2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Convolutional Deep Belief Networks for Feature Extraction of EEG Signal

Yuanfang Ren and Yan Wu
Department of Computer Science and Technology
Tongji University
Shanghai, China
yanwu@tongji.edu.cn

***Abstract*—In recent years, deep learning approaches have been successfully used to learn hierarchical representations of image data, audio data etc. However, to our knowledge, these deep learning approaches have not been extensively studied for electroencephalographic (EEG) data. Considering the properties of EEG data, high-dimensional and multichannel, we applied convolutional deep belief networks to the feature learning of EEG data and evaluated it on the datasets from previous BCI competitions. Compared with other state-of-the-art feature extraction methods, the learned features using convolutional deep belief network have better performance.**

Keywords—deep learning; EEG; convolutional deep belief networks; feature learning

I. INTRODUCTION

Brain-computer interfaces (BCI) allow people to control the external environment through direct measures of brain signal [1]. Now BCI applications have played an important role in the field of rehabilitation engineering, military, entertainment etc. Usually the EEG signal processing can be divided into three phases: preprocessing, feature extraction and classification, among which feature extraction has a great influence on the final recognition accuracy of the EEG signal. The state-of-the-art feature extraction methods which have been successfully applied into the EEG feature extraction include band power [2], multivariate adaptive autoregressive (MVAAR)[3], common spatial pattern (CSP)[4], independent component analysis (ICA)[5] etc. However, these feature representations are often hand-designed, requiring lots of domain knowledge and human labor. Thus it is better to obtain feature representations automatically through a small amount of labeled data. And Schalk et al.[6] also explicitly proposed that building a practical BCI system based on a small training set has been a challenging problem.

Given the fact that a large amount of unlabeled data is usually available, we consider the problem of learning feature representations from unlabeled data, which is called unsupervised feature learning. And fortunately the rapid developing deep learning technologies, which usually build features from unlabeled data, give us some inspiration. Deep learning methods usually use a variety of approaches such as RBMs [7], autoencoders [8], sparse coding [9] etc. The deep belief network (DBN) [7] is a classical generative probabilistic model composed of some RBMs, in which the higher layer of DBN tends to learn more complex feature representation. DBN can be trained in an unsupervised way using greedy layerwise training. Based on DBN, Lee et al. [10] proposed the convolutional deep belief network (CDBN), scaling up the algorithm to deal with high-dimensional data. And Lee et al. also demonstrated that CDBN had good performance in several vision recognition tasks [10] and audio classification tasks [11]. In BCI area, some researchers have attempted to introduce some deep learning approaches into the EEG signal processing in recent years. Wulsin et al. [12] applied DBN to model EEG waveforms for classification in a semi-supervised paradigm. And literature [13] used convolutional neural network with embedded fourier transform for EEG classification in a supervised paradigm, which is trained by backpropagation.

Considering the properties of EEG data, high-dimensional and multichannel, in this paper we apply CDBN to the feature extraction of EEG signal, hoping that it can finish the unsupervised feature learning from unlabeled data efficiently. We mainly conduct the experiments on the previous BCI competition datasets to evaluate its performance. We also make comparison with other excellent feature extraction methods.

The rest of this paper is organized as follows: Section 2 introduces the method CDBN briefly and Section 3 presents the experiments and results. Finally, conclusions and future work are discussed in Section 4.

II. METHODS

A. Convolutional Deep Belief Networks

Since CDBN is composed of convolutional restricted Boltzmann machines (CRBMs), we first briefly introduce CRBM. And as EEG data is real-valued, we mainly introduce the Gaussian convolutional RBM. The basic CRBM consists of two layers: a visible layer *V* and a hidden layer *H*. Suppose the visible layer consists of *L* channels, and each channel consists of $N_V \times N_V$ real-valued units. The hidden layer consists of *K* groups, and each group consists of $N_H \times N_H$ hidden units. Denote the convolutional filter of size $ws \times ws$ ($ws \triangleq N_V - N_H + 1$) for the *l*-th channel corresponding to the *k*-th hidden group as $\mathbf{W}^{k,l}$, and it is shared among the hidden units in the *k*-th group. To make CRBM more scalable, Lee et al. [10] further developed "probabilistic max-pooling", which can shrink the

representations of higher layers in a probabilistically sound way. Probabilistic max-pooling enables the CRBM to incorporate max-pooling-like behavior, while supporting probabilistic inference. The energy function of the probabilistic max-pooling CRBM (with real-valued visible units) is defined as follows [10][14]:

$$E(\mathbf{v},\mathbf{h})=\frac{1}{2\sigma^2}\sum_{l=1}^{L}\sum_{i,j}(v_{i,j}^{l}-c_l)^2-\sum_{k=1}^{K}\sum_{i,j}h_{i,j}^{k}\left(\sum_{l=1}^{L}\frac{1}{\sigma}(\tilde{\mathbf{W}}^{k,l}*\mathbf{v}^{l})_{i,j}+b_k\right) \quad \text{subject to} \sum_{(i,j)\in B_\alpha}h_{i,j}^{k}\le 1,\ \forall k,\alpha \tag{1}$$

where $\mathbf{v}\in\mathbb{R}^{N_V\times N_V\times L}$ denotes the visible units, $\mathbf{h}\in\mathbb{R}^{N_H\times N_H\times K}$ denotes the hidden units. The parameter σ denotes the standard deviation associated with the Gaussian visible units. The visible units in the l-th channel share the bias c_l, and the hidden units in the k-th group share the bias b_k. Moreover, $\tilde{\mathbf{W}}$ denotes flipping the array horizontally and vertically. B_α refers to a $C\times C$ block of locally neighboring hidden units $h_{i,j}^{k}$ and is pooled to a pooling node p_α^k. There is a constraint between the hidden units in the block B_α and the corresponding pooling node p_α^k: at most one hidden unit in the block can be activated at a time and the pooling node is activated if and only if a hidden unit in the block is activated. A schematic description of a probabilistic max-pooling CRBM is shown in Fig. 1.

Then we can perform block Gibbs sampling using the following conditional distributions [14]:

$$P(\mathbf{v}^l\mid\mathbf{h})=\mathcal{N}(\mathbf{v}^l;\sigma\sum_k\mathbf{W}^{k,l}*\mathbf{h}^k+c_l,\sigma^2\mathbf{I}) \tag{2}$$

$$P(h_{i,j}^{k}=1\mid\mathbf{v})=\frac{\exp(I(h_{i,j}^{k}))}{1+\sum_{(i',j')\in B_\alpha}\exp(I(h_{i',j'}^{k}))} \tag{3}$$

$$P(p_\alpha^k=1\mid\mathbf{v})=\frac{\sum_{(i',j')\in B_\alpha}\exp(I(h_{i',j'}^{k}))}{1+\sum_{(i',j')\in B_\alpha}\exp(I(h_{i',j'}^{k}))} \tag{4}$$

$$I(h_{i,j}^{k})\triangleq\sum_l\frac{1}{\sigma}(\tilde{\mathbf{W}}^{k,l}*\mathbf{v}^{l})_{i,j}+b_k \tag{5}$$

where $I(h_{i,j}^{k})$ is the signal hidden units in group k received from layer V, and $\mathcal{N}(\cdot)$ is a normal distribution.

Thus the CDBN is stacked by some probabilistic max-pooling CRBMs. Like the standard RBM, the CRBM can be trained using contrastive divergence [7]. At the same time, since the CRBM is highly overcomplete, sparsity regularization [10] is also added to constrain the hidden units to having sparse average activations. And the CDBN's training is same with DBN, accomplished by the greedy layer-wise procedure [7]: once a layer is trained, its parameters are frozen and its activations are severed as the input of next layer. CDBN is a generative model, and supports efficient bottom-up and top-down probabilistic inference. However, considering the real-time requirement of BCI systems, for inference, we use feed-forward approximation.

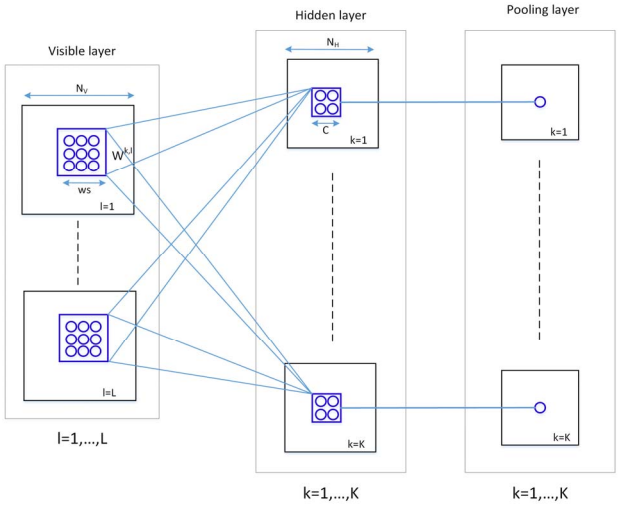


Fig. 1. A schematic description of a probabilistic max-pooling CRBM. N_V and N_H refer to the size of visible and hidden layer, and ws refers to the size of convolutional filter size. C refers to the max-pooling ratio.

B. *Application to EEG Data*

Before explaining our motivations to apply CDBN to the feature extraction of EEG signal, we first make a comparison between CDBN and DBN.CDBN is an extension of DBN and they have a lot in common, including the way of stacking building blocks, the training way of the building blocks and the network. The biggest difference between them is the structure of building blocks. CRBM is proposed mainly to address the issue of scaling models to realistic-sized images. It's main idea, weight sharing, allowing the weights between the hidden and visible layers are shared among all locations in an image, which can greatly reduce a large amount of parameters and make the representations be invariant to small translations of the input. At the same time, probabilistic max-pooling operation can further reduce the computational burden while allowing full probabilistic inference. These structures make it possible to scale CDBN to full images or high-dimensional data. Thus, it is reasonable for us to apply CDBN to EEG signal, since EEG data is high-dimensional. Moreover, CDBN can deal with data whose structure is multichannel, such as EEG data. Finally, the feature learning with CDBN is in an unsupervised way which allows us to make use of a large amount of unlabeled data.

For the application of CDBN to EEG data, as features in frequency-domain are more obvious than those in time-domain for EEG data, for each channel, we first convert time-domain signals into frequency-domain signals by fourier transform and choose the signal in 8-30Hz frequency band. As the number of channels of EEG data is large (e.g., 118 channels of dataset 2), to reduce channels, we apply PCA whitening to the signal and create lower dimensional representations, which is an important step in our experiments. As a result, the data fed into the CDBN is comprised of L channels of one-dimensional vector, where L is the number of PCA components.

Corresponding to the frequency band we choose, we build a two-layer CDBN with a filter length of 6 and a max-pooling ratio of 3 on layer 1 and a filter length of 4 and a max-pooling ratio of 3 on layer 2 for the following experiments.

III. EXPERIMENTS AND RESULTS

In this section, we conducted the experiments on three datasets from previous BCI competition. We compared the performance between CDBN and other classic feature extraction methods like CSP, band power, MVAAR. And we evaluated the features using the same classifier SVM with RBF kernel.

We have analyzed three open EEG datasets to conduct the experiments. Dataset 1 comes from dataset III in 2003 BCIC II; it contains a total of 280 groups of left and right hand Motor Imagery (MI) EEG data. Dataset 2 comes from dataset Iva in 2005 BCIC III; it includes the '*aa*', '*al*' and '*aw*' three subsets, and each subset contains a total of 280 groups of right hand and foot MI EEG data. Dataset 3 is from dataset III in 2005 BCIC III; it contains 360 groups of 4 classes (left hand, right hand, foot, tongue) MI EEG data and we omit 63 trials after checking artifacts.

Dataset 1 and dataset 2 each have 280 trails. We set different numbers of training samples, which are 80, 120, 160, 200, 240, and the remaining samples are test samples. Dataset 3 has 297 trails. And we set different numbers of training samples for dataset 3, which are 140, 160, 180, and also the remaining samples are test samples. For each number of training samples, 20 independent runs were conducted and each run corresponding training samples were selected randomly. The mean and standard deviation of the classification accuracy was recorded.

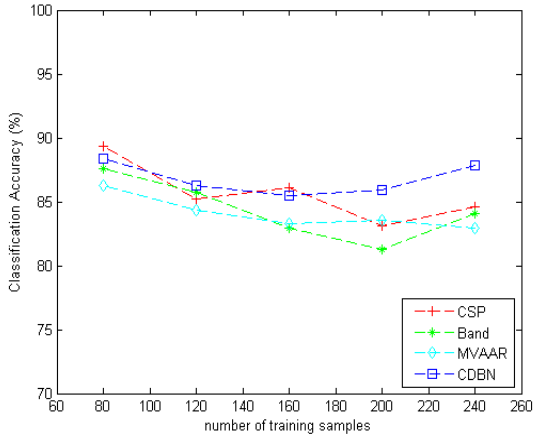
Table I shows the classification result for different feature representations on dataset 1, and the best results are shown in bold. For method CSP, we choose the most active frequency band, 9-11Hz, as our filter bank. And for method band power, we choose two active frequency bands 8-12Hz and 18-24Hz to calculate band power. As for method MVAAR, the order is set to 3. From Table I, we can clearly see that CDBN has better performance over others on the whole. Though the performance of CDBN is slightly inferior to CSP when the number of training samples is 80, it has the excellent performance on the whole especially when the number of training samples is 240.

Fig. 2 compares the classification result for different feature representations on dataset 2. Sub graph (a) for subject '*aa*', (b) for subject '*al*' and (c) for subject '*aw*'. For method CSP, we choose the most active frequency band (a) '*aa*' 13-15Hz; (b) '*al*' 12-13Hz; (c) '*aw*' 11-13Hz as our filter bank. And for method band power, we choose two active frequency band: (a) '*aa*' 13-15Hz and 25-27Hz; (b) '*al*' 12-13Hz and 24-25Hz; (c) '*aw*' 11-13Hz and 25-26Hz to calculate band power. As for method MVAAR, the order is set to3. Inspecting each sub graph, we see that the performance of CSP and CDBN is excellent on the whole. When there are few training samples (e.g., 80), CSP surpasses CDBN. However, when there are more training samples (e.g., 240), CDBN outperforms CSP clearly. This is because the unsupervised feature learning of CDBN relies heavily on the number of unlabeled data.

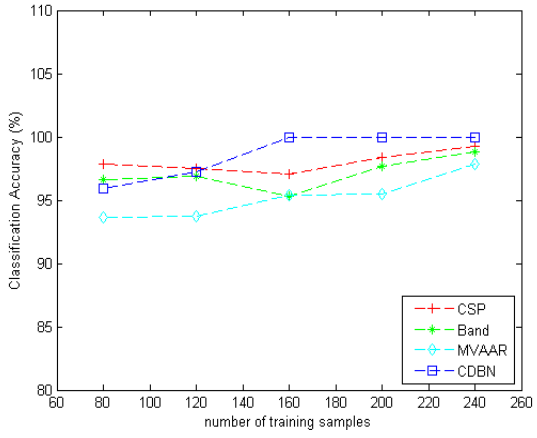
TABLE I. MEAN CLASSIFICATION ACCURACY FOR DIFFERENT FEATURES ON DATASET 1

Training Samples	Correct Rate (%)			
	CSP	***MVAAR***	***Band Power***	***CDBN***
80	**85.38±2.24**	80.35±4.07	81.90±2.75	83.63±1.82
120	85.25±1.94	84.88±3.87	83.59±2.90	**85.94±1.77**
160	85.74±2.33	85.46±2.54	85.00±2.55	**86.04±2.09**
200	85.56±3.10	84.81±4.06	84.63±4.24	**86.06±3.38**
240	85.75±6.18	85.88±5.80	86.13±6.36	**88.25±5.70**

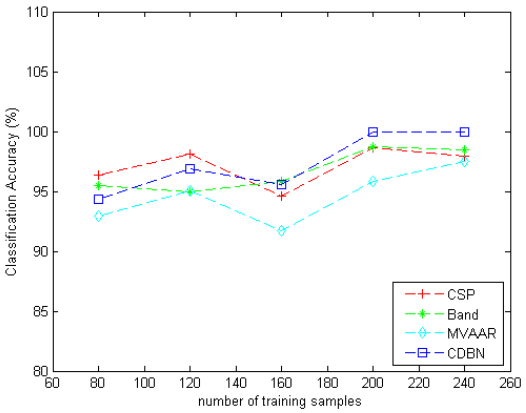
Considering the actual situation that EEG data does not restrict in 2 classes, we use dataset 3, a four-class problem to testify the algorithm. Table II shows the classification result for different feature representations on dataset 3; the best results are shown in bold. We choose the most active frequency band 11-12Hz for method CSP and band power. As for the order of method MVAAR is set to 3. From Table II, it is obvious that CDBN has the best performance over others. The most likely reason is that the more samples, the better features CDBN learns from unlabeled data.



(a) *'aa'*



(b) *'al'*



c) *'aw'*

Fig. 2. Mean classification accuracy for different features on dataset 2

TABLE II. MEAN CLASSIFICATION ACCURACY FOR DIFFERENT FEATURES ON DATASET 3

Training Samples	**Correct Rate (%)**			
	CSP	***MVAAR***	***Band Power***	***CDBN***
140	80.45±1.62	81.08±1.50	80.54±2.21	**82.02±1.88**
160	81.02±1.50	80.70±2.00	81.53±1.61	**82.41±1.44**
180	85.47±1.44	85.64±2.36	86.09±2.36	**87.33±1.74**

IV. CONCLUSION

In this paper, combining with the properties of EEG data, we apply CDBN to the feature learning of EEG data. We have conducted some experiments to compare the performance of different feature extraction methods on the datasets from previous BCIC. The results demonstrate that CDBN performs well in the feature learning of EEG signal especially when there is a large amount of unlabeled data.

Since now we only test the performance of CDBN on offline EEG data, in future we are interested in making CDBN self-adaptive so as to be adaptive to the online BCI systems.

REFERENCES

[1] J.R. Walpaw, N. Birbaumer, W.J. Heetderks, et al., "Brain-computer interface technology:a review of the first international meeting," IEEE Trans. Rehabil. Eng., vol. 8, no. 2, pp.164-173, 2000.

[2] N. Brodu, F. Lotte, A. Lecuyer, "Comparative study of band-power extraction techniques for motor imagery classification," 2011 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind and Brain (CCMB), pp.1-6.

[3] C.W. Anderson, E.A. Stolz, S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks", IEEE Trans. Biomed.Eng., vol. 45, no. 3, pp.277-286 ,1998.

[4] J. Müller-Gerking, G. Pfurtcheller, H. Flyvberg, "Designing optimal spatial filters for single-trial EEG classification in a movement task", Clinical Neurophysiology, vol.110, no.5, pp.787-789,1991.

[5] C.I. Hung, P.L. Lee, Y.T. Wu, et al., "Recognition of motor imagery electroencephalography using independent component analysis and machine classifiers", Ann. Biomed. Eng., vol.33, no.8, pp.1053-1070 , 2005.

[6] G. Schalk, B. Blankertz, S. Chiappa, et al., BCI competition III, 2004-2005, http://www.bbci.de/competition/iii/.

[7] G.E. Hinton, S. Osindero, Y.W. The, "A fast learning algorithm for deep belief nets", Neural Computing, Vol.18, no.7, pp.1527-1554, 2006.

[8] G.E. Hinton, R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks", Science, Vol.313, no.5786,pp.504-507,2006.

[9] H. Lee, A. Battle, R. Raina, and A.Y. Ng, "Efficient sparse coding algorithm", Advances in Neural Information Processing Systems (NIPS), pp.801-808, 2006.

[10] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng., "Unsupervised learning of hierarchical representations with convolutional deep belief network", Commun. ACM, vol.54, no.10, pp.95-103,2011.

[11] H. Lee, P.T. Pham, Y. Largman, A.Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks", Advances in Neural Information Processing Systems (NIPS), pp.1096-1104, 2009.

[12] D. Wulsin, J. Blanco, R. Mani, and B. Litt, "Semi-supervised anomaly detection for EEG waveforms using deep belief nets", 2010 Ninth International Conference on Machine Learning and Application (ICMLA), pp.436-441, 2010.

[13] H. Cecotti, A. Graeser, "Convolutional neural network wth embedded fourier transform for EEG classification", 19th International Conference on Pattern Recognition, pp.1-4, 2008.

[14] K. Sohn, D.Y. Jung, H. Lee, and A.O. Hero, "Efficient learning of sparse, distributed, convolutional feature representations for object recognition", In Proceedings of 13th International Conference on Computer Vision(ICCV), pp. 2643-2650, 2011.