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

Yang Yang, Xixian Wu, Qing-Yuan Jiang*

Nanjing University of Science and Technology

Speaker: Qing-Yuan Jiang

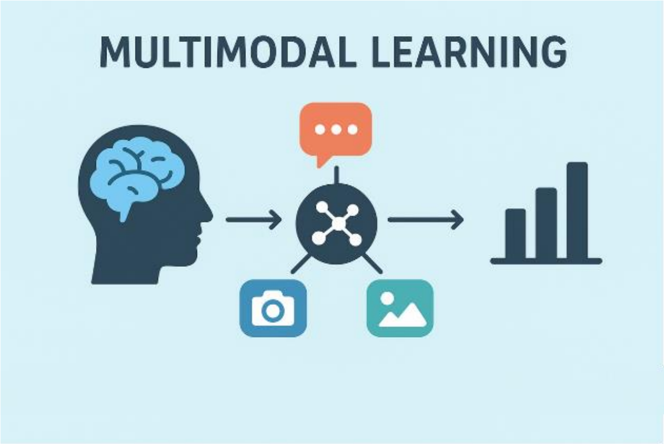
CONTENTS

Background

南京理工大学 NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

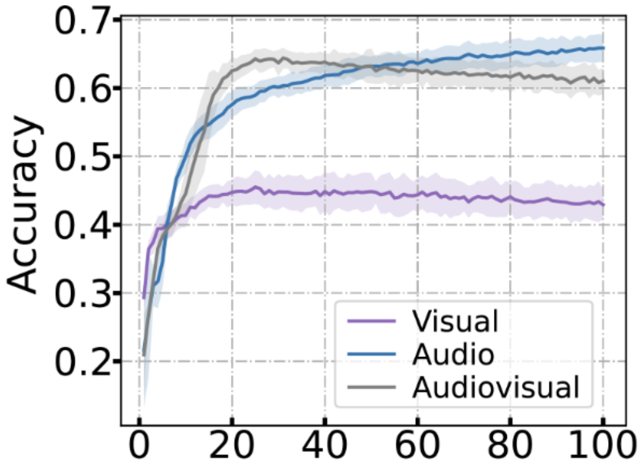
Multimodal Learning (MML) :

- Integrating data from multiple sensors.
- Making more reliable decisions.



Modality Imbalance:

- MML underperforms single-modality.
- Strong modality VS week modality.



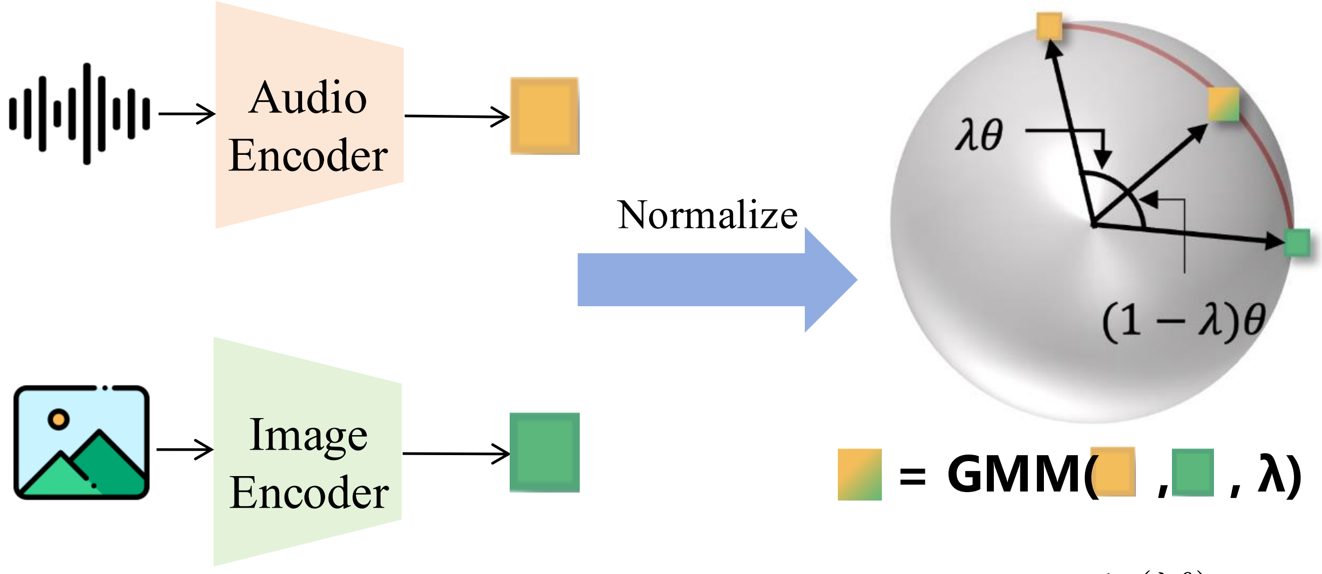
Issues in Existing Rebalancing Methods

Modality Imbalance → Imbalanced Status → Rebalancing & Learning

Deferred rebalancing strategy: addresses modal imbalance only after it has occurred !

Methodology

Multimodal Fusion with GMM



Unimodal Feature Extraction

↓

Feature Normalization

↓

Geodesic Multimodal Mixup

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

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

Enable effortless adjustment of modality strength between different modalities.

Methodology

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Instantaneous Probe-and-Rebalance for MML

Instantaneous Probing Phase:

- Extract Multimodal representation.

$$\bar{Z}_i = f_{GMM}(\bar{Z}_i^a, \bar{Z}_i^v, \lambda_t^a),$$
$$\bar{P}_i = softmax(h(\bar{Z}_i)).$$

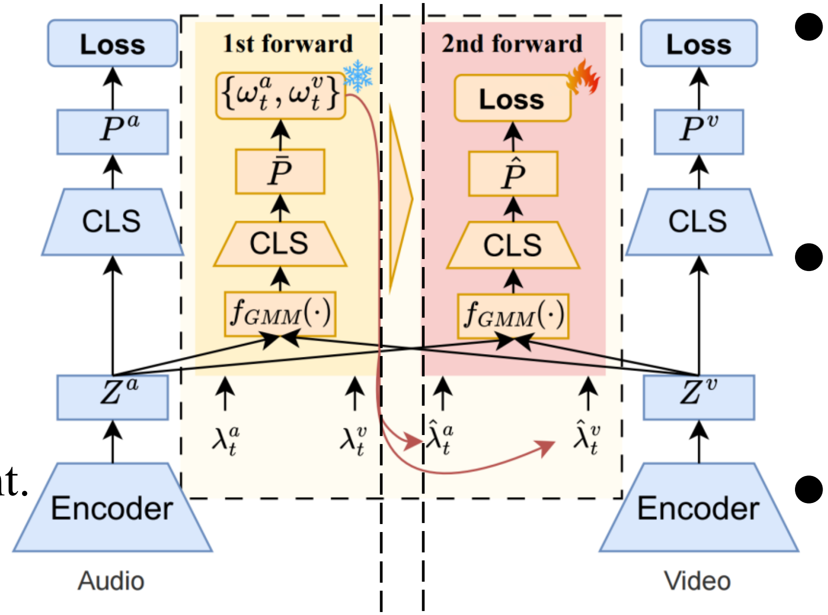
- Evaluate modality strength.

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

- Define instantaneous strength weight.

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

Probe but not learn



Rebalanced Learning Phase:

- Update modality balanced weight.

$$\forall o \in \{a, v\},$$
$$\hat{\lambda}_t^o = \omega_t^o.$$

- Obtain balanced representation.

$$\hat{Z}_i = f_{GMM}(\bar{Z}_i^a, \bar{Z}_i^v, \hat{\lambda}_t^a),$$
$$\hat{P}_i = softmax(h(\hat{Z}_i)).$$

- Update probing weight.

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

Learn under balanced status

Experiments

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Main Results

Comparison with Naive MML

Dataset	Metric	Unimodal			Naive Fusion			IPRM
		A/A/R/A/I	V/V/O/V/T	D/T	Concat	Sum	Weight	
CREMA-D	Accuracy	45.83%	63.17%	N/A	63.61%	63.44%	66.53%	**84.27%** (↑17.74%)
	MAP	58.79%	68.61%	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%)

Comparison with Rebalanced MML

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%)

IPRM achieves superior performance in almost all cases !

Experiments

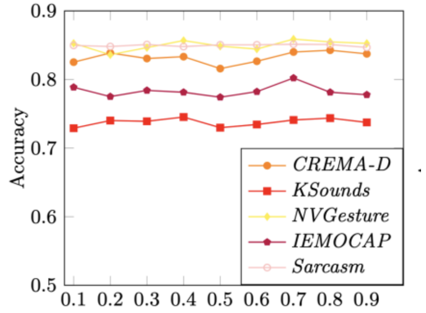
南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Additional Results

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 Anaysis to α



Computation Cost of Two-Pass Forward

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

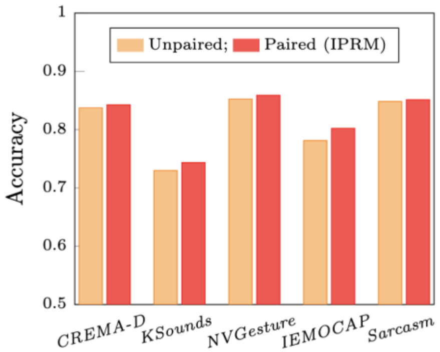
Mixup Strategy on Trimodal Dataset

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%**

Experiments

南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Unpaired GMM



Robustness of the Pretrained Model

Method	Image	Text	Multiple
CLIP	74.82%	82.15%	83.11%
CLIP+MLA	77.45%	83.19%	84.45%
CLIP+LFM	**79.78%**	83.67%	85.42%
CLIP+IPRM	77.46%	**85.43%**	**86.47%**

Conclusion

- We propose IPRM, a multimodal learning method with instantaneous probe-and-rebalance.
- GMM enables effortlessly adjustment the modality strength between different modalities.
- Two-Pass Forward strategy allows the model to learn under balanced status.
- Experiments show that IPRM achieves state-of-the-art performance on widely used datasets.

contact us