Variational Pathway Reasoning for EEG Emotion Recognition

Tong Zhang,[1] Zhen Cui,[1]* Chunyan Xu,[1] Wenming Zheng,[2] Jian Yang[1]
[1]Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
Jiangsu Key Lab of Image and Video Understanding for Social Security,
School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
[2]Research Center for Learning Science, School of Biological Science and Medical Engineering,
Southeast University, Nanjing, China.
{tong.zhang, zhen.cui, cyx}@njust.edu.cn, wenming_zheng@seu.edu.cn, csjyang@njust.edu.cn

Abstract

Research on human emotion cognition revealed that connections and pathways exist between spatially-adjacent and functional-related areas during emotion expression (Adolphs 2002a; Bullmore and Sporns 2009). Deeply inspired by this mechanism, we propose a heuristic Variational Pathway Reasoning (VPR) method to deal with EEG-based emotion recognition. We introduce random walk to generate a large number of candidate pathways along electrodes. To encode each pathway, the dynamic sequence model is further used to learn between-electrode dependencies. The encoded pathways around each electrode are aggregated to produce a pseudo maximum-energy pathway, which consists of the most important pair-wise connections. To find those most salient connections, we propose a sparse variational scaling (SVS) module to learn scaling factors of pseudo pathways by using the Bayesian probabilistic process and sparsity constraint, where the former endows good generalization ability while the latter favors adaptive pathway selection. Finally, the salient pathways from those candidates are jointly decided by the pseudo pathways and scaling factors. Extensive experiments on EEG emotion recognition demonstrate that the proposed VPR is superior to those state-of-the-art methods, and could find some interesting pathways w.r.t. different emotions.

Introduction

Emotion recognition has become an active topic of affective computing in recent years and drawn wide attention due to its huge potential applications including humanoid robots, driver monitoring, etc. Externally, human emotion is usually expressed as non-physiological signals including facial expressions, body actions, speeches, etc. In contrast, bioelectrical signals, e.g., electroencephalograph (EEG) and galvanic skin response (GSR), internally reflect the intrinsic emotion states. Specifically, EEG signals of electrodes attached on scalps are rather reliable to capture brain-emotion variations with high temporal resolution. Therefore, EEG signals have become increasingly important in analysing human emotion.

Numerous algorithms (Zheng et al. 2014; Shi, Jiao, and Lu 2013; Li et al. 2016; Zheng and Lu 2015; Li et al. 2018a; 2018b; Song et al. 2018) have been proposed to tackle with EEG emotion recognition. In early time, these methods (Zheng et al. 2014; Shi, Jiao, and Lu 2013) mostly focus on the sophisticated signal processing techniques of frequency band filtering. To boost recognition performance, recently, discriminative representation learning techniques are used to extract more effective emotion features. For instances, Zheng (Zheng 2016) formulated EEG signals as sparse channel selection to suppress those electrodes with negative effects to emotion analysis. More recently, deep networks, such as deep belif network (DBN) (Zheng and Lu 2015) and recurrent neural network (RNN) (Zhang et al. 2018; Song et al. 2019), are adopted to encode high-level features. Furthermore, in view of the non-gridded layout of EEG signals, graph convolutional neural network (GCNN) (Song et al. 2018) was introduced to model EEG signals, and achieved the state-of-the-art performance. All these methods are mostly inspired from the algorithms of machine learning, but emotion mechanisms or principles are seldom introduced in EEG emotion task.

Psychology study (Adolphs 2002a; Bullmore and Sporns 2009) on human emotion perception revealed that connections and pathways exist between spatially-adjacent and functional-related areas during emotion expression. This discovery is great interesting to encourage us to design some new-type algorithms that use pathways to boost performance. Conversely, those pathways are expected to be detected for emotion variation explanation. To this end, we need to solve two critical problems: i) how to model this perception mechanism of existing pathways into emotion recognition algorithms; ii) how to detect those salient interesting pathways w.r.t. different emotions.

To address the two problems above, here we propose a heuristic Variational Pathway Reasoning (VPR) method for EEG emotion recognition. We specially introduce random walk to generate a large number of candidate pathways along electrodes. Considering spatial adjacency and functional region, we assign possible connections to spatially adjacent electrodes and constrain pathways in local functional regions, which may largely decrease the explo-

*Corresponding author

sive magnitude of pathways. To represent each pathway, dynamic sequence models, e.g., long-short term memory (LSTM), could be used to encode their ordered connectivity that indicates between-electrode dependency. The encoded pathways around each electrode are aggregated to a pseudo maximum-energy pathway. As this pseudo pathway indicates the most salient pair-wise connections in a local region anchored around the center electrode, so the contribution strength of each candidate to this pseudo pathway determines its some saliency degree in emotion analysis. Moreover, the saliency of candidate pathways depends on the importance of pseudo pathways in the holistic brain region. To reason those salient pathways, mathematically, we formulate it as a Bayesian probabilistic learning process and propose sparse variational scaling (SVS) to learn scaling factors of pathways, which is different from previous deterministic methods suffering from instability due to the low generalization ability. Considering the high distribution variation among different subjects, we design an adaptive selection by constraining the sparse structure of those scaling factors. As a result, our proposed VPR possesses two advantages. First, the stochastic variational scaling for pathway selection endows VPR with good generalization ability, and makes the model more interpretable. Second, the sparsity on scaling factors well favors the adaptive pathway selection. We evaluate our VPR on two EEG emotion datasets. Extensive experiments demonstrate that our proposed VPR is superior to those state-of-the-art methods, and could find some interesting pathways w.r.t. different emotions.

In summary, our main contributions are three folds:

(i) The pathway mechanism is first introduced to the field of EEG emotion task, and framed into a well-constructed model to boost recognition performance.

(ii) We proposed a salient pathway reasoning method, which includes two basic modules named pathway aggregation and sparse variational scaling. It can adaptively determine salient pathways to facilitate EEG emotion recognition, and meanwhile provide some explanation for emotion analysis.

(iii) The proposed VPR achieves state-of-the-art performance on two public EEG emotion datasets, and meanwhile finds some interesting observation of pathways about different emotions.

Related work

Below we first briefly overview EEG emotion recognition methods, then we introduce literatures about random walk based graph embedding and variational auto-encoder (VAE), which are technically related to our work.

EEG emotion recognition methods. Various methods have been proposed to deal with EEG-based emotion recognition. These methods generally contain two crucial steps, feature extraction and emotion classification. In order to extract robust features from raw EEG signals with low signal-to-noise ratio, various EEG descriptors were proposed (Zheng et al. 2014; Shi, Jiao, and Lu 2013; Zheng and Lu 2015) such as power spectral density (PSD), differential entropy (DE) and differential asymmetry (DASM). According to these descriptors, the existing methods built either global or local models through feature extraction algorithms as well as classic classifiers. For instances, Zheng et al. (Zheng and Lu 2015) employed support vector machine (SVM) on features of either all electrodes (aka channels) or a part of selected ones, while group sparse canonical correlation analysis (GSCCA) (Zheng 2016) attempted to automatically select salient electrodes by using group sparse constrain. In recent years, accompanying with great successes of deep learning in various computer vision tasks, researchers also attempted to employ deep networks on EEG signals in order to extract more robust features. In the relative early stage, deep belief networks (DBN) (Zheng and Lu 2015) directly gathered all electrodes together as the input and constructed multiple hidden layers to learn high-level features.

Random walk based graph embedding. Various graph embedding techniques based on random walk have been proposed to obtain node representations, where node2vec (Grover and Leskovec 2016) and DeepWalk are two representative works. Node2vec maximizes the probability of occurrence of subsequent nodes in random paths of fixed length to preserve higher order proximity between nodes. DeepWalk (Perozzi, Al-Rfou, and Skiena 2014) learns latent representations by using local information obtained from truncated random walks, where a node and a walk path are treated as a word and a sentence in Word2Vector (Mikolov et al. 2013) respectively. Motivated by the success of DeepWalk, many subsequent studies which apply deep learning models on the sampled paths, e.g. Skip-Gram (Mikolov et al. 2013) or Long-Short Term Memory (LSTM) (Gers, Schmidhuber, and Cummins 1999), are proposed on the sampled paths for graph embedding.

VAE variants. VAE models the underlying probability distribution of observations using variational inference in a probabilistic way, which makes itself more stable and interpretable than deterministic methods. In recent years, it has drawn much attention in the field of artificial intelligence (Pu et al. 2017; Higgins et al. 2017; Li et al. 2017; Walker et al. 2016) and multiple variants, e.g. β-VAE (Higgins et al. 2017), variational graph auto-encoders (VGAE) (Kipf and Welling 2016) and Stein VAE (Pu et al. 2017), have been developed to adapt different tasks. Now VAE is flourishing in many pattern recognition tasks including object recognition (Zhao et al. 2019), recommendation system (Karamanolakis et al. 2018), and document summarization (Li et al. 2017).

Variational Pathway Reasoning

In this section, we first give an overview on the proposed VPR, and then introduce main modules and details.

Overview The entire architecture of VPR is shown in Fig. 1. The input is EEG signal produced from brain regions. In view of spherical-shape structure, EEG signals are modeled as graphs, where each electrode is regarded as one node. According to emotion perception mechanism (Adolphs 2002a; Bullmore and Sporns 2009) that there are connections and pathways during emotion activation, we introduce random walk to generate a mass of pathway

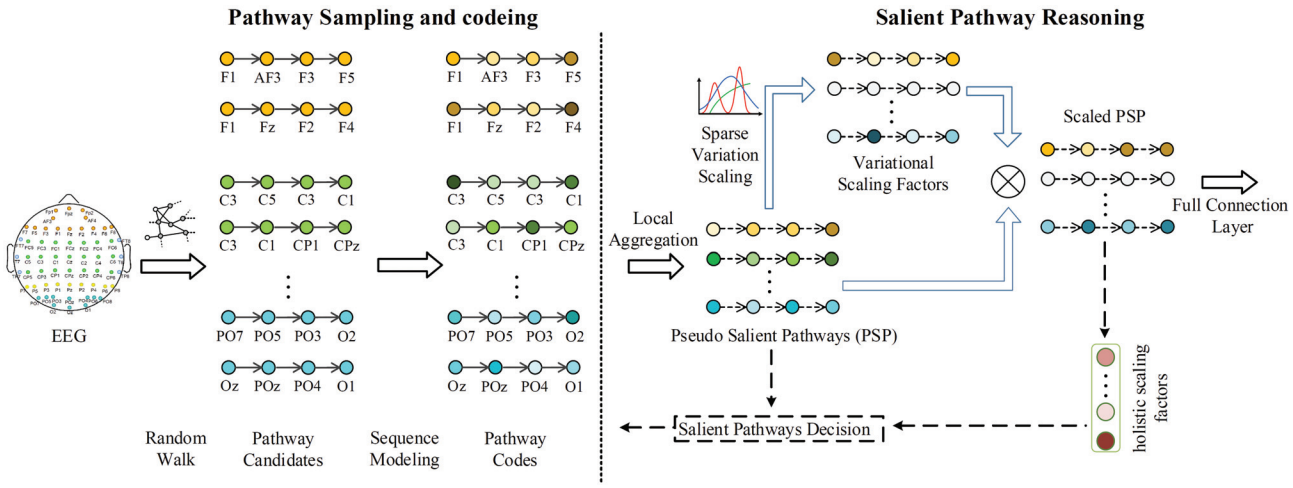

Figure 1: The proposed VPR framework. The introduction is given in the subsection "Overview". Specifically, the sparse variational factors endow the structure of sparsity to the scaled PSP which favors the adaptive pathway selection. The dotted lines mean that the PSP and holistic scaling factors are employed to reversely calculate the contribution of each candidate pathway.

candidates. To reduce the exponential magnitude of walk paths, we constrain walk scopes within local regions through the partition of five brain function regions ('Frontal' (F), 'Temporal' (T), 'Parietal' (P), 'Occipital' (O) and 'Central' (C)) (Adolphs 2002a), and define direct connections only between spatially-adjacent electrodes. For each electrode as a start node, we may sample multiple candidate pathways within a predefined walk length. Considering the sequence property of walk paths, we introduce the dynamic sequence model LSTM to encode between-electrode dependencies and further extract high-level features of pathways. For each electrode, we maximumly aggregate the encoded pathways therein to derive a pseudo salient pathway, each edge of which actually denotes the salient connection at a hopping step. To further purify pseudo pathways, we propose a sparse variational scaling (SVS) module to learn scaling factors imposed on those pathways. In contrast to the local aggregation, the variational scaling is shared on all electrodes, thus may be understood as a holistical weighting strategy. The scaling factors could be not only used for generating more discriminative features for final emotion prediction, but also fed-back into those candidate pathways to decide the final salient pathways.

Pathway Candidates Generation For the input EEG signals, multiple pathways are sampled for every electrode (node) in each local brain region. Formally, for a random walk of a certain length l, the ordered nodes denoted as $[\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_l]$ are generated by the following distribution:

$$P(\mathbf{u}_i = \mathbf{v}_2 | \mathbf{u}_{i-1} = \mathbf{v}_1) = \begin{cases} \frac{e_{(\mathbf{v}_1, \mathbf{v}_2)}}{Z}, & \text{if } (\mathbf{v}_1, \mathbf{v}_2) \in \mathcal{E}_r; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

where $e_{(\mathbf{v}_1, \mathbf{v}_2)}$ denotes the unnormalized adjacency between nodes $\mathbf{v}_1$ and $\mathbf{v}_2$ with its value being set to 1 if $\mathbf{v}_1, \mathbf{v}_2$ are spatially adjacent (more details in Fig. 2) or 0 otherwise, Z is the normalizing constant, and $\mathcal{E}_r$ represents the set of pairs of adjacent electrodes enclosed in the r-th brain region ($r \in \{F, T, P, O, C\}$). Note that, in this paper, the symbols $\mathbf{u}_i$ and $\mathbf{v}_i$ are abusively used as nodes or node signals for the simplification description. They could be easily understood according to the context statement.

Pathway Coding For the candidate pathways, the sequence recursive model LSTM (Gers, Schmidhuber, and Cummins 1999) is employed to encode the dependencies on walking nodes and obtain their embedding representation. Formally, for a given pathway $[\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_l]$, the encoding process is defined as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ui}\mathbf{u}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i), \tag{2}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{uf}\mathbf{u}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f), \tag{3}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot tanh(\mathbf{W}_{uc}\mathbf{u}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c), \tag{4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{uo}\mathbf{u}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_o), \tag{5}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{c}_t). \tag{6}$$

where $\mathbf{i}_t, \mathbf{f}_t$ denote the input and forget gates respectively, $\mathbf{c}_t$ is the memory cell, $\mathbf{h}_t$ is the hidden state, $\mathbf{o}_t$ is the output response, and t means the position of the node in the pathway. The operators $\odot$ and $\sigma(\cdot)$ denote the element-wise product and the non-linear activation function (we use sigmoid here). $\mathbf{W}_{\alpha\beta}(\alpha \in \{u, h, c\}, \beta \in \{i, f, c, o\})$ is the transformation matrix allowing LSTM to utilize both the current and past information to determine the output, and $\mathbf{b}_\beta$ is the bias. Among them above, $\mathbf{W}_{\alpha\beta}$ and $\mathbf{b}_\beta$ are the parameters to be optimized. Then, the generated output responses, denoted as $\mathbf{q} = [\mathbf{o}_1^t, \cdots, \mathbf{o}_l^T]^T$, are used as the embedding representation of the given pathway.

Local Pathways Aggregation To find those salient connections between eletrodes, we use max-aggregation to

those local candidate pathways with the same starting nodes. Concretely, for the j-th electrode, the k-th associated pathway is embedded as $\mathbf{q}_{j_k}(k \in \mathcal{F}_j, k = 1, \cdots, K)$, where $\mathcal{F}_j$ means the pathway set starting from the j-th electrode. The aggregated pathway can be calculated by taking the maximum value of each dimension among the embedded codes:

$$\mathbf{s}_j = \max_{k \in \mathcal{F}_j} \mathbf{q}_{j_k}. \tag{7}$$

Note that one aggregated pathway might not be a true walk because of inconsistency of max positions, but this pathway recorded those most salient connections between electrodes. Thus, we refer the aggregated pathways as pseudo salient pathways. After max-aggregation, all pseudo salient pathways w.r.t. electrodes are collected as the d-dimensional vector set $\mathcal{S} = \{\mathbf{s}_1, \cdots, \mathbf{s}_n\}$ where $\mathbf{s}_j \in \mathbb{R}^d, j \in [1, \cdots, n]$.

One candidate pathway highly correlates to the corresponding pseudo salient pathway therein. Formally, the local saliency of one candidate pathway is derived as

$$a_{k \to j} = < \mathbf{q}_{jk}, \mathbf{s}_j > / \sqrt{(\mathbf{q}_{jk}^T \mathbf{q}_{jk})} \sqrt{(\mathbf{s}_j^T \mathbf{s}_j)}, \quad k \in \mathcal{F}_j, \tag{8}$$

where $< \cdot, \cdot >$ denotes the inner product of two vectors, $a_{k \to j}$ is the importance degree of the k-th pathway to the local region around the j-th electrode.

Holistic Pathway Scaling The above aggregation process is used to calculate the local saliency of candidate pathways as shown in Eqn. 8. More holistically, we propose the SVS module to derive scaling factors of pseudo pathways. Instead of a deterministic way, we introduce the variational inference model with sparse constraints. Due to good mathematical property on data distribution fitting, SVS tends to be stable and interpretable. Moreover, the structure of those scaling factors is constrained to be sparse so as to adaptively select the salient pathways.

Given the pathway saliency set $\mathcal{S} = \{\mathbf{s}_1, \cdots, \mathbf{s}_n\} (\mathbf{s}_j \in \mathbb{R}^d)$, we aim to derive the latent variable set $\mathcal{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_n\} (\mathbf{z}_j \in \mathbb{R}^d)$ based on the posterior distribution $p(\mathcal{Z}|\mathcal{S})$, where $\mathbf{z}_j$ is the scaling factor to be learnt. Each pseudo pathway saliency $\mathbf{s}_j$ will be re-scaled based on the corresponding latent variable $\mathbf{z}_j$:

$$\mathbf{s}'_j = \mathbf{s}_j \odot \mathbf{z}'_j = \mathbf{s}_j \odot \sigma(\mathbf{z}_j), \tag{9}$$

where $\mathbf{s}'_j$ represents the final saliency of the pseudo pathway after scaling, $\mathbf{z}'_j = \sigma(\mathbf{z}_j)$ is the scaling vector, $\sigma(\cdot)$ is a non-linear transformation function on $\mathbf{z}_j$ and is set as the sigmoid function here, and $\odot$ means element-wise product.

Let $\mathcal{X}^{(i)} = \{\mathbf{x}_1^{(i)}, \cdots, \mathbf{x}_n^{(i)}\}$ denote the EEG signals of the i-th sample with the corresponding emotion label y_i, where $\mathbf{x}_j^{(i)}$ is the signal on the j-th electrode, we introduce how to learn the corresponding latent variable set $\mathcal{Z}^{(i)}$ based on $\mathcal{S}^{(i)}$. For notation simplicity, $\mathbf{s}_j^{(i)}, \mathbf{z}_j^{(i)}$ in $\mathcal{S}^{(i)}, \mathcal{Z}^{(i)}$ are directly denoted as $\mathbf{s}_j, \mathbf{z}_j$ below. For a given $\mathbf{s}_j$, the true posterior distribution of $\mathbf{z}_j$ is difficult to infer by the Bayes rule $p(\mathbf{z}_j|\mathbf{s}_j, y_i) = p(\mathbf{z}_j)p(\mathbf{s}_j, y_i|\mathbf{z}_j)/p(\mathbf{s}_j, y_i)$ due to the intractability. In variational inference, an approximation to the intractable true posterior $p(\mathbf{z}_j|\mathbf{s}_j, y_i)$ is introduced as $q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i)$ to solve the problem above. Also, the Kullback-Leibler (KL)-divergence is employed to estimate the distribution distance between $q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i)$ and $p(\mathbf{z}_j|\mathbf{s}_j, y_i)$. Then, minimizing KL-divergence may result into a good approximation of $p(\mathbf{z}_j|\mathbf{s}_j, y_i)$ by using $q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i)$. Considering the non-negative property of KL-divergence and supervised information of y_i, the variational lower bound (VLB) $L^{vlb}(\mathbf{s}_j, y_i, \phi)$ may be derived, thus minimizing KL-divergence is equivalent to maximizing $L^{vlb}(\mathbf{s}_j, y_i, \phi)$ as follows:

$$\begin{aligned} L^{vlb}(\mathbf{s}_j, y_i, \phi) &= \log p(y_i|\mathbf{s}_j) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i)||p(\mathbf{z}_j|\mathbf{s}_j, y_i)) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i)}[\log p(y_i|\mathbf{s}_j, \mathbf{z}_j)] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i)||p(\mathbf{z}_j|\mathbf{s}_j)) \end{aligned} \tag{10}$$

where $p(\mathbf{z}_j|\mathbf{s}_j) \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$. Here different from the previous VAE method (Walker et al. 2016), we derive the scaling factors only depending on the data distribution, which means $q_\phi(\mathbf{z}_j|\mathbf{s}_j, y_i) = q_\phi(\mathbf{z}_j|\mathbf{s}_j)$. To make the variational lower bound above be optimized through stochastic gradient strategy, the reparameterization trick (Kingma and Welling 2013) is introduced into the Gaussian case: $\mathbf{z}_j = f(\phi, \epsilon_j) = \mathbf{u}_j + \boldsymbol{\epsilon}_j \cdot \boldsymbol{\sigma}_j, \boldsymbol{\epsilon}_j \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{u}_j$ and $\boldsymbol{\sigma}_j$ are the mean and standard deviation (s.d.) vectors.

To find those most important connections between different electrodes, we impose the sparsity constraint on the scaling factors. Formally, given the factor vector set $\mathcal{Z}' = \{\mathbf{z}'_1, \cdots, \mathbf{z}'_n\}$, we define the sparsity loss as follows:

$$\mathcal{L}^{sparse}(\mathcal{Z}') = \sum_{j=1}^{n} \sum_{m=1}^{d} (z'_{jm})^2, \tag{11}$$

where z'_{jm} denotes the m-th element of $\mathbf{z}'_j$. Minimizing $\mathcal{L}^{sparse}$ may lead to a group sparse structure on $\mathcal{Z}$. In other words, it means that only a part of pathways will be retained while others are suppressed.

The Loss Function For the input $\mathcal{X}^{(i)}$, the pathway saliency set $\mathcal{S}^{(i)}$ is firstly learnt. Then, after holistic pathway scaling, the obtained final saliency vectors are concatenated spatially and fed into a fully connected layer, based on which the term denoted as $\log p(y_i|\mathbf{s}_j, \mathbf{z}_j)$ of $\mathcal{L}^{vlb}(\mathcal{X}^{(i)}, y_i, \phi)$ in Eqn. 10 can be calculated.

In training the whole VPR , we minimize the whole loss function defined as follows

$$\mathcal{L}(\mathcal{X}^{(i)}, y_i, \phi) = -\mathcal{L}^{vlb}(\mathcal{X}^{(i)}, y_i, \phi) + \lambda * \mathcal{L}^{sparse}(\mathcal{Z}'^{(i)}), \tag{12}$$

where λ is the trade-off parameter to balance the VLB term and the sparse term.

Decision of Salient Pathways To determine the salient pathways, we combine both the contribution strength of each candidate to its corresponding pseudo pathway ($a_{k \to j}$ in Eqn. 8) and the importance of pseudo pathway ($\mathbf{z}'_j = \sigma(\mathbf{z}_j)$ in Eqn. 9). Formally, for the k-th candidate pathway around the j-th anchor, the overall contribution denoted as c_{jk} to the final emotion analysis is defined as

$$c_{jk} = a_{k \to j} * ||\mathbf{z}'_j||_2. \tag{13}$$

Until now, we can select those salient pathways as an explanation of emotion activation as shown in the following experiments.

Experiments

We evaluate the VPR by conducting experiments on two EEG emotion datasets named SJTU Emotion EEG Dataset (SEED) and multi-modal physiological emotion database (MPED). To achieve a comprehensive evaluation, our proposed VPR is compared with multiple state-of-the-art methods by following the wide used subject dependent protocols on these two datasets. In the following parts, we first describe the experiment setup including the protocols and implemental details, then we compare the experimental results of our VPR with those state-of-the-art methods, finally we analyse our VPR model by conducting additional ablation experiments as well as visualizing the salient pathways.

Experiment Setup

Dataset and protocols Both SEED (Zheng and Lu 2015) and MPED (Song et al. 2019) are collected by using an ESI NeuroScan System at a sampling rate of 1000 Hz from 62-channel electrode cap according to the International 10-20 system. In total, there are fifteen subjects participating in the experiment of SEED including 7 males and 8 females, and 23 subjects participating in MPED. During the experiment, emotional film clips are shown to the participants to elicit multiple kinds of target emotion. After collection, SEED contains three emotion categories (positive (POS), negative (NEG) and neutral (NEU) while MPED contains seven emotion categories in total.

For comprehensive performance evaluation, we conduct widely employed subject-dependent experiments on both SEED and MPED datasets. For SEED, each subject conducts two times of experiments which yields totally 30 times of experiments. Following the cross session protocol in (Zheng and Lu 2015), the training and testing samples are respectively taken from different sessions of one experiment. As each time of experiment contains fifteen sessions, nine of them are used for training and the remaining six for testing. Metrics including accuracy and standard deviation are employed to evaluate the performance. Similarly for the "Protocol two" of MPED (Song et al. 2019), each subject conducts 28 trials where 7 trials are used as testing data while the rests are used as training data. Moreover, the original seven emotion categories are divided into three ones , i.e. positive, neutral and negative, as SEED. This process causes large class imbalance, which also makes the emotion recognition task rather challenging.

Preprocessing. For EEG emotion recognition, the preprocessing step is rather necessary because raw EEG signals are always with low signal-to-noise ratio. For EEG signals of both SEED and MPED, the component in five frequency bands (delta: 1-3 Hz, theta: 4-7 Hz, alpha: 8-13 Hz, beta: 14-30 Hz, gamma: 31-50 Hz) of 62 channels is first filtered and then used for feature extraction. For SEED, DE feature is extracted by applying a 256-point short-time Fourier transform with a nonoverlapped Hanning window of 1s. While for MPED, the feature named short-time Fourier transform (STFT) feature is extracted.

Regional Connection. For a given EEG signal, considering the regional functional connectivity revealed in (Adolphs 2002a), we first separate the electrodes into five regions ('F', 'T', 'P', 'O' and 'C') according to the International 10-20 system (Scharbrough et al. 1990). Then for each brain region, according to (Bullmore and Sporns 2009), spatially close electrodes are assigned with the adjacency relationship, while nodes from different regions are not connected. The detail of the regional connection is shown in Fig. 2.

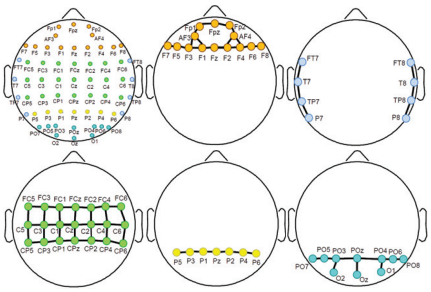

Figure 2: Exhibition of regional connection in EEG signals.

Table 1: The comparisons on SEED following cross session protocol.

Feature	Method	training / testing	accuracy / (%)
DE	SVM (Zheng and Lu 2015)	9/6	84.0 / 9.7
DE	SVM* (Zheng and Lu 2015)	9/6	86.7 / 8.3
DE	CCA (Hardoon and Szedmak 2004)	9/6	77.6/13.2
DE	GSCCA (Zheng 2016)	9/6	83.0 / 10.0
DE	DBN (Zheng and Lu 2015)	9/6	86.1 / 8.3
DE	STRNN (Zhang et al. 2018)	9/6	89.5 / 7.6
DE	GCNN (Song et al. 2018)	9/6	87.4 / 9.2
DE	DGCNN (Song et al. 2018)	9/6	90.4 /8.5
DE	BiDANN (Li et al. 2018b)	9/6	92.4 /7.0
DE	R2G-STNN (Li et al. 2019)	9/6	93.4/6.0
DE	VPR	9/6	**94.3**/6.5

Implemental details. For both SEED and MPED, the architectures of the proposed VPR framework are set the same, which are determined by cross-validation on selected validation set (i.e., a part of training set). For random walk, four pathways are generated for each starting node (yielding 268 pathways in total) with the path length of 4, where every electrode is treated as the starting node in the process of random walk. Besides, the dimension of hidden state of LSTM for pathway embedding is traversed in the range of [8, 16, 32, 64], and finally set to 16. The value of λ in Eqn.12 is set to 0.5. In the training stage, we run the our VPR model for 20 epochs with a learning rate of 0.001 for tuning the network parameters, where the batch size is set to 64.

Experiment on SEED.

The performance of our VPR is shown in Table 1, and it is also compared with various existed algorithms including SVM (Zheng and Lu 2015), canonical correlation analysis (CCA) (Hardoon and Szedmak 2004), GCNN (Song et al. 2018), etc. All these methods follow the same cross session protocol with our VPR. In general, our VPR achieves the highest performance with the accuracy of 94.1% comparing with all previous state-of-the-art methods with a rela-

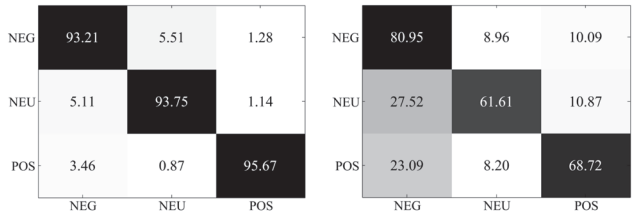

Figure 3: The confusion matrices of VPR on SEED (left) and MPED (right).

Table 2: The comparisons on MPED following "Protocol two" in (Song et al. 2019).

Feature	Method	training / testing	acc /F1
STFT	SVM (Zheng and Lu 2015)	21/7	57.06/24.43
STFT	KNN (Hochreiter 1997)	21/7	43.96/34.39
STFT	DBN (Duda, Hart, and Stork 2012)	21/7	65.98/59.19
STFT	STRNN (Zhang et al. 2018)	21/7	66.84/60.57
STFT	DGCNN (Song et al. 2018)	21/7	68.02/61.11
STFT	LSTM (Song et al. 2019)	21/7	71.92/65.12
STFT	A-LSTM (Song et al. 2019)	21/7	71.57/67.74
STFT	VPR	21/7	**75.06/68.64**

tively low standard deviation, which verifies the superiority of our VPR. Relatively high performances are also achieved by deep learning methods including R2G-STNN (Li et al. 2019), BiDANN (Li et al. 2018b) and GCNN (Song et al. 2018) with the accuracies of 93.4%, 92.4%, and 90.4% respectively. It should be specifically noted that although R2G-STNN and BiDANN seem to obtain high performance which is only a little lower than VPR, however, they belong to domain adaptive methods which involve test data for network optimization. And very different from them, our VPR does not use any information from test data during training, while still achieves better performance. This verifies the superiority and high generalization ability of our VPR.

The confusion matrix of all evaluated experiments of SEED is shown in Fig. 3(a), where the element located in the i-th row and jth column shows the percentage of those samples belonging to the class i while predicted as the class j. Diagonal elements mean the accuracies of each class and others mean the confusion. As it is shown, our algorithm performs well in recognizing all three types of emotions as the accuracies of them are more than 93.0%. In this dataset, positive emotion are more easier to be recognized, while relative high confusion appears between negative and neutral emotions.

Experiment on MPED. Table 2 shows the comparison between our VPR and other existing methods on MPED. In general, our VPR outperforms all those compared methods, including KNN (Hochreiter 1997), STRNN (Zhang et al. 2018), DGCNN (Song et al. 2018), LSTM (Song et al. 2019) and A-LSTM (Song et al. 2019). Among the comparison methods, LSTM based methods, i.e. LSTM and A-LSTM, achieve better performance with the accuracies of 71.92% and 71.57%, and our VPR outperforms them with the accuracy which is about 3 percent higher. Moreover, due to the large class imbalance, F1 score is also employed to evaluate these methods. According to Table 2, the F1 score of our VPR, which is 66.8%, is also the highest among all the compared methods, which indicates the robustness of our VPR to the variation caused by class imbalance.

Fig. 3(b) shows the confusion matrix of MPED, which exhibits a large difference comparing to that of SEED. Highest accuracy is obtained on negative emotion while high confusion appears between both pairs of neutral versus positive and neutral versus positive. This large difference may attribute to the heavy class imbalance, which may cause bias to the network during training.

Table 3: The experimental results of PSN, non-sparse VPR, and VPR on MPED and SEED datasets.

Dataset	Method	Accuracy
SEED	PSN	91.8
	non-sparse VPR	93.5
	VPR	**94.3**
MPED	PSN	72.17
	non-sparse VPR	73.95
	VPR	**75.06**

Ablation study

As promising performance has been achieved by our VPR, we want to know how the modules, e.g. SVS and sparse constraints, promote the emotion recognition. For this purpose, we conduct two additional experiments to dissect our framework based on SEED and MPED datasets as follows:

(1) Comparing the performance between VPR and the pathway saliency network (PSN). This aims to evaluate the benefit of our designed SVS module, where PSN can be easily obtained by removing the SVS module from our VPR.

(2) Comparing the performance between VPR and non-sparse VPR. To specifically evaluate the effect of sparse constraint imposed on the variational scaling factors, we test the performance of non-sparse VPR by removing the sparse term in Eqn.12 and test the performance.

The results are shown in Table 3, and we have the following observations:

(1) Our designed SVS module effectively promotes the performance. On both datasets, VPR outperforms PSN with the accuracies which are more than 2.5 percent higher.

(2) The sparse constraints on latent variables are effective and meaningful. According to Table 3, the sparse constraint brings performance gain on both datasets through an adaptive salient pathway selection.

(3) The performance of PSN verifies the effectiveness of combining regional random walk and pathway embedding (without SVS module) inspired by the perception mechanism, as it is comparable with most state-of-the-art methods shown in Table 1 and 2.

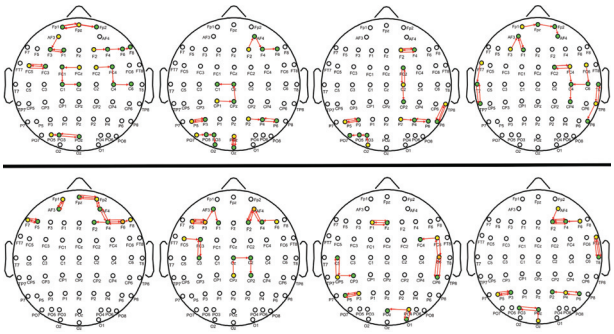

Figure 4: Visualization of salient pathways (positive emotion) of different subjects on SEED and MPED. The first row comes from different subjects in MPED, while the second row comes from different subjects in SEED. Yellow electrodes denote starting nodes, green electrodes are non-starting nodes, and orange dots are reused in different pathways as both starting and non-starting nodes.

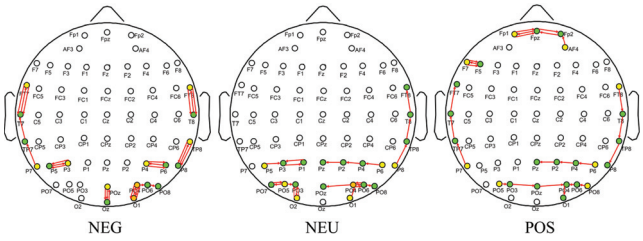

Figure 5: Visualization of salient pathways of different emotion types on SEED. Pathways of all subjects w.r.t. each emotion are gathered and salient pathways are chosen based on statistics of their contribution values.

Visualization of salient pathways.

Aforementioned experiments verify the competitive performance of our proposed VPR and the critical role that the pathways play in emotion recognition. For further intuitive understanding of the pathway selection, we here visualize those salient pathways which are adaptively highlighted by imposing scaling factors of higher values. Meantime, we exhibit the ordered connection variations among different subjects and emotion classes in Fig. 4 and Fig. 5, which may also be rather meaningful from the view of human emotion cognition. In this process, the contribution of candidate pathways to emotion recognition are calculated according to Eqn. 13 and then ranked, where pathways of top-6 contribution values are shown together with additional pathways with very close contribution to the top-6th.

According to those shown salient pathway samples, we have the following observations:

(1) Fig. 4 demonstrates large variation of salient pathways among subjects, where the pathways appear in different positions with different connection patterns.

(2) In statistics of positions, salient pathways appear frequently in emotion related brain regions, i.e. 'Frontal' (F) and 'Temporal' (T). This coincides with the previous neuroscience finding in (Adolphs 2002b).

(3) The locations of salient pathways differ in different emotion types (Adolphs 2002b). Positive emotion additionally involves salient pathways in 'Frontal' (which is related to high level cognition) comparing with negative and neutral, which is reasonable as it may involves more deliberate processings like cognitive appraisal and evaluation, which is supported by (Vytal and Hamann 2010) .

Conclusion.

In this paper, a novel perception mechanism inspired framework named VPR is proposed to deal with EEG emotion recognition. Considering spatial adjacency and functional region, random walk was introduced along spatially closed electrodes within each brain region to generate candidate pathways. To encode each pathway, considering the ordered connection, the LSTM model was employed to learn between-electrode dependencies. To capture the most important pair-wise connections, the encoded pathways around each electrode were aggregated to produce a pseudo maximum-energy pathway. For further selecting salient pathways, the SVS module was proposed using Bayesian probabilistic process and the sparsity constraint to endow the module with good generalization ability meanwhile favor the adaptive pathway selection. Finally, the salient pathways from those candidates were jointly decided by the pseudo pathways and scaling factors. Extensive experiments on EEG emotion recognition demonstrated the superiority of our proposed VPR, and some salient pathways w.r.t. different subjects and emotions are shown for intuitive understanding, which may also be rather meaningful from the view of human emotion cognition.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China (Grants Nos. 61906094, 61972204, 61902064, 61921004), the Natural Science Foundation of Jiangsu Province (Grants Nos. BK20190452, BK20190019), and the fundamental research funds for the central universities (No. 30919011232).

References

Adolphs, R. 2002a. Neural systems for recognizing emotion. *Current opinion in neurobiology* 12(2):169–177.

Adolphs, R. 2002b. Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews* 1(1):21–62.

Bullmore, E., and Sporns, O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience* 10(3):186.

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2012. *Pattern classification*. John Wiley & Sons.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 1999. Learning to forget: Continual prediction with lstm.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864. ACM.

Hardoon, D. R., and Szedmak, Sandor, e. a. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16(12):2639–2664.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR* 2(5):6.

Hochreiter, S. e. a. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Karamanolakis, G.; Cherian, K. R.; Narayan, A. R.; Yuan, J.; Tang, D.; and Jebara, T. 2018. Item recommendation with variational autoencoders and heterogeneous priors. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, 10–14. ACM.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N., and Welling, M. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Li, Y.; Zheng, W.; Cui, Z.; and Zhou, X. 2016. A novel graph regularized sparse linear discriminant analysis model for eeg emotion recognition. In *International Conference on Neural Information Processing*, 175–182. Springer.

Li, P.; Wang, Z.; Lam, W.; Ren, Z.; and Bing, L. 2017. Salience estimation via variational auto-encoders for multi-document summarization. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Li, Y.; Zheng, W.; Cui, Z.; Zhang, T.; and Zong, Y. 2018a. A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition. In *IJCAI*, 1561–1567.

Li, Y.; Zheng, W.; Zong, Y.; Cui, Z.; Zhang, T.; and Zhou, X. 2018b. A bi-hemisphere domain adversarial neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*.

Li, Y.; Zheng, W.; Wang, L.; Zong, Y.; and Cui, Z. 2019. From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. ACM.

Pu, Y.; Gan, Z.; Henao, R.; Li, C.; Han, S.; and Carin, L. 2017. Vae learning via stein variational gradient descent. In *Advances in Neural Information Processing Systems*, 4236–4245.

Scharbrough, F.; Chatrian, G.; Lesser, R.; Luders, H.; Nuwer, M.; and Picton, T. 1990. Guidelines for standard electrode position nomenclature. *Am. EEG Soc*.

Shi, L.-C.; Jiao, Y.-Y.; and Lu, B.-L. 2013. Differential entropy feature for eeg-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6627–6630. IEEE.

Song, T.; Zheng, W.; Song, P.; and Cui, Z. 2018. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*.

Song, T.; Zheng, W.; Lu, C.; Zong, Y.; Zhang, X.; and Cui, Z. 2019. Mped: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access* 7:12177–12191.

Vytal, K., and Hamann, S. 2010. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of cognitive neuroscience* 22(12):2864–2885.

Walker, J.; Doersch, C.; Gupta, A.; and Hebert, M. 2016. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, 835–851. Springer.

Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; and Li, Y. 2018. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics* 49(3):839–847.

Zhao, C.; Ni, B.; Zhang, J.; Zhao, Q.; Zhang, W.; and Tian, Q. 2019. Variational convolutional neural network pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2780–2789.

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.; Zhu, J.-Y.; Peng, Y.; and Lu, B.-L. 2014. Eeg-based emotion classification using deep belief networks. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Zheng, W. 2016. Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transactions on Cognitive and Developmental Systems* 9(3):281–290.