ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 134

2nd International Conference on Control, Automation, and Artificial Intelligence (CAAI 2017)

Improvement of GRBM Based on Activation Function

Ting Niu, Wenjing Huang and Xiang Gao

Department of Mathematics, Ocean University of China, China TSINGTAO

***Abstract*—In this paper, inspired by ReLu and Softplus activation function, we propose two improved models of GRBM, called SPC-GRBM and RPC-GRBM, to obtain better recognition results. Different from the traditional activation-function-improved models, SPC-GRBM and RPC-GRBM focus on the visual layer activation function, which is trained by CBCL database and is finally used for image classification with the help of the k-Nearest Neighbor (KNN) method. Experimental results show that the recognition accuracy of SPC-GRBM and RPC-GRBM are both enhanced and SPC-GRBM has achieved the highest recognition rate among the several models particularly, of which the recognition accuracy is 20.10% higher than the original GRBM. In addition, the reconstruction error is apparently reduced and its performance keeps well.**

Keywords-component; image recognition; Gaussian Boltzmann machine; ReLu activation function; Softplus activation function; parallel tempering

I. INTRODUCTION

Deep learning has been the most noticeable direction of machine learning during recent years. Since *A fast learning algorithm for deep belief nets* [1] was published by Hinton in 2006, the neural network research has been re-activated, which opened a new era of deep neural network. Deep learning has gradually achieved breakthrough in image recognition [2], voice recognition [3], text Processing [4] and other areas.

DBNs [5] is composed of many restricted Boltzmann machines (RBM) [6, 7], of which the parameters are learned by unsupervised methods layer by layer. RBM, as the basic component of DBNs, plays a prominent role in the improvement of the performance.

RBM is composed by a visual layer and a hidden layer, and both of them have binary neurons. We use contrast divergence (Contrastive Divergence, CD) [8, 9] method to train the preemptive value of RBM. As for image, real-valued image account for the majority of image instead of binary image. However, RBM is unable to deal with them. Therefore, the model is upgraded to Gaussian-Bernoulli Restricted Boltzmann Machines (GRBM) [10, 11]. In other words, the binary visual neurons are changed to Gaussian visual neurons, which means that the visual layer is a set of real-valued visual nodes obeying the Gaussian distribution, while hidden neurons are still binary. At the same time, to improve the sampling accuracy, Gibbs sampling is replaced by parallel tempering (Parallel Tempering, PT) [12, 13] method. Experiments show that PT provides a better training effectiveness for GRBM.

At the same time, along with the rise of sparse coding [14], some scholars combines RBM hidden layer with sparse coding. Among them, one of the most significant studies is the ReLuGRBM model proposed by Hinton [15]. This model replaces the activation function of hidden neurons with ReLu, which realizes the sparsity of the network and improves the recognition performance. Different from what the previous scholars have done, we divert our attention to the visual layer data. In the top-down sampling of the traditional GRBM, the sample value of visual layer is a random number generated by the corresponding Gaussian distribution. So it cannot overcome the problem of invalid data generation, resulted in fitted phenomenon and reducing training efficiency. The purpose of this paper is to improve the robustness of the feature by using the ReLu function to sparse the data, which will convert invalidate the data to 0. To further reduce the reconstruction error, we use the Softplus activation function to improve the model, and we achieve better results in the increase in recognition ability and reduce the reconstruction error.ve better results.

II. RESTRICTED BOLTZMANN MACHINE

The Boltzmann machine (Boltzmann Machine, BM) is a special Boltzmann machine. RBM is composed of a visible layer v and a hidden layer h, while there are weight connections between layers, there is no connection between the nodes on the same layer. The RBM is based on the energy model, nodes are all binary, that is, the node values 0 or 1. In order to meet the actual demand, it is often necessary to replace the binary visible node with a continuous real-valued number which obeys Gaussian distribution. This variant of RBM is called Gaussian Boltzmann machine (GRBM), of which the energy function is defined as:

$$E(v,h;\theta)=\sum_{i\in v}\frac{(v_i-a_i)^2}{2\sigma_i^2}-\sum_{j\in h}b_j h_j-\sum_{i\in v}\sum_{j\in h}\frac{v_i}{\sigma_i}h_j w_{ij} \qquad (1)$$

where $\theta=\{w,a,b\}$, a_i and b_j are biases corresponding to hidden and visible neurons. w_{ij} is weight connecting visible neuron v_i and hidden neurons h_j ,and σ_i is the standard deviation associated with a Gaussian visible neuron. The joint probability defined by GRBM energy function is:

$$p(v,h;\theta)=\frac{E(v,h;\theta)}{\sum E(v,h;\theta)} \qquad (2)$$

Calculate the marginal distribution over the visible layers by $p(v,h;\theta)$, then the likelihood function of observed data can

be obtained: $p(v;\theta)=\sum_h p(v,h;\theta)$ The parameters can be approximated in the processing of maximizing likelihood function. Gradient descent method is often employed to solve the optimization problem. The partial derivation formula can be obtained from log-likelihood:

$$\frac{\partial \ln P(v,\theta)}{\partial \theta} = E_{Pdata}[\frac{\partial E(V,H;\theta)}{\partial \theta}] - E_{P\theta}[\frac{\partial E(V,H;\theta)}{\partial \theta}] \tag{3}$$

Since the latter part of the partial derivative formula needs to calculate all the samples which resulted in a complex calculation, so the gradient descent method cannot be used directly. Hinton et al. proposed a CD algorithm that effectively approximates the gradient descent method by Gibbs sampling. In Gibbs sampling (1 step in general), the activation probability of visible and hidden nodes in GRBM can be deduced from $p(v,h;\theta)$:

$$\begin{aligned} P(h_j = 1 \mid v) &= sigmoid(b_j + \sum_i w_{ij} \frac{v_i}{\sigma_i}) \\ P(v_i = v \mid h) &= N(a_i + \sigma_i \sum_j w_{ij} h_j, \sigma_i^2) \end{aligned} \tag{4}$$

where $sigmoid(x) = \frac{1}{1+e^{-x}}$.

III. Improvement Of Grbm Based on Activation Function

In the negative sampling phase of GRBM model, the sampling of the visible layer is through generating a random number that obeys the corresponding Gaussian distribution as formula: $P(v_i = v \mid h) = N(a_i + \sigma_i \sum_j w_{ij} h_j, \sigma_i^2)$

In this paper, we consider adding the appropriate activation function to the visual node, so that the visual layer can remove the redundant data while preserving the data characteristics then we can improve the ability of the model to extract the feature. In the previous study, Hinton [15] and others proposed that the hidden layer can adopt the ReLu activation function, that is, $f(x) = \max(0, x)$, so that part of the hidden layer neurons output 0 which increases the network sparsity and improves the model ability to extract features. Different from previous studies, here ReLu is used in the visible layer of GRBM to accelerate convergence and improve the performance of the model. Based on this, the sampling criteria for the negative sampling phase in the improved GRBM model is defined as:

$$v_i \mid h = \max(0, N(a_i + \sigma_i \sum_j w_{ij} h_j, \sigma_i^2)) \tag{5}$$

For convenience of writing, the improved GRBM model by ReLu is abbreviated as RPC-GRBM.

Eq.(5) shows the 'fragile' defects of ReLu function, that is, neurons tend to be 0 constantly. Further, we use the smooth approximation of ReLu, Softplus function ($f(x) = \ln(1+e^x)$), to replace ReLu as the "activation function" of the visual layer, and we name this model as SPC-GRBM.

The traditional GRBM model usually adopts the contrast divergence algorithm in the training process, so when updating parameters, only the last updated data can be used. In order to make full use of the calculated parameter data, this paper adopts the improved CD algorithm [16]. The parameter-update process includes the information of the existing parameter data, and the parameter continuity of change is well enhanced. The improved parameter update formulas are:

$$\begin{aligned} w_{ij}^{(\tau+1)} &= w_{ij}^{(\tau)} + \alpha \cdot \left(\left\langle \frac{1}{\sigma_i} v_i h_j \right\rangle_{data} - \left\langle \frac{1}{\sigma_i} v_i h_j \right\rangle_{recon} \right) \\ w_{ij}^{(\tau+1)} &\leftarrow \frac{2 \times w_{ij}^{(\tau+1)} + (n+1) \times w_{ij}^{(\tau)}}{n+1} \\ a_i^{(\tau+1)} &= a_i^{(\tau)} + \alpha \cdot \left(\left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{data} - \left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{recon} \right) \\ a_i^{(\tau+1)} &\leftarrow \frac{2 \times a_i^{(\tau+1)} + (n-1) \times a_i^{(\tau)}}{n+1} \\ b_j^{(\tau+1)} &= b_j^{(\tau)} + \alpha \cdot \left(\left\langle h_j \right\rangle_{data} - \left\langle h_j \right\rangle_{recon} \right) \\ b_j^{(\tau+1)} &\leftarrow \frac{2 \times b_j^{(\tau+1)} + (n-1) \times b_j^{(\tau)}}{n+1} \end{aligned} \tag{6}$$

where α is the learning rate,and $\langle \cdot \rangle_{mean}$ represents the mathematical expectation of the reconstructed data obtained by parallel tempering (PT) sampling.

PT sampling is a more efficient sampling method than Gibbs sampling. During the GRBM training, M different temperatures correspond to M Gibbs chains, and the temperature t_i is in the range of $1 = t_1 < t_2 < \cdots < t_i < \cdots t_M$. The joint probability of the corresponding Gibbs chains at different temperatures is:

$$P_t(v,h) = \frac{1}{Z(t_i)} \exp(-\frac{1}{t_i} E(v,h;\theta)), \quad i = 1,2 \cdots M \tag{7}$$

Therefore, the parameter formula in PT sampling is temporarily changed to:

$$w_{ij}^{(t)} = \beta w_{ij}, \quad a_i^{(t)} = \beta a_i, \quad b_j^{(t)} = \beta b_j, \quad \sigma_i^{(\beta)} = \sqrt{\beta \sigma_i^2} \tag{8}$$

where $\beta = 1/t_i$.

When the Gibbs chain sampling is completed, the Gibbs chains corresponding to the adjacent temperatures are used to determine whether or not to exchange the sample values according to certain formulas. (v_τ, h_τ) and $(v_{\tau-1}, h_{\tau-1})$ are respectively sampled in adjacent temperature $(t_\tau, t_{\tau-1})$ and their exchange probability is:

$$\min\{1.\exp((\frac{1}{t_\tau} - \frac{1}{t_{\tau-1}}) \times (E(v_\tau, h_\tau) - E(v_{\tau-1}, h_{\tau-1})))\} \tag{9}$$

After several sampling exchanges, the Gibbs sample value corresponding to the temperature $t_1 = 1$ is used to train the GRBM model.

In summary, the model training process is as follows:

Step 1. Input: Training samples x_i ,number of hidden layer neurons m ,learning rate α ,maximum iterations $max\ epoch$,model parameters w, a, b ,where $0 < \beta_1 < \cdots < \beta_i < \cdots < \beta_M = 1$

Step 2. Initialization: w, a, b

Step 3 Parallel tempering sampling: For the M Gibbs chain sampling $(v_i, h_j) = ((v_1, h_1), (v_2, h_2), \cdots (v_m, h_m))$

Step 4. Exchange: According to the temperature chain exchange formula to calculate the adjacent Gibbs chain sampling value exchange probability to determine whether the exchange, and finally take $\beta = 1$ that is $t = 1$ sampling point as a training parameter value.

Step 5. Update Parameters: Update w, a, b as follows:

$$w_{ij} \leftarrow w_{ij} + \alpha \cdot \left(\left\langle \frac{1}{\sigma_i} v_i h_j \right\rangle_{data} - \left\langle \frac{1}{\sigma_i} v_i h_j \right\rangle_{recon} \right)$$

$$a_i \leftarrow a_i + \alpha \cdot \left(\left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{data} - \left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{recon} \right)$$

$$b_j \leftarrow b_j + \alpha \cdot \left(\left\langle h_j \right\rangle_{data} - \left\langle h_j \right\rangle_{recon} \right)$$

Step 6. Output: w, a, b

IV. EXPERIMENTS

A. Experimental Configuration

In order to examine the performance of models improved by ReLu and Softplus, we set the control model PC-GRBM. PC-GRBM also uses improved CD algorithm and parallel tempering sampling method, but does not change the sampling rules of visual layer.

In this section, we tested the GRBM, PC-GRBM, RPC-GRBM, and SPC-GRBM models based on the CBCL face database [17] to verify the performance of the SPC-GRBM and compare it with other models. CBCL database is divided into training set and test set, and only the faces from the training set are used in this paper. Training set contains 2429 faces and 5458 non-faces, which are all gray and in 19×19 size. The experimental process is divided into two stages. In the first stage, the four models are preliminarily trained by the randomly selected 2400 faces. In the second stage, 3000 images, which are used in the model randomly selected from Face and Non-Face, are put into the model, and the recognition results obtained by the model are classified by the nearest neighbor (KNN) [18] classification method.

In the pre-training stage, we trained four models with 361 visible neurons and 300 visible. And all the initial weights and the bias of each node were set to random numbers satisfying the equal distribution and 0, respectively. The learning rate is fixed to 0 with the maximum number of iterations set to 2000. Experiments show that when it iterates to 2000, the recognition rate of the model has been basically converged.

When training GRBM, we used a single Gibbs step in CD learning. But in other three models, we used the improved CD algorithm with a single Gibbs step. In the parallel tempering stage, the number of temperatures is set to $M = 10$, $\beta \in \{0, 0.1, 0.2 \cdots 1\}$.

B. Analysis of Experimental Results

Table I lists the recognition accuracy of different models under the same set of images, which contains 1000 faces and 2000 non-faces.

TABLE I. GRAY IMAGE RECOGNITION ACCURACY

Training model	***Recognition rate（2000）***
GRBM	75.40%
PC-GRBM	86.13%
RPC-GRBM	91.47%
SPC-GRBM	95.59%

It can be seen from Table 1 that under the same training conditions, PC-GRBM raised the recognition accuracy by 10.73% against original GRBM, with the use of improved CD algorithm combined with parallel tempering algorithm, which greatly improves the recognition effect. When PC-GRBM was improved by ReLu activation function, it achieved a further raise by 5.43%. The improved model of Softplus function had the highest recognition accuracy among the four models, which is 20.19% higher than that of the original GRBM, indicating that the SPC-GRBM model has good recognition performance.

Figure 1 shows how the recognition accuracy, which were calculated by KNN, changed with iteration. The four models had sharp contrast. The recognition accuracy of the improved model was significantly larger than that of the original GRBM and SPC-GRBM, and its recognition accuracy stayed top beyond doubt. It can be also seen from Figure 1 that the recognition accuracy of all the three models had quick convergence.

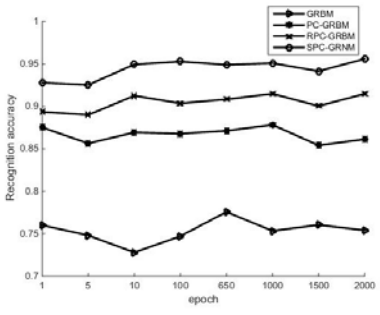

FIGURE I. CONTRAST CURVE OF RECOGNITON

Figure 2 shows a comparison of reconstruction error curves between GRBM, PC-GRBM, RPC-GRBM and SPC-GRBM. It can be seen from the figure that the reconstruction error of the four models had been controlled within a certain

range when the parameter iterates to 800, and SPC-GRBM owns the smallest reconstruction error and a quick convergence, indicating that SPC-GRBM has good reconstruction ability.

For further testing of SPC-GRBM and comparing the ability to extract features of four models, we randomly selected 12 samples from the gray image generated by the four models at the end of the training, as shown in Figure.3. Figure a is selected form the original face map randomly. Figure b-e are reconstruction images of the selected face respectively made by GRBM, PC-GRBM, RPC-GRBM and SPC-GRBM. It can be clearly seen that only the reconstruction image of SPC-GRBM model can restore the sample color and appearance changes. Fig b-d lost the original characteristics of the selected face. It

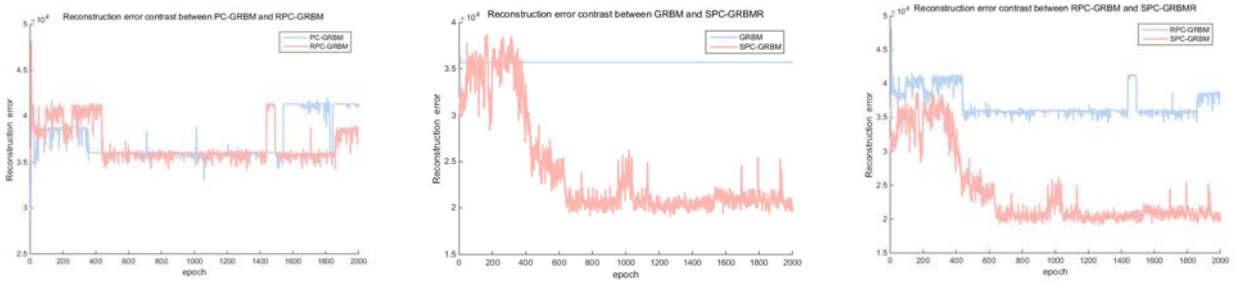

FIGURE II. CONTRAST CURVE OF RECONSTRUCTION ERROR

a b c d e

FIGURE III. VISUALIZATION OF THE LEARNED VARIANCESshows that SPC-GRBM is much better at extracting features than other three models.

V. SUMMARY AND PROSPECTS

In this paper, RPC-GRBM, SPC-GRBM, which are improved the visual layer activation function, achieved good results. Compared with the standard GRBM, SPC-GRBM can effectively simplify the operation and have a quick convergence, and effectively improve the recognition accuracy. In the future, we will try to imply this method on other types of RBM to achieve better recognition accuracy.

REFERENCES:

[1] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7):1527-1554.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] Mohamed A, Dahl G E, Hinton G, et al. Acoustic Modeling Using Deep Belief Networks [J]. IEEE Transactions on Audio, Speech and Language Processing,2012,20(1):14-22.

[4] He X, Wang D, Li Y, et al. A Novel Bearing Fault Diagnosis Method Based on Gaussian Restricted Boltzmann Machine[J]. Mathematical Problems in Engineering, 2016, 2016.

[5] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[6] Fischer, A. & Igel, C. Training restricted Boltzmann machines: an introduction. Pattern Recogn, 2014, 25, 25–39.

[7] Hinton G. A practical guide to training restricted Boltzmann machines[J]. Springer Berlin Heidelberg, 2012, 7700:599-619.

[8] Hinton G E. Training Products of Experts by Minimizing Contrastive Divergence[J]. Neural Computation, 2002, 14(8):1771-1800.

[9] Hinton G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8):1771-1800.

[10] [KH Cho,A Ilin,T Raiko. Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines[J]. Lecture Notes in Computer Science, 2011, 6791: 10-17

[11] Zhang H, Zhang S, Li K, et al. Robust shape prior modeling based on Gaussian-Bernoulli restricted Boltzmann Machine[C].//IEEE, International Symposium on Biomedical Imaging. IEEE, 2014:270-273.

[12] KyungHyun Cho, Alexander Ilin, and Tapani Raiko,etall. Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines［J］. Lecture Notes in Computer Science, 2011, 6791:10-17.

[13] Fischer A, Igel C. A bound for the convergence rate of parallel tempering for sampling restricted Boltzmann machines[J]. Theoretical Computer Science, 2015, 598:102-117.

[14] Wright J, Ma Y, Mairal J, et al. Sparse Representation for Computer Vision and Pattern Recognition[J]. Proceedings of the IEEE, 2010, 98(6):1031-1044.

[15] Nair V, Hinton G E. Rectified Linear Units Improve Restricted Boltzmann Machines[J]. Proc Icml, 2015:807-814.

[16] Zhao Caiguang, Zhang Shuqun, Lei Zhaoyi. Speech Recognition of Gaussian-Bernoulli Restricted Boltzmann Machine Based on Improved Contrastive Divergence[J].Computer Engineering, 2015,41(5):213-218.

[17] MIT Center For Biological and Computation Learning: CBCL Face Database #1,http://www.ai.mit.edu/projects/cbcl

[18] Guo G, Wang H, Bell D, et al. KNN Model-Based Approach in Classification[J]. Lecture Notes in Computer Science, 2003, 2888:986-996.