NFLB Dropout:Improve Generalization Ability by Dropping out the Best

—A Biologically Inspired Adaptive Dropout Method For Unsupervised Learning

Peijie Yin*,Lu Qi†,Xuanyang Xi†,Bo Zhang*, Hong Qiao†‡
*Institute of Applied Mathematics, Academy of Mathematics and Systems Science,
Chinese Academy of Science, Beijing 100080, China
†State Key Lab of Management and Control for Complex Systems, Institute of Automation,
Chinese Academy of Sciences Beijing 100190, China,
hong.qiao@ia.ac.cn
‡CAS Centre for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Shanghai 200031, China

***Abstract*—Generalization ability is widely acknowledged as one of the most important criteria to evaluate the quality of unsupervised models. The objective of our research is to find a better dropout method to improve the generalization ability of convolutional deep belief network (CDBN), an unsupervised learning model for vision tasks. In this paper, the phenomenon of low feature diversity during the training process is investigated. The attention mechanism of human visual system is more focused on rare events and depresses well-known facts. Inspired by this mechanism, No Feature Left Behind Dropout (NFLB Dropout), an adaptive dropout method is firstly proposed to automatically adjust the dropout rate feature-wisely. In the proposed method, the algorithm drops well-trained features and keeps poorly-trained ones with a high probability during training iterations. In addition, we apply two approximations of the quality of features, which are inspired by theory of saliency and optimization. Compared with the model trained by standard dropout, experiment results show that our NFLB Dropout method improves not only the accuracy but the convergence speed as well.**

I. Introduction

In recent years, deep neural networks (DNNs) have been successfully applied to various areas. Unsupervised learning plays an important role in this boosting trend. Serving as a pre-training module, unsupervised neural network could effectively extract intrinsic information and provide a better initialization for further supervised training [1].

However, with massive parameters in deep models, overfitting is a tough problem, especially when the training data is limited. This problem become even serious during the training procedure of an unsupervised DNN. To combat this problem, multiple regularization methods have been proposed. Dropout, one of the most popular tricks, randomly masks the input and hidden nodes during training to restrain co-adaption of feature detectors [2], [3]. The dropout technique and its extensions like DropConnect [4] have significantly improved the performance of many state-of-art models.

Dropout can also be used in unsupervised learning of neural networks. Due to the lack of label information, unsupervised learning needs to extract comprehensive features from the data. Ignorance of significant features is harmful for the performance of unsupervised model. However, even with dropout, unsupervised training of DNN may have overfitting problems especially when the training dataset is relative small. Hence, generalization ability is the key quality for features learnt from unsupervised training.

There have been many sounding works to extract more generalized features during unsupervised training. Kim *et al* try to modify the learnt features by clustering and evolution search [5]. Dosovitskiy *et al* propose a method of transforming the unsupervised problem to a faked supervised discrimination problem [6]. They firstly compose randomly picked images with many transformations as different surrogate classes, each image for one class. A convolutional network is trained with these surrogate labeled data. Features learned by the network are more suitable to discrimination task. Szerlip *et al* build a model with similar ideas [7]. Instead of focusing on reconstruction ability, the model would generate features incrementally to maximize the discriminative ability among training data.

On the other hand, as we all know, human brain has marvelous ability to learn without supervision. Actually the structures and mechanisms of human brain have provided significant and numerous inspirations for artificial intelligence algorithms. In this special case, human has extraordinary generalization ability compared with current calculation model. An infant learns how to separate and classify objects at a very young age without much supervision. Some biology researches have revealed the fact that the attention plays an important role during the whole perception process. Most of current researches adopt saliency and attention mechanisms on the detection and recognition process to find the significant area of

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

input image [8], [9], [10]. We pay more attention to the training procedure and the feature space. As [11] suggests, during learning, the brain would cost less energy on the familiar parts and be more sensitive to surprising parts. Hence, the more features are well trained, the less probability it has to activate the learning process (contrast to the cognition process when familiar features will be activated stronger and faster).

In this paper, inspired by above mechanisms, we propose a novel adaptive dropout method, No Feature Left Behind Dropout (NFLB Dropout), to train convolutional deep belief network (CDBN), a widely used unsupervised network model for vision tasks. The name No Feature Left Behind is inspired by the No Child Left Behind Act[1]. However, it would be easily generalized to other hierachical connective models, such as stacked denoising autoencoder [12] and stacked convolutional autoencoder [13].

Unsupervised convolutional deep belief network (CDBN) is firstly introduced in Lees work [14] for feature extraction tasks. In our previous work [15], [16], CDBN has been used to extract episodic information from the image. As an unsupervised model, CDBN is able to generate good local features and encode common components by minimizing the reconstruction error. CDBN is composed of stacked Convolutional Restricted Boltzmann Machine (CRBM). CRBM, as a variant of RBM, can infer the original input from the activation and minimize the reconstruction error. This gives the model an autonomous ability to recall the visual information based on the weights of network [16].

During the training iteration of CDBN, the method we propose will approximate the degree of training on each feature map and then adjust and normalize the dropout rate of each feature map in this iteration respectively. The poorly-trained features have the higher probabilities to be trained whereas the well-trained feature will be more likely to be dropped.

The rest of this paper is organized as follows: Section II describes the motivation and inspiration of biological mechanisms, Section III introduces the previous researches related to this paper, Section IV analyzes the training process of CDBN, Section V presents two algorithms to estimate the quality of features and to propose the No Feature Left Behind Dropout to adpatively adjust dropout rate, Section VI evaluates the performance of proposed model under various conditions, Section VII discusses different perspectives to interpret the proposed model, Section VIII gives the conclusion and the directions for future work.

II. Biological Motivation

The basic idea of the proposed method is motivated by the association between selective attention and learning.

In visual learning tasks, attention plays an fundamental role [17]. It is the attention system that guides the brain where to gaze and what to learn. Serving as a drastic filter, our attention systems select a little of item information from some aspect of the sensory world and ignore others.[18]. The selection process is related to the perceptual process but the learning experiences as well.

In human brain, above two contributors of selection process form two mechanisms of attention system [19]. Perceptual processing provides the bottom-up attention for salient stimuli that pop out from their surroundings. Attention can also be voluntarily directed to objects that are currently important to the observer based on the learning experiences. In the context of learning, curiosity is one of the most important motivation. [18] also reveals that the attention would select the unfamiliar or surprising features by adding activations on the related neurons. And other features which have been well learnt, come into the level of automatically encoding and do not need to draw much attention. As analyzed in [20], curiosity-driven will help the brain reduce uncertainty and be more adaptive. When humans learn an unfamiliar task, they will move eyes faster to change attention for sampling rapidly and gaining more information from sensory input.

Mechanisms introduced above give us strong inspiration and motivations. From the view of attention, dropout is a kind of random sampling selection method, while humans select and drop information more intelligently. Thus, the original dropout policy can be improved by mimicking the strategy adopted in human attention systems.

III. Related Research

Since the dropout was proposed, there have been different explanations about how and why dropout works so well. In the original paper of dropout, Hinton *et al* [2] gives a possible biological interpretation of dropout that it prevent co-adaption between feature detectors, similar to the mechanism in evolution theory. Hinton *et al* [2,3] also points out that dropout is a special form of model ensemble. During training, each time after dropout, the network model will be randomly reduced to a smaller one. These networks then ensemble together during testing by using the complete network and divide the output by the dropout rate. Wager, Wang, and Liang [21] introduce another point of view of dropout as artificial noise. They attribute dropout as a special method of artificially corrupting training data for stabilization and then formalize dropout as an adaptive regularizer. Baldi and Sadowski [22] further derive an equivalent loss function of dropout, which can approximate the dropout effect on the accuracy. Wang [23] and Maeda [24] provide a Bayesian perspective of dropout. Gal and Ghahramani [25] extend the prospective by studying the properties of model uncertainty introduced by randomly dropout. A recent work [26] reveals that dropout can also be thought as a way of data augmentation. They project the activation of a hidden layer after randomly dropout back to the input space and show that dropout on hidden layers is equivalent to giving a special variation of original data.

Based on understandings above, several adaptive or accelerated dropout approaches have been proposed to enhance generalization ability and the training speed. Wang *et al*

[1]The model is named after the famous Act since they share the same idea that resources should lean more on children (features) with lower performance

propose a fast dropout training algorithm [23], accelerating the origin one by sampling from an equivalent objective function of dropout, rather than optimizing it. Ba and Frey [27] present a model to choose dropout rate adaptively by adding an extra restricted Boltzmann machine beyond the hidden layer. Other researchers adopt the Bayesian inference rule to choose a proper rate. Zhuo, Zhu and Zhang give a Bayesian based rule to automatically adjust the dropout rate [28]. Maeda *et al* interpret dropout by a Bayesian framework and optimize the rate based on Bayesian inference.

IV. Analysis of CDBN Training Process

In this section, we first introduce the notation, the structure and the training method of CRBM and CDBN, and then applied standard dropout in CRBM training. We also analyze the role of diversity among features from multiple perspectives and proposed the solution to enhance the diversity.

A. Structure of CDBN and Energy-based Training

We firstly introduce the structure of CDBN model and the energy-based training method.

According to Section I, a standard CDBN could be regarded as stacked CRBMs. Therefore, we can train a CDBN by training the CRBMs from the bottom to the top. The output from the previous layer can be treated as the input to the next layer. The architecture of CRBM is shown in Fig 1

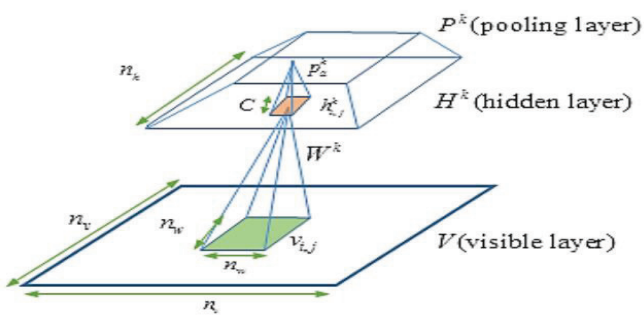

Fig. 1. A CRBM model with probabilistic max-pooling. Only the kth channels of H and P are shown for simplicity.

As shown in Fig. 3, a CRBM has three layers: visible layer $\mathbf{V}$, hidden layer $\mathbf{H}$ and pooling layer $\mathbf{P}$. n_v and n_h are the widths of $\mathbf{V}$ and $\mathbf{H}$, respectively.

The hidden layer $\mathbf{H}$ has K groups of feature maps denoted by H^k $(k = 1, 2, \cdots, K)$. It is connected with $\mathbf{V}$ by the shared local weights W^k. The width of W^k is n_w. The width of H^k is $n_h (= n_v - n_w + 1)$. We denote $v^k_{i,j}$ as an unit in layer $\mathbf{V}$ with row index i and column index j. And $h^k_{i,j}$ represents an unit in layer $\mathbf{H}^k$.

The pooling layer P also has K groups of feature maps. The unit p^k_α is obtained by pooling from a specific $c \times c$ block, denoted by B_α, of units in H^k. So the width of a feature map P^k $(k = 1, 2, \cdots, K)$ is $n_p = n_h/c$.

In mathematics, the CRBM model is a special form of energy-based models. Considering inputs V and hidden layer H with binary feature maps H^k, the energy of each possible state (v, h), where $v \in \mathbb{R}^{n_v \times n_v}$ and $h \in \mathbb{B}^{n_h \times n_h \times K}$ $(\mathbb{B} = \{0, 1\})$, is defined as

$$P(v, h; \boldsymbol{\theta}) = \frac{1}{\mathbb{Z}(\boldsymbol{\theta})} \exp(-E(v, h; \boldsymbol{\theta})) \tag{1}$$

where θ represents the parameter set $\{W, a, b\}$.

$$\begin{aligned} E(v, h) \quad = \quad & -\sum_{k=1}^{K} \sum_{i,j=1}^{n_h} h^k_{i,j} (\tilde{W}^k * v)_{i,j} - \sum_{k=1}^{K} b_k \sum_{i,j=1}^{n_h} h^k_{i,j} \\ & -a \sum_{i,j=1}^{n_v} v_{i,j} + \frac{1}{2} \sum_{i,j=1}^{n_v} v^2_{i,j}, \end{aligned} \tag{2}$$

where $h^k_{i,j}$ meets the constraint

$$\sum_{(i,j) \in B_\alpha} h^k_{i,j} \leq 1, \forall k, \alpha. \tag{3}$$

Here, $\tilde{W}^k$, representing the 180-degree rotation of matrix W^k, is the convolutional kernel, "$*$" denotes the convolution operation, b_k is the shared basis of all units in H^k, and a is the shared basis of visible layer units.

Contrastive Divergence (CD), which is an approximate Maximum-Likelihood learning algorithm [29], is used as the objective function to train the CRBM. Once the CRBMs are trained sequentially, we can combine these CRBMs as a well-trained CDBN.

B. Training CDBN with Dropout

Based on the section VI-A, this part will adopt dropout on to CDBN updating process by putting the feature with an randomly binary mask. It has been introduced in [30] about the way to application of dropout on probabilistic generative model like Restricted Boltzmann Machine (RBM). Here we will apply the dropout onto CDBN training.

In the standard Dropout RBM model, the RBM is augmented with a random binary vector $\boldsymbol{r}$.

$$P(\boldsymbol{r}, \boldsymbol{v}, \boldsymbol{h}; p, \boldsymbol{\theta}) = P(\boldsymbol{r}; p) P(\boldsymbol{v}, \boldsymbol{h} | \boldsymbol{r}; \boldsymbol{\theta}) \tag{4}$$

$$P(\boldsymbol{r}; p) = \prod_{j=1}^{F} p^{r_j} (1 - p)^{1 - r_j} \tag{5}$$

To adopt a standard dropout on CDBN, we can simply represent the convolutional network as a locally connected neural network with many nodes share the same weights.

$$P(v, h | \boldsymbol{r}; \boldsymbol{\theta}) = \frac{1}{\mathbb{Z}(\boldsymbol{\theta})} \exp(-E(v, h)) \prod_{j=1}^{F} g(h_j, r_j) \tag{6}$$

$$g(h_j, r_j) = \begin{cases} 0 & , \ r_j = 0 \\ h_j & , \ r_j = 1 \end{cases}$$

where $E(v, h)$ is calculated by (2).

C. Analysis and Possible Solutions of Feature Diversity Problem

When training the CRBM, feature diversity is harmed by the unequal training progress between feature maps. Fig. 2 visualize the process of features evolving during training process Each row includes the visualization of features and the corresponding gradient in different epochs. In training epoch 10, compared with the initialization, some of the features look more informative and like Gabor filters whereas some others are not change so much. And the gradients of the latter left behind features are also much smaller than that of the former informative features. It is reasonable if we consider the effect of momentum. For features with larger initial gradients, they are more likely break out of saddle points or plains thanks to the momentum. However, when the features are stuck, the effect of momentum on features are much insignificant because of very small initial gradients. Matthew effect occurs and the left behind features will always be trapped. Thus, those features would fail to learn efficient information and decrease the feature diversity of the whole model. Fig. 2 presents those features with red circles, whose local contrast is relative low compared to others.

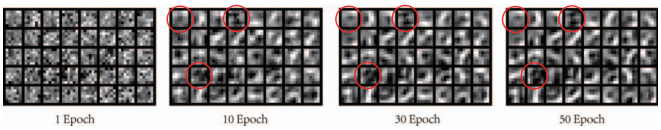

Fig. 2. An illustration about how feature evolves during training. The 40 features visualized in the figure are from the first CRBM. Even compared with the initialization, note that neurons like the top left one hardly changed.The less 'left behind' features are labeled with red circles.

The introduction of standard dropout would not eliminate the problem. Such results are not surprising. With a random mask and a uniform dropout rate, all the nodes in the hidden layer share the same probabilities to be kept. Thus, when the ratio of informative features to all features are high enough, most of gradients would be attributed to the better features. And again, Matthew effect happens.

To tackle this problem, we need a more adaptive method, as introduced in the next section, to adjust the dropout rate according the quality of features automatically.

V. NFLB Dropout CDBN

In this section, NFLB Dropout, an adaptive dropout method is proposed to deal with the problem mentioned in section IV-C. Inspired by the attention mechanism introduced in section II, the method will adjust attention to less trained features during previous step and form better and various understandings of training data. Other than the standard dropout which drops neurons randomly in hidden feature maps, the proposed method will drop an entire feature map. Each feature map shares a same dropout rate adapted to the corresponding kernel. And all the hidden features share the reconstruction gradient relatively equally in probabilistic perspective. In this section, two approximations of training progress are introduced and the procedures of NFLB Dropout are described in details.

A. Approximation of Feature Quality

To left no features behind, one firstly needs to find how the features have been trained and who are left behind. In this section, we will introduce two ways to evaluate the quality of features. The first one is inspired from theory of bottom-up saliency and attention. The second method is more consistent with learning mechanisms and top-down attention.

1) From bottom-up local contrast of reconstruction patches: As introduced in section II, attention plays an important role of human learning process. In the area of vision recognition, there have been many researches on effect of attention models. Reinagel and Zador [31] suggests that local spatial contrast is an important measure for human to decide where to gaze. Further studies in human vision system [32] and artificial intelligence [33] support this hypothesis. During the bottom-up attention model, human will be attracted more by patches are more informative.

Features of CDBN are all local features and would maximumly activated by some certain input patches. Thus, we can evaluate the information contained in those features by calculating the local contrast of reconstruction patches. The maximum activated input for a certain feature can be directly approximated by deconvolution from the top layer to the input space. Then local contrast-based filters are applied to those reconstructed image patches. The output of filters can be treated as the rates of information for features. Features drawing less attention would be the left behind ones. The details is shown in Algorithm 1

Algorithm 1 Approximate feature quality from bottom-up local contrast of reconstruction patches

Require: Parameters of Stacked CRBMs $\{\boldsymbol{\theta^i}:\{W^l, a^l, b^l\}\}$, here l denotes the lth layer, N is the number of layers, K is the number of features.

1: Reconstruct features in the hidden layer via deconvolution and unpooling layer by layer and get pathces p_i, p_i denotes the reconstrction of ith features.

2: Resize the reconstruction pacthes to a 3×3 matrix $\mathbf{M}$ by mean pooling

3: Local contrast LC(P_i) = $\sum_{p,q \in \mathbf{M}} ||I_p - I_q||$, where I_p, I_q is the pixel value at position p and q of the matrix $\mathbf{M}$.

Return: Quality vector of features, $\boldsymbol{q}$, where $q_i = LC(P_i)$

2) From top-down accumulating algorithm during training iterations: Another direction of attention is top-down adjustment. During the learning process, human would pay more attention to the unfamiliar characteristics, and less to those been learnt well. Inspired by this strategy, we can memory how much we have train a feature and then use the accumulation of gradients as a regularizer to measure the features quality historically.

The idea above is achieved by accumulating the gradients in the form used in AdaDelta [34] algorithm. For each feature, the gradients for updating gained from epochs are accumulated iteratively. Normalized all accumulated results cross features and we may got appropriate measure for quality of feature training. The details of application is shown in Algorithm 2.

Algorithm 2 Approximate feature quality from top-down accumulating gradients

Require: ρ Momentum of accumulated gradients
K Number of features
ϵ Tiny constant
Require: $g_i(t)$ Gradient on i_th feature at t epoch
Ensure: Initialize $E[g_I^2]$ at epoch 0,$E[g_i^2]_0 = \epsilon$
At epoch t
for $i = 1 : K$ **do**
$E[g_i^2]_t = \rho\ E[g_i^2]_{t-1} + (1-\rho)\ g_i^2(t)$
$RMS(g_i)_t = \sqrt{E[g_i^2]_t + \epsilon}$
end for
Return: Quality vector of features, $\boldsymbol{q}$, where $q_i = RMS(g_i)_t$

B. Adaptive Dropout For CDBN

With the approximation of feature quality, this part will introduce the detailed algorithm of NFLB Dropout.

Denote the quality approximation as $\boldsymbol{q}$. The approximated feature quality is defined as the quality ratio $\widetilde{\boldsymbol{q}}$, which is normalized by the maximum approximation $max(\boldsymbol{q})$.

We use the quality ratio to adjust dropout rate feature-wisely by multiplying them together as the dropout rate for certain features. Then we apply the adjusted dropout rate to CRBM. Note that the other than the standard dropout, NFLB Dropout directly drops the whole feature maps. Further details are included in Algorithm 3.

Algorithm 3 No Feature Left Behind Dropout

Require: p Dropout rate,
K Number of features
Ensure: Initialize $E[g_I^2]$ at epoch 0,$E[g_i^2]_0 = \epsilon$
At epoch t
Approximate $\boldsymbol{q}$ via Algorithm 1 or Algorithm 2
for $i = 1 : K$ **do**
Normalized quality: $\widetilde{q_i} = \frac{q_i}{\max_i(q_i)}, i = 1, ..., K,$
Adjusted Dropout rate: $\widetilde{p_i} = p\widetilde{q_i}$
if RAND(0,1)> $\widetilde{p_i}$ **then**
Retain feature map i in this epoch
else
Drop feature map i in this epoch
end if
end for
Return: All retained features

VI. Experiments

In this section we evaluate the improvement introduced by our model.

A. Hyperparameter Settings and Data Preparation

In this paper, we use a two layer CDBN model which have 40 feature maps in both CRBMs. The kernel sizes of two layers are set as 8x8 in the first layer CRBM and 6x6 in the second layer. The non-overlapped pooling sizes are 2 in both pooling layers. The momentums are set as 0.5 in the first 5 epochs and 0.9 in the rest epochs.

In the following experiments, we use MNIST, a standard hand-written digits dataset for training and testing. To better evaluate the generalization capacity, we randomly choose a relative small training dataset with 500 samples in total, and we do not balance different classes. Another 5000 samples are selected as test set.

B. Feature Representations on Small Training Data

We train the CDBN sequentially from bottom to top. Fig. 3 shows the evolving process of the weights of NFLB Dropout CRBM. In epoch 10, it seems that all the features are falling into the same local minimum. But it shows more and more diversity gradually. Compared to Fig. 2, we can find out that the weight still evolving after 30 epochs. It is noted that there are no obvious low quality 'left behind' features in Fig. 3 after 30 iterations. All these features contain a kind of topological information such as lines or arcs.

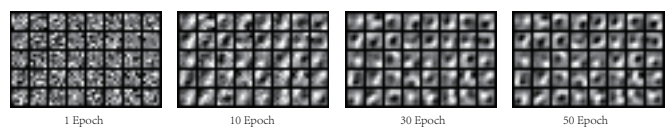

Fig. 3. An illustration about how feature evolves during NFLB Dropout training. The 40 features visualized in the figure are from the first CRBM. Note that neurons are still evolving after 30 epochs.

C. Classification Performance on Small Training Data From MNIST

We conduct NFLB Dropout CRBM with two approximation respectively on 500 MNIST training data, compared with original CRBM and CRBM with standard dropout. All the extracted features are fed into a SVM for classificaition.

The classification results are as shown in the Fig 4. More details are illustrated in Table 1. NFLB, as Fig 4 suggests, features extracted by the proposed model show more generalization ability, without sacrificing speed. Actually, since the dropout rate can be adjusted automatically, the model can bare a higher dropout rate and further reduce the complexity of model for one training epoch. Combine these effect together, NFLB Dropout method can not only improve the final precision of recognition, but also accelerate the convergence speed significantly.

As for the two approximation strategy of feature quality, we find out that estimation from the top-down process, the accumulation of gradients, performs generally better than the local contrast method in both accuracy and speed. This is consistent with the biology mechanisms introduced above that visual learning of unfamiliar task involves more active

Table 1. Classification Performance(%) on MNIST
(500 training samples for 50 epochs)

	Precision Rate (%)	Time Consumption Per Epoch (s)	Standard Deviation among features
One Layer CRBM	57.3	5.2	0.0374
CRBM w Dropout	68.2	4.3	0.041
NFLB (LC)	72.1	1.7	0.048
NFLB (AccumGrad)	**81.4**	**1.2**	**0.054**

top-down attention [18]. The relationship between standard deviation and recognition precision also prove our former analysis that the diversity in features is highly related to the generalization capacity.

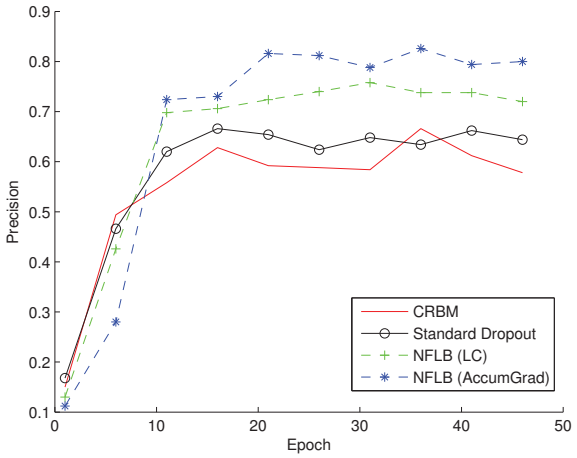

Fig. 4. Performance on test class of each model. The proposed two NFLB models are all outperform than previous standard models with the same epochs.

VII. Conclusion

In conclusion, we investigate the problem of low feature diversity during the training process and imitate the machanism in human brain, which is paying more attention to the surprising part and depressing the well-known fact. Mimicking this cognitive mechanism, we propose a novel adaptive dropout training method, No Feature Left Behind dropout method for unsupervised training. The proposed model shows outperforming generalization capacity

In the future, the method could be further applied to more hierarchical model and tested the efficiency and robustness. The theoretical analysis is also needed to find a better interpretation of model and a sounding explanation for the result. Also the relationship between feature diversity and its generalization capacity is worth to be deeply studied, which may lead to a promising direction to improve current unsupervised model.

Acknowledgment

This work is supported by the Strategic Priority Research Program of the CAS (GRANT XDB0208003).

References

[1] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent, "The difficulty of training deep architectures and the effect of unsupervised pre-training," in *International Conference on artificial intelligence and statistics*, 2009, pp. 153–160.

[2] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[4] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1058–1066.

[5] Y. Kim, W. N. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 365–369.

[6] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 766–774.

[7] P. Szerlip, G. Morse, J. Pugh, and K. Stanley, "Unsupervised feature learning through divergent discriminative feature accumulation," 2015.

[8] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 185–207, 2013.

[9] L. Wei, N. Sang, Y. Wang, and Q. Zheng, "A dynamic saliency attention model based on local complexity," *Digital Signal Processing*, vol. 22, no. 5, pp. 760–767, 2012.

[10] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 3, pp. 569–582, 2015.

[11] J. Gottlieb, "Attention, learning, and the value of information," *Neuron*, vol. 76, no. 2, pp. 281–295, 2012.

[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[13] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning–ICANN 2011*. Springer, 2011, pp. 52–59.

[14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 609–616.

[15] H. Qiao, X. Xi, Y. Li, W. Wu, and F. Li, "Biologically inspired visual model with preliminary cognition and active attention adjustment," 2014.

[16] H. Qiao, Y. Li, F. Li, W. Wu *et al.*, "Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning," 2015.

[17] T.-R. Huang and T. Watanabe, "Task attention facilitates learning of task-irrelevant stimuli," *PloS one*, vol. 7, no. 4, pp. e35 946–e35 946, 2012.

[18] D. LaBerge, "Perceptual learning and attention," *Learning and Cognitive Processes*, vol. 4, pp. 237–273, 2014.

[19] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: bottom-up versus top-down," *Current Biology*, vol. 14, no. 19, pp. R850–R852, 2004.

[20] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585–593, 2013.
[21] S. Wager, S. Wang, and P. S. Liang, "Dropout training as adaptive regularization," in *Advances in Neural Information Processing Systems*, 2013, pp. 351–359.
[22] P. Baldi and P. Sadowski, "The dropout learning algorithm," *Artificial intelligence*, vol. 210, pp. 78–122, 2014.
[23] S. Wang and C. Manning, "Fast dropout training," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 118–126.
[24] S.-i. Maeda, "A bayesian encourages dropout," *arXiv preprint arXiv:1412.7003*, 2014.
[25] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *arXiv preprint arXiv:1506.02142*, 2015.
[26] K. Konda, X. Bouthillier, R. Memisevic, and P. Vincent, "Dropout as data augmentation," *arXiv preprint arXiv:1506.08700*, 2015.
[27] J. Ba and B. Frey, "Adaptive dropout for training deep neural networks," in *Advances in Neural Information Processing Systems*, 2013, pp. 3084–3092.
[28] J. Zhuo, J. Zhu, and B. Zhang, "Adaptive dropout rates for learning with corrupted features," in *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 4126–4132.
[29] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
[30] N. Srivastava, "Improving neural networks with dropout," Ph.D. dissertation, University of Toronto, 2013.
[31] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.
[32] J. M. Henderson, "Human gaze control during real-world scene perception," *Trends in cognitive sciences*, vol. 7, no. 11, pp. 498–504, 2003.
[33] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.
[34] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.