Neural Networks 114 (2019) 119–135

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Early Alzheimer's disease diagnosis based on EEG spectral images using deep learning

Bi Xiaojun [a], Wang Haibo [b,*]

[a] School of Information Engineering, Minzu University of China, Beijing, China
[b] College of Information And Communication Engineering, Harbin Engineering University, Harbin, China

ARTICLE INFO

Article history:
Received 4 August 2018
Received in revised form 4 January 2019
Accepted 14 February 2019
Available online 11 March 2019

Keywords:
Early diagnosis of AD
Multi-task learning
Deep learning
Deep Boltzmann Machine

ABSTRACT

Early diagnosis of Alzheimer's disease (AD) is a proceeding hot issue along with a sharp upward trend in the incidence rate. Recently, early diagnosis of AD employing Electroencephalogram (EEG) as a specific hallmark has been an increasingly significant hot topic area. In consideration of the limited size of available EEG spectral images, how to extract more abstract features for better generalization still remains tremendously troubling. In this paper, we demonstrate that it can be settled well with multi-task learning strategy based on discriminative convolutional high-order Boltzmann Machine with hybrid feature maps. First, differently from our original model — Contractive Slab and Spike Convolutional Deep Boltzmann Machine (CssCDBM), we directly conduct EEG spectral image classification via inducing label layer, resulting in a discriminative version of CssCDBM, referred to as DCssCDBM. This demonstrates DCssCDBM can be extended well into the classification model instead of feature extractor alone previously. Then, the most important approach innovation is that we train our DCssCDBM with multi-task learning framework via EEG spectral images based Identification and verification tasks for overfitting reduction for the first time, which could increase the inter-subject variations and reduce the intra-subject variations respectively, both of which are essential to early diagnosis of AD. The proposed method shows the better ability of high-level representations extraction and demonstrates the advanced results over several state-of-the-art methods.

1. Introduction

Along with tendency to older population, AD is the most prevalent dementia, acting as the decline of brain function such as reasoning, thinking, memory, and so on, each of which is enough to interfere with the activities of daily living of the elderly, extensively increasing every year in the proportion of deaths (Zhou Li & Shen, 2016). Generally, it is too late to confirm since the cause of AD may be not clear. Even if timely treatment will not have much effect, early diagnosis of AD is a better approach to avoid illness or rapid deterioration at least. Mild Cognitive Impairment (MCI) (Petersen et al., 1999) may be a mediate state between AD and Healthy Control (HC). Studies have demonstrated that conversion rate from MCI to AD is about 10%–15% every year (Petersen et al., 1999), and patients with MCI are more likely to develop into AD than those who have not been confirmed MCI. MCI symptoms, which just behave memory degradation, are much lighter than AD. MCI is so hard to be found in the early stage that it is even developed into AD. Consequently, more attention should be paid to the differential diagnosis of MCI and AD. Now, the tools for early diagnosis of AD are generally brain imaging equipment, such as functional magnetic resonance imaging (FMRI) and computed tomography (CT), which need exhaustive testing sessions and experienced clinicians. In addition, it is expensive to employ such neuroimaging tools. Recently, early diagnosis of AD based on EEG signal has arisen extensive interest in the field on account of growing evidence that oscillatory electromagnetic brain activity might have a deep relation with the AD (Capecci et al., 2016; Morabito et al., 2016; O'Keeffe et al., 2017) and overcoming these above limitations. Effective solution to an early diagnosis system of AD based on EEG will allow it to be an important part of the Home Care System.

In view of moderate size of available EEG data, how to robustly learn abstract representations from EEG for better generalization still remains challenging (Omedes, Iturrate, Montesano, & Minguez, 2013; Song & Zhang, 2016). Recently, most feature extraction methods for EEG are based on handcrafted engineering (Atyabi, Luerssen, Fitzgibbon, & Powers, 2012; Temko, Nadeu, Marnane, Boylan, & Lightbody, 2011), which may be inappropriate to this complex task. Additionally, most relevant studies (Li, Guan, Zhang, & Ang, 2017; Mehmood, Du, & Lee, 2017) based

* Corresponding author.
E-mail addresses: bixiaojun@hrbeu.edu.cn (X. Bi), wanghaibo@hrbeu.edu.cn (H. Wang).

https://doi.org/10.1016/j.neunet.2019.02.005
0893-6080/

on the machine learning model may ignore the fact that extracted features should be invariant to both inter-subject and intra-subject differences. Indeed, inter- and intra-subject EEG variations are complex and highly nonlinear in high-dimension space. Better representations in favor of classification should reduce intra-class variations and enlarge inter-class variations. We should take these findings into considerations for better generalization on early diagnosis of AD. For these reasons, it is dramatically urgent to address the above concerns. In this paper, we propose a multi-task learning strategy, taking inter- and intra-subject EEG variations into consideration based on a novel deep classification model in order to solve above problems.

First, we present a classification version of a novel deep model, which is a further extension of our latest work (Bi & Wang, 2018), CssCDBM as a special deep learning model, in order to combine feature extraction and classification. Generally, Deep Boltzmann Machine (DBM) and its variants have shown to be potentially powerful alternative tools for feature extraction and have shown their validity on various image tasks in the literature (Chen, Yu, Hu, & Zeng, 2013; Leng, Zhang, Yao, & Xiong, 2015; Salakhutdinov & Hinton, 2012). However, DBM and its variants, which are simply the first move of another learning algorithm, still serve as feature extractor followed by classification models such as SVMs or a good initialization during training deep neural network classifiers. DBM and its variants cannot guarantee that the learned representations via training in an unsupervised fashion will eventually be conducive for the final supervised task. More importantly, model selection can also become extensively difficult, since we should balance jointly the space of hyper-parameters of both the DBM and the supervised learning algorithm. Based on the above-mentioned issue, we develop CssCDBM via inducing label layer for eliminating the gap between feature learning and classification. The main task of this part is to design a block, named as discriminative contractive slab and spike convolutional restricted Boltzmann Machine (DCssCRBM).

Then, we propose multi-task learning framework based on DCssCDBM model for early diagnosis of AD. On this multi-task learning framework, main task — early diagnosis of AD task is to categorize an EEG spectral image into a one correct class of three classes (HC, MCI, AD), while verification based on EEG spectral image, one of auxiliary tasks, is to determine a pair of EEG spectral images whether they belong to the same identity or not (i.e. Binary classification) and identification based on EEG, another auxiliary task, is used to classify an EEG spectral image into one special person. Identification and verification will help our deep model avoid overfitting via being invariant to inter-subject differences and intra-subject differences (Mei, Liu, Karimi, & Gao, 2014). To our best current knowledge, it is the first attempt to combine early diagnosis of AD, identification and verification together to improve early diagnosis of AD accuracy. In our experiments, we performed multi-task learning framework based on DCssCDBM for early diagnosis of AD and showed the advancement of the proposed methodology in comparison with the most competitive methods.

The rest of the paper is arranged as follows. Related work is concentrated on in Section 2. Section 3 details the proposed methodology; Experimental performances are exhibited in Section 4; Section 5 presents discussion; Section 6 concludes this paper.

2. Related work

Early diagnosis of AD is generally defined as a classification problem that involves confirming correct category of a given EEG spectral image. Conventional approaches to early diagnosis of AD involve extracting features based on handcrafted engineering from original EEG waves (Morabito et al., 2013), optionally followed by classifier. These methods may not settle well with such a complex task. Recently, deep learning has attracted huge interests from researchers due to impressive performance on various recognition tasks. In this paper, we use an advanced deep generative model with a new learning framework. We mainly describe related work about the proposed methods for easily understanding our motivations. To our best current knowledge, this is the first try to employ discriminative deep probabilistic model with multi-task learning framework based on EEG spectral image instead of FMRI, CT or multi-channel EEG waves for early diagnosis of AD.

As mentioned earlier, DBMs and its variants have been extensively utilized in the past to provide good initialization and extract useful features in an unsupervised fashion for another discriminative learning algorithm. Nevertheless, in all these cases, the learned features were not learned by discriminative model and had to be given to another supervised model which eventually conducted classification tasks. This approach appears obvious drawback that deep model may extract features inappropriate for classification tasks. The Most related work had appeared in discriminative Restricted Boltzmann Machine (DisRBM), which can be considered as an approach of addressing this issue (Larochelle, 2008). Inspired this work, we attempt to solve this issue in order to develop CssCDBM for a wider application. However, it is much more difficult than DisRBM due to the depth and high-order characteristic of the model.

Multi-task learning has recently been employed to address the risk of overfitting of deep models, where we train a model to jointly achieve the major problem of interest combining with several related tasks. Multi-task learning can well guide the model to extract better internal features. It has been successfully demonstrated to improve the performance of networks for tasks such as face alignment (Zhang, Luo, Loy, & Tang, 2014), face keypoint recognition (AbdAlmageed et al., 2016), and age estimation (Zhang & Yeung, 2010). There are several works about early diagnosis of AD with multi-task learning as well. These studies include a multi-task regression problem by treating each time point prediction as a task based on FMRI (Zhou, Liu, Narayan, Ye, & Initi, 2013), a constrained multi-modality with multi-task feature selection (Liu, Wee, Chen, & Shen, 2014; Zhang, Shen, & Neuroimaging, 2012). These works have limitations on access to such neuroimaging tools, in addition to blindly auxiliary tasks. Actually, early diagnosis of AD implicitly includes other tasks, identification and verification based on EEG spectral images. Identification and verification based on EEG spectral images are same as identification and verification based on face images. We can use these two auxiliary tasks to constraint huge number of trainable parameters of our proposed deep model in order to generalize well to unseen EEG spectral images.

3. Proposed method

3.1. DCssCDBM Descriptions

The DCssCDBM can be viewed as an extension of our latest work (Bi & Wang, 2018). We provide details of the DCssCDBM in this part. First, we define the energy function of two layers in Eq. (1).

$$
\begin{aligned}
&E\left(v, s, h, \rho, l, \boldsymbol{y}\right) \\
&= \sum_{l=1}^{L}\sum_{m=1}^{N_{vh}}\sum_{n=1}^{N_{vw}} \frac{1}{2\sigma^2}\left(v_{mn}^{l} - c_l\right)^2 - \sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}} \left(v^{l} \Theta w^{l,k}\right)_{ij} h_{ij}^{k} s_{ij}^{k} \\
&- \sum_{k=1}^{K} b^{k} \sum_{i=1,j=1}^{N_{hh},N_{hw}} h_{ij}^{k} + \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}} s_{ij}^{k} \alpha_{ij}^{k} s_{ij}^{k}
\end{aligned}
$$

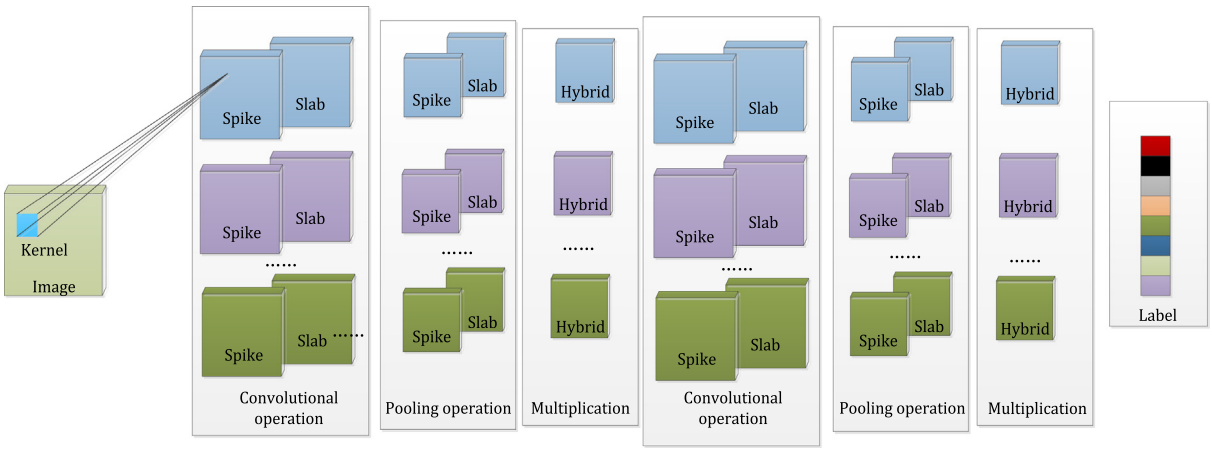


Fig. 1. Overall structure of the DCssCDBM with two-hidden layers.

$$
\begin{aligned}
&-\sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{i'j'}((ps^k\cdot ph^k)\Theta w^{k,g})_{i'j'}\ell^g_{i'j'}\rho^g_{i'j'}\\
&-\sum_{g=1}^{G}b^g\sum_{i'j'}\ell^g_{i'j'}+\frac{1}{2}\sum_{g=1}^{G}\sum_{i'j'}\rho^g_{i'j'}a^g_{i'j'}\rho^g_{i'j'}\\
&-\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}}\sum_{\boldsymbol{g},\boldsymbol{i'}\boldsymbol{j'}}\boldsymbol{y_t u}_{\boldsymbol{t},\boldsymbol{i'j'}}^{\boldsymbol{g}}\boldsymbol{\ell}^{\boldsymbol{g}}_{\boldsymbol{i'j'}}\boldsymbol{\rho}^{\boldsymbol{g}}_{\boldsymbol{i'j'}}-\sum_{\boldsymbol{t=1}}^{\boldsymbol{T}}\boldsymbol{y_t d_t}\\
&\textit{subject to}:\ \sum_{(i,j)\in B_a}h^k_{ij}\le 1,\forall k,a;\ \sum_{(i',j')\in B_a}\ell^g_{i'j'}\le 1,\forall g,a;\ \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}}\boldsymbol{y_t}=\boldsymbol{1}
\end{aligned}
\tag{1}
$$

where **notations in black bold are newly added into the original energy function**. v^l is the lth channel observation. The h^k_{ij}/s^k_{ij} and $l^g_{i'j'}/\ \rho^g_{i'j'}$ represent a spike/slab hidden unit at coordinate (i, j) of the kth feature map in the bottom layer and the coordinate (i', j') of the gth feature map in the top layer respectively. a^k_{ij} and $a^g_{i'j'}$ are employed to respectively penalize large value on slab unit s^k_{ij} and $\rho^g_{i'j'}$. L denotes the number of the observation channels. K and G are the number of convolution kernels in the bottom and top layer, respectively. $w^{l,k}/w^{k,g}$ and b^k/b^g denote the weights of the kth/gth kernels in the lth/kth channel and the bias of spike units in the kth/gth feature map on the bottom/top layer, respectively. c_l is the lth channel bias of visible units. We use $y = [y_1, \ldots, y_T]$ in order to denote labels and T represents the number of categories. The symbol $u_{t,i'j'}^{\ g}$ is weight between the t-th softmax unit and the hybrid feature layer unit with coordinate (i', j') of the gth feature map. The denotation d_t implies the bias of the tth label unit. The symbol Θstands for the convolution operation.

Here, the most of inferences in the DCssCDBM are similar to the CssCDBM. For DCssCDBM defined in Eq. (1), its probabilistic complete inferences are given in Eqs. (2)∼(10) which are given in Box I, where we denote $\otimes$ as Kronecker product. The *ones*(C, C) is a matrix with $C\times C$ scale setting all elements to 1. (a, b) denotes a Gaussian distribution with a mean value and b variance value. We also use ***Dual variable probabilistic max-pooling***. Fig. 1 shows a two-layer DCssCDBM schematic diagram. We can discover that some layer inferences in DCssCDBM combine information from the top-down layer, bottom-up layer and label information as Eqs. (6)∼(9) describe. Eqs. (11)∼(19) describe all partial derivatives of the likelihood function with respect to trainable parameters θ.

$$\nabla \mathrm{w}^{l,k}=E_{pdata}\left[v^l\Theta\left(h^k\cdot s^k\right)\right]-E_{pmodel}\left[v^l\Theta\left(h^k\cdot s^k\right)\right] \tag{11}$$

$$
\begin{aligned}
\nabla \mathrm{w}^{k,g}=&E_{pdata}\left[v^l(ps^k\cdot ph^k)\Theta\left(\rho^g\cdot\ell^g\right)\right]\\
&-E_{pmodel}\left[v^l(ps^k\cdot ph^k)\Theta\left(\rho^g\cdot\ell^g\right)\right]
\end{aligned}
\tag{12}
$$

$$\nabla b^k=E_{pdata}[\sum_{i,j=1}^{N_{hh},N_{hw}}h^k_{ij}]-E_{pmodel}[\sum_{i,j=1}^{N_{hh},N_{hw}}h^k_{ij}] \tag{13}$$

$$\nabla b^g=E_{pdata}[\sum_{i',j'=1}^{N_{\ell h},N_{\ell w}}\ell^g_{i'j'}]-E_{pmodel}[\sum_{i',j'=1}^{N_{\ell h},N_{\ell w}}\ell^g_{i'j'}] \tag{14}$$

$$\nabla a^k_{ij}=E_{pdata}[\frac{1}{2}s^k_{ij}s^k_{ij}]-E_{pmodel}[\frac{1}{2}s^k_{ij}s^k_{ij}] \tag{15}$$

$$\nabla a^g_{i'j'}=E_{pdata}[\frac{1}{2}\rho^g_{i'j'}\rho^g_{i'j'}]-E_{pmodel}[\frac{1}{2}\rho^g_{i'j'}\rho^g_{i'j'}] \tag{16}$$

$$\nabla c_l=E_{pdata}[\sum_{m,n=1}^{N_{vh},N_{vw}}\frac{1}{\sigma^2}(c_l-v^l_{mn})]-E_{pmodel}[\sum_{m,n=1}^{N_{vh},N_{vw}}\frac{1}{\sigma^2}(c_l-v^l_{mn})] \tag{17}$$

$$\nabla \boldsymbol{u}_{\boldsymbol{t},\boldsymbol{i'j'}}^{\ \boldsymbol{g}}=\boldsymbol{E_{pdata}}[\boldsymbol{y_t p\rho}^{\boldsymbol{g}}_{\boldsymbol{i'j'}}\boldsymbol{p\ell}^{\boldsymbol{g}}_{\boldsymbol{i'j'}}]-\boldsymbol{E_{pmodel}}[\boldsymbol{y_t p\rho}^{\boldsymbol{g}}_{\boldsymbol{i'j'}}\boldsymbol{p\ell}^{\boldsymbol{g}}_{\boldsymbol{i'j'}}] \tag{18}$$

$$\nabla \boldsymbol{d_t}=\boldsymbol{E_{pdata}}[\boldsymbol{y_t}]-\boldsymbol{E_{pmodel}}[\boldsymbol{y_t}] \tag{19}$$

where $E_{pmodel}[]$ and $E_{pdata}[]$ denote ***model's expectation*** and ***data-dependent expectation respectively***. Here, variational learning is employed to calculate data-dependent expectation. The mean-field approach is applied to calculate an approximation of the true posterior with $\boldsymbol{q}\left(\boldsymbol{h^k_{ij}}=\boldsymbol{1}\right)=\boldsymbol{\mu^k_{ij}}, \boldsymbol{q}\left(\boldsymbol{s^k_{ij}}\right)=\boldsymbol{^k_{ij}}, \boldsymbol{q}\left(\boldsymbol{\rho^g_{i'j'}}=\boldsymbol{1}\right)=\boldsymbol{\xi^g_{ij}}, \boldsymbol{q}\left(\boldsymbol{\ell^g_{i'j'}}\right)=\boldsymbol{\tau^g_{ij}}, \boldsymbol{q}\left(\boldsymbol{y_t}=\boldsymbol{1}\right)=\boldsymbol{\kappa_t}$. We try to maximize a limited bound with respect to the variational parameters $\{\mu, p\mu, p, \xi, p\xi, \tau, p\tau, \kappa\}$ according to Eq. (20) (see Box II).

DCssCDBM utilizes SAP to evaluate the $E_{pmodel}[]$ as well (Salakhutdinov & Hinton, 2012). $\left\|J_f(v)\right\|_F^2=\sum_l\sum_{m,n}\sum_k\sum_{i,j}\left(\frac{\partial p(h^k_{ij}(v)|v)}{\partial v^l_{mn}}\right)^2$ is also employed as a probabilistic penalty term as well in order to learn robust feature (Rifai, Muller, et al., 2011). The modified objective function is described in Eq. (21).

$$Loss=-\sum_{i=1}^{N}\log\left(p\left(v_i,y_i\right)\right)+\lambda\left\|J_f(v_i)\right\|_F^2 \tag{21}$$

Here, λ counterbalances importance of penalty term. We give a general expression of partial derivative with respect to $\theta=\{w^{l,k}, w^{k,g}, b^k, b^g, a^k_{ij}, a^g_{i'j'}, c_l, u_{t,i'j'}^{\ g}, d_t\}$ in Eq. (22).

$$\frac{\partial L(v)}{\partial\theta}=-E_{pdata}\left[\frac{\partial F(v)}{\partial\theta}\right]+E_{pmodel}[\frac{\partial F(v)}{\partial\theta}]+\lambda\frac{\partial}{\partial\theta}\left\|J_f(v)\right\|_F^2 \tag{22}$$

$$p\left(h_{ij}^{k}=1|v,\rho,\ell\right)=\frac{\exp\left(\frac{1}{2a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}^{2}+\sum_{g=1}^{G}\left[\left(\rho^{g}\cdot\ell^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right)+b^{k}\right)}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{i'j'}^{2}+\sum_{g=1}^{G}\left[\left(\rho^{g}\cdot\ell^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right)+b^{k}\right)} \tag{2}$$

$$p\left(ph_{ij}^{k}=0|v,\rho,\ell\right)=\frac{1}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{i'j'}^{2}+\sum_{g=1}^{G}\left[\left(\rho^{g}\cdot\ell^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right)+b^{k}\right)} \tag{3}$$

$$p\left(s_{ij}^{k}|v,h,\rho,\ell\right)=\therefore\left(\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}(v^{l}\Theta w^{l,k})_{ij}h_{ij}^{k}+\sum_{g=1}^{G}\left[\left(\rho^{g}\cdot\ell^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right),\frac{1}{a_{ij}^{k}}\right) \tag{4}$$

$$p\left(ps_{ij}^{k}|v,h,\rho,\ell\right)=\therefore N\left(\max_{(i,j)\in B_a}\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}(v^{l}\Theta w^{l,k})_{ij}h_{ij}^{k}+\sum_{g=1}^{G}\left[\left(\rho^{g}\cdot\ell^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right),\frac{1}{a_{ij}^{k}}\right) \tag{5}$$

$$\boldsymbol{p}\left(\ell_{ij}^{k}=\mathbf{1}|\boldsymbol{ps},\boldsymbol{ph},\rho,\boldsymbol{y}\right)=N\frac{\exp\left(\frac{1}{2a_{ij}^{g}}\sum_{k=1}^{K}\left((\boldsymbol{ps}\cdot\boldsymbol{ph})^{k}\boldsymbol{\Theta}\boldsymbol{w}^{k,g}\right)_{ij}^{2}+\left(\sum_{t=1}^{T}\boldsymbol{y}_{t}\boldsymbol{u}_{t,ij}^{g}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\rho_{ij}^{g}+\boldsymbol{b}^{g}\right)}{\mathbf{1}+\sum_{(i',j')\in\boldsymbol{B_a}}\exp\left(\frac{1}{2a_{i'j'}^{g}}\sum_{k=1}^{K}\left((\boldsymbol{ps}\cdot\boldsymbol{ph})^{k}\boldsymbol{\Theta}\boldsymbol{w}^{k,g}\right)_{i'j'}^{2}+\left(\sum_{t=1}^{T}\boldsymbol{y}_{t}\boldsymbol{u}_{t,ij}^{g}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\rho_{ij}^{g}+\boldsymbol{b}^{g}\right)} \tag{6}$$

$$\boldsymbol{p}\left(\boldsymbol{p}\ell_{a}^{g}=\mathbf{0}|\boldsymbol{ps},\boldsymbol{ph},\rho,\boldsymbol{y}\right)=\frac{\mathbf{1}}{\mathbf{1}+\sum_{(i',j')\in\boldsymbol{B_a}}\exp\left(\frac{1}{2a_{i'j'}^{g}}\sum_{k=1}^{K}\left((\boldsymbol{ps}\cdot\boldsymbol{ph})^{k}\boldsymbol{\Theta}\boldsymbol{w}^{k,g}\right)_{i'j'}^{2}+\left(\sum_{t=1}^{T}\boldsymbol{y}_{t}\boldsymbol{u}_{t,ij}^{g}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\rho_{ij}^{g}+\boldsymbol{b}^{g}\right)} \tag{7}$$

$$\boldsymbol{p}\left(\rho_{ij}^{g}|\boldsymbol{ph},\boldsymbol{ps},\ell,\boldsymbol{y}\right)=\therefore N\left(\frac{\mathbf{1}}{\boldsymbol{a}_{ij}^{g}}\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}}\left((\boldsymbol{ph}\cdot\boldsymbol{ps})\ \boldsymbol{\Theta}\boldsymbol{w}^{\boldsymbol{k},\boldsymbol{g}}\right)_{ij}\ell_{ij}^{g}+\left(\sum_{\boldsymbol{t}=\mathbf{1}}^{\boldsymbol{T}}\boldsymbol{y}_{\boldsymbol{t}}\boldsymbol{u}_{\boldsymbol{t},ij}^{\boldsymbol{g}}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\ell_{ij}^{g},\frac{\mathbf{1}}{\boldsymbol{a}_{ij}^{g}}\right) \tag{8}$$

$$\boldsymbol{p}\left(\boldsymbol{p}\rho_{a}^{g}|\boldsymbol{ph},\boldsymbol{ps},\boldsymbol{\ell},\boldsymbol{y}\right)=\therefore N\left(\mathbf{max}_{(\boldsymbol{i},\boldsymbol{j})\in\boldsymbol{B_a}}\frac{\mathbf{1}}{\boldsymbol{a}_{ij}^{g}}\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}}\left((\boldsymbol{ph}\cdot\boldsymbol{ps})\ \boldsymbol{\Theta}\boldsymbol{w}^{\boldsymbol{k},\boldsymbol{g}}\right)_{ij}\ell_{ij}^{g}+\left(\sum_{\boldsymbol{t}=\mathbf{1}}^{\boldsymbol{T}}\boldsymbol{y}_{\boldsymbol{t}}\boldsymbol{u}_{\boldsymbol{t},ij}^{\boldsymbol{g}}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\ell_{ij}^{g},\frac{\mathbf{1}}{\boldsymbol{a}_{ij}^{g}}\right) \tag{9}$$

$$\boldsymbol{p}\left(\boldsymbol{y}_{t}=\mathbf{1}|\boldsymbol{p}\rho,\boldsymbol{p}\ell\right)=\frac{\exp\left(\sum_{g,i',j'}\boldsymbol{u}_{t,i'j'}^{g}\boldsymbol{p}\rho_{i'j'}^{g}\boldsymbol{p}\ell_{i'j'}^{g}+\boldsymbol{d}_{t}\right)}{\sum_{t}\exp\left(\sum_{g,i',j'}\boldsymbol{u}_{t,i'j'}^{g}\boldsymbol{p}\rho_{i'j'}^{g}\boldsymbol{p}\ell_{i'j'}^{g}+\boldsymbol{d}_{t}\right)} \tag{10}$$

Box I.

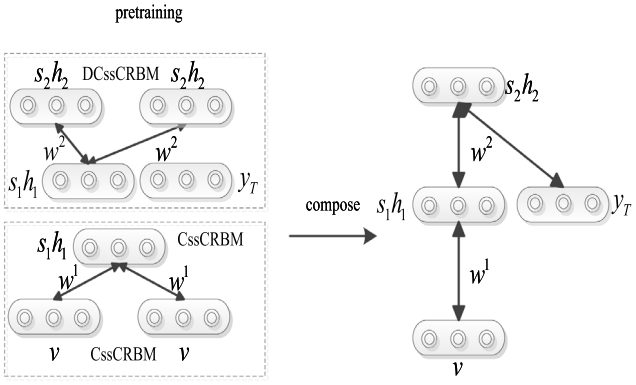


Fig. 2. Pretraining CssCRBM and DCssCRBM for initialization.

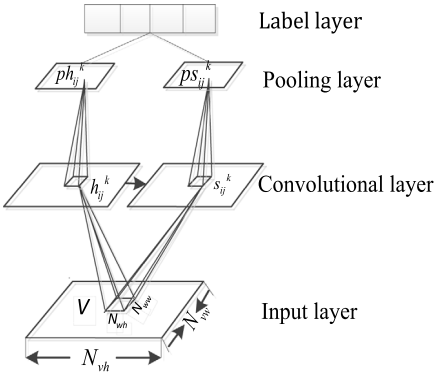


Fig. 3. Schematic diagram of DCssCRBM.

$\nabla\theta_1$ is a summarization of the first two term, and $\nabla\theta_2$ is $\frac{\partial}{\partial\theta}\left\|J_f(v)\right\|_F^2$ in Eq. (22). Furthermore, Table 1 gives the details of how to train DCssCDBM.

Differently from CssCDBM, we should design a new discriminative block to replace original CssCRBM on the top block for pretraining in Fig. 2.

The energy function of DCssCRBM can be defined in Eq. (23) and depicts Schematic diagram of DCssCRBM in Fig. 3.

$$E\left(v,s,h,y\right)=\sum_{l=1}^{L}\sum_{m=1}^{N_{vh}}\sum_{n=1}^{N_{vw}}\frac{1}{2\sigma^{2}}\left(v_{mn}^{l}-c_{l}\right)^{2}$$

$$
\begin{aligned}
\mu_{ij}^{k} &= \frac{\exp\left(\frac{1}{2a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}^{2}+\sum_{g=1}^{G}\left[\left(\xi^{g}\cdot\tau^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones(C,C)+b^{k}\right)}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{i'j'}^{2}+\sum_{g=1}^{G}\left[\left(\xi^{g}\cdot\tau^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones(C,C)+b^{k}\right)}\\
p\mu_{ij}^{k} &= \frac{1}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{i'j'}^{2}+\sum_{g=1}^{G}\left[\left(\xi^{g}\cdot\tau^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones(C,C)+b^{k}\right)}\\
v_{ij}^{k} &= \because N\left(\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}\mu_{ij}^{k}+\sum_{g=1}^{G}\left[\left(\xi^{g}\cdot\tau^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right),\frac{1}{a_{ij}^{k}}\right)\\
pv_{ij}^{k} &= \because N\left(\max_{(i,j)\in B_a}\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}\mu_{ij}^{k}+\sum_{g=1}^{G}\left[\left(\xi^{g}\cdot\tau^{g}\right)\Theta\widetilde{w^{k,g}}\right]\otimes ones\left(C,C\right),\frac{1}{a_{ij}^{k}}\right)\\
\xi_{i'j'}^{g} &= \frac{\exp\left(\frac{1}{2a_{i'j'}^{g}}\sum_{k=1}^{K}\left((p\mu\cdot p)^{k}\Theta w^{k,g}\right)_{i'j'}^{2}+(\sum_{t=1}^{T}\kappa_t u_{t,ij}^{g})\otimes\boldsymbol{ones}(C,C)\tau_{i'j'}^{g}+b^{g}\right)}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{g}}\sum_{k=1}^{K}\left((p\mu\cdot p)^{k}\Theta w^{k,g}\right)_{i'j'}^{2}+(\sum_{t=1}^{T}\kappa_t u_{t,ij}^{g})\otimes\boldsymbol{ones}(C,C)\tau_{i'j'}^{g}+b^{g}\right)}\\
p\xi_{i'j'}^{g} &= \frac{1}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{g}}\sum_{k=1}^{K}\left((p\mu\cdot p)^{k}\Theta w^{k,g}\right)_{i'j'}^{2}+(\sum_{t=1}^{T}\kappa_t u_{t,ij}^{g})\otimes\boldsymbol{ones}(C,C)\tau_{i'j'}^{g}+b^{g}\right)}\\
\tau_{i'j'}^{g} &= \because N(\frac{1}{a_{i'j'}^{g}}\sum_{k=1}^{K}\left((p\mu\cdot p)\,\Theta w^{k,g}\right)_{i'j'}\xi_{i'j'}^{g}+(\sum_{t=1}^{T}\kappa_t u_{t,ij}^{g})\otimes\boldsymbol{ones}\left(C,C\right)\xi_{i'j'}^{g},\frac{1}{a_{i'j'}^{g}})\\
p\tau_{i'j'}^{g} &= \because N(\max_{(i',j')\in B_a}\frac{1}{a_{i'j'}^{g}}\sum_{k=1}^{K}\left((p\mu\cdot p)\,\Theta w^{k,g}\right)_{i'j'}\xi_{i'j'}^{g}+(\sum_{t=1}^{T}\kappa_t u_{t,ij}^{g})\otimes\boldsymbol{ones}\left(C,C\right)\xi_{i'j'}^{g},\frac{1}{a_{i'j'}^{g}})\\
\kappa_t &= \frac{\exp\left(\sum_{g,i'j'}u_{t,i'j'}^{g}p\xi_{i'j'}^{g}p\tau_{i'j'}^{g}+d_t\right)}{\sum_{t}\exp\left(\sum_{g,i'j'}u_{t,i'j'}^{g}p\xi_{i'j'}^{g}p\tau_{i'j'}^{g}+d_t\right)}
\end{aligned}
\tag{20}
$$

Box II.

$$
\begin{aligned}
&-\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}}\left(v^{l}\Theta w^{l,k}\right)_{ij}h_{ij}^{k}s_{ij}^{k}-\sum_{k=1}^{K}b^{k}\sum_{i=1,j=1}^{N_{hh},N_{hw}}h_{ij}^{k}\\
&+\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}}s_{ij}^{k}a_{ij}^{k}s_{ij}^{k}-\sum_{t=1}^{T}\sum_{k,i,j}\boldsymbol{y_t u_{t,ij}^{k}}\boldsymbol{s_{ij}^{k}h_{ij}^{k}}-\sum_{t=1}^{T}\boldsymbol{y_t d_t}
\end{aligned}
\tag{23}
$$

Similar to CssCDBM, the visible, spike and slab hidden units are conditionally independent. The without-pooling inferences s can be written as follows.

Inference I:

$$
p\left(h_{ij}^{k}=1|v,y,s\right)=\sigma(\frac{1}{2a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}^{2}+\sum_{t=1}^{T}\boldsymbol{y_t u_{t,ij}^{k}}\boldsymbol{s_{ij}^{k}}+b^{k})
\tag{24}
$$

Inference II:

$$
p\left(s_{ij}^{k}|v,h,y\right)=\because(\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}h_{ij}^{k}+\sum_{t=1}^{T}\boldsymbol{y_t u_{t,ij}^{k}}\boldsymbol{h_{ij}^{k}},\frac{1}{a_{ij}^{k}})
\tag{25}
$$

Inference III:

$$
\boldsymbol{p}\left(\boldsymbol{y_t}=1|\boldsymbol{s},\boldsymbol{h}\right)=\frac{\exp\left(\sum_{k,i,j}\boldsymbol{u_{t,ij}^{k}}\boldsymbol{s_{ij}^{k}}\boldsymbol{h_{ij}^{k}}+\boldsymbol{d_t}\right)}{\sum_{t}\exp\left(\sum_{k,i',j'}\boldsymbol{u_{t,i'j'}^{k}}\boldsymbol{s_{i'j'}^{k}}\boldsymbol{h_{i'j'}^{k}}+\boldsymbol{d_t}\right)}
\tag{26}
$$

Inference IV:

$$
p\left(v_{mn}^{l}|h,s\right)=\because(\sigma^{2}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}}h_{ij}^{k}s_{ij}^{k}\Theta\widetilde{w^{l,k}}+c_l,\sigma^{2})
\tag{27}
$$

Where the notation $\sigma()$ is a sigmoid function.

We can also define the energy function of DCssCRBM with Dual variable probabilistic max-pooling in Eq. (28) and the inferences can be written as Eqs. (29)∼(33) describe.

$$
\begin{aligned}
&E\left(v,s,h,y\right)\\
&=\sum_{l=1}^{L}\sum_{m=1}^{N_{vh}}\sum_{n=1}^{N_{vw}}\frac{1}{2\sigma^{2}}\left(v_{mn}^{l}-c_l\right)^{2}-\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}}\left(v^{l}\Theta w^{l,k}\right)_{ij}h_{ij}^{k}s_{ij}^{k}\\
&-\sum_{k=1}^{K}b^{k}\sum_{i=1,j=1}^{N_{hh},N_{hw}}h_{ij}^{k}+\frac{1}{2}\sum_{k=1}^{K}\sum_{i=1,j=1}^{N_{hh},N_{hw}}s_{ij}^{k}a_{ij}^{k}s_{ij}^{k}\\
&-\sum_{t=1}^{T}\sum_{k,i,j}\boldsymbol{y_t u_{t,ij}^{k}}\boldsymbol{ps_{ij}^{k}}\boldsymbol{ph_{ij}^{k}}-\sum_{t=1}^{T}\boldsymbol{y_t d_t}\\
&\quad subject\ to:\ \sum_{(i,j)\in B_a}h_{ij}^{k}\le 1,\forall k,a;
\end{aligned}
\tag{28}
$$

Inference I: See Eqs. (29) and (30) which are given in Box III.

Inference II:

$$
\begin{aligned}
p\left(s_{ij}^{k}|v,ph,y\right)&=\because\left(\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}h_{ij}^{k}+(\sum_{t=1}^{T}\boldsymbol{y_t u_{t,ij}^{k}})\right.\\
&\left.\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\boldsymbol{ph_{ij}^{k}},\frac{1}{a_{ij}^{k}}\right)
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
p\left(s_{ij}^{k}|v,ph,y\right)&=\because\left(\max_{(i,j)\in B_a}\frac{1}{a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij}h_{ij}^{k}\right.\\
&\left.+(\sum_{t=1}^{T}\boldsymbol{y_t u_{t,ij}^{k}})\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\boldsymbol{ph_{ij}^{k}},\frac{1}{a_{ij}^{k}}\right)
\end{aligned}
\tag{32}
$$

Table 1
The procedure of training DCssCDBM model.

- Given : a training set of N data vectors $\{v_n, y_n\}_{n=1}^{N}$ and setting hyper parameters $\{batchsize, epoch, K, \lambda, M\}$
- Randomly initialize parameters θ^0 and M fantasy particles.
  $\{\overline{v^{0,1}}, \overline{h^{0,1}}, \overline{ph^{0,1}}, \overline{s^{0,1}}, \overline{ps^{0,1}}, \overline{\ell^{0,1}}, \overline{p\ell^{0,1}}, \overline{\rho^{0,1}}, \overline{p\rho^{0,1}}, \overline{y^{0,1}}\}$,
  $\{\overline{v^{0,M}}, \overline{h^{0,M}}, \overline{ph^{0,M}}, \overline{s^{0,M}}, \overline{ps^{0,M}}, \overline{\ell^{0,M}}, \overline{p\ell^{0,M}}, \overline{\rho^{0,M}}, \overline{p\rho^{0,M}}, \overline{y^{0,M}}\}$
- For t=0 to *epoch* (# of iterations)
  - For i=1 to int (N/*batchsize*):
    - $v = v[(i\text{-}1)*\ batchsize: i*\ batchsize]$,
      $y=y[(i\text{-}1)*\ batchsize: i*\ batchsize]$
      - Randomly initialize $\{\mu, p\mu, \nu, p\nu, \xi, p\xi, \tau, p\tau\}$ and run mean-field to update Eq.20 until convergence.
      - Set
        $$\mu^n = \mu \quad p\mu^n = p\mu \quad \nu^n = \nu \quad p\nu^n = p\nu \quad \xi^n = \xi \quad p\xi^n = p\xi \quad \tau^n = \tau \quad p\tau^n = p\tau$$
      - For each fantasy particle, m=1,..., M
        Obtain a new state:
        $\{\overline{v^{t+1,M}}, \overline{h^{t+1,M}}, \overline{ph^{t+1,M}}, \overline{s^{t+1,M}}, \overline{ps^{t+1,M}}, \overline{\ell^{t+1,M}}, \overline{p\ell^{t+1,M}}, \overline{\rho^{t+1,M}}, \overline{p\rho^{t+1,M}}, \overline{y^{t+1,M}}\}$
        by running a K CD-K Gibbs sampler using Eqs.2~10, initialized at the previous sample $\{\overline{v^{t,M}}, \overline{h^{t,M}}, \overline{ph^{t,M}}, \overline{s^{t,M}}, \overline{ps^{t,M}}, \overline{\ell^{t,M}}, \overline{p\ell^{t,M}}, \overline{\rho^{t,M}}, \overline{p\rho^{t,M}}, \overline{y^{t,M}}\}$

      End For
    - Calculate Gradients: $\nabla\theta_1$

$$\nabla \mathrm{w}^{l,k} = \frac{1}{batchsize}\sum_{bat=1}^{batchsize}(v_{bat})^l \Theta (\mu^n \bullet \nu^n)^k - \frac{1}{M}\sum_{m=1}^{M}(\overline{v^{K,m}})^l \Theta \left(\overline{h^{K,m}} \bullet \overline{s^{K,m}}\right)^k$$

$$\nabla \mathrm{w}^{k,g} = \frac{1}{batchsize}\sum_{bat=1}^{batchsize}(p\mu_{bat}^n \bullet p\nu_{bat}^n)^k \Theta (\xi^n \bullet \tau^n)^g - \frac{1}{M}\sum_{m=1}^{M}(\overline{ph^{K,m}} \bullet \overline{ps^{K,m}})^k \Theta \left(\overline{\ell^{K,m}} \bullet \overline{\rho^{K,m}}\right)^g$$

$$\nabla b^k = \frac{1}{batchsize}\sum_{bat=1}^{batchsize}\sum_{ij}(\mu_{bat}^n)_{ij}^k - \frac{1}{M}\sum_{m=1}^{M}\sum_{ij}(\overline{h^{K,m}})_{ij}^k$$

$$\nabla \mathrm{b}^{\mathrm{g}} = \frac{1}{\mathrm{batchsize}}\sum_{\mathrm{bat}=1}^{\mathrm{batchsize}}\sum_{\mathrm{i'j'}}(\xi_{\mathrm{bat}}^{\mathrm{n}})_{\mathrm{i'j'}}^{\mathrm{g}} - \frac{1}{\mathrm{M}}\sum_{\mathrm{m}=1}^{\mathrm{M}}\sum_{\mathrm{i'j'}}(\overline{\ell^{\mathrm{K,m}}})_{\mathrm{i'j'}}^{\mathrm{g}}$$

$$\nabla a_{ij}^k = \frac{1}{2 * batchsize}\sum_{bat=1}^{batchsize}\sum_{ij}(\nu_{bat}^n \bullet \nu_{bat}^n)_{ij}^k - \frac{1}{M}\sum_{m=1}^{M}\sum_{ij}(\overline{s^{K,m}} \bullet \overline{s^{K,m}})_{ij}^k$$

$$\nabla a_{i'j'}^g = \frac{1}{2 * batchsize}\sum_{bat=1}^{batchsize}\sum_{i'j'}(\xi_{bat}^n \bullet \xi_{bat}^n)_{i'j'}^g - \frac{1}{M}\sum_{m=1}^{M}\sum_{i'j'}(\overline{\rho^{K,m}} \bullet \overline{\rho^{K,m}})_{i'j'}^g$$

$$\nabla c_l = \frac{1}{batchsize}\sum_{bat=1}^{batchsize}\sum_{ij}(c_l - v_{bat})_{ij}^l - \frac{1}{M}\sum_{m=1}^{M}\sum_{ij}(c_l - \overline{v^{K,m}})_{ij}^l$$

$$\nabla u_{t,i'j'}^g = \sum_{bat=1}^{batchsize} y_{bat,t} p\xi_{i'j'}^g p\tau_{i'j'}^g - \frac{1}{M}\sum_{m=1}^{M}\overline{y_t^{K,m}}\,\overline{p\ell_{g\ i'j'}^{K,m}}\,\overline{p\rho_{g\ i'j'}^{K,m}}$$

$$\nabla d_t = \sum_{bat=1}^{batchsize} y_{bat,t} - \frac{1}{M}\sum_{m=1}^{M}\overline{y_t^{K,m}}$$

    - Calculate Gradients: $\nabla\theta_2$
    - Update parameters
      $$\theta(t+1) = \theta(t) + \nabla\theta_1 + \lambda\nabla\theta_2$$

  End For
End For

$$p\left(h_{ij}^{k}=1|v,\rho s,\mathrm{y}\right)=\frac{\exp\left(\frac{1}{2a_{ij}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{ij+}^{2}\left(\sum_{t=1}^{T}\boldsymbol{y_t}\boldsymbol{u}_{\boldsymbol{t},\boldsymbol{ij}}^{\boldsymbol{k}}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\boldsymbol{ps}_{\boldsymbol{ij}}^{\boldsymbol{k}}+b^{k}\right)}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{i'j'}^{2}+\left(\sum_{t=1}^{T}\boldsymbol{y_t}\boldsymbol{u}_{\boldsymbol{t},\boldsymbol{i'j'}}^{\boldsymbol{k}}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\boldsymbol{ps}_{\boldsymbol{i'j'}}^{\boldsymbol{k}}+b^{k}\right)} \tag{29}$$

$$p\left(ph_{a}^{k}=0|v,\rho s,\mathrm{y}\right)=\frac{1}{1+\sum_{(i',j')\in B_a}\exp\left(\frac{1}{2a_{i'j'}^{k}}\sum_{l=1}^{L}\left(v^{l}\Theta w^{l,k}\right)_{i'j'}^{2}+\left(\sum_{t=1}^{T}\boldsymbol{y_t}\boldsymbol{u}_{\boldsymbol{t},\boldsymbol{i'j'}}^{\boldsymbol{k}}\right)\otimes\boldsymbol{ones}\left(\boldsymbol{C},\boldsymbol{C}\right)\boldsymbol{ps}_{\boldsymbol{i'j'}}^{\boldsymbol{k}}+b^{k}\right)} \tag{30}$$

Box III.

Inference III:

$$\boldsymbol{p}\left(\boldsymbol{y_t}=\boldsymbol{1}|\boldsymbol{s},\boldsymbol{h}\right)=\frac{\exp\left(\sum_{\boldsymbol{k},\boldsymbol{i},\boldsymbol{j}}\boldsymbol{u}_{\boldsymbol{t},\boldsymbol{ij}}^{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{ij}}^{\boldsymbol{k}}\boldsymbol{h}_{\boldsymbol{ij}}^{\boldsymbol{k}}+\boldsymbol{d_t}\right)}{\sum_{\boldsymbol{t}}\exp\left(\sum_{\boldsymbol{k},\boldsymbol{i'},\boldsymbol{j'}}\boldsymbol{u}_{\boldsymbol{t},\boldsymbol{i'j'}}^{\boldsymbol{k}}\boldsymbol{s}_{\boldsymbol{i'j'}}^{\boldsymbol{k}}\boldsymbol{h}_{\boldsymbol{i'}}\boldsymbol{j'^{k}}+\boldsymbol{d_t}\right)} \tag{33}$$

This work is to keep consistent with our deep model for easy training. We also adopt Eq. (21) as objective function. The details of training DCssCRBM are given in Table 2.

3.2. Multi-task learning framework based on DCssCDBM

This section mainly describes our multi-task learning framework for early diagnosis of AD based on DCssCDBM in order to improve generalization ability. Generally, deep models easily face with overfitting especially under limited training data, which is a shared concern and commonly defined by a model that behaves well on the training examples, but behaves poorly on the unseen examples during the test. In order to reduce overfitting and boost the generalization performance of the early diagnosis of AD, we propose multi-task learning framework based on Siamese DCssCDBM architecture with shared weights as depicted in Fig. 4. Our multi-task learning framework contains an early diagnosis of AD, identification, and verification. What needs to be highlighted is that weights between identification label layer and the last hybrid features do not share with weights between diagnosis label layer and the last hybrid feature layer. We just need small modifications for inference. The cost function *Loss* defined in Eq. (34). The details of each loss function are provided in the next part.

$$\begin{aligned}Loss=\sum_{(v_{1,i},v_{2,i})\in X}\{&\left[loss_{main}\left(v_{1,i},y_{main,(1,i)}\right)+loss_{main}\left(v_{2,i},y_{main,(2,i)}\right)\right]\\&+c_{ide}\left[loss_{ide}\left(v_{1,i},y_{ide,(1,i)}\right)+loss_{ide}\left(v_{2,i},y_{ide,(2,i)}\right)\right]\\&+c_{ver}loss_{ver}(v_{1,i},v_{2,i},y_{ver,i})\}\end{aligned} \tag{34}$$

where X is used for training which contains many EEG spectral image pairs $(v_{1,i}, v_{2,i})$. Each pair additionally includes a verification label $y_{ver,i} \in \{0, 1\}$, representing whether the two EEG spectral images are obtained from the same or different subjects. Meanwhile, each image has an identification label denoting as $y_{ide,i} \in \{0, \ldots, n\}$. We use $y_{main,i} \in \{0, 1, 2\}$ to represent diagnosis label, denoting three states (AD, HC, and MCI) of subjects. The symbols c_{ide} and c_{ver} balance importance between auxiliary task and main task.

(1) Main Task: Early Diagnosis of AD

We convey an EEG spectral image v into DCssCDBM in order to classify this subject into the correct category from three categories. Indeed, our DCssCDBM can directly minimize the minus log joint probability $loss_{main} = -logp(v, y_{main})$. After training, we can employ Eq. (20) to determine correct label.

(2) Auxiliary Task 1: Identification

This task follows the same pipeline as main task. We use DCssCDBM to classify a given EEG spectral image v for determining this image belonging to which subject. We also can use DCssCDBM to be trained directly by minimizing minus log joint probability $loss_{ide} = -logp(v, y_{ide})$. **Different from the main task, we should add an identification label layer to DCssCDBM.** The other parts of model for this task share weights with model for main task. We follow the same procedure as main task to determine correct identification label according to Eq. (20).

(3) Auxiliary Task 2: Verification

As for the verification task, we focus DCssCDBM to determine whether two EEG spectral images (v_1, v_2) come from the same person. We use Siamese DCssCDBM architecture (see Fig. 4) to perform identity verification, in which it is made up of two copies with shared weights. **It is emphasized that we treat DCssCDBM as a special discriminative forward convolutional neural network for this task. This method weakly approximates variational learning for computation reduction. Meanwhile, it acts as regularization.** In Fig. 4, each image through forward inference of subnetwork $G(w)$ produces features $G(v_1; w)$ and $G(v_2; w)$ during training. Eq. (35) computes the Euclidean distance between the features.

$$D\left(v_1, v_2; w\right) = \left\|G\left(v_1; w\right) - G(v_2; w)\right\|_2 \tag{35}$$

We adopt the simple Euclidean distance of the features from each pair for training the DCssCDBM to complete the verification task. Considering a given EEG spectral image pair (v_1, v_2), its loss function denoted as $loss_{ver}$ decide by whether the EEG spectral images in one pair are from the identical or different subject. Our cost function should contain both cases. In Eq. (36), it represents the condition where v_1 and v_2 are EEG spectral images from the same subject.

$$loss_{vfs}\left(v_1, v_2; w\right) = \frac{1}{2}D(v_1, v_2; w)^2 \tag{36}$$

Therefore, the cost should be close to zero since the features are identical since they come from the same person. In another possibility, when v_1 and v_2 images are from different persons, the loss function can be measured as Eq. (37).

$$loss_{vfd}\left(v_1, v_2; w\right) = \frac{1}{2}(\max(0, \delta - D(v_1, v_2; w)))^2 \tag{37}$$

In this case, we call the variable δ as margin in Eq. (37). The cost value is set to zero under the condition that the distance between the features of two images is greater than δ. This margin therefore greatly helps the model to better discriminate between dissimilar and similar images. We can make further efforts to integrate the above defined loss functions $loss_{vfs}$ and $loss_{vfd}$ into a single cost function $loss_{ver}$ able to cope with diverse cases. Here, we denote $y_{ver} = 1$ when the two EEG spectral images from the same subject and $y_{ver} = 0$ for another case. Eq. (30) gives a combination loss

Table 2
The details of training DCssCRBM model.

- Given : a training set of N data vectors $\{v_n, y_n\}_{n=1}^{N}$ and setting hyper parameters $\{batchsize, epoch, K, \lambda, \eta, M\}$
- Randomly initialize trainable parameters θ
- For i=1 to $epoch$:
  - For j=1 to int ($N/batchsize$):
    - $v(0)= v[(i\text{-}1)* batchsize: i* batchsize]$
      $y(0)= y[(i\text{-}1)* batchsize: i* batchsize]$
    - Sample $h(0)$ and $s(0)$ from Eq.29 and Eq.31 or Eq.24 and Eq.25
    - For n=1 to K:
      Get samples $v(n)$ from Eq.27, get samples $h(n)$ and $s(n)$ Eq.29 and Eq.31 or Eq.24 and Eq.25
      Get samples $y(n)$ from Eq.26 or Eq.33
      End For
    - Update weights using samples from above

      $$\theta: \theta - \eta(\nabla\theta_1 + \lambda\nabla\theta_2)$$

  End For
End for

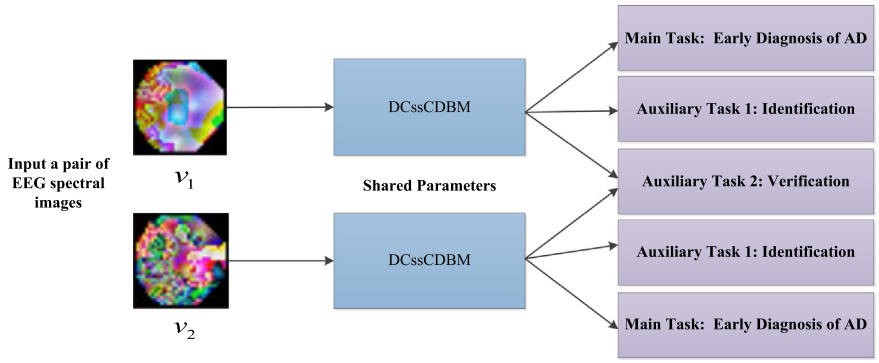


Fig. 4. Siamese DCssCDBM architecture for multi-task learning.

function.

$$loss_{ver} = y_{ver} loss_{vfs}(v_1, v_2; w) + (1 - y_{ver}) loss_{vfd}(v_1, v_2; w) \quad (38)$$

We integrate the main task and two auxiliary tasks, and further design Siamese DCssCDBM as Fig. 4 depicted to achieve multi-task learning with share weights before the last hybrid feature layer. For main task, we use $u_{t,\substack{g\\i'j'}}$ to denote weight between the last hybrid feature layer and label layer corresponding to healthy states (HC, MCI, and AD) and then we employ $u_{f,\substack{g\\i'j'}}$ to represent weight between the last hybrid feature layer and label layer corresponding to different identities. For easy explanation, θ demonstrates whole parameters, including shared parameters and unshared parameters. Last, we provide details of training DCssCDBM via Multi-task Learning in Table 3.

4. Experiments

4.1. Data acquisition and preprocessing

The EEG data used in this study were collected by our cooperative partner, Beijing Easy monitor Technology. Here, all AD and MCI patients were diagnosed by the skillful neurologists based on NINCDS-ADRDA criteria (Dubois, Jacova, et al., 2007), Mini-Mental State Examination (MMSE) and clinical dementia rating (CDR) (Cummings, 1993). Our participants for experiment can be classified into three groups. The first (HC) includes 4 cognitively healthy controls (2 females, 2 males; mean age 65.2 yr), and the second has 4 MCI patients (2 females, 2 males; mean age 72.2 yr), and the third (AD) has 4 mild-to-severe AD patients (2 females, 2 males, mean age 76.2 yr). The inclusion criteria for the three groups are described in Table 4. During data collection, In brief, all participants were required to connect with 64-channel EEG electrodes placing over the scalp at standard locations and our EEG signal was continuously obtained from these many electrodes with a sampling frequency of 500 Hz under resting with eyes closed for 1 min. Data for two of the subjects was excluded from the dataset on account of amounts of noise and artifacts in their data until their clean are enough clean for this experiment.

EEG signal is high-dimension, which consists of multiple time series corresponding to measurements across different spatial locations over the cortex. This demonstrates that EEG signal has additional spatial dimension information. Similar to other time sequence signals, such as speech signals, we usually study EEG by calculating spectrogram of the signal since the most salient features reside in the frequency domain. Fast Fourier Transform (FFT) is widely used on the time series for each trial to estimate the power spectrum of the signal. Recently, several studies have indicated that oscillatory cortical activity primarily exists in the theta band (4–7 Hz), the alpha band (8–13 Hz), and the beta band (13–30 Hz). Here, we used FFT in 0.5s time slice of each

Table 3
The details of training DCssCDBM model with multi-task learning.

- Given : a training set of N data vectors $\{v_n, y_{n,ide}, y_{n,main}\}_{n=1}^{N}$ and setting hyper parameters $\{batchsize, epoch, K, \lambda, \eta, M, c_{ide}, c_{ver}\}$
- initialize trainable parameters θ following Tab.2, where θ includes shared parameters and unshared parameters
- For i=1 to $epoch$:
  - For j=1 to int ($N/batchsize$):
    - $v(0)$= $v[(i\text{-}1)*\ batchsize: i*\ batchsize]$,
      $y_{main}(0)$= y_{main} $[(i\text{-}1)*\ batchsize: i*\ batchsize]$
      $y_{ide}(0)$= y_{ide} $[(i\text{-}1)*\ batchsize: i*\ batchsize]$
    - Given $v(0)$, $y_{main}(0)$ and related hyper parameters, optimize $loss_{main} = -logp(v, y_{main})$ following Tab.1 related procedure and obtain optimized parameters $\nabla\theta_{main}$
    - Given $v(0)$, $y_{ide}(0)$ and related hyper parameters, optimize $loss_{ide} = -logp(v, y_{ide})$ following Tab.1 related procedure and obtain optimized parameters $\nabla\theta_{ide}$
    - Split $v(0)$ into two groups equally, and randomly combine two EEG images in order to produce training pairs; last, feed these training pairs into Siamese DCssCDBM, and use back propagation to obtain optimized parameters $\nabla\theta_{ver}$
    - Update weights using
      $\theta: \theta - \eta(\nabla\theta_{main} + c_{ide} * \nabla\theta_{ide} + c_{ver} * \nabla\theta_{ver})$

  End For

End for

Table 4
Inclusion criteria for AD, MCI, and HC.

Inclusion criteria		
HC	MCI	AD
CDR 0 ∼ 0.5	CDR 0.5 ∼ 1	CDR 1 ∼ 3
MMSE>=24	MMSE>=24	MMSE<24

Table 5
The details of dataset.

Class	Instances	Images for each Instance	Total images
AD	4	1000	
HC	4	1000	12 000
MCI	4	1000	

channel and send results to corresponding low pass filter as separate measurements for each electrode. Considering the inherent structure of the data in space, frequency, and time, we used a method proposed in Pouya Bashivan, Yeasin, and Codella (2016) in order to transform the measurements into a 2-D image to save the spatial structure and used multiple color channels to represent the spectral dimension. Continuous EEG was sliced offline to equal lengths of 1 min corresponding to each trial. For each participant, we collected 1000 EEG spectral images with 32 × 32 resolutions. The whole dataset consisted of 12 000 images. Here, we show a table summarizing the dataset in Table 5. In Fig. 5, we displayed overall framework of our method.

4.2. Model configurations

The DCssCDBM structure of final system in this experiment is described in Table 6. Our final model has about 80 thousand parameters. Compared to the size of the available data, our model has more parameters. This means well regularization technology should be applied to experiments.

Table 6
DCssCDBM configuration for experiments.

First layer	Input layer, size of input(batchsize,3,32,32) Convolutional layer, size of kernels(36,3,3,3) Pooling layer, size of pooling(2,2)
Second layer	Input layer, size of input(batchsize,36,15,15) Convolutional layer, size of kernels(64,36,3,3), padding=1 Pooling layer, size of pooling(2,2)
Third layer	Input layer, size of input(batchsize,64,7,7) Convolutional layer, size of kernels(100,64,3,3), padding=1 Pooling layer, size of pooling(2,2)

4.3. Process and analysis

In order to verify advancement of DCssCDBM with multi-task learning, here, we did extensive valid experiments on this collected dataset. Here, we used 6000 samples as training data, 1500 samples as valid data, and 4500 samples as test data without data augmentation operation. These experiments can be roughly split into two parts: (1) final system and comparison with other methods; (2) parameter sensitivity analysis. Our experiment environment is equipped with two GTX1080 GPUs desktop under 64-bit Ubuntu16.04 system.

(1) Final system

We normalized the dataset for contrast normalization and applied ZCA whitening. We designed DCssCDBM configuration for final system as Table 6 described. The detailed training algorithm settings were provided as follows: (1) ***Pretraining stage:*** It is extremely necessary to pretrain our DCssCDBM in order to provide good initialization. It can be also viewed as a special regularization, which provides soft prior imposed on trainable parameters. We respectively fixed the size of batch and the step of contrastive divergence to 25 and 10. For two stacked CssCRBMs and DCssCRBM, after well-tuning, we employed the coefficients of

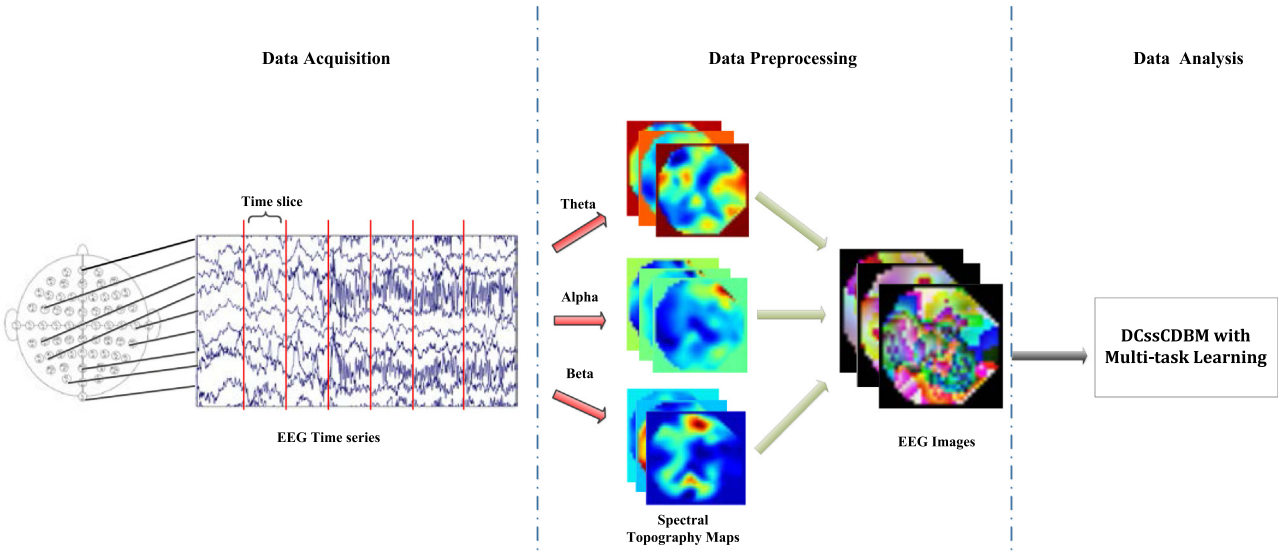


Fig. 5. Overall framework of our method.

the contractive penalty term equal to 0.1 in unity. We employed *Adam* optimizer (Kingma & Ba, 2014) to perform 500 iterations gradient descent according to the procedure described in Table 2. (2) ***Fine-tuning stage and classification:*** Then, DCssCDBM was finely tuned abiding by the process described in Table 3. Here, initial learning rates $\{\lambda, \eta\}$ were equal to 0.01. We adopted 20 fantasy particles in order to evaluate model's expectation and set the step of SAP to 10. The round of Mean-field was set to 5. The coefficients of *auxiliary tasks* $\{c_{ide}, c_{ver}\}$ were set to 0.12, 0.14 respectively. We could calculate the approximate inferences according to Eqs. (2)∼(10) after 100 iterations. **The hyperparameters of DCssCDBM are obtained through well-tuned experiments on valid dataset. Some of these are described in part 4 of this section.**

(2) Peer Competitors

In the experiments, we also conducted several current-state-of-the-art algorithms and obtained promising classification errors on the benchmark in order to collect as the peer competitors of the proposed method. **We give detailed structures and the best parameter settings for these models validated on this dataset in order to be reproducible for results.**

SVM (rbf): We used the flatten EEG images as input for SVM. Detailedly, the best SVM hyperparameters, penalty parameter $C = \{0.01, 0.1, 1, 10, 100\}$, inverse of RBF kernels standard deviation $\gamma = \{0.1, 0.2, \ldots, 10\}$ were obtained through cross-validation on training set under grid-search strategy.

DBN-3: We trained a well-designed DBN stacked by three RBMs on flatten EEG images. The first was a Gaussian-Binary Restricted Boltzmann Machine (GRBM) in order to model continuous value (Larochelle, Courville, et al., 2007) and the other two RBMs were common Binary RBMs. Then, the output of the last hidden layer was fed into a softmax layer in order to predict the class label. We also employed pretraining strategy in order to shift the initial random parameter values toward a good local minimum. We used the following empirically selected numbers of neurons in the three layers that demonstrated good performance: 512, 512, and 128. The last layer was connected to a softmax layer with 3 units. The network was fine-tuned using batch stochastic gradient descent with l_2 weight decay to reduce the overfitting during training. The detailed settings followed by default in Larochelle et al. (2007).

GDBM-2: We used a two-layer GDBM with the same hyper parameter settings literature (Cho, Raiko, & Ilin, 2013) as during training. We also adopted adaptive learning rate, parallel tempering, and enhanced gradient strategies in order to improve performance.

FitNet-10: We trained a teacher network of maxout convolutional layers as reported in Romero, Kahou, et al. (2015) and designed a FitNet with 10 maxout convolutional layers, subsequently connected with a maxout fully-connected layer and a top softmax classification layer. Most of detailed setting as same with the paper (Romero et al., 2015) and we used released codes for this experiment.

PCANet-2: First, we employed PCANet-2 (Chan et al., 2015) with filter size $k_1 = k_2 = 5$, the number of filters $L_1 = 40$; $L_2 = 8$, and 8×8 block size. We also set the overlapping region between blocks to half of the block size, and connected spatial pyramid pooling to the output layer of PCANet. This yields the 21 pooled histogram feature of dimension $L_1{}^*2^{L2}$. We employed PCA to reduce the dimension of each pooled feature to 1280 and used linear SVM classifier in the experiments at last.

Highway-10: In this experiment, we trained basic highway network (Rupesh Kumar Srivastava & Jurgen Schmidhuber, 2015) with 9 hidden layers. Each hidden layer has 1000 neurons. The last hidden layer was connected with softmax layer. We used stronger coefficient $L_2 = 0.2$ regularization for parameters and $p = 0.5$ for dropout.

VGG-A/B/C/D: VGG (Simonyan, 2015) is a well-performed model for image recognition. However, we cannot utilize original version for this task since it would be inevitable to be overfitting. We used four modified VGG versions (Simonyan, 2015) to train EEG data. Table 7 gives details of the configurations of four models. The hyper parameters are same with the original version.

GoogleNet: The original version cannot be applied directly here. We utilized two core inception modules (Szegedy et al., 2015) consisting of three information ways, (1) 1×1 convolution; (2) 1×1 convolution and 3×3 "same" convolution; (3) 2×2 max pooling and 3×3 "same" convolution. Every information pipeline has 32 feature maps, respectively. There is 2×2 average-pooling following each inception module. The hyper parameters are well-tuned by validate dataset.

ResNet-10: The employed residual architecture (Kaiming He, Ren, & Sun, 2016) for this task has three blocks and softmax layer. The basic block consists of 3×3 "same" convolutions, 3×3 "same"

Table 7
Configurations of VGGs.

ConvNet Configuration			
VGG-A	VGG-B	VGG-C	VGG-D
input (32 × 32 RGB image)			
conv3-32(same)	conv3-32(same) LRN	conv3-32(same) conv3-32(same)	conv3-64(same) conv3-64(same)
conv1-192	conv1-192	conv1-192	conv1-192
conv3-32(same)	conv3-32(same)	conv3-32(same)	conv3-64(same)
conv1-192	conv1-192	conv1-192	conv1-192
max pooling (2 × 2)			
conv3-32(same)	conv3-32(same)	conv3-32(same)	conv3-64(same)
conv1-192	conv1-192	conv1-192	conv1-192
max pooling (2 × 2)			
FC-500			
softmax			

Table 8
Diagnostic recognition rate.

Algorithms	Accuracy
SVM (*rbf*)	0.7833
DBN-3	0.842
GDBM-2	0.8566
FitNet-10	0.9193
PCANet-2	0.9244
Highway-10	0.9068
VGG-A	0.9289
VGG-B	0.906
VGG-C	0.8889
VGG-D	0.8735
GoogleNet	0.9118
ResNet-10	0.882
NIN	0.8796
CDBN	0.86
The proposed method	0.9504

convolutions, and 1 × 1 convolutions, following 2 × 2 max pooling. The output of the last block connected with softmax layer directly. This model has 10 trainable layers, called ResNet-10.

CDBN: We trained a three-layer CDBN model (Lee, Grosse, Ranganath, & Ng, 2011) with sparse constraint ($p = 0.001$). It should be noted that we used the same structure as our model described in Table 5. Differently, we used the flatten feature maps as input for SVM (*rbf*). The rule of SVM selection is the same as previously discussed ***SVM (rbf)***.

NIN: The structure used in the experiment keeps consistent with that used in Lin and Yan (2014) on the cifar-10 dataset. The detailed model hyper parameter settings were used from the caffe model zoo during training.

(3) Results Analysis

We report entire results in Table 8 and our method achieves to an accuracy of 95.04% and gains 2.5% augmentation. This is the extensively impressive signal, making a further step for application. Compared to the shallow model, like SVM (*rbf*), deep models have better performances. Taking other deep generative models into consideration, like DBN-3, GDBM-2, and CDBN, our method performs remarkable improvement. The reason may be concluded as follows: (1) more powerful ability of model images, since we modeled images using high-order statistical characteristics, such as covariance structures in the images; (2) more robust feature, because we used a contractive term to constraint parameter space; (3) more useful feature extractions due to taking label information into consideration; (4) multi-task learning framework, because it help avoid overfitting in consideration of limited data. For other deep discriminative models, such as PCANet-2, FitNet-10, Highway-10, GoogleNet, ResNet-10, NIN, these models may still suffer overfitting more or less. In consideration of limited data, strongly hard prior imposed on parameters of deep models may be inappropriate for this task. In these experiments, VGG-A performs better than other designed VGGs model. This demonstrates the theorem of No Free Lunch.

In addition, for a more detailed comparison, we calculate confusion matrix and draw receiver operating characteristic (ROC) curve respectively. In Fig. 6, we exhibit entire confusion matrix in picture form, from which we can see the difficulties lie in two aspects, including distinguishing between MCI and AD, distinguishing between MCI and HC. Most algorithms misclassify them. The proposed method gets 4.6% error rate. This is the lowest error rate among all models. For early diagnosis of AD rate, we extensively pay attention to distinguish between MCI and HC, which is the basis of our beginning of this work. Our approach obtains about 97.11% accuracy in Fig. 6(o). This result indicates that DCssCDBM with multi-task learning could most likely diagnose MCI as soon as possible and further put forward early diagnosis of AD into the application. In Fig. 6(m), NIN misclassifies MCI to HC, which is equal to the proposed model. However, NIN model misclassifies HC to others a lot, which may need additional brain imaging tools such as CT, and FMRI, for confirmation. A similar conclusion can be conducted from other models as well. From extra attention to CDBN, we can see the proposed model surpassed it over diverse classifications. Meanwhile, from most of subfigures, we can see the difficulties of classification between MCI and AD. The misclassifications between MCI and AD occupy major errors. Moreover, we draw entire competitive methods' ROC curves under a one-vs-all scheme in Fig. 7 and plot the micro-average and macro-average ROC curves in the same figure as well. We calculate the area under the curve (AUC) and plot the 95% confidence interval for each ROC. The results indicate that DCssCDBM with multi-task learning achieved the best AUC on all classes. Considering the differences of the statistical significance of the AUC on test dataset, we subsequently perform a statistical analysis after 10 000 bootstraps. We conclude the statistically significant ($p < 0.05$) advanced results of the proposed method over entire competitive approaches according to the analysis of experimental results. At last, the improvement of the proposed method after averaging over three classes is also discovered to be statistically significant ($p \ll 0.05$). These performances are consistent with the related ROC curves in Fig. 7.

The DCssCDBM with multi-task learning in previous section significantly promote early diagnosis of AD rate in comparison with all current-state-of-the-art approaches. It is of equal importance that we explore how this approach achieves such performance. Conventional way is to depict the learned kernels as images for understanding the learned representations by the model. However, we cannot employ this method in DCssCDBM since the model has the small size of kernels (3 × 3), which would not give much intuition about the learned representations. Instead, we adopt feature maps to visualize the models in Fig. 8. Fig. 8 shows three kinds of feature maps, including binary feature maps $\{h^k\}$, slab feature maps $\{s^k\}$, and hybrid feature maps $\{s^k h^k\}$. Generally, as we can see from Fig. 8(a), the first-layer feature maps have a more wide-spread input activation area, while for the second-layer feature maps the activation areas become sparser in Fig. 8(b). Different kernels play the respective role for detection. The significantly responsive activations of feature map in the first layer may be related to corresponding electrodes associated with abnormal AD (Dauwels, Vialatte, & Cichocki, 2010). In the first layer, we may confirm some feature map gripping frontal beta activity (kernel 1#) and another wide-spread theta (kernel 4#). As for the second layer, we also possibly observe detectors

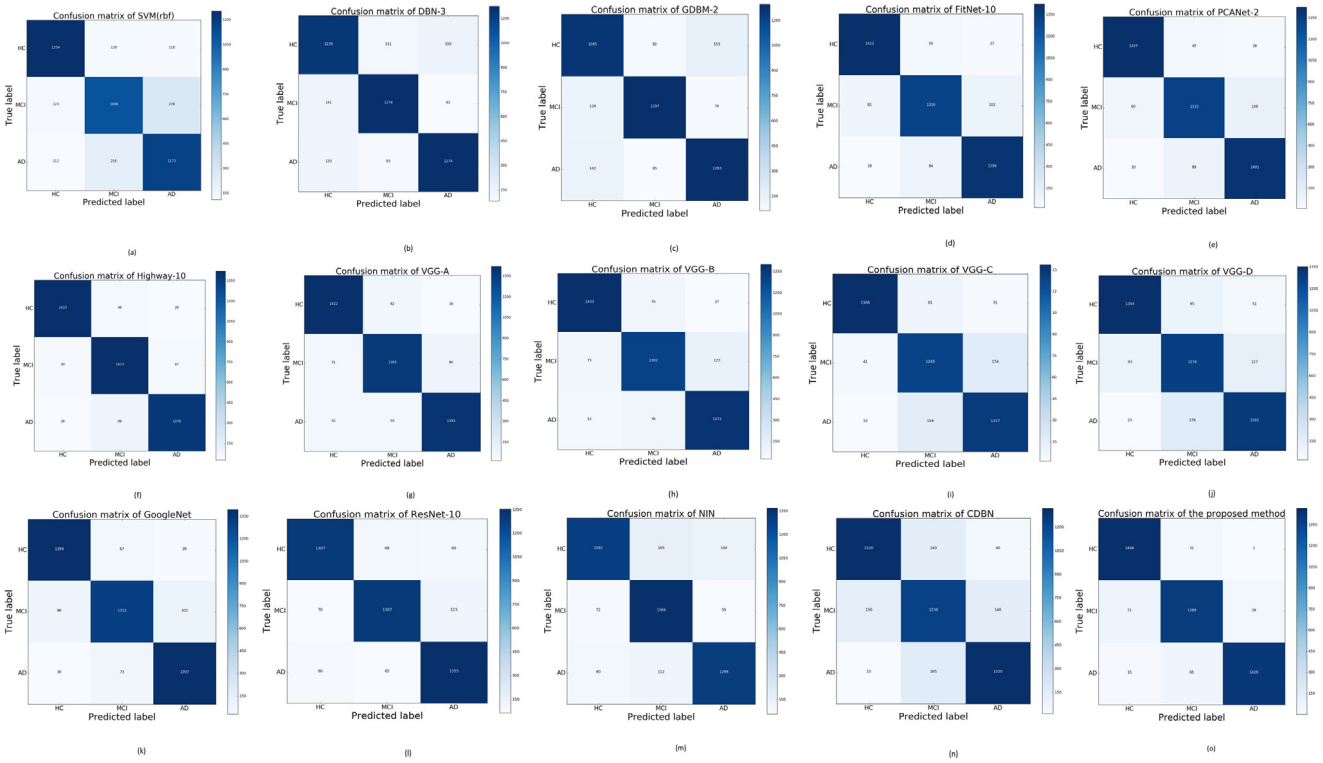


Fig. 6. Confusion Matrix, (a) SVM(rbf); (b) DBN-3; (c) GDBM-2; (d) FitNet-10; (e) PCANet-2; (f) Highway-10; (g) VGG-A; (h) VGG-B; (i) VGG-C; (j) VGG-D; (k) GoogleNet; (l) ResNet-10; (m) NIN; (n) CDBN; (o) The proposed method.

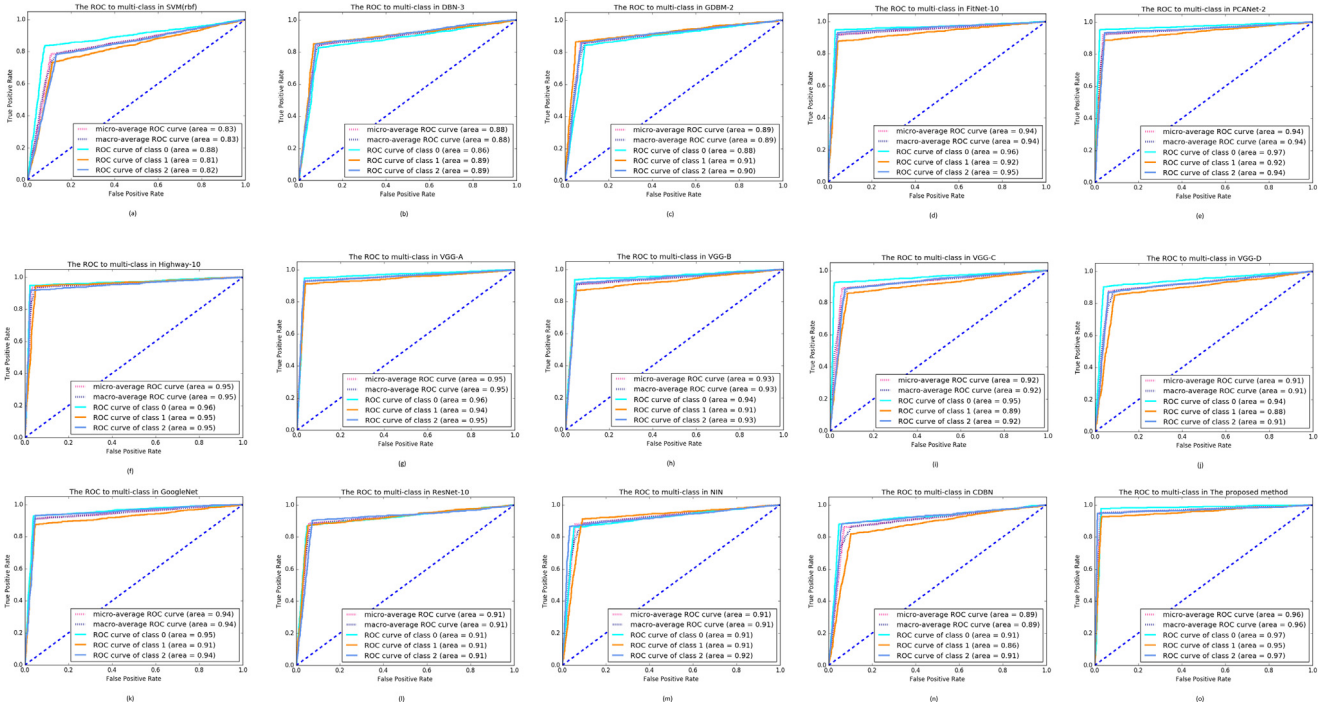
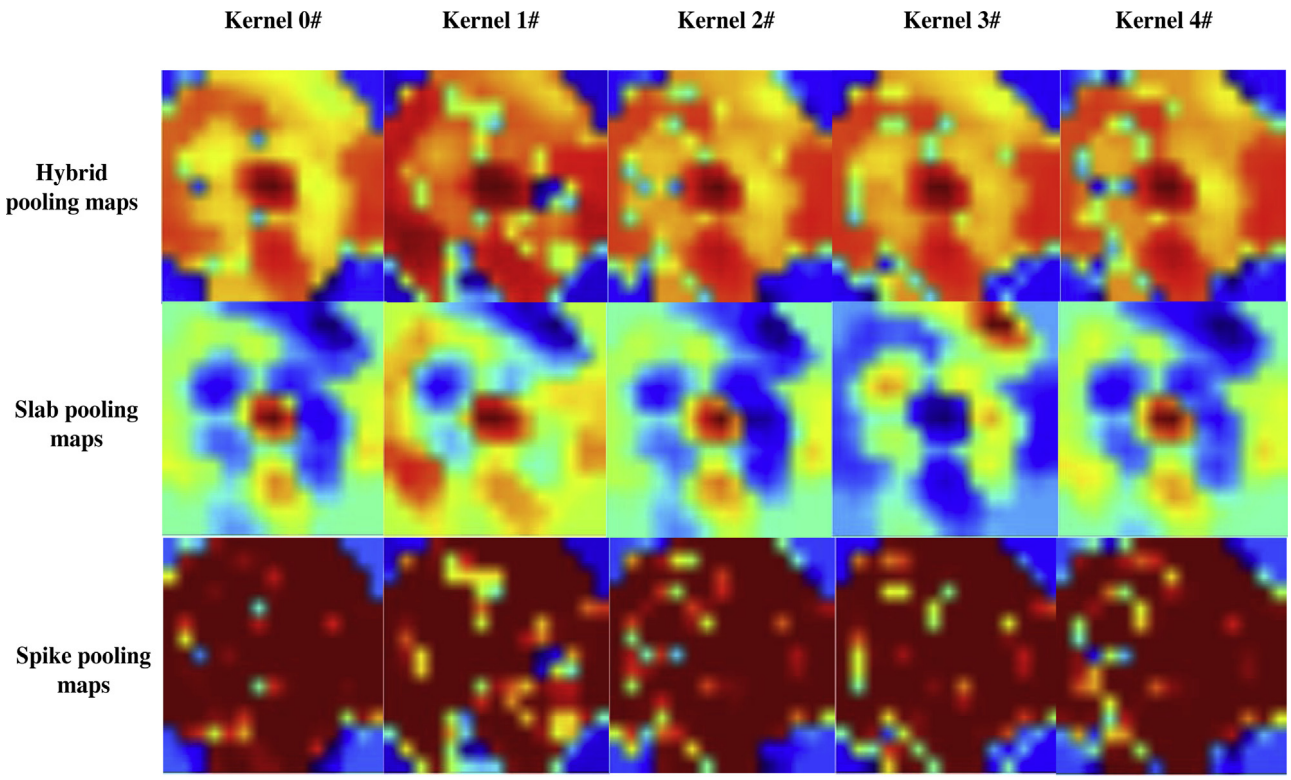
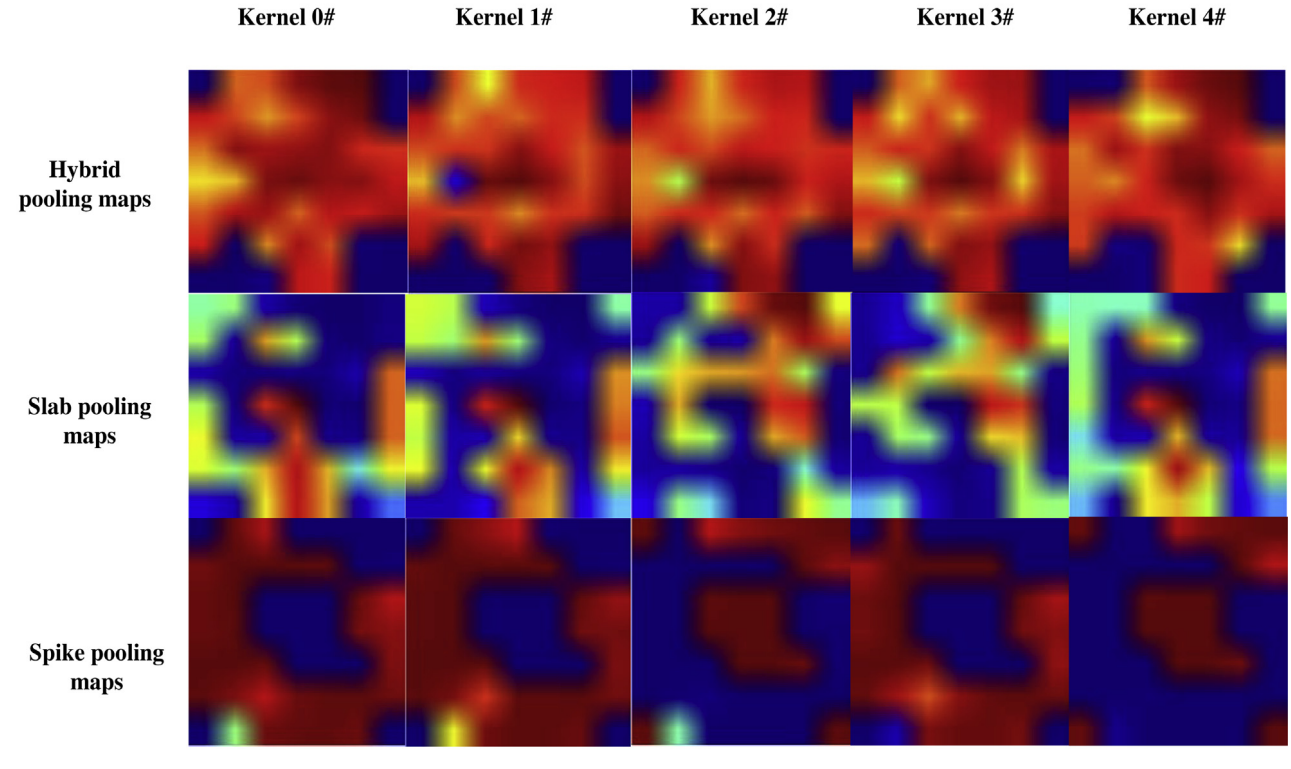


Fig. 7. ROC to multi-class, (a) SVM(rbf); (b) DBN-3; (c) GDBM-2; (d) FitNet-10; (e) PCANet-2; (f) Highway-10; (g) VGG-A; (h) VGG-B; (i) VGG-C; (j) VGG-D; (k) GoogleNet; (l) ResNet-10; (m) NIN; (n) CDBN; (o) The proposed method.



(a)



(b)

Fig. 8. Visualization of feature maps from, (a) the first layer; (b) the second layer.

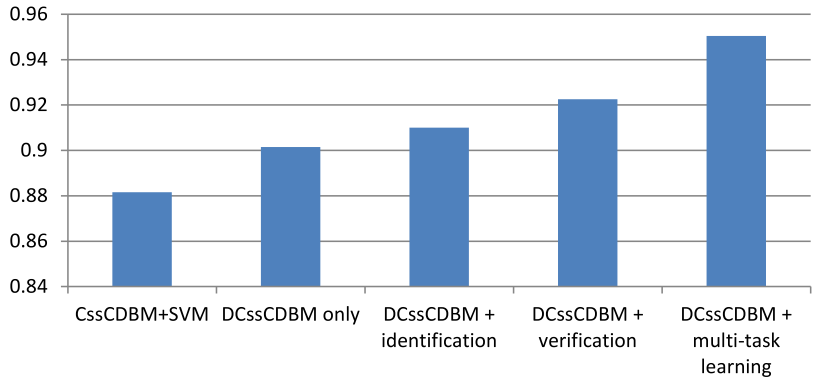


Fig. 9. Influence of multi-task learning to final performance.

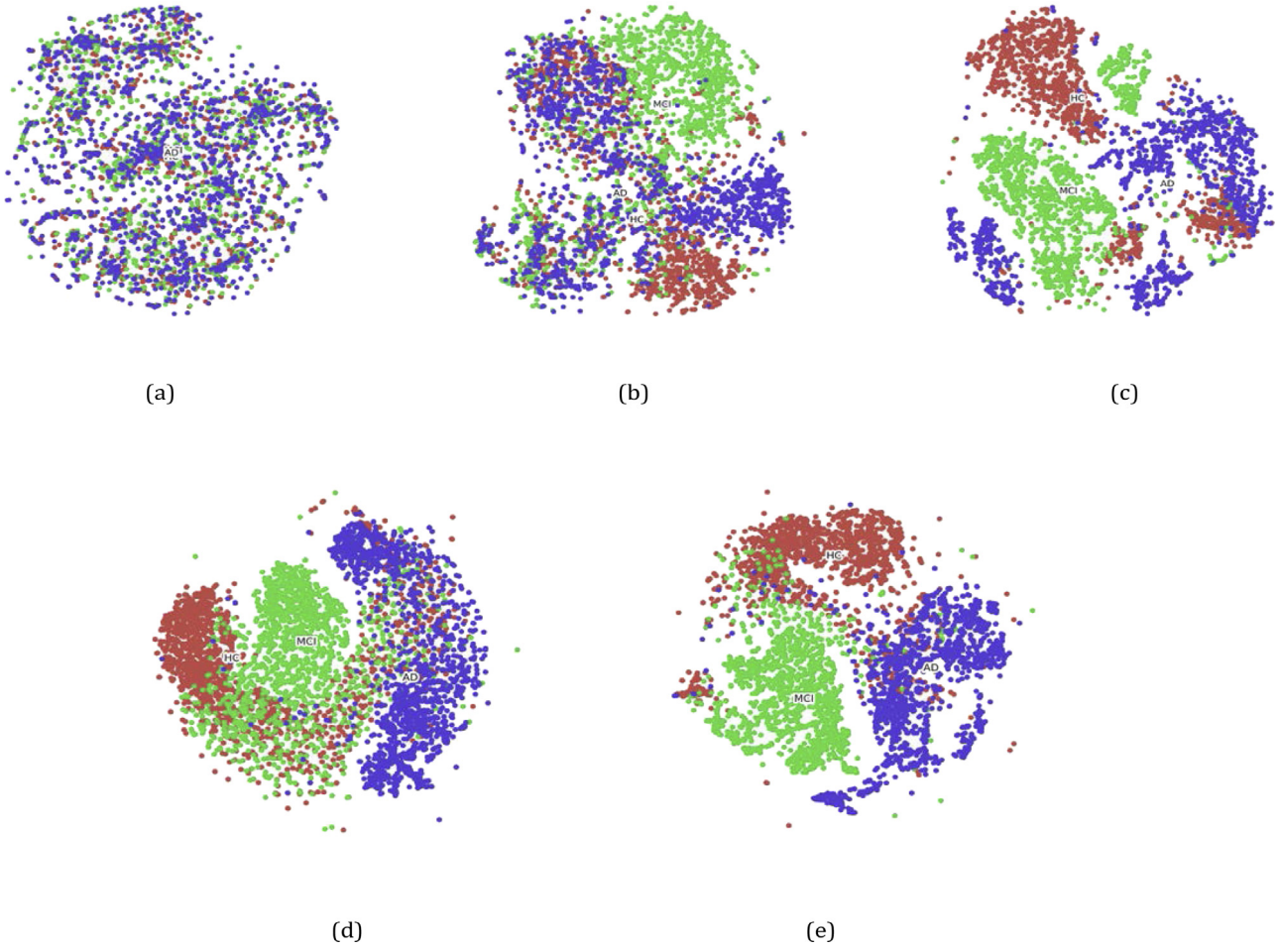


Fig. 10. Embedding representations from (a) original input; (b) the first layer of model with CssCDBM; (c) the third layer of model with CssCDBM; (d) the first layer of proposed model; (e) the third layer of proposed model.

of frontal theta/beta in addition to parietal alpha with increasing focal specificity of feature maps in the second layer. Of course, these findings still need to be further explored. One thing for sure is that these features do promote early diagnosis of AD rate. We can treat the whole model as a black box.

(4) Multi-task Learning and Parameter sensitivity analysis

By diverse experiments, we respectively confirmed the effect of multi-task learning and hyper parameters sensitivity to the final performance. The model with each condition was well-tuned by validate set. Fig. 9 clearly shows the influence of the auxiliary task. In general, DCssCDBM with multi-task learning surpass other methods, which demonstrates this strategy provided effectively soft constraints forced on parameters. DCssCDBM only can get accuracy to 90.15%. This result is better than CDBN and other generative models, but worse than discriminative models. We believe this difference will disappear if we have amounts of unlabeled data for pretraining. We discover an interesting phenomenon that DCssCDBM with verification outperformed DCssCDBM with identification. We argue that EEG images could provide much more information than labels do.

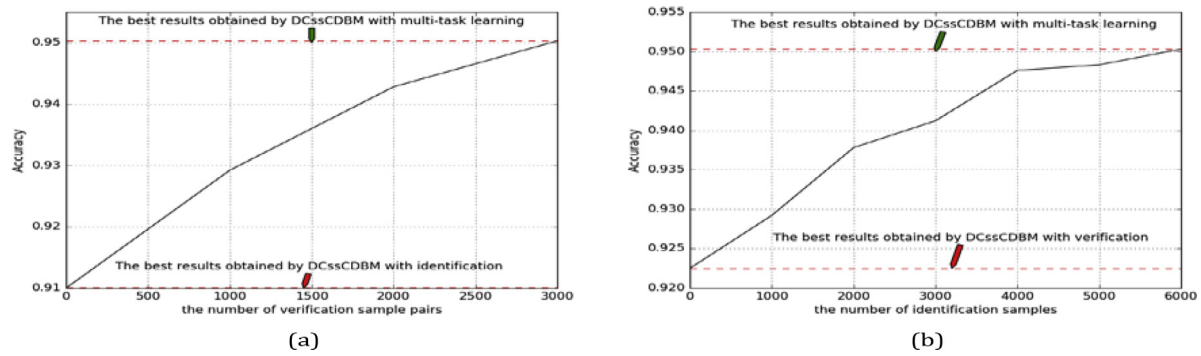


Fig. 11. Investigating the effect of (a) verification signals; (b) identification signals.

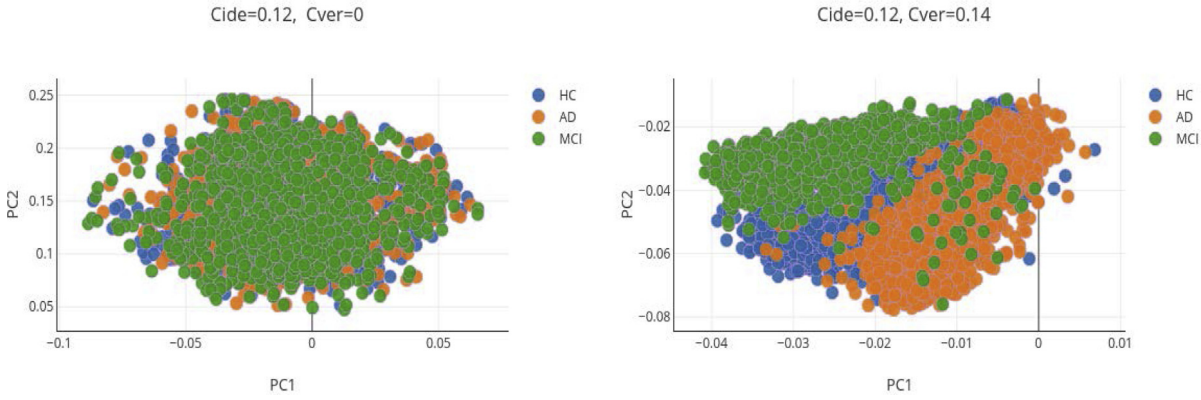


Fig. 12. The first two PCA dimensions of the proposed model with varying c_{ver}.

Most importantly, parameters $\{\lambda, c_{ver}, c_{ide}\}$ are especially significant for our model. They must be appropriate. We will discuss them one by one. In our previous work, CssCDBM has been demonstrated that the probabilistic contractive penalty term could actually improve robust feature extraction a lot than sparse constraints (Bi & Wang, 2018). We discover DCssCDBM furtherly performs well than the original model. We also employ t-SNE (van der Maaten & Hinton, 2008) to embed {3072-dimension, 8100-dimension ($36 \times 15 \times 15$), 576-dimension ($64 \times 3 \times 3$)} representations producing from the DCssCDBM on test dataset (consisting of 4500 samples) into two-dimension space for confirming the robustness to variations qualitatively. Fig. 10(a) shows the embedding result of original 3072-dimension data. Fig. 10(b) and (c) show the embedding result of representation, producing by the first and the third layer of identical-structure CssCDBM after well training, respectively. Fig. 10(d) and (e) depict the corresponding results of DCssCDBM. In comparison with Fig. 10 (a), Fig. 10(d) indicates that the representations learned from DCssCDBM are more well-separated than original data and DCssCDBM works properly. In further comparison with Fig. 10(b), Fig. 10 (d) deeply shows the advancement of robust feature extractions via DCssCDBM, which is more likely to extract more robust features. We can obtain the similar conclusion via comparing (c) with (e) in Fig. 10. From (d) to (e) in Fig. 10, features are more likely robust with the increasing depth of layer. In Rifai et al. (2011), this is explained that the higher-level features would be more invariant in their feature-specific directions of invariance. We select contractive term coefficient $\lambda = 0.1$ according to the rules in Bi and Wang (2018). Indeed, the improvement of robustness to variations may owe to introducing label layer, which give explicit signals to constraint inter-class variations, which help to produce well-separated features. This can also be explained that another regularization term associated with labels is further applied into original probabilistic contractive penalty term.

At present, we respectively consider how verification and identification supervisory signal influences feature extraction. Particularly, we conduct this experiment with a gradually increasing number of verification pairs during training from 0 to 3000, while the identification signal is generated from all the 6000 training samples all the time. Fig. 11(a) shows how early diagnosis of AD accuracies of the proposed model vary on the test set with the number of verification pairs used in the verification signal. It shows that enlarging a large number of verification pairs is crucial to extract effective EEG representations. This observation is consistent with those in the previous section. The increasing number of verification pairs provides richer information and boosts to form representations with diverse intra-class variations, making the class centers of different patients more distinguishable. In Fig. 12, we exhibit the first two PCA dimensions of the last-layer feature maps based on our model with identification task only under the condition where varying c_{ver} from two bounds. When the identification signal is just used ($c_{ver} = 0$), the extracted representations include both many inter and intra-class variations, as Fig. 12(a) depicted by the long tails of the green curves in both figures. While diverse inter-class variations boost to classify different patients, large and diverse intra-class variations are noises and make early diagnosis of AD difficult. When both the auxiliary tasks, identification and verification based on EEG spectrum images, are added into the main task with appropriate weight value ($c_{ver} = 0.14$, $c_{ide} = 0.12$) as Fig. 12(b) exhibited, the inter-class variations still keep more diverse without large changes while the variations in the main directions become larger than before to some degree. At the same time, the intra-class variations decrease in both the diversity

and magnitude. Therefore, both the beneficial changes of inter- and intra-class variations in a direction make early diagnosis of AD easier. Fig. 11(b) depicts the effect of early diagnosis of AD accuracies of the proposed model vary on the test set with the number of verification pairs used in the verification signal. We could conclude similar summary.

5. Discussion

We explain why our method is well-performed from two perspectives. The one is based on the advanced discriminative generative model and the other is multi-task learning regularization.

(1) ***Model***: Our latest work, CssCDBM, which is well-performed model for feature extraction, still need combine representation learning and classification task for the improvement of final classification performance. We argued that DCssCDBM can and should be used as stand-alone non-linear classifiers alongside other standard and more popular classifiers, instead of merely being considered as simple feature extractors. We induced label layer to CssCDBM resulting in DCssCDBM, which integrates clear label information to guide feature extraction useful to classify. Indeed, DCssCDBM works well on account of bridging the connection between feature extraction and feature selection. Indeed, CssCDBM may not extract all features suit for classification. In Fig. 9, it demonstrates DCssCDBM compares favorably with CssCDBM with SVM for early diagnosis of AD. Similar discussions can be found in DisRBM.

(2) ***Multi-task learning***: Multi-task learning can be generally considered as a soft constraint approach imposed on the parameters, which has been proved a way to improve generalization by pooling the examples arising out of several tasks. Identification and verification are implicit tasks behind the early diagnosis of AD. Conventional approach to identification and verification may pay attention to face image. In paper (Sun & Tang, 2014), Sun proposed DeepID2 based on multi-task learning to extract face representation by joint identification–verification. Indeed, similar to face image, EEG can be also seen as biometrics of identification and verification and has been extensively attracted much insight (Abo-Zahhad, Ahmed, & Abbas, 2016; Gui, Jin, & Xu, 2014; Kumari & Vaish, 2015; Rodrigues, Silva, Papa, Marana, & Yang, 2016). However, our approach differs from them in EEG signal processing. We conveyed the multi-channel EEG signal measurements with distribution information of brain into a 2-D image to utilize the spatial structure and used multiple color channels to represent the multiple spectral band information. Then we further employed the advantage of discriminative generative convolutional model. In addition, most importantly, employing identification–verification auxiliary tasks is first useful attempt for early diagnosis of AD.

6. Conclusion

In this paper, our main contribution is firstly to propose an arguably advanced discriminative deep probabilistic model with multi-task learning to classify EEG spectrum image into three classes for early diagnosis of AD. Our approach includes two parts, an advanced discriminative deep convolutional generative model and multi-task learning strategy. The designed model is well-performed in comparison to other generative model since it bridges the connection between feature extraction and classification. Another crucial point lies in overcoming overfitting by multi-task learning. Identification and verification task based on EEG image further constraint parameters by increasing the inter-subject variations and reducing the intra-subject variations respectively. This strategy is better than prior based regularization methods. The results of the proposed method outperformed several advanced models. It is the first time to employ the regularization and combine them based on EEG spectrum images instead of other data form. This demonstrated DCssCDBM with multi-task learning further develops early diagnosis of AD in Home Care System. It also can be applied computer aid system. In the future, we will explore the discriminant version of DCssCDBM to see whether it behaves much better and extend it to early diagnosis of diverse diseases, such as epilepsy detection.

Acknowledgments

Our contribution was supported by the National Natural Science Foundation of China (Grant No. 51779050) and the International S&T Cooperation Program of China (Grant No. 2015DFG12150) named as "Key Technology Elements and Demonstrator for Cloud-Assisted, Wireless Networked Ambulatory Supervision (C-Nurse)". Thanks for home cooperative partner, Beijing Easy monitor Technology, which has deeply participated in the International S&T Cooperation Program of China program and provided available data. Thanks for overseas cooperative partner, Professor Huang, the head of Integrated Systems Laboratory and director of studies at the Department of Information Technology and Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.neunet.2019.02.005.

References

AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., et al. (2016). Face recognition using deep multi-pose representations. In *Ieee Wint Conf Appl*.

Abo-Zahhad, M., Ahmed, S. M., & Abbas, S. N. (2016). A new multi-level approach to EEG based human authentication using eye blinking. *Pattern Recognition Letters*, *82*, 216–225.

Atyabi, A., Luerssen, M., Fitzgibbon, S. P., & Powers, D. M. W. (2012). Adapting subject-independent task-specific EEG feature masks using PSO. *Ieee C Evolutionary Computation*.

Bi, X. J., & Wang, H. B. (2018). Contractive slab and spike convolutional deep Boltzmann machine. *Neurocomputing*, *290*, 208–228.

Capecci, E., Doborjeh, Z. G., Mammone, N., La Foresta, F., Morabito, F. C., & Kasabov, N. (2016). Longitudinal study of Alzheimer's disease degeneration through EEG data analysis with a neucube spiking neural network model. In *2016 international joint conference on neural networks* (pp. 1360–1366).

Chan, T. H., Jia, K., Gao, S. H., Lu, J. W., Zeng, Z. N., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification? *IEEE Transactions on Image Processing*, *24*, 5017–5032.

Chen, F., Yu, H. M., Hu, R., & Zeng, X. X. (2013). Deep learning shape priors for object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1870–1877). IEEE.

Cho, K., Raiko, T., & Ilin, A. (2013). Gaussian-Bernoulli deep Boltzmann machine. In *2013 international joint conference on neural networks*.

Cummings, J. L. (1993). Mini-mental state examination. *Jama the Journal of the American Medical Association*, *269*(18), 2420–2421.

Dauwels, J., Vialatte, F., & Cichocki, A. (2010). Diagnosis of Alzheimer's disease from EEG signals: Where are we standing? *Current Alzheimer Research*, *7*, 487–505.

Dubois, F. H. H., Jacova, C., et al. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurology*, *6*(8), 734–746.

Gui, Q., Jin, Z. P., & Xu, W. Y. (2014). Exploring EEG-based biometrics for user identification and authentication. In *Ieee Sig Proc Med*.

Kaiming He, X. Z., Ren, Shaoqing, & Sun, Jian (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *Computer Science*.

Kumari, P., & Vaish, A. (2015). Brainwave based user identification system: A pilot study in robotics environment. *Robotics and Autonomous Systems*, *65*, 15–23.

Larochelle, B. Y. (2008). Classification using discriminative restricted Boltzmann machines. In *Proceedings of international conference on machine learning* (pp. 536–543).

Larochelle, E. D., Courville, A., et al. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on machine learning* (pp. 473–480).

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2011). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, *54*, 95–103.

Leng, B., Zhang, X., Yao, M., & Xiong, Z. (2015). A 3d model recognition mechanism based on deep Boltzmann machines. *Neurocomputing*, *151*, 593–602.

Li, X. Y., Guan, C. T., Zhang, H. H., & Ang, K. K. (2017). Discriminative ocular artifact correction for feature learning in EEG analysis. *Ieee Transactions on Bio-Med Engineering*, *64*, 1906–1913.

Lin, Q. C. M., & Yan, S. (2014). Network in network. In *Proceeding of international conference on learning representations*.

Liu, F., Wee, C. Y., Chen, H. F., & Shen, D. G. (2014). Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's disease and mild cognitive impairment identification. *Neuroimage*, *84*, 466–475.

Mehmood, R. M., Du, R. Y., & Lee, H. J. (2017). Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors. *Ieee Access*, *5*, 14797–14806.

Mei, J. Y., Liu, M. Z., Karimi, H. R., & Gao, H. J. (2014). Logdet divergence-based metric learning with triplet constraints and its applications. *Ieee Transactions on Image Processing*, *23*, 4920–4931.

Morabito, F. C., Campolo, M., Ieracitano, C., Ebadi, J. M., Bonanno, L., Bramanti, A., et al. (2016). Deep convolutional neural networks for classification of mild cognitive impaired and Alzheimer's disease patients from scalp EEG recordings. In *2016 Ieee 2nd international forum on research and technologies for society and industry leveraging a better tomorrow* (pp. 162–167).

Morabito, F. C., Labate, D., Bramanti, A., La Foresta, F., Morabito, G., Palamara, I., et al. (2013). Enhanced compressibility of EEG signal in Alzheimer's disease patients. *Ieee Sensors Journal*, *13*, 3255–3262.

O'Keeffe, J., Carlson, B., DeStefano, L., Wenger, M., Craft, M., Hershey, L., et al. (2017). EEG fluctuations of wake and sleep in mild cognitive impairment. In *Conf Proc IEEE Eng Med Biol Soc (Vol. 2017)* (pp. 3612–3615).

Omedes, J., Iturrate, I., Montesano, L., & Minguez, J. (2013). Using frequency-domain features for the generalization of EEG error-related potentials among different tasks. In *2013 35th annual international conference of the ieee engineering in medicine and biology society* (pp. 5263–526).

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, *56*, 303–311.

Pouya Bashivan, I. R., Yeasin, Mohammed, & Codella, Noel (2016). Learning representations from eeg with deep recurrent convolutional neural networks. In *Proceedings of the international conference on learning representations*.

Rifai, V. P., Muller, X., et al. (2011). Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on machine learning* (pp. 833–840).

Rodrigues, D., Silva, G. F. A., Papa, J. P., Marana, A. N., & Yang, X. S. (2016). EEG-Based person identification through binary flower pollination algorithm. *Expert Systems with Applications*, *62*, 81–90.

Romero, B. N., Kahou, S. E., et al. (2015). FitNets: Hints for thin deep nets. In *Proceedings of the international conference on learning representations*.

Rupesh Kumar Srivastava, K. G., & Jurgen Schmidhuber, J. (2015). Highway networks. arXiv preprint arXiv:1505.00387.

Salakhutdinov, R., & Hinton, G. (2012). An efficient learning procedure for deep Boltzmann machines. *Neural Computing*, *24*, 1967–2006.

Simonyan, Z. A. K. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the international conference on learning representations*.

Song, J. L., & Zhang, R. (2016). Automatic seizure detection using a novel EEG feature based on nonlinear complexity. In *2016 international joint conference on neural networks* (pp. 1686–1695).

Sun, W. X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. In *Proceedings of the advances in neural information processing systems* (pp. 1988–1996).

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9). IEEE.

Temko, A., Nadeu, C., Marnane, W., Boylan, G. B., & Lightbody, G. (2011). EEG Signal description with spectral-envelope-based speech recognition features for detection of neonatal seizures. *Ieee Transactions on Information Technology B*, *15*, 839–847.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Zhang, Z. P., Luo, P., Loy, C. C., & Tang, X. O. (2014). Facial landmark detection by deep multi-task learning. *Computer Vision*, *8694*, 94–108, Eccv 2014, Pt Vi.

Zhang, D. Q., Shen, D. G., & Neuroimaging, A. s. D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage*, *59*, 895–907.

Zhang, Y., & Yeung, D. Y. (2010). Multi-task warped gaussian process for personalized age estimation. In *Proc Cvpr Ieee* (pp. 2622–2629).

Zhou, J. Y., Liu, J., Narayan, V. A., Ye, J. P., & Initi, A. s. D. N. (2013). Modeling disease progression via multi-task learning. *Neuroimage*, *78*, 233–248.

Zhou Li, H.-I. S., & Shen, Dinggang (2016). Sparse multi-response tensor regression for Alzheimer's disease study with multivariate clinical assessments. *IEEE Transactions on Medical Imaging*, *35*, 1927–1936.