A Novel Semi-supervised Deep Learning Framework for Affective State Recognition on EEG Signals

Xiaowei Jia, Kang Li, Xiaoyi Li and Aidong Zhang
School of Computer Science and Engineering
State University of New York at Buffalo, Buffalo, NY, USA, 14260-1660
Email: {xiaoweij,kli22,xiaoyili,azhang}@buffalo.edu

***Abstract*—Nowadays the rapid development in the area of human-computer interaction has given birth to a growing interest on detecting different affective states through smart devices. By using the modern sensor equipment, we can easily collect electroencephalogram (EEG) signals, which capture the information from central nervous system and are closely related with our brain activities. Through the training on EEG signals, we can make reasonable analysis on people's affection, which is very promising in various areas. Unfortunately, the special properties of EEG dataset have brought difficulties for conventional machine learning methods. The main reasons lie in two aspects: the small set of labeled samples and the noisy channel problem. To overcome these difficulties and successfully identify the affective states, we come up with a novel semi-supervised deep structured framework. Compared with previous deep learning models, our method is more adapted to the EEG classification problem. We first adopt a two-level procedure, which involves both supervised label information and unsupervised structure information to jointly make decision on channel selection. And then, we add a generative Restricted Boltzmann Machine (RBM) model for the classification task, and use the training objectives of generative learning and unsupervised learning to jointly regularize the discriminative training. Finally, we extend it to the active learning scenario, which solves the costly labeling problem. The experiments conducted on real EEG dataset have shown both the convincing result on critical channel selection and the superiority of our method over multiple baselines for the affective state recognition.**

I. Introduction

With the development of cyber-physical systems and the rising interest for brain-computer interaction, the need for detecting the affective state through the human-machine interaction is ever growing [1], [2]. Recent advance in smart sensors provides the possibility to let people comfortably be equipped with machines, which facilitates better analysis and understanding of human affection.

Various kinds of studies have been conducted on human affective state recognition. The first approach focuses on the audio-visual features like facial expressions and speech [3], [4]. Even if this method brings little discomfort to the user, it may introduce lots of artifacts, which severely impact the learning process. The second method involves physiological signals like electrocardiogram (ECG) [5], skin conductance (SC) [6], etc. In addition to these periphery physiological features, recently the advance of devices and processing system enables global assessment of EEG signals, which are captured from central nervous systems. Furthermore, the information from EEG signals have been proved [7] to reveal significant characteristics regarding different affective states. The task of affective state recognition has extensive prospects for many real applications. For instance, through the real-time recognition, the consultants can make adjustment with proper topics to improve the quality of service [8]. Moreover, it can be used as the means of therapy, where doctors can detect the abnormal affective perturbation and take actions accordingly.

When it comes to the data collection, multiple electrodes are usually attached to the user's scalp in EEG assessment devices. The signals captured in each channel will record the voltage change between a pair of adjacent electrodes. The information contained in these channels, however, can be very noisy due to a variety of artifacts, which stem from different kinds of events, such as the discomfort brought by the device, the interference from outside world, or just the sudden mood change of the participant. Given the difficulty to control these factors, the irrelevant noise they brought often severely degrades the performance of conventional classification methods. To conquer the problem, we start with eliminating the redundant EEG channels. In biology, brain related activities are usually dominated by several specific areas, thus there exist several captured regions that is incoherent to emotion. Hence we can remove the irrelevant channels based on their significance to reduce the data dimensionality. In this paper we provide two reasonable options to measure the significance of each channel. The first one is based on the reconstructed error of feature extraction model, and the other is based on the distance between the extracted features from the nonactive input and the randomly features. We name this first stage as the rough channel selection procedure.

On the other hand, it is shown in recent studies [9] that different affective states are usually reflected by different scalp regions. Therefore we further propose a finer-grained channel selection to extract channels related with each affective state. In effect, these two stages of channel selection are complementary to each other, as to be elaborated later. By combining these two selection procedures we propose a novel two-level channel selection method, which can be easily integrated into our deep structured model.

After the feature selection, we adopt the deep belief network (DBN) [10] based model to handle the affective state classification problem. DBN basically consists of the stacked RBMs to extract high-level and representative features from input data. Nowadays with the rapid development of the smart sensors, the EEG data can be readily obtained by using the light and wearable devices. Despite the convenience of data acquisition, the data labeling procedure requires both professional knowledge and lots of manual efforts. Therefore in most

cases only a small set of labeled samples is available, while the majority of whole dataset is left unlabeled. For this reason the conventional learning methods that utilize only supervised information will result in severe overfitting. Hence we propose a novel semi-supervised deep structured learning model which leverages both labeled and unlabeled information. Different from the traditional DBN [10] with separate unsupervised and supervised stages, our model leverages label information in feature extraction and integrates unlabeled information to regularize the supervised training. In this way our method jointly utilizes both supervised and unsupervised information during the entire training process to reduce model variance.

Based on the model, we further propose an active learning [12] technique to make the most of our learning resources. The basic idea is to utilize the uncertainty of trained model over each unlabeled sample to decide the sample's potential contribution to the model training. The selected informative samples can guide the model to progress faster towards the optimal direction. This work will throw light on the costly labeling problem. Finally, to demonstrate the effectiveness, we conduct the experiments on the DEAP Dataset [13], where the original EEG signals are downsampled (to 128Hz) and segmented to form the input feature vectors. The results obtained greatly surpass the performance of multiple baselines and meanwhile remarkably support the effectiveness in channel selection.

II. Related Work

In recent years there has been a growing interest in detecting affective state through the human-machine interaction based applications. So far there have been several existing methods on this problem [14]–[17]. However, these works are not designated to address the special characteristics of EEG classification. At first, in [14]–[16], the noisy channel problem is not well addressed. Moreover, the limited labeled samples will severely impact the learning performance of traditional models, as used in [14], [15], [17]. For instance, in [17], the significant channels are selected based on the Fisher Criterion. However, the lack of labeled samples will greatly limit the channel selection performance, thus resulting in the mistaken removal of some critical channels. In our proposed method, we use RBM-based model to extract representative features and to reduce the data dimensionality, which greatly diminishes the impact from the scarcity of labeled data. In addition, we first adopt a rough channel selection procedure that only utilizes unsupervised information, which aims at removing irrelevant channels with little structure information and reducing data dimensionality. Based on the results from the rough selection, we propose the affection-based finer-grained channel selection procedure to handle the noisy channel problem.

Besides channel issues, we propose the semi-supervised model to overcome the scarcity of labeled data. Semi-supervised learning [11] is well known to be very effective in cases where the dataset is mixed with labeled and unlabeled data, and it has been widely used in multiple areas [18]–[21]. Unfortunately, we cannot directly use the traditional semi-supervised learning framework to solve the EEG classification problem, due to the high dimensionality of input data compared with the size of training dataset. Also the traditional feature extraction method cannot fully capture the critical factors. Hence we propose to use the semi-supervised method based on the deep learning model [22], in which the abstract, high-level features can be extracted through the consecutive training over multiple layers. The higher the layer, the more representative features it can extract from the original data.

TABLE I. Table of Notation

Notation	Meaning
$Xl = \{Xl_1, Xl_2, ..., Xl_n\}$	the labeled samples
$Y = \{Y_1, Y_2, ..., Y_n\}$	the corresponding labels for Xl
$Xu = \{Xu_1, Xu_2, ..., Xu_m\}$	the unlabeled samples
V	visible units in RBM
H	hidden units in RBM
W	weight matrix in RBM
U	weight matrix for label vector
B	bias vector for features
C	bias vector for H
D	bias vector for label vector
X	feature units in generative RBM
S	the total number of affective states
$\mathbf{e}_y$	1-out-of S vector, y^{th}element is 1
T	the total number of channels
C	the specified set of labeled samples
$k = 1, ..., T$	the index for channels
$y = 1, ..., S$	the index for affective states

There have been several existing works that apply deep learning based model in EEG classification problem. For instance, in [23], [24] DBN-based method is adopted as the reconstructor for anomaly detection in EEG signals. In [25], DBN is used for feature extraction and a generative RBM acts as the final layer for supervised training. The major drawback of these methods is that they have separate unsupervised training and supervised training phases. Even if such learning structure can utilize unlabeled data information during the unsupervised training phase, the separate final supervised phase can still be affected by the lack of labeled samples. On the other hand, without the label information, the features extracted from the unsupervised training procedure are not always reliable. Hence, in this paper we propose the deep structure that utilizes label information to guide the channel selection procedure, and regularizes the supervised training with unlabeled data and generative information in the final classification layer.

III. Deep Structured Learning Approach

In this section, we will first briefly describe how RBM works in feature extraction and classification, based on which we will propose our deep structure based learning approach. To formalize the problem, we first introduce the notation followed by this paper. We represent the matrices and vectors using the upper-case letters such as X, Y, and Z. As for a matrix D, $D_{i,j}$ denotes the $(i,j)^{th}$ element of matrix D. $D_{i.}$ and $D_{.j}$ denote the i^{th} row and j^{th} column of matrix D, respectively. We use lower-case letters such as d to represent indices or scaler values. Some notation followed by this paper is listed in Table I.

A. Restricted Boltzmann Machine

The scarcity of labeled data usually causes severe small sample problem, as each sample contains thousands of fea-

tures. To overcome the overfitting problem resulted from direct training on such datasets, we need to fully utilize the unsupervised information. In our model, the unsupervised information first assists in feature extraction, and then provides the constraints in the semi-supervised training. In this part we will discuss the RBM-based feature extraction with unsupervised information.

RBM is a restricted version of Markov Random Field. To handle the small sample problem, RBM-based deep structure aims at extracting high-level features to represent the latent characteristics, and minimizing the information loss. As for the affective state recognition, the training process conducted on the high-level features with fewer dimensions will alleviate the small sample problem. In this paper, we propose a revised DBN-based learning model which better fits this task. Before the exposition of our method, we will briefly introduce the principle of RBM in feature extraction.

RBM contains two layers of variables, V and H. V represents the set of visible units (input features), and H represents the set of hidden units (hidden features), which jointly forms a fully connected bipartite graph. The model describes the distribution of (V, H), which is defined as:

$$\begin{aligned} P(V,H) &= \frac{exp(-E(V,H))}{Z}, \\ Z &= \sum_{V,H} exp(-E(V,H)), \end{aligned} \tag{1}$$

where the energy function $E(V, H)$ is defined to be:

$$E(V,H) = -V^T W H - B^T V - C^T H. \tag{2}$$

In the above equation, W denotes the weights between V and H. Specifically, $W_{i,j}$ represents the weight between V_j and H_i, and B,C denote the biases for visible units and hidden units, respectively. The denominator serves as the normalizer for the probability distribution.

RBM is a degenerate case of Markov Random Field, which has no interconnection between units in the same layer. In this case the conditional distribution of each unit is independent of others in the same layer. Hence the $P(V|H)$ and $P(H|V)$ are fully factorial and given by:

$$\begin{aligned} P(H_i|V) &= sigm(W_{i.}V + C_i), \\ P(V_j|H) &= sigm(W_{.j}^T H + B_j), \end{aligned} \tag{3}$$

where $sigm(x) = (1 + exp(-x))^{-1}$ is the logistic sigmoid function.

From equation 1, the gradient of model parameters $\theta = \{W, B, C\}$ used for updating can be computed as follows:

$$\frac{\partial P(V,H)}{\partial \theta} = -\mathbb{E}_{H|V}[\frac{\partial}{\partial \theta} E(V,H)] + \mathbb{E}_{V,H}[\frac{\partial}{\partial \theta} E(V,H)], \tag{4}$$

where the first term on the right side of equation represents the expectation over the data, which can be computed by Equation 3. The second term stands for the expectation over the model distribution, which is derived from the term Z in Equation 1 and cannot be computed efficiently. Due to the intractability brought by the existence of Z during training, Hinton [26] proposed the Contrastive Divergence method to address the issue, which performs K steps of alternating Gibbs sampling. Usually in experiment we set the value of K to be 1, i.e., to use only one time iteration of Gibbs sampling to approximate the model expectation. This has shown to be effective enough for RBM training [26].

Due to the capacity of each RBM in feature extraction, we usually adopt the structure with stacked RBMs in deep learning model. Specifically, once we finish the training of one RBM, we will use the output of current RBM as the input for the next one, and start a new training phase. Given such structure, the RBM in higher level can extract more representative features.

B. Two-level Channel Selection Method

Channel selection problem is in effect a special case of feature extraction. Different from direct feature extraction by RBM, we are provided with very useful prior knowledge to determine the critical features by the partition of channels. In the affective state recognition, there exist many irrelevant channels which bring noise to the classification. Furthermore, as revealed in current studies [9], the different affective states may be reflected in different scalp regions. Due to these considerations, we come up with a two-level channel selection method, as shown in Figure 1 (a). This structure of channel selection can be easily integrated into the whole proposed deep learning framework.

Given the input data, we implement the channel selection procedure in two stages. At the first stage, we train different RBMs for each channel only using unsupervised information and roughly determine the relevant channels for the classification. Specifically we have two strategies to implement this procedure. The first strategy is to measure the RBM reconstructed error. The lower error will reflect the capacity of the model to successfully capture the structure of data distribution in the corresponding channel. The second strategy involves using zero-stimulus method mentioned in [25], where the all-zero features are used as input to the trained stacked RBM model. After this we measure the response, which is defined to be the distance between the value of resulted hidden units and random hidden units. With either strategy adopted, this step is based on the fact that the input data with little unsupervised structure information, thus irrelevant to the recognition, will randomly update the model parameters. In our experiment we follow the second method. To notice, here we choose a relatively large number u_1 of selected channels from the ranking list of corresponding measurement, in case we remove any potentially meaningful channels. After this, extracted features from the selected channels are collected and used as the input for the second level finer-grained channel selection procedure.

As revealed by recent studies [9], the regions that most significantly reflect each affective state are different. Hence we propose a finer-grained affection-based channel selection method. Given the output from the first level, we still train different RBMs for each selected channel. After training, we compute the extracted features of labeled samples in each class. Then we calculate the proposed significance measurement of each channel k, regarding samples from certain class C:

$$ChanSig_{k,C} = \frac{\sum_{i \in C} \|h(i) - \overline{h}\|^2}{\sum_{i \in C} \|h(i) - \overline{h_C}\|^2}, \tag{5}$$

where $h(i)$ represents the extracted feature of i^{th} sample, or the output of stacked RBMs, $\overline{h}$ is the mean vector of extracted features for all labeled samples, and $\overline{h_C}$ is the mean vector of extracted features for all samples from class C. Intuitively, the nominator represents the strength of response given the input data in the specified class. The higher the value the larger the distance of response between the data in class C and the whole dataset. On the other hand, the denominator measures the distance of response between samples in class C and the centroid of class C, which stands for the inner-class variance. Hence, the high ratio of $ChanSig_{k,C}$ displays the salience of the samples from class C in channel k.

After the calculation, we rank the value of all channels, and select top u_2 channels from each state to represent the significant characteristics of the corresponding affection. To simplify, we extract same amount of channels from all states. Assume we have totally S states, after second level we will obtain the features from $u_2 \times S$ channels.

After u_2 channels are extracted for each state, we add a consensus layer on top of previous two-level structure to combine the channels from different classes. Specifically, the outputs from $u_2 \times a$ channels are collected and jointly serve as visible units in the consensus layer. As for each hidden unit, it should not be activated by the visible units from multiple classes. Hence we add a regularizer on the loss function for the RBM layer as follows:

$$L = -P(V,H) + \lambda \sum_i \prod_C \sum_{j \in C} W_{i,j}^2, \tag{6}$$

where λ controls the weight of the regularizer, and the product of weight sums of different affective states is adopted to attain the inter-classes weight sparsity.

In our EEG problem, the procedures of rough channel selection and the affection-based selection are complementary for each other. As the first level procedure does not involve the supervised information, it cannot perfectly capture the key features that make difference on participants' affection. On the other hand, even if the second level procedure leverages the label information, it cannot be directly used on the original input, due to the scarcity of labeled samples. Hence, to select the critical channels based on a small set of labeled data, we first utilize the larger unsupervised dataset to roughly make decisions and reduce the data dimensionality as well. After this, we use the supervised information to guide a more accurate label-related selection procedure at the second level.

The whole deep learning architecture will be established on top of the above two-level channel selection structure, which results in a marked mitigation of the noise effect and a great reduction of data dimensionality. In addition, each level can be extended to deep structure with stacked RBMs.

C. Semi-supervised Generative Model

Given the high-level representation obtained from RBMs, the direct supervised learning would still result in overfitting, for the small sample problem and the relatively high dimensionality of extracted features. Therefore, we propose to involve both supervised and unsupervised information in the classification layer.

Based on the extracted features, we build the final classification layer by a generative RBM model [27], using the output from the previous stacked RBMs and the label information jointly as visible units. Thus the visible layer consists of feature units and additional S units to represent the one-out-of-S structured label information. The Figure 1 (b) depicts such RBM structure. From the generative model, we can estimate the weights and biases by minimizing the combination of discriminative loss, generative loss and unsupervised data loss, as follows:

$$\min_\theta L = L_{dis} + \alpha L_{gen} + \beta L_{unsup}, \tag{7}$$

where θ denotes the set of model parameters, L_{dis}, L_{gen} and L_{unsup} stand for the objective functions for discriminative training, generative training and unsupervised training respectively, and α and β are hyper-parameters that control the weights for generative learning and unsupervised learning, respectively. Compared with discriminative learning, generative learning usually results in smaller variance of the estimated parameters [28]. It takes into account of the data distribution during training and on the other hand, the variance is the expectation with respect to data. Hence the generative training objective can be viewed as a regularizer in Equation 7. In addition, the unsupervised learning will provide further constraint on the training and lower down the variance. As for the concrete loss function of each component, we have:

$$\begin{aligned} L_{dis} &= -\sum_{i \in L} logP(Y_i|X_i), \\ L_{gen} &= -\sum_{i \in L} logP(X_i, Y_i), \\ L_{unsup} &= -\sum_{i \in uL} logP(X_i), \end{aligned} \tag{8}$$

where L represents the set of labeled data, uL represents the set of unlabeled data, X denotes the set of training samples with each element as a feature vector, Y denotes the set of labels corresponding to L, with each element as a scalar label in $\{1, 2, ...S\}$ and S stands for the total number of affective states.

We start with the generative training, where we consider the joint probability of features and labels. During the process, the parameters are going to be updated according to the gradient of L_{gen}. As for each component in the loss function, the gradient of $logP(X_i, Y_i)$ can be computed as:

$$\begin{aligned} \frac{\partial logP(X_i,Y_i)}{\partial \theta} &= -\mathbb{E}_{H|X_i,Y_i}[\frac{\partial}{\partial \theta} E(Y_i, X_i, H)] \\ &\quad + \mathbb{E}_{y,X,H}[\frac{\partial}{\partial \theta} E(y, X, H)], \end{aligned} \tag{9}$$

and the energy function of generative model $E(y, X, H)$ is defined as:

$$E(y, X, H) = -H^T W X - H^T U \mathbf{e}_y - B^T X - C^T H - D^T \mathbf{e}_y, \tag{10}$$

where W represents the weights between hidden units H and feature units X, U represents the weights between H and label units, B, C and D serve as biases for X, H and label units respectively, and $\mathbf{e}_y$ stands for the one-out-of-S vector with the y^{th} position set as 1. Compared with Equation 2, the generative RBM model involves the connection between

H and label vector by U, and the bias for label units by D. Hence it can be viewed as jointly using X and label vector as the visible layer in traditional RBM model.

The second term on the right side of Equation 9 represents the model expectation. Due to its computational intractability, we solve it by Contrastive Divergence, which is the stochastic approximation of the gradient. As X and label vector are independent of each other, the conditional distributions are same with Equation 3, except for:

$$\begin{aligned} P(y|H) &= sigm(U_{\cdot y}^T H + D_y), \\ P(H_i|y) &= sigm(U_{i\cdot}\mathbf{e}_y + C_i). \end{aligned} \tag{11}$$

With further marginalization and derivation, we can obtain that:

$$P(y|X) = \frac{exp(D_y + \sum_j log(1 + exp(o_{y,j}(X))))}{\sum\limits_{y' \in \{1,\ldots,K\}} exp(D_{y'} + \sum_j log(1 + exp(o_{y',j}(X))))}, \tag{12}$$

where the function $o_{y,j}(x)$ is defined as: $o_{y,j}(x) = C_j + \sum_k W_{j,k}X_k + U_{j,y}$.

Then during the discriminative training, the gradient of each component in loss function, $logP(Y_i|X_i)$ can be computed from Equation 12, as:

$$\begin{aligned} \frac{\partial logP(Y_i|X_i)}{\partial \theta} &= \sum_j sigm(o_{Y_i,j}(X_i)) \frac{\partial o_{Y_i,j}(X_i)}{\partial \theta} \\ &\quad - \sum_{j,y'} sigm(o_{y',j}(X_i)) P(y'|X_i) \frac{\partial o_{y',j}(X_i)}{\partial \theta}. \end{aligned} \tag{13}$$

After the discriminative and generative training, it comes to the question how to utilize the unsupervised information. As the unlabeled data provides feature information, we can use the current trained model to infer the label value according to Equation 12. Based on the marginalization of $P(X, Y)$, the gradient of $logP(X_i)$ can be computed as:

$$\begin{aligned} \frac{\partial logP(X_i)}{\partial \theta} &= -\mathbb{E}_{y|X_i}[\mathbb{E}_{H|y,X_i}[\frac{\partial}{\partial \theta} E(y, X_i, H)]] \\ &\quad + \mathbb{E}_{y,X,H}[\frac{\partial}{\partial \theta} E(y, X, H)], \end{aligned} \tag{14}$$

where the first term on the right side can be calculated either as the weighted average over $P(y|X_i)$, or from the simple sampling of $P(y|X_i)$.

D. Semi-supervised EEG Classification with Two-level Channel Selection

Given a set of T-channel labeled training samples $Xl = \{Xl_1, Xl_2, ..., Xl_n\}$ with the corresponding labels $Y = \{Y_1, ..., Y_n\}$, and a set of unlabeled samples $Xu = \{Xu_1, Xu_2, ..., Xu_m\}$, we use them jointly as the input to the training procedure through the channel selection and the classification. With the trained model, we can predict the label for test data in terms of Equation 12. The whole procedure is given in Algorithm 1. In the EEG affective state classification, the conventional training methods are usually plagued by the large variance due to the scarcity of labeled data. In our proposed framework, we mitigate the overfitting on two stages, the feature extraction stage and the final classification stage. During the first stage, we train RBMs to learn features from each channel, and simultaneously implement our two-level channel selection procedure. After we select the critical channels and add the consensus layer to combine different classes, the dimensionality has been greatly reduced. Thus at this point we can directly use RBMs to extract high-level features. On the other hand, the final classification stage mainly involves the training on a generative RBM model with the combinatorial objectives. This stage not only depends on the supervised information, including the discriminative and the generative learning, but it also uses the unsupervised data to regularize the training, thus to further alleviate the overfitting problem brought by the scarcity of labeled data.

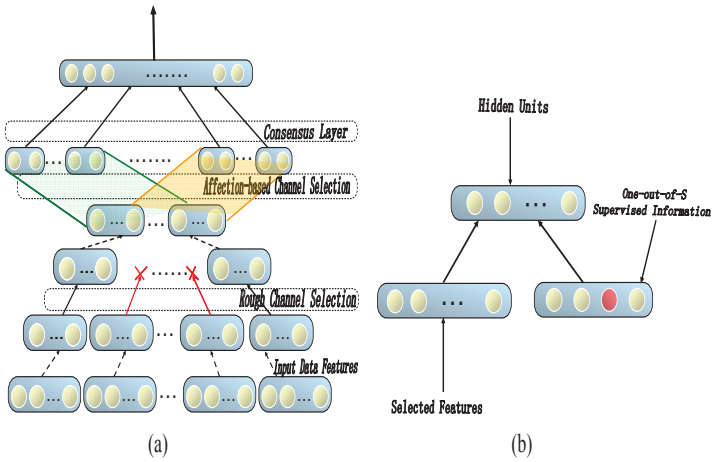

Fig. 1. The Framework of the Proposed Model. The two-level channel selection procedure (a) and the final generative classifer (b) are included. The model basically consists of RBM layers, with the dashed line denotes the structure of stacked RBMs.

Algorithm 1: Semi-supervised EEG Classification

Input: Labeled training set Xl and Y, unlabeled samples Xu, the number of selected channels u_1 at the rough selection stage, and u_2 at the finer-grained selection stage
Output: The labels Yu for samples in Xu

```
for i ← 1 to T do
    Train the stacked RBMs on the i^th channel;
    Calculate the zero-stimulus response;
Select the top u_1 channels with the highest response;
Collect the extracted features from selected u_1 channels;
for i ← 1 to u_1 do
    Train stacked RBMs on the i^th selected channels;
    for j ← 1 to S do
        Calculate the Significance Measurement of i^th channel to
        j^th state by Equation 5;
Select the top u_2 channels for each state;
Collect the extracted features from selected u_2 * S channels;
Train the consensus layer with respect to Equation 6;
Train stacked RBMs on top of the consensus layer;
Use the extracted features from last layer to train the combinatorial
model following Equations 9, 13 and 14;
Classify Xu based on trained model following Equation 12, and store
the results at Yu;
return Yu;
```

E. Reinforced Training Using Active Learning

Now we are going to extend our work a little bit further. With the collected EEG signals, usually it is expensive to hire experts to label manually. Assume we have some budget on labeling, then the question arises whether the different sample selection will result in different contribution to the training. The answer is definitely positive. For instance, if all the chosen samples lie in the central area of the same class, close with each other, they can provide little useful knowledge to the training. Hence, in this part we propose an extended application of our method in active learning scenario, to make the most of our limited budget.

After the training stage, we can quickly predict the label for a test sample by following Equation 12. With careful scrutinization, we can notice that the result is sometimes unreliable, that is, the value of $P(y|X)$ for different label y can be very close. On the contrary, if the probability value for certain label y dominates the others, we can agree that the model is quite confident with its decision. In this way, we conclude that the sample in the former case contains more uncertainty, and thus the learning with its true label can result in a faster advance to the more accurate decision boundary. To measure such kind of uncertainty, we can utilize the function of entropy:

$$Uncertainty(X) = -\sum_{y=1,\ldots,S} P(y|X) log P(y|X). \quad (15)$$

To involve the active learning procedure in the proposed model, it is first trained with the currently available dataset. After this, we rank all the unlabeled samples based on its uncertainty, and select a batch in proper size from the top. Then true labels are requested for the selected samples. Once we obtain the additional supervised information, the model can be retrained with a larger labeled dataset. We repeat this process until running out of budget. The whole procedure is given in Algorithm 2. As the "uncertain" samples can guide the decision boundary to progress towards the right direction, the active procedure can greatly speed up the model training. Notice that if each time we only selected one sample for inquiry, it would become a degenerative case with only a single sample as new added information in each round. Even if it may slow down the training, it is still desirable with small dataset and very limited budget.

Algorithm 2: The Model with Active Sample Selection

Input: Labeled training set Xl and Y, unlabeled samples Xu, available budget M, batch size BS
Output: The labels Yu for samples in Xu, enlarged labeled set Xl and Y

while *budget enough for another round* **do**
  Train the model with Xl, Y and Xu, according to Algorithm 1;
  Select the top BS samples from Xu based on uncertainty, according to Equation 15;
  Ask for labels regarding the selected samples, transfer them from Xu to Xl, add information to Y;
  deduct the cost from M;
Train the model with Xl, Y and Xu, according to Algorithm 1;
return Yu,Xl *and* Y;

IV. Experiments

In this section we implement our method on the DEAP Dataset [13], which provides EEG data especially for emotional analysis. The data was collected from 32 participants as they watched 40 one-minute long music videos. The labels were obtained from the surveys, with each of 40 videos rated according to arousal, valence, dominance and like/dislike. To fit the input of our affective state recognition problem, we preprocess the data according to [13] and obtain the downsampled version(128Hz), with 8064 features at each channel, and totally 40 channels. As for the EEG classification, we focus our view on whether the participants like or dislike the videos. As this is a binary classification problem, we name two states hereinafter as positive class and negative class for simplicity.

A. Evaluation on Affective State Recognition

During the process of data acquisition, it is commonly found that some participants in good mood may rate most videos with strongly positive scores while some others in really bad mood may give lower ratings. In this way, the individual variability and data acquisition errors will finally result in label imbalance. Different from traditional datasets, the measurement of accuracy is not able to well capture the classification performance for such skewed dataset, as the ratio value will be dominated by the majority class. Hence, we adopt the area-under-the-curve (AUC), i.e., the area under receiver operating characteristic (ROC) curve, to evaluate the classification result. ROC curve demonstrates the relationship between the true positive ratio and the false positive ratio, with the AUC value ranging from 0 to 1. The higher AUC value indicates the better classification performance and the greater robustness. Especially, we divide data into two halves and hide the labels for one half, which serves as the unsupervised information. To show the superior effectiveness, we compare our method with the following baselines:

Support Vector Machine (SVM): SVM serves as the first baseline method in our experiment. The features from EEG signals are directly used as the input to supervised SVM classifier. Through the comparison with SVM, we hope to show the necessity of feature extraction in EEG classification.

Principle Component Analysis (PCA)+SVM (PSVM): PCA [29] is an unsupervised feature extraction method by maximizing the data variance. In this baseline method, the EEG features are first processed using PCA on each channel. Then SVM is conducted as the supervised classifier for the affective state recognition. By comparing with PSVM, we can demonstrate the effectiveness of channel selection on improving the classification performance.

PCA+Fisher Criterion+SVM (PFSVM): Similarly we first use PCA to extract features on each channel. After this Fisher Criterion [17] is implemented to select critical channels. Finally SVM is adopted as the supervised classifier for the affective state recognition. The comparison with PFSVM can show the remarkable performance of our method on feature extraction procedure using RBMs, as well as the effectiveness of our final semi-supervised classifier model.

DBN+Fisher Criterion+RBM (DFRBM): To compare with previous deep learning model, in this implementation we use

TABLE II. THE PERFORMANCE OF AFFECTIVE STATE RECOGNITION BY AUC SCORE

ID	SVM	PSVM	PFSVM	DFRBM	DLM	semi-DLM
s01	0.677	0.670	0.631	0.720	0.830	**0.837**
s02	0.692	0.738	0.637	0.714	0.810	**0.822**
s03	0.680	0.517	0.615	0.724	0.745	**0.780**
s04	0.636	0.535	0.690	0.667	0.755	**0.775**
s05	0.604	0.631	0.510	0.792	0.808	**0.820**
s06	0.729	0.594	0.515	0.688	**0.782**	0.775
s07	0.657	0.656	0.516	0.688	0.820	**0.822**
s08	0.547	0.552	0.702	0.627	0.750	**0.762**
s09	0.616	0.530	0.515	0.708	0.832	**0.832**
s10	0.670	0.525	0.747	0.599	0.766	**0.772**
s11	0.708	0.606	0.615	0.626	0.770	**0.815**
s12	0.596	0.566	0.560	0.720	0.765	**0.796**
s13	0.643	0.625	0.536	0.641	0.737	**0.820**
s14	0.657	0.531	0.571	0.590	0.744	**0.762**
s15	0.637	0.748	0.560	0.681	0.766	**0.806**
s16	0.667	0.556	0.600	0.712	0.720	**0.742**
s17	0.715	0.625	0.630	0.596	0.714	**0.774**
s18	0.583	0.531	0.566	0.729	0.765	**0.830**
s19	0.546	0.657	0.604	0.676	0.770	**0.844**
s20	0.619	0.563	0.635	0.725	0.752	**0.826**
s21	0.626	0.625	0.594	0.678	0.721	**0.788**
s22	0.657	0.692	0.637	0.630	0.744	**0.770**
s23	0.604	0.660	0.707	0.687	0.710	**0.769**
s24	0.600	0.640	0.596	0.635	0.700	**0.781**
s25	0.677	0.531	0.697	0.710	**0.712**	0.705
s26	0.677	0.552	0.635	0.673	0.752	**0.832**
s27	0.657	0.680	0.546	0.625	0.760	**0.820**
s28	0.667	0.606	0.771	0.682	0.722	**0.793**
s29	0.570	0.636	0.531	0.768	0.776	**0.810**
s30	0.707	0.626	0.525	0.736	0.769	**0.848**
s31	0.681	0.797	0.680	0.662	0.774	**0.782**
s32	0.687	0.667	0.576	0.635	0.797	**0.852**

DBN to lower the dimension of data in each channel. Then Fisher Criterion is used to select critical channels, which are then combined and fed into a supervised RBM for the classification.

Besides the above baseline methods, we also implement our method in two versions, *deep learning model (DLM)* and the *semi-supervised deep learning model (semi-DLM)*. The semi-DLM is the complete model, while the DLM uses only supervised information in the training of final classification layer, that is, with only the discriminative and generative training objectives in Equation 7. We hope to testify the positive impact from the unsupervised regularization through the comparison with DLM .

The experimental results with AUC scores of our method and baseline methods are listed in Table II. We can observe that our semi-DLM method very well outperforms baseline methods. Given the results of SVM, we can conclude that feature extraction is necessary for EEG classification. On the other hand, the performance of PSVM and PFSVM shows that the method of PCA and Fisher Criterion cannot successfully extract the meaningful features from the small dataset. For some of the participants, these two methods even generate worse results than simply SVM. This is resulted from the mistakenly selected critical features. On the contrary, by using deep learning model, we can obtain the representative features that is crucial for the successful affective state recognition. In addition, with the comparison of three columns, we arrive at a conclusion that the model with two-level channel selection and unsupervised regularization surpasses the traditional deep models. It also shows that the unsupervised structure is beneficial for both feature extraction and classification with less overfitting.

Then we implement our method involving active learning procedure. Specifically, for each participant, we fix 10 samples as the test data, and use 10 of the remaining 30 for the initial training. In each round we determine which samples to label, by selecting top 2 most beneficial ones as a batch according to Equation 15. Then we retrain the model with the new added samples, and repeat this procedure five times. Totally 5 batches of 10 samples are selected as the additional labeled training set. To demonstrate its effectiveness, we compare it to the case with random selection. We still fix the test set, and use the identical data with active case for the initial training. Besides, we randomly pick 10 others from the remaining 20 samples, and conduct training. We repeat this random training process 10 times and calculate the average. The average AUC scores over all participants are given in Table III. The first row denotes the maximum AUC score generated from the random selection in 10 times, while the second row denotes the averaged AUC score in 10 times. The third and fourth rows denote the same measurement with 4 more samples used for the initial training. However, the additional training set is decided by the first two batches of active learning, and after this we only select 6 more labeled samples for training. As for our active-DLM model, which does not involve randomness, we only provide the final performance in the row of average. From the Table III we can easily observe the outstanding performance with the active sample selection.

TABLE III. THE PERFORMANCE OF THE ACTIVE LEARNING

Methods	DLM	semi-DLM	active-DLM
max	0.788	0.800	-
avg	0.767	0.789	**0.808**
max2	0.794	0.806	-
avg2	0.779	0.796	**0.808**

B. Evaluation on Channel Selection

In this part we conduct the experiment respectively on each level of channel selection, to demonstrate the effectiveness of our proposed two-level structure.

Starting form the first level rough channel selection, we visualize the result in Figure 2. The method of zero-stimulus response is adopted as the measurement for channel selection and we pick top two most critical channels. As for each visualized channel, we choose two most significant features from the RBM output as the coordinate axis. Here the measurement of feature significance is defined as: $Sig(i) = (\sum_{X \in L1} R(X)_i - \sum_{X \in L2} R(X)_i)^2$, where $L1$ denotes the set of samples in the positive class, $L2$ denotes the set of samples in the negative class, and $R(X)$ represents the extracted features of sample X from RBM. Basically this measurement describes the difference of feature i between samples in positive class and in negative class. By using the most significant features, we hope to clearly show the contribution of channel selection procedure in the affective state recognition.

From Figure 2 (a) and (b) we can observe that, even if the unsupervised channel selection process can automatically capture the salient data layout, it may not be able to pick the

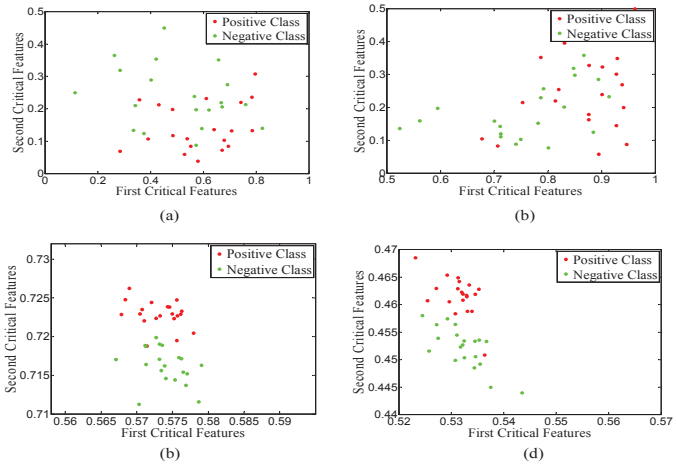

Fig. 2. The Visualization of Selected Channels. The Figure(a) and (b) represent the feature distribution in the detected top two most critical channels at rough selection level. Figure(c) and (d) shows the feature distribution in the detected top two most critical channels for positive class at second selection level. The vertices in red color and green color represent the samples in positive class and negative class, respectively.

channels where the samples from different classes are well separated. This is due to the incapability of the unsupervised learning in catching the crucial difference between different classes, which necessitates the guidance by labels during the second level channel selection.

After the supervised information is leveraged, we show the second level selection results in Figure 2 (c) and (d). we can notice that our finer-grained selection method can easily determine the channels that contain representative features for each affective state.

To recap, the results reveal that the channel selection procedure is meaningful and necessary for the ultimate classification task. Besides, the comparison between Figure 2 (a), (b) and Figure 2 (c), (d) reflects that the performance from the rough selection has been very well enhanced by the affection-based selection. Reversely, the first level selection is also necessary for the second level selection, as it greatly reduces the data dimensionality, and thus enables the second level procedure to progress even with limited number of labeled samples.

V. Conclusions

In this paper, we propose a novel semi-supervised method for the affective state recognition using deep learning model. The method very well combines the supervised and unsupervised information for both feature extraction and affective state classification. During the feature extraction procedure, we come up with a two-level channel selection structure. Due to the lack of labeled samples, at first level we only use unsupervised information to roughly make decisions. Then we conduct a finer-grained selection with the guidance of label information. After we successfully extract the representative features from the constructed deep layers, we build a generative RBM model as the final classifier, jointly regularized by generative and unsupervised training objectives. Finally, we extend our proposed model to the active learning scenario, which solves the costly labeling problem. The experimental results reveal that our model surpasses extensive baselines in classification and our proposed reinforced process outpaces the random labeling training by a decent margin.

References

[1] R. W. Picard, "Toward computers that recognize and respond to user emotion," *IBM Systems Journal*, vol. 39, no. 3.4, pp. 705–719, 2000.

[2] R. W. Picard, *Affective computing*. MIT press, 2000.

[3] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *ANNIE*. Citeseer, 1999, pp. 7–10.

[4] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 568–573.

[5] D. Jun, X. Miao, Z. Hong-hai, and L. Wei-feng, "Wearable ecg recognition and monitor," in *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*. IEEE, 2005, pp. 413–418.

[6] A. Nakasone, H. Prendinger, and M. Ishizuka, "Emotion recognition from electromyography and skin conductance," in *BSI*. Citeseer, 2005, pp. 219–222.

[7] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using eegs and peripheral physiological signals," in *Multimedia content representation, classification and security*. Springer, 2006, pp. 530–537.

[8] H. Prendinger, J. Mori, and M. Ishizuka, "Recognizing, modeling, and responding to users affective states," in *User Modeling 2005*. Springer, 2005, pp. 60–69.

[9] R. Khosrowabadi, A. Wahab, K. K. Ang, and M. H. Baniasad, "Affective computation on eeg correlates of emotion from musical and vocal stimuli," in *IJCNN*. IEEE, 2009, pp. 1590–1594.

[10] G. E. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.

[11] O. Chapelle, B. Schölkopf, A. Zien *et al.*, *Semi-supervised learning*. MIT press Cambridge, 2006, vol. 2.

[12] A. Krogh, J. Vedelsby *et al.*, "Neural network ensembles, cross validation, and active learning," *Advances in neural information processing systems*, pp. 231–238, 1995.

[13] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

[14] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time eeg-based emotion recognition and its applications," in *Transactions on computational science XII*. Springer, 2011, pp. 256–277.

[15] R. Khosrowabadi, H. C. Quek, A. Wahab, and K. K. Ang, "Eeg-based emotion recognition using self-organizing map for boundary detection," in *ICPR*. IEEE, 2010, pp. 4242–4245.

[16] D. O. Bos, "Eeg-based emotion recognition," *The Influence of Visual and Auditory Stimuli*, pp. 1–17, 2006.

[17] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Scholkopf, "Support vector channel selection in bci," *TBME*, vol. 51, no. 6, pp. 1003–1010, 2004.

[18] X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.

[19] X. Zhu, "Semi-supervised learning literature survey," *University of Wisconsin-Madison*, vol. 2, p. 3, 2006.

[20] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *Learning theory*. Springer, 2004, pp. 624–638.

[21] X. Zhu, J. Lafferty, and R. Rosenfeld, "Semi-supervised learning with graphs," Ph.D. dissertation, Carnegie Mellon University, 2005.

[22] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[23] D. Wulsin, J. Blanco, R. Mani, and B. Litt, "Semi-supervised anomaly detection for eeg waveforms using deep belief nets," in *ICMLA*. IEEE, 2010, pp. 436–441.

[24] D. Wulsin, J. Gupta, R. Mani, J. Blanco, and B. Litt, "Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement," *Journal of neural engineering*, vol. 8, no. 3, p. 036015, 2011.

[25] K. Li, X. Li, Y. Zhang, and A. Zhang, "Affective state recognition from eeg with deep belief networks," in *BIBM*, Dec 2013, pp. 305–310.

[26] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[27] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *The Journal of Machine Learning Research*, vol. 13, pp. 643–669, 2012.

[28] P. Liang and M. I. Jordan, "An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators," in *ICML*. ACM, 2008, pp. 584–591.

[29] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.