EEG-based Emotion Recognition Using Domain Adaptation Network

Yi-Ming Jin, Yu-Dong Luo, Wei-Long Zheng, Bao-Liang Lu*
Center for Brain-Like Computing and Machine Intelligence
Department of Computer Science and Engineering
Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Eng.
Brain Science and Technology Research Center
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai 200240, China
Email: {jinyiming, miyunluo, weilong, bllu}@sjtu.edu.cn

***Abstract*—This paper explores a fundamental problem of eliminating the differences between source subject and target subject in EEG-based emotion recognition. The major limitation of traditional classification methods is that the lack of domain adaptation and subspace alignment will degrade the performance of cross-subject emotion recognition. To address this problem, we adopt Domain Adaptation Network (DAN) for knowledge transfer, which maintains both feature discriminativeness and domain-invariance during training stage. A feed-forward neural network is constructed by augmenting a few standard layers and a gradient reversal layer. Compared with five traditional methods, DAN outperforms its counterparts and achieves the mean accuracy of 79.19%. Moreover, a visualization of the features learned by DAN is represented in this paper, which intuitively describes the transfer virtue of domain adaptation network.**

***Keywords*-EEG; emotion recognition; domain adaptation network; transfer learning;**

I. Introduction

Emotion is normal but essential to humans. It not only contributes to interaction between humans, but also plays a critical role in rational and intelligent behavior. With the development of wearable EEG devices and affective Brain-Computer Interface (aBCI) [1], researchers begin to use EEG for emotion recognition.

A variety of classifiers have been constructed for EEG-based emotion recognition, such as *k*-nearest neighbor (kNN), logistic regression (LR), support vector machine (SVM), and deep belief network (DBN) [2]. However, these classifiers are not suitable for cross subject emotion recognition due to the variability between subjects and sessions.

Domain adaptation (DA) scenario is investigated to address this problem [3]. Typically, training samples with class labels are defined as source domain and testing samples with different distributions are called target domain [4]. In this work, we denote $X \in \mathcal{X}$ as an input of (X, y) from the EEG recording, where $y \in \mathcal{Y}$ is the corresponding emotion label. Let C be the number of channels and d be the number of time series samples, then $\mathcal{X} = \mathcal{R}^{C*d}$ in this circumstance. According to [4], $\mathcal{D} = \{\mathcal{X}, P(X)\}$ is a domain, where $P(X)$ is the marginal probability distribution of X. In our case, this domain contains given subjects from which we recorded the EEG signals. The source and target domains in this paper share the same feature space, $\mathcal{X}_S = \mathcal{X}_T$, but differ in marginal probability distributions, which means $P(X_S) \neq P(X_T)$. The key assumption in most domain adaptation methods is that the conditional probability distributions are the same, i.e. $P(Y_S \| X_S) = P(Y_T \| X_T)$ [5].

A crucial issue for subject transfer is how to reduce the discrepancy between source domain and target domain. Several approaches are proposed to minimize the variability of EEG signals among subjects. Zheng *et al.* [6] applied transfer component analysis (TCA) [7] and kernel principle component analysis (KPCA) [8] for feature selection and reduction. Besides, Zheng *et al.* [5] adopted transductive parameter transfer (TPT) [9] method to personalize EEG-based affective models, which achieves a significant improvement in subject transfer.

Inspired by the transferability of deep neural networks, in this paper we propose a EEG-based emotion recognition method by adopting domain adaptation network (DAN) which learns domain-invariant features by backpropagation. We compare the performance of our work with three state-of-the-art approaches, TCA, KPCA and TPT.

II. Method

A. Domain Adaptation Network

We use a domain adaptation network (DAN) to achieve transfer learning. This framework is proposed by Yaroslav Ganin *el at.* [10] for image classification. This architecture consists of three parts as shown in Figure 1. We denote $\mathcal{S}(x)$ and $\mathcal{T}(x)$ as the marginal distributions of source and target domain, respectively. Given the training dataset $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ concatenated from both source and target data, we introduce a binary variable $d_i \in \{0, 1\}$ serving as the domain label for the i-th sample. It points out which domain distribution x_i comes from, i.e. $\boldsymbol{x}_i \sim \mathcal{S}(x)$ if $d_i = 0$ and $\boldsymbol{x}_i \sim \mathcal{T}(x)$ if $d_i = 1$.

For each input data $\boldsymbol{x}$, a feature vector $f \in \mathbf{R}^D$ is firstly learned by a feature extractor M_f, i.e. $f = M_f(\boldsymbol{x}, \theta_f)$, where θ_f denotes parameters of all layers in this mapping.

*Corresponding author: Bao-Liang Lu. This work was supported by the National Key Research and Development Program of China (2017YFB1002501), the National Natural Science Foundation of China (61272248), the Major Basic Research Program of Shanghai Science and Technology Committee (15JC1400103), the ZBYY-MOE Joint Funding (6141A02022604), and the Technology Research and Development Program of China Railway Corporation (2016Z003-B).

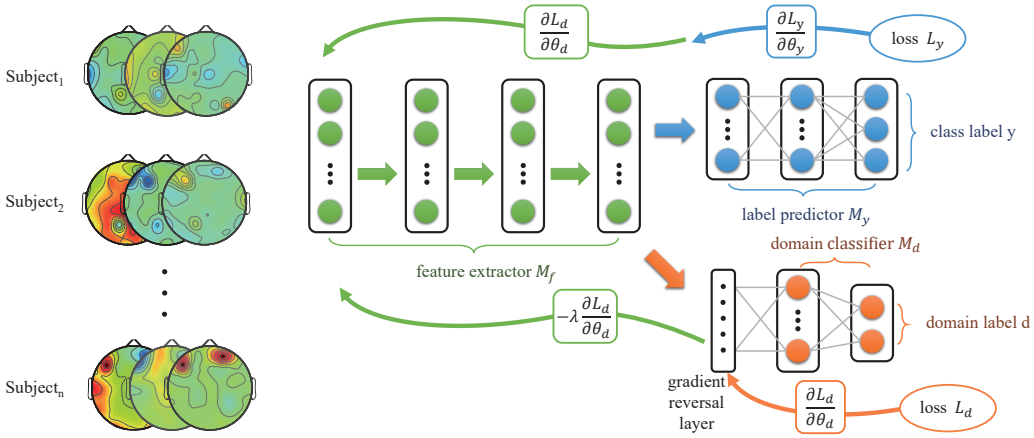

Figure 1. Domain Adaptation Network contains a feature extractor (green), a label predictor (blue), a domain classifier (red) and a gradient reversal layer (black). Feature extractor and label predictor together form a traditional feed-forward neural network. These two parts can be used to predict the label. Domain adaptation is achieved by feature extractor and domain classifier by backpropagation. The gradient conveyed from domain classifier to feature extractor is multiplied by a negative parameter in gradient reversal layer. In this way, traditional gradient descent method can be applied. Label predictor ensures the data discrimitiveness and domain classifier maintains the domain-invariance.

Then the feature vector f is mapped by two parallel networks. The label predictor M_y maps f with parameters θ_y and the domain classifier M_d maps f with parameters θ_d.

The label predictor M_y is used to predict labels of the source domain. In order to get a good prediction performance, we intend to minimize the loss of label predictor on the the source domain. Consequently, both parameters θ_f and θ_y are adjusted during the training time aiming to minimize the experiential loss for source domain samples.

The domain classifier M_d indicates the domain label d for each input data. The network learns domain-invariant feature $\boldsymbol{f}$ from input data, which means the distributions $\mathcal{S}(\boldsymbol{f}) = \{M_f(\boldsymbol{x}, \theta_f) \| \boldsymbol{x} \sim \mathcal{S}(\boldsymbol{x})\}$ and $\mathcal{T}(\boldsymbol{f}) = \{M_f(\boldsymbol{x}, \theta_f) \| \boldsymbol{x} \sim \mathcal{T}(\boldsymbol{x})\}$ should be similar. According to *covariance shif* assumption [11], the features are transferable and the label prediction accuracy on target domain matches that on source domain in this case. To tackle this problem, the loss of domain classifier M_d is used to measure the domain discrepancy.

The loss function of the architecture can be computed as

$$\begin{aligned} E(\theta_f, \theta_y, \theta_d) &= \sum_{\substack{i=1\ldots N \\ d_i=0}} L_y(M_y(M_f(x_i;\theta_f);\theta_y), y_i) - \\ &\quad \lambda \sum_{i=1\ldots N} L_d(M_d(M_f(x_i;\theta_f);\theta_d), y_i) \\ &= \sum_{\substack{i=1\ldots N \\ d_i=0}} L_y^i(\theta_f, \theta_y) - \lambda \sum_{i=1\ldots N} L_d^i(\theta_f, \theta_d) \end{aligned} \tag{1}$$

where $L_y(\cdot,\cdot)$ is the loss of class label prediction, $L_d(\cdot,\cdot)$ represents the loss of domain label classification, L_y^i and L_d^i denote the corresponding loss functions of i-th training example, and λ is the trade-off parameter that balances the two objectives.

During learning process, a method similar to the standard stochastic gradient descent (SGD) is adopted to search the saddle point: $\theta_f \leftarrow \theta_f - \mu(\frac{\partial L_y^i}{\partial \theta_f} - \lambda\frac{\partial L_d^i}{\partial \theta_f})$, $\theta_y \leftarrow \theta_y - \mu\frac{\partial L_y^i}{\partial \theta_y}$, $\theta_d \leftarrow \theta_d - \mu\frac{\partial L_d^i}{\partial \theta_d}$, where μ denotes the learning rate.

The updating process is similar to SGD except for parameters θ_f which combine gradients from label predictor and domain classifier. However, directly implementing it using SGD is difficult. Ganin *et al.* [10] adds a gradient reversal layer (GRL) between feature extractor and domain classifier. During forward propagation, GRL acts as an identity transformation which does not change the input parameters. While in backpropagation, GRL multiplies $-\lambda$ to the gradient from subsequent layer and passes the result to the preceding layer. More precisely, the GRL can be regarded as a pseudo-function R_λ: $R_\lambda = x$, $\frac{dR_\lambda}{dx} = -\lambda \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix. With the help of such transformation, the loss function E can be further written as:

$$\begin{aligned} \tilde{E}(\theta_f, \theta_y, \theta_d) = &\sum_{\substack{i=1\ldots N \\ d_i=0}} L_y(M_y(M_f(\boldsymbol{x_i};\theta_f);\theta_y), y_i) + \\ &\sum_{i=1\ldots N} L_d(M_d(R_\lambda(M_f(\boldsymbol{x}_i;\theta_f));\theta_d), y_i) \end{aligned} \tag{2}$$

After optimization, the feature extractor $M_f(\cdot,\cdot)$ and label predictor $M_y(\cdot,\cdot)$ can be used in target domain to predict the label.

III. Experiment Setup

A. EEG Dataset

In this paper, we evaluate the performance of domain adaptation network (DAN) using a publicly available EEG dataset for emotion recognition called SJTU Emotion EEG dataset (SEED)[1] downloaded from the project website. Different from other existing EEG dataset, each participant

[1]http://bcmi.sjtu.edu.cn/~seed/

performs experiments three times at intervals of one week or longer and is asked to assess his or her real emotional reactions right after watching a movie chip. Only the data with explicit corresponding emotions are used for analysis. For more details of SEED, we recommend readers to refer to the literature [2].

B. Signal Preprocessing and Feature Extraction

For signal preprocessing, since raw EEG signals are often contaminated by electromyography (EMG) signals and electrooculogram (EOG) signals [12], a bandpass filter between 1 Hz and 75 Hz is used to filter out noise and artifacts. To reduce the data size, EEG signals are further down-sampled to 200 Hz.

For feature extraction, differential entropy (DE) features are extracted from raw data since DE outperforms power spectral density (PSD) [13]. DE features contain five frequency bands: delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz), and gamma (31-50 Hz). Each frequency band contains the features of 62 EEG channels, thus the total dimensions of extracted EEG features are 310.

C. Parameter Details

All the evaluations are conducted using leave-one-subject-out cross validation. Each time, one subject is selected as target domain out of 15 subjects and the rest 14 remain as source domain. Classification accuracy is evaluated on the target domain.

Baselines. We adopt Support Vector Machine (SVM) as one of our baseline. We use a linear kernel and search the parameter C from -10 to 10 with step of 1. To compare with DAN, we also adopt traditional multilayer perceptron named NN as the other baseline. The number of hidden layers is 3, the number of neurons in each layer is set by searching [50, 100, 200, 400] and the learning rate is set to 0.1.

TCA & KPCA. Due to the limitation of time and memory, we randomly select 5000 samples from 14 subjects as source domain data while the target domain data remain the same. The kernel function is linear kernel and the regularization parameter μ is set to 1.

TPT. Multiple classifiers are learned on each source domain data using linear Support Vector Machine (lSVM), and the regularization parameter is set to 0.1. Multioutput Support Vector Regression (M-SVR) framework is utilized to learn the final mapping function $f(\cdot)$.

DAN. The feature extractor has two hidden layers with 100 neurons in each layer. The label predictor has one hidden layer with 100 neurons and a output layer. The domain classifier has one hidden layer with 100 neurons, one gradient reversal layer and one output layer. The learning rate is set to 0.1 and the parameter λ in gradient layer is set to 1.

IV. Experiment Results

In this part, we will discuss experiment results of six different methods on SEED. We adopt leave-one-subject-out evaluation scheme in each case.

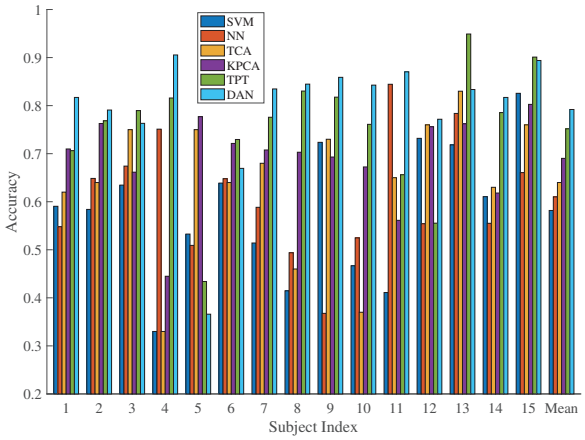

Figure 2. The accuracies of six methods (SVM, NN, TCA, KPCA, TPT, DAN) for each subject and the mean accuracies.

Table I
MEAN ACCURACIES AND STANDARD DEVIATIONS OF THE SIX DIFFERENT METHODS

Methods	SVM	NN	TCA	KPCA	TPT	DAN
Mean	0.5818	0.6101	0.6400	0.6902	0.7517	**0.7919**
Std.	0.1385	0.1238	0.1466	**0.0925**	0.1283	0.1314

Figure 2 shows the accuracies of six different methods (SVM, NN, TCA, KPCA, TPT and DAN) for 15 subjects and Table I presents mean accuracies and standard deviations in detail. SVM has the lowest accuracy of 58.18%. Because the EEG signals vary between subjects, individual classifiers learned by SVM in source domain cannot distinguish emotion states in target domain precisely. By comparison, TCA and KPCA extract latent common components from both source and target data with low domain variance. These extracted features can improve the performance of classification. TPT achieves accuracy of 75.17%. It learns a mapping function that can transfer the classifiers trained in source domain to target domain instead of looking for a latent feature space between source and target data. Our DAN outperforms these method with the mean accuracy of **79.19**%. The standard deviation of DAN is not the smallest, because the result of subject 5 is extremely low.

To illustrate the transferability and discriminativeness of DAN learned features, we visualize the output of the second hidden layer from DAN and NN by projecting features into a 3-dimensional space using t-SNE [14] as shows in Figure 3. We then make the following observations: (1) NN features from different subjects are scattered in three emotion states, while DAN features are aligned much better between source and target domains. (2) With NN features, the datapoints are not discriminative enough, while with DAN features, all of the three emotion states can be evidently distinguished. These observations explain the superior performance of DAN over NN: (1) indicates that DAN can reduce the differences between source and target domain to learn domain-invariant features, which makes source classifiers more suitable for target data. (2)

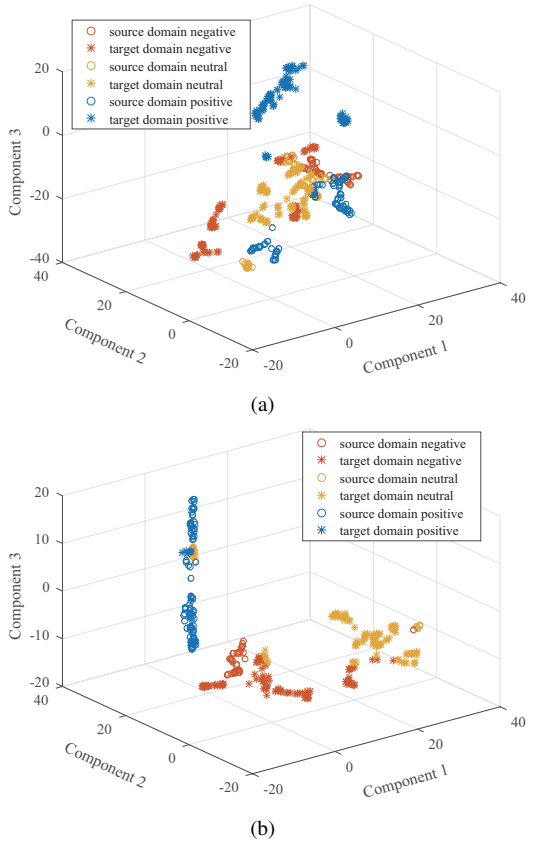

Figure 3. The effect of domain adaptation on the distributions of extracted features. Figures (a) and (b) denote NN and DAN, respectively, which both present all three different emotion states together from different domains.

implies that different emotion states are more easily identified with DAN features, which guarantees the accuracy of label predictor.

V. Conclusion and Future Work

In this paper, we have proposed a neural network architecture for EEG-based emotion recognition with transfer learning techniques. Domain-invariant features are learned by backpropogation after adding a gradient reversal layer to domain classifier. We have compared the performance with five different methods: SVM, NN, TCA, KPCA, and TPT for SEED. The experimental results show that domain adaptation network (DAN) exceeds the other approaches in terms of accuracy, and achieves a 21.01% increase compared with SVM and a 18.18% increase compared with NN. We also search different parameters for DAN to find the optimal values. Our future work will focus on applying our proposed model to more categories of emotions, as well as to domain adaptation on subjects from different culture backgrounds.

References

[1] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[2] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[3] L. Duan, D. Xu, and I. W.-H. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 504–518, 2012.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2732–2738.

[6] W.-L. Zheng, Y.-Q. Zhang, J.-Y. Zhu, and B.-L. Lu, "Transfer components between subjects for EEG-based emotion recognition," in *International Conference on Affective Computing and Intelligent Interaction*. IEEE, 2015, pp. 917–922.

[7] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[8] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[9] E. Sangineto, G. Zen, E. Ricci, and N. 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*. ACM, 2014, pp. 357–366.

[10] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International Conference on Machine Learning*, 2015, pp. 1180–1189.

[11] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[12] M. Fatourechi, A. Bashashati, R. K. Ward, and G. E. Birch, "EMG and EOG artifacts in brain computer interface systems: A survey - clinical neurophysiology," *Clinical Neurophysiology*, vol. 118, no. 3, pp. 480–494, 2007.

[13] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *6th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2013, pp. 81–84.

[14] L. Van Der Maaten, "Barnes-hut-sne," in *Proceedings of the First International Conference on Learning Representations*, 2013.