SUPPLEMENTARY MATERIAL FOR 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]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]Tencent AI Lab, Bellevue, WA, USA

ABSTRACT

This supplementary material provides the theoretical details to the derivation of the lower bound of scaling factor s in the multi-modal cosine loss, *i.e.*, Equation 4 in the main paper.

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.

2. SENSITIVITY OF THE SCALING PARAMETER

To reveal the influence of various scaling parameters, we typically use 0.99 as the value of p to calculate the lower bound s. The sensitivity analysis is conducted on CREMAD, SSW60 and Voxceleb2 by merely changing the value of s.

From Tab. 1, our method outperforms the baseline across a wide range of s values above the lower bound, and the optimal value is dependent on specific datasets. On the large-scale dataset Voxceleb2, the performance increases when reducing s, which could be an empirical trick for other datasets.

s	CD ↑	s	SSW ↑	s	VC2 ↓
vanilla	60.08	vanilla	73.32	vanilla	6.13
2	**63.44**	10	73.58	10	**4.13**
5	61.69	20	**75.95**	20	4.91
10	59.82	30	74.87	30	5.50
30	62.09	40	73.85	40	5.91

Table 1. Results of different s.

3. GENERALITY TO OTHER MODALITIES

We propose our method under AVFG tasks and name it MMCosine for its universe form for any modalities. To verify whether our work can be generalized to other multi-modal scenarios, we conduct additional experiments on UCF-101 for coarse-grained action recognition using RGB, flow, and RGB difference [2].

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

Table 2. Results of other modalities on UCF-101.

As shown in Tab. 2, our method yields improvement across various modality combinations, including triple modalities due to its simple and universal form.

4. COMPARISON WITH OTHER MULTI-MODAL LOSS

In the original paper, we compare our method with other multi-modal losses, such as G-blending, which is composed of multiple loss terms. Our method outperforms G-blending both in imbalance mitigation and performance enhancement. We also provide results comparing the multi-modal loss in [3]. It combines classification loss with auxiliary contrastive loss, and we follow the same setting including temperature and weights of loss components.

Loss Method	Vanilla	MMCosine
Softmax+CE	60.08	63.44
CCL [3]	61.56	**64.02**

Table 3. Results of other multi-modal loss on CREMA-D.

From Tab. 3, our method directly surpasses other multi-modal losses and can easily be stacked with them to further boost the performance. It is worth emphasizing that versatility and imbalance mitigation are key aspects of our contributions.

5. REFERENCES

[1] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[2] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu, "Tdn: Temporal difference networks for efficient action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1895–1904.

[3] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata, "Distilling audio-visual knowledge by compositional contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7016–7025.