CONTRASTIVE-CENTER LOSS FOR DEEP NEURAL NETWORKS

Ce Qi, Fei Su

School of Information and Communication Engineering

Beijing University of Posts and Telecommunications, Beijing

1. Introduction

- The key of generic object, scene or action recognition is the separable of features. Because the classes are constant.
- For face recognition task, the deeply learned features need to be not only separable but also discriminative. Because the classes of faces are inconstant.
- The deeply learned features are required to be generalized enough for identifying new unseen classes without label prediction.

2. Discriminative Feature Learning

- SOFTMAX LOSS: encourages the separability of features.
- CENTER LOSS: learns a center for deep features of each class, but only penalizes the distances between the deep features and their corresponding class centers.
- 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.

3. Formula of contrastive-center loss

- Formula of center loss:
- $$L_c = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{y_i}\right\|^2$$
- Formula of contrastive-center loss:
- $$L_{ct-c} = \frac{1}{2}\sum_{i=1}^{m}\frac{\left\|x_i - c_{y_i}\right\|^2}{\left(\sum_{j=1,j\neq y_i}^{k}\left\|x_i - c_j\right\|^2\right)+\delta}$$
- Final formula(joint supervision):
- $$L = L_s + \lambda L_{ct-c}$$

$\boldsymbol{L}$, $\boldsymbol{L_s}$, $\boldsymbol{L_c}$, $\boldsymbol{L_{ct-c}}$ denote the total loss, softmax loss, center loss and contrastive-center loss respectively.
$\boldsymbol{\lambda}$ denotes the scalar used for balancing the two loss functions.
$\boldsymbol{m}$ denotes the number of training samples in a mini-batch.
$\boldsymbol{x_i} \in \boldsymbol{R^d}$ denotes the ith training sample with dimension d.
$\boldsymbol{d}$ is the feature dimension.
$\boldsymbol{y_i}$ denotes the label of x_i.
$\boldsymbol{c_{y_i}} \in \boldsymbol{R^d}$ denotes the y_ith class center of deep features with dimension d.
$\boldsymbol{k}$ denotes the number of class.
$\boldsymbol{\delta}$ is a constant used for preventing the denominator equal to 0.
In our experiments, we set $\delta = 1$ by default.

4. Details of contrastive-center loss

- Derivative:
- $$\frac{\partial L_{ct-c}}{\partial x_i} = \frac{x_i - c_{y_i}}{\left(\sum_{j=1,j\neq y_i}^{k}\left\|x_i - c_j\right\|^2\right)+\delta} - \frac{\left\|x_i - c_{y_i}\right\|^2 \sum_{j=1,j\neq y_i}^{k}(x_i - c_j)}{\left[\left(\sum_{j=1,j\neq y_i}^{k}\left\|x_i - c_j\right\|^2\right)+\delta\right]^2}$$
- $$\frac{\partial L_{ct-c}}{\partial c_n} = \sum_{i=1}^{m}\begin{cases}\dfrac{c_{y_i} - x_i}{\left(\sum_{j=1,j\neq y_i}^{k}\left\|x_i - c_j\right\|^2\right)+\delta} & if\ y_i = n \\ \dfrac{(x_i - c_n)\left\|x_i - c_{y_i}\right\|^2}{\left[\left(\sum_{j=1,j\neq y_i}^{k}\left\|x_i - c_j\right\|^2\right)+\delta\right]^2} & if\ y_i \neq n\end{cases}$$

About code optimazation on GPU:

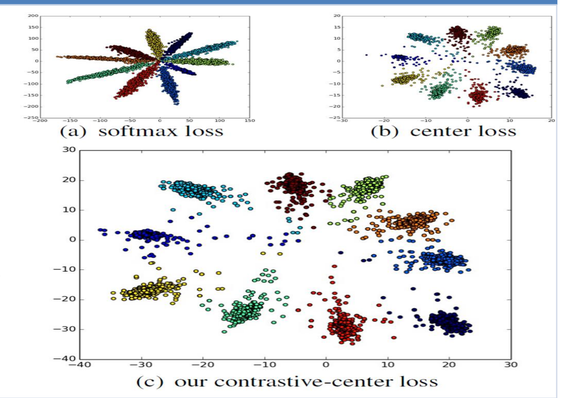
- For 128,512,10 vs 128,512,100000 (batch size, feature dimension and number of class), the former uses 128*512 kernels , while the latter uses 512*100000 kernels.
- Keep the variables with high time complexity used in forward and backward in memory to save computing time

5. Details of contrastive-center loss

- Centers are updated based on mini-batch with an adjustable learning rate.
- Contrastive-center loss enjoys the same requirement as the softmax loss and needs no complex sample mining and recombination, which is inevitable in contrastive loss and triple loss.
- Contrastive-center loss does help on tasks of not only face recognition but also image classification.
- In general, for tasks using class labels, contrastive-center can do help.

6. Visualization on MNIST

- Softmax loss: features are separable, but not discriminative.
- Softmax loss+center loss: features are separable and have better intra-class compactness.
- Softmax loss + contrastive-center loss: features are separable and have better intra-class compactness and inter-class separability (30 vs 10~15 , 2~3 times than center loss).



7. Experiments on MNIST and Cifar-10

- MNIST:
(use a small network for visualization and accuracy comparison)

	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

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

- Cifar-10:
(use 20-layer ResNet)

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**

8. Experiments on LFW

- LFW:
(use a **small** network for accuracy comparison)

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**