Learning Towards the Largest Margins

Xiong Zhou[1], Xianming Liu[1 2], Deming Zhai[1], Junjun Jiang[1 2], Xin Gao[3 2 4], Xiangyang Ji[5]

[1] *Harbin Institute of Technology*

[2] *Peng Cheng Laboratory*

[3] *King Abdullah University of Science and Technology*

[5] *Gaoling School of Artificial Intelligence, Renmin University of China*

[5] *Tsinghua University*

ICLR | 2022

Tenth International Conference on Learning Representations

Motivations

- One of the main challenges for feature representation in deep learning-based classification is the design of appropriate loss functions that exhibit strong discriminative power.
- The classical softmax loss does not explicitly encourage discriminative features.

Motivations

- One of the main challenges for feature representation in deep learning-based classification is the design of appropriate loss functions that exhibit strong discriminative power.
- The classical softmax loss does not explicitly encourage discriminative features.
- A popular direction of research is to incorporate margins in well-established losses.
- We explain the margin-based losses by formulating it as *learning towards the largest margins*.

Motivations

- One of the main challenges for feature representation in deep learning-based classification is the design of appropriate loss functions that exhibit strong discriminative power.
- The classical softmax loss does not explicitly encourage discriminative features.
- A popular direction of research is to incorporate margins in well-established losses.
- We explain the margin-based losses by formulating it as *learning towards the largest margins*.
- In this work, we introduce two measures: class margin and sample margin.
- The loss function should promote the largest possible margins for both classes and samples.

Motivations

- One of the main challenges for feature representation in deep learning-based classification is the design of appropriate loss functions that exhibit strong discriminative power.
- The classical softmax loss does not explicitly encourage discriminative features.
- A popular direction of research is to incorporate margins in well-established losses.
- We explain the margin-based losses by formulating it as *learning towards the largest margins*.
- In this work, we introduce two measures: class margin and sample margin.
- The loss function should promote the largest possible margins for both classes and samples.
- Furthermore, we derive a generalized margin softmax loss to draw general conclusions for the existing margin-based losses, which can also guide the design of new tools, including *sample margin regularization* and *largest margin softmax loss* for class-balanced cases, and *zero-centroid regularization* for class-imbalanced cases.

The Softmax Loss

- With a Labeled dataset $D = \{(x_i, y_i)\}_{i=1}^N$, the softmax loss for a k-classification problem is formulated as

$$L = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{\exp(w_{y_i}^T z_i)}{\sum_{j=1}^{k}\exp(w_j^T z_i)} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{\exp(\|w_{y_i}\|_2 \|z_i\|_2 \cos\theta_{iy_i})}{\sum_{j=1}^{k}\exp(\|w_j\|_2 \|z_i\|_2 \cos\theta_{ij})}$$

- where $z_i = \phi_\Theta(x_i) \in \mathbb{R}^d$ (usually $\mathrm{k} \leq d+1$) is the learned feature representation vector , ϕ_Θ denotes the feature extraction sub-network, $W = (w_1, \ldots, w_k) \in \mathbb{R}^{d\times k}$ denotes the linear classifier which is implemented with a linear layer at the end of the network , θ_{ij} denotes the angle between z_i and w_j, and $\|\cdot\|_2$ denotes the Euclidean norm, where $w_1, \ldots, w_k$ can be regarded as the class centers or prototypes. For simplicity, we use prototypes to denote the weight vectors in the last layer.

Theorem 0. $\forall\varepsilon \in \left(0, \frac{\pi}{2}\right)$, if the domain of $w_1, \ldots, w_k, z_1, \ldots, z_N$ is $\mathbb{R}^d$, then there exist prototypes that achieve the infimum of the softmax loss and have the class margin ε.

Class Margin

For the prototypes $w_1, \dots, w_k \in \mathbb{R}^d$, we define the class margin as the minimal pairwise angle distance, i.e.,

$$m_c(\{w_i\}_{i=1}^k) = \min_{i \neq j} \angle(w_i, w_j) = \arccos\left(\max_{i \neq j} \frac{w_i^T w_j}{\|w_i\|_2 \|w_j\|_2} \right),$$

where $\angle(w_i, w_j)$ denotes the angle between the vectors w_i and w_j. Notice that we omit the magnitudes of the prototypes in the definition, since the magnitudes tend to be very close.

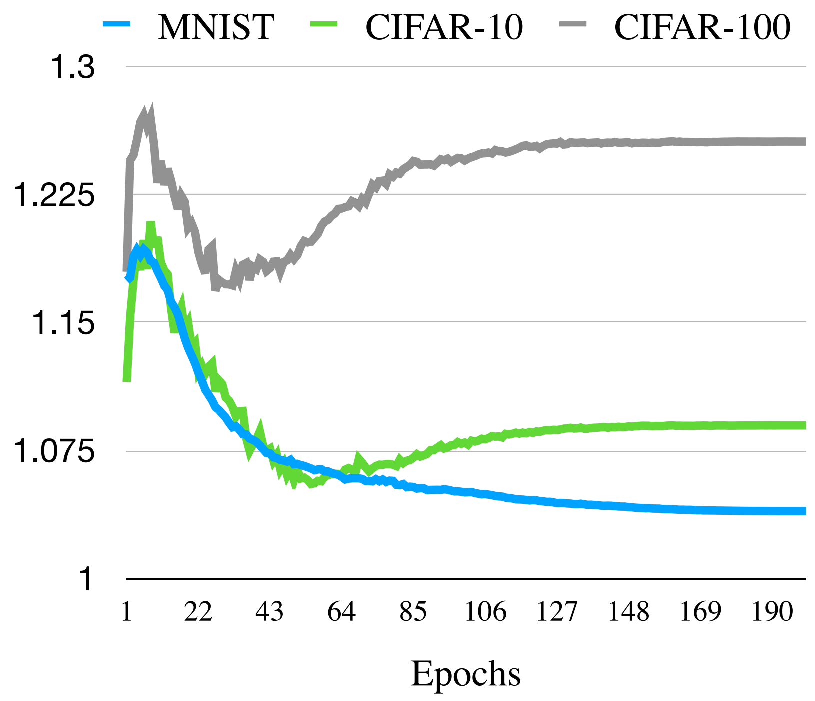

Figure 1: *The curves of ratio between maximum and minimum magnitudes of prototypes on MNIST and CIFAR-10/-100 using the softmax loss. The ratio is roughly close to 1 ($<$ 1.3).*

Class Margin

For the prototypes $w_1, \ldots, w_k \in \mathbb{R}^d$, we define the class margin as the minimal pairwise angle distance, i.e.,

$$m_c(\{w_i\}_{i=1}^k) = \min_{i \neq j} \angle(w_i, w_j) = \arccos\left(\max_{i \neq j} \frac{w_i^T w_j}{\|w_i\|_2 \|w_j\|_2}\right),$$

where $\angle(w_i, w_j)$ denotes the angle between the vectors w_i and w_j. Notice that we omit the magnitudes of the prototypes in the definition, since the magnitudes tend to be very close.

To obtain better inter-class separability, we seek the largest class margin, which can be formulated as

$$\max_{\{w_i\}_{i=1}^k} m_c(\{w_i\}_{i=1}^k) = \max_{\{w_i\}_{i=1}^k} \min_{i \neq j} \angle(w_i, w_j).$$

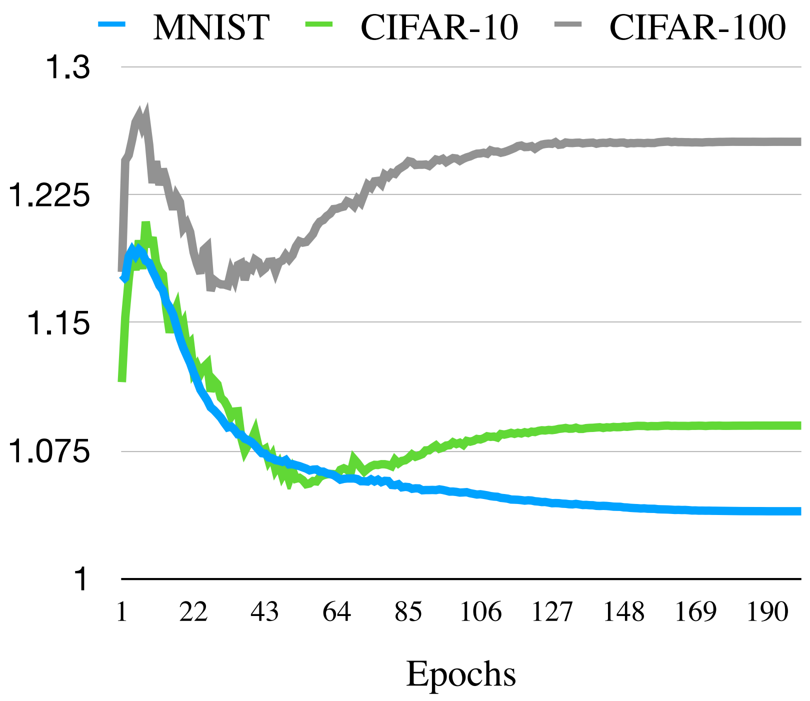

Figure 1: *The curves of ratio between maximum and minimum magnitudes of prototypes on MNIST and CIFAR-10/-100 using the softmax loss. The ratio is roughly close to 1 ($<$ 1.3).*

Maximization of Class Margin

- W perform ℓ_2 normalization to effectively restrict the prototypes on the unit sphere $\mathbb{S}^{d-1}$. Under this constraint, the maximization of the class margin is equivalent to the configuration of k points on $\mathbb{S}^{d-1}$ to maximize their minimum pairwise distance:

$$\arg\max_{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}} \min_{i \neq j} \angle(w_i, w_j) = \arg\max_{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}} \left\| w_i - w_j \right\|_2 .$$

- The right-hand side is well known as the k –points best-packing problem on spheres, whose solution leads to the optimal separation of points. And the best-packing problem turns to be the limiting case of the minimal Riesz energy problem:

$$\min_{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}} \lim_{t \to \infty} \sum_{i \neq j} \frac{1}{\left\| w_i - w_j \right\|_2^t} = \arg\max_{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}} \left\| w_i - w_j \right\|_2$$

[1] Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

Maximization of Class Margin

- W perform ℓ_2 normalization to effectively restrict the prototypes on the unit sphere $\mathbb{S}^{d-1}$. Under this constraint, the maximization of the class margin is equivalent to the configuration of k points on $\mathbb{S}^{d-1}$ to maximize their minimum pairwise distance:

$$\underset{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}}{\arg\max} \ \min_{i \neq j} \angle(w_i, w_j) = \underset{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}}{\arg\max} \left\| w_i - w_j \right\|_2 .$$

- The right-hand side is well known as the k –points best-packing problem on spheres, whose solution leads to the optimal separation of points. And the best-packing problem turns to be the limiting case of the minimal Riesz energy problem:

$$\min_{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}} \lim_{t \to \infty} \sum_{i \neq j} \frac{1}{\left\| w_i - w_j \right\|_2^t} = \underset{\{w_i\}_{i=1}^{k} \subset \mathbb{S}^{d-1}}{\arg\max} \left\| w_i - w_j \right\|_2$$

- **Lemma 1.[Optimality of Maximizing Class Margin]** For any $w_1, \ldots, w_k \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, the solution of minimal Riesz t-energy and k–points best-packing configurations are uniquely given by the vertices of regular $(k-1)$-simplices inscribed in $\mathbb{S}^{d-1}$. Furthermore, $w_i^T w_j = -\frac{1}{k-1}, \forall i \neq j$.

[1] Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*. Springer, 2019.

Sample Margin

According to the definition in Koltchinskii et al.[2], for the network $f(x;\Theta,W)=W^T\phi_\Theta(x):\mathbb{R}^m\to\mathbb{R}^k$ that outputs k logits, the margin of a sample (x,y) is defined as

$$\gamma(x,y)=f(x)_y-\max_{j\neq y} f(x)_j=w_y^T z-\max_{j\neq y} w_j^T z,$$

where $z=\phi_\Theta(x)$ denotes the corresponding feature. Let n_j be the number of samples in class j and $S_j=\{i:y_i=j\}$ denote the sample indices corresponding to class j. We can define the sample margin for samples in class j as

$$\gamma_j=\min_{i\in S_j}\gamma(x_i,y_i),$$

and the minimal sample margin over the entire dataset is $\gamma_{min}=\min\{\gamma_1,\dots,\gamma_k\}$.

[2] Koltchinskii V, Panchenko D. Empirical margin distributions and bounding the generalization error of combined classifiers[J]. The Annals of Statistics, 2002, 30(1): 1-50.

Sample Margin

According to the definition in Koltchinskii et al.[2], for the network $f(x;\Theta,W) = W^T\phi_\Theta(x): \mathbb{R}^m \to \mathbb{R}^k$ that outputs k logits, the margin of a sample (x,y) is defined as

$$\gamma(x,y) = f(x)_y - \max_{j\neq y} f(x)_j = w_y^T z - \max_{j\neq y} w_j^T z\,,$$

where $z = \phi_\Theta(x)$ denotes the corresponding feature. Let n_j be the number of samples in class j and $S_j = \{i: y_i = j\}$ denote the sample indices corresponding to class j. We can define the sample margin for samples in class j as

$$\gamma_j = \min_{i\in S_j} \gamma(x_i,y_i)\,,$$

and the minimal sample margin over the entire dataset is $\gamma_{min} = \min\{\gamma_1,\dots,\gamma_k\}$.

Theorem 2. For any $w_1,\dots,w_k \in \mathbb{S}^{d-1}$ (where n_j>0), the optimal solution $\{w_i^*\}_{i=1}^k$, $\{z_i^*\}_{i=1}^N$ of maximizing γ_{min} is obtained if and only if $\{w_i^*\}_{i=1}^k$ maximizes the class margin $m_c\left(\{w_i\}_{i=1}^k\right)$, and $z_i^* = \frac{w_{y_i}^* - \bar{w}_{y_i}^*}{\left\|w_{y_i}^* - \bar{w}_{y_i}^*\right\|_2}$, where $\bar{w}_{y_i}^*$ denotes the centroid of the vectors $\{w_j: j\ maximizes\ w_j^T w_{y_i}^*, j \neq y_i\}$.

[2] Koltchinskii V, Panchenko D. Empirical margin distributions and bounding the generalization error of combined classifiers[J]. The Annals of Statistics, 2002, 30(1): 1-50.

Maximization of Sample Margin

Proposition 3. For any $w_1, \dots, w_k, z_1, \dots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, the maximum of γ_{min} is $\frac{k}{k-1}$, which is obtained if and only if $\forall i \neq j, w_i^T w_j = -\frac{1}{k-1}$, and $z_i = w_{y_i}$.

Theorem 2 and **Proposition 3** show that the best separation of prototypes is obtained when maximizing the minimal sample margin γ_{min}.

On the other hand, let $L_{\gamma,j}[f] = \Pr[\max_{j' \neq j} f(x)_{j'} > f(x)_j - \gamma]$ denote the hard margin loss on samples from class j, and let $\hat{L}_{\gamma,j}$ denote its empirical variant. When the training dataset is separable, Cao et al.[3] provide a class-balanced generalization error bound, i.e., for $\gamma_j > 0$ and all $f \in \mathcal{F}$, with a high probability we have

$$\Pr\left[\max_{j' \neq j} f(x)_{j'} > f(x)_y\right] \leq \frac{1}{k} \sum_{j=1}^{k} \left(\hat{L}_{\gamma,j}[f] + \frac{4}{\gamma_j} \hat{\mathfrak{R}}_j(\mathcal{F}) + \varepsilon_j(\gamma_j) \right).$$

where $\hat{\mathfrak{R}}_j(\mathcal{F})$ denotes the empirical Rademacher complexity.

[3] Cao K, Wei C, Gaidon A, et al. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss[J]. Advances in Neural Information Processing Systems, 2019, 32: 1567-1578.

Margin-based Losses

What loss can learn towards the largest margins? Can CE?

Theorem 4. $\forall \varepsilon \in \left(0, \frac{\pi}{2}\right)$, if the domain of $w_1, \dots, w_k, z_1, \dots, z_N$ is $\mathbb{R}^d$, then there exists prototypes that achieve the infimum of the softmax loss and have the class margin ε.

This theorem reveals that, the original softmax loss may produce an arbitrary small class margin.

Therefore, many works emphasize the normalization of both features and prototypes.

A unified framework[8] that covers A-Softmax[4] with feature normalization, NormFace[5], CosFace[6] /AM-Softmax[7] and ArcFace[8] as a special cases can be formulated with hyper-parameters m_1, m_2, m_3:

$$L_i' = -\log \frac{\exp\left(s\left(\cos(m_1\theta_{iy_i} + m_2)\right) - m_3\right)}{\exp\left(s\left(\cos(m_1\theta_{iy_i} + m_2)\right) - m_3\right) + \sum_{j\neq y_i} \exp\left(s\cos\theta_{ij}\right)}.$$

[4] Liu et al. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

[5] Wang et al. NormFace: L2 hypersphere embed- ding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1041–1049, 2017.

[6] Wang et al. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018b.

[7] Wang et al. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018a.

[8] Deng et al. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Margin-based Losses

The setting of these hyper-parameters always guarantees that $\cos(m_1\theta_{iy_i} + m_2) \leq \cos m_2 \cos\theta_{iy_i}$, and m_2 usually set to satisfy $\geq \frac{1}{2}$. Let $\alpha = \cos m_2$ and $\beta = -m_3 < 0$, we have

$$L_i' \geq -\log \frac{\exp\left(s\left(\alpha\cos\theta_{iy_i}\right) + \beta\right)}{\exp\left(s\left(\alpha\cos\theta_{iy_i}\right) + \beta\right) + \sum_{j\neq y_i}\exp\left(s\cos\theta_{ij}\right)},$$

which indicates that the existing well-designed normalized softmax loss functions are all considered as the upper bound of the RHS, and the equality holds if and only if $\theta_{iy_i} = 0$.

Generalized Margin Softmax Loss. We can derive a more general formulation:

$$L_i = -\log \frac{\exp\left(s\left(\alpha_{i1}\cos\theta_{iy_i}\right) + \beta_{i1}\right)}{\exp\left(s\left(\alpha_{i2}\cos\theta_{iy_i}\right) + \beta_{i2}\right) + \sum_{j\neq y_i}\exp\left(s\cos\theta_{ij}\right)},$$

where α_{i1}, α_{i2}, β_{i1} and β_{i2} are the hyper-parameters to handle the margins in training, which are set specifically for each sample. We also require that $\alpha_{i1} \geq \frac{1}{2}$, $\alpha_{i2} \leq \alpha_{i1}$, $s > 0$, $\beta_{i1}, \beta_{i2} \in \mathbb{R}$.

Learning Towards the Largest for Class-balanced Cases

Theorem 5. For balanced datasets (i.e., each class has the same number of samples), $w_1, \dots, w_k$, $z_1, \dots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, learning with GM-Softmax (where $\alpha_{i1} = \alpha_1$, $\alpha_{i2} = \alpha_2$, $\beta_{i1} = \beta_1$, $\beta_{i2} = \beta_2$) leads to maximizing both the class margin and the sample margin. More specifically, the optimal solutions

$$\{w_i^*\}_{i=1}^{k}, \{z_i^*\}_{i=1}^{N} = \underset{\{w_j\},\{z_i\}\subset\mathbb{S}^{d-1}}{\arg\min} \frac{1}{N}\sum_{i=1}^{n} -\log \frac{\exp\left(s\left(\alpha_{i1}\cos\theta_{iy_i}\right)+\beta_{i1}\right)}{\exp\left(s\left(\alpha_{i2}\cos\theta_{iy_i}\right)+\beta_{i2}\right)+\sum_{j\neq y_i}\exp\left(s\cos\theta_{ij}\right)}$$

has the largest class margins $m_c^* = \arccos\frac{-1}{k-1}$, and the largest sample margin $\gamma_{min} = \frac{k}{k-1}$. The lower bound of the risk is $\log\left[\exp\left(s(\alpha_1+\beta_1-\alpha_2-\beta_2)\right)+(k-1)\exp(-s(\frac{-1}{k-1}+\alpha_1+\beta_1))\right]$, which is obtained if and only if $\forall i \neq j$, $w_i^T w_j = -\frac{1}{k-1}$, and $z_i = w_{y_i}$.

Learning Towards the Largest for Class-balanced Cases

ICLR

Theorem 5. For balanced datasets (i.e., each class has the same number of samples), $w_1, \ldots, w_k$, $z_1, \ldots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, learning with GM-Softmax (where $\alpha_{i1} = \alpha_1$, $\alpha_{i2} = \alpha_2$, $\beta_{i1} = \beta_1$, $\beta_{i2} = \beta_2$) leads to maximizing both the class margin and the sample margin. More specifically, the optimal solutions

$$\{w_i^*\}_{i=1}^k, \{z_i^*\}_{i=1}^N = \underset{\{w_j\},\{z_i\}\subset\mathbb{S}^{d-1}}{\arg\min} \frac{1}{N}\sum_{i=1}^{n} -\log \frac{\exp\left(s\left(\alpha_{i1}\cos\theta_{iy_i}\right)+\beta_{i1}\right)}{\exp\left(s\left(\alpha_{i2}\cos\theta_{iy_i}\right)+\beta_{i2}\right)+\sum_{j\neq y_i}\exp\left(s\cos\theta_{ij}\right)}$$

has the largest class margins $m_c^* = \arccos\frac{-1}{k-1}$, and the largest sample margin $\gamma_{min} = \frac{k}{k-1}$. The lower bound of the risk is $\log\left[\exp\left(s(\alpha_1+\beta_1-\alpha_2-\beta_2)\right)+(k-1)\exp(-s(\frac{-1}{k-1}+\alpha_1+\beta_1))\right]$, which is obtained if and only if $\forall i \neq j$, $w_i^T w_j = -\frac{1}{k-1}$, and $z_i = w_{y_i}$.

Proposition 6. For the balanced dataset, $w_1, \ldots, w_k, z_1, \ldots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, learning with the loss functions A-Softmax with feature normalization, NormFace, CosFace/AM-Softmax, and ArcFace share the same optimal solution.

Sample Margin Regularization

ICLR

In order to encourage learning towards the largest margins, we try to explicitly leverage the sample margin as the loss function, which is defined as

$$R_{sm}(x, y) = -\left(w_y^T z - \max_{j \neq y_i} w_j^T z\right).$$

The empirical risk $\frac{1}{N}\sum_{i=1}^{N} R_{sm}(x_i, y_i)$ is a lower-bound surrogate of $-\gamma_{min}$, i.e., $-\gamma_{min} \geq \frac{1}{N}\sum_{i=1}^{N} R_{sm}(x_i, y_i)$, while directly minimizing $-\gamma_{min}$ is too difficult to optimize a neural network. When $k \leq d + 1$, learning with new loss would promote the learning towards the largest margins:

Theorem 7. For the balanced dataset, $w_1, \dots, w_k, z_1, \dots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d + 1$, learning with R_{sm} leads to the maximization of the class margin and the sample margin.

Although learning with R_{sm} theoretically achieves the largest margins, in practical implementation, the optimization by the gradient-based methods shows unstable and non-convergent results for large scale datasets. Alternatively, we turn to combine R_{sm} as a regularization or complementary term with commonly-used losses.

Sample Margin Regularization

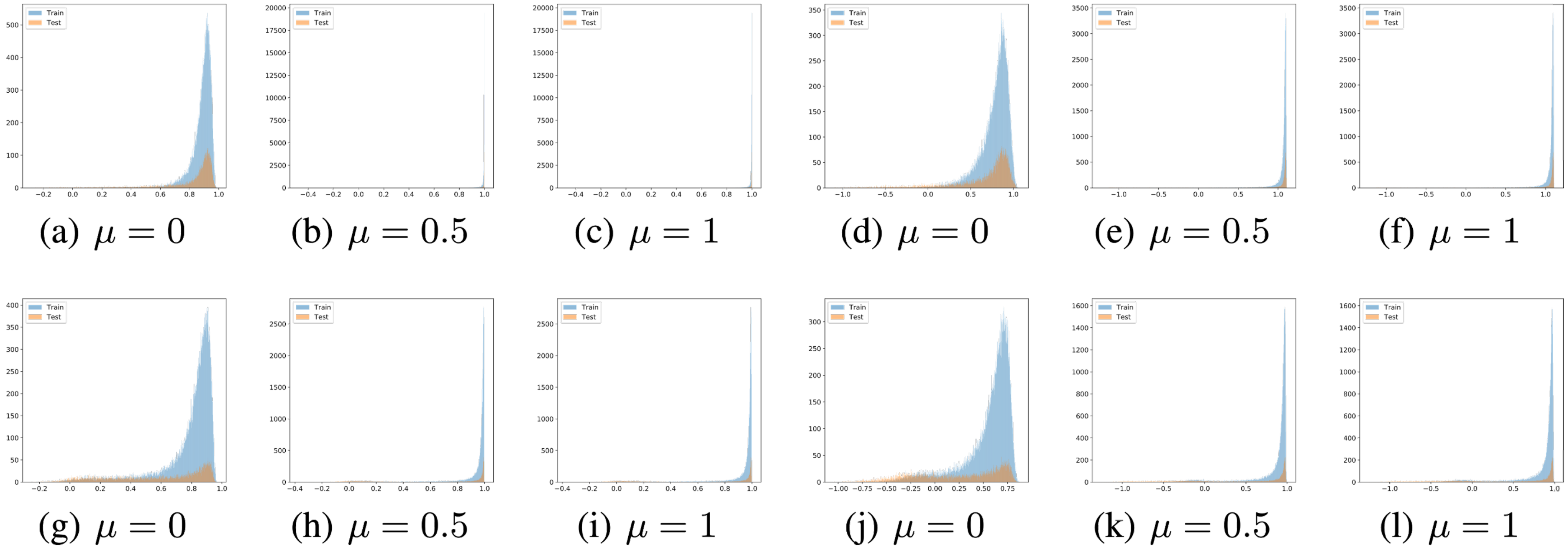

Figure 9: Histogram of similarities and sample margins for CosFace ($s = 20$) with/without sample margin regularization R_{sm} on CIFAR-10 and CIFAR-100. (a-c) and (g-i) denote the cosine similarities on CIFAR-10 and CIFAR-100, respectively. (d-f) and (j-l) denote the sample margins on CIFAR-10 and CIFAR-100, respectively.

Largest Margin Softmax Loss

ICLR

Theorem 2 provides a theoretical guarantee that maximizing γ_{min} would lead to maximizing the class margin regardless of the feature dimension, the class number, and class balancedness. However, directly maximizing γ_{min} is difficult to optimize a neural network with only one sample margin. As a consequence, we introduce an appropriate surrogate loss, which is called Largest Margin Softmax (LM-Softmax) loss:

$$L(x, y; s) = -\frac{1}{s}\log\frac{\exp\left(s w_y^T z\right)}{\sum_{j\neq y}\exp\left(s w_j^T z\right)}.$$

Actually, based on the limiting case of the *log-sum-exp* operator, we have

$$-\gamma_{min} = \lim_{s\to\infty}\frac{1}{s}\log\sum_{i=1}^{N}\sum_{j\neq y_i}\exp\left(s\left(w_j^T z_i - w_{y_i}^T z_i\right)\right).$$

Since log is strictly concave, we can derive the following inequality:

$$\frac{1}{s}\log\sum_{i=1}^{N}\sum_{j\neq y_i}\exp\left(s\left(w_j^T z_i - w_{y_i}^T z_i\right)\right) \geq \frac{1}{sN}\sum_{i=1}^{N} L(x_i, y_i; s) + \frac{1}{s}\log N.$$

Thus, we can achieve the maximizing of γ_{min} by learning with $L(x, y; s)$.

Largest Margin Softmax Loss

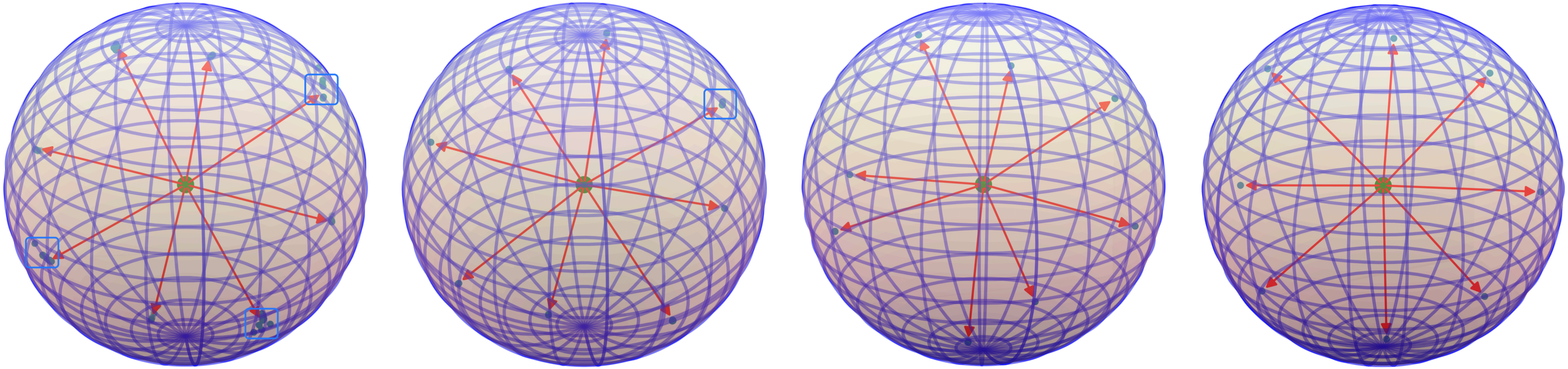

Figure 1: *Visualization of the learned prototypes (red arrows) and features (green points) using NormFace, CosFace, ArcFace and LM-Softmax on* $\mathbb{S}^2$ *for eight classes. The optimal solution of Tammes problem[3] for* $N = 8$ *have the class margin* $74.86°$*[4], where the class margin of learning with the losses NormFace, CosFace, ArcFace and LM-Softmax are* $68.50°$, $71.86°$, $67.74°$ *and* $74.46°$*, respectively.*

[9]"Tammes Problem" (2021) Wikipedia. Available at https://en.wikipedia.org/wiki/Tammes_problem (Accessed: 8 March 2022)
[10] L. L. Whyte. Unique arrangements of points on a sphere. *The American Mathematical Monthly*, 59 (9):606–611, 1952. ISSN 00029890, 19300972.

Learning Towards the Largest for Class-Imbalanced Cases

Theorem 8. For balanced or imbalanced cases, $w_1, \dots, w_k, z_1, \dots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, if $\sum_{i=1}^{k} w_i = 0$, then learning with GM-Softmax leads to maximizing both the class margin and the sample margin. More specifically, the optimal solutions $\{w_i^*\}_{i=1}^{k}, \{z_i^*\}_{i=1}^{N}$ has the largest class margins $m_c^* = \arccos\frac{-1}{k-1}$, and the largest sample margin $\gamma_{min} = \frac{k}{k-1}$. The lower bound of the risk is $\frac{1}{N}\sum_{i=1}^{N} \log[\exp\left(s(\alpha_{i1} + \beta_{i1} - \alpha_{i2} - \beta_{i2})\right) + (k-1)\exp(-s(\frac{1}{k-1} + \alpha_{i1} + \beta_{i1}))]$, which is obtained if and only if $\forall i \neq j, w_i^T w_j = -\frac{1}{k-1}$, and $z_i = w_{y_i}$.

Learning Towards the Largest for Class-Imbalanced Cases

Theorem 8. For balanced or imbalanced cases, $w_1, \dots, w_k, z_1, \dots, z_N \in \mathbb{S}^{d-1}$, $d \geq 2$, and $2 \leq k \leq d+1$, if $\sum_{i=1}^{k} w_i = 0$, then learning with GM-Softmax leads to maximizing both the class margin and the sample margin. More specifically, the optimal solutions $\{w_i^*\}_{i=1}^{k}, \{z_i^*\}_{i=1}^{N}$ has the largest class margins $m_c^* = \arccos\frac{-1}{k-1}$, and the largest sample margin $\gamma_{min} = \frac{k}{k-1}$. The lower bound of the risk is $\frac{1}{N}\sum_{i=1}^{N} \log[\exp\big(s(\alpha_{i1} + \beta_{i1} - \alpha_{i2} - \beta_{i2})\big) + (k-1)\exp(-s(\frac{1}{k-1} + \alpha_{i1} + \beta_{i1}))]$, which is obtained if and only if $\forall i \neq j, w_i^T w_j = -\frac{1}{k-1}$, and $z_i = w_{y_i}$.

Zero-Centroid Regularization. As a consequence, we propose a straight regularization term as follows, which can be combined with commonly-used losses to remedy the class-imbalanced problem:

$$R_W\left(\{w_j\}_{j=1}^{k}\right) = \lambda \left\| \frac{1}{k} \sum_{j=1}^{k} w_j \right\|_2^2$$

The zero centroid regularization only applies to the prototypes at the last inner-product layer.

Experiments

Table 1: Test accuracies (acc), class margins (m_{cls}) and sample margins (m_{samp}) on MNIST, CIFAR-10 and CIFAR-100 using loss functions with/without R_{sm} in (3.4). The results with positive gains are **highlighted**.

Dataset	MNIST			CIFAR-10			CIFAR-100		
Metric	acc	m_{cls}	m_{samp}	acc	m_{cls}	m_{samp}	acc	m_{cls}	m_{samp}
CE	99.11	87.39°	0.5014	94.12	81.73°	0.6203	74.56	65.38°	0.1612
CE + $0.5R_{\text{sm}}$	**99.13**	**95.41°**	**1.026**	**94.45**	**96.31°**	**0.9744**	**74.96**	**90.00°**	**0.4955**
CosFace ($s = 10$)	98.98	95.93°	0.9839	94.39	96.00°	0.9168	74.44	83.31°	0.4578
CosFace ($s = 20$)	99.06	93.24°	0.8376	94.13	91.22°	0.7955	73.26	79.17°	0.3078
CosFace ($s = 64$)	99.25	89.50°	0.7581	93.53	64.14°	0.6969	73.87	72.56°	0.2233
CosFace ($s = 10$) + $0.5R_{\text{sm}}$	**99.16**	95.56°	**1.033**	**94.42**	**96.26°**	**0.9675**	73.76	**90.21°**	**0.5089**
CosFace ($s = 20$) + $0.5R_{\text{sm}}$	**99.24**	95.41°	**1.030**	**94.27**	**96.18°**	**0.9490**	74.41	**89.02°**	**0.4780**
CosFace ($s = 64$) + $0.5R_{\text{sm}}$	**99.27**	95.35°	**1.019**	**94.20**	**95.48°**	**0.9075**	**74.53**	**85.31°**	**0.3817**
ArcFace ($s = 10$)	99.05	94.64°	0.8225	94.50	91.23°	0.8501	73.96	76.91°	0.4313
ArcFace ($s = 20$)	99.11	90.84°	0.6091	94.11	53.98°	0.5707	74.74	60.91°	0.3010
ArcFace ($s = 64$)	99.21	82.63°	0.4038	—	—	—	—	—	—
ArcFace ($s = 10$) + $0.5R_{\text{sm}}$	**99.14**	**95.42°**	**1.034**	94.21	**96.27°**	**0.9651**	**74.47**	**90.13°**	**0.5143**
ArcFace ($s = 20$) + $0.5R_{\text{sm}}$	**99.19**	**91.38°**	**1.030**	**94.32**	**96.15°**	**0.9571**	74.64	**88.73°**	**0.4804**
ArcFace ($s = 64$) + $0.5R_{\text{sm}}$	99.14	**95.29°**	**1.019**	—	—	—	—	—	
NormFace ($s = 10$)	99.06	94.34°	0.7750	94.16	94.40°	0.8004	74.23	79.10°	0.4250
NormFace ($s = 20$)	99.09	89.27°	0.5263	94.09	74.32°	0.6001	73.87	77.47°	0.2498
NormFace ($s = 64$)	99.00	82.08°	0.2621	94.01	36.50°	0.2633	73.42	52.37°	0.0993
NormFace ($s = 10$) + $0.5R_{\text{sm}}$	**99.16**	**95.38°**	**1.034**	**94.23**	**96.28°**	**0.9650**	**74.54**	**90.10°**	**0.5160**
NormFace ($s = 20$) + $0.5R_{\text{sm}}$	**99.19**	**95.37°**	**1.031**	**94.38**	**96.17°**	**0.9519**	**74.75**	**88.86°**	**0.4773**
NormFace ($s = 64$) + $0.5R_{\text{sm}}$	**99.34**	**95.29°**	**1.021**	**94.42**	**93.87°**	**0.9508**	**74.33**	**76.02°**	**0.3665**

Table 3: The results on Market-1501 and DukeMTMC for person re-identification task. The best three results are **highlighted**.

Dataset	Market-1501			DukeMTMC		
Method	mAP	Rank1	Rank@5	mAP	Rank@1	Rank@5
CE	82.8	92.7	97.5	**73.0**	83.5	**93.0**
ArcFace ($s = 10$)	67.5	84.1	92.1	37.7	58.7	72.7
ArcFace ($s = 20$)	79.1	90.8	96.5	61.4	78.3	88.6
ArcFace ($s = 64$)	80.4	92.6	97.4	67.6	83.4	91.4
CosFace ($s = 10$)	68.0	84.9	92.7	39.3	60.6	73.1
CosFace ($s = 20$)	80.5	92.0	97.1	64.2	81.3	89.7
CosFace ($s = 64$)	78.7	92.0	97.1	68.2	83.1	92.5
NormFace ($s = 10$)	81.2	91.6	96.3	63.7	79.3	88.5
NormFace ($s = 20$)	83.2	**93.5**	**97.9**	71.6	83.8	93.3
NormFace ($s = 64$)	77.5	90.0	96.9	60.1	75.2	88.1
LM-Softmax ($s = 10$)	**83.3**	92.8	97.1	72.2	**85.8**	92.4
LM-Softmax ($s = 20$)	**84.7**	**93.8**	**97.6**	**74.1**	**86.4**	**93.5**
LM-Softmax ($s = 64$)	**84.6**	**93.9**	**98.1**	**74.2**	**86.6**	**93.5**

Table 2: Test accuracies (acc) and class margins (m_{cls}) on imbalanced CIFAR-10. The results with positive gains are **highlighted** (where * denotes coupling with zero-centroid regularization term).

Dataset	Imbalanced CIFAR-10								Imbalanced CIFAR-100							
Imbalance Type	long-tailed				step				long-tailed				step			
Imbalance Ratio	100		10		100		10		100		10		100		10	
Metric	acc	m_{cls}	acc	m_{cls}	acc	m_{cls}	acc	m_{cls}	acc	m_{cls}	acc	m_{cls}	acc	m_{cls}	acc	m_{cls}
CE	70.88	77.41°	88.17	79.63°	62.21	76.50°	85.06	82.24°	40.38	64.73°	60.42	66.24°	42.36	60.32°	56.88	62.82°
Focal	66.30	74.14°	87.33	74.48°	60.55	63.31°	84.49	75.16°	38.04	54.67°	60.09	59.29°	41.90	55.98°	57.84	55.72°
CosFace	69.28	58.77°	87.02	81.61°	53.64	19.78°	84.86	75.96°	34.91	4.731°	60.60	70.81°	40.36	0.764°	47.56	8.559°
CosFace*	**69.52**	**91.90°**	**87.55**	**95.46°**	**62.49**	**95.86°**	**85.59**	**96.12°**	**40.98**	**80.93°**	**60.77**	**84.97°**	**41.17**	**41.59°**	**57.97**	**83.93°**
ArcFace	72.20	65.86°	89.00	85.23°	62.48	54.29°	86.32	80.51°	42.77	13.22°	63.21	67.73°	41.47	0.497°	58.89	0.369°
ArcFace*	**72.23**	**92.30°**	**89.22**	**96.23°**	**64.38**	**93.51°**	**86.65**	**96.23°**	**44.68**	**56.60°**	**63.80**	**73.45°**	**44.26**	**32.10°**	**60.79**	**79.85°**
NormFace	72.37	62.72°	89.19	82.60°	63.69	51.00°	86.37	77.82°	43.71	16.11°	63.50	71.26°	41.93	1.363°	59.85	21.32°
NormFace*	72.07	**94.95°**	**89.30**	**94.50°**	**64.07**	**93.06°**	**86.49**	**96.28°**	**44.25**	**64.85°**	**63.81**	**79.85°**	**44.51**	**36.30°**	**60.22**	**80.83°**
LDAM	72.86	73.30°	88.92	88.19°	63.27	61.42°	87.04	85.21°	43.28	7.733°	63.62	73.19°	41.65	0.852°	58.32	6.085°
LDAM*	72.86	**91.75°**	**89.51**	**96.26°**	**64.99**	**96.04°**	86.74	**96.26°**	**45.23**	**70.96°**	**64.18**	**85.03°**	**44.48**	**43.26°**	**60.83**	**75.22°**
LM-Softmax	65.32	4.420°	88.69	68.91°	50.47	0.452°	86.08	52.20°	41.52	4.500°	63.26	68.31°	41.53	0.467°	55.44	1.372°
LM-Softmax*	**73.21**	**92.57°**	**89.12**	**95.73°**	**65.91**	**93.84°**	**87.07**	**96.05°**	**45.28**	**69.53°**	**63.77**	**81.99°**	**46.23**	**43.15°**	**60.73**	**74.78°**

Table 3: The results on Market-1501 and DukeMTMC for person re-identification task. The best three results are **highlighted**.

Dataset	Market-1501			DukeMTMC		
Method	mAP	Rank1	Rank@5	mAP	Rank@1	Rank@5
CE	82.8	92.7	97.5	**73.0**	83.5	**93.0**
ArcFace ($s = 10$)	67.5	84.1	92.1	37.7	58.7	72.7
ArcFace ($s = 20$)	79.1	90.8	96.5	61.4	78.3	88.6
ArcFace ($s = 64$)	80.4	92.6	97.4	67.6	83.4	91.4
CosFace ($s = 10$)	68.0	84.9	92.7	39.3	60.6	73.1
CosFace ($s = 20$)	80.5	92.0	97.1	64.2	81.3	89.7
CosFace ($s = 64$)	78.7	92.0	97.1	68.2	83.1	92.5
NormFace ($s = 10$)	81.2	91.6	96.3	63.7	79.3	88.5
NormFace ($s = 20$)	83.2	**93.5**	**97.9**	71.6	83.8	93.3
NormFace ($s = 64$)	77.5	90.0	96.9	60.1	75.2	88.1
LM-Softmax ($s = 10$)	**83.3**	92.8	97.1	72.2	**85.8**	92.4
LM-Softmax ($s = 20$)	**84.7**	**93.8**	**97.6**	**74.1**	**86.4**	**93.5**
LM-Softmax ($s = 64$)	**84.6**	**93.9**	**98.1**	**74.2**	**86.6**	**93.5**

Thanks for your attention!

Any question? Please contact us!

Xiong Zhou: cszx@hit.edu.cn
Xianming Liu: csxm@hit.edu.cn