中国人民大学 RENMIN UNIVERSITY OF CHINA

ICASSP 2023 4 - 10 JUNE, RHODES ISLAND, GREECE

高瓴人工智能学院 Gaoling School of Artificial Intelligence

Tencent 腾讯

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

Background

■ Common Framework for discriminative multi-modal learning

- Mid-concatenation of uni-modal features
- MLP for label-wise logit scores
- Optimization by Softmax and Cross-entropy

$f(x_i)_j = W_j^{aT}\phi_i^a + W_j^{vT}\phi_i^v + b_j$ Logit score of sample x_i for label j

$W_j^a \quad W_j^v$ Modality-relevant weight in MLP

$\phi_i^a \quad \phi_i^v$ Uni-modal features

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

Background

■ Imbalance in joint multi-modal learning

- Training multiple modalities jointly with a single optimization objective is **sub-optimal**[1]
- The potential of the weak modality is not fully exploited[2]

■ Previous work

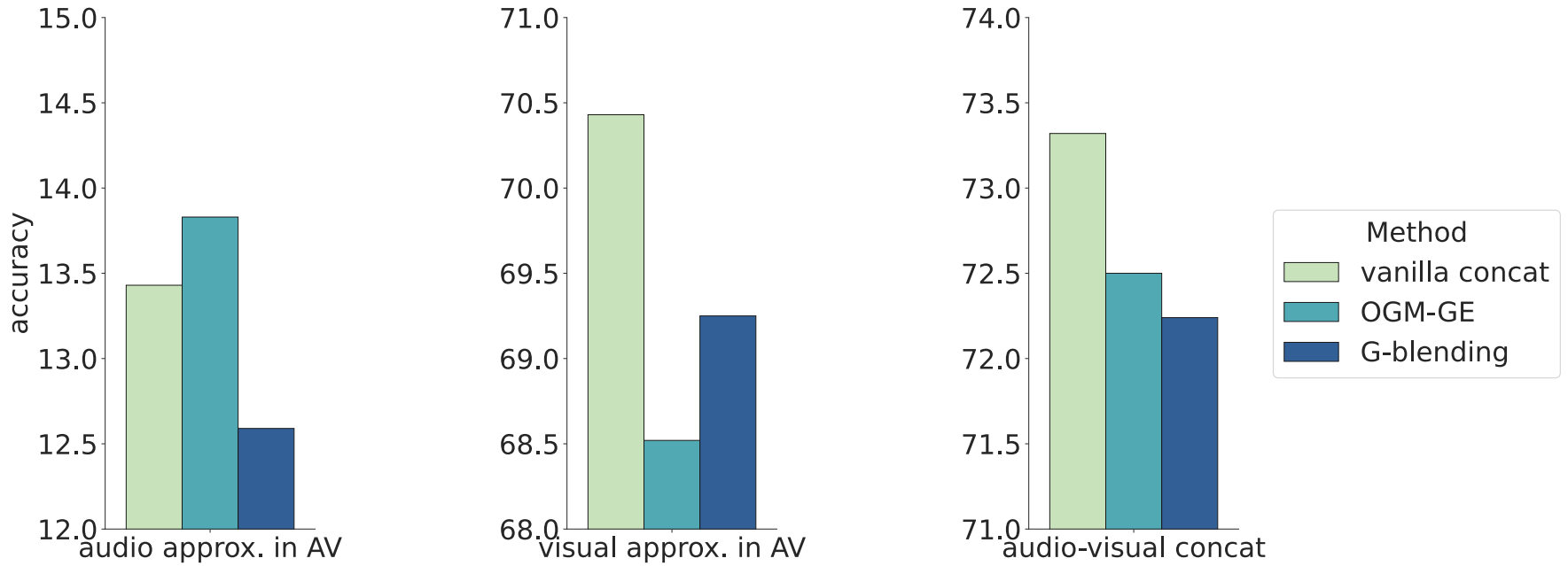
- OGM-GE[2]: Modality-specific dynamic learning rate
- G-blending[1]: Additional uni-modal classifiers and loss terms

[1] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal classification networks hard?," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.
[2] xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8238–8247.

Background

高瓴人工智能学院
Gaoling School of Artificial Intelligence

■ Limitation of Previous work



- These methods fail to enhance the discriminability of the entire model on harder **audio-visual fine-grained tasks**

Motivation

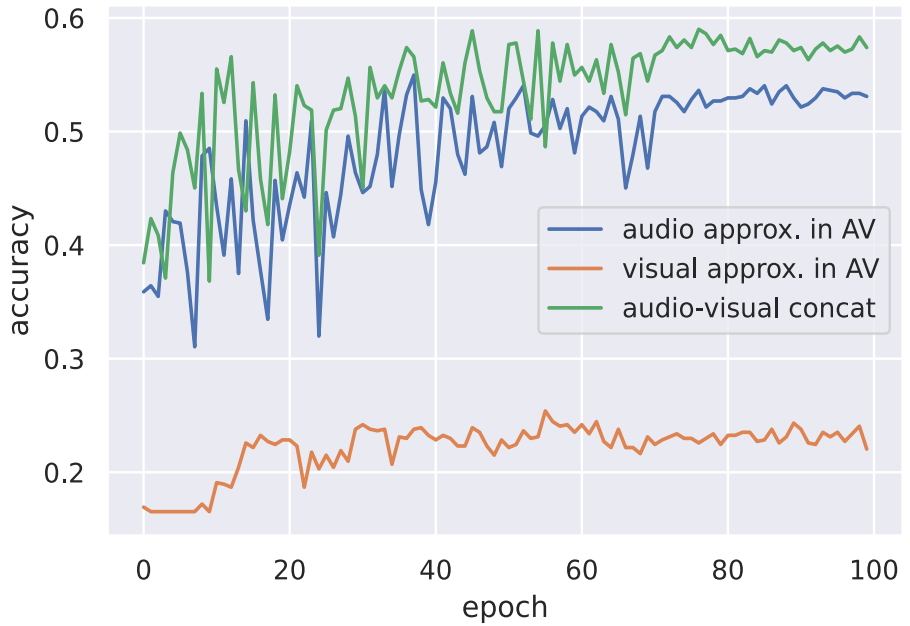
- **Mitigate the imbalance between uni-modalities**

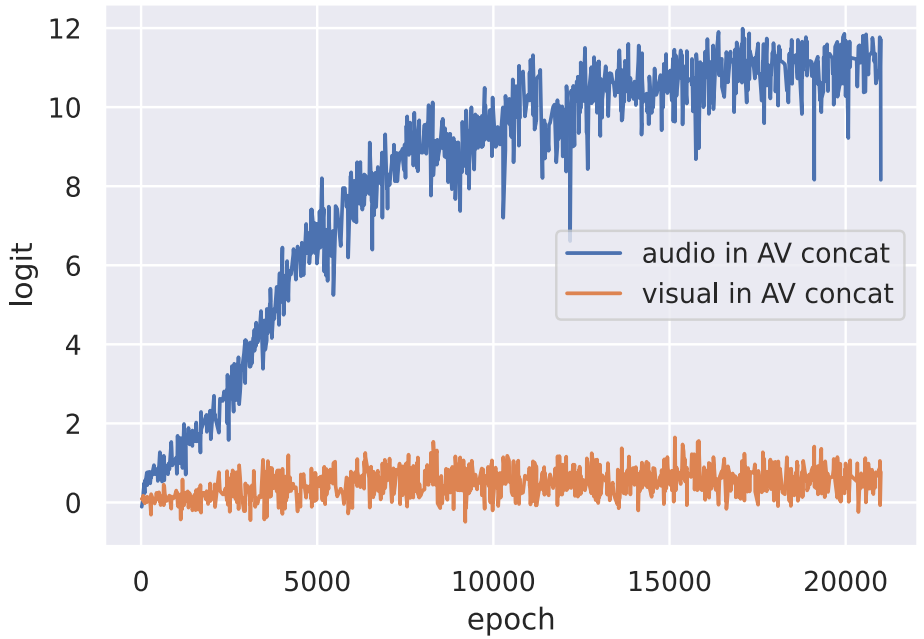
- **Boost the discriminability of the weak modality and the joint model**

Method

高瓴人工智能学院
Gaoling School of Artificial Intelligence

■ Analysis of imbalance from the perspective of weight norm



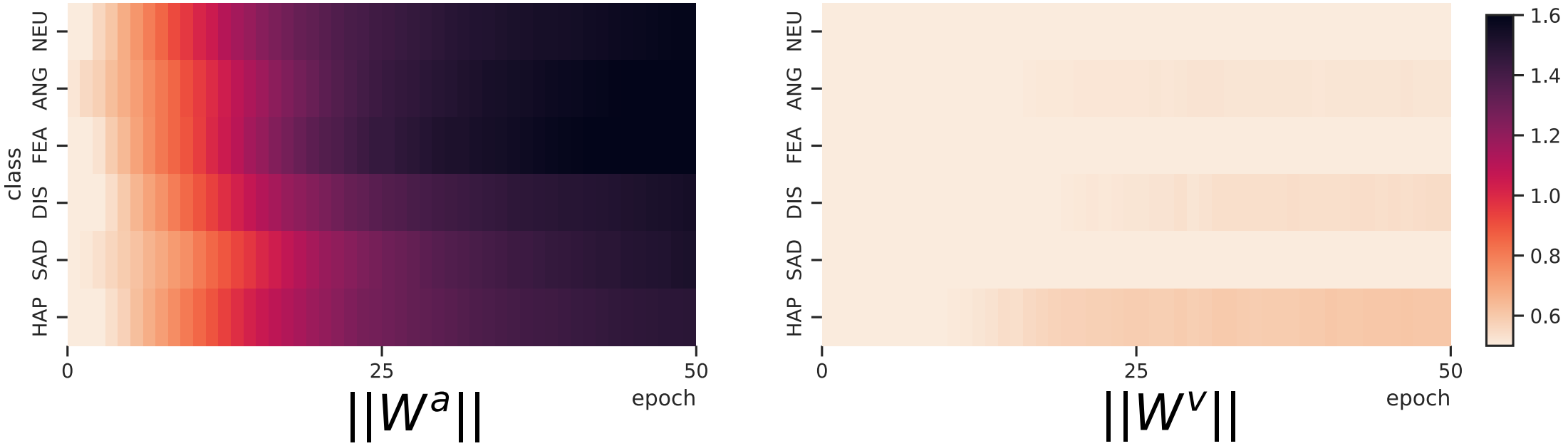


Uni-modal performance and logit scores in the end-to-end trianing

Method

高瓴人工智能学院
Gaoling School of Artificial Intelligence

■ Analysis of imbalance from the perspective of weight norm



$||W^a||$ $||W^v||$

Uni-modal weight norm in the end-to-end training

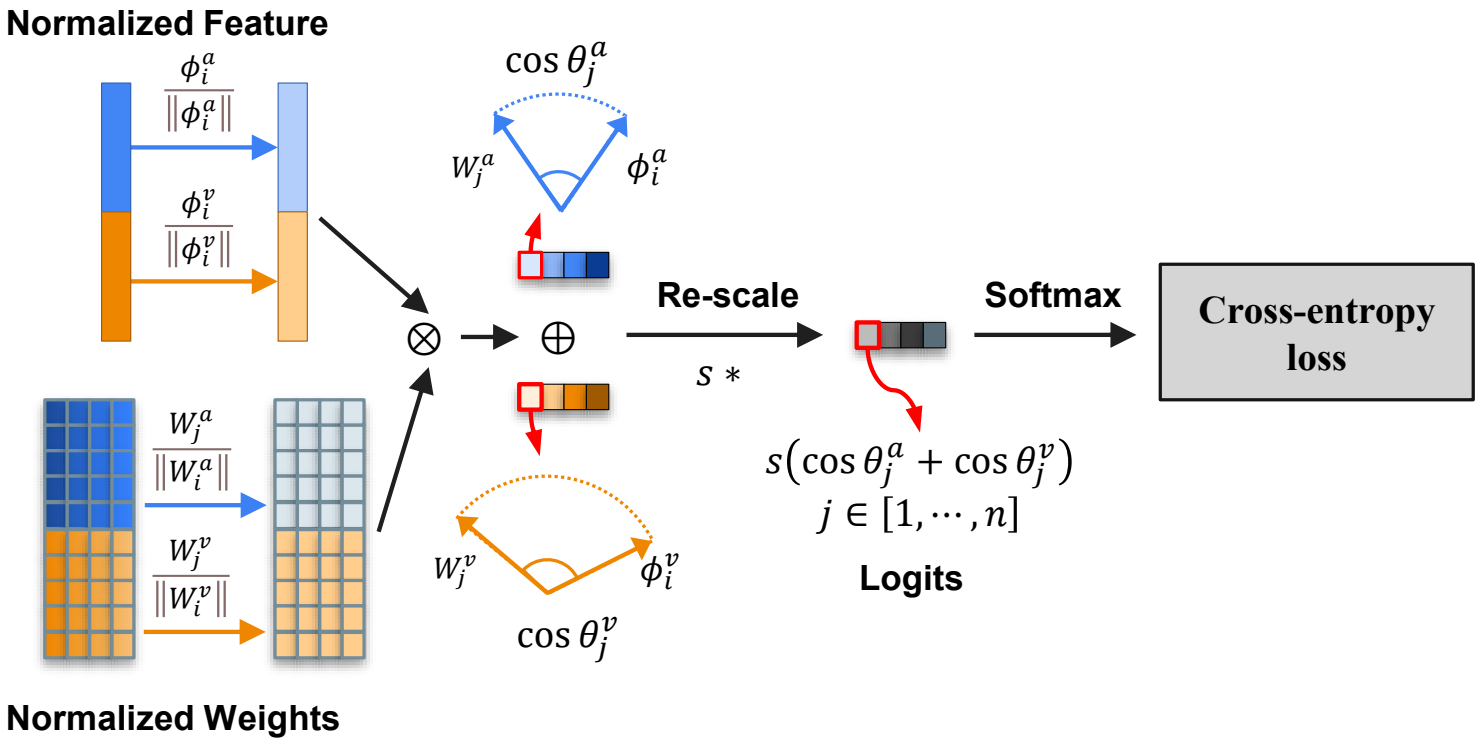
One uni-modality dominates the overall model performance and logit scores by its **fast-growing weight norm**

Method

高瓴人工智能学院
Gaoling School of Artificial Intelligence

■ Multi-modal cosine loss

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



Method

■ Multi-modal cosine loss

Modality-wise L2 normalization

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

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

Lower bound of s

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

Experiments

■ Audio-visual fine-grained tasks and datasets

- Speaker Verification: Voxceleb
- Emotion Recognition: Crema-D
- Bird categorization: SSW60

Uni-modal backbone: Similar Resnet-like network for all branches

Experiments

高瓴人工智能学院
Gaoling School of Artificial Intelligence

■ 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**

Table 1. Performance of various fusion strategies on three AVFG tasks. † indicates MMCosine is applied. Combined with MMCosine, most of the fusion methods gain considerable improvement.

Experiments

➢ 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

Uni-modal accuracy by linear-probing

Method	Vanilla Softmax	MMCosine
Mid-concat	73.32	75.95
OGM-GE	72.50	74.30
G-blending	72.24	74.51

Comparison with previous imbalance-mitigating methods

➢ Extension to other modalities

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

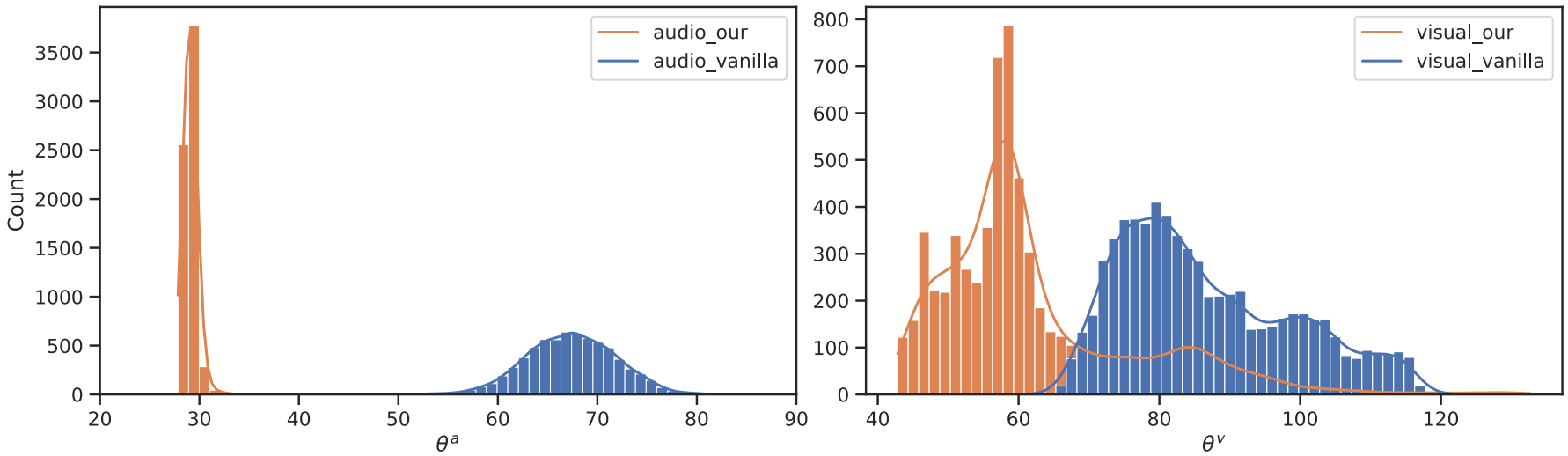
Experiments on coarse-grained dataset UCF-101.

Qualitive Analysis

- **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

高瓴人工智能学院
Gaoling School of Artificial Intelligence

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

Thank You for listening!

Visit the **Project Homepage** for paper, code, and supplementary materials

Contact: xrz0315@ruc.edu.cn

Lab Page: https://gewu-lab.github.io

1. PROOF OF EQUATION 4

Equation 4: The Lower Bound of The Scaling Parameter. Denoting C as the total class number and p as the expected posterior probability for the ground-truth class, the lower bound of s in MMCosine can be given as:

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

We follow the demonstration of [1] in single-modality scenario and hypothesize that the learned features of audio and visual encoder lie on a modality-specific hypersphere. The corresponding weight vectors serve as the learned uni-modal class centers. We denote $\tilde{W}_j = [\tilde{W}_j^a; \tilde{W}_j^v]$ as the weight after modality-wise L_2 normalization. It should be noted that $\tilde{W_j}^T \tilde{W}_j = \tilde{W_j^a}^T \tilde{W}_j^a + \tilde{W_j^v}^T \tilde{W}_j^v = 2$. Denoting p_y as the predicted probability for class center $\tilde{W}_y$ for the ground-truth label y, we have:

$$\begin{aligned} p_y &= \frac{e^{s\tilde{W}_y^T \tilde{W}_y}}{e^{s\tilde{W}_y^T \tilde{W}_y} + \sum_{j \neq y} e^{s(\tilde{W}_y^T \tilde{W}_j)}} \\ &= \frac{e^{2s}}{e^{2s} + \sum_{j \neq y_i} e^{s(\tilde{W}_y^T \tilde{W}_j)}} \end{aligned} \tag{2}$$

Further, to satisfy $p_{y_i} \geq p$, we have:

$$\begin{aligned} 1 + e^{-2s} \sum_{j \neq y_i} e^{s(\tilde{W}_y^T \tilde{W}_j)} &\leq \frac{1}{p}, \\ \sum_{y=1}^{C} (1 + e^{-2s} \sum_{j \neq y} e^{s(\tilde{W}_y^T \tilde{W}_j)}) &\leq \frac{C}{p}, \\ 1 + \frac{e^{-2s}}{C} \sum_{y=1}^{C} \sum_{j \neq y} e^{s(\tilde{W}_y^T \tilde{W}_j)} &\leq \frac{1}{p}. \end{aligned} \tag{3}$$

With Jensen's inequality, we have:

$$\frac{1}{C(C-1)} \sum_{y=1}^{C} \sum_{j \neq y} e^{s(\tilde{W}_y^T \tilde{W}_j)} \geq e^{\frac{s}{C(C-1)} \sum_{y=1}^{C} \sum_{j \neq y} \tilde{W}_y^T \tilde{W}_j}. \tag{4}$$

Then we can simplify (4) further by:

$$\sum_{y=1}^{C} \sum_{j \neq y} \tilde{W}_y^T \tilde{W}_j = (\sum_{y}^{C} \tilde{W}_y)^2 - \sum_{y}^{C} (\tilde{W_y}^2) \geq -4C. \tag{5}$$

$$\frac{1}{C(C-1)} \sum_{y=1}^{C} \sum_{j \neq y} e^{s(\tilde{W}_y^T \tilde{W}_j)} \geq e^{\frac{-4s}{C-1}}. \tag{6}$$

We plug (6) into (3) and get:

$$1 + (C-1) e^{-2s \frac{C+1}{C-1}} \leq \frac{1}{p}. \tag{7}$$

By further simplification, we get the final formulation of the lower bound as:

$$s \geq \frac{C-1}{2(C+1)} log \frac{(C-1)p}{1-p}. \tag{8}$$

This formula provides a theoretical view that the scaling parameter should be enlarged with higher expectation of p and larger class numbers. Considering s as the radius of each hypersphere, larger radius allows features of more labels to distribute in a compact space, which is associated with higher p. It should also be noticed that formula (4) is a loose scaling without constraints to the combined uni-modal weight and might not be the best.