MagFace: A Universal Representation for Face Recognition and Quality Assessment (Supplementary Material)

Qiang Meng, Shichao Zhao, Zhida Huang, Feng Zhou
Algorithm Research, Aibee Inc.
{qmeng,sczhao,zdhuang,fzhou}@aibee.com

A. Proofs for MagFace

Recall the MagFace loss for a sample i is

$$L_i = -\log \frac{e^{s\cos\left(\theta_{y_i}+m(a_i)\right)}}{e^{s\cos\left(\theta_{y_i}+m(a_i)\right)}+\sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}} + \lambda_g g(a_i) \tag{1}$$

Let $A(a_i) = s\cos(\theta_{y_i} + m(a_i))$ and $B = \sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}$ and rewrite the loss as

$$L_i = -\log \tfrac{e^{A(a_i)}}{e^{A(a_i)}+B} + \lambda_g g(a_i) \tag{2}$$

We first introduce and prove Lemma 1.

Lemma 1. *Assume that f_i is top-k correctly classified and $m(a_i) \in [0, \pi/2]$. If the number of identities n is much larger than k (i.e., $n \gg k$), the probability of $\theta_{y_i} + m(a_i) \in [0, \pi/2]$ approaches 1.*

Proof. Denote the angle between feature f_i and center class $W_j, j \in \{1, \cdots, n\}$ as θ_j. Assuming the distribution of θ_j is uniform, it's easy to prove $P\left(\theta_j + m(a_i) \in [0, \pi/2]\right) = \frac{\pi/2-m(a_i)}{\pi}$. Let $p = \frac{\pi/2-m(a_i)}{\pi}$. If f_i is top-k correctly classified, the probability of $\theta_{y_i} + m(a_i) \in [0, \pi/2]$ is the same as the probability of there are at least k θ to satisfy $\theta + m(a_i) \in [0, \pi/2]$. Then the probability is

$$\begin{aligned} P\left(\theta_{y_i} + m(a_i) \in [0, \pi/2]\right) &= \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{(n-i)} o \\ &= 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{(n-i)} \end{aligned} \tag{3}$$

When n is a large integer and $n \gg k$, each $\binom{n}{i} p^i (1-p)^{(n-i)}, i = 1, 2, \cdots k-1$ converges to 0. Therefore, probability of $\theta_{y_i} + m(a_i) \in [0, \pi/2]$ approaches 1. □

Lemma 1 is fundamental for the following proofs. The number of identities is large in real-world applications (*e.g.*, 3.8M for MS1Mv2 [2, 1]). Therefore, the probability of $\theta_{y_i} + m(a_i) \in [0, \pi/2]$ approaches 1 in most cases.

A.1. Requirements for MagFace

In MagFace, $m(a_i), g(a_i), \lambda_g$ are required to have the following constraints:

1. $m(a_i)$ is an increasing convex function in $[l_a, u_a]$ and $m'(a_i) \in (0, K]$, where K is a upper bound;
2. $g(a_i)$ is a strictly convex function with $g'(u_a) = 0$;
3. $\lambda_g \geq \frac{sK}{-g'(l_a)}$.

A.2. Proof for Property of Convergence

We prove the property of convergence by showing the strictly convexity of the function L_i (Property 1) and the existence of the optimum (Property 2).

Property 1. *For $a_i \in [l_a, u_a]$, L_i is a strictly convex function of a_i.*

Proof. The first and second deriviates of $A(a_i)$ are

$$\begin{aligned} A'(a_i) &= -s\sin(\theta_{y_i} + m(a_i))m'(a_i) \\ A''(a_i) &= -s\cos(\theta_{y_i} + m(a_i))(m'(a_i))^2 \\ &\quad - s\sin(\theta_{y_i} + m(a_i))m''(a_i) \end{aligned} \tag{4}$$

According to Lemma 1, we have $\cos(\theta_{y_i} + m(a_i)) \geq 0$ and $\sin(\theta_{y_i} + m(a_i)) \geq 0$. Because we define $m(a_i)$ to be convex and $g(a_i)$ to be strictly convex when $a_i \in [l_a, u_a]$, $m''(a_i) \geq 0$ and $g''(a_i) > 0$ always hold. Therefore, $A''(a_i) \leq 0$.

The first and second order derivatives of the loss L_i are

$$\begin{aligned} \frac{\partial L_i}{\partial a_i} &= -\frac{B}{e^{A(a_i)}+B}A'(a_i) + \lambda_g g'(a_i) \\ \frac{\partial^2 L_i}{(\partial a_i)^2} &= -\frac{B}{(e^{A(a_i)}+B)^2}\left((e^{A(a_i)}+B)A''(a_i) - Be^{A(a_i)}A'(a_i)^2\right) \\ &\quad + \lambda_g g''(a_i) \\ &= -\frac{B}{e^{A(a_i)}+B}A''(a_i) + \frac{B^2}{(e^{A(a_i)}+B)^2}e^{A(a_i)}A'(a_i)^2 \\ &\quad + \lambda_g g''(a_i) \end{aligned}$$

Method	Hyperparameters					Margin			CFP-FP	IJB-C (TAR@FAR)			
	l_m	u_m	λ_g	l_a	u_a	mean	max	min		1e-6	1e-5	1e-4	1e-3
ArcFace	-	-	-	-	-	0.50	-	-	97.32	83.88	91.59	95.00	96.86
MagFace	0.45	0.65	35	10	110	0.50	0.49	0.52	97.23	81.12	91.44	94.95	96.96
	0.40	0.80	35	10	110	0.50	0.46	0.53	**97.47**	**85.82**	**92.06**	**95.12**	96.92
	0.35	1.00	35	10	110	0.50	0.42	0.54	97.40	84.35	91.65	95.05	**97.02**
	0.25	1.60	35	10	110	0.50	0.35	0.61	97.30	81.64	91.09	94.91	96.87

Table A1: Verification accuracy (%) on CFP-FP and IJB-C with different ditributions of margins. Backbone network: ResNet50.

It's easy to prove that $B > 0, e^{A(a_i)} + B > 0$, the first two part of $\frac{\partial^2 L_i}{(\partial a_i)^2}$ is non-negative while the third part is always positive. Therefore, $\frac{\partial^2 L_i}{(\partial a_i)^2} > 0$ and L_i is a strictly convex function with respect to a_i. □

Property 2. *A unique optimal solution a_i^* exists in $[l_a, u_a]$.*

Proof. Because the loss function L_i is a strictly convex function, we have $\frac{\partial L_i}{\partial a_i^1} > \frac{\partial L_i}{\partial a_i^2}$ if $u_a \geq a_i^1 > a_i^2 \geq l_a$. Next we prove that there exist a optimal solution $a_i^* \in [l_a, u_a]$. If it exists, then it is unique because of the strictly convexity.

As $\frac{\partial L_i}{\partial a_i}(a_i) = \frac{Bs}{e^{A(a_i)}+B} \sin(\theta_{y_i} + m(a_i))m'(a_i) + \lambda_g g'(a_i)$ and considering the constraints $m'(a_i) \in (0, K]$, $g'(u_a) = 0$, $\lambda_g \geq \frac{sK}{-g'(l_a)}$, the values of derivatives of l_a, u_a are

$$\begin{aligned}
\frac{\partial L_i}{\partial a_i}(u_a) &= \frac{Bs}{e^{A(a_i)} + B} \sin(\theta_{y_i} + m(a_i))m'(u_a) > 0 \\
\frac{\partial L_i}{\partial a_i}(l_a) &= \frac{Bs}{e^{A(a_i)} + B} \sin(\theta_{y_i} + m(a_i))m'(l_a) + \lambda_g g'(l_a) \\
&< sK + \lambda_g g'(l_a) \leq 0
\end{aligned} \tag{5}$$

As $\frac{\partial L_i}{\partial a_i}$ is monotonically and strictly increasing, there must exist a unique value in $[l_a, u_a]$ which have a 0 derivative. Therefore, a optimal solution exists and is unique. □

A.3. Proof for Property of Monotonicity

To prove the property of monotonicity, we first show that optimal a_i^* increases with a samller cosine-distance to its class center (Property 3). As B can reveal the overall cos-distances to other class centers, we further prove that increaing B can lead to a larger optimal feature magnitude (Property 4).

Property 3. *With fixed f_i and $W_j, j \in \{1, \cdots, n\}, j \neq y_i$, the optimal feature magnitude a_i^* is monotonically decreasing if the cosine-distance to its class center W_{y_i} increases.*

Proof. Assuming there are two class center $W_{y_i}^1, W_{y_i}^2$ and their cosine distance with feature f_i are $\theta_{y_i}^1, \theta_{y_i}^2$. Assuming $\theta_{y_i}^1 < \theta_{y_i}^2$ (*i.e.*, class center $W_{y_i}^1$ has a smaller distance with feature f_i) and the corresponding optimal feature magnitudes are $a_{i,1}^*, a_{i,2}^*$.

The first deriviate of L_i is

$$\begin{aligned}
\frac{\partial L_i}{\partial a_i} &= -\frac{B}{e^{A(a_i)} + B} A'(a_i) + \lambda_g g'(a_i) \\
&= \frac{Bsm'(a_i)}{e^{s\cos(\theta_{y_i} + m(a_i))} + B} \sin(\theta_{y_i} + m(a_i)) + \lambda_g g'(a_i)
\end{aligned} \tag{6}$$

For $\theta_{y_i} + m(a_i) \in (0, \pi/2]$, we have $\cos(\theta_{y_i}^1 + m(a_i)) > \cos(\theta_{y_i}^2 + m(a_i))$ and $\sin(\theta_{y_i}^1 + m(a_i)) < \sin(\theta_{y_i}^2 + m(a_i))$. With $m'(a_i) > 0$, it's obvious that

$$\frac{Bsm'(a_i)}{e^{s\cos(\theta_{y_i}^1 + m(a_i))} + B} \sin(\theta_{y_i}^1 + m(a_i)) < \frac{Bsm'(a_i)}{e^{s\cos(\theta_{y_i}^2 + m(a_i))} + B} \sin(\theta_{y_i}^2 + m(a_i)).$$

Therefore, we have $\frac{\partial L_i(\theta_{y_i}^1)}{\partial a_i} < \frac{\partial L_i(\theta_{y_i}^2)}{\partial a_i}$. Based on the property of optimal solution for strictly convex function, we have $0 = \frac{\partial L_i(\theta_{y_i}^1)}{\partial a_{i,1}^*} = \frac{\partial L_i(\theta_{y_i}^2)}{\partial a_{i,2}^*} > \frac{\partial L_i(\theta_{y_i}^1)}{\partial a_{i,2}^*}$, which leads to $a_{i,1}^* > a_{i,2}^*$. □

Property 4. *With other things fixed, the optimal feature magnitude a_i^* is monotonically decreasing with a decreasing inter-class distance B.*

Proof. Assume $B_1 > B_2 > 0$ with optimal $a_{i,1}^*, a_{i,2}^*$. Similar to the proof before, we have

$$\frac{B_1 sm'(a_i)}{e^{s\cos(\theta_{y_i} + m(a_i))} + B_1} \sin(\theta_{y_i} + m(a_i)) > \frac{B_2 sm'(a_i)}{e^{s\cos(\theta_{y_i} + m(a_i))} + B_2} \sin(\theta_{y_i} + m(a_i)).$$

Therefore, we have $\frac{\partial L_i(B_1)}{\partial a_i} > \frac{\partial L_i(B_2)}{\partial a_i}$. Based on the property of optimal solution for strictly convex function, we have $0 = \frac{\partial L_i(B_1)}{\partial a_{i,1}^*} = \frac{\partial L_i(B_2)}{\partial a_{i,2}^*} < \frac{\partial L_i(B_1)}{\partial a_{i,2}^*}$, which leads to $a_{i,1}^* < a_{i,2}^*$. □

B. Experimental Settings

B.1. Training settings for Figure 3

We adpot ResNet50 as the backbone network. Models are trained on MS1Mv2 [2, 1] for 20 epochs with batch size 512 and initial learning rate 0.1, dropped by 0.1 every 5 epochs. 512 samples of the last iteration are used for visualization.

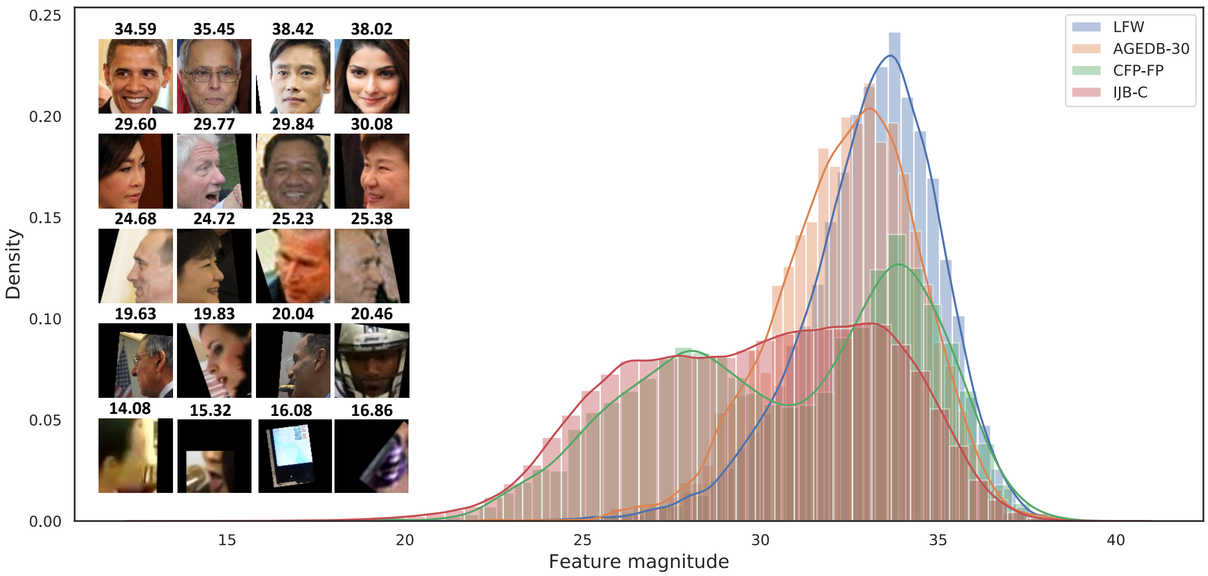

Figure A1: Extended Visualization of Figure 6.

B.2. Settings of $m(a_i)$, $g(a_i)$ and λ_g

In our experiments, we define function $m(a_i)$ as a linear function defined on $[l_a, u_a]$ with $m(l_a) = l_m, m(u_a) = u_m$ and $g(a_i) = \frac{1}{a_i} + \frac{1}{u_a^2} a_i$. Therefore, we have

$$\begin{aligned} m(a_i) &= \frac{u_m - l_m}{u_a - l_a}(a_i - l_a) + l_m \\ \lambda_g &\geq \frac{sK}{-g'(l_a)} = \frac{s u_a^2 l_a^2}{(u_a^2 - l_a^2)} \frac{u_m - l_m}{u_a - l_a} \end{aligned} \tag{7}$$

C. Ablation Study on Margin Distributions

In this section, effects of the feature distributions during training are studied. With (λ_g, l_a, u_a) fixed to $(35, 10, 110)$, we carefully select various combinations of l_m, u_m to align the mean margin on the training dataset to ArcFace (0.5) in our implementation. Features are distributed more separated if with a larger maximum margin and a smaller minimum margin.

Table A1 shows the recognition results with various hyperparameters. With $(l_m, u_m) = (0.45, 0.65)$, the penalty of magnitude loss degrades the performance of the recognition. With $(l_m, u_m) = (0.25, 1.60)$, the performance is also worse than then baseline as hard samples are assigned to small margins (*a.k.a.*, hard/noisy samples are down-weighted). Parameter $(0.40, 0.80)$ balances the feature distribution and margins for hard/noisy samples, and therefore achieves a significant improvement on benchmarks.

D. Extended Visualization of Figure 6

We present a extended visualization of figure 6 in figure A1 which has more examples of faces with feature magnitudes. All the faces are sample from the IJB-C benchmark. It can be seen that faces with magnitudes around 28 are mostly profile faces while around 35 are high-quality and frontal faces. That is consistent with the profile/frontal peaks in the CFP-FP benchmark and indicates that faces with similar magnitudes show similar quality patterns across benchmarks. In real applications, we can set a proper threshold for the magnitude and should be able to filter similar low-quality faces, even under various scenarios.

Besides directly served as qualities, our feature magnitudes can also be used as quality labels for faces, which avoids human labelling costs. These labels are more suitable for recognition, and therefore can be used to boost other quality models.

E. Authors' Contributions

Shichao Zhao and Zhida Huang contribute similarly to this work. Besides involved in discussions and help polish the work, shichao zhao mainly conductsed experiments on clustering and zhida huang implemented baselines as well as evaluation metrics in quality experiments.

References

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 2

[2] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 1, 2