MMCOSINE: MULTI-MODAL COSINE LOSS TOWARDS BALANCED AUDIO-VISUAL FINE-GRAINED LEARNING

Ruize Xu[1], **Ruoxuan Feng**[1], **Shi-xiong Zhang**[2], **Di Hu**[1,*]

[1]GeWu-Lab, Gaoling School of Artificial Intelligence, Renmin University of China, [2]Tencent AI Lab

ICASSP 2023 4 - 10 JUNE, RHODES ISLAND, GREECE

Project Homepage

Background

➢ **Common framework of discriminative multi-modal learning:**

- Mid-concatenation of uni-modal features
- MLP for label-wise logit scores $f(x_i)_j = W_j^{aT}\phi_i^a + W_j^{vT}\phi_i^v + b_j$
- Optimization by Softmax and Cross-entropy

$$L_{vani} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{f(x_i)_{y_i}}}{\sum_{j=1}^{n} e^{f(x_i)_j}}$$

➢ **Phenomenon of imbalance**

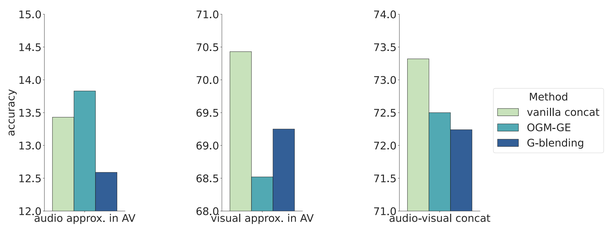
- The uni-modal encoders of the joint model converge at different rates and are under-optimized with a unified objective.
- The potential of the weak modality is not fully exploited.

Limitation of Previous Work

➢ **Previous imbalance-mitigating work**

- OGM-GE: Modality-wise dynamic learning rate
- G-blending: Additional uni-modal classifiers and loss terms

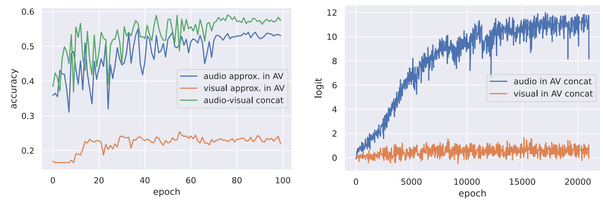
➢ **Failure on fine-grained tasks**



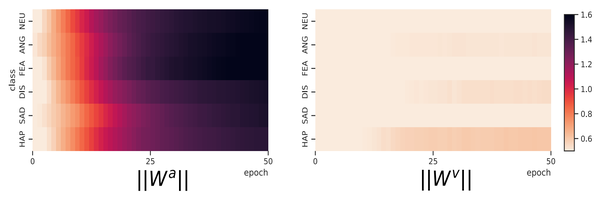
- These methods constrain the gap between uni-modal encoders but fail to enhance the discriminability of the entire model on harder **audio-visual fine-grained tasks**.

Analysis of Imbalance

➢ **Uni-modal performance and logit scores**



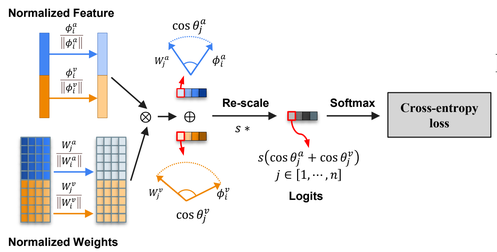
➢ **Uni-modal weight norm**



- One uni-modality dominates the overall model performance and logit scores by its **fast-growing weight norm**.

Multi-modal Cosine Loss

- Modality-wise L_2 normalization → Remove the radial variance
- Rescaling by s → Guarantee the convergence of the network



$$L_{mmcosine} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\cdot(cos\theta_{y_i}^a + cos\theta_{y_i}^v)}}{\sum_{j=1}^{n} e^{s\cdot(cos\theta_j^a + cos\theta_j^v)}}$$

Modality-wise normalization

$$cos\theta_j^a = \frac{W_j^{aT}\phi_i^a}{||W_j^a||\cdot||\phi_i^a||}$$

Lower bound of s

$$s \geq \frac{C-1}{2(C+1)} log\frac{(C-1)p}{1-p}$$

Quantitative Results

➢ **Universal enhancement with various fusion methods**

Method	CREMA-D	SSW60	Voxceleb			
	Top1-Accuracy(%)	Top1-Accuracy(%)	VC1 EER(%)	VC1 minDCF	VC2 EER(%)	VC2 minDCF
Mid-concat	60.08	73.32	6.81	0.578	6.21	0.580
FiLM	59.68	71.67	11.50	0.537	8.31	0.644
Gated	60.48	70.64	10.39	0.567	7.76	0.640
Mid-concat†	63.44	**75.95**	**4.26**	0.461	**4.13**	0.371
FiLM†	61.42	74.30	8.03	0.373	4.58	0.342
Gated†	**66.40**	75.70	5.34	**0.335**	4.30	**0.322**

† means MMCosine is applied.

➢ **Imbalance mitigation**

Metric	Vanilla Softmax	MMCosine
A-probe	56.32	48.66
V-probe	32.26	42.40
A-V gap	24.06	6.26

➢ **Extension to other modalities**

Modality	Softmax+CE	MMCosine
RGB+flow	81.15	**82.02**
RGB+flow+diff	82.29	**83.22**

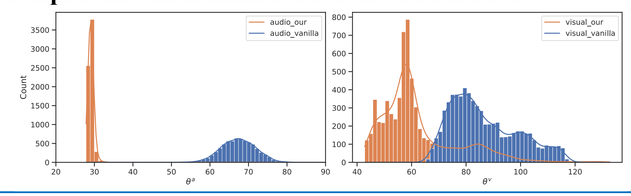
Experiments on coarse-grained dataset UCF-101.

Qualitive Evaluation

➢ **Symmetric constraints on cooperation and discrepancy**

$$\tilde{f}(x_i)_j = cos\theta_j^a + cos\theta_j^v = 2cos(\frac{\theta_j^a + \theta_j^v}{2})\cdot cos(\frac{\theta_j^a - \theta_j^v}{2})$$

➢ **More compact and discriminative feature distribution**



Conclusion

- Explained **imbalance from a view of weight-norm**.
- Proposed a **plug-and-use substitute** for cross-entropy.
- Mitigated the imbalance and boosted the entire joint model.

E-mail: xrz0315@ruc.edu.cn **Lab Page:** https://gewu-lab.github.io