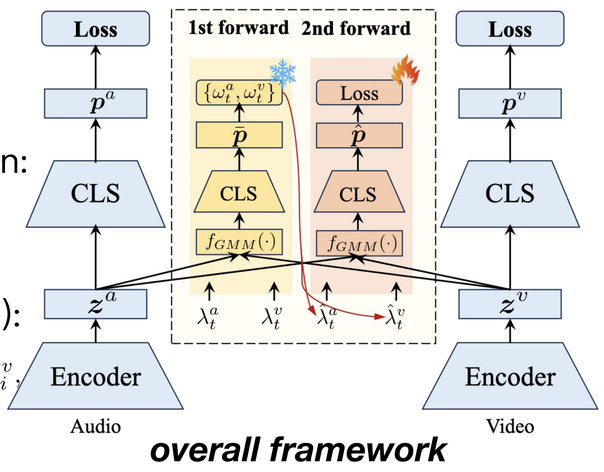
- Generate normalized representation:

$$\forall o \in \{a, v\}, \bar{Z}_i^o = \frac{Z_i^o}{\|Z_i^o\|_2}$$

- Geodesic multimodal mixup (GMM):

$$f_{GMM}(\bar{Z}_i^a, \bar{Z}_i^v, \lambda) = \frac{sin(\lambda\theta)}{sin(\theta)}\bar{Z}_i^a + \frac{sin((1-\lambda)\theta)}{sin(\theta)}\bar{Z}_i^v,$$

$$\text{where } \theta = arccos(\langle \bar{Z}_i^a, \bar{Z}_i^v \rangle)$$



overall framework

Instantaneous Probe–and–Rebalance for MML

Instantaneous Probing Phase: During this phase, we probe the strength of modality imbalance based on multimodal and unimodal predictions.

- **Step one:** We leverage GMM to obtain the fusion representation $\bar{\boldsymbol{z}}_i = f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \lambda_t^a)$ and prediction $\bar{\boldsymbol{p}}_i = softmax(h(\bar{\boldsymbol{z}}_i))$.
- **Step two:** We measure Kullback–Leibler (KL) divergence to evaluate the strength of each modality:

$$\forall o \in \{a, v\}, \mathcal{D}_{\mathrm{KL}}(\mathcal{P}^o|\bar{\mathcal{P}}; \mathcal{T}_t) = \sum_{\boldsymbol{x}_i \in \mathcal{T}_t} \boldsymbol{p}_i^o \log\left(\frac{\boldsymbol{p}_i^o}{\bar{\boldsymbol{p}}_i}\right)$$

- **Step three:** We define the instantaneous strength weight of a specific modality based on the proportion of the KL divergence from another modality:

$$\omega_t^a \triangleq \frac{\mathcal{D}_{\mathrm{KL}}(\mathcal{P}^v|\mathcal{P}; \mathcal{T}_t)}{\mathcal{D}_{\mathrm{KL}}(\mathcal{P}^a|\mathcal{P}; \mathcal{T}_t) + \mathcal{D}_{\mathrm{KL}}(\mathcal{P}^v|\mathcal{P}; \mathcal{T}_t)}, \qquad \omega_t^v \triangleq 1 - \omega_t^a.$$

Rebalanced Learning Phase: At this stage, we perform rebalanced learning under the balanced status.

- **Step one:** Update the balanced weights for each modality at t–th iteration: $\forall o \in \{a, v\}, \hat{\lambda}_t^o = \omega_t^o$.
- **Step two:** Obtain fusion representation $\hat{\boldsymbol{z}}_i = f_{GMM}(\bar{\boldsymbol{z}}_i^a, \bar{\boldsymbol{z}}_i^v, \hat{\lambda}_t^a)$ and prediction $\hat{\boldsymbol{p}}_i = softmax(h(\hat{\boldsymbol{z}}_i))$ under balanced status.
- **Step three:** Update the initial weights for the next iteration to adjust the intervention intensity between modalities:

$$\forall o \in \{a, v\}, \lambda_{t+1}^o = \begin{cases} \omega_t^o, & t = 0, \\ \alpha\lambda_t^o + (1-\alpha)\omega_t^o, & t > 0. \end{cases}$$

Overall Loss function

$$\ell(\boldsymbol{x}_i, \boldsymbol{y}_i) = \ell_m(\boldsymbol{x}_i, \boldsymbol{y}_i; \Phi) + \sum_{o \in \{a,v\}} \ell_u(\boldsymbol{x}_i^o, \boldsymbol{y}_i; \Theta_o, \Phi_o).$$

Experiments

Main Results

Comparision with Naive MML Methods

Dataset	Metric	Unimodal			Naive Fusion			IPRM
		A/A/R/A/I	V/V/O/V/T	D/T	Concat	Sum	Weight	
CREMA-D	Accuracy	63.17%	45.83%	N/A	63.61%	63.44%	66.53%	**84.27%** (↑17.74%)
	MAP	68.61%	58.79%	N/A	68.41%↓	69.08%	71.34%	**90.66%** (↑19.32%)
KSounds	Accuracy	54.12%	55.62%	N/A	64.55%	64.90%	65.33%	**74.37%** (↑9.04%)
	MAP	56.69%	58.37%	N/A	71.30%	71.03%	71.10%	**80.63%** (↑9.33%)
NVGesture	Accuracy	78.22%	78.63%	81.54%	82.37%	80.50%↓	78.42%↓	**85.89%** (↑3.52%)
	Macro-F1	78.33%	78.65%	81.83%	82.70%	80.67%↓	79.39%↓	**86.34%** (↑3.64%)
IEMOCAP	Accuracy	58.45%	30.71%	70.55%	75.97%	76.06%	69.29%↓	**80.22%** (↑4.16%)
	Macro-F1	58.29%	11.75%	69.93%	75.88%	76.03%	68.91%↓	**80.63%** (↑4.60%)
Sarcasm	Accuracy	71.81%	81.36%	N/A	82.86%	82.94%	82.65%	**85.14%** (↑2.20%)
	Macro-F1	70.73%	80.56%	N/A	82.40%	82.47%	82.19%	**84.41%** (↑1.94%)

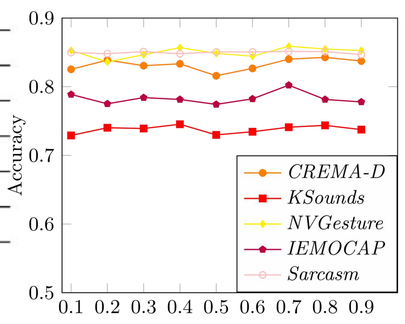
Comparision with Rebalancing MML Methods

Dataset	Metric	OGR-GB	MSLR	OGM	PMR	AGM	MMPareto	ReconBoost	MLA	LFM	IPRM
CREMA-D	Accuracy	64.65%	68.68%	66.12%	66.59%	67.33%	74.87%	75.57%	79.43%	83.62%	**84.27%** (↑0.65%)
	MAP	73.92%	74.12%	73.72%	70.58%	78.07%	85.35%	81.40%	85.72%	90.06%	**90.66%** (↑0.60%)
KSounds	Accuracy	67.22%	67.56%	65.82%	66.75%	67.91%	70.00%	68.55%	70.04%	72.53%	**74.37%** (↑1.84%)
	MAP	72.74%	72.82%	71.59%	72.74%	73.88%	78.50%	76.62%	79.45%	78.97%	**80.63%** (↑1.66%)
NVGesture	Accuracy	82.99%	82.37%	N/A	N/A	82.79%	83.82%	83.86%	83.40%	84.36%	**85.89%** (↑1.53%)
	Macro-F1	83.05%	82.84%	N/A	N/A	82.84%	84.24%	84.34%	83.72%	84.68%	**86.34%** (↑1.66%)
IEMOCAP	Accuracy	70.10%	76.69%	N/A	N/A	77.51%	77.69%	76.87%	79.31%	78.41%	**80.22%** (↑0.91%)
	Macro-F1	69.90%	76.77%	N/A	N/A	77.29%	77.89%	77.08%	79.73%	78.51%	**80.63%** (↑0.90%)
Sarcasm	Accuracy	82.86%	84.39%	83.60%	83.10%	83.06%	83.48%	84.37%	84.26%	84.97%	**85.14%** (↑0.17%)
	Macro-F1	82.15%	83.78%	82.93%	82.56%	82.93%	82.84%	83.17%	83.48%	**84.57%**	84.41% (↓0.16%)

Ablation Study

Dataset	w/ L-Mixup	w/o EMA	One-Pass	IPRM
CREMA-D	75.53%	83.06%	83.47%	**84.27%**
KSounds	71.94%	73.91%	73.64%	**74.37%**
NVGesture	84.85%	85.27%	84.44%	**85.89%**
IEMOCAP	75.79%	78.05%	77.60%	**80.22%**
Sarcasm	84.52%	84.81%	84.10%	**85.14%**

Sensitivity Analysis

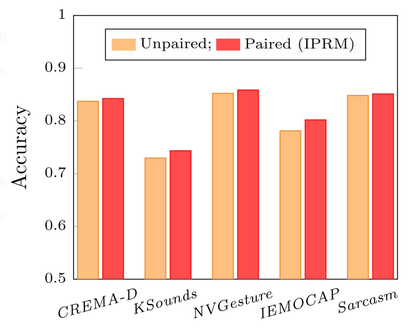


Further Analysis

Computation Cost

Method	Accuracy	Training time (second/epoch)
Naive MML	63.61%	55.08 ± 0.2729
MLA	79.43%	71.12 ± 0.7025
LFM	83.62%	60.14 ± 0.0920
IPRM	84.27%	57.03 ± 0.2138

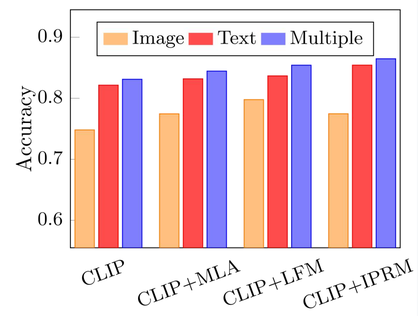
Unpaired GMM



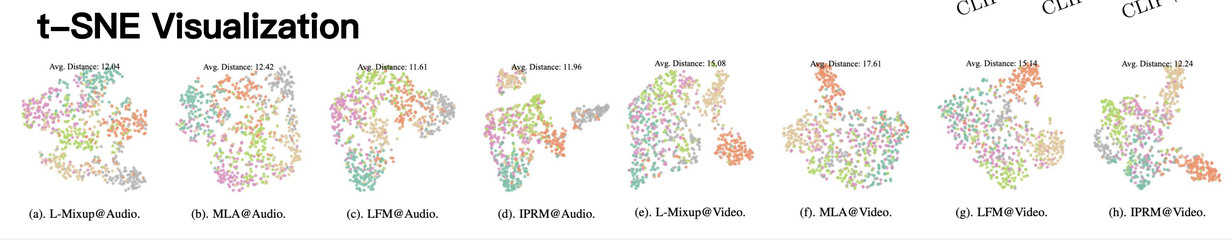
Mixup Strategy for Trimodal Dataset

Dataset	Modality	Single-CLS	Tri-CLS
NVGesture	RGB	**78.84%**	77.80%
	OF	79.25%	**81.12%**
	Depth	**82.78%**	82.16%
	Multi	**85.89%**	**85.89%**
IEMOCAP	Audio	**58.27%**	54.20%
	Video	**32.07%**	30.80%
	Text	**71.91%**	**71.91%**
	Multi	78.95%	**80.22%**

Pretrained Model



t–SNE Visualization



(a). L-Mixup@Audio. (b). MLA@Audio. (c). LFM@Audio. (d). IPRM@Audio. (e). L-Mixup@Video. (f). MLA@Video. (g). LFM@Video. (h). IPRM@Video.

Conclusion

- Proposed IPRM: An instantaneous probe-and-rebalance framework for multimodal learning.
- Key Techniques: Two-forward phase strategy and geodesic multimodal mixup for dynamic modality probing and weight adjustment.
- Effectiveness: Achieves consistent improvements over state-of-the-art methods on multiple benchmark datasets.

Contact Info

yyang@njust.edu.cn
xixianwu@njust.edu.cn
jiangqy@njust.edu.cn
KMG Group
WeChat