Supplementary Material for CosFace: Large Margin Cosine Loss for Deep Face Recognition

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou,
Zhifeng Li*, and Wei Liu*

Tencent AI Lab

{hawelwang,yitongwang,encorezhou,denisji,sagazhou,michaelzfli}@tencent.com

gongdihong@gmail.com wliu@ee.columbia.edu

This supplementary document provides mathematical details for the derivation of the lower bound of the scaling parameter s (Equation 6 in the main paper), and the variable scope of the cosine margin m (Equation 7 in the main paper).

Proposition of the Scaling Parameter s

Given the normalized learned features x and unit weight vectors W, we denote the total number of classes as C where $C > 1$. Suppose that the learned features separately lie on the surface of a hypersphere and center around the corresponding weight vector. Let P_w denote the expected minimum posterior probability of the class center (*i.e.*, W). The lower bound of s is formulated as follows:

$$s \geq \frac{C-1}{C} \ln \frac{(C-1)P_W}{1-P_W}$$

Proof:

Let W_i denote the i-th unit weight vector. $\forall i$, we have:

$$\frac{e^s}{e^s + \sum_{j,j\neq i} e^{s(W_i^T W_j)}} \geq P_W, \tag{8}$$

$$1 + e^{-s} \sum_{j,j\neq i} e^{s(W_i^T W_j)} \leq \frac{1}{P_W}, \tag{9}$$

$$\sum_{i=1}^{C} (1 + e^{-s} \sum_{j,j\neq i} e^{s(W_i^T W_j)}) \leq \frac{C}{P_W}, \tag{10}$$

$$1 + \frac{e^{-s}}{C} \sum_{i,j,i\neq j} e^{s(W_i^T W_j)} \leq \frac{1}{P_W}. \tag{11}$$

Because $f(x) = e^{s \cdot x}$ is a convex function, according to Jensen's inequality, we obtain:

$$\frac{1}{C(C-1)} \sum_{i,j,i\neq j} e^{s(W_i^T W_j)} \geq e^{\frac{s}{C(C-1)} \sum_{i,j,i\neq j} W_i^T W_j}. \tag{12}$$

Besides, it is known that

$$\sum_{i,j,i\neq j} W_i^T W_j = (\sum_i W_i)^2 - (\sum_i W_i^2) \geq -C. \tag{13}$$

Thus, we have:

$$1 + (C-1)e^{-\frac{sC}{C-1}} \leq \frac{1}{P_W}. \tag{14}$$

Further simplification yields:

$$s \geq \frac{C-1}{C} \ln \frac{(C-1)P_W}{1-P_W}. \tag{15}$$

The equality holds if and only if every $W_i^T W_j$ is equal ($i \neq j$), and $\sum_i W_i = 0$. Because at most $K+1$ unit vectors are able to satisfy this condition in the K-dimension hyper-space, the equality holds only when $C \leq K+1$, where K is the dimension of the learned features.

Proposition of the Cosine Margin m

Suppose that the weight vectors are uniformly distributed on a unit hypersphere. The variable scope of the introduced cosine margin m is formulated as follows :

$$0 \leq m \leq 1 - \cos\frac{2\pi}{C}, \quad (K = 2)$$

$$0 \leq m \leq \frac{C}{C-1}, \quad (K > 2, C \leq K+1)$$

$$0 \leq m \ll \frac{C}{C-1}, \quad (K > 2, C > K+1)$$

*Corresponding authors

where C is the total number of training classes and K is the dimension of the learned features.

Proof:

For $K = 2$, the weight vectors uniformly spread on a unit circle. Hence, $\max(W_i^T W_j) = \cos\frac{2\pi}{C}$. It follows $0 \leq m \leq (1 - \max(W_i^T W_j)) = 1 - \cos\frac{2\pi}{C}$.

For $K > 2$, the inequality below holds:

$$\begin{aligned} C(C-1)\max(W_i^T W_j) &\geq \sum_{i,j,i\neq j} W_i^T W_j \\ &= (\sum_i W_i)^2 - (\sum_i W_i^2) \\ &\geq -C. \end{aligned} \tag{16}$$

Therefore, $\max(W_i^T W_j) \geq \frac{-1}{C-1}$, and we have $0 \leq m \leq (1 - \max(W_i^T W_j)) \leq \frac{C}{C-1}$.

Similarly, the equality holds if and only if every $W_i^T W_j$ is equal ($i \neq j$), and $\sum_i W_i = 0$. As discussed above, this is satisfied only if $C \leq K + 1$. On this condition, the distance between the vertexes of two arbitrary W should be the same. In other words, they form a regular simplex such as an equilateral triangle if $C = 3$, or a regular tetrahedron if $C = 4$.

For the case of $C > K + 1$, the equality cannot be satisfied. In fact, it is unable to formulate the strict upper bound. Hence, we obtain $0 \leq m \ll \frac{C}{C-1}$. Because the number of classes can be much larger than the feature dimension, the equality cannot hold in practice.