Hyperspherical Classification with Dynamic Label-to-Prototype Assignment

Supplementary Material

6. Softmax Cross-Entropy Gradient Analysis

With the identity mapping as the label-to-prototype assignment $A(y_i) = y_i$, the widely used softmax Cross-Entropy (CE) loss can be presented as follows:

$$L_{CE} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{e^{\mathbf{z}_i^\top \mathbf{w}_{y_i}}}{e^{\mathbf{z}_i^\top \mathbf{w}_{y_i}} + \sum_{j=1, j\neq y_i}^{c} e^{\mathbf{z}_i^\top \mathbf{w}_j}}, \quad (10)$$

where B represents the batch size. The gradient with respect to the features can be expressed as:

$$\frac{\partial L_{CE}}{\partial \mathbf{z}_i} = -(1-p_{y_i})\mathbf{w}_{y_i} + \sum_{j=1, j\neq y_i}^{c} p_j \mathbf{w}_j = \mathbf{g}^+ + \mathbf{g}^-, \quad (11)$$

where $p_j = \frac{e^{\mathbf{z}_i^\top \mathbf{w}_j}}{\sum_{k=1}^{c} e^{\mathbf{z}_i^\top \mathbf{w}_k}}$. Equation 11 shows that from a feature perspective, the backbone optimization aims to align features with their corresponding ground-truth prototype, $\mathbf{g}^+$, while ensuring separation from all other prototypes, $\mathbf{g}^-$. Note that $\mathbf{g}^-$ is necessary for a learnable classifier to impose the inter-class separability. However, in a fixed classifier framework, $\mathbf{g}^-$ does not necessarily lead to separability and may cause performance degradation [54].

The gradient with respect to the prototypes can be expressed as:

$$\frac{\partial L_{CE}}{\partial \mathbf{w}_j} = -\sum_{i\in\mathbb{B}^+}(1-p_i)\mathbf{z}_i + \sum_{i\in\mathbb{B}^-} p_i \mathbf{z}_i = \hat{\mathbf{g}}^+ + \hat{\mathbf{g}}^-, \quad (12)$$

$\mathbb{B}^+$ represents samples belonging to the j-th class in mini-batch, and $\mathbb{B}^-$ denotes the samples from other classes. Consequently, from a classifier perspective, the prototypes belonging to the j-th class are updated towards the features of samples from their own class, $\hat{\mathbf{g}}^+$, while being pushed away from features of other classes, $\hat{\mathbf{g}}^-$. Note that in Equation 12, the $\hat{\mathbf{g}}^-$ is available even when no sample from the class j is present in the mini-batch (*passive update*). Therefore, given an unbalanced dataset, the optimization of prototypes of minority classes is predominantly influenced by $\hat{\mathbf{g}}^-$ [10]. Consequently, the prototypes of the minority classes gradually drift away from the feature space [10, 54]. Additionally, $\hat{\mathbf{g}}^-$ is approximately uniform across all minority classes, and forces prototypes of minority classes to the same subspace, leading to less separability among them [10].

7. Training Setups

7.1. Balanced

ImageNet-200 is a subset of ImageNet, containing 110K images from 200 classes. CIFAR-100 consists of 60K images from 100 classes. For both sets, 10K instances are randomly selected as the testing set and are held out from the training. We use SGD optimizer with batch size set as 128, initial learning rate as 0.01, momentum as 0.9, and weight decay as $1e-4$. All models are trained from scratch for 250 epochs, where in epochs 100 and 200, the learning rate is decreased by a factor of 10. For data augmentation, random cropping, and horizontal flips are conducted.

In ImageNet-1K experiments, ResNet-50 is trained using SGD optimizer with an initial learning rate of 0.5 with five epochs of linear warm-up and cosine decay learning rate. The batch size was set to 1024, weight decay to 2×10^{-5} 0.00002, and total training epoch to 100. As the data augmentation, we employed Mixup with β set as 1.0. For Swin-T, we employed AdamW as the optimizer with a batch size of 1024, an initial learning rate of 0.001, and a weight decay of 0.00002.

	d =16	d =32	d =64
Fixed $\mathbf{W}$ + CE	55.30 ± 0.76	56.46 ± 0.43	57.03 ± 0.34
Ours	**62.72** ± 0.36	**65.09** ± 0.24	**65.94** ± 0.11

Table 9. Abblation study on the effect of utilizing L_{IPM} when the classifier prototypes are predefined and fixed during the training. The classification accuracy (%) on CIFAR-100 using ResNet-32 is reported.

7.2. Long-tailed

Models are trained from scratch for 200 epochs, using SGD optimizer with an initial learning rate of 0.1, a batch size of 128, momentum of 0.9, and weight decay of $2e-4$. The learning rate is decayed by a factor of 10 at epochs 160 and 180. Following ETF [54], the loss is weighted by the inverse ratio of the number of samples per class and in addition to random cropping and horizontal flipping, we used Mixup with hyperparameter β set as 1.0.

8. Additional Ablations

9. Impact of L_{IPM}

In Table 9, we study the effect of the L_{IPM}. To this end, we substitute the PSC prototypes with our optimized prototypes. In this experiment, the prototypes of PSC are fixed during the training. Employing CE as the training objective drastically degrades performance. This degeneration is due to the unnecessary pushing force in the gradient of CE as discussed in Section 6 of this Supplementary Material. Also, this performance degradation emphasizes the important role of dynamic label-to-prototype assignment in better metric-space exploitation.