Motor Imagery Classification via Temporal Attention Cues of Graph Embedded EEG Signals

Dalin Zhang, *Student Member, IEEE,* Kaixuan Chen, *Student Member, IEEE,* Debao Jian, and Lina Yao, *Member, IEEE*

***Abstract*—Motor imagery classification from EEG signals is essential for motor rehabilitation with a Brain-Computer Interface (BCI). Most current works on this issue require a subject-specific adaptation step before applied to a new user. Thus the research of directly extending a pre-trained model to new users is particularly desired and indispensable. As brain dynamics fluctuate considerably across different subjects, it is challenging to design practical hand-crafted features based on prior knowledge. Regarding this gap, this paper proposes a Graph-based Convolutional Recurrent Attention Model (G-CRAM) to explore EEG features across different subjects for motor imagery classification. A graph structure is first developed to represent the positioning information of EEG nodes. Then a convolutional recurrent attention model learns EEG features from both spatial and temporal dimensions and emphasizes on the most distinguishable temporal periods. We evaluate the proposed approach on two benchmark EEG datasets of motor imagery classification on the subject-independent testing. The results show that the G-CRAM achieves superior performance to state-of-the-art methods regarding recognition accuracy and ROC-AUC. Furthermore, model interpretation studies reveal the learning process of different neural network components and demonstrate that the proposed model can extract detailed features efficiently.**

***Index Terms*—EEG, Motor Imagery, Deep Learning**

I. Introduction

Motor imagery classification is the basic to a BCI, which supports motor rehabilitation of post-stroke patients [1]. The EEG signals, which are captured from a human's scalp and thus reflect the electrical activities of human the cortex, is one of the most active physiological cues to build a BCI system. Researchers have widely explored the EEG-based BCI due to its zero clinical risks as well as portable and cost-effective acquisition devices.

In recent years, there have been substantial achievements in EEG-based motor imagery classification. However, the outstanding works generally focus on the subject-dependent scenario, where training and test data are from the same group of subjects [2]. In this condition, a brief calibration session is essential before a BCI system is ready to be used by a new user [3]. This adaptation process needs to be performed on each new subject and in each usage, which is labor-intensive and time-consuming, resulting in limited usability and scalability of BCI systems. It is essential to overcome this subject-independent issue. However, the apparent changes in EEG signals across different subjects cause enormous challenges in solving such a problem [4].

Traditional EEG analysis methods depend on hand-crafted features and subsequent machine learning algorithms. One of the most popular hand-crafted features is the power spectral density (PSD). The phenomenon of EEG power in some frequency bands increasing/decreasing, which is called event-related synchronization/desynchronization (ERS/ERD), are widely observed when analyzing the PSD patterns of motor imagery EEG signals [5]. In the meantime, not all EEG nodes can provide distinguishable information in terms of PSD features. An EEG channel selection approach is usually preferred to choose the most discriminative EEG nodes [6]. C3, C4, and Cz are three commonly reported channels that are most useful for motor imagery classification. However, there are some drawbacks when using the hand-crafted features. First, previous studies disagreed with the range of some frequency bands. For example, [7] defined the mu band between 8-13Hz, while [8] defined the mu band between 8-12Hz. Second, the amount of effective EEG nodes that are selected by a channel selection algorithm is generally decided by an expert's knowledge and experience. Third, in these traditional works, all steps are separated, which is meaningless and not only wastes time but also prevents the potential that different steps may promote each other during the feature learning process. Even though powerful state-of-the-art classifiers have been used on hand-crafted features and achieved partial improvements in performance [9], human-designed features may neglect the critical information within raw EEG signals [10].

In contrast to hand-crafted features, deep learning methods can learn the underlying information across different subjects automatically [11]. Considerable effort has been devoted to developing EEG analysis approaches using deep learning techniques and achieved promising results [12]–[15]. Reference [13] presents a compact convolutional neural network (CNN) and demonstrates its success on different EEG paradigms. To make use of the temporal dynamics efficiently, [14] proposes to adopt a recurrent neural network (RNN) of long short-term memory (LSTM) cells besides CNNs. Some works also combine traditional spectral features and deep learning methods [12], [15]. Despite the success of deep learning in EEG analysis, few deep learning works build a motor imagery classification model demonstrating generalization abilities on new subjects [16], [17].

To solve the subject-independent problem of EEG-based

D. Zhang, K. Chen, D. Jian, and L. Yao are with the School of Computer Science and Engineering, University of New South South, Australia. Email: {dalin.zhang, d.jian, lina.yao}@unsw.edu.au, kaixuan.chen@student.unsw.edu.au

Manuscript received xx xx, 2019.

motor imagery classification, this work proposes a novel Graph-based Convolutional Recurrent Model (*G-CRAM*) that efficiently learns the spatial information with the aid of graph representations of EEG nodes and extracts attentional temporal dynamics using a recurrent attention network. First, we utilize the spatial positioning of EEG nodes to form the EEG graph to explicitly exhibit the spatial information of EEG node connections. Different from previous spatial filtering methods, the proposed graph embedding does not rely on subjects or tasks, thus being more robust on new subjects. Then, a sliding window technique is used to split the EEG representations into multiple consecutive temporal slices, and a specifically designed CNN structure is designed to learn spatio-temporal traits within an EEG temporal slice. Lastly, we employ a recurrent attention network to acquire the temporal dependencies across different EEG temporal slices. In the standard recurrent module, the temporal cues are usually accumulated to the last time step and consequently used for classification that some critical information in early time steps may be omitted. In contrast, the proposed model assigns weights to different temporal cues and aggregates all information for the final classification. This study employs two benchmark EEG motor imagery datasets to validate the proposed method in a subject-independent manner and demonstrates its superiority to a series of comparison approaches. Besides the performance improvement, understanding how the model works is also a critical and attractive research topic. Besides the overall performance evaluation, insights of how each part of the presented model works are also investigated and discussed in details. A preliminary version of this work has been reported [18]. The implementation code is made publicly available[1]. The main contributions of this work are summarized as follows:

- This work designs a novel deep learning method for the EEG motor imagery classification task. The proposed model uses a graph embedding to represent EEG spatial information, which differs from other spatial filtering methods by its independence of subjects and tasks. A recurrent attention module is then developed to assign weights to different temporal cues, instead of relying on accumulative temporal information by a standard recurrent network;
- We carry out comprehensive experiments on two benchmark datasets with the subject-independent setting, which is rarely reported in previous studies. The results show the superior generalization performance of the proposed method on new subjects;
- We provide detailed insight discussions on the model interpretation of the neural network and performance impact of the graph representation. The CNN prefers to focus on small brain areas, and the recurrent attention network not only focuses on the last temporal step but also gives high weights to early steps. The results also indicate that the graph embedding impacts more on the dataset of a larger number of EEG nodes.

[1]https://github.com/dalinzhang/GCRAM

II. Related Work

A. EEG Motor Imagery Classification

EEG-based motor imagery classification is the basis of many synchronous BCIs, and abundant approaches have been published on this task. Common Spatial Pattern (CSP) is one of the most popular and effective feature extraction methods in motor imagery EEG classification [19]. It is a spatial filtering approach that tries to find a linear combination of EEG channels that the power difference of different motor imagery classes is maximized [20]. There have been lots of works reported to extend CSP and achieve remarkable improvement [21], [22]. One of its most successful variants is the Filter Bank CSP (FBCSP), which addresses CSP's drawback of depending on a particular frequency band by applying CSP to different frequency bands and selecting subject-specific features by a feature selection method [22]. FBCSP was the state-of-the-art method in motor imagery EEG classification and has provided excellent results [23]. In terms of classifiers, traditional algorithms like SVM and LDA are commonly used in many EEG motor imagery studies [22], [24].

Due to its end-to-end structure and superior performance, deep learning has been applied to classifying motor imagery in some studies. [24] proposes a carefully designed CNN with and a crop training strategy and achieves better performance than FBCSP. [13] designs a lightweight CNN that shows competitive performance to state-of-the-art methods in diverse BCI paradigms, such as motor imagery and P300. RNN is also extensively used to extract temporal features in the motor imagery classification task. [25] proposes to fuse the CNN and RNN with fuzzy integral and leverage the reinforcement learning technique to optimize the fuzzy measures. Feature engineering is also used to improve the performance of deep learning models. Power spectral features are popular for motor imagery classification due to previous evidence of its discriminative ability [26]. [27] proposes to use the CNN and the power spectral density features for motor imagery classification and achieves competitive accuracy.

B. Graph theory and attention model

Graph theory can be used to model many types of relations and has been applied to different areas such as human action recognition [28], recommendation system [29] and anomaly detection [30]. Kipf et al. propose a Graph Convolutional Networks (GCN) for semi-supervised citation-based document classification [31]. Given the labels of some nodes, a neural network tries to infer the labels of the rest nodes with the aid of graph theory which represents the relations between all the nodes. Considering that the spatial relations between different EEG nodes are crucial to successful EEG analysis [7], [12], [32], we propose to utilize the graph theory to represent the spatial relationships of EEG nodes in this work.

The attention-based neural network architecture has shown promising power on various tasks, like speech recognition [33], machine translation [34], and activity recognition [35], [36]. The self-attention mechanism is a specific and powerful soft attention mechanism [34]. Ashish et al. entirely rely on the self-attention mechanism for the sequence to sequence task

without using traditional CNN or RNN structures and achieve the state-of-the-art results [37]. It has also shown promising performance in various tasks like abstractive summarization [38], intention recognition [39], and textual entailment [40]. Regarding the characteristic of the self-attention mechanism and human's attention is focused on different periods, we adopt the self-attention mechanism to emphasize on different EEG temporal periods.

III. Methodology

A. Problem Definition

Before going into the methodology details, we first briefly introduce the motor imagery experiment process and formally define the research problem. Fig. 1 shows the timing scheme of a typical motor imagery experiment. The beep and cue are used to notice and indicate the subject to perform the motor imagery task. The duration of motor imagery is of research interest.

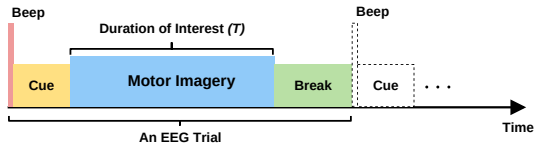


Fig. 1. A timing scheme of a motor imagery EEG acquisition experiment

Formally, the duration of interest is T-second long. Each of the n EEG nodes has a sensor recording sequence $\mathbf{r}_{i\in[1,n]} = [s_1^i,\ s_2^i, ..., s_k^i] \in \mathbb{R}^k$ through $k = T \times f$ time points, where f is the sampling frequency and s_t^i is the measurement of the ith EEG sensor at the time point t. Thus the raw EEG features of the trial T is a two-dimensional (2D) tensor $\mathbf{X}_T = [\mathbf{r}_1;\ \mathbf{r}_2;\ \ldots\ ;\mathbf{r}_n] \in \mathbb{R}^{n\times k}$ with one dimension representing EEG node and the other representing time series. Our goal is to make motor imagery classification of the EEG trials $\mathbf{X}_T$. The following experiment results are based on single-trial subject-independent testing, where training and testing trials are drawn from different groups of subjects.

B. Pipeline Overview

Fig. 2 shows an overview of our proposed approach. The EEG signals are first embedded by a graph representation. Three different graph embedding schemes are developed based on different considerations. The embedded EEG signals are then cut into slices along the time dimension and fed into the attention-based neural network, which can better. extract the temporal features for the motor imagery classification. The model is an end-to-end framework that can be trained by standard back-propagation.

C. Represent EEG Node Connections

In the node dimension of $\mathbf{X}_T$, one EEG node at most has two neighbors. Such a representation is limited to reflect the real-world situation where an EEG node usually has multiple neighboring nodes acquiring EEG signals of a certain brain area. Thus representing the relations of different EEG nodes is essential to successful EEG analysis. In our work, we leverage the EEG node positioning to form graph representations of EEG nodes, which include spatial information of the natural EEG node connections. In particular, we construct an undirected spatial graph $G = (V, E)$ on the EEG node positioning. The node set $V = \{s^i | i \in [1, n]\}$ includes all the EEG nodes in an experiment. Depending on the structure of the adjacency matrix of EEG nodes, we design three EEG representation graphs: N-Graph (NG), D-Graph (DG), and S-Graph (SG). The graph definition enhances the brain area representation ability of EEG signals but decreases the effect of noise on each EEG node by combining neighboring nodes to represent the central one. This design also empowers the EEG representations to be robust to missing value issues by embedding each EEG node with the assist of its neighboring nodes instead of only relying on the measurement of itself.

1) N-Graph: Fig. 3 shows an example positioning of 64-channel EEG nodes. In the 2D position projection (Fig. 3 (b)), each node has several naturally neighbors (up, down, left, right, up-left, up-right, down-left, and down-right); for example, the node s^{11} has eight neighbouring nodes (s^3, s^4, s^5, s^{12}, s^{19}, s^{18}, s^{17}, s^{10}). Based on this observation, we build a connection between two naturally neighboring EEG nodes. Formally, the edge set can be denoted as $\mathrm{E}_v = \{s^i s^j | (i, j) \in \mathrm{H}\}$, where H is the set of naturally neighboring EEG nodes. We also regard each node as connecting to itself. We can define the adjacency matrix of the N-Graph as a square matrix $|\mathrm{V}| \times |\mathrm{V}|$ with its binary element representing whether two EEG nodes are neighboring to each other:

$$A_{ij} = \begin{cases} 1 & if\ \ s^i s^j \in E_v \\ 0 & else. \end{cases}$$

We then follow the spectral graph theory [31] to normalize the adjacency matrix: $\hat{A}_v = \widetilde{D}_v^{-\frac{1}{2}} \widetilde{A}_v \widetilde{D}_v^{-\frac{1}{2}}$, where $\widetilde{A}_v = A_v + I_n$, $\widetilde{D}_v = \mathrm{diag}(\sum_j A_{1j}, \sum_j A_{2j}\ \ldots\ \sum_j A_{|V|j})$ is the diagonal node degree matrix, and $\widetilde{D}_v^{-\frac{1}{2}} = \mathrm{diag}(\frac{1}{\sqrt{\sum_j A_{1j}}}, \frac{1}{\sqrt{\sum_j A_{2j}}}\ \ldots\ \frac{1}{\sqrt{\sum_j A_{|V|j}}})$. Then the N-Graph representation Z_v of raw EEG signals is the matrix product of the normalized N-Graph adjacency matrix $\hat{A}_v$ and the raw EEG trial X_T:

$$Z_v = \hat{A}_v X_T,\ \ Z_v \in \mathbb{R}^{n\times k}.$$

2) D-Graph: The adjacency matrix of N-Graph, is a simple binary embedding that roughly represents the spatial information without refined depiction of EEG node spatial relationships. The binary adjacency matrix considers all neighboring nodes contribute equally to the central node, while the real-world situation is that those relatively distant neighboring nodes have less influence and the relatively adjacent neighboring nodes have more influence on the central nodes. In Fig. 3 (b) for example, the central node s^{11} may be influenced to different degrees by its eight neighboring nodes (s^3, s^4, s^5, s^{12}, s^{19}, s^{18}, s^{17}, s^{10}) based on the spatial distance between neighboring nodes to the central node s^{11}. The simple binary adjacency matrix is not flexible and not able to convey such kind information.

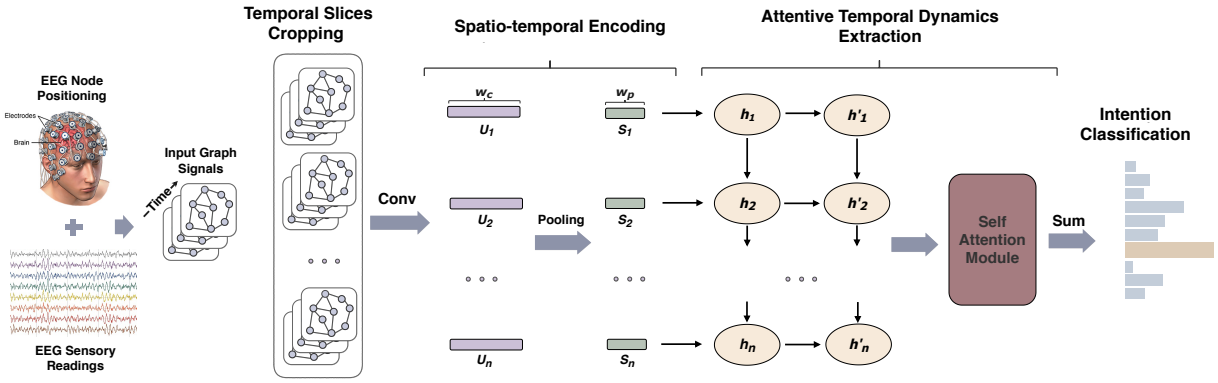
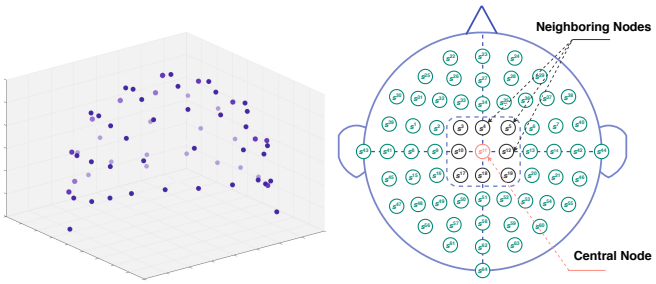


Fig. 2. Overview of the Graph Convolutional Recurrent Attention Model (G-CRAM) on EEG motor imagery classification. We first represent the raw EEG measurement by a spatial graph drawn from EEG node positions; then we apply a sliding window technique to crop continuous EEG sequences into temporal slices and utilize a CNN layer to extract spatio-temporal features of each slice; a recurrent attention layer is used to extract the attentive temporal dynamic features; lastly the extracted features are classified to the target using a dense layer and a standard softmax classifier.



(a) 3D positioning of EEG nodes (b) 2D position projection

Fig. 3. An example positioning of 64-channel EEG nodes.

Considering the above disadvantages, we define a distance-based EEG graph called D-Graph, which uses the real-world 3D distance between EEG nodes rather than the binary connections between naturally neighboring nodes. The adjacency matrix of D-Graph has the distance between two neighboring EEG nodes as its element instead of binary elements indicating neighboring or not. First, we define the set of the distance between any two EEG nodes as $L = \{d_{ij} | (s^i, s^j) \in V^2, \ i \neq j\}$, where d_{ij} is the Euclidean distance between node s^i and s^j. The locations of the EEG nodes are from the international 10-10 system [41] with the three-dimensional Talairach coordinate representation. In practice, two issues should be addressed before constructing the adjacency matrix: 1) how to define neighboring nodes; 2) how to define the distance between a node and itself. For the first problem, we regard the two EEG nodes are neighboring if the distance between two nodes is smaller than the average value of the distance set L. For the second problem, the distance between a node and itself is defined as the average distance of other neighboring nodes to this node. Therefore, we define the elements of the adjacency matrix A_d as:

$$A_{ij} = \begin{cases} \frac{1}{d_{ij}} & if \ \ d_{ij} < E(L) \\ 0 & if \ \ d_{ij} \geqslant E(L) \\ \frac{1}{E(\{d_{iq} | d_{iq} < E(L), \ \ q \in [1,n]\})} & if \ \ i = j \end{cases}$$

where $E(L)$ is the average of distance set L. Similar to the N-Graph, the D-Graph adjacency matrix is also normalized to $\hat{A}_d = \widetilde{D}_d^{-\frac{1}{2}} A_d \widetilde{D}_d^{-\frac{1}{2}}$, where $\widetilde{D}_d = \text{diag}(\sum_j A_{1j}, \sum_j A_{2j} \ \ldots \ \sum_j A_{|V|j})$ is the diagonal degree matrix, and $\widetilde{D}_d^{-\frac{1}{2}} = \text{diag}(\frac{1}{\sqrt{\sum_j A_{1j}}}, \frac{1}{\sqrt{\sum_j A_{2j}}} \ \ldots \ \frac{1}{\sqrt{\sum_j A_{|V|j}}})$.

The raw EEG trial X_T represented with the D-Graph spatial information is:

$$Z_d = \hat{A}_d X_T, \ \ Z_d \in \mathbb{R}^{n \times k}.$$

3) S-Graph: In the definition of D-Graph, the distance between a node and itself is defined as the average distance of other neighboring nodes to this node. Another strategy to defining the self-distance is to use the shortest distance from a node's neighbors to itself. We call this graph definition S-Graph. Similarly, its adjacency matrix element is defined as:

$$A_{ij} = \begin{cases} \frac{1}{d_{ij}} & if \ \ d_{ij} < E(L) \\ 0 & if \ \ d_{ij} \geqslant E(L) \\ \frac{1}{\text{Min}(d_{iq} | q \in [1,n]\})} & if \ \ i = j \end{cases}$$

The S-Graph is also normalized in the same way to avoid changing the scale of X_T. The final representation of the S-Graph is:

$$Z_s = \hat{A}_s X_T, \ \ Z_s \in \mathbb{R}^{n \times k}.$$

Different from other spatial filtering like CSP [42], the graph embedding only relies on the real-world node placement, so it is independent of subjects, targeting tasks, and manually-set parameters. Therefore, it is a supplement to the following neural network, which is a data-driven feature learning approach.

D. Spatio-temporal Encoding

After embedding raw EEG signals, a sliding window is applied to cut the EEG representations along the time dimension into several temporal slices $Q_i \in \mathbb{R}^{n \times w}$, where w is the temporal slice length. Let the interval between two neighbouring slices be p, then $m = \text{int}((k - w)/p)$ slices are obtained from one EEG trial. We specifically design a CNN

TABLE I
THE CONFIGURATIONS OF THE SPATIO-TEMPORAL ENCODING NETWORK

Layer	Kernel Size/Stride	Kernel #	Padding	Activation
Convolution	$n \times 45/1$	40	valid	ELU
MaxPool	$1 \times 75/10$	40	valid	-

to encode the spatio-temporal information within a temporal slice.

Although deep networks have strong learning abilities, deeper is not always better for EEG analysis [43]. TABLE I gives the detailed configuration of the proposed spatio-temporal encoding network. We use one CNN layer and one pooling layer. The height of the CNN kernel is set to n, same to the amount of EEG nodes, for considering all EEG nodes at once. The width of the kernel is extended to 45 for exploring long temporal dynamics. The output amount of CNN filters is empirically set to 40. The convolutional filtering thus can uncover the spatio-temporal information across different EEG nodes. Each temporal slice is encoded to higher-level representations $\{U_i \in \mathbb{R}^{w_c} | U_i = \text{Conv}(Q_i),\ i \in [1, m]\}$. The activation function used in the convolutional operations is the *Exponential Linear Unit* (*ELU*) function. We use the $valid$ padding option. Thus the output of the CNN layer has the height of 1. A maxpooling layer is then applied to reduce the number of parameters and extract important information. The final encoded representation is $\{S_i \in \mathbb{R}^{w_p} | S_i = \text{MaxPool}(U_i),\ i \in [1,\ m]\}$.

E. Exploring Attentive Temporal Dynamics

Following the spatio-temporal feature extraction within single EEG temporal slices, a recurrent attention network is introduced to discover the attentive temporal dependencies across different EEG temporal slices. In traditional recurrent networks, the features that are accumulated from the previous time step are usually adopted for further analysis. However, some crucial information in early steps may be forgotten inevitably due to the structural limitation of recurrent cells. To overcome this issue, we propose to utilize a self-attention module to assign adaptive weights to different recurrent time steps, so that information in early time steps is preserved and flexibly incorporated.

The Long Short-Term Memory (LSTM) units are used to build two stacked RNN layers. After flattening the output of the previous spatio-temporal encoding, m 1D vectors are obtained and input into the RNN. Therefore, each RNN layer has m LSTM cells. The output of the RNN is the hidden states of the second recurrent layer $\{h_i^{'} \in \mathbb{R}^l | h_i^{'} = LSTM(S_i),\ i = 1...m\}$, where l is the hidden state size.

Because a subject usually concentrates on the experiment at some time but is distracted at the other time and different subjects pay attention a different time within a trial, emphasizing on the EEG temporal slices when a subject concentrates on the experiment while neglecting the other slices is necessary for successful EEG analysis. A self-attention mechanism [34], as illustrated in Fig. 4, is used to allocate adaptable weights to different input elements according to their values and aggregate

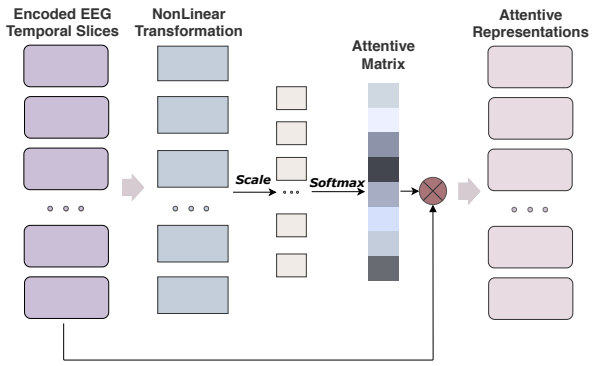


Fig. 4. Illustration of the self-attention module. A nonlinear encoding layer first transforms the encoded EEG temporal slices and the results are scaled and normalized to get the attention weight of each temporal slice. Lastly, the attention weight is multiplied with its corresponding encoded features.

this information to form a final representation. One important feature of the self-attention mechanism is that the weight values are adapted according to the input values, thus meets the demand of subject-independent EEG signal analysis where different subjects concentrate on different temporal periods. Each slice representation $h_i^{'}$ is first non-linearly transformed into a latent space:

$$H_i = tanh(W_i h_i^{'} + b_i),\ \ H_i \in \mathbb{R}^{h_a}$$

where $W_i \in \mathbb{R}^{l \times h_a}$ and $b_i \in \mathbb{R}^{h_a}$ are the input-to-hidden weight matrix and bias for a hidden layer of size h_a. The softmax activation function, defined as $\text{softmax}(x_i)=\frac{1}{\mathcal{Z}}\exp(x_i)$ with $\mathcal{Z}=\sum_i \exp(x_i)$, is applied to the nonlinear latent representation H_i to obtain the weight of importance for each slice:

$$V_i = \frac{exp(H_i^{\intercal} v_i)}{\sum_i exp(H_i^{\intercal} v_i)}.$$

The slice attention vector $v_i \in \mathbb{R}^{h_a}$ is randomly initialized and jointly learned during the training process. The $softmax()$ function guarantees that all the computed weights sum to 1. This weight matrix will focus on specific temporal slices that are more distinguishable than others. Lastly, in the interest of computational efficiency, a weighted sum of all EEG temporal slices is computed to a slice-focused representation:

$$A = \sum_i V_i h_i^{'},\ \ A \in \mathbb{R}^l.$$

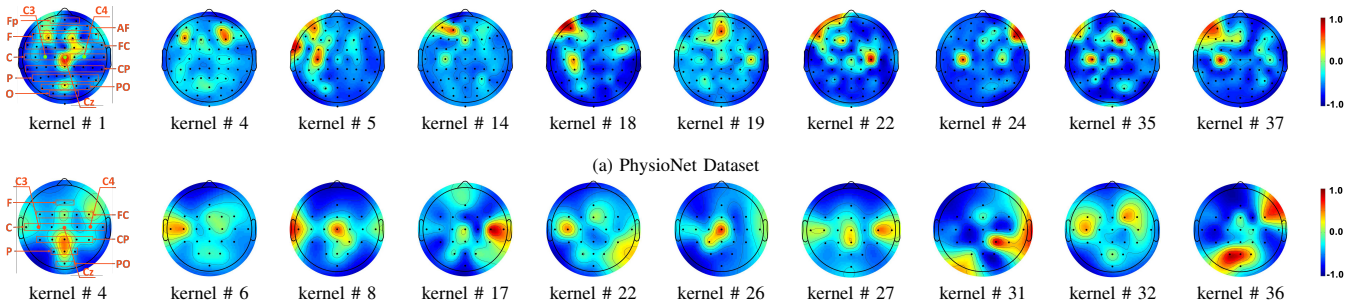
The attentive temporal dynamic representation A is fed into a standard softmax classifier:

$$P = softmax(WA + b),$$

where W and b are weight and bias matrices respectively of the motor imagery classification layers. Then the cross-entropy error over all labeled samples is evaluated:

$$\mathcal{L} = -\sum_c \hat{Y}_c log(P_c),$$

where $\hat{Y}_c$ and P_c is the label and the classification probability of motor imagery strategy c respectively. The network weights and biases are trained with batch gradient descent. The final classification result is defined as the motor imagery strategy with max classification probability.



(a) PhysioNet Dataset

(b) BCICIV2a Dataset

Fig. 5. Visualization of the convolutional feature maps in topographic plots. We select 10 representative feature maps for each dataset. The feature values of all EEG nodes are normalized to range [-1, 1]. A large value indicates a major impact on subsequent motor imagery classification; in contrast a small value represents minor impact. # represents the digital label of the feature map.

IV. Experiment and Results

A. Dataset and Implementation Details

The proposed method is evaluated on two widely used benchmark EEG dataset: PhysioNet EEG Motor Imagery Dataset [44] and BCI Competition IV dataset 2a [45].

1) PhysioNet Dataset: The PhysioNet dataset comprises 109 healthy subjects executing left/right fist open and close imagery. The EEG data is collected using BCI2000 instrumentation with 64 EEG nodes and a 160Hz sampling rate. Each trial lasts about 3.1 seconds resulting in 497 recording time steps. After data inspection, we remove the data of subject #88, #89, #92, and #100 because of the damaged recordings with multiple consecutive "rest" sections. As a result, we have 105 subjects in total.We prepared nine subject-independent evaluation sets (A01-A09), each of which contains ten randomly selected subjects as a test and the remaining 95 subjects as training. Because each subject has around 43 trials with a roughly balanced ratio in the right and left fist motor imagery, there are about 4085 trials in one training set and 430 trials in one test set.

2) BCICIV2a Dataset: The BCICIV2a dataset contains EEG signals of 22 nodes recorded with nine healthy subjects and two sessions on two different days. Each session consists of 288 four-second trials of motor imagery per subject (imagining the movement of the left hand, the right hand, the feet, and the tongue). The signals were sampled with 250 Hz and bandpass-filtered between 0.5Hz and 100Hz by the dataset provider before release. The original dataset uses the 288 trials of the first session as training and the 288 trials of the second session as a test. However, in the subject-independent scenario, the original dataset needs to be re-split by subject with the leave-one-subject-out manner. Consequently, nine evaluation datasets (A01-A09) are achieved, each of which has 576 trials (288 trials $\times$ 2 sessions) of one subject as a test and 4608 trials (288 trials $\times$ 2 sessions $\times$ 8 subjects) of the remaining eight subjects as training.

3) Implementation Details: One of the crucial advantages of the deep learning technique lies in its no need for hand-crafted features. By following the conventions [43], [46], we directly feed the raw EEG data into the proposed framework pipeline without any filtering. The sliding window size for both datasets is 400 and the step of is 10 and 50 for the PhysioNet and BCICIV2a dataset respectively. We make use of the TensorFlow framework for a GPU-based implementation using matrix multiplications. The stochastic gradient descent with Adam update rule is used to minimize the cross-entropy loss function. The network parameters are optimized with a learning rate of 10^{-5}. Dropout regularization is applied after the CNN layer and the recurrent network layer with the dropout probability of 0.5. The hidden state size of the LSTM cell l is 64. The non-linear transformation size of the self-attention is 512. The proposed model has 16 hyper-parameters and 420,356 trainable parameters.

TABLE II
Comparison with state-of-the-art and baseline models

Dataset	Comparison Model	Evaluation Criterion	
		Accuracy	ROC-AUC
PhysioNet	EEGNet [13]	0.6994±0.0337	0.7738±0.0377
	CTCNN [43]	0.7128±0.0346	0.7542±0.0413
	EEG Image [15]	0.5600±0.0171	0.5591±0.0163
	Cascade Model [46]	0.6129±0.0252	0.6573±0.0305
	Parallel Model [46]	0.5858±0.0183	0.6187±0.0259
	FBCSP [22]	0.5902±0.0301	0.6193±0.0388
	PSD-SVM [47]	0.5542±0.0156	0.5552±0.0150
	CNN	0.6877±0.0233	0.7249±0.0420
	RNN	0.5389±0.0176	0.5324±0.0233
	CRAM	0.7291±0.0488	0.8008±0.0565
	NG-CRAM	0.7420±0.0449	0.8133±0.0464
	DG-CRAM	**0.7471±0.0419**	**0.8164±0.0465**
	SG-CRAM	0.7408±0.0443	0.8128±0.0512
BCICIV2a	EEGNet [13]	0.5130±0.0518	0.7722±0.0442
	CTCNN [43]	0.4767±0.1506	0.7703±0.1241
	EEG Image [15]	0.3270±0.0430	0.5695±0.0466
	Cascade Model [46]	0.3183±0.0399	0.5796±0.0438
	Parallel Model [46]	0.3267±0.4499	0.5908±0.0478
	FBCSP [22]	0.3569±0.0853	0.6553±0.1093
	PSD-SVM [47]	0.3611±0.0817	0.5752±0.0522
	CNN	0.4720±0.0477	0.7176±0.0426
	RNN	0.3548±0.0228	0.6069±0.0291
	CRAM	0.5910±0.1085	0.8186±0.0805
	NG-CRAM	**0.6011±0.0996**	**0.8208±0.0709**
	DG-CRAM	0.5964±0.0947	0.8154±0.0652
	SG-CRAM	0.5900±0.1015	0.8108±0.0681

B. Experimental Results

1) Comparison Results: The PhysioNet dataset and BCICIV2a dataset we used are roughly balanced. Thus we evaluate

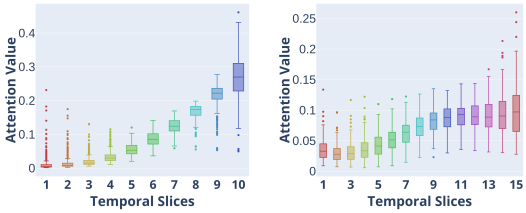
the proposed model with classification accuracy and the Area Under ROC Curve (ROC-AUC). TABLE II presents the overall comparison results and the detailed results can be found in the supporting documents. Because deep learning is an advanced technique that relies on proper structure design, we compare with several deep learning approaches with various model structures and feature embedding strategies. To make a fair comparison and show the superior structure of the proposed approach, the most recent state-of-the-art approaches whose implementation code is available online are selected for comparison. We first make a comparison with the recently published EEGNet [13], which encapsulates well-known EEG feature extraction concepts for BCI to construct a uniform approach for different BCI paradigms. Then the proposed model is compared with the CTCNN (CroppedTrainingCNN) [43] method, as this work reports comprehensive research on various CNN architectures and proposes a crop training strategy which outperforms the traditional trial-based training manner. A further comparison with the EEG-Image [15] approach is performed. The EEG-Image model selects three widely explored aspects of EEG signals: spectral, spatial, and temporal as prior features and proposes a carefully designed convolutional recurrent architecture for the mental workload classifcation. The Cascade and Parallel CRNN reported in [46] are also used for comparison, as this work also reports to preserve EEG spatial information by considering adjacent EEG nodes. It provides state-of-the-art results on the PhysioNet dataset in the subject-dependent scenario. Lastly, the proposed G-CRAM is compared with two traditional EEG analysis methods, PSD-SVM [47] and FBCSP [22]. The PSD provides time-frequency features that are commonly used in traditional EEG motor imagery analysis. FBCSP is a widely used traditional method and has won several BCI competitions. Apart from the state-of-the-art approaches, the proposed model is further compared with two self-built baseline models: CNN and RNN. The CNN model has three CNN layers directly applied on the raw EEG trials. The RNN model has two LSTM-based RNN layers to find the temporal relationships between different slices in an EEG trial. We also compare the proposed method with the CRAM model to demonstrate the effectiveness of the graph representation of EEG signals.

In TABLE II, the NG-CRAM, DG-CRAM, and SG-CRAM represent the G-CRAM with N-Graph, D-Graph, and S-Graph respectively. All three models outperform comparison methods in terms of accuracy and ROC-AUC on the subject-independent testing. The primary reason that the proposed methods surpass traditional methods like FBCSP is the multiple nonlinear transformation process which is a main advantage of deep learning frameworks. When compared with deep learning models, our proposed methods have two main advantages that help to produce superior performance. The first advantage is our proposed graph embedding method. Compared with the pure deep learning models which do not have particular data representations, like EEGNet and CTCNN, our proposed graph representation embeds the spatial relationship of EEG nodes, which facilitates the following neural network to analyze EEG signals. On the other hand, compared with the EEG-Image model, which has a spatial representation, our graph representation scheme does not rely on data implanting, so that is free of the risk of introducing noises. Finally, different from the cascade model and parallel model, which only adopt naive channel re-arrangement, our graph scheme introduces an adjacency matrix to optimize the raw data to a more effective embedding. The second advantage is the recurrent attention module. Compared with the naive recurrent network, the recurrent-attention module not only takes the temporal information into consideration but also assigns adaptive weights to different time periods within an EEG trial. The recurrent-attention module has already been widely demonstrated more powerful than the traditional recurrent network in exploring temporal features [48].

2) How does the G-CRAM encode spatial information?: In this section, we analyze the learning process of the CNN layer to show how EEG features are learned for motor imagery classification. Specifically, instead of summing up the convolutional results along the EEG node dimension, we retain the convolutional results in the EEG node dimension and average the results along the time dimension. Therefore, a feature vector of size n is achieved after ELU activation with each element representing the extracted features of each EEG node from the CNN layer. The feature vector is then normalized to [-1, 1] and visualized with topographic scalp plots.

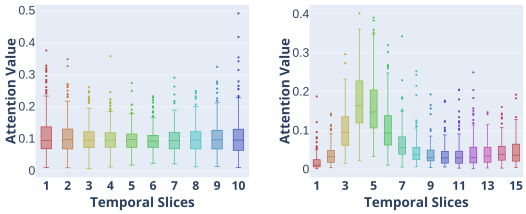
Fig. III-E presents 10 representative topographic scalp plots of convolutional feature maps for each evaluation datasets. As shown in Fig. III-E, the CNN layer focuses on relatively small detailed brain areas, which is important for successful EEG feature extractions [12]. Furthermore, it is consistent with previous reports [49], [50] that the CNN layer emphasizes on the central (FC, C, and CP) and frontal (F)/pre-frontal (Fp) areas of a human brain. More specifically, some convolutional feature maps activate at the three EEG nodes C3, C4, and Cz, which are widely demonstrated holding the most distinguishable information regarding EEG-based motor imagery classification in previous studies [47], [51]. For example, as presented in Fig. III-E (a) of the PhysioNet dataset, kernel # 1 focuses on Cz; kernel # 22 focuses on C4; kernel # 5, # 24, # 35 and # 37 focus on C3. For the BCICIV2a dataset presented in Fig. III-E (b), kernel # 8, # 26 and # 27 focus on Cz; kernel # 17 and # 27 focus on C4; kernel # 6, # 22, # 27 and # 32 focus on C3. Besides, the CNN layer also learns to target other EEG nodes, which is helpful to discriminate different motor imagery tasks as well [52], [53], especially for different subjects and paradigms. Therefore, the spatio-temporal encoding layer is able to act as a spatial filter to extract features of the most distinguishable EEG nodes.

3) How does the recurrent attention module work?: In order to understand how the recurrent attention module learns EEG features, we collect the attention matrices of the correctly classified test samples of the DG-CRAM model and plot the statistical results in Fig. 6. The elements in the attention matrix indicate the weight values assigned to the corresponding RNN output. In Fig. 6, larger number on X axis means later in time. It is obvious that most temporal slices that are later in time have larger weight values, suggesting more influence on the final classification results. This trend shows that the recurrent attention module tends to focus more on later temporal slices.



(a) PhysioNet A05 (10 subjects)(b) BCICIV2a A09 (1 subjecst)

Fig. 6. Box plot of the attention matrices of the correctly classified test samples **with RNN layers**. Larger number on X axis means later in time.



(a) PhysioNet A05 (10 subjects) (b) BCICIV2a A09 (1 subject)

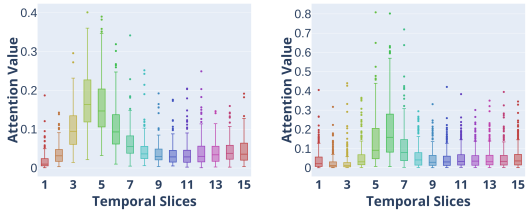
Fig. 7. Box plot of the attention matrices of the correctly classified test samples **without RNN layers**. Larger number on X axis means later in time.

Intuitively, different subjects have different ways of thinking and would concentrate on different temporal periods. Thus effective attention would focus on different temporal periods. However, considering the input of the self-attention module is the features from the previous RNN layers that accumulate the information from early time step and aggregate gradually to the final time step, the later time steps would have more information than earlier time steps. Hence the self-attention module would give larger weights to later RNN output.

To make further exploration, we build a comparison model with the self-attention module directly after the pooling layer without RNN layers in between. We collect the comparison model's attention matrices of the correctly classified test samples and plot the statistical results in Fig. IV-B2. It is found that there is no such a "later-higher" trend in the attention matrices, demonstrating the RNN layer is the cause of the "later-higher" trend in the attention matrices. In the PhysioNet results, the attention matrices show an even distribution on different temporal slices (Fig. IV-B2(a)), while in the BCICIV2a results, there is a peaking trend throughout the attention matrices (Fig. IV-B2(b)). The main reason is that there are ten subjects in the PhysioNet evaluation set; thus, the attention matrix distribution tends to be even. By contrast, there is only one subject in each BCICIV2a evaluation set, so the attention matrix exhibits a clear subject-specific pattern. As shown in Fig. 8(a), (b), and (c), different subjects have different attention patterns indicating strong variability across subjects.

C. Discussion

1) Robustness to Artifacts: Traditional methods use frequency filtering to remove high- and low-frequency artifacts before classification. In contrast, as deep learning is a new technique that combines feature extraction and classification



(a) BCICIV2a A09 (b) BCICIV2a A06

Fig. 8. Box plot of the attention matrices of the correctly classified test samples **without RNN layers** of **individual subjects**

into an end-to-end framework, it is powerful to process raw data directly. Besides, [43] argued that CNN could learn to work as a frequency filter to extract the band power in frequency bands relevant to motor imagery. Therefore, CNN in the proposed model can help to minimize artifacts. From the view of experiment results, the proposed method outperforms the FBCSP, which uses frequency filtering to remove artifacts. As a result, the proposed model is at least robust to artifacts as using a frequency filter qualitatively. The quantitative analysis of the robustness against artifacts is critical but rarely studied, so we leave it for future work.

2) Statistical Significance Test: Following the conventions in previous EEG studies [12], [43], we perform the Wilcoxon Signed-Rank test to analyze the statistical significance of performance improvement of the proposed approaches. The detailed results are summarized in the supporting document. It is demonstrated that on the PhysioNet dataset, our proposed G-CRAM models outperform all the state-of-the-art and baseline methods significantly ($p < 0.05$). On the other hand, considering the BCICIV2a dataset, the performance improvement of the proposed models to the comparison methods is significant ($p < 0.05$) except for the baseline CRAM model. This result indicates that the graph representation does not significantly improve the performance of the proposed models on the BCICIV2a dataset ($p > 0.05$ when comparing CRAM with NG-CRAM, DG-CRAM, and SG-CRAM). The reason may be that the BCICIV2a dataset has fewer EEG nodes that the benefits imported by the graph embedding are limited. The difference between the three graph representation methods is also not significant ($p > 0.05$), suggesting that the exact distance between EEG nodes is not essential in current graph schemes. However, the distance-based method is more easily adaptive to various EEG nodes.

3) Effect of the Number of EEG Nodes: The proposed graph embedding strategy can be directly applied to EEG data with any number of EEG nodes. As the graph representation is location-based and the locations of EEG nodes are recognized internationally (such as 10-10 system), given an EEG headset, the coordinates (locations) of its EEG nodes would be fixed, and consequently, the graph representation could be achieved. Therefore, the proposed graph representation approach is adaptive to different amounts of EEG nodes.

The graph embedding aims at explicitly representing the spatial distribution of EEG nodes to help the following neural network extract powerful features. However, if there were only

a few EEG nodes (such as three or single nodes), the network would easily find the node correlations and extract useful features without the help of spatial embedding. In the evaluation, two datasets with different amounts of EEG nodes (PhysioNet 64 *vs* BCICIV2a 22) are used. The significance test results show that the graph representation has a significant impact on the dataset of 64 nodes ($p < 0.05$) but an insignificant impact on the dataset of 22 nodes ($p > 0.05$). In addition to the number of EEG nodes, the locations of EEG nodes are also critical to model performance. If EEG nodes were not placed on the active brain areas, even though there were lots of EEG nodes, the graph representation would not work.

4) Effect of Temporal Slice Size: The size of the temporal slice is an important hyper-parameter. In light of the evidence that the EEG signal presents multiple time scales, such as both local and global oscillations in time [43], [54], [55], we design the temporal slices for local temporal feature extraction. Then the embedded local temporal features are input into a recurrent attention module to obtain attentive global temporal features. A large or small size of the temporal slice would degrade the model performance. We carefully tuned the size of temporal slices and reported the best results.

V. Conclusion and Future Direction

A. Conclusion

This paper targets the EEG motor imagery classification task and proposes a novel deep learning approach. The deep learning model leverages an original graph representation to embed the spatial information, which is different from other spatial filtering methods, due to its independence of both subjects and tasks. A recurrent attention network is used to assign weights to different temporal cues instead of using a standard recurrent network to accumulate temporal information. Comprehensive experiments on two benchmark datasets show the superior performance of the proposed model on new subjects subjects, which is rarely reported in previous studies. Detailed insights of feature extraction and the impact of EEG nodes are also investigated by interpretation experiments and statistical significance tests. It is revealed that the EEG graph with a more considerable amount of nodes improves the overall performance more significantly.

B. Future Direction

There are several future research directions to further developing the proposed method. The first direction is to explore the G-CRAM on other BCI modalities, such as P300 and SSVEP. Due to its task-irrelevant scheme, the proposed graph representation can be directly applied to other BCI modalities. Meanwhile, the recurrent attention module would also benefit the extraction of the most discriminative temporal cues, such as that of P300. As most current works focus on a particular EEG task, the potential of G-CRAM being adaptive to different EEG tasks would be of great interest to researchers. The second research direction is to incorporate the proposed method into a real-world BCI and evaluate its online performance. The presented model can be efficiently incorporated into an online BCI with the off-the-shelf deep learning framework, like Tensorflow. Since only a few works have been reported to incorporate a deep learning model into a real-world BCI application, developing such an online system would be a remarkable contribution to the community. Taking power spectral features into consideration is also an exciting research opportunity. The power spectral is another widely used feature for EEG motor imagery classification. The proposed G-CRAM can be used to represent the power distribution over the scalp and find the distinguishable power changes over time.

References

[1] A. Berger, F. Horst, S. Müller, F. Steinberg, and M. Doppelmayr, "Current state and future prospects of eeg and fnirs in robot-assisted gait rehabilitation: A brief review," *Frontiers in Human Neuroscience*, vol. 13, pp. 172:1–17, 2019.

[2] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent eeg signal analysis," *IEEE Signal Processing Letters (SPL)*, vol. 26, no. 5, pp. 715–719, 2019.

[3] J. Van Erp, F. Lotte, and M. Tangermann, "Brain-computer interfaces: beyond medical applications," *Computer*, vol. 45, no. 4, pp. 26–34, 2012.

[4] H.-I. Suk and S.-W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35, no. 2, pp. 286–299, 2013.

[5] M. Hamedi, S.-H. Salleh, and A. M. Noor, "Electroencephalographic motor imagery brain connectivity analysis for bci: a review," *Neural computation*, vol. 28, no. 6, pp. 999–1041, 2016.

[6] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, "Subject-specific time-frequency selection for multi-class motor imagery-based bcis using few laplacian eeg channels," *Biomedical Signal Processing and Control*, vol. 38, pp. 302–311, 2017.

[7] Y. Kim, J. Ryu, K. K. Kim, C. C. Took, D. P. Mandic, and C. Park, "Motor imagery classification using mu and beta rhythms of eeg with strong uncorrelating transform based complex common spatial patterns," *Computational Intelligence and Neuroscience*, vol. 2016, pp. 1 489 692:1–13, 2016.

[8] D. J. McFarland and J. R. Wolpaw, "Sensorimotor rhythm-based brain-computer interface (bci): feature selection by regression improves performance," *IEEE Transactions on Neural Systems and Rehabilitation Engineering (TNSRE)*, vol. 13, no. 3, pp. 372–379, 2005.

[9] C. Ieracitano, N. Mammone, A. Bramanti, A. Hussain, and F. C. Morabito, "A convolutional neural network approach for classification of dementia stages based on 2d-spectral representation of eeg recordings," *Neurocomputing*, vol. 323, pp. 96–107, 2019.

[10] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on eeg data," *Pattern Recognition*, vol. 76, pp. 582–595, 2018.

[11] K. Chen, L. Yao, D. Zhang, X. Chang, G. Long, and S. Wang, "Distributionally robust semi-supervised learning for people-centric sensing," in *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, 2019, pp. 3321–3328.

[12] P. Zhang, X. Wang, W. Zhang, and J. Chen, "Learning spatial–spectral–temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering (TNSRE)*, vol. 27, no. 1, pp. 31–42, 2019.

[13] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: A compact convolutional network for eeg-based brain-computer interfaces," *Journa of Neural Engineering*, vol. 15, no. 15, pp. 056 013:1–17, 2018.

[14] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, and R. Boots, "Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface," in *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[15] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from eeg with deep recurrent-convolutional neural networks," in *International Conference on Learning Representation (ICLR)*, 2016, pp. 1–15.

[16] X. Zhu, P. Li, C. Li, D. Yao, R. Zhang, and P. Xu, "Separated channel convolutional neural network to realize the training free motor imagery bci systems," *Biomedical Signal Processing and Control*, vol. 49, pp. 396–403, 2019.

[17] M. Riyad, M. Khalil, and A. Adib, "Cross-subject eeg signal classification with deep neural networks applied to motor imagery," in *International Conference on Mobile, Secure, and Programmable Networking*. Springer, 2019, pp. 124–139.

[18] D. Zhang, K. Chen, D. Jian, L. Yao, S. Wang, and P. Li, "Learning attentional temporal cues of brainwaves with spatial embedding for motion intent detection," in *The 19th IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 1–6.

[19] I. Xygonakis, A. Athanasiou, N. Pandria, D. Kugiumtzis, and P. D. Bamidis, "Decoding motor imagery through common spatial pattern filters at the eeg source space," *Computational intelligence and neuroscience*, vol. 2018, pp. 1–10, 2018.

[20] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[21] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering (TBME)*, vol. 58, no. 2, pp. 355–362, 2010.

[22] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2008, pp. 2390–2397.

[23] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on bci competition iv datasets 2a and 2b," *Frontiers in neuroscience*, vol. 6, p. 39, 2012.

[24] A. Schlögl, F. Lee, H. Bischof, and G. Pfurtscheller, "Characterization of four-class motor imagery eeg data for the bci-competition 2005," *Journal of Neural Engineering*, vol. 2, no. 4, p. L14, 2005.

[25] D. Zhang, L. Yao, S. Wang, K. Chen, Z. Yang, and B. Benatallah, "Fuzzy integral optimization with deep q-network for eeg-based intention recognition," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, 2018, pp. 156–168.

[26] P. Herman, G. Prasad, T. M. McGinnity, and D. Coyle, "Comparative analysis of spectral approaches to feature extraction for eeg-based motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering (TNSRE)*, vol. 16, no. 4, pp. 317–326, 2008.

[27] A. Pérez-Zapata, A. F. Cardona-Escobar, J. A. Jaramillo-Garzón, and G. M. Díaz, "Deep convolutional neural networks and power spectral density features for motor imagery classification of eeg signals," in *International Conference on Augmented Cognition*. Springer, 2018, pp. 158–169.

[28] Y. Yi and M. Lin, "Human action recognition with graph-based multiple-instance learning," *Pattern Recognition*, vol. 53, pp. 148–162, 2016.

[29] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *The 23rd International Conference on Information and Knowledge Management (CIKM)*. ACM, 2014, pp. 659–668.

[30] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[31] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representation (ICLR)*, 2017, pp. 1–14.

[32] D. Zhang, L. Yao, K. Chen, S. Wang, P. D. Haghighi, and C. Sullivan, "A graph-based hierarchical attention model for movement intention detection from eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering (TNSRE)*, vol. 27, no. 11, pp. 2247–2253, 2019.

[33] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representation (ICLR)*, 2015, pp. 1–15.

[35] K. Chen, L. Yao, X. Wang, D. Zhang, T. Gu, Z. Yu, and Z. Yang, "Interpretable parallel recurrent neural networks with convolutional attentions for multi-modality activity modeling," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[36] K. Chen, L. Yao, D. Zhang, B. Guo, and Z. Yu, "Multi-agent attentional activity recognition," in *The 28th International Joint Conferences on Artificial Intelligence (IJCAI)*, 2019, pp. 4031–4038.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.

[38] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *International Conference on Learning Representation (ICLR)*, 2018, pp. 1–12.

[39] D. Zhang, L. Yao, K. Chen, and S. Wang, "Ready for use: Subject-independent movement intention recognition via a convolutional attention model," in *The 27th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM, 2018, pp. 1763–1766.

[40] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2016, pp. 2249–2255.

[41] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, 2007.

[42] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[43] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[44] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[45] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set a," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[46] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE Transactions on Cybernetics (TCYB), Early Access*, pp. 1–12, 2019.

[47] V. P. Oikonomou, K. Georgiadis, G. Liaros, S. Nikolopoulos, and I. Kompatsiaris, "A comparison study on eeg signal processing techniques using motor imagery eeg data," in *2017 IEEE 30th international symposium on computer-based medical systems (CBMS)*. IEEE, 2017, pp. 781–786.

[48] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2016, pp. 606–615.

[49] J. Shin, J. Kwon, and C.-H. Im, "A ternary hybrid eeg-nirs brain-computer interface for the classification of brain activation patterns during mental arithmetic, motor imagery, and idle state," *Frontiers in Neuroinformatics*, vol. 12, pp. 5:1–9, 2018.

[50] B. Shrestha, I. Vlachos, J. A. Adkinson, and L. Iasemidis, "Distinguishing motor imagery from motor movement using phase locking value and eigenvector centrality," in *32nd Southern Biomedical Engineering Conference*. IEEE, 2016, pp. 107–108.

[51] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. L. Da Silva, "Mu rhythm (de) synchronization and eeg single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, 2006.

[52] A. Ghaemi, E. Rashedi, A. M. Pourrahimi, M. Kamandar, and F. Rahdari, "Automatic channel selection in eeg signals for classification of left or right hand movement in brain computer interfaces using improved binary gravitation search algorithm," *Biomedical Signal Processing and Control*, vol. 33, pp. 109–118, 2017.

[53] H. Shan, H. Xu, S. Zhu, and B. He, "A novel channel selection method for optimal classification in different motor imagery bci paradigms," *BioMedical Engineering Online*, vol. 14, no. 1, pp. 93:1–18, 2015.

[54] S. Monto, S. Palva, J. Voipio, and J. M. Palva, "Very slow eeg fluctuations predict the dynamics of stimulus detection and oscillation amplitudes in humans," *Journal of Neuroscience*, vol. 28, no. 33, pp. 8268–8272, 2008.

[55] S. Vanhatalo, J. M. Palva, M. Holmes, J. Miller, J. Voipio, and K. Kaila, "Infraslow oscillations modulate excitability and interictal epileptic activity in the human cortex during sleep," *Proceedings of the National Academy of Sciences*, vol. 101, no. 14, pp. 5053–5057, 2004.