A Instantiations of Negative-Margin Loss

Eqn. 1 provides a general formulation for the margin softmax loss. Here, we provide detailed formulation of the two instantiations: negative-margin softmax loss and negative-margin cosine softmax loss as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\beta\cdot\left(W_{y_i}^T z_i - m\right)}}{e^{\beta\cdot\left(W_{y_i}^T z_i - m\right)} + \sum\limits_{j=1,j\neq y_i}^{C} e^{\beta\cdot\left(W_j^T z_i\right)}}, \tag{7}$$

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\beta\cdot\left(\cos\left(W_{y_i}, z_i\right) - m\right)}}{e^{\beta\cdot\left(\cos\left(W_{y_i}, z_i\right) - m\right)} + \sum\limits_{j=1,j\neq y_i}^{C} e^{\beta\cdot\cos\left(W_j, z_i\right)}}, \tag{8}$$

where $m \leq 0$ is the margin parameter. Note that the formulations are the same as the *large*-margin softmax loss in [22] and the large-margin cosine loss in [47], but we restrict the margin m as a non-positive value ($m \leq 0$) while the original formulations restrict the margin m as a non-negative value ($m \geq 0$).

We follow the pre-training and fine-tuning pipeline introduced in Sec. 3.5, and the proposed negative-margin (cosine) softmax loss is applied in the *pre-training* stage, as illustrated in Figure 8. Note we do not apply the negative-margin loss to the *pre-training* stage, where we find regular (cosine) softmax loss (margin $m = 0$) performs well, as detailed in the following section.

B Framework

Here we show the framework of our proposed approach in Figure 8, which follows a two-stage training pipeline for few-shot classification, including pre-training stage to perform metric learning on the abundant labeled data in base classes, and fine-tuning stage to learn a classifier to recognize novel classes.

C Theoretical Proof

Proof. Denote $z(x_i, m) = \frac{f_{\theta(m)}(x_i)}{\|f_{\theta(m)}(x_i)\|}$. Substituting Equation (2) into Equation (3), and take expectation, for any pair of data $(x_i, y_i), (x_l, y_l) \in I$,

$$E(\|z(x_i, m) - z(x_l, m)\|_2^2) = \begin{cases} 2D_{\text{intra}}(I, m), & y_i = y_l \\ 2D_{\text{inter}}(I, m), & y_i \neq y_l \end{cases}. \tag{9}$$

For any pair of data in the same novel classes $(x_i', y'), (x_l', y') \in I^n$,

$$\begin{aligned} D_{\text{intra}}(I^n, m) &= \frac{1}{2}E(\|z(x_i', m) - z(x_l', m)\|_2^2) \\ &= P^s D_{\text{intra}}(I, m) + (1 - P^s) D_{\text{inter}}(I, m). \end{aligned} \tag{10}$$

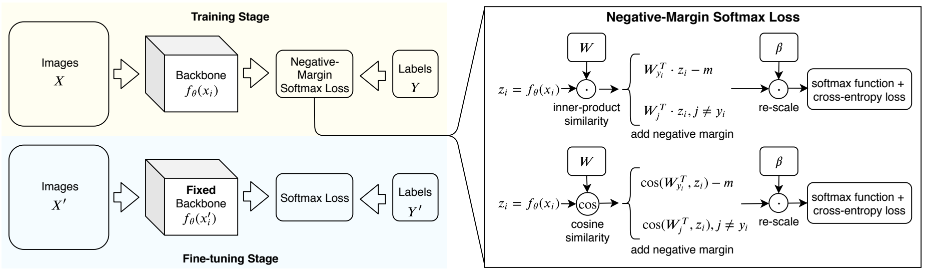

Fig. 8. Overview of our proposed approach, which consists of two stages, pre-training (learning metrics with sufficient annotated examples on training classes) and fine-tuning (learning a classifier on novel classes with few labeled examples). The negative-margin softmax loss is integrated in pre-training for learning more transferable features for novel classes, with two types of similarities, inner-product and cosine similarity.

Substituting it into Eqn (4), we have

$$\phi(I^n, m) = \frac{D_{\text{inter}}(I^n, m)}{P^s D_{\text{intra}}(I^b, m) + (1 - P^s) D_{\text{inter}}(I^b, m)}. \tag{11}$$

Substituting it into Eqn (6), we have

$$\begin{aligned}
&\phi(I^n, m_2) < \phi(I^n, m_1) \\
\Leftrightarrow &P^s < \frac{\frac{D_{\text{inter}}(I^n, m_1)}{D_{\text{inter}}(I^b, m_1)} - \frac{D_{\text{inter}}(I^n, m_2)}{D_{\text{inter}}(I^b, m_2)}}{\frac{D_{\text{inter}}(I^n, m_1)}{D_{\text{inter}}(I^b, m_1)} \cdot \left(1 - \frac{D_{\text{intra}}(I^b, m_2)}{D_{\text{inter}}(I^b, m_2)}\right) - \frac{D_{\text{inter}}(I^n, m_2)}{D_{\text{inter}}(I^b, m_2)} \cdot \left(1 - \frac{D_{\text{intra}}(I^b, m_1)}{D_{\text{inter}}(I^b, m_1)}\right)} \\
\Leftrightarrow &P^s < \frac{\psi(m_1) - \psi(m_2)}{\psi(m_1)(1 - \phi^{-1}(I^b, m_2)) - \psi(m_2)(1 - \phi^{-1}(I^b, m_1)))} \\
\Leftrightarrow &P^s < \frac{t}{t(1 - \phi^{-1}(I^b, m_1)) + r\psi(m_1)}.
\end{aligned} \tag{12}$$

D The effects of margin in the fine-tuning stage

In the main part of the paper, we show that applying the negative-margin softmax loss into the pre-training stage leads to better discrimination of novel classes. In this section, we investigate the effects of the margin parameter on the fine-tuning stage. As shown in Fig. 9, the 5-shot classification accuracy in the validation classes is insensitive when varying margin values, while that of 1-shot classification increases marginally when increasing margin values. Such behavior is different from the effects of the margin parameter in the pre-training stage, indicating that the negative margin mainly effects in open-set scenarios. Also note the accuracy is much more insensitive w.r.t margin parameter than that of pre-training and fine-tuning both using base classes (the blue curves in Fig 1), probably because the feature is fixed in the fine-tuning stage and different margin values can all find good discrimination planes well.

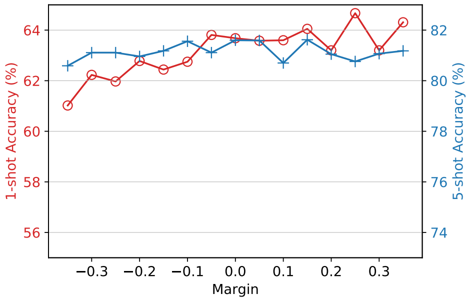

Fig. 9. The one-shot (in red) and five-shot (in blue) accuracy on validation classes w.r.t different margins in cosine softmax loss applied in fine-tunning stage on mini-ImageNet. The margin parameter in the pre-training stage is fixed to -0.3 and the backbone is ResNet-18.

E Relationship between negative margin and label smoothing

The same as negative margin loss, the label smoothing technique also aims at facilitating the difficulty of equalising softmax outputs and the binary ground-truth labels. Technically, while the label smoothing technique changes the ground truth target labels from binary values to soft ones which alleviate the equalising of softmax outputs and the binary ground-truth labels, the negative margin loss modify the other side of softmax outputs. Although resulting in similarly smaller loss than regular training, they perform very differently on training classes and novel classes.

In softmax based image classification, the final prediction is the class with the largest softmax probability and the accuracy is 100% if all predictions match the ground truth label. The softmax output for the ground truth label is not necessarily as 1, but is correct as long as it is larger than the softmax outputs of other classes. In another word, the cross entropy loss which encourages the softmax output of the right class to be exactly 1 is stricter than the real target of doing right classification. The label smoothing technique relaxes the ground-truth labels as soft ones, which shrinks the gap between cross-entropy loss and the real classification target. As a result, it usually achieves higher accuracy on the validation images of training classes, indicating better discrimination of training classes. On the contrary, the negative margin technique enlarges the gap between loss and real classification accuracy. It usually results in lower accuracy on the validation images of training classes, indicating lower discrimination of training classes.

Label	**Negative**	**Base**		**Val**	
Smoothing	**Margin**	**1 shot**	**5 shot**	**1 shot**	**5 shot**
		85.45 ± 0.67	93.72 ± 0.31	59.49 ± 0.90	78.90 ± 0.61
✓		**88.93 ± 0.62**	**93.93 ± 0.30**	58.02 ± 0.97	75.87 ± 0.71
	✓	79.92 ± 0.81	92.34 ± 0.38	**63.68 ± 0.86**	**81.60 ± 0.56**

Table 6. Comparison with label smoothing on the 5-way 1-shot and 5-shot accuracy for the base and validation classes in the mini-ImageNet dataset with ResNet-18 as backbone. For label smoothing, ϵ is set as 0.05. For negative margin, m in pre-training stage is fixed to -0.3.

Table 6 investigate the effects of two techniques by applying few-shot fine-tuning on either validation images (ILSVRC-2012 images of the base classes but not in mini-ImageNet) of base classes or images of val classes. For label smoothing, ϵ is set as 0.05. For negative margin, $m = -0.3$. It can be seen that the label smoothing technique improves the performance on the base classes by 3.48% and 0.21% for the 1-shot and 5-shot setting respectively, but has 1.47% and 3.03% accuracy drop on the validation classes for the 1-shot and 5-shot settings, respectively. On the contrary, the negative margin technique has lower performance on the validation images of training classes but benefits both the 1-shot and 5-shot classification accuracy on the validation classes.

These results are in accord with the previous analysis that the different effects of the negative margin technique and the label smoothing technique. While the label smoothing technique tends to improve the discriminability of base classes and harm the transferability to novel classes, the negative margin technique has lower discriminability of base classes but can improve the feature transferability to novel classes.