Negative Margin Matters: Understanding Margin in Few-shot Classification

Bin Liu[1⋆], Yue Cao[2*], Yutong Lin[2,3†], Qi Li[1], Zheng Zhang[2], Mingsheng Long[1], and Han Hu[2]

[1]Tsinghua University [2]Microsoft Research Asia [3]Xi'an Jiaotong University
{liubinthss,liqi17thu}@gmail.com mingsheng@tsinghua.edu.cn
{yuecao,v-yutlin,zhez,hanhu}@microsoft.com

Abstract. This paper introduces a negative margin loss to metric learning based few-shot learning methods. The negative margin loss significantly outperforms regular softmax loss, and achieves state-of-the-art accuracy on three standard few-shot classification benchmarks with few bells and whistles. These results are contrary to the common practice in the metric learning field, that the margin is zero or positive. To understand why the negative margin loss performs well for the few-shot classification, we analyze the discriminability of learned features w.r.t different margins for training and novel classes, both empirically and theoretically. We find that although negative margin reduces the feature discriminability for training classes, it may also avoid falsely mapping samples of the same novel class to multiple peaks or clusters, and thus benefit the discrimination of novel classes. Code is available at https://github.com/bl0/negative-margin.few-shot.

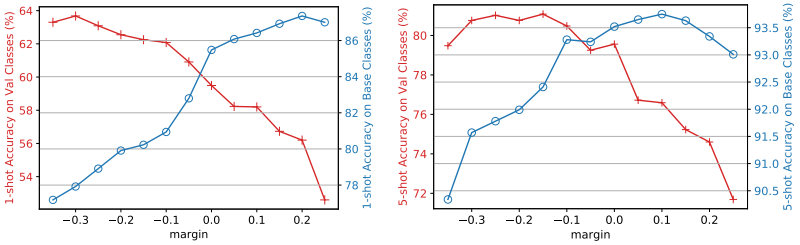
Keywords: Few-shot Classification, Metric Learning, Large Margin Loss

1 Introduction

Recent success on visual recognition tasks [17, 42, 13, 38, 4, 2] heavily relies on the massive-scale manually labeled training data, which is too expensive in many real scenarios. In contrast, humans are capable of learning new concepts with only a few examples, yet it still remains a challenge for modern machine learning systems. Hence, learning to generalize the knowledge in base classes (with sufficient annotated examples) to novel classes (with a few labeled examples), also known as few-shot learning, has attracted more and more attention [3, 19, 25, 7, 45, 43, 9, 37, 44, 36, 35, 10].

An important direction of few-shot classification is meta learning, which aims to learn a meta-learner on base classes and generalizes it to novel classes. Metric learning based methods [3, 7, 25], are an important series of the meta-learning methods, and perform metric learning in the base classes and then transfer the learned metrics to the novel classes. For example, [3] proved that simply using

⋆ Equal contribution. † The work is done when Yutong Lin is an intern at MSRA.



(a) 1-shot accuracy (b) 5-shot accuracy

Fig. 1. The one-shot and five-shot accuracy on novel classes (in red) and base classes (in blue) w.r.t different margins in cosine softmax loss on mini-ImageNet. As we expect, applying larger margin to softmax loss can achieve better accuracy on base classes. But surprisingly, applying appropriate negative margin to softmax loss can achieve state-of-the-art few-shot accuracy on novel classes.

standard softmax loss or cosine softmax loss for learning metrics in base classes can achieve the state-of-the-art few-shot classification performance via learning a linear classifier on novel classes.

In the metric learning area, a common view is that the standard softmax loss is insufficient for discrimination on different training classes. Several previous approaches integrate the large and positive margin to the softmax loss [22] or the cosine softmax loss [6, 47] so as to enforce the score of ground truth class larger than that of other classes by at least a margin. This could help to learn highly-discriminative deep features and result in remarkable performance improvement on visual recognition tasks, especially on face recognition [22, 6, 47].

Consequently, it inspires us to adopt this large-margin softmax loss to learn better metrics for few-shot classification. As we expected, shown as the blue curves in Fig. 1, the metrics learned by large-margin softmax with positive margin are more discriminative on training classes, resulting in higher few-shot accuracy on the validation set of training classes. But in the standard open-set setting of few-shot classification, shown as red curves in Fig. 1, we surprisingly find out that adding the positive margin in softmax loss would hurt the performance.

From our perspective, the positive margin would make the learned metrics more discriminative to training classes. But for novel classes, positive margin would map the samples of the same class to multiple peaks or clusters in base classes (shown in Fig. 3 and Fig. 7) and hurt their discriminability. We then give a theoretical analysis that the discriminability of the samples in the novel classes is monotonic decreasing w.r.t the margin parameter under proper assumption. Instead, appropriate negative margin could achieve a better tradeoff between the discriminability and transferability for novel classes, and achieves better performance on few-shot classification.

The main contributions of this paper are summarized as follows:

1. This is the first endeavor to show that softmax loss with negative margin works surprisingly well on few-shot classification, which breaks the inherent understanding that margin can only be limited to positive values [6, 22, 47].
2. We provide insightful intuitive explanation and the theoretical analysis about why negative margin works well for few-shot classification.
3. The proposed approach with negative margin achieves state-of-the-art performance on three widely-used few-shot classification benchmarks.

2 Related Work

Few-Shot Classification. The existing representative few-shot learning methods can be broadly divided into three categories: *gradient-based* methods, *hallucination-based* methods, and *metric-based* methods.

Gradient-based methods tackle the few-shot classification by learning the task-agnostic knowledge. [9, 39, 31, 29, 27] focus on learning a suitable initialization of the model parameters which can quickly adapt to new tasks with a limited number of labeled data and a small number of gradient update steps. Another line of works aims at learning an optimizer, such as LSTM-based meta learner [37] and weight-update mechanism with an external memory [28], for replacing the stochastic gradient descent optimizer. However, it is challenging to solve the dual or bi-level optimization problem of these works, so their performance is not competitive on large datasets. Recently, [19, 1] alleviate the optimization problem by closed-form model like SVM, and achieve better performance on few-shot classification benchmark of large dataset.

Hallucination-based methods attempt to address the limited data issue by learning an image generator from base classes, which is adopted to hallucinate new images in novel classes [12, 48]. [12] presents a way of hallucinating additional examples for novel classes by transferring modes of variation from base classes. [48] learns to hallucinate examples that are useful for classification by the end-to-end optimization of both classifier and hallucinator. As hallucination-based methods can be considered as the supplement and are always adopted with other few-shot methods, we follow [3] to exclude these methods in our experimental comparison and leave it to future work.

Metric-based methods aim at learning a transferable distance metric. MatchingNet [45] computes cosine similarity between the embeddings of labeled images and unlabeled images, to classify the unlabeled images. ProtoNet [43] represents each class by the mean embedding of the examples inside this class, and the classification is performed based on the distance to the mean embedding of each class. RelationNet [44] replaces the non-parametric distance in ProtoNet to a parametric relation module. Recently, [3, 7, 25] reveal that the simple pre-training and fine-tuning pipeline (following the standard transfer learning paradigm) can achieve surprisingly competitive performance with the state-of-the-art few-shot classification methods.

Based on this simple paradigm, our work is the first endeavor towards explicitly integrating the margin parameter to the softmax loss, and mostly importantly breaks the inherent understanding that the margin can be only restricted

as positive values, with both intuitive understanding and theoretical analysis. With an appropriate negative margin, our approach could achieve the state-of-the-art performance on three standard few-shot classification benchmarks.

Margin based Metric Learning. Metric learning aims to learn a distance metric between examples, and plays a critical role in many tasks, such as classification [49], clustering [51], retrieval [20] and visualization [24].

In practice, the margin between data points and the decision boundary plays a significant role in achieving strong generalization performance. [16] develops a margin theory and shows that the margin loss leads to an informative generalization bound for classification task. In the past decades, the idea of margin-based metric learning has been widely explored in SVM [40], k-NN classification [49], multi-task learning [33], etc. In the deep learning era, many margin-based metric learning methods are proposed to enhance the discriminative power of the learned deep features, and show remarkable performance improvements in many tasks [20, 19, 30], especially in face verification [41, 21, 47, 6]. For example, SphereFace [21], CosFace [47], and ArcFace [6] enforce the intra-class variance and inter-class diversity by adding the margin to cosine softmax loss.

However, as the tasks of previous works are based on close-set scenarios, they limit the margin parameter as positive values [21, 47, 6], where making the deep features more discriminative could be generalized to the validation set and improve the performance. For open-set scenarios, such as few-shot learning, increasing the margin would not enforce the inter-class diversity but unfortunately enlarge the intra-class variance for novel classes, as shown in Fig. 2, which would hurt the performance. In contrast, an appropriate negative margin would better tradeoff the discriminability and transferability of deep features in novel classes, and obtain better performance for few-shot classification.

3 Methodology

In a few-shot classification task, we are given two sets of data with different classes, formulated as $I^b = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N^b}$ as the base training set with C^b base classes for the first training stage, and $I^n = \{(\mathbf{x}'_i, y'_i)\}_{i=1}^{N^n}$ as the novel training set with C^n novel classes for the second training stage. For the novel training set, each class has K samples, where $K = 1$ or 5, and $C^n = 5$ is the standard setting [3, 19, 25, 7, 44, 36, 35]. This is called C^n-way K-shot learning. Few-shot classification aims to learn both discriminative and transferable feature representations from the abundant labeled data in base classes, such that the features can be easily adapted for the novel classes with few labeled examples.

3.1 Negative-margin Softmax Loss

In image classification, the softmax loss is built upon the feature representation of deep networks $\mathbf{z}_i = f_\theta(\mathbf{x}_i) \in \mathbb{R}^D$ ($f_\theta(\cdot)$ denotes the backbone network with the parameters θ), its corresponding label y_i and the linear transform matrix $\mathbf{W} = [W_1, W_2, ..., W_{C^b}] \in \mathbb{R}^{D \times C^b}$. Recently, introducing the **large and positive**

margin parameter to the softmax loss is widely explored in metric learning [22, 47, 6]. Hence, we directly integrate the margin parameter to the softmax loss to learn the transferable metrics, aiming at benefiting the few-shot classification on novel classes. The general formulation of large-margin softmax loss is defined as

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

where m is the margin parameter, β denotes the temperature parameter which defines how much strength to enlarge the gap between the largest logit and other logits. And $\mathtt{s}(\cdot,\cdot)$ denotes the similarity function between two input vectors.

It's worth noting that all the previous works on large-margin softmax loss restrict the margin as positive values [22, 47, 6]. This is because that previous works focus on the close-set scenarios, the loss with larger margin leads to the smaller intra-class variance and the larger between-class variance, which will help to classify examples in the same classes. This is also validated in Figure 1, that the softmax loss with larger margin could improve the classification accuracy on the validation set of training classes.

However, the situations are different in the open-set scenarios. Learned metrics which are too discriminative to training classes may hurt their transferability to the novel classes. So applying appropriate negative margin to softmax loss aims to tradeoff the discriminability on training classes and the transferability to novel classes of the learned metrics.

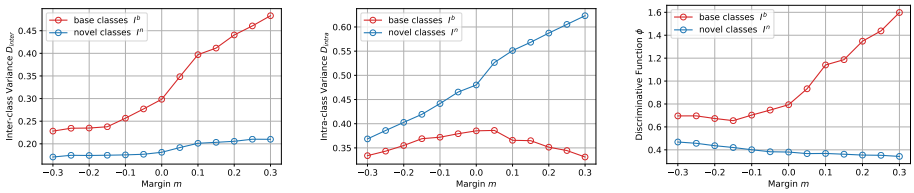
Here we formulate two instantiations of Eqn. 1 with different similarity functions. By taking the inner-product similarity $\mathtt{s}\left(\mathbf{z}_i,W_j\right)=W_j^T\mathbf{z}_i$ into Eqn. 1, the **negative-margin softmax loss** (abbreviated as Neg-Softmax) could be obtained. By taking the cosine similarity $\mathtt{s}\left(\mathbf{z}_i,W_j\right)=\frac{W_j^T\mathbf{z}_i}{\|\mathbf{z}_i\|\|W_j\|}$ into Eqn. 1, we can formulate the **negative-margin cosine softmax loss** (abbreviated as Neg-Cosine). The detailed loss functions could be found at the Appendix. These two loss functions are adopted at the pre-training stage.

3.2 Discriminability analysis of deep features w.r.t different margins

We analyze the discriminability of the deep features extracted by the deep model with different margins, to understand why negative margin works well on novel classes. For simplicity, we only analyze the cosine softmax loss, and it is direct to extend the analysis and conclusion to standard softmax loss.

We denote the pre-trained backbone network trained with margin parameter m as $f_{\theta(m)}$. For class j in base classes or novel classes, denote the set of examples labeled with class j as $I_j=\{(x_i,y_i)|y_i=j\}$. We compute the class center $\mu(I_j,m)$ for class j as the mean of the L2-normalized feature embeddings as

$$\mu(I_j,m)=\frac{1}{|I_j|}\sum_{(\mathbf{x}_i,y_i)\in I_j}\frac{f_{\theta(m)}\left(\mathbf{x}_i\right)}{\|f_{\theta(m)}\left(\mathbf{x}_i\right)\|_2}. \tag{2}$$



(a) D_{inter} (b) D_{intra} (c) ϕ

Fig. 2. Inter-class variance D_{inter}, intra-class variance D_{intra}, and discriminative function ϕ w.r.t margin m on both base and novel classes of mini-ImageNet. As the margin increases, the features of base classes is more discriminative, while that of novel classes is less discriminative.

The dataset $I = I_1 \cup I_2 \cup \cdots \cup I_C$ with C classes could be base dataset I^b with a large number of base classes or novel dataset I^n with small number of novel classes (such as 5 for 5-way few shot learning). Then we define the inter-class variance $D_{\text{inter}}(I, m)$, and intra-class variance $D_{\text{intra}}(I, m)$ as

$$\begin{aligned} D_{\text{inter}}(I, m) &= \frac{1}{C(C-1)} \sum_{j=1}^{C} \sum_{k=1, k\neq j}^{C} \|\mu(I_j, m) - \mu(I_k, m)\|_2^2, \\ D_{\text{intra}}(I, m) &= \frac{1}{C} \sum_{j=1}^{C} \Big(\frac{1}{|I_j|} \sum_{(\mathbf{x}_i, y_i) \in I_j} \Big\| \frac{f_{\theta(m)}(x_i)}{\|f_{\theta(m)}(x_i)\|} - \mu(I_j, m) \Big\|_2^2 \Big). \end{aligned} \tag{3}$$

For every two classes, the inter-class variance is the squared L2 distance between their class centers. For each class, the intra-class variance is the squared L2 distance between every sample in this class and the class center.

If inter-class variance becomes larger or intra-class variance becomes smaller, the deep features would be more discriminative. So we follow [26] to define the discriminative function $\phi(I, m)$ as the inter-class variance divided by the intra-class variance:

$$\phi(I, m) = \frac{D_{\text{inter}}(I, m)}{D_{\text{intra}}(I, m)}. \tag{4}$$

To measure the discriminability of the deep features with different margins, we plot the inter-class variance D_{inter}, intra-class variance D_{intra}, and discriminative function ϕ w.r.t margin m on both the base and novel classes of mini-ImageNet, respectively. As shown in Fig. 2, for base classes (red curves), as the margin increases, the inter-class variance increases a lot, meanwhile the intra-class variance does not change much, so the features of base classes become more discriminative. This is widely observed in previous works [6, 47, 22], and motivates them to introduce large and also positive margin to softmax loss for close-set scenarios.

But for novel classes (blue curves), the situation is just on the contrary. As the margin increases, the inter-class variance does not change much, but the intra-class variance increases a lot, so the features of base classes become less

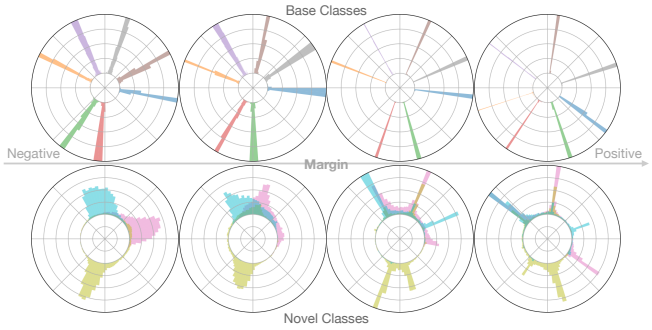


Fig. 3. The visualizations of the data distributions on angular space with different margins, on base classes (the first row) or novel classes (the second row) of MNIST. Plots from left to right denotes the margins from negative to positive. For each figure, we plot the histogram of the occurrence for each angle. Different colors denote the data points belonging to different classes.

discriminative. This indicates that larger margin may hurt the classification on the novel classes. This is also verified in the real few-shot classification task, shown as red curves in Fig. 1, larger and positive margin will achieve worse performance of few-shot classification on novel classes. Instead, the appropriate negative margin could achieve the best performance, which may lead to a better tradeoff on discriminability and transferability for novel classes.

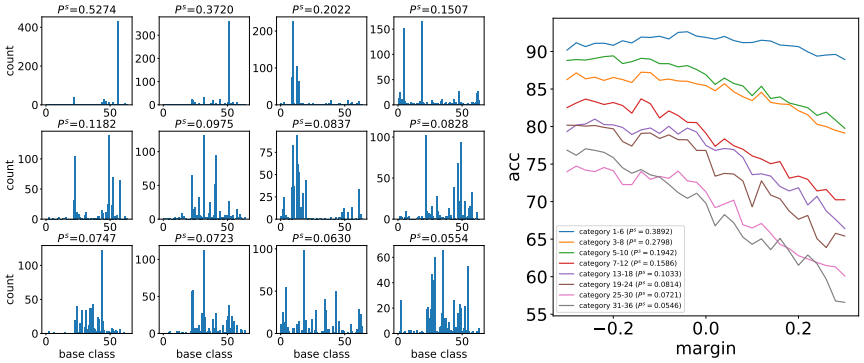
3.3 Intuitive Explanation

To better understand how the margin works, we perform the visualization on the data distributions in the angular space trained on MNIST[1], as shown in Fig. 3. We choose seven classes as the base classes for pre-training, and adopt the other three classes as the novel classes. We first train this deep model with 2-dimensional output features using cosine softmax loss with different margins on the base classes. Then we normalize the 2-D features to obtain the direction of each data point, and visualize the count of each direction (also known as the data distributions in angular space) on both base (first row) and novel classes (second row) using the models trained with different margins.

As shown in the first row in Fig. 3, with larger and even positive margin (from left to right), the clusters for each training class are getting thinner and higher, and the angle differences between different class centers are getting larger. This matches our previous observation in Fig. 2, that enlarging the margin leads to the smaller intra-class variance and larger inter-class variance on the base classes.

However, with larger margin, less data points would lie in the space far from all centers, which to some extent makes the output space much narrower. As

[1] This technique is widely used to characterize the feature embedding under the softmax-related objectives [47, 21, 54].



(a) (b)

Fig. 4. We first sort the 36 novel classes according to the probability of sample pairs in the same novel class j classified into the same base class P_j^s (one of every 3 categories are plotted for clarity) on mini-ImageNet. For each novel class, (a) shows the histogram of samples in this class to be classified to 64 base classes. (b) shows the accuracy curves w.r.t different margins for novel classes with different averaged P^s.

shown on the right side of the second row in Fig. 3, as novel classes are different to base classes, model with large margin may map the data points of the same class in novel classes to multiple peaks or clusters belonging to different base classes. Then the intra-class variance for novel classes would increase accordingly, making the classification of novel classes more difficult. Instead, as shown on the left side of second row in Fig. 3, the appropriate negative margin would not enforce the data points in novel classes too close to the training center, and may alleviate the multi-peak issue, which could benefit the classification on novel classes.

3.4 Theoretical Analysis

After giving the intuitive explanation that why negative margin works well on novel classes, we then prove this claim theoretically. Denote the parameter of the classifier joint pre-trained with backbone on base classes with margin m as $\mathbf{W}(m)$, the probability of a sample in the novel category j classified by pre-trained backbone $f_{\theta(m)}$ and classifier $W(m)$ as a base category k is

$$P_{jk}(m) = \frac{1}{|I_j|} \sum_{(x_i,y_i)\in I_j} \frac{\exp\left(\beta \mathtt{s}(f_{\theta(m)}(x), W_k(m))\right)}{\sum_{k'=1}^{C^b} \exp\left(\beta \mathtt{s}(f_{\theta(m)}(x), W_{k'}(m))\right)}, \tag{5}$$

where $\mathtt{s}(\cdot,\cdot)$ denotes the similarity function. The probability of a pair of samples in the same novel category j classified into the same base class is $P_j^s(m) = \sum_{k=1}^{C^b} P_{jk}^2(m)$. And the average probability of $P_j^s(m)$ is $P^s(m) = \frac{1}{|C^n|}\sum_{j=1}^{C^n} P_j^s(m)$.

Proposition. Assuming discriminative function for the base classes $\phi(I^b, m)$ is a monotonic increasing function w.r.t margin parameter m, and then we denote

$\phi^{-1}(I^b, m_1) - \phi^{-1}(I^b, m_2) = r \cdot (m_2 - m_1)$, where $m_2 > m_1$ and $r > 0$ is a scale variable. $\psi(m) = D_{\text{inter}}(I^n, m)/D_{\text{inter}}(I^b, m)$ is a monotonic decreasing function and we denote $\psi(m_1) - \psi(m_2) = t \cdot (m_2 - m_1), t > 0$.

Then $\forall 0 < P^s < \frac{t}{t(1-\phi^{-1}(I^b,m_1))+r\psi(m_1)}$, we have [2]:

$$\phi(I^n, m_2) < \phi(I^n, m_1). \tag{6}$$

The above proposition proves that the discriminative function on the novel classes $\phi(I^n, m)$ is a monotonic decreasing function w.r.t m under proper assumption and a measurable condition about the similarity between base and novel classes using P^s. The proposition indicates that an appropriate value of "negative" margin could work well for discriminating the samples in novel classes.

Fig. 4 shows the actual behavior of mini-ImageNet dataset. We first sort the 36 novel classes according to the probability of sample pairs in the same novel class j classified into the same base class P_j^s (one of every 3 categories are plotted for clarity) on mini-ImageNet. And the histograms of the samples in novel classes to be classified to 64 base classes is shown in Fig. 4(a). Fig. 4(b) shows the accuracy curves w.r.t different margins for novel classes with different averaged P^s. With smaller P^s, the histograms of novel classes become more diverse (shown in Fig. 4(a)) and their accuracies become lower (shown in Fig. 4(b)). Importantly, most subsets of novel classes favor negative margins, implying the condition in the Proposition is not hard to reach.

3.5 Framework

Following the standard transfer learning paradigm [52, 8], we adopt 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. This pipeline is widely adopted in recent few-shot learning methods [3, 7, 25].

In the pre-training stage, we aim at training the backbone network $f_\theta(\cdot)$ with abundant labeled data I^b in base classes, driven by metric learning loss, such as softmax loss in [3]. In our paper, we adopt the negative-margin softmax loss, which could learn more transferable representations for few-shot learning. In the fine-tuning stage, as there are only few labeled samples in I^n for training (e.g. 5-way 1-shot learning only contains 5 training samples), we follow [3] to fix the parameters of the backbone $f_\theta(\cdot)$, and only train a new classifier from scratch by the softmax loss. Note that, the computation of similarity (such as inner-product similarity or cosine similarity) in softmax loss is the same as that in the pre-training stage.

[2] Proof is attached in the supplemental material.

4 Experiments

4.1 Setup

Datasets and scenarios. Following [3], we address the few-shot classification problem under three different scenarios: **(1)** generic object recognition; **(2)** fine-grained image classification; and **(3)** cross-domain adaptation.

For the generic scenario, the widely-used few-shot classification benchmark: mini-ImageNet, is used to evaluate the effectiveness of the proposed Negative-Margin Softmax Loss. The mini-ImageNet dataset, firstly proposed by [45], consists of a subset of 100 classes from the ILSVRC-2012 [5], and contains 600 images for each classes. Following the commonly-used evaluation protocol of [37], we split the 100 classes into 64 base, 16 validation, and 20 novel classes for pre-training, validation, and testing. To validate the effectiveness of our model on the large dataset, we further conduct ablation study on the ImageNet-1K dataset following the setting in [12, 48].

For the fine-grained image classification, we use CUB-200-2011 dataset [46] (hereinafter referred as CUB), which consists of 200 classes and 11,788 images in total. Followingit stg the standard setting of [14], we split the classes in the dataset into 100 base classes, 50 validation classes, and 50 novel classes.

For the cross-domain adaptation scenario, we use mini-ImageNet $\rightarrow$ CUB [3], in which the 100 classes in mini-ImageNet, the 50 validation and 50 novel classes in CUB are adopted as base, validation and novel classes respectively, to evaluate the performance of the proposed Negative-Margin Softmax Loss in the presence of domain shift.

Implementation details. For fair comparison, we evaluate our model with four commonly used backbone networks, namely Conv-4 [45], ResNet-12 [32], ResNet-18 [3] and WRN-28-10 [25, 53]. Besides the differences in network depth and architecture, the expected input size of Conv-4 and ResNet-12 is 84×84, and that of ResNet-18 is 224×224, while WRN-28-10 takes 80×80 images as input.

Our implementation is based on PyTorch [34]. In the training stage, the backbone network and classifier are trained from scratch, with a batch size of 256. The models are trained for 200, 400 and 400 epochs in the CUB, mini-ImageNet and mini-ImageNet $\rightarrow$ CUB, respectively. We adopt the Adam [15] optimizer with initial learning rate 3e-3 and cosine learning rate decay [23]. We apply the same data argumentation as [3], including random cropping, horizontal flipping and color jittering.

In the fine-tuning stage, each episode contains 5 classes and each class contains 1 or 5 support images to train a new classifier from scratch and 16 query images to test the accuracy. The final performance is reported as the mean classification accuracy over 600 random sampled episodes with the 95% confidence interval. Note that all the hyper-parameters are determined by the performance on the validation classes.

Table 1. Few-shot classification results on the mini-ImageNet dataset. † indicates the method using the combination of base and validation classes to train the meta-learner

Backbone	Method	1 shot	5 shot
Conv-4	MAML [9]	48.70 ± 1.84	63.11 ± 0.92
	ProtoNet [43]	49.42 ± 0.78	68.20 ± 0.66
	MatchingNet [45]	48.14 ± 0.78	63.48 ± 0.66
	RelationNet [44]	50.44 ± 0.82	65.32 ± 0.70
	MAML+Meta-dropout [18]	51.93 ± 0.67	67.42 ± 0.52
	R2D2 [1]	51.20 ± 0.60	68.80 ± 0.10
	Neg-Softmax (ours)	47.65 ± 0.78	67.27 ± 0.66
	Neg-Cosine (ours)	**52.84 ± 0.76**	**70.41 ± 0.66**
ResNet-12	SNAIL [27]	55.71 ± 0.99	68.88 ± 0.92
	TADAM [32]	58.50 ± 0.30	76.70 ± 0.30
	MetaOptNet-SVM [19]	62.64 ± 0.61	78.63 ± 0.46
	Neg-Softmax (ours)	62.58 ± 0.82	80.43 ± 0.56
	Neg-Cosine (ours)	**63.85 ± 0.81**	**81.57 ± 0.56**
ResNet-18	SNCA [50]	57.80 ± 0.80	72.80 ± 0.70
	Baseline [3]	51.75 ± 0.80	74.27 ± 0.63
	Baseline++ [3]	51.87 ± 0.77	75.68 ± 0.63
	Neg-Softmax (ours)	59.02 ± 0.81	78.80 ± 0.61
	Neg-Cosine (ours)	**62.33 ± 0.82**	**80.94 ± 0.59**
WRN-28-10	Activation to Parameter† [36]	59.60 ± 0.41	73.74 ± 0.19
	LEO† [39]	61.76 ± 0.08	77.59 ± 0.12
	Fine-tuning [7]	57.73 ± 0.62	78.17 ± 0.49
	Cosine + rotation [11]	**62.93 ± 0.45**	79.87 ± 0.33
	Neg-Softmax (ours)	60.04 ± 0.79	80.90 ± 0.60
	Neg-Cosine (ours)	61.72 ± 0.81	**81.79 ± 0.55**

4.2 Results

Results on mini-ImageNet. For the generic object recognition scenario, we evaluate our methods on the widely-used mini-ImageNet dataset. For fair comparison with existing methods which uses different network architecture as backbone, we evaluate our methods with all four commonly used backbone networks. The 5-way 1-shot and 5-shot classification results on the novel classes of the mini-ImageNet dataset are listed in Table 1. We find that by simply adopting appropriate negative margin in standard softmax loss, our Neg-Softmax achieves competitive results with the existing state-of-the-art methods. It is worth noting that our Neg-Cosine achieves the state-of-the-art performance for both 1-shot and 5-shot settings on almost all four backbones on mini-ImageNet.

Results on CUB. On the fine-grained dataset CUB, we compared the proposed method with several state-of-the-art methods with ResNet-18 as backbone. The results are showed in Table 2, in which the results of the comparison methods are directly borrowed from [3]. It shows that the proposed Neg-Cosine outperforms all the comparison methods on both 1-shot and 5-shot settings. Furthermore,

Table 2. The few-shot classification accuracy on the novel classes (also known as test classes) of the CUB dataset and cross-domain setting with ResNet-18 as the backbone

Method	CUB		mini-ImageNet→CUB
	1 shot	5 shot	5 shot
MAML [9]	69.96 ± 1.01	82.70 ± 0.65	51.34 ± 0.72
ProtoNet [43]	71.88 ± 0.91	87.42 ± 0.48	62.02 ± 0.70
MatchingNet [45]	72.36 ± 0.90	83.64 ± 0.60	53.07 ± 0.74
RelationNet [44]	67.59 ± 1.02	82.75 ± 0.58	57.71 ± 0.73
Baseline [3]	65.51 ± 0.87	82.85 ± 0.55	65.57 ± 0.70
Baseline++ [3]	67.02 ± 0.90	83.58 ± 0.54	62.04 ± 0.76
Neg-Softmax (ours)	71.48 ± 0.83	87.30 ± 0.48	**69.30 ± 0.73**
Neg-Cosine (ours)	**72.66 ± 0.85**	**89.40 ± 0.43**	67.03 ± 0.76

Neg-Softmax also achieves highly competitive performance on both 1-shot and 5-shot settings.

Results on mini-ImageNet → CUB. In the real-world applications, there may be a signification domain shift between the base and novel classes. So we evaluate our methods on a cross domain scenario: mini-ImageNet → CUB, where we pre-train the backbone on a generic object recognition dataset, and transfer it to a fine-grained dataset. We follow [3] to report the 5-shot results with ResNet-18 backbone, as shown in Table 2. We can observe that both Neg-Softmax and Neg-Cosine are significantly better than all the comparison methods. Specifically, Neg-Softmax outperforms Baseline [3], the state-of-the-art method on the mini-ImageNet → CUB, by a large margin of 3.73%.

Results on ImageNet 1K dataset. To validate that negative margin works well on large dataset, we follow [48, 12] to run an ablation study on the ImageNet-1K dataset. We train ResNet-10 with standard cosine softmax loss and proposed Neg-Cosine for 90 epochs on the base classes. The learning rate starts at 0.1 and is divided by 10 every 30 epochs. The weight decay is 0.0001 and the temperature factor is 15. In the fine-tuning stage, we train a new linear classifier using SGD for 10000 iterations. The top-5 accuracy is reported in Table 4, which shows that the accuracies of Neg-Cosine are consistently better than standard cosine softmax loss with margin = 0 and LogReg [48].

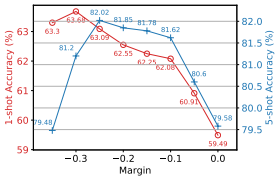
4.3 Analysis

This section presents a comprehensive analysis of the proposed approach. In the following experiments, we use Neg-Cosine with ResNet-18 backbone as default.

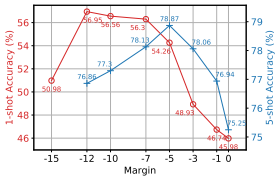
Effects of negative margin. Table 3 shows the 1-shot and 5-shot accuracy of the standard softmax, cosine softmax and our proposed Neg-Softmax, Neg-Cosine on the validation classes of mini-ImageNet, CUB and mini-ImageNet → CUB. By adopting appropriate negative margin, Neg-Softmax and Neg-Cosine yields significant performance gains over standard softmax loss and cosine softmax loss on all three benchmarks. Interestingly, Neg-Cosine outperforms Neg-Softmax in the in-domain setting, such as mini-ImageNet and CUB, while Neg-

Table 3. The few-shot accuracy of standard softmax, cosine softmax and our proposed Neg-Softmax, Neg-Cosine on validation classes of three standard benchmarks

Method	mini-ImageNet		CUB		mini-ImageNet→CUB	
	1 shot	5 shot	1 shot	5 shot	1 shot	5 shot
Softmax	45.98±0.79	75.25±0.61	58.32±0.87	80.21±0.59	46.87±0.78	67.68±0.71
Neg-Softmax	56.95±0.82	78.87±0.57	59.54±0.88	80.60±0.57	47.74±0.73	68.58±0.70
Cosine	59.49±0.90	79.58±0.59	66.39±0.93	82.17±0.58	42.96±0.76	61.99±0.75
Neg-Cosine	63.68±0.86	82.02±0.57	69.17±0.85	85.60±0.56	44.51±0.85	64.04±0.75



(a) Neg-Cosine



(b) Neg-Softmax

Fig. 5. The 1-shot (on red) and 5-shot (on blue) accuracy on validation classes of mini-ImageNet w.r.t different margins in Neg-Cosine and Neg-Softmax

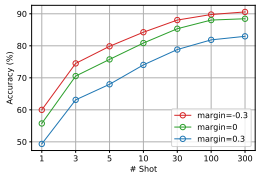


Fig. 6. Accuracy w.r.t # shots of validation classes on the mini-ImageNet dataset for margin = -0.3, 0 and 0.3

Softmax could achieve better performance than Neg-Cosine in the cross-domain setting. This is also observed in [3].

Accuracy w.r.t different margins. Figure 5 shows the 1-shot accuracy and 5-shot accuracy on validation classes of mini-ImageNet dataset w.r.t different margins in Neg-Cosine and Neg-Softmax. As we expect, as the margin gets negative and smaller, both the 1-shot accuracy and 5-shot accuracy of Neg-Cosine and Neg-Softmax first increase and then decrease, demonstrating a desirable bell-shaped curve. Hence, adopting appropriate negative margin yields significant performance gains over both standard softmax loss and cosine softmax loss on 1-shot and 5-shot classification of mini-ImageNet.

Various regularization techniques. Table 5 shows the importance of regularizations on Neg-Cosine, which reveals that integrating various regularization techniques steadily improves the 1-shot and 5-shot test accuracy on mini-ImageNet benchmark. Firstly, by simply adopting negative margin, the test accuracy increased by 5.74% and 4.37% on the 1-shot and 5-shot settings, respectively. Based on our approach, weight decay and DropBlock could further improve the performance. After integrating all regularizations together, our method achieves state-of-the-art accuracy of 62.33% and 80.94% for the 1-shot and 5-shot settings respectively on novel classes of mini-ImageNet.

More shots. We conduct an experiment by varying the number of shots from 1 (few shot) to 300 (many shot) and report the classification accuracy of the validation classes on the mini-ImageNet dataset in Figure 6. It shows that the test accuracy of margin=−0.3 is consistently higher than that of margin=0 from

Table 4. Top-5 accuracy on ImageNet 1K dataset with ResNet-10 as backbone.

Mehod	1 shot	5 shot
LogReg [48]	38.4	64.8
Cosine	42.1	64.0
Neg-Cosine	43.8	66.3

Table 5. Test accuracy on 5-way mini-ImageNet of various regularization techniques

negative margin	weight decay	drop block	1 shot	5 shot
			54.51 ± 0.79	75.70 ± 0.62
✓			60.25 ± 0.81	80.07 ± 0.58
✓	✓		62.21 ± 0.83	80.81 ± 0.59
✓	✓	✓	62.33 ± 0.82	80.94 ± 0.59

1-shot to 300-shot settings, which prove that the negative margin could benefit the open-set scenarios with more shots.

T-SNE visualization. Fig. 7 shows the t-SNE [24] visualizations. As shown in the first row, compared with negative margin, the feature embedding of zero and positive margin exhibit more discriminative structures and achieve better 1-shot accuracy on the base classes. However, the second row shows that enlarging the margin parameter would break the cluster structure of the novel classes and make the classification of novel classes harder. Instead, the appropriate negative margin retain the better cluster structure for novel classes. Thus the few-shot classification accuracy of negative margin is better than that of zero and positive margin.

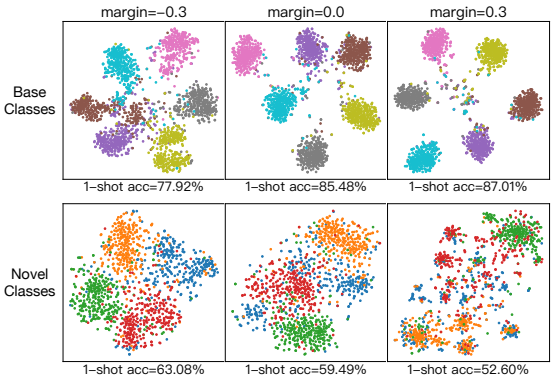


Fig. 7. The t-SNE visualizations of the feature embeddings and the corresponding 1-shot accuracy in the base and novel classes of mini-ImageNet dataset for the softmax loss with negative, zero and positive margin respectively

5 Conclusion

In this paper, we unconventionally propose to adopt appropriate negative-margin to softmax loss for few-shot classification, which surprisingly works well for the open-set scenarios of few-shot classification. We then provide the intuitive explanation and the theoretical proof to understand why negative margin works well for few-shot classification. This claim is also demonstrated via sufficient experiments. With the negative-margin softmax loss, our approach achieves the state-of-the-art performance on all three standard benchmarks of few-shot classification. In the future, the negative margin may be applied in more general open-set scenarios that do not restrict the number of samples in novel classes.

References

1. Bertinetto, L., Henriques, J.F., Torr, P., Vedaldi, A.: Meta-learning with differentiable closed-form solvers. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=HyxnZh0ct7
2. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
3. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773 (2017)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: 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)
7. Dhillon, G.S., Chaudhari, P., Ravichandran, A., Soatto, S.: A baseline for few-shot image classification. arXiv preprint arXiv:1909.02729 (2019)
8. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1126–1135. JMLR. org (2017)
10. Garcia, V., Bruna, J.: Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043 (2017)
11. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. arXiv preprint arXiv:1906.05186 (2019)
12. Hariharan, B., Girshick, R.: Low-shot visual recognition by shrinking and hallucinating features. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3018–3027 (2017)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
14. Hilliard, N., Phillips, L., Howland, S., Yankov, A., Corley, C.D., Hodas, N.O.: Few-shot learning with metric-agnostic conditional embeddings. arXiv preprint arXiv:1802.04376 (2018)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Koltchinskii, V., Panchenko, D., et al.: Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics **30**(1), 1–50 (2002)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
18. Lee, H.B., Nam, T., Yang, E., Hwang, S.J.: Meta dropout: Learning to perturb latent features for generalization. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=BJgd81SYwr

19. Lee, K., Maji, S., Ravichandran, A., Soatto, S.: Meta-learning with differentiable convex optimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10657–10665 (2019)
20. Liu, B., Cao, Y., Long, M., Wang, J., Wang, J.: Deep triplet quantization. 2018 ACM Multimedia Conference on Multimedia Conference - MM '18 (2018)
21. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
22. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. arXiv preprint arXiv:1612.02295 (2016)
23. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
24. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. JMLR (Nov 2008)
25. Mangla, P., Singh, M., Sinha, A., Kumari, N., Balasubramanian, V.N., Krishnamurthy, B.: Charting the right manifold: Manifold mixup for few-shot learning. arXiv preprint arXiv:1907.12087 (2019)
26. Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Mullers, K.R.: Fisher discriminant analysis with kernels. In: Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468). pp. 41–48. Ieee (1999)
27. Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: A simple neural attentive meta-learner. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=B1DmUzWAW
28. Munkhdalai, T., Yu, H.: Meta networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 2554–2563. JMLR. org (2017)
29. Munkhdalai, T., Yuan, X., Mehri, S., Trischler, A.: Rapid adaptation with conditionally shifted neurons. In: ICML. pp. 3661–3670 (2018)
30. Narayanaswamy, V.S., Thiagarajan, J.J., Song, H., Spanias, A.: Designing an effective metric learning pipeline for speaker diarization. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5806–5810. IEEE (2019)
31. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)
32. Oreshkin, B., López, P.R., Lacoste, A.: Tadam: Task dependent adaptive metric for improved few-shot learning. In: Advances in Neural Information Processing Systems. pp. 721–731 (2018)
33. Parameswaran, S., Weinberger, K.Q.: Large margin multi-task metric learning. In: Advances in neural information processing systems. pp. 1867–1875 (2010)
34. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017)
35. Qi, H., Brown, M., Lowe, D.G.: Low-shot learning with imprinted weights. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5822–5830 (2018)
36. Qiao, S., Liu, C., Shen, W., Yuille, A.L.: Few-shot image recognition by predicting parameters from activations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7229–7238 (2018)
37. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016), https://openreview.net/forum?id=rJY0-Kcll

38. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
39. Rusu, A.A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., Hadsell, R.: Meta-learning with latent embedding optimization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=BJgklhAcK7
40. Schölkopf, B., Smola, A.J., Bach, F., et al.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2002)
41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
43. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems. pp. 4077–4087 (2017)
44. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1199–1208 (2018)
45. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: Advances in neural information processing systems. pp. 3630–3638 (2016)
46. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
47. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5265–5274 (2018)
48. Wang, Y.X., Girshick, R., Hebert, M., Hariharan, B.: Low-shot learning from imaginary data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7278–7286 (2018)
49. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb), 207–244 (2009)
50. Wu, Z., Efros, A.A., Yu, S.X.: Improving generalization via scalable neighborhood component analysis. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 685–701 (2018)
51. Xing, E.P., Jordan, M.I., Russell, S.J., Ng, A.Y.: Distance metric learning with application to clustering with side-information. In: Advances in neural information processing systems. pp. 521–528 (2003)
52. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)
53. Zagoruyko, S., Komodakis, N.: Wide residual networks. Procedings of the British Machine Vision Conference 2016 (2016). https://doi.org/10.5244/c.30.87, http://dx.doi.org/10.5244/C.30.87
54. Zheng, Y., Pal, D.K., Savvides, M.: Ring loss: Convex feature normalization for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5089–5097 (2018)