Classification on ADHD with Deep Learning

Deping Kuang
Department of Computer Science and Technology
Tongji University
Shanghai, China
Kuangdp1990@163.com

Lianghua He*
Department of Computer Science and Technology
Tongji University
Shanghai, China
helianghua@tongji.edu.cn

* Corresponding author

***Abstract*—Effective discrimination of attention deficit hyperactivity disorder (ADHD) using imaging and functional biomarkers would have fundamental influence on public health. In usual, the discrimination is based on the standards of American Psychiatric Association. In this paper, we modified one of the deep learning method on structure and parameters according to the properties of ADHD data, to discriminate ADHD on the unique public dataset of ADHD-200. We predicted the subjects as control, combined, inattentive or hyperactive through their frequency features. The results achieved improvement greatly compared to the performance released by the competition. Besides, the imbalance in datasets of deep learning model influenced the results of classification. As far as we know, it is the first time that the deep learning method has been used for the discrimination of ADHD with fMRI data.** *(Abstract)*

***Index Terms*—ADHD, fMRI, Deep Learning, Deep Belief Network.** *(key words)*

I. Introduction

Attention deficit hyperactivity disorder is also known as ADHD, which is one of the most common mental disorder among children. The primary symptoms of them are distractibility, poor concentration, excessive activity or weak self-control [1]. According to the biased towards of the symptoms, the American Psychiatric Association divided ADHD into three subtypes, which were mixed ADHD, ADHD inattentive, ADHD hyperactive. The mixed ADHD has the symptoms of both inattentive and hyperactive. The cause and pathogenesis of ADHD yet to verify. So far, the diagnosis of ADHD is defective. The sensitivity of the diagnosis of American Psychiatric Association's Diagnostic and Statistical Manual on ADHD achieves 70 percent to 90 percent. The development of more quantified and objective analysis for classifying and diagnosing the ADHD is an important goal of the neuroscience research.

Along with the development of magnetic resonance, neuroimaging is usually used for the research of mental disorder. Magnetic resonance imaging is also called MRI, which is a neuroimaging technology with defining activity in the healthy and diseased human brain based on blood oxygenation level dependent (BOLD) [2]. For the non-invasive and non-radiative property of the fMRI, it rapidly became the mainstream of methods for the study of human brain. Recently, ADHD has exploited neuroimaging for research. The structural magnetic resonance imaging research shows that there exists abnormality between ADHD and the healthy on the frontal and parietal lobe, occipital lobe and thalamus [3]. Zhu et al. [4] proposed a PC-FDA classifier using features of regional homogeneity (ReHo) based on resting-state fMRI to discriminate 20 subjects as ADHD or healthy. Wolf et al. [5] used independent component analysis in a working memory task. The participants of the task are 12 ADHD and 12 healthy. Methods including seed-based approaches, independent component analysis, graph classifiers are used to analyze resting state data. With the belief that a community-wide effort focused on advancing functional and structural imaging examinations of the developing brain will accelerate the rate at which neuroscience can inform clinical practice, the 1000 Functional Connectomes Project provided a model which includes large-scale datasets [6], which overcomes the defects of fewer subjects. Eloyan et al. [7] used decomposition of CUR and gradient boosting with motor network segmentation and random forest for prediction based on rs-fc-fMRI. They achieved the best score among the participants of the global ADHD-200 competition. Despite the score is highest among the participants, the performance of the prediction is yet to improve to reach the perfect. But the competition shows that using functional magnetic resonance images can classify ADHD from the healthy.

In recent years, an increasing number of attention has been aroused on the application of pattern recognition and machine learning techniques in brain image analysis. Classification feature and learning algorithm are two important aspects of a classification system. Shadow models such as support vector machine (SVM) or neural network are usually used for the prediction. There is a consistent view that a simple machine learning model is more effective than complex model under the background of big data. Since 2006 [8], along with the presentation of the greedy training of deep learning model by restricted Boltzmann machine (RBM), deep structured learning, or more commonly called hierarchical learning, has emerged as a new area of machine learning research. Deep learning method has been applied for many areas including image processing [8,9], audio classification [10], object recognition [11,12], natural language processing [13] and so on. All of

978-1-4799-6621-9/14 $31.00 © 2014 IEEE
DOI 10.1109/CCBD.2014.42

them achieved good performance. Deep learning method simulates the mechanisms of visual information system of human brain to process low-level features to high-level abstract features. It prompts us to rethink the issue that complex model may exploit the wealth of hidden information in massive data. Since the recovery of the deep learning method, research mainly focused on the problem of classification on MNIST handwriting and the performance improved remarkably. This paper will introduce deep learning method on analyzing massive fMRI data considering its powerful learning ability and advantage to dig out the cognitive significance of brain.

In this paper, the frequency feature on the voxels of the brain are used as a vector of feature for the raw input for the deep learning. Deep learning is applied for the feature extraction and classification. Results show that the method proposed in this paper got better performance than the competition. The sections below are our learning method and experiments. Section 2 mainly describes the DBN method and application of DBN for ADHD data. Section 3 presents our experiments and results on ADHD dataset.

II. METHOD

Common deep learning network architectures include deep belief network (DBN), convolutional neural network (CNN) and its variants and so on. In this paper, DBN is applied for classification. Besides, in order to overcome the imbalance of the training datasets, a trick of copying by self is used.

A. Deep Belief Network

Deep belief network (DBN) is composed of a stack of restricted Boltzmann machines (RBM), see fig.1. The raw input can be seen as the visible units of the first RBM and by training the RBM, the hidden units can be derived. The results connect to the upper hidden units to build a new RBM. As the same, the DBN is built by stacking multiple RBMs. The pre-training of DBN is to pretrain the weight and biases of each RBM.

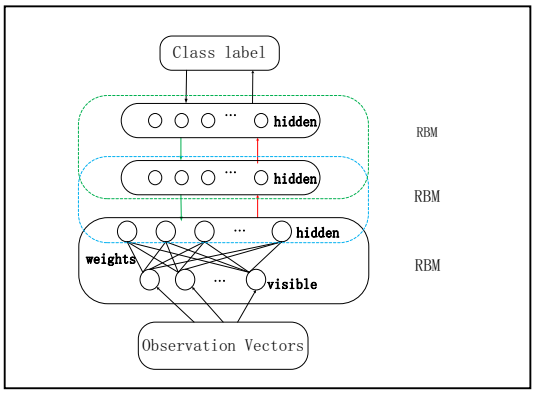

Fig. 1. Architecture of DBN

RBM (see fig.2) is a two-layer architecure which is visible units and hidden units respectively. The visible and hidden units connect to each other but no connections are built within the visible units or hidden units. While using RBM, there are some important description which are the weights W_{ij} between visible units v_i and hidden units h_j, the biases of visible units b_i and the biases of hidden units c_j.

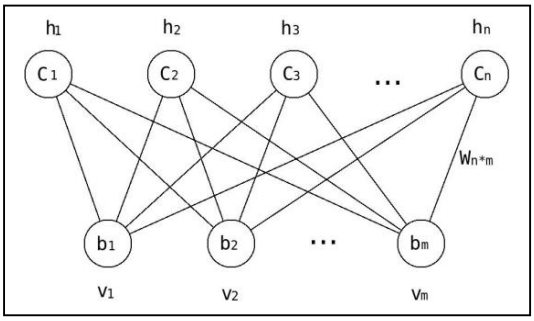

Fig. 2. Structure of RBM

Given this parameters, the energy of a pair of visible and hidden unit (v,h) is defined as

$$E(v,h) = -\sum_{i=1}^{V}\sum_{j=1}^{H} v_i h_j w_{ij} - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} c_j h_j$$

In general Boltzmann machines, the probability distributions over hidden and visible units are defined in terms of the energy function:

$$p(v,h) = \frac{1}{z} e^{-E(v,h)}$$

Where Z is defined as the sum of $e^{-E(v,h)}$. Based on the joint probability of v and h, the conditional probability of v given h and of h given v is easily obtained. So the individual activation probabilities are given by

$$p(h_j = 1|v) = \sigma\left(\sum_{i=1}^{V} wi_j v_i + c_j\right)$$

$$p(v_i = 1|h) = \sigma\left(\sum_{j=1}^{H} wi_j h_j + b_i\right)$$

Where $\sigma(t) = (1 + e^{-t})^{-1}$.

A RBM is pre-trained to maximize the log-likelihood logP(v). Introducing the Contrastive Divergence (CD) [14] approximation to the gradient, the update rule of W is given by

$$\Delta w_{ij} = \varepsilon\left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}\right)$$

Where ε is the learning rate and the angle brackets manifests the expectations relative to the distribution specified in the subscript. The updating rule makes it the reconstruction of hidden units equals to the data. Only by this way, the hidden unit is approximate to the visible units. Then they can be seen as the exact expression of the data.

B. Imbalanced Class Distrubution

Imbalanced class distribution of a data set has encountered a serious difficulty to most learning algorithms. The imbalanced data is characterized as having many more instances of certain classes than others. As data with more samples occur frequently, classification rules that predict the prevalent classes tend to be usual and frequent; consequently, test samples belonging to the small classes are misclassified more often than those belonging to the prevalent classes. Imbalance of data for training in the DBN model is a problem that may have influence in the prediction. Against to the imbalance in SVM, cost is applied. But how to choose the cost is a difficult task. In addition to the cost, there are two methods to deal with the problem, one is to repeat the less data to the popularity of the large datasets and the other is to select from the large dataset the equal number of samples in the little dataset. In this paper, the previous method copying the data by self is used for overcome the data imbalance of DBN.

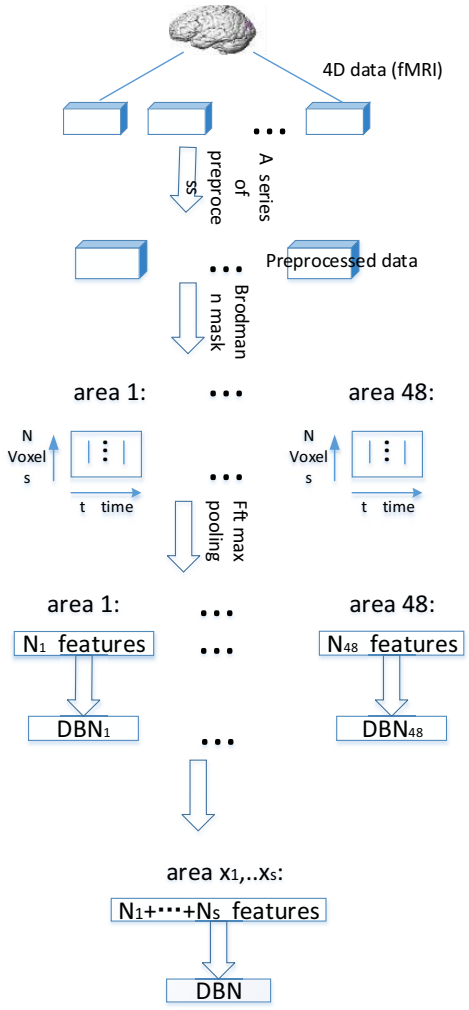

Fig. 3. Application of DBN on ADHD

C. Application of DBN for ADHD Dataset

DBN is applied to predict the ADHD subtypes in this paper. But before appling DBN, some preprocessing are needed for the specific of the fMRI data.

First, in order to analysis data effectively, a series of preprocessing are conducted in spm8 [15] such as realign, slice time, co-register, normalize, smooth which are described in [16] in detail.

In addition, fMRI data used in this paper is 4D data which is a time series of 3D images. To reduce the dimension, some strategies are conducted. First, according to brodmann template to divide the 3D images to 48 areas, brodmann template is a region of human cerebral cortex defined on its cytoaerchitectonics, or structure and organization of cells. And it is always thought that data in frequency domain may indicate more information. Then, the fast Fourier transform algorithm (FFT) is used to transform data from time domain to frequency omain. At last, execute max-pooling of frequencies in each voxel to select the frequency which has the maximum value of amplitude due to the assumption that the highest frequency of voxels in some areas may be different during the scanning procedure and the frequency is higher when the voxel is more active. After FFT and max-pooling every voxels have only one property and the number of properties in each area depends on the number of voxels.

It can be seen from fig.3 that the data after preprocessing can be invested into the DBN architecture with three hidden layers for training the model. In the program, repeat the training and testing for classification for 20 times. In each procedure, the trainning data is used for pretraining the DBN model by greedy training RBM of every two neighboring layers and is used for backproping to tune the model. The experiments are conducted both in every areas of brodmann and combining of some interesting areas. The pseudo-code for application of DBN on ADHD data can be seen as below.

Procedure 1: Training DBN for ADHD

```
// give the max epoch for training rbm.
maxepoch =100;
// give the number of hidden units
s_numhid = [500,200,50];
// give the number of target classes
targetsnum = 4;
// training DBN for 20 times.
for cv = 1:20
    // make batches of training data
    maketraindata;
    // for every rbm to train the weights
    for k=1:3
        // train rbm in greedy way
        [vishid,visbiases,hidbiases,batchposhidprobs] = rbm(batchdata);
        // the results of previous hidden units for the visible units of next visible units
        batchdata = batchposhidprobs;
    end
    // back adjust the whole weights of DBN
    backprop;
    // test the DBN architecture on test dataset
    testDBN;
end
```

Fig. 4. Pseudo-code of DBN on ADHD

III. Experiments and Results

A. Data

The data used in this paper can be downloaded from the ADHD-200 Global Competition website. DBN model is built upon the ADHD dataset for NYU, Neuro and OHSU respectively.

For NYU, the training subjects are 222, and test subjects are 41; for Neuro dataset, the training subjects are 48, and test subjects are 25; and for OHSU the training subjects and test subjects are 79 and 34 respectively. In this paper, DBN model is pretrained and tuned according to the training dataset, then tested it on test dataset.

B. Results

1) The Influence of Dataset in DBN

It is assumed that the imbalance of training data batches may result in the favoritism of the DBN model. The training procedure of DBN is composed of many batches, the batch may include different size of data of each class. In this way, the model may prefer the class with more samples. To verify the assumption, some experiments are conducted on the NYU dataset by keeping the size of one class fixed and change the size of another class. The result is shown in the figure below which keeps the size of normal subjects 95 and improve the size of ADHD subjects from 45 to 115. From the figure, the error rate of ADHD increased while the rate of the normal decreased.

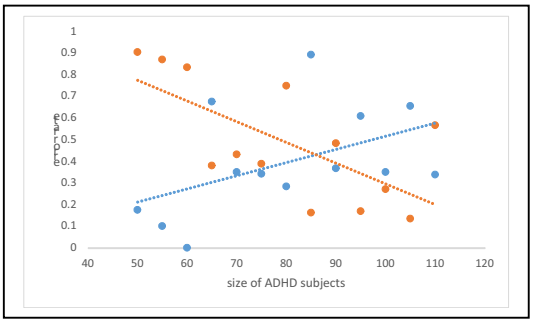

Fig. 5. Example of a figure caption. *(figure caption)*

2) Comparision of Balanced and Imbalanced Dataset in DBN

To make it balance, copy the data which is less to the size of prevalent dataset. The experiment of overcoming the imbalance of training dataset is tested on the NYU dataset. The experiment is conducted 50 times by making every batch containing the same number of samples for every class. The prediction accuracy of balanced data batch ranges from 31.71% to 53.66%. The average prediction is 42.20% which improved 4.79 percents compared to the imbalanced DBN. The comparision of released performance, imbalanced batches and balanced batches can be seen from the Figure below.

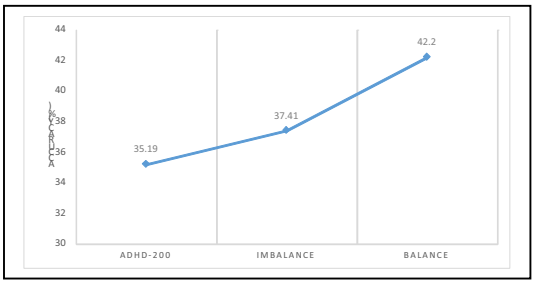

Fig. 6. Comparision of Performance

3) Accuracies of Three Functional Areas on NYU

The average accuracy for four classes including control, mixed ADHD, ADHD inattentive and ADHD hyperactive on NYU dataset is shown in fig.7. The results consists of three functional areas mentioned above. It can be seen that the accuracy of the method proposed above is better than the discrimination accuracy achieved by the team in the ADHD-200 competition, which is 35.19%, for NYU in prefrontal cortex and cingulate cortex but in visual cortex the accuracy is less than the competition.

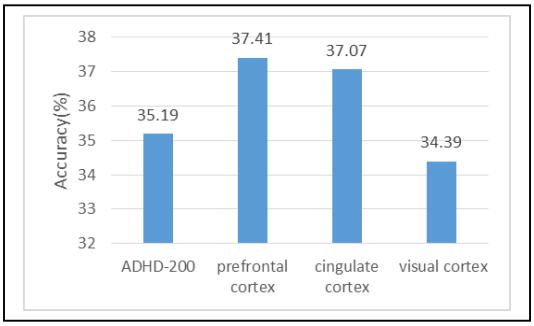

Fig. 7. Accuracy of DBN in Three Functional Areas

4) Accuracies of Neuro, OHSU and Pittsburgh on area of prefrontal cortex

The experiments are also excuted on the Neuro, OHSU and Pittsburgh dataset. The results released by ADHD-200 competition are 56.95% for Neuro, 65.37% for OHSU and 40.74% for Pittsburgh respectively. Performing DBN on prefontal cortex, visual cortex and cingulate cortex areas for the three datasets. The prediction accuracies of them are shown in fig.7. For Neuro the accuracies are 44.4%, for OHSU are 80.88% and for Pittsburgh the accuracies are 55.56% respectively. From the figure, it is obvious that the performance are better than the released result on average. It is believed that the data on prefrontal cortex can be used for classification of ADHD.

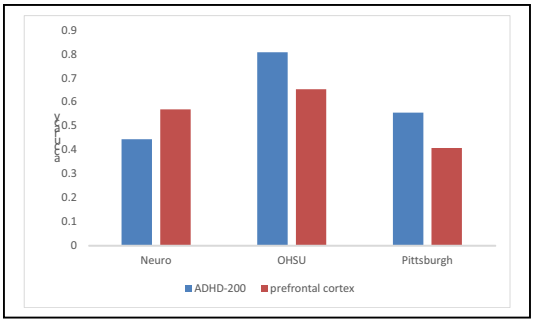

Fig. 8. Accuracies of Neuro, HSU and Pittsburgh

5) Prediction of Using whole Brain Data

In order to make use of the data from the whole brain to predict ADHD subtypes and control. Frequency features are combimed together by conducting pca variants kpca-st. Frequency features are produced not only by FFT but also by wavelet transform. The size of input for the input of DBN is shown in the table below, there are three feature extract methods used in the experiments and the size of vector after the extraction is different from each other. The vector m*n stands for the subjects*feature spots. Sym5 and Coif3 are two kinds of wavelet transform method.

TABLE I. SIZE OF FEATURES

Extraction Method	FFT	Sym5	Coif3
size	257*9275	257*9178	257*9177

Unlike the way above that training and test data given by the competition, these experiments are conducted by mixing the subjects all together and select the 41 as test data set randomly and the rest data are used as train data. The prediction accuracies of conducting the deep belief network as feature learning and classification are show in the fig.8.

The accuracies shown in the figure indicates that the data of the whole brain as the feature for the DBN has better performance than parts of brain. The highest score comes from the coif3 which is 44.63% and it is 2.43 percents higher than the balanced data set.

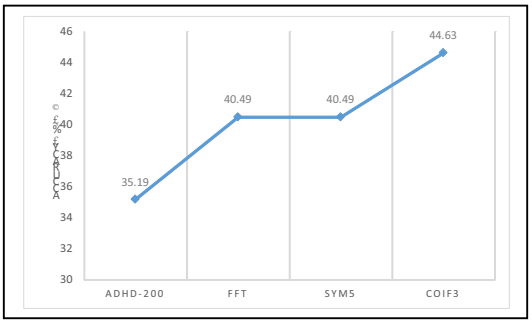

Fig. 9. Accuracy of DBN by Combining Whole Brain

IV. CONCLUSION

In this paper, one of the deep learning models-deep belief network-is applied for extracting feature and classification. The proposed method was proved to be effective in discriminating ADHD from control and subtypes. The accuracy of prediction has improved in some degree compared with the results published in ADHD-200 competition. In addition, the performance improved after overcoming the drawback of imbalance in dataset as well as using the whole brain data as feature.

In the future, we will verify the effect of imbalance in DBN in detail. What's more, it is considering to take the personal characteristic data as well as fMRI-based information into analysis as described in [19]. By considering the above factors, there is a reason to believe that the discrimination performance based on the proposed method can be more effective.

REFERENCES

[1] Kooij S J J, Bejerot S, Blackwell A, et al. European consensus statement on diagnosis and treatment of adult ADHD: The European Network Adult ADHD[J]. BMC psychiatry, 2010, 10(1): 67.

[2] Huettel S A, Song A W, McCarthy G. Functional magnetic resonance imaging[M]. Sunderland, MA: Sinauer Associates, 2004.

[3] Seidman L J, Valera E M and Makris N et al. Dorsolateral prefrontal and anterior cingulate cortex volumetric abnormalities in adults with attention-deficit/hyperactivity disorder identified by magnetic resonance imaging[J]. Biol. Psychiatry , 2006, 60: 1071–1080.

[4] Zhu C Z, Zang Y F, Cao Q J, et al. Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder [J]. Neuroimage, 2008, 40(1): 110-120.

[5] Wolf R C, Plichta M M, Sambataro F, et al. Regional brain activation changes and abnormal functional connectivity of the ventrolateral prefrontal cortex during working memory processing in adults with attention‐deficit/hyperactivity disorder[J]. Human brain mapping, 2009, 30(7): 2252-2266.

[6] Milham M P. Open neuroscience solutions for the connectome-wide association era[J]. Neuron, 2012, 73(2): 214-218.

[7] Eloyan A, Muschelli J, Nebel M B, et al. Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging [J]. 2012.

[8] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

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

[10] Lee H, Pham P T, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]//NIPS. 2009, 9: 1096-1104.

[11] Lee H, Grosse R, Ranganath R, et al. Unsupervised learning of hierarchical representations with convolutional deep belief networks[J]. Communications of the ACM, 2011, 54(10): 95-103.

[12] Ranzato M, Huang F J, Boureau Y L, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition[C]//Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-8.

[13] Sarikaya R, Hinton G E, Deoras A. Application of Deep Belief Networks for Natural Language Understanding[J].

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

[15] Statistical parametric mapping, version 8. http://www.fil.ion.ucl.ac.uk/spm/, 2009.

[16] Huettel S A, Song A W, McCarthy G. Functional magnetic resonance imaging[M]. Sunderland, MA: Sinauer Associates, 2004.

[17] Bush G, Frazier J A, Rauch S L, et al. Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fMRI and the Counting Stroop[J]. Biological psychiatry, 1999, 45(12): 1542-1552.

[18] Bush G, Frazier J A, Rauch S L, et al. Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fMRI and the Counting Stroop[J]. Biological psychiatry, 1999, 45(12): 1542-1552.

[19] ADHD-200 Consortium. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience[J]. Frontiers in systems neuroscience, 2012, 6.