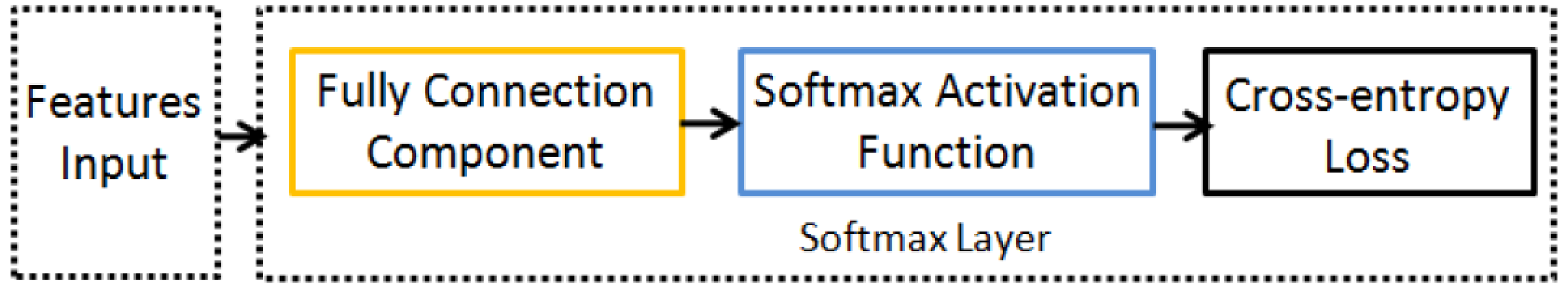
Noisy Softmax: Improving the Generalization Ability of DCNN via Postponing the Early Softmax Saturation

Binghui Chen, Weihong Deng, Junping Du

Beijing University of Posts and Telecommunications

IEEE 2017 Conference on Computer Vision and Pattern Recognition

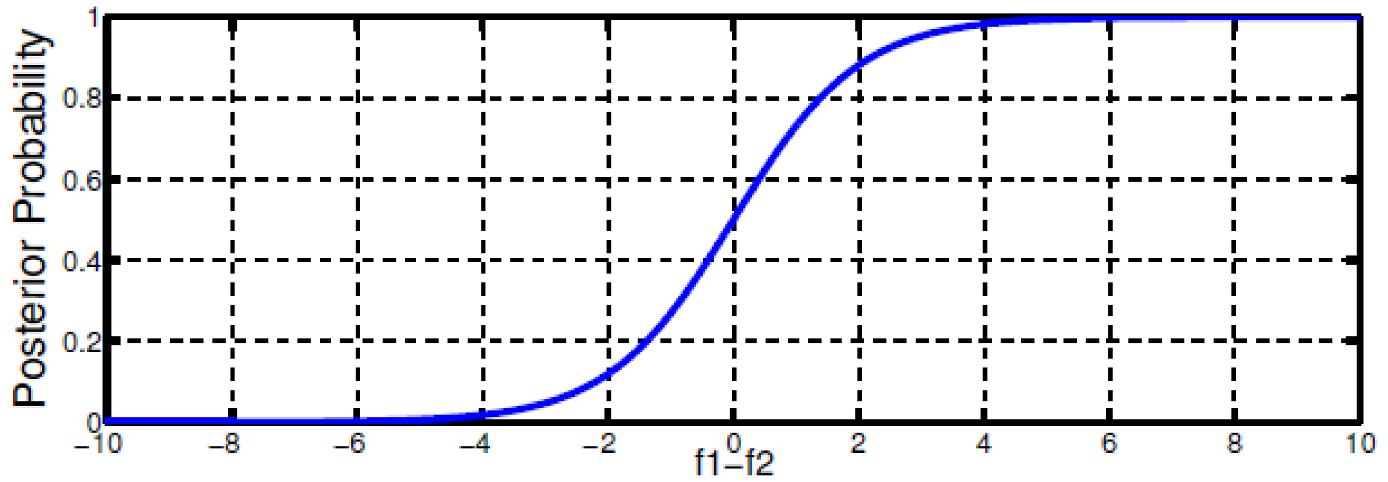


Introduction:

The commonly used Softmax layer is composed of the fully connected layer, Softmax activation function and the cross-entropy loss. While in standard Softmax activation, the saturation behavior like in sigmoid is always omitted. Inspired by [1], then we propose a desaturation strategy of injecting annealed noise to address this problem. Our Noisy Softmax improves the generalization ability of DCNN by giving the SGD solver more chances to explore a better solution.

Saturation behavior:

Softmax function: $\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}$



The gradients of standard Softmax:

$$\frac{\partial L}{\partial f_j} = P(y_i = j|x_i) - 1\{y_i = j\} = \frac{e^{f_j}}{\sum_k e^{f_k}} - 1\{y_i = j\}$$

http://bhchen.cn

Noisy Softmax:

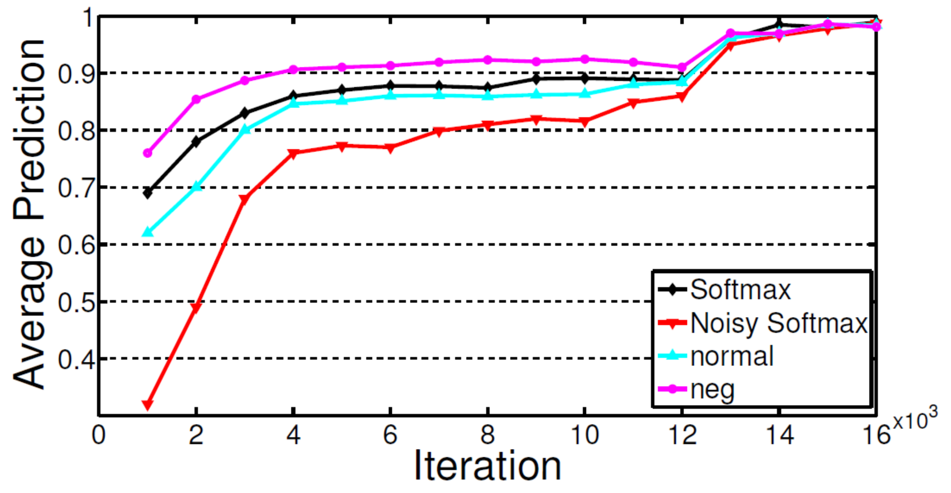
To postpone the early softmax saturation, we modify the input of the original Softmax. Considering the fact that the inner product of two vectors can be rewritten into the dot product of amplitude and angular, we construct our noisy input as follows:

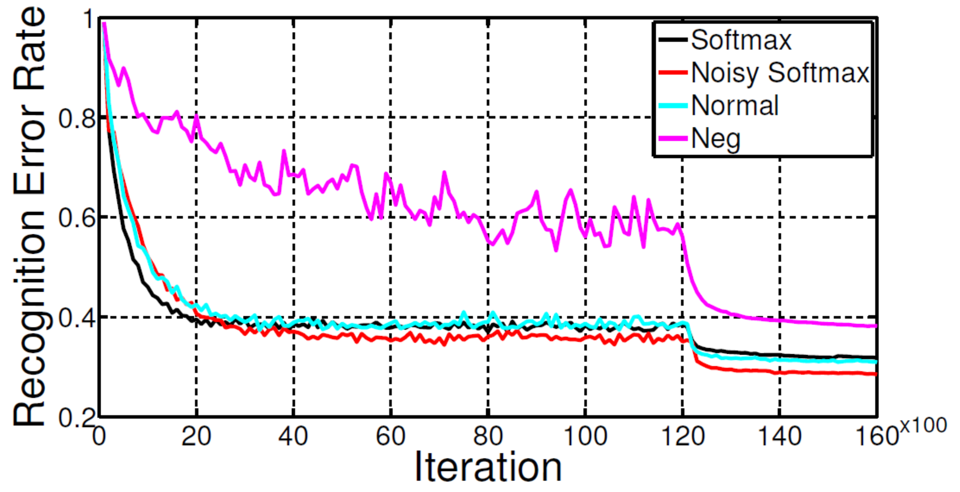
$$f_{y_i}^{noise} = f_{y_i} - \underbrace{\alpha \|W_{y_i}\| \|X_i\| (1 - \cos\theta_{y_i}) |\xi|}_{\text{noise term}}$$

where f_{y_i} is the input to original Softmax, we leverage $\|W_{y_i}\|\|X_i\|$ to make the magnitude of the noise and that of the original softmax input to be comparable, and use $(1 - \cos\theta_{y_i})$ to adaptively anneal the noise.

Discussion:

The saturation behavior comparison.





Training and testing procedure.

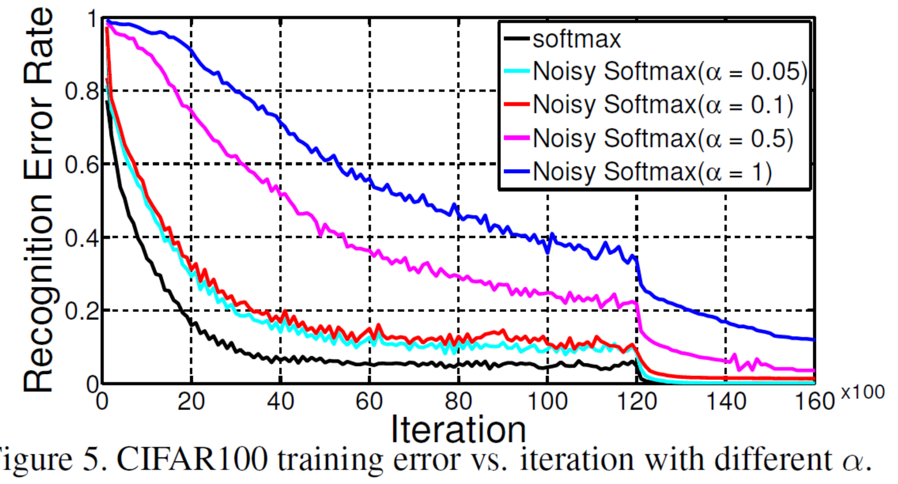


Figure 5. CIFAR100 training error vs. iteration with different α.

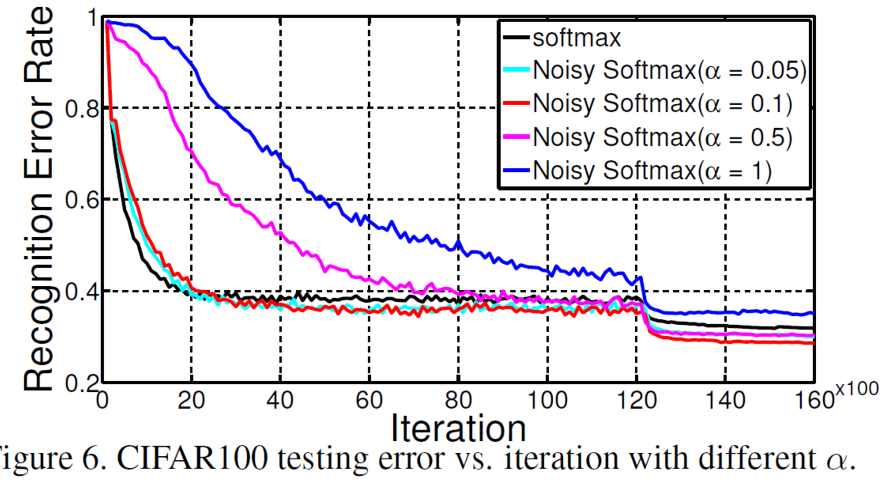


Figure 6. CIFAR100 testing error vs. iteration with different α.

Experimental Results:

Method	MNIST
CNN [19]	0.53
NiN [29]	0.47
Maxout [7]	0.45
DSN [27]	0.39
R-CNN [28]	0.31
GenPool [26]	0.31
DisturbLabel [50]	0.33
Softmax	0.43
Noisy Softmax ($\alpha^2 = 1$)	0.42
Noisy Softmax ($\alpha^2 = 0.5$)	**0.33**
Noisy Softmax ($\alpha^2 = 0.1$)	**0.33**
Noisy Softmax ($\alpha^2 = 0.05$)	0.37

Table 3. Recognition error rates (%) on MNIST.

Method	CIFAR10	CIFAR10+	CIFAR100
NiN [29]	10.47	8.81	35.68
Maxout [7]	11.68	9.38	38.57
DSN [27]	9.69	7.97	34.57
All-CNN [39]	9.08	7.25	33.71
R-CNN [28]	8.69	7.09	31.75
ResNet [13]	N/A	6.43	N/A
DisturbLabel [50]	9.45	6.98	32.99
Softmax	8.11	6.98	31.77
Noisy Softmax ($\alpha^2 = 1$)	9.09	8.77	35.23
Noisy Softmax ($\alpha^2 = 0.5$)	7.84	7.13	30.22
Noisy Softmax ($\alpha^2 = 0.1$)	**7.39**	**6.36**	**28.48**
Noisy Softmax ($\alpha^2 = 0.05$)	7.58	6.61	29.99

Table 4. Recognition error rates (%) on CIFAR datasets. + denotes data augmentation.

Method	Images	Models	LFW	Rank-1	DIR@FAR=1%	FGLFW	YTF
FaceNet [35]	200M*	1	99.65	-	-	-	95.18
DeepID2+ [43]	300k*	1	98.7	-	-	-	91.90
DeepID2+ [43]	300k*	25	99.47	95.00	80.70	-	93.20
Sparse [44]	300k*	1	99.30	-	-	-	92.70
VGG [32]	2.6M	1	97.27	74.10	52.01	88.13	92.80
WebFace [51]	WebFace	1	97.73	-	-	-	90.60
Robust FR [5]	WebFace	1	98.43	-	-	-	-
Lightened CNN [49]	WebFace	1	98.13	89.21	69.46	91.22	91.60
Softmax	WebFace$^+$	1	98.83	91.68	69.51	92.95	94.22
Noisy Softmax($\alpha^2 = 0.1$)	WebFace$^+$	1	**99.18**	**92.68**	**78.43**	**94.50**	**94.88**
Noisy Softmax($\alpha^2 = 0.05$)	WebFace$^+$	1	99.02	92.24	75.67	94.02	94.51

Table 5. Recognition accuracies (%) on LFW, FGLFW and YTF datasets. * denotes the images are not publicly available and $^+$ denotes data expansion. In LFW, closed-set and open-set accuracies are evaluated by Rank-1 and DIR@FAR=1 respectively.

References:

[1] Gulcehre, Caglar, et al. "Noisy activation functions." *International Conference on Machine Learning*. 2016.