Emphasizing Closeness and Diversity Simultaneously for Deep Face Representation Supplementary Material

Chaoyu Zhao[1], Jianjun Qian[1(✉)], Shumin Zhu[2], Jin Xie[1], and Jian Yang[1]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{cyzhao,csjqian}@njust.edu.cn

[2] AiDLab, Laboratory for Artificial Intelligence in Design, School of Fashion and Textiles, The Hong Kong Polytechnic University

1 Distribution Visualization Results

From the distribution view, we provide visualization results to demonstrate that our model can achieve both closeness and diversity, as shown in Fig. A1. For the positive distribution (the green one), our model obtains a desirable boundary value, indicating closeness is effectively achieved. The negative distribution (the red one) is compact and close to zero, indicating that we can enforce higher diversity at the same time.

It is hard to achieve both closeness and diversity for the competed methods. The negative distributions of MV-softmax[4] and CurricularFace[1] are also compact; however, the boundary values of MV-softmax and CurricularFace are not high enough to obtain sufficient closeness. For MagFace[2], the negative distribution is not so compact as MV-Softmax and CurricularFace, introducing a higher risk of misclassification.

2 Further Discussion for Diversity Loss

In this part, we present the specific reasons for our diversity loss design. In [3], $\mathcal{L}_{uniform}$ is proposed to preserve the maximal information in the feature space. However, the original loss relies on a large batch size(i.e., 768) to achieve desirable results. Therefore, we convert the comparison between sample pairs into the comparison between class centers and samples to save graphic memory.

The exponent of Eq. 7 contains two different components: the truncated scale term $s \cdot \max(0, \operatorname{sgn}(\cos\theta))$ and the similarity term $(\cos\theta)^2$. In the following part, we will explain why we employ $(\cos\theta)^2$ as the similarity term instead of other formats(e.g., $-\|x-y\|_2^2$ or $\cos\theta$) and why we use the truncated scale term in the diversity loss. Given only one sample and the class center, gradient of diversity loss is degraded as follows:

$$\frac{\partial \mathcal{L}_{diversity}}{\partial \boldsymbol{x}} = T(s, \boldsymbol{x}, \boldsymbol{w}) \cdot \frac{\partial D(\boldsymbol{x}, \boldsymbol{w})}{\partial \boldsymbol{x}} \tag{1}$$

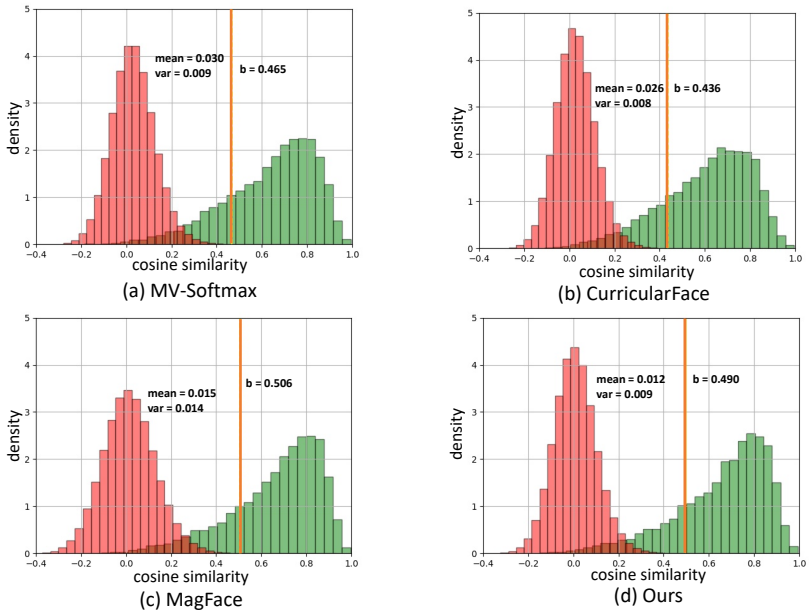

Fig. A1. Visualization results on the distribution view.

where $\boldsymbol{x}$ and $\boldsymbol{w}$ are the normalized feature and class center, s is the scale term. $D(\boldsymbol{x}, \boldsymbol{w}) = (\cos\theta)^2$, where θ is the angular distance between $\boldsymbol{x}$ and $\boldsymbol{w}$. $T(\cdot)$ is the truncated scale term formulated as follows:

$$T(s, \boldsymbol{x}, \boldsymbol{w}) = \begin{cases} s, & \boldsymbol{w}^T\boldsymbol{x} > 0 \\ 0, & \boldsymbol{w}^T\boldsymbol{x} \leq 0 \end{cases} \tag{2}$$

Intuitively, a desirable diversity loss function should satisfy the following property: the gradient should get magnified when $\boldsymbol{x}$ is closer to $\boldsymbol{w}$. Let us set $D_1(\boldsymbol{x}, \boldsymbol{w}) = -\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$, $D_2(\boldsymbol{x}, \boldsymbol{w}) = \cos\theta$, $D_3(\boldsymbol{x}, \boldsymbol{w}) = (\cos\theta)^2$. Their gradients are calculated as follows:

$$\boldsymbol{g}_1 = \frac{\partial D_1(\boldsymbol{x}, \boldsymbol{w})}{\partial \boldsymbol{x}} = 2 \cdot (\boldsymbol{w} - \boldsymbol{x}) \tag{3}$$

$$\boldsymbol{g}_2 = \frac{\partial D_2(\boldsymbol{x}, \boldsymbol{w})}{\partial \boldsymbol{x}} = \boldsymbol{w} \tag{4}$$

$$\boldsymbol{g}_3 = \frac{\partial D_3(\boldsymbol{x}, \boldsymbol{w})}{\partial \boldsymbol{x}} = 2\cos\theta \cdot \boldsymbol{w} \tag{5}$$

As shown in Fig. A2, $\|\boldsymbol{g}_1\|$ gets larger as θ increases and $\|\boldsymbol{g}_2\|$ is identical when θ varies from 0 to π. Therefore, both $D_1 = -\|\boldsymbol{x} - \boldsymbol{w}\|_2^2$ and $D_2 = \cos\theta$ are unsuitable in diversity loss. By contrast, $\|\boldsymbol{g}_3\|$ meets the demand when $\theta < \frac{\pi}{2}$, but it still remains a considerable value when $\theta \geq \frac{\pi}{2}$. Therefore, we further

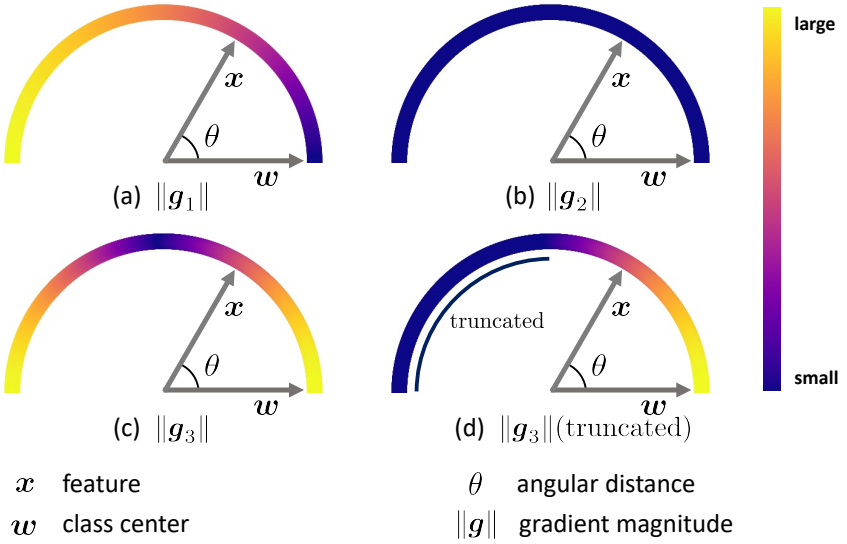

Fig. A2. The comparison of gradient magnitudes among D_1(a), D_2(b), D_3(c) and the truncated version of D_3(d). (Best view in colors)

employ $T(\cdot)$ to truncate the gradient when θ is greater than $\frac{\pi}{2}$ for training stability.

3 Further discussion on λ_1 and λ_2

Table A1. Verification comparisons on different benchmarks. We conduct the experiment on the MS1MV2's subset containing 10K unique identities with ResNet18.

Models	Settings		Verification Accuracy			IJB	
	λ_1	λ_2	LFW	CFP-FP	AgeDB	IJB-B	IJB-C
1	0.5	2.0	99.28	88.59	94.10	83.09	86.32
2	0.1	2.0	99.15	88.64	94.06	83.12	86.33
3	2.0	2.0	99.40	88.43	93.85	82.55	86.02
4	5.0	2.0	99.37	88.52	93.88	82.13	85.89
5	0.5	1.0	99.25	88.53	94.12	82.85	86.10
6	0.5	5.0	99.15	88.70	94.17	83.15	86.37
7	0.5	10.0	99.12	88.61	94.08	83.21	86.40

In our Soft Mining Scheme(SMS), we set $\lambda_1 = 0.5$, $\lambda_2 = 2.0$. In Table A1, we vary λ_1 and λ_2 in a certain range and provide the results on several bench-

marks. In Model 2-4, we adjust λ_1 and fix λ_2. With the increase of λ_1, the performance increases on LFW but degrades significantly on IJB-B/C. In Model 5-7, we adjust λ_2 and fix λ_1. Although the performance on IJB gets better as λ_2 increases, it sightly degrades on LFW and AgeDB when λ_2 reaches a relatively large value(e.g., 10.0).

References

1. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
2. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14234 (2021)
3. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
4. Wang, X., Zhang, S., Wang, S., Fu, T., Shi, H., Mei, T.: Mis-classified vector guided softmax loss for face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12241–12248 (2020)