



Learning discriminative features via weights-biased softmax loss

Xiaobin Li, Weiqiang Wang*

University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:

Received 22 April 2019

Revised 1 January 2020

Accepted 28 April 2020

Available online 31 May 2020

Keywords:

Classification

Softmax

CNNs

Fully connected layer units

Classifier weights

ABSTRACT

Loss functions play a key role in training superior deep neural networks. In convolutional neural networks (CNNs), the popular cross entropy loss together with softmax does not explicitly guarantee minimization of intra-class variance or maximization of inter-class variance. In the early studies, there is no theoretical analysis and experiments explicitly indicating how to choose the number of units in fully connected layer. To help CNNs learn features more fast and discriminative, there are two contributions in this paper. First, we determine the minimum number of units in FC layer by rigorous theoretical analysis and extensive experiment, which reduces CNNs' parameter memory and training time. Second, we propose a negative-focused weights-biased softmax (W-Softmax) loss to help CNNs learn more discriminative features. The proposed W-Softmax loss not only theoretically formulates the intra-class compactness and inter-class separability, but also can avoid overfitting by enlarging decision margins. Moreover, the size of decision margins can be flexibly controlled by adjusting a hyperparameter α . Extensive experimental results on several benchmark datasets show the superiority of W-Softmax in image classification tasks.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

In the early studies about image classification based on the convolutional neural networks (CNNs), there is no theoretical analysis and experiments explicitly indicating how to choose the number of units in FC layer. If the number of units in FC layer is too big, continued training can result in overfitting of the training data, increasing redundant parameter and training time, and if too small, training can result in underfitting of the training data. FC layer is a form of Artificial Neural Networks (ANN). Early works about determining the number of hidden units for an ANN model [1,2] mainly focus on the size of the training set and the number of input variables, which does not provide theoretical analysis. Without theoretical foundation in the number of nodes of CNNs' FC layer, researchers tend to choose a larger number of nodes. In YOLO [3], the number of units in FC layer is 1000 for dataset COCO [4] with 80 classes and RCNN series [5–7] network have 4096 units for dataset PASCAL VOC07+12 with 20 classes and dataset COCO, where the number of units is up to 200 times the number of classes. In this paper, we determine the minimum of FC layer number of units by rigorous theoretical analysis and extensive experiments for various classes tasks, which can reduce CNNs' parameter memory and training time.

In recent years, CNNs have been widely applied in many vision tasks like object recognition and segmentation [8–11], face verification [12] and hand-writing character recognition [13]. In the CNNs, the convolution layers together with pooling layers are generally used to extract discriminative feature representations, then fully connected layers implement the regression map from features to target labels, i.e., they involve two stages, features extraction and classification, as shown in Fig. 1.

In the aspect of feature representation learning, many effective techniques have been presented during the past decade. For example, the deeper and wider network architectures are built to improve the performances of CNNs [14,15]; different feature normalizations are adopted, like batch normalization [16], layer normalization [17], instance normalization [18] and group normalization [19]; diverse non-linear activation functions are exploited, like PReLU [20]; weights regularization [21] and stochastic pooling [22] are also investigated. However, all these techniques play a supporting role in extracting features fast and accurately, since the training of network is driven by loss calculated by loss function, whose inputs are the output of the last dense layer and the ground truth label. Now, overfitting is still a challenge to be addressed for CNNs.

In features classification subtask, FC layer with softmax loss is the mainstream where softmax loss tends to makes CNNs early stop in training. Actually, softmax function is sensitive to the size of input values, which is the main weakness of softmax loss. For example, considering the binary classification (referring

* Corresponding author.

E-mail addresses: lixiaobin161@mails.ucas.ac.cn (X. Li), wqwang@ucas.ac.cn (W. Wang).

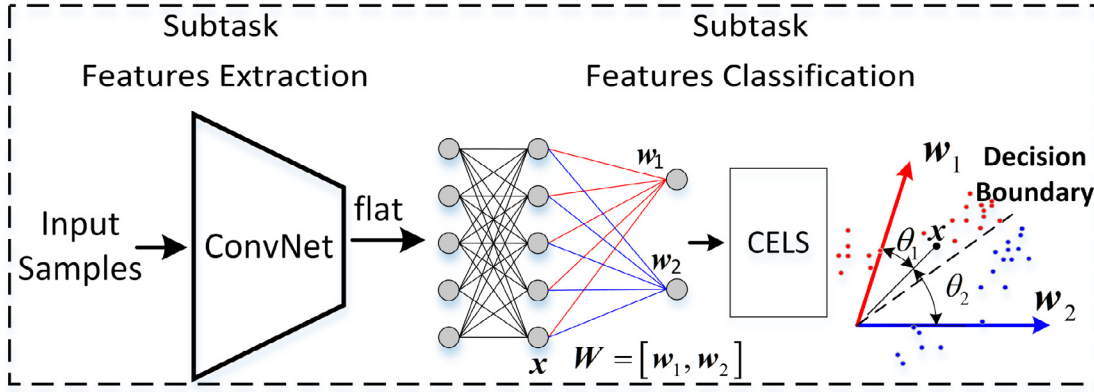


Fig. 1. Classification task with CNNs. 'CELS' denotes cross entropy loss with softmax.

to Fig. 3 $C = 2$ case), the decision boundary of conventional Softmax loss is depicted as $y_1 = y_2 (y_1 = \|\mathbf{w}_1\| \|\mathbf{x}\| \cos(\theta_1), y_2 = \|\mathbf{w}_2\| \|\mathbf{x}\| \cos(\theta_2))$, where \mathbf{w}_1 and \mathbf{w}_2 denote the weight vectors of two classes, \mathbf{x} denotes the feature representation for a given instance, θ_1 and θ_2 are the angles between weight vectors and feature. Here we suppose $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$, $\cos(\theta_1) = 0.05$ and $\cos(\theta_2) = -0.05$, which means feature \mathbf{x} is very close to the decision boundary. When we increase the norm of feature \mathbf{x} , e.g., let $\|\mathbf{x}\|$ equal 1, 10, 30, 50 respectively, $\text{softmax}([y_1, y_2]) = [0.52, 0.48]$, $[0.73, 0.27]$, $[0.95, 0.05]$, $[0.99, 0.01]$ correspondingly. So, feature \mathbf{x} with a large norm makes the Softmax loss decrease easily to zero, even if θ_1 is approximately equals to θ_2 . The distribution of features makes CNNs train easily and perform poorly in testing and that is where the inertia of CNNs exists.

We propose a new Softmax-like loss function, called the negative-focused weights-biased softmax (W-Softmax) loss, which has no extra trainable parameters compared with the conventional Softmax loss. By increasing the probabilities of all the negative classes in the softmax output, W-Softmax loss can help CNNs learn more discriminative features. Generally, while training c th class instances in the multi-class classification, each $\mathbf{w}_i (i \neq c)$ is replaced by normalized $\alpha \mathbf{w}_c + \mathbf{w}_i$ so that the decision boundaries between any two classes get separated and the decision margins are enlarged. Fig. 2 illustrates the idea via the example of two-class classification, where the decision margin increases when hyper-parameter α gets bigger. When training CNNs using the conventional softmax loss, the decision boundary between any two classes is coincident, and it brings premature convergence of CNNs in training when features distribute around the decision boundaries. However, by using the proposed the W-Softmax loss, the problem can be addressed, since the loss of CNNs will be enlarged if features locate around the original decision boundaries, separating the decision boundaries and enlarging the decision margins.

The W-Softmax loss can force features to draw close to the weight vectors of their corresponding class by increasing the value of $\alpha (\alpha \geq 0)$. A bigger α corresponds to a larger decision boundary margin, and the strong constraint tends to make intra-class variance decrease and inter-class variance increase. Compared with other works [23,24], the proposed loss function does not need to calculate the cosine values and use multiple-angle formula, thus it is computationally very efficient in the training and optimization, just as the conventional Softmax loss. In fact, the softmax loss is a special case of W-Softmax loss when $\alpha = 0$. The contributions of this work are summarized as follows:

1. We determine the minimum number of units in FC layer by rigorous theoretical analysis and extensive experiments for various classes tasks, which reduces CNNs' parameter memory and training time.
2. We present a new W-Softmax loss to make CNNs learn more discriminative features, and it can effectively improve the classification performance by avoiding premature convergence.
3. The size of decision margins can be optionally adjusted by a positive real-value parameter α . By increasing the value of α , CNNs can maximize inter-class variance and minimize intra-class variance. Extensive experiments on benchmark datasets show the effectiveness of W-Softmax loss.

2. Related works

2.1. Units in artificial neural networks

In image classification task based on convolutional neural networks (CNNs), fully connected (FC) layer is a common method in feature classification. In the early studies without CNNs, artificial neural networks (ANNs) are the focus in artificial intelligence and pattern recognition, where FC layer is a form of ANNs. Shai et al. [25] discussed the relationship between VC dimensions and hypothesis space and provided lots of theoretical proof in machine learning. Shaham et al. [26] applied VC dimension theory to derive a lower bound on the size of SpectralNet, a deep learning approach to spectral clustering. Murata et al. [1] studied the relation between the training error and the generalization error in terms of the number of the training examples and the complexity of a network which reduces to the number of parameters in the ordinary statistical theory of Akaike's information criterion (AIC). The number of hidden units is selected based on a given training set. Fletcher et al. [2] developed an algorithm to optimize the number of hidden nodes in feedforward artificial neural network by minimizing the mean square error over noisy training data, where network minimized the number of training sessions necessary for optimization of the number of hidden nodes. All these works focused on optimizing the number of hidden nodes of the whole ANNs. Differently, this paper is aimed at determining the number of units in FC layer for image classification task based on CNNs. The minimum number of FC layer units is determined by rigorous theoretical analysis and extensive experiments, which shows the minimum number varies from various classes in image classification.

2.2. Loss functions

The design of loss functions plays a significant role in training deep networks. Various loss functions have been presented and applied to learn discriminating feature representations. Contrastive loss [27] and triplet loss [28] need to carefully select instance pairs and triplet instances as the input of network in the train stage, since the performance of CNNs heavily depends on selected train-

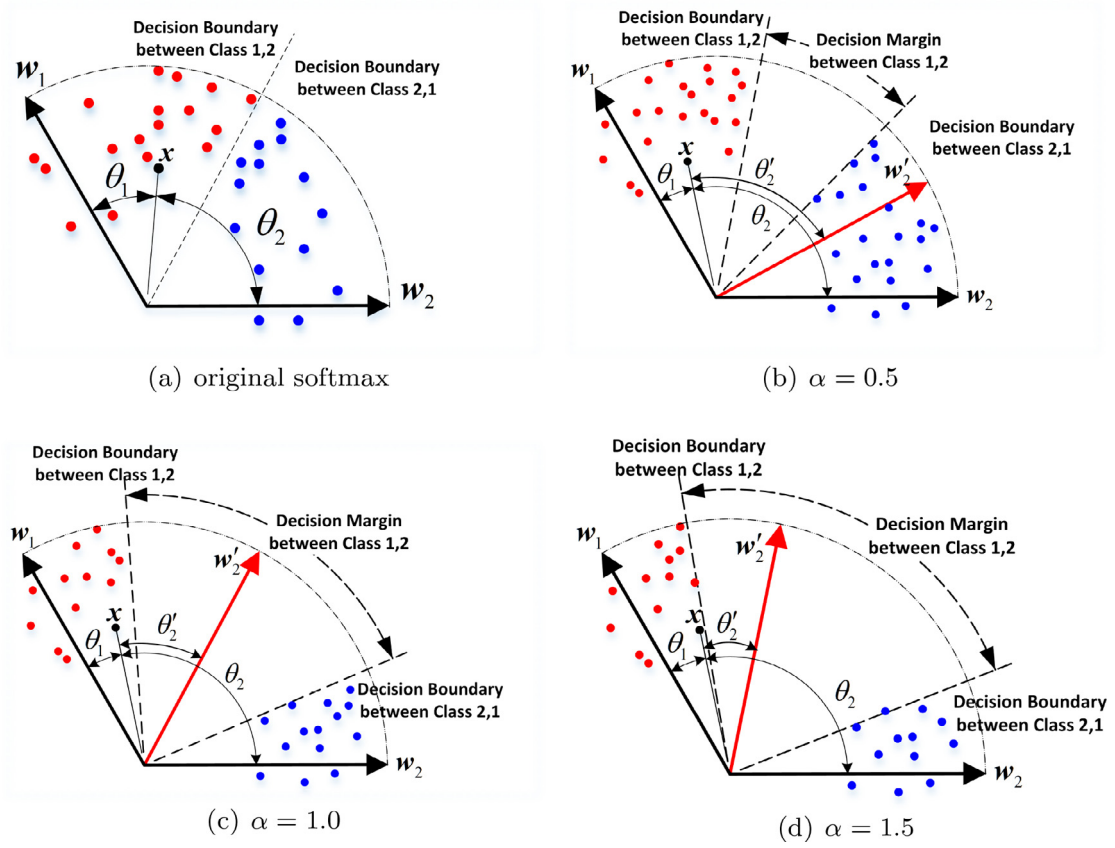


Fig. 2. The comparison of original softmax loss and W-Softmax loss when training instances with label 1. (a) is original softmax loss, where the decision boundary is coincident, and (b)-(d) are W-Softmax loss, where the decision boundaries get separated and $\mathbf{w}'_2 = \frac{\alpha \mathbf{w}_1 + \mathbf{w}_2}{\|\alpha \mathbf{w}_1 + \mathbf{w}_2\|}$.

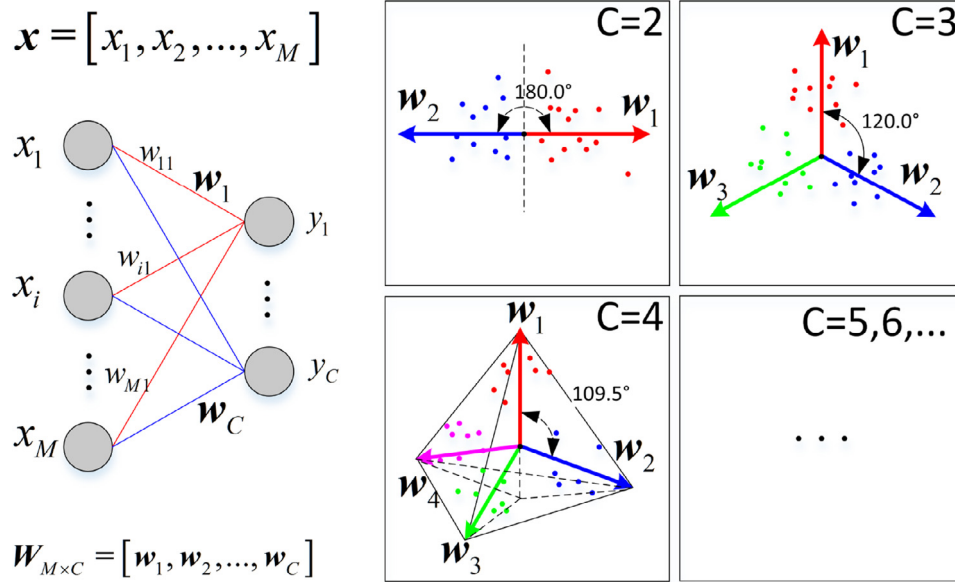


Fig. 3. The optimal weight vectors' distribution in classification tasks with various classes, where the angle between any two weight vectors is equal. C is the number of classes and M is the length of feature vector \mathbf{x} . $\mathbf{W}_{M \times C}$ is classifier weight matrix and column vector \mathbf{w}_i is classifier weight vector for class i . $y_i (y_i = \mathbf{w}_i^T \mathbf{x})$ is the output of class i .

ing instances. Similar to contrastive loss and triplet loss in increasing the Euclidean margin, Yang et al. [29] is a kind of k-nearest-neighbor (K-NN) method, which totally abandons the softmax layer and increases the burden of storages space and computation requirement. Center loss [30] together with softmax loss can help

CNNs reduce the intra-class variance and learn more discriminative features. Liu et al. [23] proposed a large-margin softmax loss and designed an angle function $\psi(\theta)$ related to m to decrease the probability of positive instances, which improves the feature discrimination. Liu et al. [24] used A-Ssoftmax loss to encourage a large an-

gular margin similar to L-Softmax. Differently, the A-Softmax loss normalized the weights by L_2 -norm, which has demonstrated its effectiveness on a series of open-set face recognition benchmarks. Both L-Softmax loss and A-Softmax loss are positive-focused softmax loss since both of them decrease the probability of positive class by enlarging the angle between features and weight vectors of positive class. Liu et al. [31] proposed a novel deep hyperspherical convolution network called SphereNet, adopting SphereConv to replace conventional convolution operator. In [31], the generalized angular softmax (GA-Softmax) loss is proposed based on weight-normalized softmax loss, where A-Softmax loss is the cosine GA-Softmax loss. A generic regularization called minimum hyperspherical energy (MHE) was proposed in [32] for neural networks. The theoretical analysis and experiments shown the effectiveness of MHE. Further, [32] proposed SphereFace+ by applying MHE to A-Softmax loss to improve the accuracy for face recognition. Similarly, our proposed weights-biased softmax (W-Softmax) loss is also based on the weight-normalized softmax loss. However, limited by the multiple-angle formula, the hyper parameter m in the losses based on L-Softmax and A-Softmax should be a positive integer. The hyper parameter \hat{A} in our proposed W-Softmax loss can be any positive real number.

Compared with the L-Softmax and A-Softmax losses, the proposed W-Softmax loss is a negative-focused softmax loss. We first remove the biases from the last FC layer and normalize weight vectors of all the classes by L_2 -norm and then evaluate the weight vectors of each negative class by Eq. (1).

$$\mathbf{w}'_i = \frac{\alpha \mathbf{w}_c + \mathbf{w}_i}{\|\alpha \mathbf{w}_c + \mathbf{w}_i\|}, (i \neq c) \quad (1)$$

where c is the index of positive classifier weight vector, i is the index of negative classifier weight vector, \mathbf{w}_c is positive classifier weight vector and \mathbf{w}_i is negative classifier weight vector. When training instance with label c , the weight matrix in the last FC layer is transformed as $\mathbf{W}' = [\mathbf{w}'_1, \dots, \mathbf{w}'_{c-1}, \mathbf{w}_c, \mathbf{w}'_{c+1}, \dots, \mathbf{w}'_C]$, where only the positive weight vector with true label c is not transformed. After the inner product between \mathbf{W}' and \mathbf{x} , we get the output of the last FC layer $\mathbf{f} = \mathbf{W}'^T \mathbf{x}$. And then \mathbf{f} is input into the softmax layer and the softmax loss is calculated the same as original softmax loss. In testing time, we use original classifier weight matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_{c-1}, \mathbf{w}_c, \mathbf{w}_{c+1}, \dots, \mathbf{w}_C]$ instead. In this case, we encourage the negative classes probabilities in softmax and increase their loss, which makes CNNs more stricter with the positive class and learn more discriminating features.

3. Determining the number of units in FC layer

For C -classes classification task, the feature vector extracted from convolutional network is \mathbf{x} with length M and classifier weight matrix without biased in FC layer is $\mathbf{W}_{M \times C} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_C]$ denoted as \mathbf{W}_C seen in Fig. 3. Geometrically, vector \mathbf{x} and \mathbf{w}_i are the points in \mathbb{R}^M (i.e. $\mathbf{x}, \mathbf{w}_i \in \mathbb{R}^M$). Theoretically, to encourage the inter-class separability, the distribution of weight vectors is optimal when the weight vectors are uniformly distributed in space, which means the angle between any two weight vectors is a constant value. To facilitate analysis, all the weight vectors are normalized by L_2 . For $i = 1, 2, \dots, C$, $\mathbf{w}_i = [w_{1i}, w_{2i}, \dots, w_{Mi}]^T$ and $\|\mathbf{w}_i\| = 1$. In the feature vector space \mathbb{R}^M , all the weight vectors are points on the hyper unit sphere and for all index i, j and k ($i \neq j, i \neq k, j \neq k$), there exists $d(\mathbf{w}_i, \mathbf{w}_j) = d(\mathbf{w}_i, \mathbf{w}_k)$, where $d(\cdot)$ is Euclidean distance function. So the solution is turned to how to determine the range of variable M to ensure that the problem that the angle between any two weight vectors is a constant value has a solution.

The problem is formulated by

$$\mathbf{w}_2^T \mathbf{w}_1 = \mathbf{w}_3^T \mathbf{w}_1 = \dots = \mathbf{w}_C^T \mathbf{w}_1$$

$$\begin{aligned} &= \mathbf{w}_3^T \mathbf{w}_2 = \dots = \mathbf{w}_C^T \mathbf{w}_2 \\ &\vdots \\ &= \mathbf{w}_{C-1}^T \mathbf{w}_C, \end{aligned} \quad (2)$$

The minimum value of M denoted as M_{\min} for C -classes task can be determined by mathematical induction. For continuously increasing number C , the weight matrix \mathbf{W}_C is constructed from a special solution \mathbf{W}_2 of Eq. (2), which is

$$\mathbf{W}_C = \begin{cases} \begin{bmatrix} 1 & -1 \end{bmatrix}_{1 \times 2} & C = 2 \\ \begin{bmatrix} \sqrt{\frac{(C-1)^2 - 1}{(C-1)^2}} \mathbf{W}_{C-1} & \mathbf{0} \\ -\frac{1}{C-1} & 1 \end{bmatrix}_{(C-1) \times C} & C > 2 \end{cases} \quad (3)$$

where each weight matrix \mathbf{W}_C satisfies the condition in Eq. (2) and for all i , $\|\mathbf{w}_i\| = 1$, and there exists

$$\mathbf{w}_i^T \mathbf{w}_j = \begin{cases} 1 & i = j, \\ -\frac{1}{C-1} & i \neq j. \end{cases} \quad (4)$$

In Eq. (3), the distribution of \mathbf{W}_2 , \mathbf{W}_3 and \mathbf{W}_4 are the cases $C = 2, 3, 4$ respectively shown in Fig. 3. Because the special case $\mathbf{W}_2 (M = 1)$ in Eq. (3) is the simplest case and the size of weight matrix \mathbf{W}_C is $M \times C$, we can determine the minimum value of M as $M_{\min} = C - 1$, which means if $M \geq C - 1$, the Eq. (2) has solution.

Next, we will prove that the $C - 1$ is the minimum value for M in construction. Reductio ad absurdum is adopted to prove the assumption.

Assumption 1. There is no unit vector $\mathbf{w}_{C+1} (C > 2)$ with length $C - 1$ making new weight matrix $\mathbf{W}'_{C+1} = [\mathbf{W}_C, \mathbf{w}_{C+1}]$ satisfy Eq. (2), where \mathbf{W}_C is from Eq. (3).

Proof. Suppose there is a unit vector \mathbf{w}_{C+1} with length $C - 1$ making new weight matrix $\mathbf{W}'_{C+1} = [\mathbf{W}_C, \mathbf{w}_{C+1}] = [\mathbf{w}_1, \dots, \mathbf{w}_C, \mathbf{w}_{C+1}]$ satisfy Eq. (2), described as

$$\mathbf{w}_1^T \mathbf{w}_{C+1} = \mathbf{w}_2^T \mathbf{w}_{C+1} = \dots = \mathbf{w}_{C-1}^T \mathbf{w}_{C+1} = \mathbf{w}_C^T \mathbf{w}_{C+1}. \quad (5)$$

According the construction, the rank of weight matrix \mathbf{W}_C is $C - 1$, there exists nonzero vector $\mathbf{a} = [a_1, a_2, \dots, a_{C-1}]$ satisfying $\mathbf{w}_C = a_1 \mathbf{w}_1 + a_2 \mathbf{w}_2 + \dots + a_{C-1} \mathbf{w}_{C-1}$. The same as \mathbf{w}_C , there exists nonzero vector $\mathbf{b} = [b_1, b_2, \dots, b_{C-1}]$ satisfying

$$\mathbf{w}_{C+1} = b_1 \mathbf{w}_1 + b_2 \mathbf{w}_2 + \dots + b_{C-1} \mathbf{w}_{C-1}. \quad (6)$$

Substituting Eq. (6) into Eq. (5) and then simplifying the equation by Eq. (4), we gets

$$\begin{aligned} &b_1 - \frac{1}{C-1} (b_2 + b_3 + \dots + b_{C-1}) \\ &= b_2 - \frac{1}{C-1} (b_1 + b_3 + \dots + b_{C-1}) \\ &\vdots \\ &= b_{C-1} - \frac{1}{C-1} (b_1 + b_2 + \dots + b_{C-2}) \\ &= -\frac{1}{C-1} (b_1 + b_2 + \dots + b_{C-1}). \end{aligned} \quad (7)$$

Because $C > 2$, Eq. (7) has only one solution $b_1 = b_2 = \dots = b_{C-1} = 0$, which is inconsistent with the assumption. Therefore, according to Reductio ad absurdum, assumption 1 is correct, which means that the length of weight vector \mathbf{w}_{C+1} is at least C . Under the condition of 2, the construction in Eq. (3) ensures that each weight vector \mathbf{w}_i in weight matrix \mathbf{W}_C has the minimum length $(C - 1)$, in other words $M_{\min} = C - 1$. \square

Extensive experiments on many benchmark datasets validate our conclusion.

4. Weights-biased softmax loss

4.1. Review of conventional softmax loss

In this section, we review the conventional softmax loss. Suppose we have a C -classes classification task. For a given instance with label c , its feature is \mathbf{x} . The probability for every class can be evaluated by

$$p_i = \frac{\exp(\mathbf{w}_i^T \mathbf{x} + b_i)}{\sum_{j=1}^C \exp(\mathbf{w}_j^T \mathbf{x} + b_j)}, \quad (8)$$

where \mathbf{w}_i and b_i denote the weights and biases of the last FC layer. In the prediction stage, an instance is classified to label i if $p_i > p_j$ (for all j and $j \neq i$). It can be converted as $\mathbf{w}_i^T \mathbf{x} + b_i > \mathbf{w}_j^T \mathbf{x} + b_j$, i.e., $\|\mathbf{w}_i\| \|\mathbf{x}\| \cos \theta_i + b_i > \|\mathbf{w}_j\| \|\mathbf{x}\| \cos \theta_j + b_j$, where θ_i denotes the angle between \mathbf{w}_i and \mathbf{x} , $0 \leq \theta_i \leq \pi$. The decision boundary of two classes i and j is defined by $\|\mathbf{w}_i\| \|\mathbf{x}\| \cos \theta_i + b_i = \|\mathbf{w}_j\| \|\mathbf{x}\| \cos \theta_j + b_j$. If we let $\|\mathbf{w}_i\| = 1$ and remove the biases, the decision boundaries become $\cos \theta_i = \cos \theta_j$, so the angle between weight vector \mathbf{w}_i of each class and feature \mathbf{x} is very important for classification.

The multi-class softmax loss for an instance \mathbf{x} can be formulated by

$$\begin{aligned} L &= -\log p_c = -\log \left(\frac{\exp(\mathbf{w}_c^T \mathbf{x} + b_c)}{\sum_{j=1}^C \exp(\mathbf{w}_j^T \mathbf{x} + b_j)} \right) \\ &= -\log \left(\frac{\exp(\|\mathbf{w}_c\| \|\mathbf{x}\| \cos \theta_c + b_c)}{\sum_{j=1}^C \exp(\|\mathbf{w}_j\| \|\mathbf{x}\| \cos \theta_j + b_j)} \right), \end{aligned} \quad (9)$$

where c is the class label of instance \mathbf{x} . The decision boundary for class c and class i can be defined by $\|\mathbf{w}_c\| \|\mathbf{x}\| \cos \theta_c + b_c = \|\mathbf{w}_i\| \|\mathbf{x}\| \cos \theta_i + b_i$ ($i \neq c$). Because the decision boundary between two classes is coincident, the conventional softmax loss cannot make CNNs learn a more discriminative feature representation. To encourage the ability of feature representation, we propose a new weights-biased softmax loss.

4.2. Weights-biased softmax loss

Positive Probability and Negative Probabilities. To obtain a large decision margin, we present a new loss called the weights-biased Softmax loss (W-Softmax), which utilizes parameter α in Eq. (1) to control the size of expected decision margin. Fig. 2 illustrates the basic principle of the proposed loss via an example of two-class classification, and it is also true for the case of multi-class classification. In our CNN network, we first remove the biases in the last FC layer of CNN and normalize the corresponding weight vectors, i.e., let $\|\mathbf{w}_i\| = 1$. For a given class c corresponding to \mathbf{w}_c , other classes are called negative class and each negative class i ($i \neq c$) has a corresponding weight \mathbf{w}_i' evaluated by Eq. (1), which is specifically used for evaluating the loss of instance \mathbf{x} in class c . It should be noted that, for two vectors \mathbf{p}, \mathbf{q} with angle $\theta \in [0, \pi]$ between them, if $\alpha > 0$ and $\mathbf{z} = \alpha \mathbf{p} + \mathbf{q}$, then vector must fall into angle θ , and vector \mathbf{z} will get closer to vector \mathbf{p} as α gets larger. So, \mathbf{w}_i' must fall into the included angle of \mathbf{w}_c and \mathbf{w}_i . Further, the angular bisector of \mathbf{w}_c and \mathbf{w}_i' forms a new decision boundary, which makes instances from class c become closer to \mathbf{w}_c and far away from \mathbf{w}_i in training. In C -classes classification, for input feature \mathbf{x} to the last FC layer with label c , the positive probability p_c and negative probabilities p_i ($i \neq c$) are evaluated by

$$\begin{aligned} p_c &= \frac{\exp(\|\mathbf{x}\| \cos \theta_c)}{\exp(\|\mathbf{x}\| \cos \theta_c) + \sum_{j \neq c}^C \exp(\|\mathbf{x}\| \cos \theta_j')}, \\ p_i &= \frac{\exp(\|\mathbf{x}\| \cos \theta_i')}{\exp(\|\mathbf{x}\| \cos \theta_c) + \sum_{j \neq c}^C \exp(\|\mathbf{x}\| \cos \theta_j')}, \end{aligned} \quad (10)$$

where θ_c and θ_i' denote the angles between weight vector \mathbf{w}_c and feature \mathbf{x} , as well as \mathbf{w}_i' and \mathbf{x} , respectively.

Decision Boundaries for Class c . In the training phase of learning features, for instance in class c , we use $\mathbf{W}' = [\mathbf{w}_1', \dots, \mathbf{w}_{c-1}', \mathbf{w}_c, \mathbf{w}_{c+1}', \dots, \mathbf{w}_C']$ to characterize the boundaries between class c and other classes. Concretely, let $p_c = p_i$ ($i \neq c$), we can easily derive $\theta_c = \theta_i'$, which means the decision boundary between class c and class i is the angular bisector of angle between \mathbf{w}_c and \mathbf{w}_i' . In testing phase, \mathbf{W}' is replaced by original weights $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c, \dots, \mathbf{w}_{c-1}, \mathbf{w}_C]$. Fig. 2 illustrates the decision boundaries for two classes, where the class 1 is considered as the positive class. In the conventional Softmax loss, the decision boundary margin is zero, so learned features from two classes likely distribute very close on both sides of their common decision boundary. In W-Softmax loss, the decision margins are magnified by parameter α in Eq. (1). It is clear that there exists $\theta_2 = \theta_2' + \theta_{\mathbf{w}_2', \mathbf{w}_2}$, where $\theta_{\mathbf{w}_2', \mathbf{w}_2}$ is the angle between weight vector \mathbf{w}_2' and \mathbf{w}_2 . Only if $\alpha = 0$, $\theta_{\mathbf{w}_2', \mathbf{w}_2} = 0$ and if $\alpha > 0$, $\theta_{\mathbf{w}_2', \mathbf{w}_2} > 0$. The following discussion is based on $\alpha > 0$. If the instance with label 1 is classified correctly, there exists $\|\mathbf{w}_1\| \|\mathbf{x}\| \cos \theta_1 > \|\mathbf{w}_2'\| \|\mathbf{x}\| \cos \theta_2'$, equivalent to $\cos \theta_1 > \cos \theta_2'$. Because of $\theta_2 = \theta_2' + \theta_{\mathbf{w}_2', \mathbf{w}_2} > \theta_2'$, it is satisfied that $\cos \theta_2' > \cos \theta_2$. Hence, in testing phase, it's satisfied that $\cos \theta_1 > \cos \theta_2' > \cos \theta_2$ by a large angular margin. As a result, there are two decision boundaries between any two classes with a large margin.

Weights-biased Softmax Loss. Based on previous discussion, for an instance \mathbf{x} from class c , we evaluate its W-Softmax loss by

$$\begin{aligned} L &= -\log \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\exp(\mathbf{w}_c^T \mathbf{x}) + \sum_{j \neq c}^C \exp(\mathbf{w}_j^T \mathbf{x})} \\ &= -\log \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\exp(\mathbf{w}_c^T \mathbf{x}) + \sum_{j \neq c}^C \exp\left(\frac{\alpha \mathbf{w}_c^T + \mathbf{w}_j^T}{\|\alpha \mathbf{w}_c + \mathbf{w}_j\|} \mathbf{x}\right)}, \end{aligned} \quad (11)$$

and it can be simplified as

$$L = -\log \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\sum_{j=1}^C \exp\left(\frac{\alpha \mathbf{w}_c^T + \mathbf{w}_j^T}{\|\alpha \mathbf{w}_c + \mathbf{w}_j\|} \mathbf{x}\right)}. \quad (12)$$

Further, for a set of instances $\{\mathbf{x}_k, k = 1, 2, \dots, N\}$, we can evaluate their average W-Softmax loss by

$$L = \frac{1}{N} \sum_k -\log \frac{\exp(\mathbf{w}_{c_k}^T \mathbf{x}_k)}{\sum_j \exp\left(\frac{\alpha \mathbf{w}_{c_k}^T + \mathbf{w}_j^T}{\|\alpha \mathbf{w}_{c_k} + \mathbf{w}_j\|} \mathbf{x}_k\right)}, \quad (13)$$

where c_k denotes the label of instance \mathbf{x}_k . Algorithm 1 summarizes the inference algorithm of CNNs training with W-Softmax loss in one batch input case, where CELS denotes cross entropy loss with softmax.

4.3. Discussion of hyperparameter α

In the proposed W-Softmax loss, parameter α plays an important role in regulating the decision angular margin. As shown in Fig. 2, the decision margin will be zero when $\alpha = 0$, and in this case, W-Softmax loss becomes the conventional softmax loss. As

Algorithm 1 CNNs training with W-Softmax loss in one batch input case.

Input: Training images \mathbf{I} with batch size B , labels $\mathbf{y} = [y_1, \dots, y_B]$;
 1: Gets extraction network N_{extr} and feature classification network N_{cls} where classifier weight matrix $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_C]$;
 2: Extracting features $\mathbf{X} = N_{extr}(\mathbf{I})$ as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_B]$;
 3: Initialize total loss $L \leftarrow 0$;
 4: **for** $i = 1 \rightarrow B$ **do**
 5: Transform classifier weight matrix, gets \mathbf{W}_i as $\mathbf{W}_i = [\mathbf{w}'_1, \dots, \mathbf{w}'_{y_i-1}, \mathbf{w}_{y_i}, \mathbf{w}'_{y_i+1}, \dots, \mathbf{w}'_C]$;
 6: Calculate instance loss $L_i \leftarrow CELS(\mathbf{W}_i^T \mathbf{x}_i, y_i)$;
 7: Add up instance loss $L \leftarrow L + L_i$;
 8: **end for**
 9: Update all weights in CNNs, $w \leftarrow w - learning_rate * \frac{\partial L}{\partial w}$.

the value of α increases, the decision margin also increases and the decision boundaries among different classes become more separated. It should be noted that, for L-Softmax loss and A-Softmax loss, hyper-parameter m is an integer, i.e., $m = 2, 3, 4, \dots$, while hyper-parameter α in our W-Softmax loss is a positive real number ($\alpha \geq 0$). Although a big value of α can make learned features more discriminative, it also increases the difficulty of training convergence, because it imposes a stronger constraint on the spatial distribution of learned features.

5. Experiments and results

5.1. Experimental setting

Datasets. We carry out the experiments on several standard benchmark datasets, MNIST [13], CIFAR10 [33], CIFAR100 [33] and LFW dataset [34]. MNIST dataset consists of 60,000 binary training images and 10,000 binary testing images, and the size of images is 28×28 . Each of datasets CIFAR10 and CIFAR100 consists of 50,000 color training images and 10,000 color testing images with image size 32×32 . LFW dataset is mainly for face recognition and face verification. In this paper, we focus on face verification part. LFW dataset has 13,233 training images covering 5749 people, only 1680 people with two or more images and 6000 pair images for testing.

CNN Setup. In order to compare expediently with the conventional original softmax loss and other existing losses, we use the CNN architecture presented by Liu et al. [23] as the backbone. Our experiments are carried with a Quadro P5000 GPU on TensorFlow. In convolution layers, the stride is 1 and PReLU [20] is chosen as the activation function. Momentum optimizer is used in training and the momentum is set to 0.9. We set the initial learning rate as 0.01 for MNIST and CIFAR10/CIFAR100, its exponential decay rate is 0.9 and the decay step is 6000. The weights are initialized by xavier initializer and the weight parameter of weights regularization is 0.0005. Batch normalization is used after PReLU and the dropout is not adopted for the sake of fair comparison.

Training Detail. To make the network converge quickly in training, we first train the CNNs using the conventional softmax loss, and refine the network using the proposed W-Softmax loss. For LFW dataset, there is an alignment process before training. In our experiment, the faces cropped from all the images are set to 160×160 and the alignment algorithm is adopted from MTCNN [35].

5.2. Effect of Units' number M in FC layer

Fig. 4 (a)–(c) illustrate the relation between the number of units in FC layer and classification accuracy on various datasets

Table 1

Test accuracy(%) of different methods on MNIST, where * denotes our proposed method.

Method	test accuracy(%)
CNN [36]	99.47
DropConnect [21]	99.43
FitNet [37]	99.49
NiN [38]	99.53
Maxout [39]	99.55
DSN [40]	99.61
R-CNN [41]	99.69
GenPool [42]	99.69
Hinge Loss	99.53
original softmax	99.58
L-Softmax [23]	99.69
W-Softmax ($\alpha=0.5$)*	99.64
W-Softmax ($\alpha=1$)*	99.67
W-Softmax ($\alpha=1.5$)*	99.69

with different classes. Fig. 4(d) is the comparison of the parameter memory in FC layer with different number of units. To visualize the relation detailedly, the scale on the horizontal axis is uneven. The value M_{\min} is 9 for MNIST dataset and CIFAR10 dataset and 99 for CIFAR100. Experimental results in Fig. 4 show that when $M < M_{\min}$, the accuracies increase with the increase of M and when $M \geq M_{\min}$, accuracies reach its maximum and fluctuate slightly around it, however, the parameter memory rapidly increases when M gets bigger. Notably, when $M = M_{\min}$, accuracies are close to or even maximum. In MNIST dataset, if M gets to $10x \sim 30x M_{\min}$, overfitting arises and accuracies with original softmax loss decline, while the accuracies in CNNs with W-Softmax loss remain almost unchanged, which distinctly eliminate effectively the overfitting.

5.3. Experiments and analysis on MNIST

In the experiments on MNIST, the batch size is set to 50. Table 1 lists the best results of different methods on MNIST. The results in Table 1 and Fig. 4(a) show the proposed W-Softmax loss has better performance than the conventional Softmax loss based on the same network architecture and can achieve the state-of-the-art performance compared with the other methods. The larger α in the W-Softmax loss can bring the higher accuracy to the trained CNN network.

Learned Features Comparison. We calculate the angles between learned features and the weight vectors of classifiers corresponding to their true categories, and then get the mean of these angles, i.e.,

$$\bar{\theta}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \arccos \frac{\mathbf{w}_i^T \mathbf{x}_i^{(j)}}{\|\mathbf{w}_i\| \|\mathbf{x}_i^{(j)}\|}, \forall i = 1, 2, \dots, C \quad (14)$$

where C is the number of classes in dataset, N_i is the number of instances with label i , $\mathbf{x}_i^{(j)}$ denotes the j th instance of class i , \mathbf{w}_i is the weight vector of classifier responsible for class i and $\bar{\theta}_i$ denotes the mean of angles between \mathbf{w}_i and all $\mathbf{x}_i^{(j)}$. The statistical results are shown in Fig. 5, where lower mean angle corresponds to the compactness of intra-class. We can see that the mean angle of each class with original softmax loss is larger than those in W-Softmax loss with different α , which denotes conventional softmax loss can not encourage the intra-class compactness. It's conspicuous that the mean angle gets smaller when α increases. The mean angle of training dataset is slightly less than the one of testing dataset for each class, and larger α in the W-Softmax loss can encourage the intra-class compactness and inter-class separability.

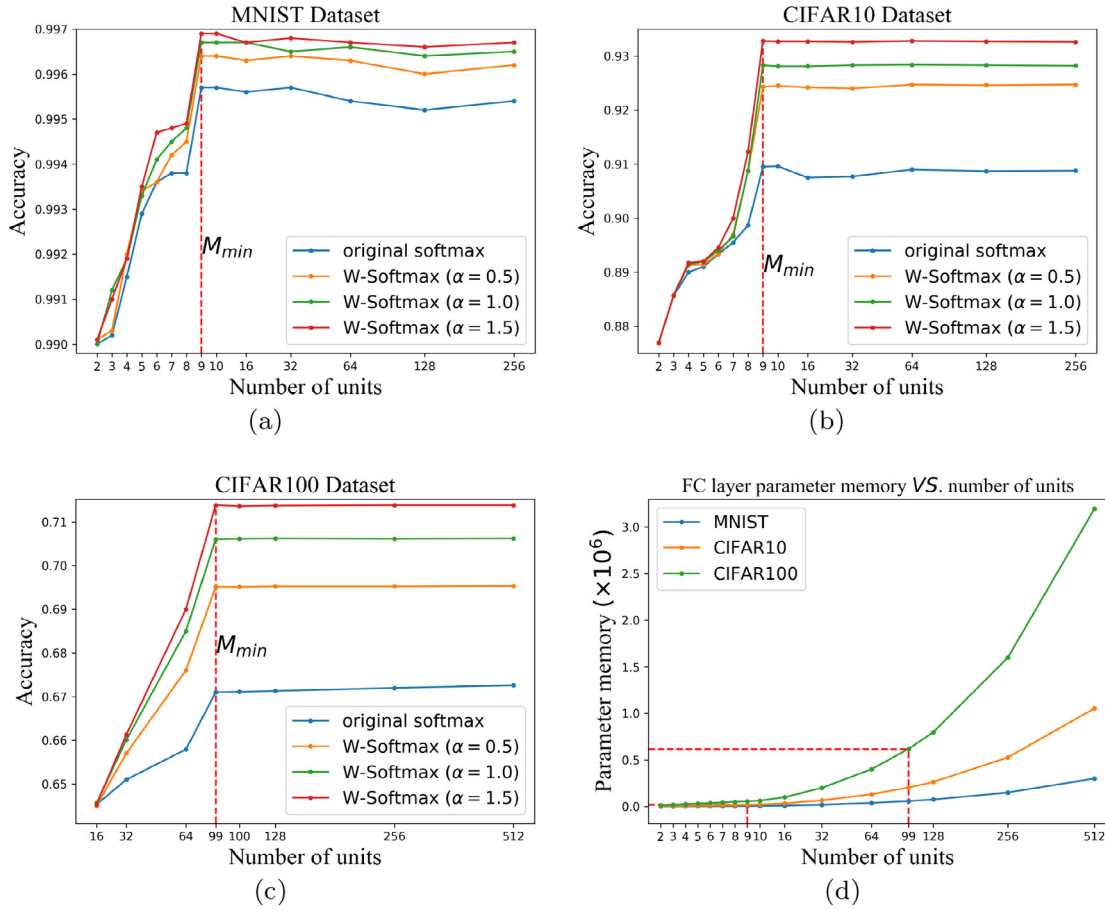


Fig. 4. Accuracy vs. the number of units M on different datasets with original softmax loss and W-Softmax loss (subfigures (a)–(c)). Subfigure (d) is the parameter memory in FC layer with various number of units. M_{min} is the theoretical minimum proved in previous section.

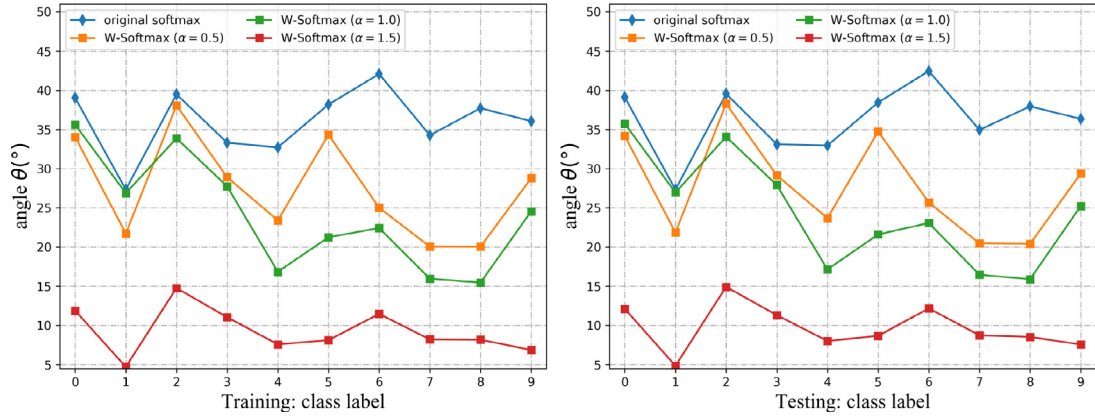


Fig. 5. Learned features comparison between original softmax loss and W-Softmax loss on MNIST dataset. The value in vertical axis is the mean angle $\bar{\theta}_i$ between \mathbf{w}_i and all \mathbf{x}_i for class label i . The left figure is the result on training dataset and the right one is on testing dataset.

5.4. Experiments on CIFAR10 and CIFAR100

Both CIFAR10 and CIFAR100 have 50,000 training instances and 10,000 testing instances. But, CIFAR10 has 5000 training instances for each class, and CIFAR100 only has 500. In the training, batch size is 256 for CIFAR10 and CIFAR100. The experimental results in Table 2 show that the W-Softmax loss achieves 2%–3% improvement on CIFAR10 and improves more than 4% accuracy on CIFAR100 compared with the conventional softmax loss.

5.5. Experiments on LFW dataset

The faces in LFW dataset are detected and aligned by MTCNN [35] and then cropped to 160x160. Before training and testing, each face image is normalized to $[-1, 1]$ by subtracting 127.5 and then dividing by 128. The feature extraction network is trained on a small training dataset that is the publicly available CASIA-WebFace [44] dataset containing 0.49M face images from 10,575 subjects. When training feature extraction network, batch size is set to 128 and the learning rate is initially 0.1 and divided by 10 for every

Table 2

Test accuracy(%) of different methods on CIFAR10 and CIFAR100, where N/A means the lack of comparative results.

Method	CIFAR10(%)	CIFAR100(%)
DropConnect [21]	90.59	N/A
FitNet [37]	N/A	64.96
NiN [38]	89.53	64.32
Maxout [39]	88.32	61.43
DSN [40]	90.31	65.43
All-CNN [43]	90.92	66.29
R-CNN [41]	91.31	68.25
GenPool [42]	92.38	67.63
Hinge Loss	90.09	67.10
original softmax	90.95	67.26
L-Softmax [23]	92.42	70.47
W-Softmax($\alpha=0.5$)*	92.47	69.53
W-Softmax($\alpha=1$)*	92.84	70.62
W-Softmax($\alpha=1.5$)*	93.28	71.38

Table 3

Face verification (%) on the LFW dataset, where * denotes the outside data is private (not publicly available). The base architecture is the same as [23] with 17 convolution layers.

Method	Outside Data	Accuracy(%)
FaceNet [28]	200M*	99.65
Deep FR [45]	2.6M	98.95
DeepID2+ [46]	300K*	98.97
L-Softmax [23]	WebFace	98.71
original softmax	WebFace	96.53
W-Softmax ($\alpha=0.5$)	WebFace	97.98
W-Softmax ($\alpha=1$)	WebFace	98.86
W-Softmax ($\alpha=1.5$)	WebFace	98.91

Table 4

Comparison in face verification (%) on the LFW dataset. The base architecture is the ResNet called SphereFace-64 [24] with 64 convolution layers.

Method	Accuracy(%)
Softmax Loss	97.88
Triplet Loss [28]	99.70
Softmax+Contrastive [27]	98.78
L-Softmax Loss [23]	99.05
A-Softmax Loss [24]	99.42
CosFace [47]	99.10
SphereFace+ [32]	99.47
W-Softmax ($\alpha=0.5$)	99.11
W-Softmax ($\alpha=1$)	99.41
W-Softmax ($\alpha=1.5$)	99.48

10k iterations, and training is stopped at 30k iterations. The cosine distance of features is adopted as the similarity score. The result is shown in Tables 3 and 4. Compared with the original softmax loss, the accuracy of W-SoftMax loss is greatly improved, which proves that encouraging intra-class compactness and inter-class separability is more conducive to improving the accuracy of face verification. The network with W-Softmax Loss has comparable performance as SphereFace+[32]. SphereFace+ applies MHE to A-Softmax loss which increases the complexity of the network and makes network training more difficult.

5.6. Experiments with multi-class

To explore the effect of W-Softmax when the number of classes increases, we design the experiments on MNIST and CIFAR10. Concretely, we randomly select k classes from 10 classes. Since the accuracy on MNIST almost reaches 100% when $k=2$, we choose k from 5 to 10, and specially select the first k classes from 10

Table 5

Effect of W-Softmax to multi-class number on MNIST and CIFAR10. The 'softmax' in table denotes original softmax loss.

class number	MNIST		CIFAR10	
	softmax	W-Softmax	softmax	W-Softmax
$k=5$	99.92	99.93	94.35	95.23
$k=6$	99.83	99.88	91.55	93.08
$k=7$	99.67	99.78	91.26	92.65
$k=8$	99.67	99.76	91.27	92.73
$k=9$	99.63	99.73	91.24	92.72
$k=10$	99.58	99.69	90.95	92.44

classes and set $\alpha=1$ in our experiments. The experimental results in Table 5 show that (1) the classification accuracy decreases as k increases; (2) when k is small, the advantage of W-Softmax loss over conventional Softmax is not obvious, and when k is big enough for a specific dataset, a big gain can be obtained.

6. Conclusion

In this paper, we theoretically determine the minimum number of nodes of classifier weight and verify this by experiments, which reduces the CNNs' parameter and training time. We present a new weights-biased Softmax(W-Softmax) loss, which is useful to build high-performance CNNs by learning highly discriminative features. By applying it, the decision margin can be flexibly adjusted by parameter α . The preliminary experiments show W-Softmax loss can achieve obvious improvement over conventional Softmax loss and obtain comparable or better classification accuracy in CNNs training compared with state-of-the-art loss functions.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled

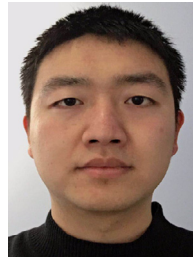
Acknowledgements

This work is supported by National Key R&D Program of China under contract No. 2017YFB1002203, NSFC projects under Grant 61976201, NSFC Key Projects of International (Regional) Cooperation and Exchanges under Grant 61860206004, and Ningbo 2025 Key Project of Science and Technology Innovation with No. 2018B10071.

References

- [1] N. Murata, S. Yoshizawa, S. Amari, Network information criterion-determining the number of hidden units for an artificial neural network model, IEEE Trans. Neural Netw. 5 (6) (1994) 865–872, doi:[10.1109/72.329683](https://doi.org/10.1109/72.329683).
- [2] L. Fletcher, V. Katkovnik, F.E. Steffens, A.P. Engelbrecht, Optimizing the number of hidden nodes of a feedforward artificial neural network, in: IJCNN, volume 2, 1998, pp. 1608–1612, doi:[10.1109/IJCNN.1998.686018](https://doi.org/10.1109/IJCNN.1998.686018).
- [3] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: CVPR, 2017, pp. 6517–6525, doi:[10.1109/CVPR.2017.690](https://doi.org/10.1109/CVPR.2017.690).
- [4] T.-Y. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, 2014, pp. 740–755, doi:[10.1007/978-3-319-10602-1_48](https://doi.org/10.1007/978-3-319-10602-1_48).
- [5] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: CVPR, 2014, pp. 580–587, doi:[10.1109/CVPR.2014.81](https://doi.org/10.1109/CVPR.2014.81).
- [6] R. Girshick, Fast R-CNN, in: ICCV, 2015, pp. 1440–1448, doi:[10.1109/ICCV.2015.169](https://doi.org/10.1109/ICCV.2015.169).
- [7] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, PAMI 39 (6) (2017) 1137–1149, doi:[10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031).

- [8] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: NIPS - Volume 1, Curran Associates Inc., USA, 2012, pp. 1097–1105. <http://dl.acm.org/citation.cfm?id=2999134.2999257>.
- [9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F.F. Li, Large-scale video classification with convolutional neural networks, in: CVPR, 2014, pp. 1725–1732, doi:10.1109/CVPR.2014.223.
- [10] R.Q. Charles, H. Su, M. Kaichun, L.J. Guibas, PointNet: deep learning on point sets for 3d classification and segmentation, in: CVPR, 2017, pp. 77–85, doi:10.1109/CVPR.2017.16.
- [11] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, PAMI (2018), doi:10.1109/TPAMI.2018.2844175.
- [12] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, DeepFace: closing the gap to human-level performance in face verification, in: CVPR, 2014, pp. 1701–1708, doi:10.1109/CVPR.2014.220.
- [13] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, doi:10.1109/5.726791.
- [14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778, doi:10.1109/CVPR.2016.90.
- [15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S.E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: CVPR, 2015, pp. 1–9, doi:10.1109/CVPR.2015.7298594.
- [16] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: ICML - Volume 37, in: ICML'15, JMLR.org, 2015, pp. 448–456. <http://dl.acm.org/citation.cfm?id=3045118.3045167>.
- [17] L.J. Ba, R. Kiros, G.E. Hinton, Layer normalization, CoRR (2016) arXiv:1607.06450.
- [18] D. Ulyanov, A. Vedaldi, V.S. Lempitsky, Instance normalization: the missing ingredient for fast stylization, CoRR, arXiv:1607.08022 (2016).
- [19] Y. Wu, K. He, Group normalization, in: ECCV, 2018, pp. 3–19, doi:10.1007/978-3-030-01261-8_1.
- [20] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: ICCV, 2015, pp. 1026–1034, doi:10.1109/ICCV.2015.123.
- [21] L. Wan, M.D. Zeiler, S. Zhang, Y. LeCun, R. Fergus, Regularization of neural networks using dropconnect, in: ICML 2013, Atlanta, GA, USA, 16–21 June 2013, 2013, pp. 1058–1066. <http://jmlr.org/proceedings/papers/v28/wan13.html>.
- [22] M.D. Zeiler, R. Fergus, Stochastic pooling for regularization of deep convolutional neural networks, ICLR, 2013 arXiv:1301.3557.
- [23] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: ICML, 2016, pp. 507–516. <http://jmlr.org/proceedings/papers/v48/liu16.html>.
- [24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: deep hypersphere embedding for face recognition, in: CVPR, 2017, pp. 6738–6746, doi:10.1109/CVPR.2017.713.
- [25] S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, USA, 2014.
- [26] U. Shaham, K.P. Stanton, H. Li, R. Basri, B. Nadler, Y. Kluger, SpectralNet: spectral clustering using deep neural networks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018. https://openreview.net/forum?id=HJ_aoCyRZ.
- [27] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, in: NIPS, 2014, pp. 1988–1996.
- [28] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: CVPR, 2015, pp. 815–823, doi:10.1109/CVPR.2015.7298682.
- [29] H.-M. Yang, X.-Y. Zhang, F. Yin, C.-L. Liu, Robust classification with convolutional prototype learning, in: CVPR, 2018, pp. 3474–3482, doi:10.1109/CVPR.2018.00366. http://openaccess.thecvf.com/content_cvpr_2018/html/Yang_Robust_Classification_With_CVPR_2018_paper.html.
- [30] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: ECCV, 2016, pp. 499–515, doi:10.1007/978-3-319-46478-7_31.
- [31] W. Liu, Y.-M. Zhang, X. Li, Z. Liu, B. Dai, T. Zhao, L. Song, Deep hyperspherical learning, in: I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA, 2017, pp. 3950–3960. <http://papers.nips.cc/paper/6984-deep-hyperspherical-learning>.
- [32] W. Liu, R. Lin, Z. Liu, L. Liu, Z. Yu, B. Dai, L. Song, Learning towards minimum hyperspherical energy, in: S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3–8 December 2018, Montréal, Canada, 2018, pp. 6225–6236. <http://papers.nips.cc/paper/7860-learning-towards-minimum-hyperspherical-energy>.
- [33] A. Krizhevsky, G. Hinton, *Learning Multiple Layers of Features from Tiny Images*, Tech. Rep. 1, Computer Science Department, University of Toronto, 2009.
- [34] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, Technical Report, University of Massachusetts, Amherst, 2007.
- [35] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503, doi:10.1109/LSP.2016.2603342.
- [36] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition? in: ICCV, 2009, pp. 2146–2153, doi:10.1109/ICCV.2009.5459469.
- [37] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, FitNets: hints for thin deep nets, ICLR, 2015 arXiv:1412.6550.
- [38] M. Lin, Q. Chen, S. Yan, Network in network, ICLR, 2014 arXiv:1312.4400.
- [39] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A.C. Courville, Y. Bengio, Maxout networks, in: ICML, 2013, pp. 1319–1327. <http://jmlr.org/proceedings/papers/v28/goodfellow13.html>.
- [40] C.-Y. Lee, S. Xie, P.W. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, AIS-TATS, 2015. <http://jmlr.org/proceedings/papers/v38/lee15a.html>.
- [41] M. Liang, X. Hu, Recurrent convolutional neural network for object recognition, in: CVPR, 2015, pp. 3367–3375, doi:10.1109/CVPR.2015.7298958.
- [42] C.-Y. Lee, P.W. Gallagher, Z. Tu, Generalizing pooling functions in convolutional neural networks: mixed, gated, and tree, in: AISTATS, 2016, pp. 464–472. <http://jmlr.org/proceedings/papers/v51/lee16a.html>.
- [43] J.T. Springenberg, A. Dosovitskiy, T. Brox, M.A. Riedmiller, Striving for simplicity: the all convolutional net, ICLR, 2015 arXiv:1412.6806.
- [44] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, CoRR (2014) arXiv:1411.7923.
- [45] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: BMVC, BMVA Press, 2015, pp. 41.1–41.12, doi:10.5244/C.29.41.
- [46] Y. Sun, X. Wang, X. Tang, Deeply learned face representations are sparse, selective, and robust, in: CVPR, 2015, pp. 2892–2900, doi:10.1109/CVPR.2015.7298907.
- [47] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, CosFace: large margin cosine loss for deep face recognition, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 5265–5274, doi:10.1109/CVPR.2018.00552. http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_CosFace_Large_Margin_CVPR_2018_paper.html.



XiaoBin Li is a Ph.D. student of School of Computer and Control Engineering in University of Chinese Academy of Sciences. He obtained his B.E. degree from NanKai University in 2016. His research interests include machine learning, pattern recognition, image processing and computer vision.



Weiqiang Wang received the B.E. and M.E. degrees in computer science from Harbin Engineering University, Harbin, China, in 1995 and 1998, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, in 2001. He is currently a Professor with the School of Computer and Controlling Engineering, University of CAS. His research interests include multimedia content analysis and computer vision.