Thirty-Eighth Conference on Neural Information Processing Systems
2024

# Facilitating Multimodal Classification via Dynamically Learning Modality Gap

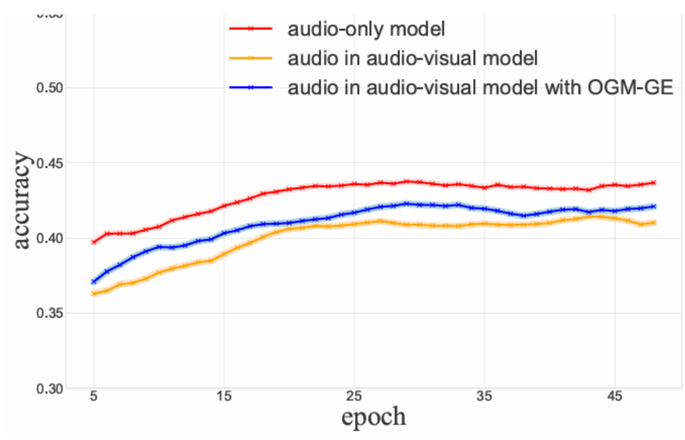

**Yang Yang**[1], Fengqiang Wan[1], Qingyuan Jiang[1*], Yi Xu[2]

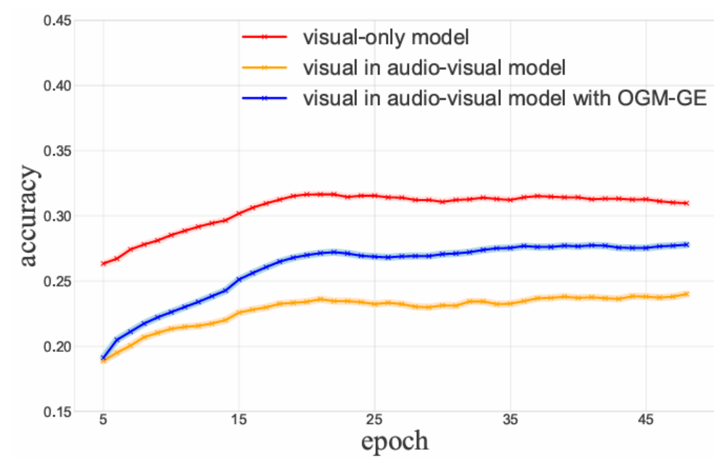
[1]Nanjing University of Science and Technology, Nanjing, China;
[2]Dalian University of Technology, Liaoning, China.


## Background

➢ Multimodal Classification Leverage multimodal data to improve comprehension and processing of complex tasks.

➢ Different modalities converge at different speeds [Peng, et al, Wang, *et al*], causing strong modalities to dominate while weak ones are ignored.




## Motivation

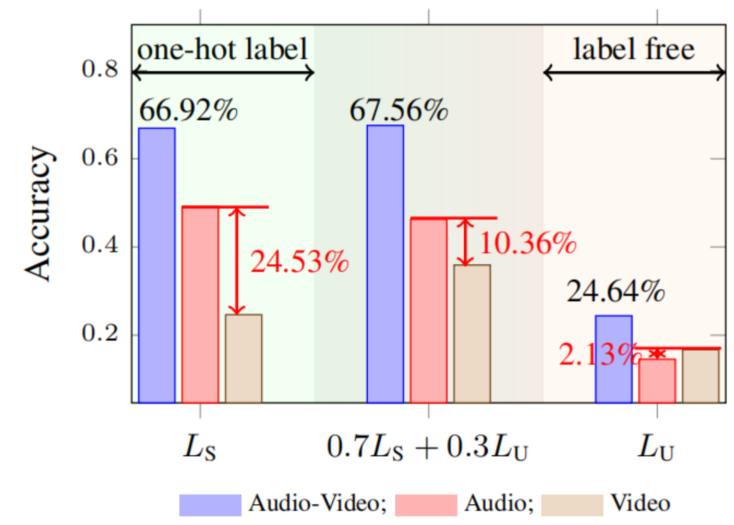



Figure 1: The influence of labels fitting on performance gaps (best view in color), where $L_S$ and $L_U$ denote the loss with one-hot labels and uniform labels (label free).

What are the core causes of modality imbalance?

➢ The key cause of modality imbalance is the bias introduced during label fitting, where over-reliance on one-hot labels amplifies differences in learning dynamics between modalities.

## Proposed Method

Unsupervised contrastive learning

➢ Learn similar representations of different modalities

➢ Cross-modal similarity as a key learning signal, reducing reliance on one-hot labels.

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

Supervised multimodal learning

➢ Focuses on optimizing the **fit** of class labels

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

**Dynamic** integration

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

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

➢ Heuristic: Focus on alignment first, then classification.

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

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

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

## Experiments

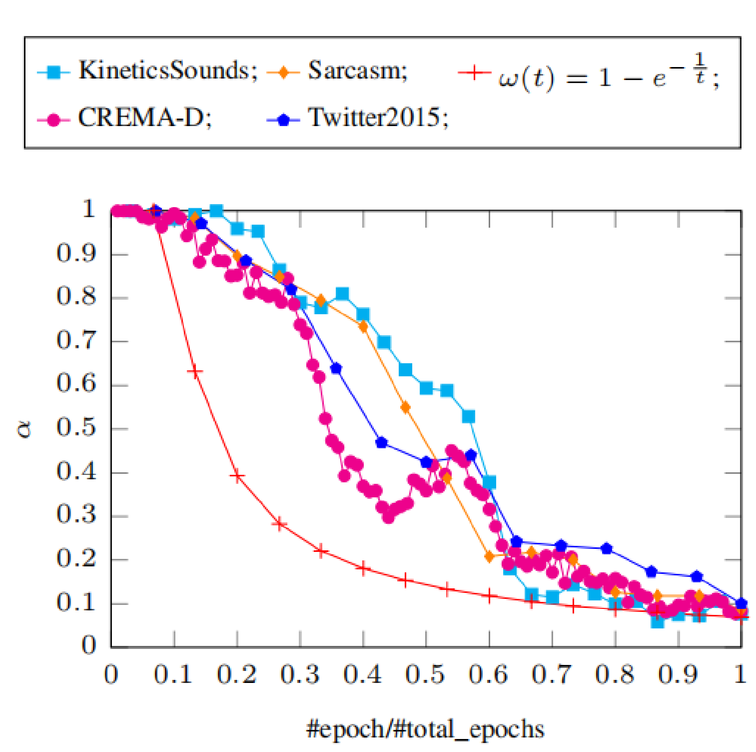
### Classification performance

| 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% | 74.92% | 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% | 79.43% | 85.72% | 84.26% | 83.48% | 73.52% | 67.13% | 83.73% | 83.87% |
| ReconBoost | 70.85% | 74.24% | 74.84% | 81.24% | 84.37% | 83.17% | 74.42% | 68.34% | 84.13% | **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% | 83.98% ±0.22% | 73.87% ±0.35% | 69.17% ±0.26% | 83.24% ±0.07% | 83.87% ±0.18% |
| Ours-LB | **72.53%** ±0.31% | 78.38% ±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% | 84.68% ±0.24% |

From the results, it reveals that:

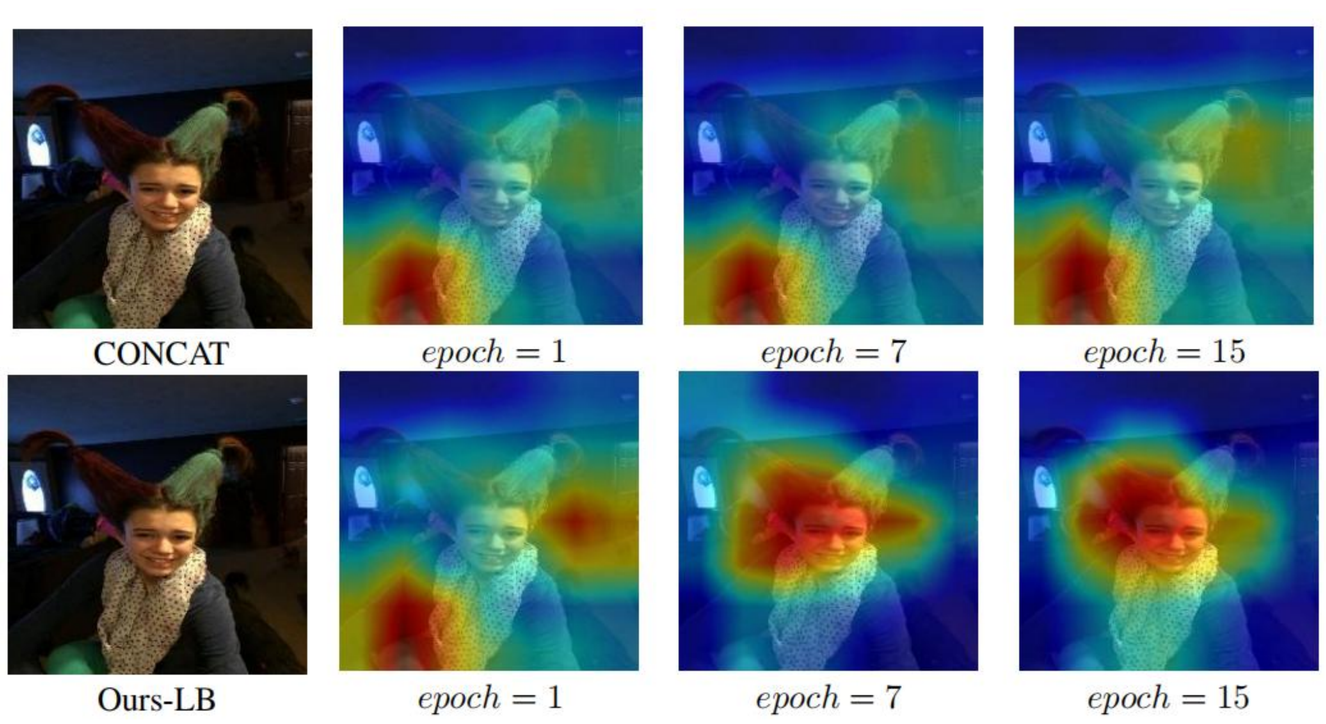
➢ Our learning-based strategy consistently outperforms baselines across datasets.

### Alpha trend




➢ Focus first on alignment and then on classification throughout the training process

### Visualization




➢ 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 key cause of modality imbalance and proposes dynamically combining unsupervised contrastive learning with supervised multimodal learning.
Future work will explore whether some labels inherently favor specific modalities.

## Contact


yyang@njust.edu.cn
fqwan@njust.edu.cn
jiangqy@njust.edu.cn
yxu@dlut.edu.cn