A Novel Neural Network Model based on Cerebral Hemispheric Asymmetry for EEG Emotion Recognition

Yang Li[1,2], **Wenming Zheng**[1,*], **Zhen Cui**[3], **Tong Zhang**[1,2] and **Yuan Zong**[1]
[1] Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, China
[2] School of Information Science and Engineering, Southeast University, China
[3] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
wenming_zheng@seu.edu.cn

Abstract

In this paper, we propose a novel neural network model, called bi-hemispheres domain adversarial neural network (BiDANN), for EEG emotion recognition. BiDANN is motivated by the neuroscience findings, $i.e.$, the emotional brain's asymmetries between left and right hemispheres. The basic idea of BiDANN is to map the EEG data of both left and right hemispheres into discriminative feature spaces separately, in which the data representations can be classified easily. For further precisely predicting the class labels of testing data, we narrow the distribution shift between training and testing data by using a global and two local domain discriminators, which work adversarially to the classifier to encourage domain-invariant data representations to emerge. After that, the learned classifier from labeled training data can be applied to unlabeled testing data naturally. We conduct two experiments to verify the performance of our BiDANN model on SEED database. The experimental results show that the proposed model achieves the state-of-the-art performance.

1 Introduction

Emotion plays a crucial role in human being's learning and communications, and has been one of the most attractive topic in affective computing area. Psychologists have conducted a lot of studies about the definition, constitution, property and function of emotion [Izard, 1991; Storbeck and Clore, 2005]. However, emotion is still hard to be understood by machines. Emotion recognition, as a popular topic, receives substantial attentions in computer vision and pattern recognition researches [Picard and Picard, 1997].

The responses of emotion can be facial expression, speech and other physiologic signals such as skin conductance response, heart rate, blood pressure, cortisol level, electromyography and respiration rate. However, from a neuroscience point of view [Lotfi and Akbarzadeh-T, 2014], human's emotion is closely related to a variety of brain subregions, such as the orbital frontal cortex, ventral medial prefrontal cortex, amygdala [Britton *et al.*, 2006; Etkin *et al.*, 2011]. Thus it is a direct means to study emotion by collecting human's brain activity signals under deferent moods. The technology of electroencephalograph (EEG) can measure the changes of this brain electrical activities. It places electrodes on the head of participants non-invasively, and has high temporal resolution and can directly reflect the potential activity of the nerve. Therefore, EEG signal can be used to decode emotions. As a novel research means for emotion, EEG advances the research of emotion recognition.

EEG emotion recognition task can be roughly partitioned into two steps: feature extraction and classifier design. First, features are extracted from time domain, frequency domain or time-frequency domain [Jenke *et al.*, 2014]. Then a set of EEG feature vectors are chosen to train a classifier and the other EEG data are tested based on it. Many researchers construct models and introduce methods to deal with EEG emotion recognition tasks [Musha *et al.*, 1997]. In [Kim *et al.*, 2013], Kim $et\ al.$ reviewed the computational methods that have been developed to deduct EEG indices of emotion, to extract emotion-related features, or to classify EEG signals into one of many emotional states. In [Jenke *et al.*, 2014], Jenke $et\ al.$ made a lot of experiments to compare the existing features using machine learning techniques for feature selection on a self recorded data set. In [Li *et al.*, 2016], Li $et\ al.$ proposed a novel regression model, called graph regularized sparse linear discriminant analysis (GraphSLDA), to deal with EEG emotion recognition problem. In [Zheng, 2017], Zheng $et\ al.$ proposed a novel group sparse canonical correlation analysis (GSCCA) method for simultaneous EEG channel selection and emotion recognition. Recently, deep learning methods have shown better performance than traditional methods to deal with EEG emotion recognition problem. For example, in [Zheng and Lu, 2015], Zheng $et\ al.$

*Corresponding author

used Deep Belief Network(DBN) to extract high-level features of EEG emotion data.

Although there have been so many algorithms or models to deal with EEG emotion recognition problem, most of them have not considered the dependencies between training and testing data. Specifically, for traditional deep learning methods, the label information of testing data is predicted directly based on the classifier trained based on the training data samples. The effect of this learning strategy is mainly based on a assumption that the distributions of training data and testing data are similar. For EEG signals, however, data distribution shift is tremendous for different people or even same people but under different circumstances. Thus, in EEG emotion recognition algorithm, we should ensure the data representations invariable referring to source (training) or target (testing) domain by removing the domain identification information to decrease this data distribution shift. In other words, we should find a data representation space where the data coming from source and target domains is indistinguishable as more as possible, while preserving a low risk on the source labeled data.

On the other side, from the view of neuroscience, it is more valuable to consider the nature or characteristics of brain in the construction of model for EEG emotion recognition. In fact, although anatomy of human brain looks like symmetric, the left and right hemispheres are not entirely symmetrical. Asymmetry, both in structure and function, exists throughout the neocortex and cortical substructures [Zatorre *et al.*, 1992; Greve *et al.*, 2013].

Recently, in [Ganin *et al.*, 2016], Ganin $et\ al.$ proposed a Domain Adversarial Neural Networks (DANN) to deal with domain adaptation problems. It adopted a domain discriminator to distinguish which domain the input comes from. Without any class information from testing data, DANN modified the data representation space to generate domain-invariable data features. Benefiting from this, we can consider to fit the asymmetry of emotional brain into the DANN framework, meanwhile utilize time information for EEG emotion recognition tasks.

Thus, in this paper, we propose a novel deep neural network model, called bi-hemispheres domain adversarial neural network (BiDANN) model, which considers distribution shift between training and testing data and cerebral hemisphere asymmetry, to deal with two general EEG emotion recognition tasks. BiDANN learns discriminative information in regard to emotion while in-discriminative information referring to domains. This is achieved by jointly optimizing three modules: (1) $feature\ extractors$. Two $feature\ extractors$ learn the inter-information on each cerebral hemisphere separately, and map the original EEG data into deep feature space which has more discriminative emotional information. (2) $classifier$. It predicts the emotion class label by mapping the feature into label space. (3) $domain\ discriminators$. A global domain discriminator is used to distinguish which domain (training data or testing data) the input comes from so as to decrease the distribution shift, while two hemispheric local discriminators further narrow the left and right hemispheric data distributions separately in either source or target domain, which are components of the entire cerebral EEG data. The global discriminator constrains the entire data distribution similar, meanwhile the local discriminator further constrains one part of the entire data distribution similar because of the large difference between two hemispheres. The parameters of $feature\ extractors$ are optimized to minimize the loss of $classifier$ but maximize the loss of $domain\ discriminators$, which will lead to an adversarial learning between $classifier$ and $domain\ discriminators$ to encourage emotion-related but domain-invariant data representation appeared.

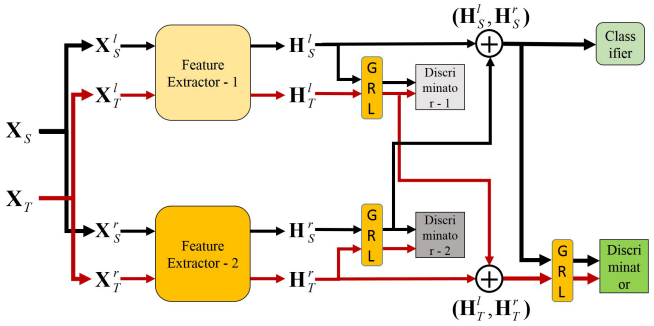


Figure 1: The framework of BiDANN. The black lines and arrows refer to source domain path, while the red lines and arrows refer to target domain path.

The major contribution of this paper can be summarized as follows:

- This is the first work to consider the dependence between left and right hemispheres in emotion recognition research, and integrate this neuroscience finding of cerebral asymmetry into deep learning model;
- Besides constraining the global distribution similarity between training and testing data, we consider the local distribution dependency between left and right hemispheres.

2 The Proposed Model for EEG Emotion Recognition

In this section, we introduce the proposed BiDANN, and then use it to deal with EEG emotion recognition.

2.1 The BiDANN Model

Fig. 1 shows the overall framework of BiDANN, which consists of two feature extractors (Feature Extractor -1 and Feature Extractor -2), a global discriminator (Discriminator) and two local discriminators (Discriminator -1 and Discriminator -2), three gradient reversal layers (GRL), and a classifier. The two feature extractors capture the dynamic features of two hemispheric EEG signal separately. Three discriminators are trained on a binary domain label set $\mathcal{D} = \{0, 1\}$, in which the domain labels of source samples are set to 0 while the domain labels of target samples are set to 1. The global discriminator narrows the features' distribution gap between source and target domains, while the two local discriminators complementally eliminate the left and right hemispheric features' distribution difference within source and target domains respectively. GRL can maximize the loss of discriminators by

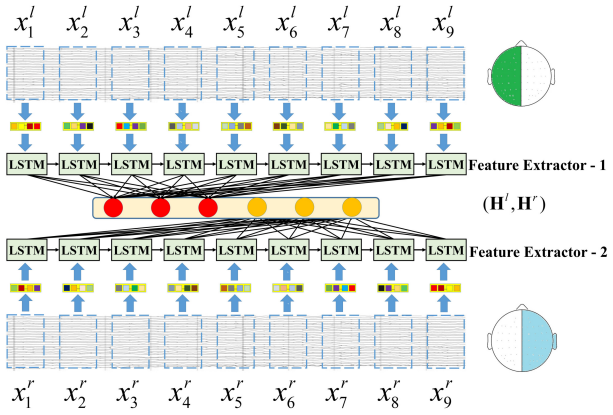


Figure 2: The feature extractors of BiDANN.

leaving the input unchanged during forward propagation and reversing the gradient by multiplying it by a negative scalar during backpropagation [Ganin *et al.*, 2016]. The classifier maps the emotion-related and domain-invariable features into class label space to predict the class labels.

Overall, the complete objective function of the proposed BiDANN is as follows:

$$L(\mathbf{X}_{\mathcal{R}}; \theta_f^l, \theta_f^r, \theta_c, \theta_d^l, \theta_d^r, \theta_d) = L_c(\mathbf{X}_{\mathcal{S}}; \theta_f^l, \theta_f^r, \theta_c) - L_d^l(\mathbf{X}_{\mathcal{R}}^l; \theta_f^l, \theta_d^l) - L_d^r(\mathbf{X}_{\mathcal{R}}^r; \theta_f^r, \theta_d^r) - L_d(\mathbf{X}_{\mathcal{R}}; \theta_f^l, \theta_f^r, \theta_d). \quad (1)$$

Here $\mathbf{X}$ denotes an EEG sequence. $\mathcal{R} \in (\mathcal{S}, \mathcal{T})$, $\mathcal{S}$ and $\mathcal{T}$ denote the source and target domains separately. L_c, L_d^l, L_d^r and L_d are the loss functions of classifier, left and right hemispheric local discriminators, and global discriminator. θ_c, θ_d^l, θ_d^r and θ_d are their corresponding parameters. θ_f^l and θ_f^r are the parameters of left and right hemispheric feature extractors. Here l and r denote the left and right hemisphere separately.

The detailed operations of BiDANN on left and right hemispheric EEG data come from three aspects:

(1) Feature extractor for single hemispheric data. Let $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_{t-1}, \mathbf{x}_t, \cdots, \mathbf{x}_T\} \in \mathbb{R}^{d_x \times T}$, where d_x is the dimension and T is the length of sequence. To make full use of the temporal dependence in sequence, we construct a Long Short-Term Memory (LSTM) framework to learn the context information and transfer the input to another space which has more effective and high level components. Given an input $\mathbf{x}_t \in \mathbb{R}^{d_x \times 1}$, it is encoded to a latent state $\mathbf{h}_t^f \in \mathbb{R}^{d_h \times 1}$, which locates in another space and has more discriminability that we expect. Here d_h is the dimension of hidden states.

Through repeating the LSTM module recurrently, we obtain a sequence of hidden states that fully represents the input sequence. For the sequence of hidden states, we map it into another compressed sequence $\mathbf{H}^c = \{\mathbf{h}_1^c, \cdots, \mathbf{h}_k^c, \cdots, \mathbf{h}_K^c\} \in \mathbb{R}^{d_h \times K}$ with a projection matrix $\mathbf{G}^c = [G_{ik}^c]_{T \times K}$ which scans the whole hidden sequence and represents them with simplified states which has a higher level features, *i.e.*,

$$\mathbf{h}_k^c = \sigma(\sum_{i=1}^{T} G_{ik}^c \mathbf{h}_i^f + \mathbf{b}^c), \quad k = 1, 2, \cdots, K, \quad (2)$$

where $\mathbf{b}^c \in \mathbb{R}^{d_h \times 1}$ is a bias and K is the length of the compressed sequence. Then we get dynamic features with more discriminability about emotions to represent the input states. The above feature extraction process can be simply formulated as $\mathbf{H}^c = \mathcal{L}(\mathbf{X})$.

For an EEG sequence $\mathbf{X}$, we split it into left and right hemispheric data sequences, *i.e.*, $\mathbf{X} = [\mathbf{X}^l, \mathbf{X}^r] = \{[\mathbf{x}_1^l, \mathbf{x}_2^l, \cdots, \mathbf{x}_t^l, \cdots, \mathbf{x}_T^l], [\mathbf{x}_1^r, \mathbf{x}_2^r, \cdots, \mathbf{x}_t^r, \cdots, \mathbf{x}_T^r]\}$. The features of both hemispheres on source (training) data $\mathbf{X}_{\mathcal{S}} = [\mathbf{X}_{\mathcal{S}}^l, \mathbf{X}_{\mathcal{S}}^r]$ and target (testing) data $\mathbf{X}_{\mathcal{T}} = [\mathbf{X}_{\mathcal{T}}^l, \mathbf{X}_{\mathcal{T}}^r]$ can be formulated as:

$$\begin{aligned}\mathbf{H}_{\mathcal{R}}^l &= [\mathbf{H}_{\mathcal{S}}^l, \mathbf{H}_{\mathcal{T}}^l] = [\mathbf{h}_{1\mathcal{S}}^l, \cdots, \mathbf{h}_{K\mathcal{S}}^l, \mathbf{h}_{1\mathcal{T}}^l, \cdots, \mathbf{h}_{K\mathcal{T}}^l] \\ &= E_f(\mathbf{X}_{\mathcal{S}}^l, \mathbf{X}_{\mathcal{T}}^l; \theta_f^l) = [\mathcal{L}_1(\mathbf{X}_{\mathcal{S}}^l), \mathcal{L}_1(\mathbf{X}_{\mathcal{T}}^l)], \quad (3)\end{aligned}$$

$$\begin{aligned}\mathbf{H}_{\mathcal{R}}^r &= [\mathbf{H}_{\mathcal{S}}^r, \mathbf{H}_{\mathcal{T}}^r] = [\mathbf{h}_{1\mathcal{S}}^r, \cdots, \mathbf{h}_{K\mathcal{S}}^r, \mathbf{h}_{1\mathcal{T}}^r, \cdots, \mathbf{h}_{K\mathcal{T}}^r] \\ &= E_f(\mathbf{X}_{\mathcal{S}}^r, \mathbf{X}_{\mathcal{T}}^r; \theta_f^r) = [\mathcal{L}_2(\mathbf{X}_{\mathcal{S}}^r), \mathcal{L}_2(\mathbf{X}_{\mathcal{T}}^r)]. \quad (4)\end{aligned}$$

The complete feature extraction process is shown in Fig. 2.

(2) Local and global discriminators. We set domain label sets $\mathcal{D}_{\mathcal{S}} = 0_{,i=1,2,\cdots,N}$ and $\mathcal{D}_{\mathcal{T}} = 1_{,j=1,2,\cdots,M}$ to source and target samples separately, where N and M are the number of source and target domain samples. This is used to calculate the loss of discriminators. We can denote the loss function of discriminator as:

$$\mathcal{L}(G_d(E_f(\mathbf{X}_{\mathcal{R}}; \theta_f); \theta_d), \mathcal{D}_{\mathcal{R}}), \quad (5)$$

where $\mathcal{L}$ is the classification loss such as cross-entropy loss function, G_d is the domain label classifier and E_f is the feature extractor function. To coincide the feature distributions of source and target domains, the parameters of feature extractors are updated to strive to generate data representation to confuse the discriminator to distinguish which domain the input comes from by maximizing the discriminator loss function.

Furthermore, because the special nature of human cerebral function brings the left hemispheric channels' distribution of EEG signal has a gap with the channels' distribution on the right, we can not narrow this gap only use the global discriminator in reality although it looks feasible in theory, which is different from traditional data's constitutions such as images or audio. Thus we use two hemispheric data domain discriminators to constrain the local data distribution similarity between source and target domains. The experimental results show that the operations on split left and right cerebral hemispheric data of our model is useful to get a good performance. In summary, we can formulate the loss functions of local and global discriminators as:

$$L_d^l(\mathbf{X}_{\mathcal{R}}^l; \theta_f^l, \theta_d^l) = \mathcal{L}(G_d(E_f(\mathbf{X}_{\mathcal{R}}^l; \theta_f^l); \theta_d^l), \mathcal{D}_{\mathcal{R}}), \quad (6)$$

$$L_d^r(\mathbf{X}_{\mathcal{R}}^r; \theta_f^r, \theta_d^r) = \mathcal{L}(G_d(E_f(\mathbf{X}_{\mathcal{R}}^r; \theta_f^r); \theta_d^r), \mathcal{D}_{\mathcal{R}}), \quad (7)$$

$$L_d(\mathbf{X}_{\mathcal{R}}; \theta_f^l, \theta_f^r, \theta_d) = \mathcal{L}(G_d(E_f(\mathbf{X}_{\mathcal{R}}; \theta_f^l, \theta_f^r); \theta_d), \mathcal{D}_{\mathcal{R}}). \quad (8)$$

(3) Discriminative prediction. Like most supervised models, we add a supervision term into the network so as to enhance the model's discriminability. Concretely, we use softmax function on the transformed hidden states to predict the class labels, *i.e.*,

$$\mathbf{q}_i = [{\mathbf{h}^l_{1\mathcal{S}}}^T, \cdots, {\mathbf{h}^l_{K\mathcal{S}}}^T, {\mathbf{h}^r_{1\mathcal{S}}}^T, \cdots, {\mathbf{h}^r_{K\mathcal{S}}}^T]^T, \quad (9)$$

$$P(y_i = c|\mathbf{q}_i, \mathbf{G}, \mathbf{b}) = \frac{\exp(\mathbf{G}\mathbf{q}_i + \mathbf{b})}{\sum_k \exp(\mathbf{G}\mathbf{q}_k + \mathbf{b})}, \quad (10)$$

$$\widetilde{y}_i = arg\max P(y_i = c|\mathbf{q}_i, \mathbf{G}, \mathbf{b}), \quad (11)$$

where $\mathbf{q}_i \in \mathbb{R}^{2Kd_h \times 1}$, the variables $\mathbf{G} \in \mathbb{R}^{d_L \times 2Kd_h}$ and $\mathbf{b} \in \mathbb{R}^{d_L \times 1}$ are respectively the transform matrix and bias, c is the c-th class, y_i is the ground-truth label of i-th training data, d_L is the number of class. The loss function of class label prediction can be expressed as:

$$L_c(\mathbf{X}_\mathcal{S}; \theta^l_f, \theta^r_f, \theta_c) = \mathcal{L}(G_c(E_f(\mathbf{X}_\mathcal{S}; \theta^l_f, \theta^r_f); \theta_c), y_i)$$
$$= -\sum_t \log(P(\widetilde{y}_i = c|\mathbf{q}_i, \mathbf{G}, \mathbf{b}), \quad (12)$$

where G_c denotes the class label classifier of source domain.

2.2 Optimization of BiDANN

Through minimizing L_c and maximizing L^l_d, L^r_d, L_d, we optimize the objective function of Eq. (1) to achieve a saddle point by:

$$(\hat{\theta}^l_f, \hat{\theta}^r_f, \hat{\theta}_c) = arg\min_{\theta^l_f, \theta^r_f, \theta_c} L(\mathbf{X}_\mathcal{R}; (\theta^l_f, \theta^r_f, \theta_c), \hat{\theta}_d, \hat{\theta}^l_d, \hat{\theta}^r_d), \quad (13)$$

$$\hat{\theta}_d = arg\max_{\theta_d} L(\mathbf{X}_\mathcal{R}; \hat{\theta}^l_f, \hat{\theta}^r_f, \hat{\theta}_c, \theta_d, \hat{\theta}^l_d, \hat{\theta}^r_d), \quad (14)$$

$$\hat{\theta}^l_d = arg\max_{\theta^l_d} L(\mathbf{X}^l_\mathcal{R}; \hat{\theta}^l_f, \hat{\theta}^r_f, \hat{\theta}_c, \hat{\theta}_d, \theta^l_d, \hat{\theta}^r_d), \quad (15)$$

$$\hat{\theta}^r_d = arg\max_{\theta^r_d} L(\mathbf{X}^r_\mathcal{R}; \hat{\theta}^l_f, \hat{\theta}^r_f, \hat{\theta}_c, \hat{\theta}_d, \hat{\theta}^l_d, \theta^r_d). \quad (16)$$

We use stochastic gradient descent (SGD) to update θ^l_f, θ^r_f, θ_c to the direction of minimizing L_c and maximizing L^l_d, L^r_d, L_d, and update θ^l_d, θ^r_d, θ_d to the direction of minimizing L^l_d, L^r_d, L_d. This max-minimum goal can be converted to a minimum function, *i.e.*, $\min L = \min L_c + L^l_d + L^r_d + L_d$, by three gradient reversal layers (GRL) shown in Fig. 1, which keep the gradient sign at forward-propagation but reverse it while performing back-propagation. The optimization procedure of BiDANN is shown in Algorithm 1.

We iteratively train the classifier and three discriminators and update the parameters same with standard deep learning methods by chain rule. But the difference is that the parameters before the GRL module, *i.e.*, the parameters of feature extractors, will minus the gradients with opposite sign that come from GRL at back-propagation. Thus the feature extractors will generate data representations that minimize the loss of classifier while maximize the loss of discriminators. In addition, for BiDANN, we update the classifier more times than the discriminators, because that our goal is to classify the EEG emotion data instead of wiping out the domain-related information thoroughly.

Algorithm 1 Optimization of BiDANN.

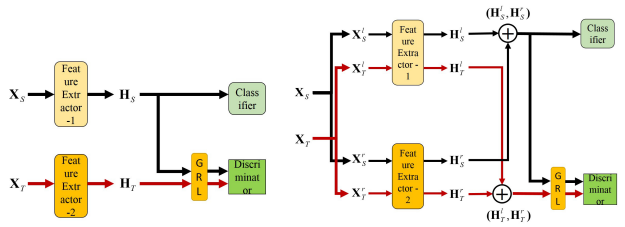
Input:

Training data set $\mathbf{X}_S$ and Testing data set $\mathbf{X}_T$;
Ground-truth label set $\mathbf{L}_S$ of training data set;
Training (source) domain label set $\mathcal{D}_\mathcal{S} = [\mathcal{D}^l_\mathcal{S}, \mathcal{D}^r_\mathcal{S}] = \{0\}$ and testing (target) domain label set $\mathcal{D}_\mathcal{T} = [\mathcal{D}^l_\mathcal{T}, \mathcal{D}^r_\mathcal{T}] = \{1\}$;
Initial learning rate α;

Output:

Parameter: $\hat{\theta}^l_f, \hat{\theta}^r_f, \hat{\theta}_c, \hat{\theta}_d, \hat{\theta}^l_d, \hat{\theta}^r_d$.

1: Input $\mathbf{X}_\mathcal{S}$ and $\mathbf{L}_\mathcal{S}$ to update the parameters of Classifier: $\theta_c \leftarrow \theta_c - \alpha\frac{\partial L_c}{\partial \theta_c}$, $\theta^l_f \leftarrow \theta^l_f - \alpha\frac{\partial L_c}{\partial \theta^l_f}$, $\theta^r_f \leftarrow \theta^r_f - \alpha\frac{\partial L_c}{\partial \theta^r_f}$;
2: Input $\mathbf{X}_\mathcal{S}$, $\mathbf{X}_\mathcal{T}$, $\mathcal{D}_\mathcal{S}$ and $\mathcal{D}_\mathcal{T}$ to update the parameters of global Discriminator: $\theta_d \leftarrow \theta_d - \alpha\frac{\partial L_d}{\partial \theta_d}$, $\theta^l_f \leftarrow \theta^l_f + \alpha\frac{\partial L_d}{\partial \theta^l_f}$, $\theta^r_f \leftarrow \theta^r_f + \alpha\frac{\partial L_d}{\partial \theta^r_f}$;
3: Input $\mathbf{X}^l_\mathcal{S}$, $\mathbf{X}^l_\mathcal{T}$, $\mathcal{D}^l_\mathcal{S}$ and $\mathcal{D}^l_\mathcal{T}$ to update the parameters of left hemispheric local Discriminator: $\theta^l_d \leftarrow \theta^l_d - \alpha\frac{\partial L^l_d}{\partial \theta^l_d}$, $\theta^l_f \leftarrow \theta^l_f + \alpha\frac{\partial L^l_d}{\partial \theta^l_f}$;
4: Input $\mathbf{X}^r_\mathcal{S}$, $\mathbf{X}^r_\mathcal{T}$, $\mathcal{D}^r_\mathcal{S}$ and $\mathcal{D}^r_\mathcal{T}$ to update the parameters of right hemispheric local Discriminator: $\theta^r_d \leftarrow \theta^r_d - \alpha\frac{\partial L^r_d}{\partial \theta^r_d}$, $\theta^r_f \leftarrow \theta^r_f + \alpha\frac{\partial L^r_d}{\partial \theta^r_f}$;
5: If algorithm has scanned all data 100 times, then $\alpha \leftarrow 0.9 \times \alpha$ and goto step1;
6: **return** $\hat{\theta}^l_f, \hat{\theta}^r_f, \hat{\theta}_c, \hat{\theta}_d, \hat{\theta}^l_d, \hat{\theta}^r_d$.



(a) BiDANN-R1 (b) BiDANN-R2

Figure 3: The baseline methods in our experiments.

3 Experiments

3.1 Setting Up

The baseline methods in the experiments are DANN [Ganin *et al.*, 2016] and two reduced versions of our proposed BiDANN, *i.e.*, BiDANN-R1 and BiDANN-R2 shown in Fig. 3 in our experiments. BiDANN-R2 reduces the local discriminators compared with BiDANN, and further BiDANN-R1 extracts source and target domain samples' feature ignoring hemispheric difference. The feature extractors' structure of these methods are same with the Feature Extractor -1 (or Feature Extractor -2) of BiDANN shown in Fig. 2. These compared methods and our BiDANN are implemented with the popular Theano.

We verified our BiDANN method on SEED database, which contains 15 subjects' EEG data with 62 channels sorted according to 10-20 system. The EEG recording experiments

are conducted using ESI NeuroScan at a sampling rate of 1000 Hz, and each subject has done the experiments twice separately. During the recording experiments, the participants watched three kinds of film clips that are related to emotions (positive, neutral, negative). Each emotion contains 5 sessions and each session has 185-238 samples [Zheng and Lu, 2015].

Following the same features released in [Zheng and Lu, 2015], we use differential entropy (DE) of EEG signals as the input to feed into our model, which is equivalent to the logarithm energy spectrum in a certain frequency band. DE is calculated in five bands (δ: $1{\sim}3Hz, \theta$: $4{\sim}7Hz, \alpha$: $8 \sim 13Hz, \beta$: $14{\sim}30Hz, \gamma$: $31{\sim}50Hz$), thus it has a dimension of 310.

In any session, we use a slicing window of 9s to temporally scan the sequences by one step. For each step, the sequences in the slicing window are used as the representation of the point which is in the center of the slicing window. By doing this, the temporal dependencies can be involved while recognizing the human emotion at a specific moment. This is quite different from [Zheng and Lu, 2015] which just focuses on recognizing the average energy within a short time ignoring the temporal variation information. Then the input data can be represented as a sequence, $i.e.$, $\mathbf{X}{=}\{\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{d_x \times T}$, where the length of the sequence $T = 9$ and the feature dimension $d_x = 310$. We use the left hemispheric electrodes (FP1, AF3, F7, F5, F3, F1, FT7, FC5, FC3, FC1, T7, C5, C3, C1, TP7, CP5, CP3, CP1, P7, P5, P3, P1, PO7, PO5, PO3, CB1, O1, FPZ, FCZ, CPZ, POZ) of 10-20 system as left hemispheric data $\mathbf{X}^l$, and the symmetric right hemispheric electrodes (FP2, AF4, F8, F6, F4, F2, FT8, FC6, FC4, FC2, T8, C6, C4, C2, TP8, CP6, CP4, CP2, P8, P6, P4, P2, PO8, PO6, PO4, CB2, O2, FZ, CZ, PZ, OZ) as right hemispheric data $\mathbf{X}^r$. Then both the training (source) and testing (target) data are split into two components separately, $i.e.$, $\mathbf{X}_{\mathcal{S}}^l, \mathbf{X}_{\mathcal{S}}^r, \mathbf{X}_{\mathcal{T}}^l, \mathbf{X}_{\mathcal{T}}^r \in \mathbb{R}^{155 \times 9}$. In the experiments, the dimension of hidden state d_h and the length of compressed sequence K are set to 150 and 3 respectively. These parameters are roughly set without elaborate traversal.

3.2 EEG Emotion Recognition on SEED Database

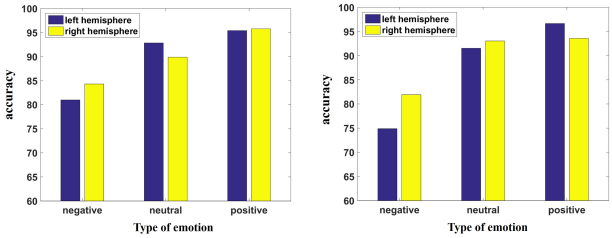
Conventional (subject-dependent) EEG Emotion Recognition

In this experiment, we obey the protocol of Zheng $et\ al.$ [Zheng and Lu, 2015] strictly, which makes 9 sessions of EEG data as training data while 6 sessions as testing data from a same subject. Thus there are totally 1938 samples in training data and 1336 samples in testing data. We calculate the mean accuracy of 15 subjects as the evaluation criterion in our experiment.

Here we compare our BiDANN with the linear SVM [Suykens and Vandewalle, 1999], Canonical Correlation Analysis (CCA) [Thompson, 2005], Group Sparse Canonical Correlation Analysis (GSCCA) [Zheng, 2017], and Deep Believe Network (DBN). These methods have been used in the classification of EEG signals [Zheng and Lu, 2015]. We compare BiDANN with the baseline methods, $i.e.$, DANN, BiDANN-R1 and BiDANN-R2. And we conduct additional experiments using BiDANN-R1 frameworks trained on single hemispheric data to investigate which hemisphere prefers to process emotions.

Method	ACC/STD(%)
SVM [Suykens and Vandewalle, 1999]	83.99/09.72
CCA [Thompson, 2005]	77.63/13.21
GSCCA [Zheng, 2017]	82.96/09.95
DBN [Zheng and Lu, 2015]	86.08/08.34
GraphSLDA [Li *et al.*, 2016]	87.39/08.64
DANN [Ganin *et al.*, 2016]	91.36/08.30
BiDANN-R1	90.29/08.02
BiDANN-R1 (Left)	88.98/08.00
BiDANN-R1 (Right)	89.70/07.63
BiDANN-R2	91.60/08.47
BiDANN	**92.38/07.04**

Table 1: The mean accuracies (ACC) and standard deviations (STD) on SEED database for conventional EEG emotion recognition experiment.



(a) DANN (b) BiDANN-R1

Figure 4: The investigation of hemisphere usage.

Table 1 shows the performance on SEED database. Our BiDANN achieves the state-of-the-art performance. Even compared with DBN, which also is a deep learning method, our method improves 6.3 percent in classification accuracy. In addition, from Table 1, we can see that the methods with domain discriminator (including DANN and BiDANN-R1 frameworks) improve performance compared with other methods. This shows that, for conventional EEG emotion recognition, even the training and testing data come from a same subject, the domain gap always disturbs the decision of classifier. It may be a small difference between EEG signal and other data such as images. Moreover, we can see that the results of BiDANN-R1 (Right) using right hemispheric data get better classification accuracy than BiDANN-R1 (Left) using left hemispheric data, which shows that right hemisphere is superior to the left hemisphere in emotion recognition process. But nether the experimental results of BiDANN-R1 (Right) nor BiDANN-R1 (Left) surpass the performance of BiDANN-R1. This shows that human left and right cerebral hemispheres indeed have dependency in emotion processing.

In addition, we investigate the effect of left and right hemisphere on three types of emotion using DANN and BiDANN-R1 methods. The results are shown in Fig. 4. For negative

Method	ACC/STD(%)
SVM [Suykens and Vandewalle, 1999]	56.73/16.29
KPCA [Schölkopf *et al.*, 1998]	61.28/14.62
TCA [Pan *et al.*, 2011]	63.64/14.88
T-SVM [Collobert *et al.*, 2006]	72.53/14.00
TPT [Sangineto *et al.*, 2014]	76.31/15.89
DANN [Ganin *et al.*, 2016]	75.08/11.18
BiDANN-R1	76.97/11.08
BiDANN-R2	82.22/07.61
BiDANN	**83.28/09.60**

Table 2: The mean accuracies (ACC) and standard deviations (STD) on SEED database for personalized EEG emotion recognition experiment.

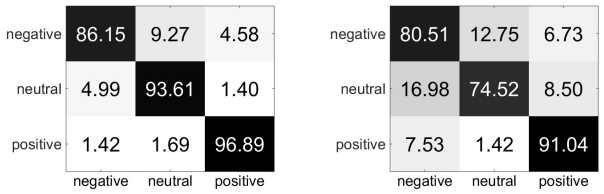
emotion, we can see that using the right hemispheric data has a much better performance than the left ether in DANN or BiDANN-R1, which shows that the right hemisphere can better process negative emotion than left hemisphere does. For positive emotion, the performance of left hemispheric EEG data approximates that of right hemispheric data in the experimental result of DANN method, and in BiDANN-R1 method, the performance of left hemispheric EEG data improves 3% compared with right hemispheric data, which shows that the left hemisphere can better process positive emotion than right hemisphere.

Personalized (Subject-Independent) EEG Emotion Recognition

In this experiment, we adopt a leave-one-subject-out cross validation strategy to evaluate the performance of our model, which is same with the protocol of Zheng $et\ al.$ [Zheng and Lu, 2016]. This strategy takes one subject's EEG as the testing data while the rest 14 subjects' EEG as training data. We calculate the mean accuracy of 15 times experiments as the evaluation criterion.

Here we compare our BiDANN with linear SVM [Suykens and Vandewalle, 1999], KPCA [Schölkopf *et al.*, 1998], TCA [Pan *et al.*, 2011], T-SVM [Collobert *et al.*, 2006], TPT [Sangineto *et al.*, 2014] and the baseline methods, $i.e.$, DANN, BiDANN-R1 and BiDANN-R2. TCA and KPCA are infeasible to include all the training EEG data due to limits of memory and time cost for singular value decomposition. Thus in the experiment, we use the randomly selected 5000 samples as the training data for TCA and KPCA.

Table 2 shows the performance on SEED database. We can see that BiDANN-R2 improves 5.2 percent compared with BiDANN-R1, which shows the importance of considering the discrepancy between left and right cerebral hemispheric data for EEG emotion recognition. Furthermore, BiDANN with local discriminators has a better performance than BiDANN-R2 about 1 percent. This reveals that local discriminators are useful to further narrow the distribution difference between source and target domains on both hemispheres.



(a) The conventional EEG emotion recognition experiment.

(b) The personalized EEG emotion recognition experiment.

Figure 5: The confusion matrices in our experiments.

3.3 Confusion Matrix

To see the results of recognizing each emotion, we depict the confusion matrices corresponding to the experimental results of our BiDANN. Fig. 5 shows the confusion matrices of conventional and personalized EEG emotion recognition experiments on SEED database respectively. From these two figures, we can obtain two observations:

(1) Our BiDANN method performs well in recognizing all three types of emotion, especially the positive emotion as the accuracies are more than 90% either in conventional or personalized EEG emotion recognition tasks. This shows that there indeed exists similarities in the same emotion of EEG signal. It is efficient to use the EEG emotion signal to decode human emotion.

(2) The mean accuracies of three types of emotion in all subjects are negative 86.15%, neutral 93.61%, positive 96.89% in conventional EEG emotion recognition task from Fig. 5(a) and negative 80.51%, neutral 74.51%, positive 91.04% in personalized EEG emotion recognition from Fig. 5(b). We can observe that positive emotion is much easier than negative and neutral emotions to be recognized. In addition, negative and neutral emotions are much more likely to be confused than positive emotion. Maybe the positive emotion stimulus materials cause more resonance in participants.

4 Conclusion

Emotion is a basic and common phenomenon which exists in every human being. The technology of EEG provides a direct means to study emotion by measuring the signal of nerve activity in brain. EEG emotion recognition models should consider the neurophysiology nature of brain and the statistics characteristics of EEG signal. In this paper, we utilize cerebral hemispheric asymmetry to deal with EEG emotion recognition problem and propose a novel EEG emotion recognition framework called BiDANN. BiDANN first extracts time dynamic features of left and right hemispheric EEG data separately and then narrows the distribution gap between training and testing data by using local and global discriminators. The experimental results show that our BiDANN is superior to the baselines even other deep learning methods. In the future work, we will investigate the effect of hemisphere data on more types of emotion.

Acknowledgments

This work was supported by the National Basic Research Program of China under Grant 2015CB351704, the National Natural Science Foundation of China under Grant 61572009, Grant 61772276, and Grant 61602244, and the Jiangsu Provincial Key Research and Development Program under Grant BE2016616.

References

[Britton *et al.*, 2006] Jennifer C Britton, K Luan Phan, Stephan F Taylor, Robert C Welsh, Kent C Berridge, and I Liberzon. Neural correlates of social and nonsocial emotions: An fmri study. *Neuroimage*, 31(1):397–409, 2006.

[Collobert *et al.*, 2006] Ronan Collobert, Fabian Sinz, Jason Weston, and Léon Bottou. Large scale transductive svms. *Journal of Machine Learning Research*, 7(Aug):1687–1712, 2006.

[Etkin *et al.*, 2011] Amit Etkin, Tobias Egner, and Raffael Kalisch. Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, 15(2):85–93, 2011.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[Greve *et al.*, 2013] Douglas N Greve, Lise Van der Haegen, Qing Cai, Steven Stufflebeam, Mert R Sabuncu, Bruce Fischl, and Marc Brysbaert. A surface-based analysis of language lateralization and cortical asymmetry. *Journal of Cognitive Neuroscience*, 25(9):1477–1492, 2013.

[Izard, 1991] Carroll E Izard. *The psychology of emotions*. Springer Science & Business Media, 1991.

[Jenke *et al.*, 2014] Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 5(3):327–339, 2014.

[Kim *et al.*, 2013] Min-Ki Kim, Miyoung Kim, Eunmi Oh, and Sung-Phil Kim. A review on the computational methods for emotional state estimation from the human eeg. *Computational and Mathematical Methods in Medicine*, 2013, 2013.

[Li *et al.*, 2016] Yang Li, Wenming Zheng, Zhen Cui, and Xiaoyan Zhou. A novel graph regularized sparse linear discriminant analysis model for eeg emotion recognition. In *International Conference on Neural Information Processing*, pages 175–182. Springer, 2016.

[Lotfi and Akbarzadeh-T, 2014] Ehsan Lotfi and M-R Akbarzadeh-T. Practical emotional neural networks. *Neural Networks*, 59:61–72, 2014.

[Musha *et al.*, 1997] Toshimitsu Musha, Yuniko Terasaki, Hasnine A Haque, and George A Ivamitsky. Feature extraction from eegs associated with emotions. *Artificial Life and Robotics*, 1(1):15–19, 1997.

[Pan *et al.*, 2011] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[Picard and Picard, 1997] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[Sangineto *et al.*, 2014] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 357–366. ACM, 2014.

[Schölkopf *et al.*, 1998] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[Storbeck and Clore, 2005] Justin Storbeck and Gerald L Clore. With sadness comes accuracy; with happiness, false memory: Mood and the false memory effect. *Psychological Science*, 16(10):785–791, 2005.

[Suykens and Vandewalle, 1999] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[Thompson, 2005] Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[Zatorre *et al.*, 1992] Robert J Zatorre, Marilyn Jones-Gotman, Alan C Evans, and Ernst Meyer. Functional localization and lateralization of human olfactory cortex. *Nature*, 360(6402):339–340, 1992.

[Zheng and Lu, 2015] Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.

[Zheng and Lu, 2016] Wei-Long Zheng and Bao-Liang Lu. Personalizing eeg-based affective models with transfer learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2732–2738. AAAI Press, 2016.

[Zheng, 2017] Wenming Zheng. Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3):281–290, 2017.