Combination Normal Sparse and Discriminative Deep Belief Networks

Faisal Khalid
Department of Computer Science
University of Indonesia
Depok, West Java, Indonesia
faisal.khalid.si@gmail.com

Mohamad Ivan Fanany
Department of Computer Science
University of Indonesia
Depok, West Java, Indonesia
ivan@cs.ui.ac.id

***Abstract*—Problem in deep models is time consuming and may be trapped into local minima. One useful tool for dealing with his problem to use DBN (Deep Belief Network). Sparse representations are more efficient than non-sparse. This paper aim to find the best structure of combination sparsity and discriminative DBN. We use DBN architecture with 784 as input, 500 hidden unit, 500 hidden unit, 2000 hidden unit, and 10 as output. We took 3 step in this experiment, preliminary, intermediate and final result. Every analysis of each step be a background to the next step. We use normal sparse for sparse generative and sparse discriminative. Experimental studies on MNIST dataset show that the best structure on combination normal sparse and discriminative deep belief networks is with input- generative (Contrastive Divergence) - generative (Contractive Divergence) normal sparse discriminative (Contractive Divergence).**

***Keywords*—Deep Belief Network, Normal Sparse Discriminative Deep Belief Network**

I. Introduction

Artificial neural network have been used in pattern recognition, voice recognition and natural language processing. For some theoretical reason deep architecture were suggested [1]. Problem in deep models is time consuming and may be trapped into local minima. To solve that problem, we can use DBN (Deep Belief Network). DBN can create a neural networks which have several hidden layers [2]. Computing connection between hidden layer and visible layer in DBN is so difficult. Gibbs sampling can be used, although it takes a long time. So it is impossible to do.Contrastive Divergence (CD)[3], PCD [4], and FEPCD [5] are used to solved that problem. In recent years there has been an increasing interest in automatic feature extraction. Many models were then developed for such as deep learning with interest in sparse coding[6] methods. Sparsity has recently become a concept of great interest and has become a key ingredient in Deep Belief Network. Bengio has argued that sparse representations are more efficient than non-sparse in fixed size representations [7]. Several previous studies have done in sparse and discriminative deep belief network. The studies in[4] proposed semi-supervise learning by using discriminative Deep Belief Network(DDBN). [8] proposed a sparse deep belief networks and denoising autoencoder to a new dataset proposed in the ICDAR 2013 handwritten digit competition.[9] using discriminative RBM as classifiers can improve performance in a semi-supervised learning. The studies in[10] present a theoretical approach for sparse constraints in the DBN using the mixed norm for both- overlapping and overlapping groups. We argue that combination sparsity and discriminative DBN will increase the accuracy, but no one ever proposed what structure of that combination will give the best accuracy. This paper aim to find the best structure of combination sparsity and discriminative DBN. For ease of reproduction of results, this paper are performed on publicly available dataset (MNIST Dataset). This paper organized as follows. Section II explain literature study. Section III describes proposed methods. Section IV describe the experimental setup. Section V explain the experimental result and discussions. Finally conclusions are given in section VI.

II. Literature Review

A. Normal Sparse RBM

Essentially RBMs learn non-sparse distributed representations. In all proposed methods, learning algorithm in RBM has been changed to enforce RBM to learn sparse representation. The goal of sparsity in RBM is that most of hidden units have zero values and this equivalent to force activation probability of hidden units to zero. Normal sparse regulation term has a variance parameter that can control the force degree of sparseness. The regulation term used normal function properties[1]. In another paper based on rate distortion theory, the penalty factor is activation probability in hidden unit[11]. Algorithm sparse RBM learning algorithm[12]

1) Update the parameters using approximation to the gradient of log likelihood like CD, PCD or FEPCD.
2) Update the parameter using gradient of the regulation term.
3) Repeat steps 1 and 2 until convergence on reach to max epoch.

In [1] proposed a new regulation term that has different behavior according to deviation of the activation of the hidden units from a (low) fixed level p. They used normal probability density function as regulation term.

$$L_{sparsity} = \sum_{j=1}^{n} f(q_j, p, \sigma^2) \tag{1}$$

According to 1 the gradient if regularization term must be used in step 2. The gradient of regularization term can be computed as follows :

$$\frac{\partial}{\partial b_j} L_{sparsity} \propto \frac{1}{m}\left(p - \frac{1}{m}\sum_{l=1}^{m} q_j^{(l)}\right) f\left(q_j, p, \sigma^2\right) \times \sum_{l=1}^{m} q_j^{(l)}\left(1 - q_j^{(l)}\right) \quad (2)$$

B. Discriminative Restricted Boltzmann Machines (DRBM)

The important thing in classification is have a good and correct classification. we can optimize directly $p\left(x|y\right)$ below[9]

$$L_{disc}\left(D_{train}\right) = -\sum_{i=1}^{|D_{train}|} logp\left(y_i|x_i\right) \quad (3)$$

Based on L_{disc}, discriminative RBM (DRBM) can refer to RBMs trained. Contrastive divergence can be used in DRBM to training[13]. Gradient can be compute as:

$$\frac{\partial logp\left(y_i|x_i\right)}{\partial\theta} = \sum_j sigm\left(o_{yj}\left(x_i\right)\right)\frac{\partial o_{yj}\left(x_i\right)}{\partial\theta} - \sum_{i,y^*} sigm\left(o_{y^*}\left(x_i\right)\right)p\left(y^*|x_i\right)\frac{\partial o_{yj}\left(x_i\right)}{\partial\theta} \quad (4)$$

Where $o_{yj}\left(x\right) = c_j + \sum_k W_{jkxk} + U_{jy}$. Stochastic gradient descent optimization can be done by computed this gradient with efficient. Fine-tuning the top RBM of DBN is using this approach[14].

C. Free Energy in Persistent Contrastive Divergence

Gibbs sampling is an appropriate method as in RBM because each unit in RBM is independent from other units. CD, PCD, or FEPCD[5] can be used to obtain proper samples from the model because impossible to use Gibbs sampling. In PCD methods, many persistent chain can be run in parallel. If we can define the criteria for good chain, and for computing the gradient sample would be more accurate. Defining the best chain by using the free energy is as follows [15] :

$$F\left(v\right) = -\sum_i v_i a_i - \sum_j q_j I_j + \sum_j \left(q_j log q_j + \left(1 - q_j\right) log\left(1 - q_j\right)\right) \quad (5)$$

Sum of input to hidden unit j is equal with $I_j = b_j + \sum v_i w_{ij}$ and $q_j = q(I_j)$ is equal to activation probability of hidden unit h_j given v and q is logistic function. An equivalent and simpler equation for computing $F(v)$ is as follows:

$$F(v) = -\sum_i v_i a_i - \sum_j log(1 + e^{I_j}) \quad (6)$$

III. Proposed Method

Fig. 1. Show proposed method in this paper. The proposed method including three steps as follows

1) Train the first RBM by inputting the original data and fixing up the parameters of this RBM. Then we use these output as the input of the second RBM and the rest can be done in the same manner.
2) We use modified the RBM with normal sparse.
3) Fine-tuning: using Back Propagation (BP). We use gradient-descent algorithm to revise the weight matrix of the whole network.

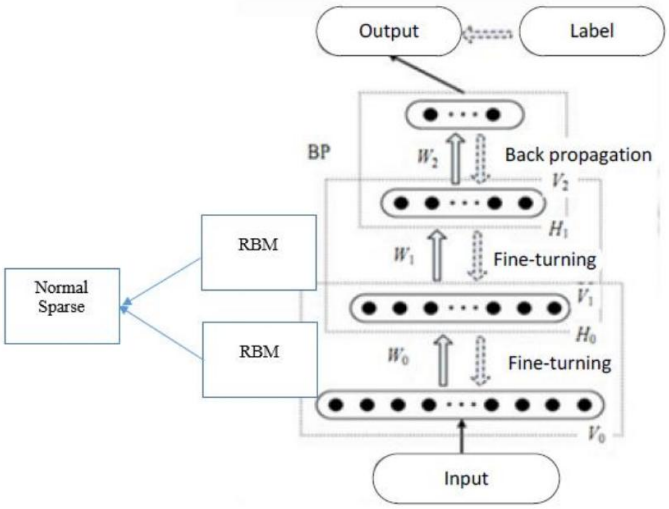

Fig. 1 Proposed Method

IV. Experimental Setup

The experimental study is carried out by used DeeBNet Matlab toolbox[16]. Also we used the new FEPCD method to approximate the gradient of the log likelihood. We use DBN architecture with 784 as input, 500 hidden unit, 500 hidden unit, 2000 hidden unit, and 10 as output. This architecture we used in each experiment.

A. Data

To evaluate the different DBN models on the the problem of classifying images of digits, we used small MNIST and MNIST dataset. MNIST dataset includes images of handwritten digits[17] (10 classes of digits 0-9). Each digit was cared to be located in the center of each 28*28 image. The image pixels have discrete values between 0 and 255 hat most of them have the values at the edge of this interval. The dataset was divided to train and test part including 60,000 and 10,000 images respectively. Inour experiment, these discrete values have been mapped to interval [0-1] using min-max normalization method. For preliminary and intermediate we use small MNIST. Small MNIST is part of MNIST dataset but only 6000 training data and 1000 testing data.

V. Experiment Result And Analysis

In this section we describe the performance of DBN by comparing with other modified DBN. We took 3 step in this experiment, preliminary, intermediate and final result. Every analysis of each step be a background to the next step. We use normal sparse for sparse generative and sparse discriminative.

A. Preliminary Result

In this step, we combine generative RBM, normal sparse generative RBM, discriminative RBM, and normal sparse generative RBM to find the best architecture to obtain best accuracy. Table I shown the accuracy obtained. Fig 2 sparsity feature for 2^{nd} Fig 3 sparsity feature for 4^{th} experiment. Fig 4 Shown reconstruction for 1^{st} experiment. Fig 5 reconstruction for 4^{th} experiment.

TABLE I Accuracy Obtained In Preliminary Result

Exp. No	DBN Structure	Error Before BP	Error After BP	Accuracy (%)	Epoch
1	Input- SG-SG-SD	0.568	0.055	99.945	200
2	Input-G-G-D	0.063	0.044	99.956	175
3	Input-SG-SG-D	0.063	0.095	99.905	200
4	Input-G-G-SD	0.067	0.038	99.962	124
5	Input-SG-G-SD	0.31	0.045	99.955	200
6	Input-G-SG-D	0.115	0.048	99.952	200

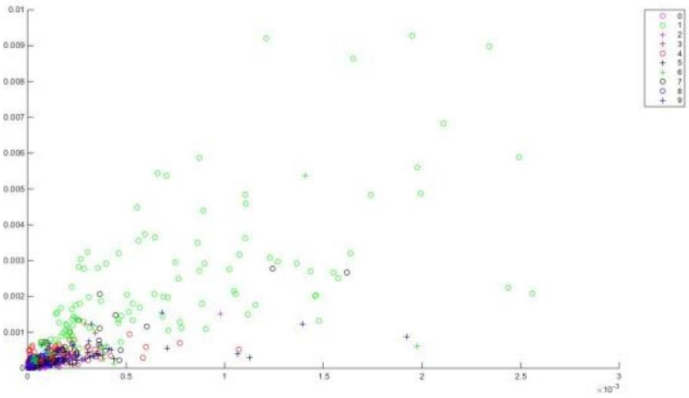

Fig. 2 Sparsity feature 2^{nd} experiment

From table I, we obtained 99.96.2with 124 epoch. Accuracy obtained using input-generative- generative-Sparse Discriminative. Based on fig 2 and fig 3, sparse feature can obtained by using sparse distributive RBM. Fig 4 and fig 5 shown that from reconstruct, generative RBM make feature more general than sparse generative RBM.

B. Intermediate Result

Based on preliminary result, we know that the best result is using input-generative-generative-Sparse Discriminative with 99.962we do experiment with another likelihood method, named FEPCD. In this experiment we implement FEPCD in our DBN structure and compare with CD. Experiment result shown in table II. Fig 6,7,8,9 shown that the confusion matrix for each experiment.

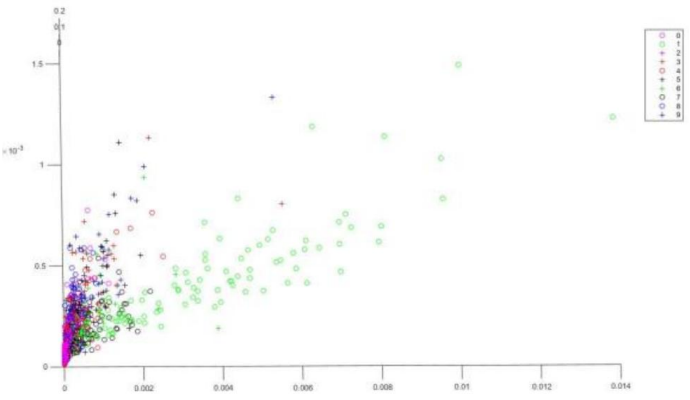

Fig. 3 Sparsity feature 4^{th} experiment

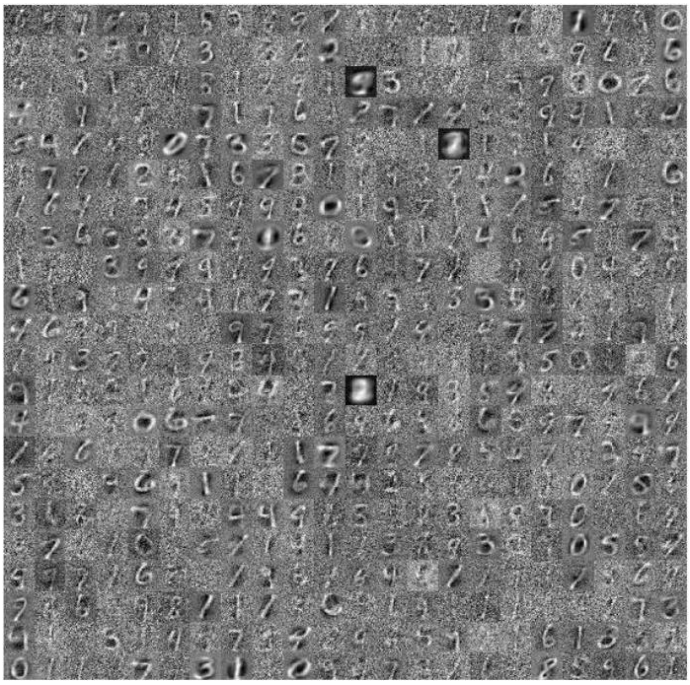

Fig. 4 RBM hidden layer in 1^{st} experiment

From the all experiments, we got conclusion that best accuracy obtained is 99.8generative (FEPCD)-sparse generative (FEPCD) and 116 epoch.

C. Final Result

From intermediate result we got conclusion that used FEPCD in all of RBM is better that CD, but its consuming time. For final result we did experiment for:

- Input- Generative(CD) - Generative(CD) - Sparse Generative(CD)
- Input-Generative(CD)-Generative(CD)-Sparse Generative(FEPCD)

We choose Input - Generative(CD) - Generative(CD) - Sparse Generative (FEPCD) because based on intermediate result, this structure give good accuracy and smallest epoch. The experiment result shown in table III. Confusion matrix for

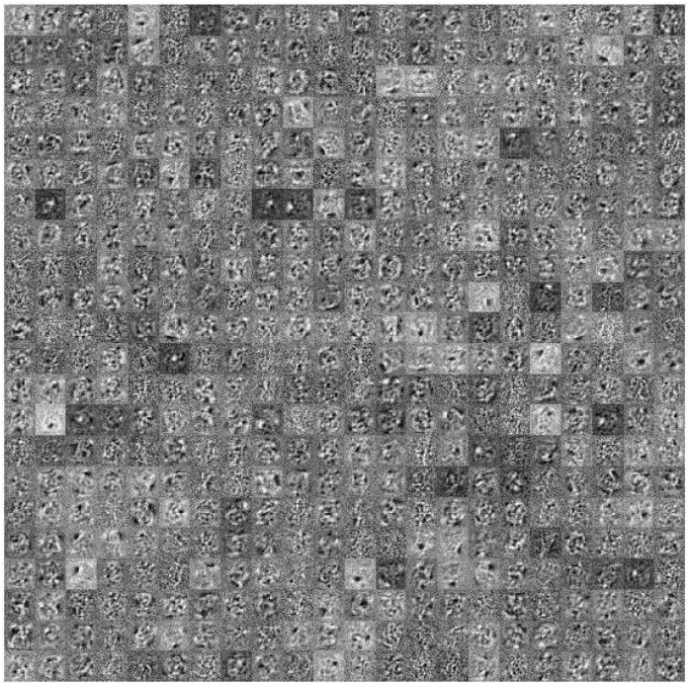

Fig. 5 RBM hidden layer in 4^{th} experiment

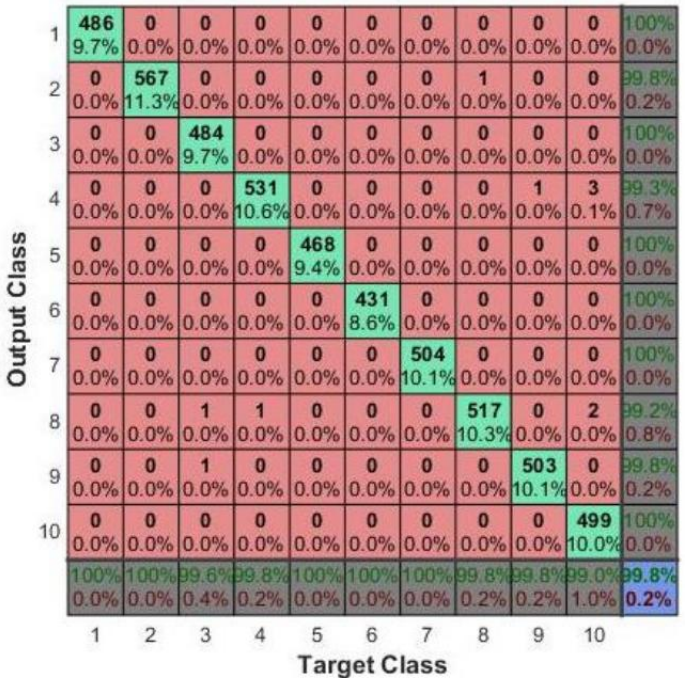

Fig. 6 1^{st} intermediate experiment confusion matrix

TABLE II Accuracy Obtained In Intermediate Result

Exp. No	DBN Structure	Error Before BP	Acc (%)	Epoch
1	Input- G(CD)-G(CD)-SD(CD)	0.0850	99.8	154
2	Input-G(FEPCD)-G(FEPCD) -SD(FEPCD)	0.1000	99.9	200
3	Input- G(CD)-G(CD)-SD(FEPCD)	0.0890	99.8	116
4	Input- G(FEPCD)-G(FEPCD) -SD(CD)	0.1090	99.8	125

all experiment shown in fig 10 and fig 11. For final experiment, we use full MNIST dataset.

TABLE III Accuracy Obtained In Final Result

Exp. No	DBN Structure	Error Before BP	Accuracy (%)	Epoch
1	Input- G(CD)-G(CD)-SD(CD)	0.0850	99.8	154
2	Input-G(FEPCD)-G(FEPCD) -SD(FEPCD)	0.1000	99.9	200
3	Input- G(CD)-G(CD)-SD(FEPCD)	0.0890	99.8	116
4	Input- G(FEPCD)-G(FEPCD) -SD(CD)	0.1090	99.8	125

VI. Conclusion

This paper investigates the best structure in combination normal sparse and discriminative deep belief network. Our experimental result show that DBN with input- generative (Contrastive Divergence) - generative (Contractive Divergence) normal sparse discriminative (Contractive Divergence) give the best accuracy. It is obtained 99.9862Sparse Discriminative make sparse feature. Sparse feature give better accuracy that non sparse feature.

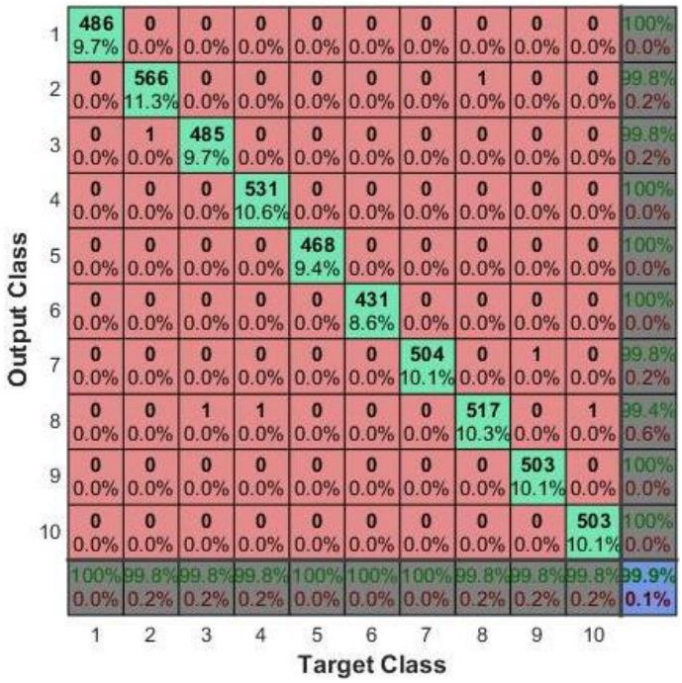

Fig. 7 2^{nd} intermediate experiment confusion matrix

Acknowledgment

The authors would like to thank...

References

[1] M. A. Keyvanrad and M. M. Homayounpour. Normal sparse deep belief network. In *2015 International Joint*

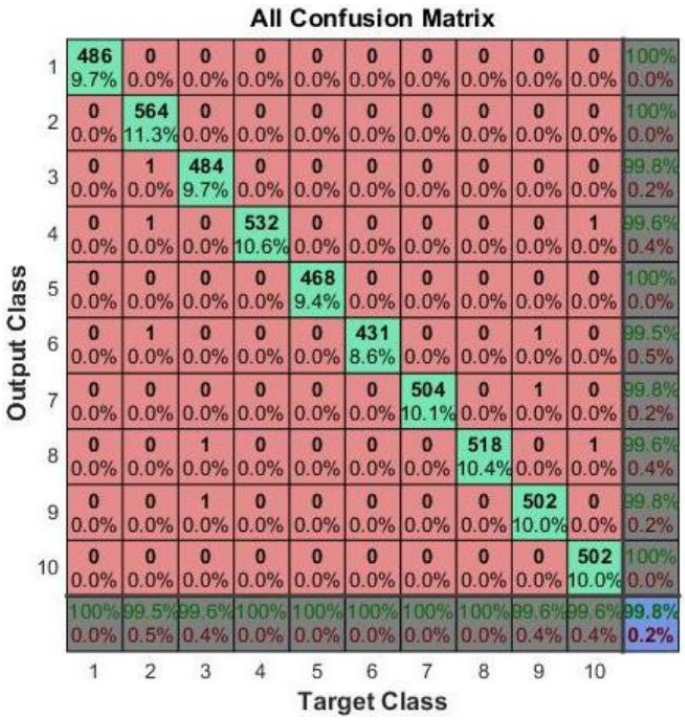

Fig. 8 3^{rd} intermediate experiment confusion matrix

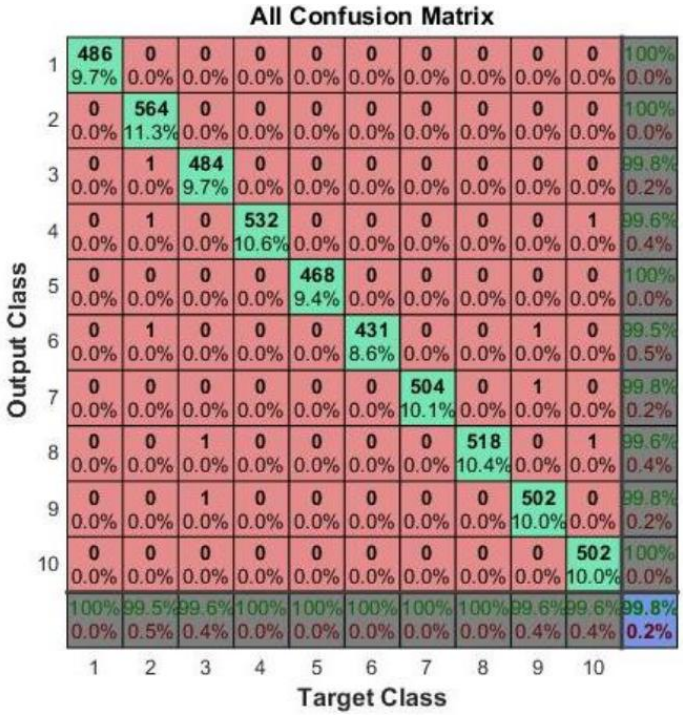

Fig. 9 4^{th} intermediate experiment confusion matrix

Conference on Neural Networks (IJCNN), pages 1–7, July 2015.

[2] Yan Liu, Shusen Zhou, and Qingcai Chen. Discriminative deep belief networks for visual data classification. *Pattern Recognition*, 44(1011):2287 – 2296, 2011. Semi-Supervised Learning for Visual Content Analysis and Understanding.

[3] Miguel A. Carreira-Perpinan and Geoffrey E. Hinton. On contrastive divergence learning. 2005.

[4] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1064–1071, New York, NY, USA, 2008. ACM.

[5] Mohammad Ali Keyvanrad and Mohammad Mehdi Homayounpour. Deep belief network training improvement using elite samples minimizing free energy. *CoRR*, abs/1411.4046, 2014.

[6] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[7] Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[8] R. Walid and A. Lasfar. Handwritten digit recognition using sparse deep architectures. In *Intelligent Systems: Theories and Applications (SITA-14), 2014 9th International Conference on*, pages 1–6, May 2014.

[9] Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 536–543, New York, NY, USA, 2008. ACM.

[10] Xanadu Halkias, Sébastien Paris, and Hervé Glotin. Sparse penalty in deep belief networks: Using the mixed norm constraint. *CoRR*, abs/1301.3533, 2013.

[11] Nan-Nan Ji, Jiang-She Zhang, and Chun-Xia Zhang. A sparse-response deep belief network based on rate distortion theory. *Pattern Recognition*, 47(9):3179 – 3191, 2014.

[12] Honglak Lee, Chaitanya Ekanadham, and Andrew Y. Ng. Sparse deep belief net model for visual area v2. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 873–880, USA, 2007. Curran Associates Inc.

[13] Graham W. Taylor, Geoffrey E. Hinton, and Sam Roweis. Modeling human motion using binary latent variables. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 1345–1352, Cambridge, MA, USA, 2006. MIT Press.

[14] Geoffrey E. Hinton. To recognize shapes, first learn to generate images. In Trevor Drew Paul Cisek and John F. Kalaska, editors, *Computational Neuroscience: Theoretical Insights into Brain Function*, volume 165 of *Progress in Brain Research*, pages 535 – 547. Elsevier, 2007.

[15] Geoffrey E. Hinton. *A Practical Guide to Training Restricted Boltzmann Machines*, pages 599–619. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[16] Mohammad Ali Keyvanrad and Mohammad Mehdi Homayounpour. A brief survey on deep belief networks

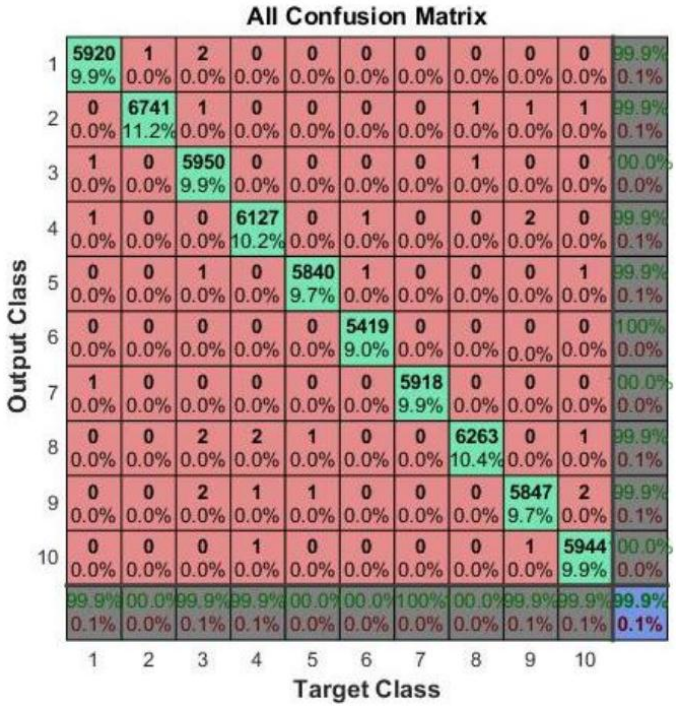

Fig. 10 1^{st} final experiment confusion matrix

and introducing a new object oriented MATLAB toolbox (deebnet). *CoRR*, abs/1408.3264, 2014.

[17] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

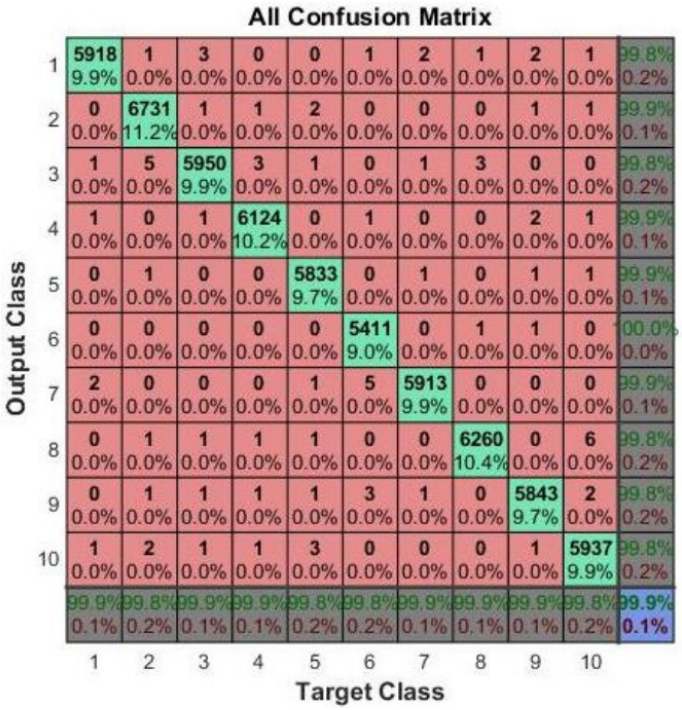

Fig. 11 2^{nd} final experiment confusion matrix