CONTRASTIVE-CENTER LOSS FOR DEEP NEURAL NETWORKS

Ce Qi[1], *Fei Su*[1,2]

[1]School of Information and Communication Engineering
[2]Beijing Key Laboratory of Network System and Network Culture
Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

The deep convolutional neural network(CNN) has significantly raised the performance of image classification and face recognition. Softmax is usually used as supervision, but it only penalizes the classification loss. In this paper, we propose a novel auxiliary supervision signal called contrastive-center loss, which can further enhance the discriminative power of the features, for it learns a class center for each class. The proposed contrastive-center loss simultaneously considers intra-class compactness and inter-class separability, by penalizing the contrastive values between: (1)the distances of training samples to their corresponding class centers, and (2)the sum of the distances of training samples to their non-corresponding class centers. Experiments on different datasets demonstrate the effectiveness of contrastive-center loss.

Index Terms— Class center, Auxiliary loss, Deep convolutional neural networks, Image classification and face recognition

1. INTRODUCTION

Recently, deep neural networks have achieved state of the art performance on different tasks such as visual object classification [1–5] and recognition [6–13], showing the power of the discriminative features.

In general visual classification and recognition task, deep convolutional neural networks(CNN) [1–3,5] are usually chosen. For the discriminative features extracted from CNN, the performance is usually much higher than other traditional machine learning algorithms. Usually, the CNN maps images to high dimension space to let the softmax or SVM easy to classify the images to a certain class. The softmax loss only penalizes the classification loss, and does not consider the intra-class compactness and inter-class separability explicitly.

Recently, there are some works learning with even more discriminative features to further improve the performance of CNN. Some researchers use deeper, wider and more complex network structures to obtain better features, such as [5], in which the authors train a very deep neural network with some training tricks to make the network converge to get more discriminative features, but the training is relatively harder and not that effective. There are also some other efforts on new non-linear activations [4, 14], dropout [1] and batch normalization [15] to make the network perform better.

Another kind of strategy to obtain more discriminative features is to use auxiliary loss to train the neural network, such as contrastive loss [8], triplet loss [10] and center loss [13]. The three new losses are proposed for the purpose of enforcing better intra-class compactness and inter-class separability. The contrastive loss and triplet loss do really improve the quality of features extracted from the network. The triplet needs carefully pre-selected triple samples consisted of two same people's face images and one different person's face image. And the selection of triple samples is significant for it will influence the result of training. The contrastive loss chooses couple sample pairs to get the loss, so contrastive loss needs careful pre-selection, too. What's more, if all possible training samples combinations are chosen, the number of training pairs and triplets would theoretically go up to O(N^2), where N is the total number of training samples. The center loss [13], which learns a center for each class and penalizes the distances between the deep features and their corresponding class centers, is a new novel loss to enforce extra intra-class compactness. However, the center loss does not consider the inter-class separability.

In this paper, we propose the contrastive-center loss, which learns a center for each class. This new loss will simultaneously consider intra-class compactness and inter-class separability by penalizing the contrastive values between: (1)the distances of training samples to their corresponding class centers, and (2)the sum of the distances of training samples to their non-corresponding class centers. The training process is simple because the contrastive-center loss does not need pre-selected sample pairs or triples.

Experiments and visualizations show the effectiveness of our proposed contrastive-center loss. The experiments on MNIST [16] and CIFAR10 [17] demonstrate the effectiveness of contrastive-center loss on classification task. And the experiments of face recognition on LFW [18] demonstrate the

This work is supported by Chinese National Natural Science Foundation(61372169, 61532018) and Special Funds of Beijing Municipal Co-construction Project.

effectiveness of contrastive-center loss on recognition task.

2. PROPOSED METHOD

In this section, we introduce center loss and indicate its weakness of only considering intra-class compactness. Then the proposed contrastive-center loss is described, which simultaneously considers the intra-class compactness and inter-class separability. Using softmax loss assisted with our contrastive-center loss to train a deep neural network will do really boost the performance of the network.

2.1. Center loss

The features extracted from the deep neural network trained under the supervision of softmax loss are separable but not that discriminative enough, since they show significant intra-class variations, as shown in Fig. 1(a). Based on the phenomenon, authors in [13] develop an effective loss function to improve the power of the deep features extracted from deep neural networks. Center loss minimizes the intra-class distances while keeping the features can be classified into right classes by softmax. Eq. (1) gives the center loss function.

$$L_c = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{1}$$

Where L_c denotes the center loss. m denotes the number of training samples in a min-batch. $x_i \in R_d$ denotes the ith training sample. y_i denotes the label of x_i. $c_{y_i} \in R_d$ denotes the y_ith class center of deep features. d is the feature dimension.

When training the deep neural networks, authors in [13] adopt the joint supervision of softmax loss and center loss to train the networks, as formulated in Eq. (2).

$$L = L_s + \lambda L_c \tag{2}$$

Where L denotes the total loss of deep neural network. L_s denotes the softmax loss. L_c denotes the center loss. λ denotes the scalar used for balancing the two loss functions.

The weakness of center loss: Discriminative features should have better intra-class compactness and inter-class separability. The center loss uses loss function Equation 1 to penalize big intra-class distances. However, center loss does not consider the inter-class separability. It will make the distances of different classes not that far, as show in Fig. 1(b). As we know, if the distances of different classes is far enough, the features will more discriminative for the better inter-class separability. In addition, for the center loss just penalizes big intra-class distances, does not consider inter-class distances, the changing of inter-class is small, meaning the positions of class centers will be slightly changed through all the training process. As a result, if the network initializes the class centers using a relatively smaller variance, it will result in the smaller distances between class centers after training because the center loss function only penalize the big intra-class distances without considering the inter-class distances.

2.2. Contrastive-center loss

As mentioned in section 2.1, the weakness of center loss is that it does not consider the inter-class separability. So, we propose a new loss function to consider the intra-class compactness and inter-class separability simultaneously by penalizing the contrastive values between: (1)the distances of training samples to their corresponding class centers, and (2)the sum of the distances of training samples to their non-corresponding class centers. Formally, it is defined as illustrated in Eq.3.

$$L_{ct-c} = \frac{1}{2}\sum_{i=1}^{m} \frac{\|x_i - c_{y_i}\|_2^2}{(\sum_{j=1, j\neq y_i}^{k} \|x_i - c_j\|_2^2) + \delta} \tag{3}$$

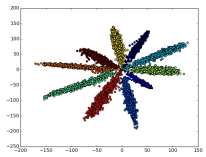
Where L_{ct-c} denotes the contrastive-center loss. m denotes the number of training samples in a min-batch. $x_i \in R_d$ denotes the ith training sample with dimension d. d is the feature dimension. y_i denotes the label of x_i. $c_{y_i} \in R_d$ denotes the y_ith class center of deep features with dimension d. k denotes the number of class. δ is a constant used for preventing the denominator equal to 0. In our experiments, we set $\delta = 1$ by default.

Obviously, the contrastive-center loss can be used in deep neural network directly and the network will be trained as general deep neural network. The class centers c_{y_i} will be updated through the training process. Comparing with those in the center loss, the class centers of our proposed contrastive-center loss will be updated to a more discrete distribution for the existence of penalization for too small distances between different class centers.

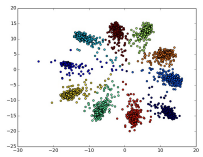
In this method, we update the class centers based on mini-batch, for it is not possible to update the centers based on the entire training set. And to make the training process is more stable, we use a scalar α to control the learning rate of class centers.

In each iteration, the deep neural network updates the class centers and network parameters simultaneously. The derivative of L_{ct-c} with respect to x_i and derivative of L_{ct-c} with respect to c_n is illustrated in Eq. (4) and Eq. (5) respectively. The two derivatives are used to update the parameters of deep neural networks and class centers respectively.

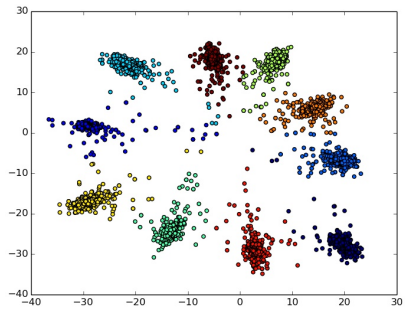
$$\begin{aligned}\frac{\partial L_{ct-c}}{\partial x_i} =& \frac{x_i - c_{y_i}}{(\sum_{j=1, j\neq y_i}^{k} \|x_i - c_{y_j}\|_2^2) + \delta} \\ &- \frac{\|x_i - c_{y_i}\|_2^2 \sum_{j=1, j\neq y_i}^{k} (x_i - c_{y_j})}{[(\sum_{j=1, j\neq y_i}^{k} \|x_i - c_{y_j}\|_2^2) + \delta]^2}\end{aligned} \tag{4}$$



(a) softmax loss



(b) center loss



(c) our contrastive-center loss

Fig. 1. Visualization of MNIST. Note: The domain of coordinates axis of the visualization of contrastive-center loss is bigger than center loss's. The average L2 distance of class centers to the center of class centers is about 10 to 15. The average L2 distance of class centers to the center of class centers is about 50. The average L2 distance of class centers to the center of class centers is obvious, which indicates the contrastive-center loss's is about 3.3 to 5 times of the center loss's.

$$\frac{\partial L_{ct-c}}{\partial c_n} = \sum_{i=1}^{m} \begin{cases} \frac{c_{y_i} - x_i}{(\sum_{j=1, j\neq y_i}^{k} \|x_i - c_{y_j}\|_2^2) + \delta} & \text{if } y_i = n \\ \frac{(x_i - c_n)\|x_i - c_{y_i}\|_2^2}{[(\sum_{j=1, j\neq y_i}^{k} \|x_i - c_{y_j}\|_2^2) + \delta]^2} & \text{if } y_i \neq n \end{cases} \quad (5)$$

In Eq. (4) and Eq. (5), the meaning of symbols are the same as those in Eq. (3), except that $n = 1, ..., m$ denotes the current class center's serial number.

3. EXPERIMENTS

To verify the effectiveness of the contrastive-center loss, we evaluate the experiments on two typical visual tasks: visual classification and face recognition. The experiment results demonstrate our contrastive-center loss can not only improve the accuracy on classification, but also boost the performance on visual recognition. In visual classification, we use two wildly used dataset(MNIST [16] and CIFAR10 [17]). In face recognition, the LFW [18] dataset is used. We implement the contractive-center loss and do the experiments using the Caffe library [19].

Table 1. The CNNs architecture we use for MNIST and visualization is same as [13], called LeNets++. $(5, 32)_{/1,2} \times 2$ denotes 2 cascaded convolution layers with 32 filters of size 5×5, where the stride and padding are 1 and 2 respectively. $2_{/2,0}$ denotes the max-pooling layers with grid of 2×2, where the stride and padding are 2 and 0 respectively. In LeNets++, Parametric Rectified Linear Unit (PReLU) [4] is used as the nonlinear unit.

	stage 1		stage 2		stage 3		stage 4
Layer	conv	pool	conv	pool	conv	pool	FC
LeNets	$(5, 20)_{/1,0}$	$2_{/2,0}$	$(5, 50)_{/1,0}$	$2_{/2,0}$			500
LeNets++	$(5, 32)_{/1,2} \times 2$	$2_{/2,0}$	$(5, 64)_{/1,2} \times 2$	$2_{/2,0}$	$(5, 128)_{/1,2} \times 2$	$2_{/2,0}$	2

Table 2. Classification accuracy (%) on MNIST dataset.

Method	Accuracy(%)
Softmax	98.8
Center loss	98.94
Our contrastive-center loss	**99.17**

3.1. Experiments on MNIST and visualiztion

The MNIST [16] are consisted of $60,000$ training images and $10,000$ testing images in total. The images are all hand written digits $0 - 9$ in 10 classes which are 28×28 in size.

The network used in this experiments are the same as the network used for MNIST in [13]. The network are modified from LeNets [20] to a deeper and wider network, but reduce the output number of the last hidden layer to 2, meaning the dimension of the deep features is 2, which is easy to be plot in 2-D surface for visualization. The details of the network architecture are given in Table 1. Note that we set loss weight $\lambda = 0.1$ for L_{ct-c} there.

When training and testing LeNet++, we only use original training images and original testing images without any data augmentation. The result is shown in Table 2. Contrastive-center loss boosts accuracy of 0.37% compared to softmax loss and 0.23% compared to center loss respectively.

We then visualize the deep features of the last hidden layer(the output number is 2) of LeNet++. All the features are extracted using the $10,000$ testing images as input. The visualization is shown in Fig. 1. We can observe that:

(1) Under the single supervision signal of softmax loss, the features are separable, but with significant intra-class variations.

(2) The center loss makes the deep features have better intra-class compactness. But the inter-class separability is not good enough. The average L2 distance of class centers to the center of class centers is about 10 to 15.

(3) The contrastive-center loss simultaneously achieves good intra-class compactness and inter-class separability. The average L2 distance of class centers to the center of class centers is about 50.

(4) The contrastive-center loss's average L2 distance of class centers to the center of class centers is about 3.3 to 5

times of the center loss's, showing that the contrastive-center loss gets better inter-class separability than the center loss.

3.2. Experiments on CIFAR10

The CIFAR10 dataset [17] is consisted of 10 classes of natural images with $50,000$ training images and $10,000$ testing images. Each image is RGB image of size 32×32.

We use 20-layer ResNet [13] in the experiments. Following the commonly used strategy, we do data augmentation in training, and in testing, there is no data augmentation. We follow the standard data augmentation in [13] for training: 4 pixels are padded on each side, and a 32×32 crop is randomly sampled from the padded images or its horizontal flip. In testing, we only evaluate the single view of the original 32×32 testing images. Note that we set loss weight $\lambda = 0.1$ for L_{ct-c} there.

The result is shown in Table 3. We can observe that:

(1) The center loss makes the net's accuracy increased by 0.4% compared with the net's only supervised under softmax loss.

(2) Our contrastive-center loss makes the net's accuracy increased by 1.2% compared with the net only supervised under softmax loss.

(3) Our contrastive-center loss gets better result than the center loss with accuracy gain of 0.35% on CIFAR10.

Table 3. Classification accuracy (%) on CIFAR10 dataset.

Method	Accuracy(%)
20-layer ResNet [5]	91.25
20-layer ResNet(our implementation based on center loss [13])	92.1
20-layer ResNet(our contrastive-center loss)	**92.45**

3.3. Experiments on LFW

To further demonstrate the effectiveness of our contrastive-center loss, we conduct the experiments on LFW dataset [18]. The dataset collects $13,233$ face images from 5749 persons from uncontrolled conditions. Following the unrestricted with labeled outside data protocol [18], we train on the publicly available CASIA-WebFace [21] outside dataset (490k labeled face images belonging to over $10,000$ individuals) and test on the $6,000$ face pairs on LFW. The training data is cleaned for wrong collected images. People overlapping between the outside training data and the LFW testing data are excluded. As preprocessing, we use MTCNN [22] to detect the faces and align them based on 5 points.

Then we train a single network for feature extraction. For good comparison, we use the network released by center loss [13], which is called FRN(or FaceResNet) in later. Note that the network released by center loss is not the network they use in the paper [13]. Based on the network publicly available, we re-implement the training process of center loss and get better result than the models released by the author. We also train the networks under the supervision of softmax loss and our contrastive-center loss jointly. In feature extraction, like in [13], the original image and its flip one are used to get two feature vectors and concatenate them as the final feature. Note that we set loss weight $\lambda = 1$ for L_{ct-c} there.

Table 4. Verification accuracy (%) on LFW dataset. $*$ denotes the outside data is private (not publicly available).

Method	Images	Networks	Accuracy(%)
DeepFace [6]	$4M$	3	97.35
Fusion [11]	$10M$	5	98.37
SeetaFace [23]	$0.5M$	1	98.60
SeetaFace(Full) [23]	$0.5M$	1	98.62
DeepID-2+ [9]	–	1	98.70
DeepFR [24]	$2.6M$	1	98.95
Yi et al., 2014 [21]	$0.494, 414M$	1	97.73
Ding & Tao, 2015 [25]	$0.494, 414M$	1	98.43
FRN(trian with softmax loss only)	$0.455, 594M$	1	97.47
FRN(model released by center loss [13])	$0.494, 414M$	1	98.43
FRN(retrain with center loss [13])	$0.455, 594M$	1	98.55
FRN(our contrastive-center loss)	$\mathbf{0.455, 594M}$	**1**	**98.68**

The result is shown in Table 4. We can observe that:

(1) We train the FRN(FaceResNet) only with small data(CASIA webface cleaned, $0.455, 594M$). And the accuracy is comparable to the current state-of-art CNNs.

(2) FRN trained with our contrastive-center loss boosts the accuracy on LFW of 1.21%, 0.25% and 0.13% compared respectively with FRN trained with softmax loss only, model released by center loss [13] and our re-implementation of center loss.

4. CONCLUSION

We proposed a contrastive-center loss for deep neural networks. The contrastive-center loss simultaneously considers intra-class compactness and inter-class separability, by penalizing the contrastive values between (1)the distances of training samples to their corresponding class centers, and (2)the sum of the distances of training samples to their non-corresponding class centers. More appealingly, the contrastive-center loss has very clear intuition and geometric interpretation. The experimental results on several benchmark datasets prove the effectiveness of the proposed contrastive-center loss.

5. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan,

Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[6] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[7] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[8] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[9] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deeply learned face representations are sparse, selective, and robust," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2892–2900.

[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[11] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Web-scale training for face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2746–2754.

[12] Yandong Wen, Zhifeng Li, and Yu Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4893–4901.

[13] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[14] Ian J Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C Courville, and Yoshua Bengio, "Maxout networks.," *ICML (3)*, vol. 28, pp. 1319–1327, 2013.

[15] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[16] Yann LeCun, Corinna Cortes, and Christopher JC Burges, "The mnist database of handwritten digits," 1998.

[17] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.

[18] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007.

[19] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[22] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[23] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen, "Viplfacenet: An open source deep face recognition sdk," *arXiv preprint arXiv:1609.03892*, 2016.

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015, vol. 1, p. 6.

[25] Changxing Ding and Dacheng Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.