IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2019, accepted February 1, 2019, date of publication February 5, 2019, date of current version February 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2897692

Large-Margin Regularized Softmax Cross-Entropy Loss

XIAOXU LI, DONGLIANG CHANG, TAO TIAN, AND JIE CAO
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

Corresponding author: Jie Cao (caoj@lut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61563030 and Grant 61763028, in part by the Natural Science Foundation of Gansu Province, China, under Grant 17JR5RA125, and in part by the Hong-Liu Outstanding Youth Talents Foundation of the Lanzhou University of Technology.

ABSTRACT Softmax cross-entropy loss with L2 regularization is commonly adopted in the machine learning and neural network community. Considering that the traditional softmax cross-entropy loss simply focuses on fitting or classifying the training data accurately but does not explicitly encourage a large decision margin for classification, some loss functions are proposed to improve the generalization performance by solving the problem. However, these loss functions enhance the difficulty of model optimization. In addition, inspired by regularized logistic regression, where the regularized term is responsible for adjusting the width of decision margin, which can be seen as an approximation of support vector machine, we proposed a large-margin regularization method for softmax cross-entropy loss. The advantages of the proposed loss are twofold as follows: the first is the generalization performance improvement, and the second is easy optimization. The experimental results on three small-sample datasets show that our regularization method achieves good performance and outperforms the existing popular regularization methods of neural networks.

INDEX TERMS Neural networks, cross-entropy loss, large-margin regularization.

I. INTRODUCTION

Over the past several years, deep learning has gained great success [1]–[6], especially in computer vision; convolutional neural networks (CNNs) have boosted the state-of-the-art performance in many visual recognition tasks [7]–[12]. Thus far, there also are many works that have improved classical CNNs from the aspects of data augmentation, loss function, network structure, optimization algorithm, activation function, decorrelation of neutral vector variables, and so on [5], [13]–[18].

Among studies on loss function improvement, many focus on learning features that simultaneously maximized their intra-class compactness and inter-class separability [19], [20]. Wen *et al.* [21] proposed center loss that is used to add into the normal supervision signal, such as cross-entropy (CE) loss. The loss aims to simultaneously learn a center for the deep features of each class and penalizes the distances between the deep features and their corresponding class centers for face recognition. Liu *et al.* [22] introduced cosine distance to replace the linear classification score in original CE loss and proposed congenerous cosine loss for person recognition.

Similar to congenerous cosine loss, large-margin loss also tried to change the classification score. The loss reduced compulsively classification score so that the learned features in each class can be more compact; therefore, the classification margin can be enlarged [23]. Furthermore, Schroff *et al.* [24] proposed that the triplet loss for face recognition minimizes the distance between a class center, a positive sample and samples with the same identity; the loss maximizes the distance between the class center and its negative samples.

Liu *et al.* [23] built on L-Softmax loss proposed that the angular Softmax loss pushes convolutional neural networks (CNNs) to learn angularly discriminative features. The main difference between L-Softmax loss and A-Softmax loss is that the classification score of L-Softmax is a linear score, while the classification score of the A-Softmax [25] is an angular score. GM-loss assumes that the deep features of all sample points in a dataset follow a Gaussian mixture distribution, and the sample points belonging to different classes follow different Gaussian distributions. The GM-loss involves

The associate editor coordinating the review of this manuscript and approving it for publication was Zhanyu Ma.

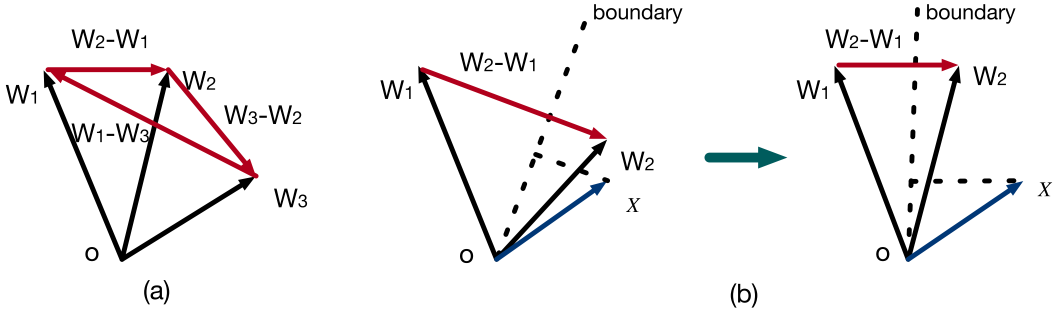

FIGURE 1. Subfigure (a) shows the regularized vectors in the proposed *large-margin regularized S-CE loss*. The regularized vectors are indicated in red. W_1, W_2 and W_3 represent the weight vectors of three classes in Softmax regression, and $W_i - W_j$ represents the regularized regression, which includes the vector difference of W_i and W_j, $i, j \in \{1, 2, 3\}$. Subfigure (b) illustrates how the regularization term enlarges the classification margin by taking the example of two-class classification.

two terms, a classification loss and a log likelihood regularization term, which is responsible for pushing the network to generate GM distributed features [26]. Actually, many real datasets do not follow Gaussian distribution [14], [27], [28]; therefore, the loss still has some limitations.

In addition, Softmax cross-entropy (S-CE) loss with L2 or L1 regularization is commonly adopted for alleviating overfitting in the machine learning and neural network community, especially when the training samples are insufficient. Regularization [29] is a method of introducing additional information to solve an ill-posed problem or prevent overfitting by limiting the solution space of a model. L2 regularization is equivalent to placing a Gaussian prior to weight parameters in the Softmax regression classifier, while L1 regularization is equivalent to placing a Laplace prior to weight parameters in the Softmax regression classifier. They limit the L2 or L1 norm of weight parameters, where the regularization term coefficient can adjust the effecting main loss function of the regularization term. Except for L1 and L2 regularization [30], Dropout [31]–[33] randomly drops or freezes some network neurons during the training process, so it can constrain the norm of some weights. DropConnect [34] randomly drops or freezes some connections of a network during the training process and can be seen as a general version of Dropout. Both Dropout and DropConnect do not have great effects on small-sample classification due to randomness.

This work builds on the idea that a regularized logistic regression can be seen as an approximation of a support vector machine (SVM) [35], [36], which has been proven in [37]. A SVM is a large-margin classifier, in which the regularized term is responsible for adjusting the width of the decision margin. Inspired by the relationship of regularized logistic regression and SVM, we proposed a large-margin regularization for S-CE loss to alleviate the overfitting of the neural network or other learning algorithms that used S-CE loss. The advantages of this proposed loss are the following: (*i*) it obtains a large-margin for classification and improve the generalization ability of the learning algorithm that used S-CE loss, and (*ii*) its optimization is without any constraint on parameter initialization and optimization algorithm of the model. The experimental results on three small-sample datasets show that the proposed regularized loss achieves good performance and outperforms the existing popular regularization methods of neural networks.

II. CE LOSS WITH A LARGE-MARGIN REGULARIZATION

Before we introduce the proposed loss, we first review the S-CE loss that is usually used as the loss function of neural networks. Denoting the training dataset as $D = \{(x_i, y_i)|i \in \{1, 2, ..., N\}\}$ and Y_i as one-hot vector of C classes, the nonzero dimension records the class label of sample x_i. W_j and b_j, $j \in \{1, 2, ..., C\}$ represent the parameter vector and bias of the jth class in Softmax regression, respectively, and the S-CE loss can then be written as follows:

$$L_{S-CE} = -\sum_{i=1}^{N} \log\left(\frac{exp({W_{y_i}}^T x_i + b_{y_i})}{\sum_{j=1}^{C} exp({W_j}^T x_i + b_j)}\right). \tag{1}$$

The optimization goal of a neural network is to minimize the loss. When the loss function reaches the minimum, 0, it simply means the model optimized by the loss can classify the training data 100 percent correctly and does not explicitly enlarge the decision margin.

A. LARGE-MARGIN REGULARIZED S-CE LOSS

A previous study [37] proved that regularized logistic regression can be seen as an approximation of SVM. Motivated by this theory, we focus on extending the regularized logistic regression to multiple class classification. Since the margin between a class and any other class needs to be enlarged, we simultaneously constrain the L2 norm difference vector of weight vectors of a class and any other class, and we propose *Large-margin Regularized S-CE Loss*, which is shown in (2).

$$L_{RCE} = L_{S-CE} + \beta \sum_{i \neq j} ||W_i - W_j||^2, \tag{2}$$

where L_{S-CE} is responsible for classifying the training samples correctly. W_i and W_j, $i, j \in \{1, 2, ..., C\}$, are the parameter vectors of the ith and jth class in Softmax regression, and $W_i - W_j$, the vector difference of W_i and W_j,

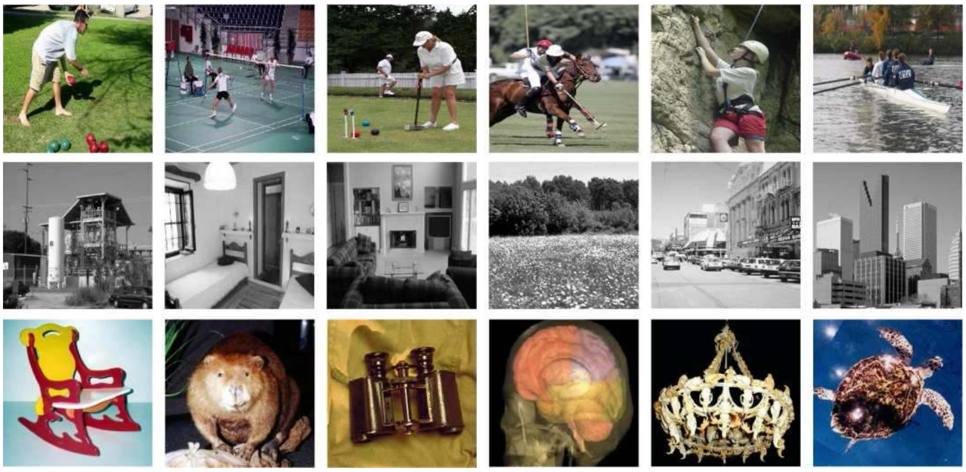

FIGURE 2. Examples of the UIUC-Sport dataset (the upper row), the 15Scenes dataset (the second row) and the Caltech101 dataset (the bottom row). The UIUC-Sport dataset is an event classification dataset, the 15Scenes dataset is a scene classification dataset, and the Caltech101 dataset is an object classification dataset.

are the regularized vectors. Figure 1 shows the regularized vectors and how the regularized vectors could enlarge the classification margin.

B. DISCUSSION

In this section, we comment on the proposed loss. First, our loss function is easy to optimize due to the quadratic regularized term. Compared with L-Softmax loss, it has no special requirements on optimization algorithm and parameters initialization.

Second, the proposed loss is a general version of the regularized logistic regression [37]. When the number of classes equals 2, the proposed loss will degrade into the regularized logistic regression. The regularization term limits the L2 norm of the difference vector of any parameter vectors in Softmax regression; thus, it enlarges the distances from sample points to decision boundary between the class that the sample points belong to other classes. It is noted that generalizing the regularized logistic regression is not straightforward because the correspondence between the regularized parameter vector in logistic regression and the one in Softmax regression cannot be derived easily.

Third, the proposed loss is more adapted to the situation in which features have been fixed or compact features are difficult to learn. There exists an implicit assumption under our loss, that is, the feature embedding of data points is totally frozen or has little change, which is different from L-Softmax loss, GM-Softmax loss and A-Softmax loss. In these losses, the underground assumption is that the features of the sample points are learnable. Unlike these loss functions, the proposed loss focuses on constraint weight vectors in the Softmax classification layer to obtain a large decision margin.

III. EXPERIMENTAL RESULTS

For the experiments, in order to fully evaluate the proposed method, we compare it with four methods and use three challenging small-sample datasets. We especially compare these methods with the following four aspects: classification accuracies, paired Student's t-test, feature visualization, and the effect of varying the activation function and initialization.

TABLE 1. Dataset statistics in this paper: category, training data and test data sizes.

Datasets	#Category	#Training	#Test
UIUC	8	749	749
15Scenes	15	2,240	2,240
Caltech101	101	4,310	4,310

A. DATASETS AND FEATURES

1) DATASETS

We conduct experiments on three challenging small-sample datasets, including the UIUC-Sports dataset (UIUC) [38], the 15 Scenes dataset (15Scenes) [39], and the Caltech 101 dataset (caltech101) [40], which are widely used to evaluate small-sample image classification. The detail statistics with category numbers and data splits of the three datasets are summarized in Table 1, and the example images are shown in Figure 2.

2) FEATURES

Since discriminative features are quite important for image classification, we adopt a convolutional neural network feature extractor, VGG16 [41], which is pre-trained on the ImageNet dataset. First, we resize the images into identical sizes of 256 × 256 and extract the image features using the pre-trained VGG16 network. Second, we reserve the features of the last convolutional layer and simply flatten them. Finally, the features dimension of each image is $512 \times 8 \times 8 = 32768$.

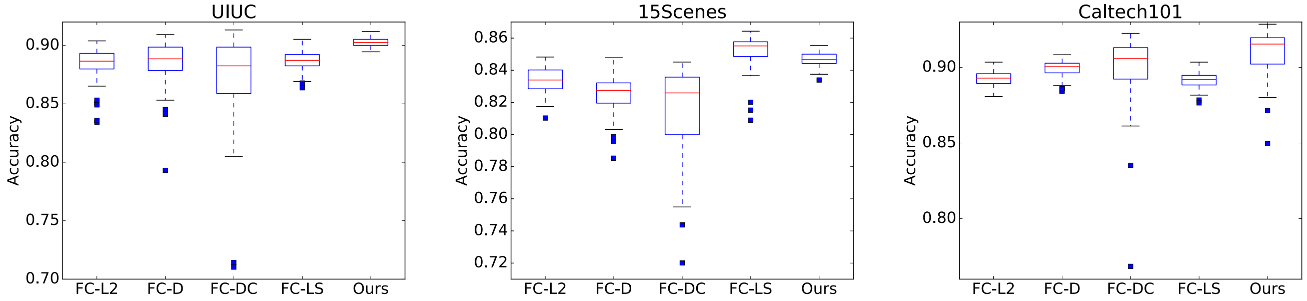

FIGURE 3. Comparison of accuracies obtained by the methods: fully connected network with L2 regularization (FC-L2), fully connected network with Dropout (FC-D), fully connected network with DropConnect (FC-DC), and fully connected network with large-margin Softmax loss (FC-L), and our loss (Ours) via boxplots on the UIUC, 15Scenes, and Caltech101 datasets. In each box plot, each method runs 60 rounds.

B. COMPARED METHODS AND THEIR IMPLEMENTATION

To evaluate the classification performance of the proposed loss on the aforementioned three datasets, we compare a fully connected network (FC) with the proposed loss, short for *Ours*, with the following four baseline methods:

1) FC-L2

fully connected network with L2 regularization. We use a fully connected network with two layers, where the activation functions of the first and second layers are rectified linear unit function (Relu) and Softmax, respectively.

2) FC-D

fully connected network with Dropout. We add a Dropout layer after the hidden layer of FC-L2; the probability that a neuron unit is dropped is 0.5. Except for this difference, the other settings of FC-Dropout are identical to FC-L2.

3) FC-DC

fully connected network with DropConnect. We add a layer after the hidden layer of FC; the probability that a neuron unit is dropped is 0.5, and the other settings of FC-L2 are kept unchanged.

4) FC-LS

fully connected network with large-margin Softmax loss. We use the large-margin Softmax loss [23] instead of the Softmax cross-entropy loss in FC-L2. Specifically, since the large-margin Softmax loss is very difficult to optimization, following [23], we replaced Relu with PReLU and initialize the network with Kaiming initialization [42].

C. IMPLEMENTATION DETAILS

To make it fair, we adopt the same settings for all the compared methods. Specifically, we use the same features extracted with the pre-trained VGG16 and use the minibatch stochastic gradient descent. The optimization algorithm is the RMSprop with the initial learning rate of 0.001, the coefficient of L2 norm is $5e-4$, the batch size is 32, and the number of epochs is 200. All compared methods are implemented by the Pytorch method.

TABLE 2. Comparison of the classification performance on the UIUC, 15Scenes, and Caltech101 datasets. The methods include the following: fully connected network with L2 regularized S-CE loss (FC-L2), fully connected network with Dropout (FC-D), fully connected network with DropConnect (FC-DC), and fully connected network with large-margin Softmax loss (FC-LS). Each method runs 60 rounds, and the mean values and standard deviations of the classification accuracies are reported.

Datasets	Methods	Mean	Std.
UIUC	FC-L2	0.8837	0.0151
	FC-D	0.8846	0.0189
	FC-DC	0.8723	0.0399
	FC-LS	0.8869	0.0096
	Ours	**0.9028**	**0.0040**
15Scenes	FC-L2	0.8331	0.0080
	FC-D	0.8254	0.0121
	FC-DC	0.8147	0.0278
	FC-LS	**0.8518**	0.0105
	Ours	0.8468	**0.0041**
Caltech101	FC-L2	0.8927	**0.0046**
	FC-D	0.8990	0.0056
	FC-DC	0.8995	0.0244
	FC-LS	0.8914	0.0053
	Ours	**0.9093**	0.0151

D. CLASSIFICATION ACCURACIES

We run FC-L2, FC-D, FC-DC, FC-LS and Ours on the UIUC, 15Scenes, and Caltech101 datasets for 60 rounds each. The mean values and standard deviations of the classification accuracies of 60 rounds are shown in Table 2. A larger mean and a smaller standard deviation indicate better performance, which are labeled in bold. The box plot of the classification accuracies of 60 rounds is shown in Figure 3.

Table 2 shows that on the three datasets, FC-L2 is easy to overfit and has quite unstable performance. FC-D performs slightly better than FC-L2 on the UIUC and Caltech101 datasets but has worse performance with FC-L2 on the 15Scenes dataset. FC-DC has competitive performance with FC-L2 on the Caltech101 dataset but performs worse than FC-L2 on the UIUC and 15Scenes datasets. FC-LS performs better than FC-L2 on the UIUC and 15Scenes datasets but slightly worse than FC-L2 on the Caltech101 dataset.

Our method achieves the best results on the UIUC and Caltech101 datasets and has a competitive performance with FC-LS, the best performance, on the 15Scenes dataset. Especially, our method has the smallest standard deviations of the

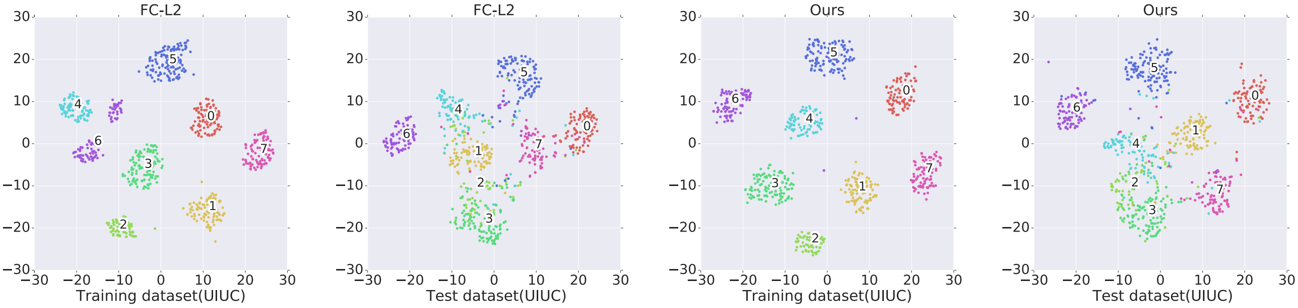

FIGURE 4. Feature visualization of FC-L2 and the proposed method on the UIUC dataset. The first two subfigures are the feature visualization of the Softmax loss on training data and test data, respectively, and the accuracy on the test data is 83.44%. The points with different colors denote the features from different classes. The last two subfigures are the feature visualization of our method on training data and test data, respectively, and the accuracy on the test data is 89.45%.

TABLE 3. ***p*-values of the compared methods, FC-L2, FC-Dropout (FC-D), FC-DropConnect (FC-DC), and FC-LSoftmax (FC-LS), and the proposed method by paired Student's t-test. Each method runs 60 rounds on each dataset. The scope of *p*-values is listed in the table (0.005 is the significance level).**

Datasets	FC-L2	FC-D	FC-DC	FC-LS
UIUC	<0.005	<0.005	<0.005	<0.005
15Scenes	<0.005	<0.005	<0.005	<0.005
Caltech101	<0.005	<0.005	<0.005	<0.005

classification accuracies on the UIUC and 15Scenes datasets, which means that our method is more stable than other methods.

In addition, Figure 3 shows that on the UIUC dataset, the box plot of our method is more compact than all the compared methods, and both the central mark and edges of the boxes are higher than all the other compared methods; especially, it has no bad-performing outlier. On the 15Scenes dataset, the box plot of our method is more compact than FC, FC-D and FC-DC, but it is close to FC-L. On the Caltech101 dataset, though several of the compared methods are more compact than our method, our method has higher central mark and edges of the boxes, which means that our method could have better performance on this dataset.

E. PAIRED STUDENT'S T-TEST

The experimental results in the previous sections show that the proposed method obtains better performance. To confirm that the improvement is not by chance, we perform a paired Student's t-test [43] for the proposed method and other compared methods, and the p-values are listed in Table 3. Following [44], the significance level is set as 0.005 in the paired Student's t-test. According to Table 3, all the p-values are much smaller than the significance level. Thus, the null hypothesis that the compared method has the identical mean value to the proposed method is always rejected on the UIUC, 15Scenes and Caltech101 datasets.

F. FEATURE VISUALIZATION

In previous experiments, we ran FC-L2 and our method 60 rounds each on the UIUC dataset. In this section, for the two methods, we select their lowest performance in 60 rounds, reduce the feature dimensions (the input of the Softmax layer) as 2 by T-SNE [44] and plot the reduced features of the training data and test data in Figure 4, where different colors represent different classes.

From Figure 4, we can observe that on the training dataset, the features of the fifth and seventh categories learned by our method are more separable than the ones learned by FC-L2. That is, the inter-class distance of the features learned by our method is larger than FC-L2, and the features of the sixth category learned by our method is more compact than the one learned by FC-L2. This means that the intra-class distance of the features learned by our method is smaller than FC-L2, which shows that in our method has better generalization performance on the test dataset and indicates our method learns a large decision margin for classification.

G. EFFECT OF VARYING THE ACTIVATION FUNCTION AND INITIALIZATION

To show that the proposed loss can be optimized easily, we evaluate the effect on network parameters of different activation functions and initialization methods on FC with the proposed loss (our method) and FC-LS. In particular, we select two activation functions, ReLU and PReLU, and two initialization methods, Uniform initialization and Kaiming initialization. Under four combinations of the activation functions and initialization methods, our method and FC-LS runs on the UIUC, 15Scenes, and Caltech101 datasets for 60 rounds each, and the corresponding mean values are reported in Table 4.

According to Table 4, we can observe the following: first, when the activation function ReLU is selected, FC-LS cannot be optimized regardless of which initialization method is selected. However, our method always has good performance regardless of which the activation function is selected. Second, the network parameter initialization method has slight effect on performance for both our method and FC-LS. Third, our method has a good and stable performance. In summary, the proposed loss is a more general loss function that does not have more requirements on the activation function and initialization method of network parameters.

TABLE 4. Comparison of the classification accuracies obtained by different methods on the UIUC, 15Scenes, and Caltech101 datasets when changing the activation function and network initialization. Each setting for FC with the proposed loss (our method) and FC-LS runs 60 rounds; the mean values are listed in the table cells.

Datasets	Methods	Uniform		Kaiming	
		ReLU	PReLU	ReLU	PReLU
UIUC	FC-LS	N/A	0.8871	N/A	0.8869
	Ours	0.9028	0.9040	0.9024	0.9038
15Scenes	FC-LS	N/A	0.8518	N/A	0.8518
	Ours	0.8468	0.8475	0.8355	0.8485
Caltech101	FC-LS	N/A	0.8931	N/A	0.8914
	Ours	0.9093	0.9108	0.9061	0.9059

H. DISCUSSION

The experimental results on the UIUC, 15Scenes, and Caltech101 datasets show that among the compared methods, both Dropout and DropConnect did not show an obvious advantage compared with a network without Dropout or DropConnect. The reason is that these two methods train many subnetworks of the original network randomly in the training phase and in the test phase, no neuron or no connection is dropped. Due to large randomness, the accuracy of the subnetworks and the ambiguity among the subnetworks cannot be ensured, so that they cannot work well on these small-sample datasets.

FC-LSoftmax performs better than Dropout and DropConnect. The method introduces a large margin idea into the S-CE loss function so as to learn more discriminative features and obtain better generalization ability. The important issue of the method is that it does not easily converge and has requirements on activation function and initialization method of network parameters. In addition, we find that FC-LSoftmax cannot perform well on image data including more objects, such as the UIUC and Caltech101 datasets. In contrast, our method converges easily and can work well for different initialization methods of network parameters and activation functions.

Our method shows higher accuracy, better stability, and easy optimization, which are mainly attributed to the introduction of a regularized term that facilitates a large decision margin between classes for network or model. A regularized term is quadratic, which does not add difficulty to parameter optimization. In contrast, since the regularized term limits parameter space, the optimization speed is increased. Furthermore, our method obtains the best performance on the UIUC and Caltech101 datasets and has a competitive performance with FC-LS on the 15Scenes dataset. The experimental results on these three datasets suggest that our method obtains good performance regardless of the image data include more objects or few objects.

IV. CONCLUSION

In the paper, we proposed a new large-margin regularization method with easy optimization for Softmax cross-entropy loss of neural networks. The experimental results on three small-sample datasets confirmed that our ensemble method (*i*) can obtain good generalization performance and outperforms the existing popular regularization methods of neural networks, and (*ii*) is adapted to optimize the neural network compared with some newly proposed loss functions. Future work includes increasing the number of networks and experimenting on different types of networks as well as different kinds of data, such as speech and text, to evaluate the effectiveness of the proposed loss.

REFERENCES

[1] B. Li, W. Xiong, W. Hu, and B. Funt, "Evaluating combinational illumination estimation methods on real-world images," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1194–1209, Mar. 2014.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[3] X. Ding, B. Li, W. Xiong, W. Guo, W. Hu, and B. Wang, "Multi-instance multi-label learning combining hierarchical context and its application to image annotation," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1616–1627, Aug. 2016.

[4] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[5] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, Mar. 2017.

[6] Y. Zhou, Z. M. Fadlullah, B. Mao, and N. Kato, "A deep-learning-based radio resource assignment technique for 5G ultra dense networks," *IEEE Netw.*, vol. 32, no. 6, pp. 28–34, Nov./Dec. 2018.

[7] T. Zhang *et al.*, "Predicting functional cortical ROIs via DTI-derived fiber shape models," *Cerebral Cortex*, vol. 22, no. 4, pp. 854–864, 2011.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[9] J. Han *et al.*, "Representing and retrieving video shots in human-centric brain imaging space," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2723–2736, Jul. 2013.

[10] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019.

[11] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[12] K. Zhang, M. Sun, T. X. Han, X. Yuan, L. Guo, and T. Liu, "Residual networks of residual networks: Multilevel residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1303–1314, Jun. 2018.

[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[14] Z. Ma, J.-H. Xue, A. Leijon, Z.-H. Tan, Z. Yang, and J. Guo, "Decorrelation of neutral vector variables: Theory and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 129–143, Jan. 2018.

[15] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[16] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognit.*, vol. 47, no. 9, pp. 3143–3157, 2014.

[17] K. Zhang, L. Guo, C. Gao, and Z. Zhao, "Pyramidal RoR for image classification," *Cluster Comput.*, to be published.

[18] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.

[19] Q. Miao, Y. Cao, G. Xia, M. Gong, J. Liu, and J. Song, "RBoost: Label noise-robust boosting algorithm based on a nonconvex loss function and the numerically stable base learners," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2216–2228, Nov. 2016.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 499–515.

[22] Y. Liu, H. Li, and X. Wang. (2017). "Learning deep features via congenerous cosine loss for person recognition." [Online]. Available: https://arxiv.org/abs/1702.06890

[23] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 507–516.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[25] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jul. 2017, pp. 6738–6746.

[26] W. Wan, Y. Zhong, T. Li, and J. Chen, "Rethinking feature distribution for loss functions in image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9117–9126.

[27] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, Apr. 2015.

[28] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019.

[29] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, 1998.

[30] A. Y. Ng, "Feature selection, *L*1 vs. *L*2 regularization, and rotational invariance," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 78.

[31] S. Wager, S. Wang, and P. S. Liang, "Dropout training as adaptive regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 351–359.

[32] S. Wang and C. Manning, "Fast dropout training," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 118–126.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1058–1066.

[35] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[36] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.

[37] J. Zhang, R. Jin, Y. Yang, and A. G. Hauptmann, "Modified logistic regression: An approximation to SVM and its applications in large-scale text categorization," in *Proc. Int. Mach. Learn. Conf.*, vol. 3, 2003, pp. 888–895.

[38] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[39] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 2169–2178.

[40] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," *Comput. Vis. Image Understand.*, vol. 106, no. 1, pp. 59–70, Jan. 2007.

[41] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[43] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. New York, NY, USA: Springer, 1998, pp. 199–213.

[44] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

XIAOXU LI received the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2012. She is currently an Associate Professor with the School of Computer and Communication, Lanzhou University of Technology. Her research interest includes machine learning fundamentals with a focus on applications in image and video understanding. She is a member of the China Computer Federation.

DONGLIANG CHANG received the B.E. degree in network engineering from Zhoukou Normal University, China, in 2016. He is currently pursuing the degree with the Lanzhou University of Technology. His research interests include machine learning and computer vision.

TAO TIAN graduated from the Lanzhou Polytechnical College, in 2010. He is currently pursuing the master's degree with the Lanzhou University of Technology. His research interests include small-sample learning and image understanding.

JIE CAO received the M.E. degree from Xi'an Jiaotong University, China, in 1994. She is currently a Professor and a Vice President of the Lanzhou University of Technology. Her research interests include machine learning, pattern recognition, speech and speaker recognition, information fusion, and computer vision.