Facilitating Multimodal Classification via Dynamically Learning Modality Gap

Yang Yang[1], Fengqiang Wan[1], Qingyuan Jiang[1*], Yi Xu[2]

[1]Nanjing University of Science and Technology

[2]Dalian University of Technology

NeurIPS2024

Code

NJUST-KMG

CONTENT

Part 01 *Background*

Part 02 *Method*

Part 03 *Experiments*

Part 04 *Conclusion*

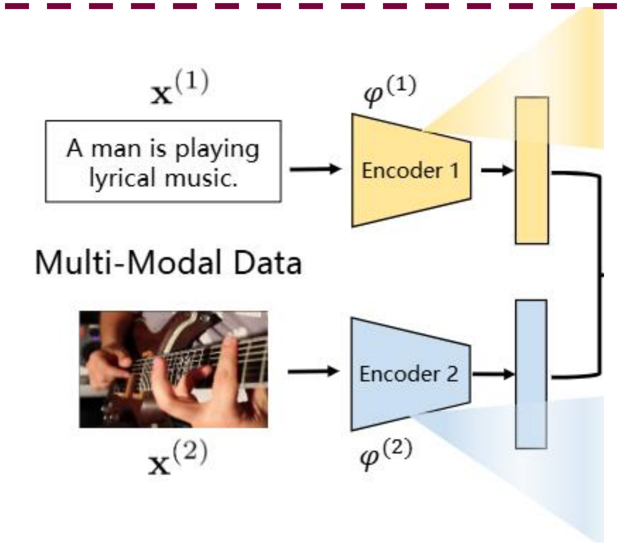
Background

Multimodal Classification

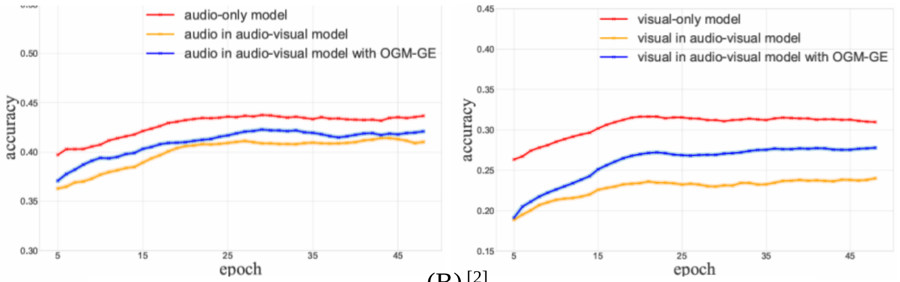
Use Multimodal data to **enhance** understanding and processing of complex tasks.

Modality Imbalance

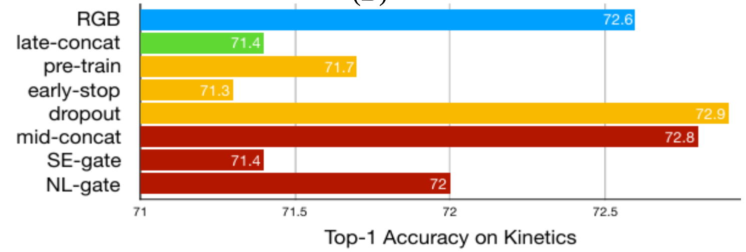
Different modalities converge at **different** speeds[2,3], causing strong modalities to **dominate** while weak ones are ignored.



(A) [1]



(B) [2]



(C) [3]

[1] Yang, Yang, *et al*. "Learning to Rebalance Multi-Modal Optimization by Adaptively Masking Subnetworks." *arXiv*. 2024.

[2] Peng, Xiaokang, *et al*. "Balanced multimodal learning via on-the-fly gradient modulation." *CVPR*. 2022.

[3] Wang, Weiyao, *et al*. “What makes training multi-modal classification networks hard? " *CVPR*. 2020.

Motivation

What are the core causes of modality imbalance?

Tiny Experiment

- One-hot Labels $y = [0, \mathbf{1}, 0, 0, 0, \dots, 0]$
- Lable Smoothing $y = [\frac{1}{31}, \frac{\mathbf{10}}{\mathbf{31}}, \frac{1}{248}, \frac{1}{124}, \frac{1}{62}, \dots, \frac{1}{31}]$
- Lable Free $y = [\frac{1}{31}, \frac{1}{31}, \frac{1}{31}, \frac{1}{31}, \frac{1}{31}, \dots, \frac{1}{31}]$

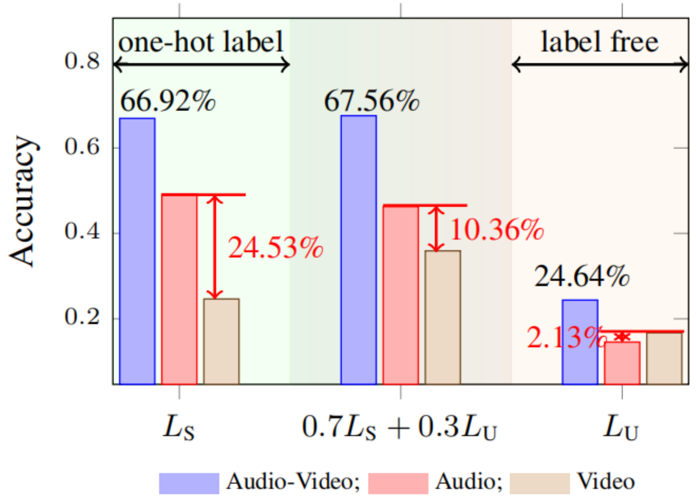


Figure 1: The influence of labels fitting on performance gaps (best view in color), where L_S and L_U denote the loss with one-hot labels and uniform labels (label free).

Appropriate intervention label fitting can alleviate the difference in the learning ability of different modalities

Motivation

How do we impose positive intervention?

Contrastive learning

- Learn **similar representations** for data pairs of different modalities
- **Cross-modal similarity** as a key learning signal, reducing reliance on one-hot labels.

Dynamic integration

- Gradually finds the optimal combination of modal alignment and classification accuracy.

Method

◆ Unsupervised contrast learning focuses on feature representation between **aligned** modalities

$$L_{MM}(X) = -\frac{1}{2n_b}\sum_{i}^{n_b}[log(\frac{exp(s(x_i^{(j)}, x_i^{(l)})/\tau)}{\sum_k exp(s(x_i^{(j)}, x_k^{(l)})/\tau)}) + log(\frac{exp(s(x_i^{(j)}, x_i^{(l)})/\tau)}{\sum_k exp(s(x_k^{(j)}, x_i^{(l)})/\tau)})]$$

◆ Supervised multimodal learning focuses on optimizing the **fit** of class labels

$$L_{CLS}(X, Y) = -\frac{1}{n}\sum_{i=1}^{n} y_i^T log\widehat{y}_i$$

Method

Integrating classification and modality matching losses gives the following objective function:

$$L_{Total} = (1 - \alpha)L_{CLS}(\theta) + \alpha L_{MM}(\theta)$$

To meet the model's evolving needs, the importance of different objectives should adapt at each stage.

Heuristic: Focus on alignment first, then classification.

$$\alpha(t) = 1 - e^{-\frac{1}{t}}$$

Learning-based: Find optimal classification within feasible regions across tasks.

$$\min_{0 \le \alpha \le 1} L_{CLS}(\theta^*(\alpha)) \;\; s.t. \, \theta^*(\alpha) \in \underset{\theta}{\operatorname{argmin}}\{(1-\alpha)L_{CLS}(\theta) + \alpha L_{MM}(\theta)\}$$

Algorithm 1: The Proposed Algorithm.

Input :Training set $\mathcal{X}$, labels $\mathcal{Y}$, method.
Output :Learned parameters $\{\theta\}$ of all models.
INIT initialize parameters θ, parameter α, maximum iterations T, learning rate η_α.

```
for t = 1 to T do
    /* updating neural network parameters θ. */
    for i = 1 to Inner_Iters do
        Calculate total loss L_Total by forward phase.
        Update parameters θ according to its gradient.
    end
    /* updating weighting parameters α based on the chosen method. */
    if method == 'learning-based' then
        Calculate gradient apprixmation:
          ∇L_CLS(θ(α)) = −∇²_{α,θ} L_Total · [∇²_{θ,θ} L_Total]^{-1} · ∇_θ L_CLS(X, Y).
        Update α according to: α = α − η_α ∇L_CLS(θ(α)).
        Clip α into [0, 1]: α := max(0, min(1, α)).
    else if method == 'heuristic' then
        Update α according to: α = 1 − e^{-1/t}.
    end
end
```

Experiments

Table 1: Comparison with SOTA multimodal learning methods. The best results are highlighted in bold. The underlining symbol denotes the second best performance. The results with gray background are based on MML but perform worse than the best unimodal approach.

Method	KineticsSounds		CREMA-D		Sarcasm		Twitter2015		NVGesture	
	ACC	MAP	ACC	MAP	ACC	F1	ACC	F1	ACC	F1
Unimodal-1	54.12%	56.69%	63.17%	68.61%	81.36%	80.65%	73.67%	68.49%	78.22%	78.33%
Unimodal-2	55.62%	58.37%	45.83%	58.79%	71.81%	70.73%	58.63%	43.33%	78.63%	78.65%
Unimodal-3	–	–	–	–	–	–	–	–	81.54%	81.83%
Concat	64.55%	71.31%	63.31%	68.41%	82.86%	82.43%	70.11%	63.86%	81.33%	81.47%
Affine	64.24%	69.31%	66.26%	71.93%	82.47%	81.88%	72.03%	59.92%	82.78%	82.81%
Channel	63.51%	68.66%	66.13%	71.75%	–	–	–	–	81.54%	81.57%
ML-LSTM	63.84%	69.02%	62.94%	64.73%	82.05%	70.73%	70.68%	65.64%	83.20%	83.30%
Sum	64.97%	71.03%	63.44%	69.08%	82.94%	82.47%	73.12%	66.61%	82.99%	83.05%
Weight	65.33%	71.33%	66.53%	73.26%	82.65%	82.19%	72.42%	65.16%	83.42%	83.57%
ETMC	65.67%	71.19%	65.86%	71.34%	83.69%	83.23%	73.96%	67.39%	83.61%	83.69%
MSES	64.71%	72.52%	61.56%	66.83%	84.18%	83.60%	71.84%	66.55%	81.12%	81.47%
G-Blend	67.12%	71.39%	64.65%	68.54%	83.35%	82.71%	74.35%	68.69%	82.99%	83.05%
OGM	66.06%	71.44%	66.94%	71.73%	83.23%	82.66%	<u>74.92%</u>	68.74%	–	–
Greedy	66.52%	72.81%	66.64%	72.64%	–	–	–	–	82.74%	82.69%
DOMFN	66.25%	72.44%	67.34%	73.72%	83.56%	82.62%	74.45%	68.57%	–	–
MSLR	65.91%	71.96%	65.46%	71.38%	84.23%	83.69%	72.52%	64.39%	82.86%	82.92%
PMR	66.56%	71.93%	66.59%	70.36%	83.61%	82.49%	74.25%	68.62%	–	–
AGM	66.02%	72.52%	67.07%	73.58%	84.28%	83.44%	74.83%	69.11%	82.78%	82.82%
MLA	70.04%	74.13%	<u>79.43%</u>	<u>85.72%</u>	84.26%	83.48%	73.52%	67.13%	83.73%	83.87%
ReconBoost	<u>70.85%</u>	74.24%	74.84%	81.24%	<u>84.37%</u>	83.17%	74.42%	68.34%	<u>84.13%</u>	**86.32%**
MMPareto	70.00%	**78.50%**	74.87%	75.15%	83.48%	82.84%	73.58%	67.29%	83.82%	84.24%
Ours-H	69.05% ±0.15%	72.97% ±0.43%	72.15% ±0.32%	80.45% ±0.85%	84.12% ±0.17%	<u>83.98%</u> ±0.22%	73.87% ±0.35%	<u>69.17%</u> ±0.26%	83.24% ±0.07%	83.87% ±0.18%
Ours-LB	**72.53%** ±0.31%	<u>78.38%</u> ±0.37%	**83.62%** ±0.11%	**90.06%** ±1.09%	**84.97%** ±0.27%	**84.57%** ±0.18%	**75.01%** ±0.16%	**70.57%** ±0.28%	**84.36%** ±0.14%	<u>84.68%</u> ±0.24%

Table 2: Results on VGGSound dataset.

Method	ACC	MAP
AGM	47.11%	51.98%
MLA	<u>51.65%</u>	54.73%
ReconBoost	50.97%	53.87%
MMPareto	51.25%	<u>54.74%</u>
Ours-H	50.42%	53.62%
Ours-LB	**52.74%**	**55.98%**

Table 5: Results on the Sarcasm and Twitter2015 datasets achieved by using the CLIP pre-trained model as encoders.

Method	Sarcasm			Twitter2015		
	Image	Text	Multiple	Image	Text	Multiple
CLIP	74.82%	82.15%	83.11%	54.48%	71.75%	72.52%
CLIP+MLA	77.45%	83.19%	84.45%	56.53%	72.37%	73.95%
CLIP+Ours	**79.78%**	**83.67%**	**85.42%**	**64.67%**	**72.59%**	**74.43%**

◆ Our learning-based strategy consistently **outperforms** baselines across datasets.

◆ Our method leads on VGGSound and excels with CLIP integration on Sarcasm and Twitter2015.

Experiments

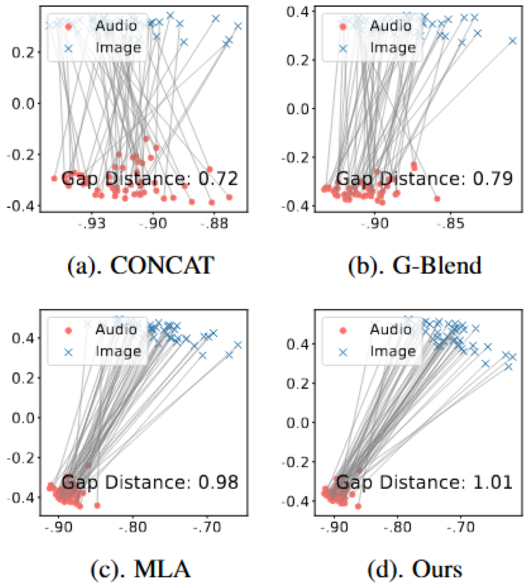


Figure 2: Visualizations of the modality gap distance on the CREMA-D dataset.

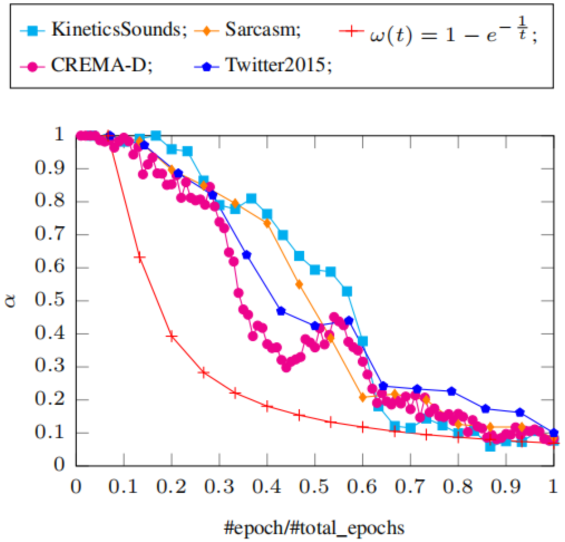


Figure 3: Change of α on different datasets. We illustrate the value of the heuristic integration strategy for comparison.

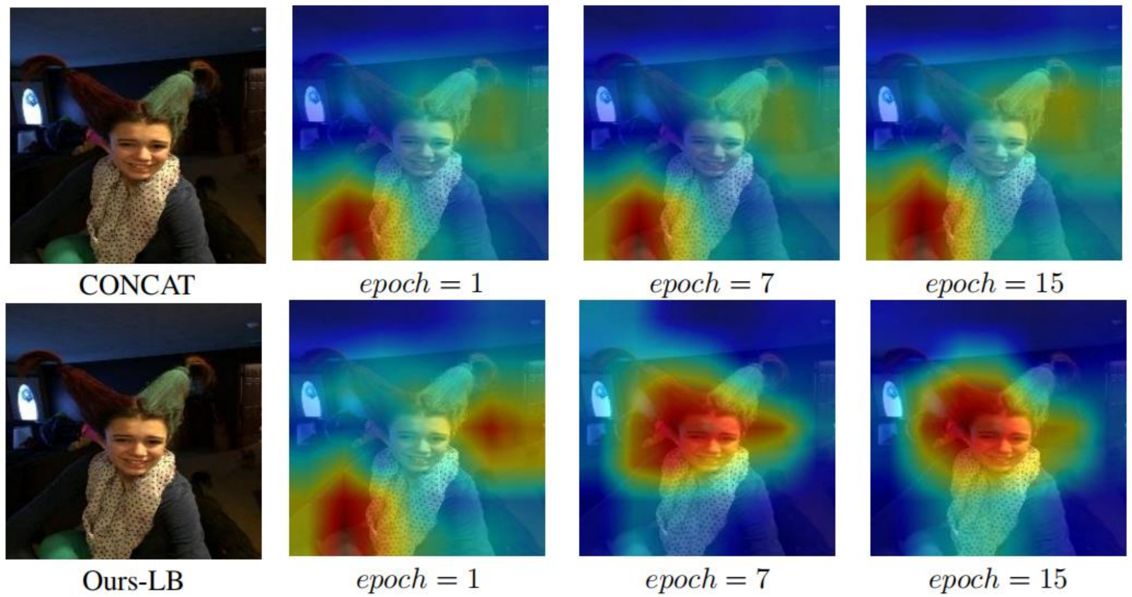


Figure 4: Visualization on Twitter2015 dataset. Our proposed method tends to perform feature learning first and then fit the learned features to the category labels.

- The learning-based strategy adapts α effectively across datasets, with **a polynomial approximation** of heuristic adjustments further enhancing performance.
- Larger modality gaps in our method lead to **more discriminative representations** and higher accuracy.
- Compared to CONCAT, our method better **aligns features with category labels** by focusing on relevant modality details.

Conclusion

◆This study identifies label fitting as a core cause of modality imbalance in multimodal learning.

◆We propose a method that dynamically combines unsupervised contrastive learning with supervised multimodal learning to mitigate this imbalance.

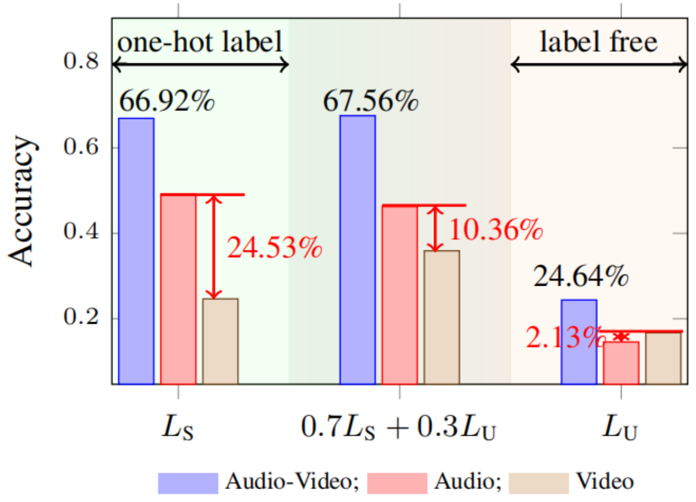


Figure 1: The influence of labels fitting on performance gaps (best view in color), where L_S and L_U denote the loss with one-hot labels and uniform labels (label free).

◆Future work will explore whether certain category labels inherently favor specific modalities to better address modality imbalance.

THANK YOU !
Q&A

Code

NJUST-KMG