Plug-and-Play Domain Adaptation for Cross-Subject EEG-based Emotion Recognition

Li-Ming Zhao,[1] **Xu Yan,** [3] **Bao-Liang Lu** [1 2*]

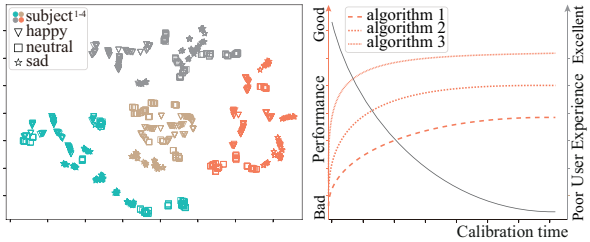
[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, 200240
[2] Center for Brain-Machine Interface and Neuromodulation,
Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China, 200020
[3] Department of Linguistics, University of Washington, Seattle, WA, USA, 98195
{lm_zhao, bllu}@sjtu.edu.cn, xyan3@uw.edu

Abstract

Human emotion decoding in affective brain-computer interfaces suffers a major setback due to the inter-subject variability of electroencephalography (EEG) signals. Existing approaches usually require amassing extensive EEG data of each new subject, which is prohibitively time-consuming along with poor user experience. To tackle this issue, we divide EEG representations into *private components* specific to each subject and *shared emotional components* that are universal to all subjects. According to this representation partition, we propose a plug-and-play domain adaptation method for dealing with the inter-subject variability. In the training phase, subject-invariant emotional representations and private components of source subjects are separately captured by a shared encoder and private encoders. Furthermore, we build one emotion classifier on the shared partition and subjects' individual classifiers on the combination of these two partitions. In the calibration phase, the model only requires few unlabeled EEG data from incoming target subjects to model their private components. Therefore, besides the shared emotion classifier, we have another pipeline to use the knowledge of source subjects through the similarity of private components. In the test phase, we integrate predictions of the shared emotion classifier with those of individual classifiers ensemble after modulation by similarity weights. Experimental results on the SEED dataset show that our model greatly shortens the calibration time within a minute while maintaining the recognition accuracy, all of which make emotion decoding more generalizable and practicable.

Introduction

The emerging affective computing aims at detecting, recognizing, processing, and responding to people's affected states. It has broad prospects in many fields of applications in daily life, ranging from specific scenarios such as medical treatment, intelligent education, and entertainment to general affect-sensitive systems like brain-computer interfaces (BCIs), among which emotion recognition is the primary step and the milestone (Brunner et al. 2015). Recently, EEG-based emotion recognition has greatly attracted researchers' interest for its information sufficiency (Alarcao and Fonseca 2017) and stable neural patterns over time (Zheng, Zhu, and Lu 2019). However, EEG data is highly subject-dependent due to the structural and functional variability between subjects (Samek, Meinecke, and Müller 2013), like mental states, electrode impedance, head shapes, etc. Figure 1(a) illustrates the inter-subject variability of emotional EEG data, which brings great challenges of constructing practical EEG-based affective models. This shortage definitely hinders developments and applications of affective computing on a large scale.



(a) (b)

Figure 1: Illustration of the domain-shift challenges and the dilemma of constructing practical EEG-based affective models. (a) Individual differences of EEG signals of four subjects with three classes of emotions. (b) A rough display of the trade-off dilemma between algorithm performance and user experience.

The obstacle mentioned above has motivated many researchers to develop practical emotion recognition algorithms. The conventional method to deal with this problem is to collect a large amount of data from the new subject, label them, and use them to customize the classifier parameters before the test stage. Unfortunately, this demand is time-consuming and causes poor user experience, which makes the model less practical. Another route is using transfer learning methods (Pan and Yang 2009) to deal with the individual differences. The transfer learning can be roughly divided into domain adaptation (DA) and domain generalization (DG) according to whether the data of the target domain is used in the model training phase. For the practical application of emotion recognition, DA is inefficient due to the use of all target data, and DG may suffer from its generalization ability as it does not rely on any information from

*Corresponding author

target subjects. Contrary to the extremes of the DA and DG, it is acceptable and necessary to introduce a short-term calibration stage before the real-time recognition starts. Figure 1(b) subjectively exhibits a trade-off dilemma between algorithm performance and user experience. However, existing research has indicated that if the number of training data is small compared to the dimensions of the feature vectors, the model will most probably break down (Lotte et al. 2007). Thus, it is challenging to achieve good DA results with limited target training data.

To address the problems above, we propose a plug-and-play domain adaptation method that can calibrate with few unlabeled target data without sacrificing the recognition accuracy. We hypothetically divide EEG representations into shared emotional components that are universal to all subjects and private components specific to each subject. We use Long Short-Term Memory Auto-Encoder Neural Network and explicit loss functions to separate the private components, and in the process produce representations that are more meaningful for emotion recognition. However, we believe the single shared classifier built in shared emotional space still has limited ability for the new subject never been seen. Thus, we additionally build a series of individual classifiers for existing source subjects, the purpose of which is to provide a reference for new subjects. By reconstructing the few calibration data, we can quickly construct new subjects' private encoder together with the trained shared encoder and decoder enforced. Thus, the target subjects can borrow the knowledge from source individual classifiers through the similarity of private components and strengthen emotion prediction together with the shared classifier. The private components are the main reasons for the inter-subject variability and remain unchanged within one collection once the EEG sensors are set up, which is the key to shorten calibration time. In addition, we hope to use the attention mechanism to automatically learn the critical EEG channels and frequency bands most relevant to emotion recognition.

Related Work

The debut of transfer learning has received a surge of attention and has quickly become an important method to deal with inter-subject variability in BCI. There are two main branches in transfer learning that can help reduce the subject-invariability. One is domain adaptation (DA) (Ben-David et al. 2010). DA methods increase the accuracy of the target data by minimizing domain shifts between source and target domains, indicating that during the training phase, we must have got the data from the target domain. Zheng and Lu (2016) first introduced the transductive parameter transfer (TPT) (Sangineto et al. 2014), a kind of DA method, to EEG-based cross-subject emotion recognition and got a decent result. Especially, the Domain-Adversarial Neural Networks (DANNs) (Ganin et al. 2016), trying to find the shared representations among all domains specific to the task, pushed the accuracy up to another record high (Li et al. 2018). Bousmalis et al. (2016) noticed that in the core idea of DANNs, the shared representations are sensitive to the noise correlated with the underlying shared distribution, and proposed the novel Domain Separation Networks (DSN) to learn domain-invariant representations. Nevertheless, no matter how DA achieves knowledge transfer, all methods demand all target information, which is applicable to the offline datasets transfer, but cannot be reached in real-time BCI applications.

This implicit shortage urges researchers to turn to domain generalization (DG) for help. DG methods can extract domain-invariant features by exploiting domain differences across multiple source subjects without the need to acquire any data from the target subjects (Blanchard, Lee, and Scott 2011). Domain-Invariant Component Analysis (DICA) (Muandet, Balduzzi, and Schölkopf 2013) and Scatter Component Analysis (SCA) (Ghifary, Balduzzi, Kleijn, and Zhang 2016) are two examples. Ma et al. (2019) compared the results of these two approaches and proposed a novel framework called Domain Residual Network (DResNet). In this model, weights are explicitly divided into biased weights that are exclusive to each domain and unbiased weights that are shared by all domains. DResNet, a robust model with better generalization ability for incoming domains gained from unbiased weights, can speak for the state-of-the-art performance of DG methods that are often used in EEG-based emotion recognition tasks.

While DG methods seem more likely to be implemented in the real world, some questions are still worth thinking about. Is the DG methods' restriction of no demand for target data the most suitable for real-time application scenarios? Although long-term calibration will result in poor user experience, we can still collect few target data through shorter-term calibration to adapt to target subject quickly. Li et al. (2019) utilized the style transfer mapping (STM) method to reduce the domain differences with the support of a small amount of labeled target data. They used three sessions of labeled EEG data about ten minutes in total during calibration stage for three-kind emotion recognition task. Li et al. have achieved excellent results, but the 10-minute calibration time is still long for practical use. What's more, STM requires labeled data to cover all categories, which means that the calibration time will be further increased with the expansion of emotion categories.

Anothor key point that influences the adaptation time is that EEG signals contain too much information. Ahern and Schwartz (1985) found that our brain shows lateralization in emotions, and some regions and frequency bands are more informative than others with different emotions. Zheng and Lu (2015) confirmed the existence of relevance between neural signatures and different emotions. Besides, they also assigned the critical channels and frequency bands according to the weights given by the trained Deep Belief Networks in emotion recognition tasks. By contrast, the prevalent attention mechanism originated from image processing (Mnih et al. 2014) and natural language processing (Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017) offers the new possibility to solve the problem.

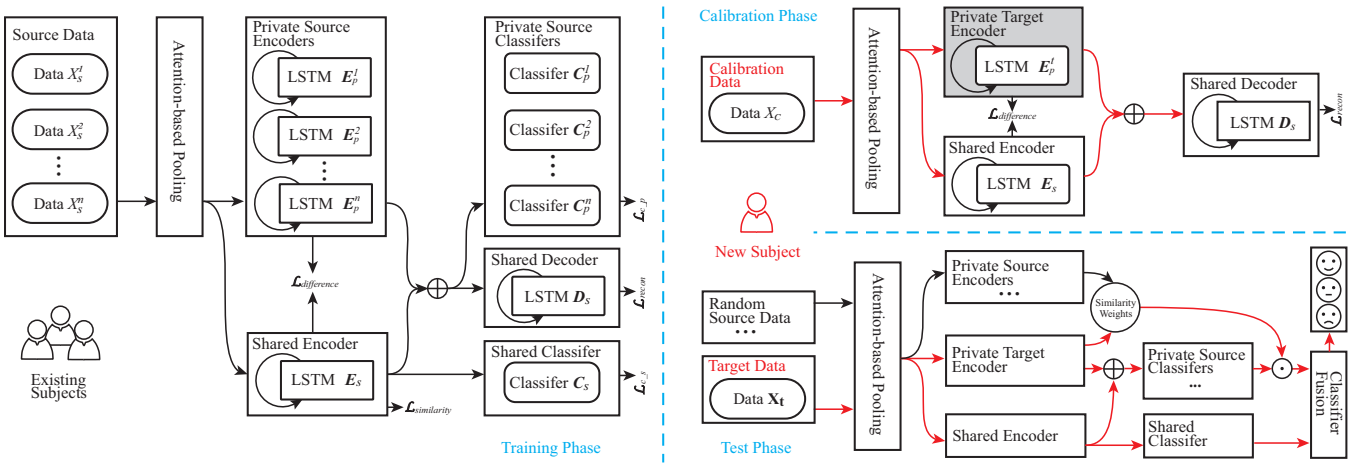


Figure 2: The framework of proposed PPDA. The whole structure can be divided into training phase, calibration phase, and test phase. The submodules in the training phase will be optimized enforced by the combination of several loss functions. In the calibration phase, only the private target encoder, highlighted in gray, will be updated. In the test phase, the final predictions will be made by two pipelines. The red directional lines mark the data flow of new subject, while dark mark the source.

Methods

Overview

To get over the inter-subject variability of EEG signals and execute the task of rapid emotion recognition on new subjects, we propose a novel plug-and-play domain adaptation (PPDA) method. The framework of PPDA is depicted in Figure 2. The whole structure can be divided into a training phase, a calibration phase, and a test phase. In the training phase, the attention-based pooling is first applied to utilize the spatial information of critical channels and bands of EEG signals. Then, Long Short-Term Memory based encoder-decoder scheme is adopted to explore the temporal dependency. We raise a shared encoder $\boldsymbol{E}_s$ and private encoders $\boldsymbol{E}_p^{1\sim n}$ to capture the subject-invariant emotional representations and private components, respectively. By using the outputs of the encoders, we further build a shared classifier $\boldsymbol{C}_s$ and individual classifiers $\boldsymbol{C}_p^{1\sim n}$ to recognize emotions simultaneously. In this stage, only labeled source data are adopted to train the model. In the calibration phase, we use the data from the very beginning of one collection to model the private component of the new subject with the help of a trained $\boldsymbol{E}_s$ and decoder $\boldsymbol{D}_s$, which we call the calibration phase. Therefore, in the test phase, we can not only use the pipeline of the shared classifier as domain generalization methods do, but also obtain knowledge from the private classifiers through the similarity with private source components. Finally, the classifier fusion strategy is applied to integrate the two recognition results. To make the description clearer and avoid confusion, we summarize the notations and our algorithm in Table 1 and Algorithm 1, respectively.

Attention-based Pooling

Inspired by the biological evidence from Ahern and Schwartz (1985) and the previous work of Zheng and Lu (2015), we attempt to introduce the attention mechanism to let the model automatically explore the critical channels and bands for emotion recognition. We use $x_t \in R^m$ as the annotation of one EEG feature vector at time t, where m is the feature dimensionality. Each dimension of x_t represents information from a specific channel at a band. We obtain the weighted EEG feature vector $\widetilde{x}_t$ with $\widetilde{x}_t$=$AT(x_t)$, where AT means the Attention-based Pooling. Specifically, x_t is input into a single layer fully connected neural network, and the normalized weight vector $\alpha_t \in R^m$, representing the importance of each dimension of x_t, is measured through a softmax function as:

$$\alpha_t = softmax(W_a x_t + b_a). \tag{1}$$

After that, we calculate $\widetilde{x}_t$ as the weighted new EEG feature as:

$$\widetilde{x}_t = \alpha_t \cdot x_t. \tag{2}$$

For each element in α_t, the larger the value, the more crucial its corresponding dimension, i.e. the channel at this band. The weight matrix $W_a \in R^{m\times m}$ and the bias vector $b_a \in R^m$ are randomly initialized and fine-tuned during the training process.

Notation	Meaning
$\mathbf{X}_s$=$\{X_s^j, Y_s^j\}_{j=1}^n$	source labeled dataset of n subjects
$\boldsymbol{X}_c, \boldsymbol{X}_t$	unlabeled calibration data and target data
$\boldsymbol{E}_s, \boldsymbol{D}_s, \boldsymbol{C}_s, \boldsymbol{C}_d$	shared encoder, decoder, emotion classifier, and domain classifier
$\boldsymbol{E}_p^j, \boldsymbol{C}_p^j$	private source encoders and emotion classifiers, with $j = 1, 2, ..., n$
$\boldsymbol{E}_p^t$	private encoder of the target subject

Table 1. Notation Summary

Algorithm 1: Plug-and-play domain adaptation

Input:
Source data $\mathbf{X}_s$=$\{X_s^j, Y_s^j\}_{j=1}^n$.
Target calibration data X_c from time 0 to T.
Target test data X_t.
Output: Recognition accuracy of target subject data.
Training Phase:
Randomly initialize $\boldsymbol{E}_p^{1\sim n}$, $\boldsymbol{E}_s$, $\boldsymbol{C}_p^{1\sim n}$, $\boldsymbol{C}_s$, and $\boldsymbol{D}_s$.
for *j=1:n* **do**
  Optimize $\boldsymbol{AT}$, $\boldsymbol{E}_p^j$, $\boldsymbol{E}_s$, $\boldsymbol{C}_p^j$, $\boldsymbol{C}_s$, and $\boldsymbol{D}_s$ by minimizing Equation (5).
end
return $\boldsymbol{AT}$, $\boldsymbol{E}_p^{1\sim n}$, $\boldsymbol{E}_s$, $\boldsymbol{C}_p^{1\sim n}$, $\boldsymbol{C}_s$, and $\boldsymbol{D}_s$.
Calibration Phase:
Randomly initialize $\boldsymbol{E}_p^t$.
Obtain the trained $\boldsymbol{AT}$, $\boldsymbol{E}_s$, and $\boldsymbol{D}_s$.
Calculate $\widetilde{X}'$=$\boldsymbol{D}_s(\boldsymbol{E}_s(\boldsymbol{AT}(X_c)) + \boldsymbol{E}_p^t(\boldsymbol{AT}(X_c)))$
Optimize $\boldsymbol{E}_p^t$ through minimizing Equation (8).
return $\boldsymbol{E}_p^t$.
Test Phase :
for *target data series* $\mathbf{x}_t$ *in* X_t **do**
  Randomly select $X_{rand}^{1\sim n}$ from $\mathbf{X}_s$.
  Calculate the similarity weight w_s between $\boldsymbol{E}_p^{1\sim n}(\boldsymbol{AT}(X_{rand}^{1\sim n}))$ and $\boldsymbol{E}_p^t(\boldsymbol{AT}(\mathbf{x}_t))$
  Prediction of weighted private source classifiers:
  $\hat{y}_p^t$=$w_s \cdot \boldsymbol{C}_p^{1\sim n}(\boldsymbol{E}_p^t(\boldsymbol{AT}(\mathbf{x}_t)) + \boldsymbol{E}_s(\boldsymbol{AT}(\mathbf{x}_t)))$
  Prediction of shared classifier:
  $\hat{y}_s^t$=$\boldsymbol{C}_s(\boldsymbol{E}_s(\boldsymbol{AT}(\mathbf{x}_t)))$
  Integrate predictions: $\hat{y}^t$=$\boldsymbol{CF}(\hat{y}_p^t, \hat{y}_s^t)$
end
return $\hat{y}^t$.

LSTM-based Encoder-Decoder

As LSTM neural networks have demonstrated their effectiveness for extracting temporal dependencies in EEG-based emotion recognition (Kim and Jo 2020), we choose LSTM to construct the Encoder-Decoder architecture. For each element in the input series, the LSTM unit computes the following functions in Equation (3):

$$\begin{aligned} i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}), \\ f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}), \\ g_t &= tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}), \\ o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}), \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ h_t &= o_t \odot tanh(c_t), \end{aligned} \tag{3}$$

where i_t, f_t, g_t, o_t are the input, forget, cell, and output gates. h_t and c_t are the hidden state and the cell state at time t, while h_{t-1} is the hidden state of the layer at time $t-1$ or the initial hidden state at the very beginning. σ represents the sigmoid function, and $\odot$ is the Hadamard product.

Two kinds of encoders are designed to extract the shared emotional components and private components of EEG representations separately. Depending on the concatenation of these two components, a shared decoder is applied to reconstruct the input EEG representations.

Consider an EEG series $\widetilde{\mathbf{x}} = \{\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_l\}$ of time step l, where each point $\widetilde{x}_i \in R^m$ is an m-dimensional EEG feature modulated by the attention mechanism of one subject. Figure 3 depicts the inference steps in an LSTM Encoder-Decoder reconstruction model for an EEG series with $l = 4$. The EEG feature $\widetilde{x}_i$ at time t_i and the hidden state $h_{E_s}^{i-1}$ of the shared encoder at time t_{i-1} are used to calculate $h_{E_s}^i$. The hidden state $h_{E_p}^i$ of the private encoder is simultaneously calculated. The combination of hidden states $h_{E_s}^4$ and $h_{E_p}^4$ initializes $h_{D_s}^4$ of the shared decoder as:

$$h_{D_s}^4 = h_{E_s}^4 + h_{E_p}^4. \tag{4}$$

Just like Sutskever, Vinyals, and Le (2014) did, the decoder reconstructs the EEG feature series in a reverse order, i.e. $\{\widetilde{x}_4', \widetilde{x}_3', \widetilde{x}_2', \widetilde{x}_1'\}$. A linear layer on top of the decoder is used to establish a mapping between $h_{D_s}^i$ and $\widetilde{x}_i'$ to ensure the reconstructed EEG feature $\widetilde{x}_i'$ has the same dimensions with the input EEG feature $\widetilde{x}_i$. Then, the shared decoder uses $h_{D_s}^i$ and $\widetilde{x}_i'$ to infer $\widetilde{x}_{i-1}'$. In particular, $\widetilde{x}_4$ is used to fire the decoding process and thereby form the entire series gradually. In order to distinguish it from $\widetilde{x}_i$, we use $\widetilde{\mathbf{x}}_i$ to represent the EEG series $\{\widetilde{x}_{i-l+1}, \widetilde{x}_{i-l+2}, ..., \widetilde{x}_i\}$ with a time step of l at time t_i, which is the basic input unit of LSTM. To clarify, $\widetilde{\mathbf{x}}_i^j$ stands for the EEG series from subject j at time t_i, and we use the emotion label y_i^j of $\widetilde{x}_i$ as the label of the entire series.

Learning Loss

In the training phase, only labeled EEG data of existing source subjects are utilized to train the model, which aims to minimize to following loss:

$$\mathcal{L} = \mathcal{L}_{\text{c\_s}} + \alpha\mathcal{L}_{\text{c\_p}} + \beta\mathcal{L}_{\text{recon}} + \gamma\mathcal{L}_{\text{difference}} + \delta\mathcal{L}_{\text{similarity}}, \tag{5}$$

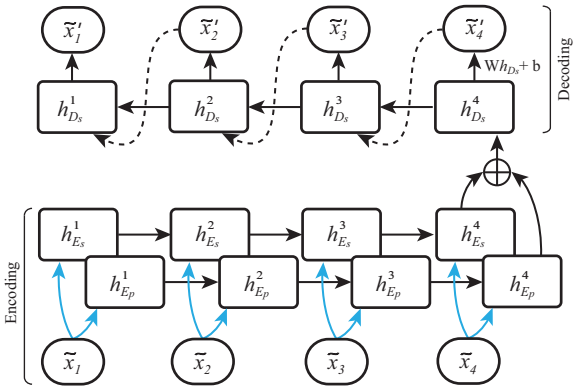


Figure 3: The LSTM-based Encoder-Decoder reconstruction with EEG series input $\{\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3, \widetilde{x}_4\}$ and output $\{\widetilde{x}_1', \widetilde{x}_2', \widetilde{x}_3', \widetilde{x}_4'\}$. The blue lines and black dashed lines are applied to avoid unsightly intersections.

where $\alpha, \beta, \gamma, \delta$ are trade-offs that control the synergy of the loss terms. We minimize the cross-entropy loss of emotion classifiers as:

$$\begin{aligned}\mathcal{L}_{\text{c\_s}} &= -\sum_{i,j} y_i^j \log \hat{y}_{i,s}^j, \\ \mathcal{L}_{\text{c\_p}} &= -\sum_{i,j} y_i^j \log \hat{y}_{i,p}^j,\end{aligned} \tag{6}$$

where y_i^j is the ground truth emotion label for input x_i^j from specific j^{th} subject. $\hat{y}_{i,s}^j$ and $\hat{y}_{i,p}^j$ are the softmax predictions of the shared classifier and the corresponding private classifier:

$$\begin{aligned}\hat{y}_{i,s}^j &= \boldsymbol{C}_s(\boldsymbol{E}_s(\widetilde{\mathbf{x}}_i^j)), \\ \hat{y}_{i,p}^j &= \boldsymbol{C}_p^j(\boldsymbol{E}_s(\widetilde{\mathbf{x}}_i^j) + \boldsymbol{E}_p^j(\widetilde{\mathbf{x}}_i^j)).\end{aligned} \tag{7}$$

We use the mean squared error to calculate the reconstruction loss $\mathcal{L}_{\text{recon}}$:

$$\mathcal{L}_{\text{recon}} = \frac{1}{k} \left\| \widetilde{X} - \widetilde{X}' \right\|_2^2, \tag{8}$$

where k is the number of EEG features, and $\|\cdot\|_2^2$ is the squared L_2-norm. The difference loss $\mathcal{L}_{\text{difference}}$ is applied to encourage the shared and private encoders to encode different aspects of the inputs:

$$\mathcal{L}_{\text{difference}} = \frac{1}{n} \sum_{j=1}^{n} \left\| \mathbf{H}_s^{j\top} \mathbf{H}_p^j \right\|_F^2, \tag{9}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm, $\mathbf{H}_s^j = \boldsymbol{E}_s(\widetilde{X}^j)$, and $\mathbf{H}_p^j = \boldsymbol{E}_p^j(\widetilde{X}^j)$. Driven by the idea of extracting the subject-invariant emotional representations, we train a domain classifier $\boldsymbol{C}_d$ to confuse the shared encoder via a Gradient Reversal Layer (GRL). The GRL works as an identity function during forward propagation, but reverses the gradient direction in backward. The $\mathcal{L}_{\text{similarity}}$ is calculated as:

$$\mathcal{L}_{\text{similarity}} = \sum_i d_i \log(\hat{d}_i), \tag{10}$$

where d_i is the ground truth domain label and $\hat{d}_i = \boldsymbol{C}_d(\boldsymbol{C}_s(\widetilde{\mathbf{x}}_i))$.

Calibration and Test

Since the EEG data is chronologically recorded, we can only take the data from the very beginning as the calibration data. We first initialize the parameters randomly of the private target encoder $\boldsymbol{E}_p^t$, and optimize them with the reconstruction loss and the different loss by Equation (8) and Equation (9) using the calibration data. We believe that once the task is settled, the shared encoder $\boldsymbol{E}_s$ is generalized enough and would have the ability to extract subject-invariant emotion components, and $\boldsymbol{D}_s$ would work well in data reconstruction. Therefore, the parameters of $\boldsymbol{E}_s$ and $\boldsymbol{D}_s$ will stay unchanged during the back-propagation and when the joint loss reaches the minimum, $\boldsymbol{E}_p^t$ characterizes the individual differences of the current subject the best.

In the test phase, once a target series $\mathbf{x}_t$ is collected, we randomly choose data in the same length from each $\widetilde{X}_s^j$ as $\widetilde{X}_{rand}^{1\sim n}$, simultaneously. The performance of our model is ensured by two pipelines. One uses the trained shared classifier like most domain generalization methods do to guarantee the generalization ability as $\hat{y}_s^t$=$\boldsymbol{C}_s(\boldsymbol{E}_s(\boldsymbol{AT}(\mathbf{x}_t)))$. For the other pipeline, we calculate the cosine similarity w_s between $\boldsymbol{E}_p^{1\sim n}(\boldsymbol{AT}(X_{rand}^{1\sim n}))$ and $\boldsymbol{E}_p^t(\boldsymbol{AT}(\mathbf{x}_t))$ to make use of the private information. The higher weight indicates that the distribution is more similar with target data, thus, more trust can be given to the corresponding classifier. Then we get the prediction $\hat{y}_p^t$ by the dot product of the weight vector and the result vector of $\boldsymbol{C}_p^{1\sim n}$. The final result is determined after the integration of these two labels through a classifier fusion strategy following Lu et al. (2015).

Experiments

Dataset and Protocols

We verify the performance of our PPDA model on SEED (Zheng and Lu 2015), a public affective EEG dataset for emotion recognition. Fifteen rigorously screened Chinese movie clips are used to elicit the desired target emotion among happy, sad, and neutral. 15 subjects (8 females, mean: 23.27, std: 2.37) participated in the experiments three times on different days. During the experiment, subjects are encouraged to immerse themselves in the video to arouse corresponding emotions. The 62-channel EEG signals are recorded during the movie watching with the international 10-20 system using the ESI Neuroscan system. The preprocessed data are downsampled to 200 Hz and filtered with a bandpass of 0-75 Hz. Different entropy (DE) features are extracted within a non-overlapping one-second time window from 5 frequency bands (namely δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz) of every sample (Duan, Zhu, and Lu 2013). Therefore, in total, there are 3394 samples of 310 features per subject in one experiment, calculated by 62 channels multiplied by 5 bands.

Implementation Details

In order to compare with the state-of-the-art results, we align with their assessment details in this paper. Specifically, only one experiment from each subject is involved in the leave-one-subject-out cross strategy to study the inter-subject variability. In each iteration, we select one subject as the target new subject and the other 14 as the existing source subjects. It should be noted that although we do not use labeled emotion data in the calibration phase, all the 3394 sample points in SEED are labeled. Therefore, in the calibration phase, we take the first T second data as our calibration data after discarding the emotional tag.

The layer number, the hidden size, and the time step of the LSTM are fixed to 2, 64, and 15, respectively. Emotion classifiers and domain classifier are single-layer fully connected networks with hidden dimensions of 64. The calibration time T is set to 45 s. T For the trade-offs that control the synergy of the loss terms, the parameters are randomly sought, i.e. $\alpha \in \{k * 10^{-1} | k \in \{1, ..., 9\}$, $\beta \in$

Methods	#ATD	Avg.	Std.
Baseline SVM (Zheng and Lu 2016)	None	0.567	0.163
DICA (Ma et al. 2019)	None	0.694	0.078
DResNet (Ma et al. 2019)		0.853	0.080
TCA (Zheng and Lu 2016)	All	0.640	0.146
TPT (Zheng and Lu 2016)		0.752	0.128
DANN (Li et al. 2018)		0.792	0.131
DAN (Li et al. 2018)		0.838	0.086
WGANDA (Luo et al. 2018)		0.871	0.071
PPDA_NC	None	0.854	0.071
PPDA	Few	0.867	0.071

Table 2. Results of different methods running on SEED. #ATD is the abbreviation for the amount of target data used for model training.

$\{k * 10^{-4} | k \in \{1, ..., 5\}$, $\gamma \in \{k * 10^{-5} | k \in \{1, ..., 3\}$ and $\delta \in \{k * 10^{-2} | k \in \{1, ..., 3\}$. Adam optimizer is applied as the optimizing function, and the learning rate is selected in $\{2^k * 10^{-4} | k \in [-5, 5]\}$. The whole model is implemented by PyTorch.

Experiment Results

Comparison with state-of-the-art methods To validate the efficiency of our PPDA model, we compare the performance with that of the other seven popular methods on the cross-subject emotion recognition task on SEED. The mean accuracies (avg.) and standard deviations (std.) are reported in Table 2. At the same time, the methods are grouped according to the number of target data used in the model training. The traditional support vector machine (SVM) is taken as the baseline, in which 14 source data are regarded as training data to train one SVM and test the remaining target data. As shown in the above, the traditional SVM has poor recognition results due to the inter-subject variability. In all, PPDA gets a stable and decent result with about 86.7% accuracy with the standard deviation around 0.071. For DG methods that don't rely on any target data during model training, such as DICA and DResNet (Ma et al. 2019), our model improves the accuracy by 22.71% and 1.41%, respectively, which demonstrates that even using a small amount of target data will improve the recognition performance of the model. When compared with the DA methods, our model outperforms all of them except for WGANDA with a slight decrease (Zheng and Lu 2016; Luo et al. 2018). Although the recognition performance of our method is not the optimal one, it greatly shortens the calibration time while maintaining the recognition accuracy, which is of great practical significance. We also omit the calibration session of PPDA, noted as PPDA_NC, with other parts unchanged and run on the SEED experiment again to check its generalization ability. The reduced recognition performance demonstrates the importance of the calibration phase, which will be discussed in detail in the next section.

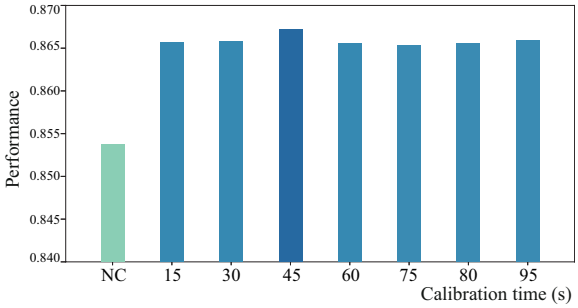


Figure 4: Performance of PPDA with different calibration data amounts. NC means without calibration.

Calibration data amount The purpose of employing the calibration phase is to enhance the model performance by tailoring the parameters with incoming data. However, it's hard to determine how many data is optimal since we hope to keep a balance between the buffer time it generates and the model accuracy. In order to study the effect of the calibration time on model performance and find the appropriate length of calibration data, we depict the accuracy change with respect to the increase of calibration data amount in Figure 4. As presented in Figure 4, once the calibration process is added, the performance of the model significantly improves as expected, which emphasizes the importance of the calibration. Generally speaking, the performance of the model will increase with the extension of calibration time. However, we do not see a significant growth as the calibration time becomes longer. The discrepancy is mainly attributed to the unique properties of EEG signals as we discussed at the beginning. Indeed, the EEG data are highly sensitive to external factors like the electrode impedance and head shapes, and internal variables such as the mental states. Despite the fact that it is easy to be affected, in one acquisition, the private components remain basically unchanged because those factors are relatively stable during a certain period. In other words, we can model the private components of the target subject well even with few calibration data and use it to improve the overall performance. Given that few data have

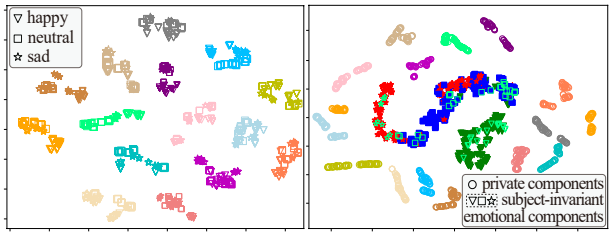


Figure 5: Feature visualization of PPDA. (a) The color represents different subjects, and the shapes represent the emotional categories. (b) The circle stand for the private components and dots in the middle represents the shared emotional components. The bright green in both pictures represents the target subject.

brought enough information we need, the performance of the model does not rise with the increase of calibration data.

Verification of LSTM-based encoder-decoder The encoder-decoder is supposed to separate the shared components that are constant and the private components various among subjects. To certify that the structure works properly, we randomly pick out 50 EEG samples from each subject to visualize them with t-SNE (Vazquez et al. 2013) via a scatter plot as displayed in Figure 5(a). Figure 5(b) exhibits the output of the shared encoder and private encoders. As can be seen, after PPDA codes, EEG representations reshape from a mess with personal characteristics to shared emotional components and private components. The solid dots with different shapes in Figure 5(b) represent the shared emotional components of the source subjects and the hollow circles are the private components specific to each subject. When new data come in, as the bright green shows, it is interesting that some dots perfectly fall into the places where the shared components gather while the others allocate randomly elsewhere. The same distribution of these two with that of source data indicates that the shared encoder successfully extracts the components that the most relevant to emotions and eliminates the inter-subject variability. Notice that the distance between any two private groups are not equivalent. It is certain that the output of $\boldsymbol{E}_p^t$ will be closer to some than others since they may have similar private components like identical gender, age, cultural background, etc. Therefore, we measure the distance and give them a weight to increase the reliability.

Attention mechanism The attention mechanism is adopted here to automatically assign a weight to each dimension of the EEG feature. To intuitively study the relevance of channels and bands to emotions, we visualize the weights as shown in Figure 6. In Figure 6(a), it is obvious that the Beta and the Gamma bands are much more activated than other three bands. Moreover, the distribution of the color depth even exposes the feasibility of using it to recognize emotion. For example, in Beta band, the blue goes deeper when happy emotion is aroused. So does the Gamma band in both neutral and sad feelings. In Figure 6(b), we plot the topographical EEG maps reflecting the distributions of crucial channels. The darker the brain region, the more important the channels in this area. Lateral temporal lobes where FT7, FT8, T7, and T8 channels lie are triggered intensively. The consistency between our findings and the existing observations on critical bands and channels for EEG-based emotion recognition Zheng and Lu (2015) confirms that the attention mechanism has the ability to catch the emotion-sensitive properties, which is more reliable and interpretable than previous ways of assigning weights, particularly subjective manual labeling. Besides, the centralization of decisive components offers a possibility to develop compact EEG devices to make real-world applications practicable. Meanwhile, the reduction of channels will also significantly reduce model's calculation time, which is crucial to real-time applications.

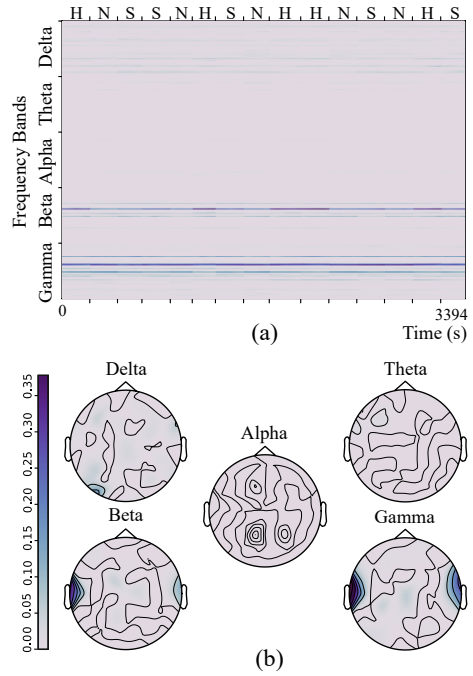


Figure 6: Visualization of attention weight. (a) The distribution of weights over time in one experiment. The vertical axis shows the arrangement of bands and channels in 310-dimension EEG feature. The capital letters at the top represent the ground truth label at the corresponding time (H: happy, N: neutral, S: sad). (b) The weight distribution of different brain regions in five frequency bands.

Conclusion

In this paper, we devise a methodology called plug-and-play domain adaptation for cross-subject EEG-based emotion recognition, aiming to allow everyone immediately use it without waiting while maintaining the recognition accuracy. It has managed to shorten the calibration time within a minute with the accuracy over 86.7%, a comparable result to the state-of-the-art domain adaptation performance. This technique can be used to enhance user experience and make EEG-based affective computing applications more practicable. Moreover, the critical frequency channels and bands discovered through the attention mechanism sheds lights on the development of wearable EEG devices and real-time emotion recognition. Our future work will concentrate on the real-time test under various actual environments to see its practicability since the current results are still based on simulations of real applications on offline datasets.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (Grant 2017YFB1002501), the National Natural Science Foundation of China (Grant No. 61673266 and No. 61976135), SJTU Trans-med Awards Research (WF540162605), the Fundamental Research Funds for the Central Universities, and the 111 Project.

References

Ahern, G. L.; and Schwartz, G. E. 1985. Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis. *Neuropsychologia* 23(6): 745–755.

Alarcao, S. M.; and Fonseca, M. J. 2017. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing* 10(3): 374–393.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. W. 2010. A theory of learning from different domains. *Machine Learning* 79(1-2): 151–175.

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, 2178–2186.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, 343–351.

Brunner, C.; Birbaumer, N.; Blankertz, B.; Guger, C.; Kübler, A.; Mattia, D.; Millán, J. d. R.; Miralles, F.; Nijholt, A.; Opisso, E.; et al. 2015. BNCI Horizon 2020: towards a roadmap for the BCI community. *Brain-computer Interfaces* 2(1): 1–10.

Duan, R.-N.; Zhu, J.-Y.; and Lu, B.-L. 2013. Differential entropy feature for EEG-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 81–84. IEEE.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 17(1): 2096–2030.

Ghifary, M.; Balduzzi, D.; Kleijn, W. B.; and Zhang, M. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(7): 1414–1430.

Kim, B. H.; and Jo, S. 2020. Deep Physiological Affect Network for the Recognition of Human Emotions. *IEEE Transactions on Affective Computing* 11(2): 230–243.

Li, H.; Jin, Y.-M.; Zheng, W.-L.; and Lu, B.-L. 2018. Cross-subject emotion recognition using deep adaptation networks. In *International Conference on Neural Information Processing*, 403–413. Springer.

Li, J.; Qiu, S.; Shen, Y.-Y.; Liu, C.-L.; and He, H. 2019. Multisource transfer learning for cross-subject EEG emotion recognition. *IEEE Transactions on Cybernetics* .

Lotte, F.; Congedo, M.; Lécuyer, A.; Lamarche, F.; and Arnaldi, B. 2007. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering* 4(2): R1.

Lu, Y.; Zheng, W.-L.; Li, B.; and Lu, B.-L. 2015. Combining eye movements and eeg to enhance emotion recognition. In *IJCAI*, volume 15, 1170–1176. Citeseer.

Luo, Y.; Zhang, S.-Y.; Zheng, W.-L.; and Lu, B.-L. 2018. WGAN domain adaptation for EEG-based emotion recognition. In *International Conference on Neural Information Processing*, 275–286. Springer.

Ma, B.-Q.; Li, H.; Zheng, W.-L.; and Lu, B.-L. 2019. Reducing the subject variability of EEG signals with adversarial domain generalization. In *International Conference on Neural Information Processing*, 30–42. Springer.

Mnih, V.; Heess, N.; Graves, A.; et al. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, 2204–2212.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, 10–18.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10): 1345–1359.

Samek, W.; Meinecke, F. C.; and Müller, K.-R. 2013. Transferring subspaces between subjects in brain–computer interfacing. *IEEE Transactions on Biomedical Engineering* 60(8): 2289–2298.

Sangineto, E.; Zen, G.; Ricci, E.; and Sebe, N. 2014. 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*, 357–366.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Vazquez, D.; Lopez, A. M.; Marin, J.; Ponsa, D.; and Geronimo, D. 2013. Virtual and real world adaptation for pedestrian detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(4): 797–809.

Zheng, W.-L.; and Lu, B.-L. 2015. 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.

Zheng, W.-L.; and Lu, B.-L. 2016. Personalizing EEG-based affective models with transfer learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2732–2738.

Zheng, W.-L.; Zhu, J.-Y.; and Lu, B.-L. 2019. Identifying Stable Patterns over Time for Emotion Recognition from EEG. *IEEE Transactions on Affective Computing* 10(3): 417–429.