An Attentional-LSTM for Improved Classification of Brain Activities Evoked by Images

Sheng-hua Zhong*†
Ahmed Fares*‡
csshzhong@szu.edu.cn
ahmd.fares@szu.edu.cn
Institute for Future Media Computing
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, China

Jianmin Jiang§
Institute for Future Media Computing
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, China
jianmin.jiang@szu.edu.cn

ABSTRACT

Multimedia stimulation of brain activities is not only becoming an emerging area for intensive research, but also achieved significant progresses towards classification of brain activities and interpretation of brain understanding of multimedia content. To exploit the characteristics of EEG signals in capturing human brain activities, we propose a region-dependent and attention-driven bi-directional LSTM network (RA-BiLSTM) for image evoked brain activity classification. Inspired by the hemispheric lateralization of human brains, the proposed RA-BiLSTM extracts additional information at regional level to strengthen and emphasize the differences between two hemispheres. In addition, we propose a new attentional-LSTM by adding an extra attention gate to: (i) measure and seize the importance of channel-based spatial information, and (ii) support the proposed RA-BiLSTM to capture the dynamic correlations hidden from both the past and the future in the current state across EEG sequences. Extensive experiments are carried out and the results demonstrate that our proposed RA-BiLSTM not only achieves effective classification of brain activities on evoked image categories, but also significantly outperforms the existing state of the arts.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Neural networks**.

KEYWORDS

EEG, brain activities classification, region-level information, attention-driven LSTM, bi-directional computational model.

ACM Reference Format:
Sheng-hua Zhong, Ahmed Fares, and Jianmin Jiang. 2019. An Attentional-LSTM for Improved Classification of Brain Activities Evoked by Images. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350886

1 INTRODUCTION

Over the past decades, analyzing EEGs evoked by specific stimuli patterns has been widely researched across a number of areas, including neuroscience, artificial intelligence, neural computation etc. The primary efforts, however, are focused on (i) designing of various experimental paradigm based on multimedia materials, including images, sounds, texts etc., as stimuli to activate human brains and expose their cognitive activities for computerized recognition and classification, such as human face with clean background etc. [9]. (ii) artificial intelligence algorithm development, particularly deep learning models, for content understanding, pattern recognition and classifications towards intelligent interpretation of human brains or brain intelligence [12, 15, 17, 31]. While most of these methods researched and reported in the literature follow a similar roadmap, where the original EEG sequences of the extracted time-frequency features are used as the input and all sorts of machine learning models then follow, the unique characteristics of human brains are not sufficiently explored, especially among the extensive research across brain sciences, such as hemispheric lateralization etc.

While the macrostructure of the left and right hemispheres of human brains appears to be similar, research in brain sciences indicate that the level of sensitivities to the differences in stimuli patterns remains variable, and structural designs of neural networks can benefit from individually specialized function in each hemisphere [1]. Hemispheric lateralization

*Sheng-hua Zhong and Ahmed Fares are joint first authors.
†Sheng-hua Zhong is also with the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China.
‡Ahmed Fares is also with the Department of Electrical Engineering, the Computer Engineering branch, Faculty of Engineering at Shoubra, Benha University, Cairo.
§Jianmin Jiang, the corresponding author.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350886

refers to the tendency for some neural functions or cognitive processes to be specialized to the right or left hemispheres of the brain. Although a growing body of evidences have suggested that cognitive tasks in humans rely on a number of related processes whose neural loci are largely lateralized to one hemisphere or the other [34], most of current researches focus on studying the lateralization in different tasks [5, 6, 23], or developing better tools and models for assessing lateralization [7, 34]. To the best of our knowledge, no research has been attempted to integrate the hemispheric lateralization into the deep learning model to extract the region-level information from brain signals.

Our extensive literature survey indicates that brain activity analysis other than EEGs, such as fMRI-based etc., could give us new inspirations for further improvement. Considering the fact that EEGs are generated by a number of different channels, and each individual channel is located in a particular spatial position on the cap, their locations reflect the characteristics of the brain responses across the EEG sequences, and the time flow hides the dynamic correlation of the brain activations. While deep learning models have been reported to achieve performance improvement for EEG-based object classification [2, 4, 12, 15, 31], most of these models have not considered jointly both the spatial and dynamic correlations embedded inside the EEG data sequences. In addition, these models only consider each channel as a independent flow, and thus the spatial information and the correlations across different channels are basically ignored. On the other hand, EEG-based brain activity classifications evoked by images are widely known as high-level cognitive tasks, and the electrical activations from both the past and the future provide important dynamics and correlations for the current spontaneous responses and the state of the test subjects. To this end, it becomes desirable to consider these attributes and factors in developing next generation deep learning models for brain activity analysis and understanding.

In addition, the attention mechanism, which allows a deep network to pay attention to only part of the input information, becomes one of the most powerful and influential ideas in deep learning. Xu et al. proposed a deep learning model with attention mechanism for image captioning [39]. As EEGs are complex and channel-based temporal-spatial signal sequences, some parts of human brains are more deeply involved than others, leading to the oscillations in EEG signals and further spaces for further research and improvement. To tackle these challenges, we propose a bi-directional and attention driven deep network in this paper for classifications of brain activities. In comparison with the existing state-of-the-arts, our proposed model achieves a number of advantages and novelties, which can be highlighted as: (i) inspired by the phenomenon of hemispheric lateralization of human brains, we introduce a new concept of region-level computation into the deep learning framework to strengthen and emphasize the differences between two hemispheres with low dimension; (ii) By selectively focusing on parts of input channels which are useful in classifying EEG signals, we are the first to propose a new attentional-LSTM, to integrate the attention mechanism in our framework to measure and seize the importance of different EEG channels, and hence propose a RA-BiLSTM (region-dependent, and bi-directional attention-driven LSTM) deep learning framework to achieve significantly improved performances for classifications of brain activities evoked by natural images.; (iii) we propose a new deep brain analytics method to capture the dynamic correlations hidden from both the past and the future to the current state in EEG sequences; and finally (iv) we carry out extensive experiments to support that our deep framework achieves superior performances in comparison with the existing state-of-the-arts.

The rest of the paper is organized as follows. In section 2, we present a literature survey about the existing methods that use deep learning models for EEG-based image classifications. In Section 3, we describe the details of our proposed region-dependent, and bi-directional attention-driven deep network for EEG-based visual object classification. In Section 4, we report our extensive experimental results and validate the superiority and effectiveness of our proposed framework, compared with the existing state of the arts, and finally Section 5 provides concluding remarks and future work.

2 RELATED WORK

At present, all the existing research on EEG-based analysis can be summarized in two steps: feature extraction and pattern recognition or machine learning-based methods to complete the signal analysis[13, 22]. With the extensive application and in-depth promotion of deep learning, an ever-increasing number of brain science and neuroscience research teams are exploiting its strength in designing algorithms to achieve intelligent understanding and analysis of brain activities via EEGs, leading to an end-to-end model by integrating feature extraction and classification/clustering.

Jiao et al. [18] proposed a multi-channel deep convolution network to classify mental loads. Wang et al. [37] used LSTM network to classify motor imagery tasks, and used a one-dimensional aggregation approximation method to extract the network's effective features. Cole et al. [8] used a predictive modelling approach based on CNN for predicting brain ages. Their analysis showed that the brain-predicted age is highly reliable. Gao et al. [14] proposed a spatiotemporal deep convolution model, which significantly improved the accuracy of detecting driver fatigue by emphasizing the importance of spatial information and time dependence of EEGs. Yuan et al. [41] proposed an end-to-end multi-view deep learning framework to automatically detect epileptic seizures in EEG signals. Li et al. [26] tried to incorporate transfer learning into the construction of convolutional neural networks and successfully applied the model to the clinical diagnosis of mild depression. Dong et al. [12] used a rectified linear unit (ReLU) activation function and a mixed neural network of LSTM on time-frequency-domain features to classify sleep stages. Lawhern et al. [24] proposed a compact full convolutional network as the EEG-specific model (EEGNet) and applied it to four different brain-machine interface classification tasks.

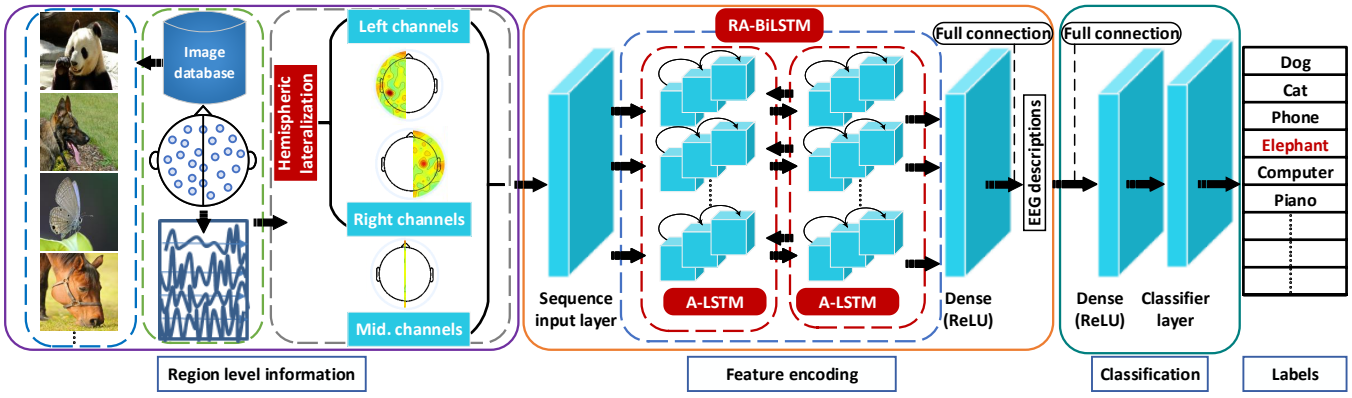

Figure 1: Structural illustration of the proposed deep framework.

Zhang et al. [42] proposed a cascaded and parallel convolution recurrent neural network model to accurately identify human expected motion instructions by effectively learning the spatiotemporal representation of the original EEG signal. Tan et al. [33] converted EEG data into EEG-based video and optical flow information, classified them by convolution neural network and recurrent neural network (RNN), and established an effective rehabilitation support system based on BCI.

Multimedia data, which contain a large amount of content information and rich visual characteristics, are considered to be a very suitable stimuli material and widely used in the acquisition and analysis of EEG signals [21, 30, 31]. Researchers tried to identify and classify the content information of multimedia data viewed by users through the analysis of EEG signals [9, 28, 36]. Spampinato et al. [31], used LSTM network to learn an EEG data representation based on image stimuli and constructed a mapping relationship from natural image features to EEG representation. Finally, they used the new representation of EEG signals for classification of natural images. Compared with traditional methods, these deep learning-based approaches have achieved outstanding classification results. Recent studies have shown that it is possible to reconstruct multimedia content information itself by mining EEG data. Kavasidis et al. [21] proposed a method for reconstructing visual stimuli content information through EEGs. By using a variable-valued autoencoder (VAE) and generative adversarial networks (GANs), they found that EEG data contain patterns related to visual content, and the content can be used to generate images that are semantically consistent with the input visual stimuli. While these methods have demonstrated the capability of using deep learning framework for EEG-based image classification, the original EEG data or the extracted time-frequency features based on signal analysis algorithms are often used as the input, and some characteristics of human brains have not been seriously considered, such as hemispheric lateralization, the spatial and dynamic correlations in the EEG data sequences have not been jointly considered, and the classification accuracy achieved to date by Spampinato et al. was 82.9% [31], leaving significant space for further research and improvement.

3 METHODOLOGY

Given the extensive survey on existing research in the previous section, we propose a region-dependent and bi-directional attention-driven LSTM framework for automated classifications of brain activities evoked by natural images. Specifically, our approach consists of three stages, i.e., the region-level information extraction stage, the feature encoding stage, and the classification stage. A structural illustration is given in Fig. 1.

3.1 The Lateralization Effect

Although extensive research in neuroscience has revealed that, in some cognitive processes, the neural loci are largely lateralized to one hemisphere or the other, no existing research for EEG-based classification of brain activities evoked by images has ever attempted to exploit this concept during the development of deep learning models. For channel i, the raw EEG signal denoted as $\mathbf{E} = [\mathbf{e}_i]_{i=1}^{l_{ch}}$ is regarded as input to the region-level information extraction, where $i \in [1,\ l_{ch} = 128]$ is the index for channels, and l_{ch} is the number of channels. To achieve the desired lateralization effect, we further split the EEG data into three groups, including the left hemisphere, the right hemisphere, and the middle part. By denoting the left hemisphere group, the right hemisphere group, and the middle group as, $\mathbf{E}^{[l]}$, $\mathbf{E}^{[r]}$, and $\mathbf{E}^{[m]}$, respectively, each channel $\mathbf{e}_i$ can be linked to one group based on the corresponding electrode physical location. Each channel in the left hemisphere group has a corresponding channel in the right hemisphere group, and as a result, the difference, $\mathbf{d}_j$, can be calculated according to the following equation:

$$\mathbf{d}_j = \mathbf{E}_j^{[l]} - \mathbf{E}_j^{[r]} \tag{1}$$

where $\left(\mathbf{E}_j^{[l]},\ \mathbf{E}_j^{[r]}\right)$ is regarded as the corresponding pair. $j \in [1,\ l_g]$ is the index for the left hemisphere group, the right hemisphere group, and the difference, and l_g is the number of channels linked to the left hemisphere group or

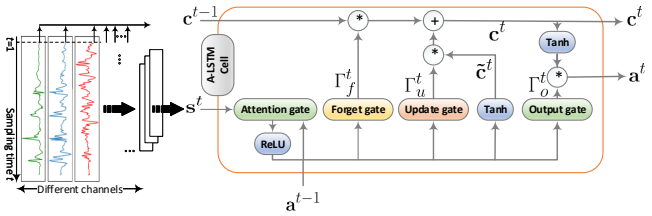

Figure 2: Structural illustration of the proposed A-LSTM cell.

the right hemisphere group. The output of the region-level information extraction stage is obtained when the difference, $\mathbf{D} = [\mathbf{d}_j]_{j=1}^{l_g}$, is combined with the middle group, $\mathbf{E}^{[m]}$, into one variable, $\mathbf{S}$, and then this output is passed to the sequence layer as an input according to the following equation:

$$\mathbf{S} = \left[\mathbf{D}^{\mathrm{T}}\ \mathbf{E}^{[m]^{\mathrm{T}}}\right] \tag{2}$$

3.2 The Proposed Attentional-LSTM and RA-BiLSTM

The attention mechanism was firstly proposed by Bahdanau et al. in machine translation [3], which is utilized to select the reference words in sentences. Its further applications include syntactic constituency parsing by Vinyals et al. [35], natural language question answering by Sukhbaatar et al. [32], and image question answering by Yang et al. [40]. The concept of "attention" has obtained popularity in training neural networks, and it can work collaboratively with different modalities, e.g., the attention mechanism is used to select the relevant image regions when generating words in the captions by Xu et al. [39]. Recent progress on self-attention mechanism is represented by computing vector-space descriptions and characterizations for both input and output, and improved results have been reported [11].

To reward those channels that provide more clues for correct classification and analysis of the EEG sequences, we introduce a new attention-driven LSTM for EEG-based classification of brain activities evoked by natural images. In details, we propose a channel-level attention gate and use this gate to measure the level of importance for each channel and hence optimize their collective contributions for brain activity analysis.

The feature encoding stage extracts the EEG descriptions from the region-level information via RA-BiLSTM network. The RA-BiLSTM learns long-term dependencies across different timing steps of the sequence data. It not only solves the vanishing gradient problem, which appears in recurrent neural network (RNN) through the forget gate and the update gate, but also measures and seizes the importance of the information from different channels through the attention gate and catches the dynamic correlations inside the EEG sequences. In contrast to the unidirectional LSTM, the RA-BiLSTM calculates the output $\mathbf{y}^t$ at any time t by taking information from both the earlier and the later states inside the sequences.

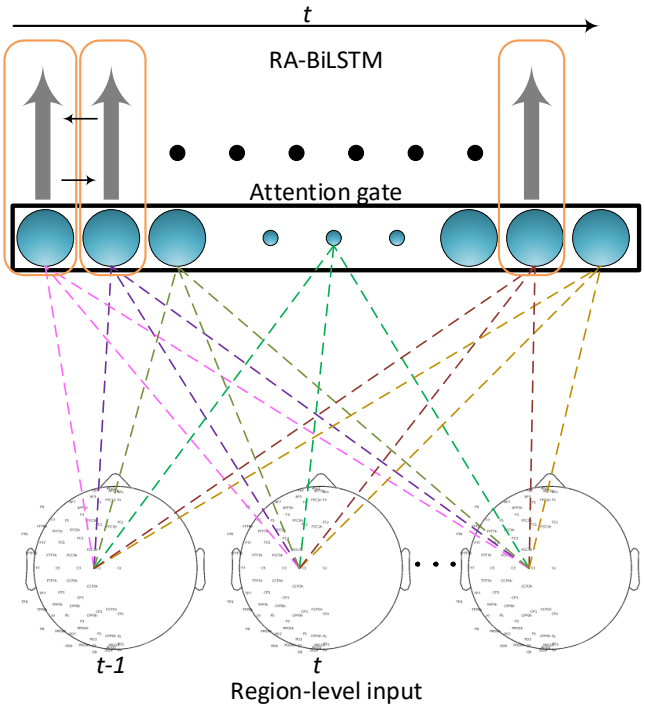

Figure 3: Structural illustration of the proposed soft attention gate, where the input is $\mathbf{s}^t$, and the connection with the same color means that they share the same parameters.

The structure of the layer in RA-BiLSTM, including a forward A-LSTM layer and a backward A-LSTM layer, is illustrated in Fig. 1. As seen, the forward layer output sequence, $\overrightarrow{\mathbf{a}}^t$, is iteratively calculated using the inputs in a sequence from time 1 to time $t-1$, while the backward layer output sequence, $\overleftarrow{\mathbf{a}}^t$, is calculated using the inputs from the end of sequence to time $t+1$. Specifically, given the input $\mathbf{s}$ from all channels at time t, the attention gate Γ_a^t, the update gate Γ_u^t, the forget gate Γ_f^t, and the output gate Γ_o^t, which are represented by colorful boxes in the A-LSTM cell in Fig. 2, can be calculated from the the region-level information $\mathbf{S} = [\mathbf{s}^t]_{t=1}^{l_s}$, where l_s is the length of the sequence, and the previous layer output $\mathbf{a}^{t-1}$ according to the following equation:

$$\begin{pmatrix} \Gamma_a^t \\ \Gamma_f^t \\ \Gamma_u^t \\ \Gamma_o^t \end{pmatrix} = g \begin{pmatrix} \mathbf{W}_a & \mathbf{U}_a & 0 \\ 0 & \mathbf{U}_f & \mathbf{W}_f \\ 0 & \mathbf{U}_u & \mathbf{W}_u \\ 0 & \mathbf{U}_o & \mathbf{W}_o \end{pmatrix} \begin{pmatrix} \mathbf{s}^t \\ \mathbf{a}^{t-1} \\ \Gamma_a^t \end{pmatrix} + g \begin{pmatrix} \mathbf{b}_a \\ \mathbf{b}_f \\ \mathbf{b}_u \\ \mathbf{b}_o \end{pmatrix} \tag{3}$$

where, for $k \in \{a, f, u, o\}$, $\mathbf{W}_k$ is the weight matrix mapping the layer input to the four gates, $\mathbf{U}_k$ is the weight matrix connecting the previous cell output state to the four gates, and $\mathbf{b}_k$ is the bias vector. The function $g()$ is designed as ReLU activation function for Γ_a^t and element-wise sigmoid for Γ_f^t, Γ_u^t, and Γ_o^t, respectively, and the state of the attention gate Γ_a^t is fed through the three gates.

To optimize the process of adding the attention-driven mechanism, we propose two different versions of attention mechanism, soft attention gate and hard attention gate. For the soft attention gate, as shown in Fig. 3, the input EEG signals of different channels are fully connected with the nodes in the gate. As a result, the size of $\mathbf{W}$ depends on the number of channels and the number of nodes in the attention gate. For the hard attention gate, we prune the connections in attention gate based on the physical locations of the electrodes. In other words, the input channels are split into m nearest-neighbor groups according to their physical locations, and the channels in each group are fully connected with a separate node in the attention gate. In this way, $\mathbf{W}_a$ in (3) will be changed to $\boldsymbol{\alpha}\ \mathbf{W}_a$ as given in the following equation (4), where $\boldsymbol{\alpha}$ is the pruning parameter. Alpha is 1 if there exists a connection between the channel and their corresponding node in the attention gate. Otherwise, it is 0.

$$\mathbf{x}^t = g(\mathbf{s}^t\ \boldsymbol{\alpha}\ \mathbf{W}_a + \mathbf{a}^{t-1}\ \mathbf{U}_a + \mathbf{b}_a) \tag{4}$$

Based on the results of (3), the cell output state $\mathbf{c}^t$ and the layer output $\mathbf{a}^t$ (both the forward and the backward outputs) can be calculated from the state of the attention gate Γ_a^t and the previous layer output $\mathbf{a}^{t-1}$, details of which are given below:

$$\mathbf{c}^t = \Gamma_f^t * \mathbf{c}^{t-1} + \Gamma_u^t * (\tanh(\mathbf{U}_c\mathbf{a}^{t-1} + \mathbf{W}_c\Gamma_a^t + \mathbf{b}_c)) \tag{5}$$

$$\mathbf{a}^t = \Gamma_o^t * \tanh(\mathbf{c}^t) \tag{6}$$

where $\mathbf{W}_c$ is the weight matrix mapping the layer input to the candidate for replacing the memory cell. While $\mathbf{U}_c$ is the weight matrix connecting the previous cell output state to the candidate for replacing the memory cell, $\mathbf{b}_c$ is the bias vector. The function tanh() indicates a hyperbolic tangent.

The final output of a RA-BiLSTM layer is a vector of all outputs, represented by $\mathbf{Y} = \left[\mathbf{y}^t\right]_{t=1}^{l_s}$. At each time of iteration t, $\mathbf{y}^t$ can be calculated according to (7). Taking the EEG-based object classification problem as an example, only the last element of the output vector, $\mathbf{y}^{l_s}$, is taking into account when making the predictions.

$$\mathbf{y}^t = \sigma_y(\mathbf{W}_y[\overrightarrow{\mathbf{a}}^t, \overleftarrow{\mathbf{a}}^t] + \mathbf{b}_y) \tag{7}$$

where $\mathbf{W}_y$ is the weight matrix from the RA-BiLSTM hidden layer to the output layer, $\mathbf{b}_y$ is the bias vector of the output layer and $\sigma_y()$ is the sigmoid activation function of the output layer.

The classification stage takes the EEG descriptions from the RA-BiLSTM network as an input and returns the prediction as an output, where a softmax classifier is adopted.

4 EXPERIMENTS

To evaluate our proposed RA-BiLSTM, we have carried out extensive experiments which are arranged in three stages. In the first stage, the EEG-based classification performance of our proposed deep learning framework is assessed. In the second stage, we visualize the weights in attention gate and analyze the contribution from different channels and regions. In the third stage, we study the relationships between the neural activations and the specific emotional states.

Table 1: The classification performance comparisons among our proposed RA-BiLSTM, the RNN-based method, siamese network, and the RS-LDA.

Models	Accuracy
Proposed RA-BiLSTM	98.4%
RNN-based model [31]	82.9%
Siamese network [29]	93.7%
RS-LDA [20]	13.0%

4.1 Experimental Settings

For the convenience of comparatively analyzing the experimental results against the existing efforts, we adopt ImageNet-EEG, which is a publicly available EEG dataset for brain imaging classification proposed by Spampinato et al. [31]. For benchmarking purposes, the proposed framework is compared with the EEG-based object classification methods [29, 31], which are the most recent deep learning methods on the same dataset and the baseline method: representational similarity based linear discriminant analysis (RS-LDA) [20].

For our method, concerning parameters for the attentional-LSTM for the hard attention gate, we split the input channels to $m = 17$ nearest-neighbor groups according to the physical locations, each group contains 4 channels, for the soft attention gate, we have assigned the number of nodes to 68. The iteration limit is set to 2500, and the batch size is set to 440 for the feature encoding stage of the RA-BiLSTM.

4.2 Image-Stimulated Brain Activity Classification

In the first stage of experiments, the effectiveness of our RA-BiLSTM deep network is validated for EEG-based object classification. All the experimental setting follows that of the existing work [31].

Table 1 summarizes the experimental results in terms of the classification precisions for our proposed RA-BiLSTM deep network, the most recent deep learning methods, including the state-of-the-art RNN-based method [31] and siamese network [29], and the RS-LDA method [20]. As seen, while the precision rate accomplished by our proposed RA-BiLSTM deep network is 98.4%, the existing state-of-the-art, siamese network, and the RS-LDA compared are 82.9%, 93.7%, and 13.0%, respectively.

To quantify and analyze the contribution of each stage designed in our proposed RA-BiLSTM deep network, further experiments are conducted to explore the effectiveness of different configurations made by individual stages. At the feature encoding stage, individual elements considered include: (i) selection of different channel-based attention gate structures, including soft attention gate and the hard attention gate; (ii) choice of different feature-encoding techniques, including unidirectional LSTM, bi-directional LSTM, and bi-directional attentional-LSTM (A-BiLSTM). In the hard attention gate, two designs have been tested, including the

Table 2: Comparative assessment of the proposed framework upon different configurations.

	Framework				
Configurations	1	2	3	4	5
Lateralization effect	✓	✓	✓	✓	✓
Soft attention gate			✓		
Hard attention gate w/ shared parameters				✓	
Hard attention gate w/o shared parameters					✓
LSTM	✓				
BiLSTM		✓			
A-BiLSTM			✓	✓	✓
Accuracy %	92.9	97.0	98.4	98.0	97.9

hard attention gate without sharing the learnable parameters and the hard attention gate with sharing the learnable parameters. While the hard attention gate without sharing the learnable parameters utilizes a separate network for each group of channels, the hard attention gate sharing the learnable parameters utilizes the same network for all groups.

Table 2 summarizes the experimental results in classification precision rates with all the various configurations, from which we can notice and reach a number of conclusions that can be described as follows:

First, the performance of utilizing the lateralization effect plus the attention gate is always better than utilizing the lateralization effect alone at the stage of the region-level information extraction. These results are presented by configurations $2-5$ in Table 2. If the bi-directional LSTM is chosen as the feature encoder, the best performance accomplished by the lateralization effect alone is 97.0% by configuration 2, and the best performance accomplished by combining both the lateralization effect and the attention gate (RA-BiLSTM) is 98.4% by configuration 3.

Second, the performance of utilizing the soft attention gate is always better than utilizing the hard attention gate at the feature extraction stage in A-LSTM. If the A-BiLSTM is chosen as the feature encoder, the best performance accomplished by the hard attention gate is 98.0% by configuration 5, and the best performance accomplished by the soft attention gate is 98.4% configuration 3.

Third, the performance of utilizing the hard attention gate in sharing the learnable parameters is better than utilizing the hard attention gate without sharing the learnable parameters at the feature extraction stage in A-LSTM. If the A-BiLSTM is chosen as the feature encoder, the best performance accomplished by the hard attention gate without sharing the learnable parameters is 97.9% by configuration 5, and the best performance accomplished by the hard attention gate with sharing of the learnable parameters is 98.0% by configuration 4.

Fourth, the performance of bi-directional attentional-LSTM is always better than that of the unidirectional LSTM as the feature encoder in the feature extraction stage. These results are presented by configurations 1 and 3 in Table 2. While the best performance accomplished by the unidirectional LSTM is 92.9% by configuration 1, the best performance accomplished by the bi-directional attentional-LSTM is 94.8% by configuration 3.

Fifth, the performance of bi-directional attentional-LSTM (A-BiLSTM) is always better than that of the bi-directional LSTM (BiLSTM) as the feature encoder in the feature extraction stage. These results are presented by configurations 2 and 3 in Table 2. While the best performance accomplished by the BiLSTM is 97.0% by configuration 1, the best performance accomplished by the A-BiLSTM is 94.8% by configuration 3.

4.3 Visualization of the Attention Weights

To further enhance the revelation of the roles played by attention weights, we in this section visualize the connection weights of the attention gate and analyze the contribution of different channels and regions. In Fig. 4(a), we provide the visualization of the connection weights between the input channels and the nodes in soft attention gate. From this figure, we can see that some of channels are more "active" than others. It means the weights of these channels are far away from 0. It also means these channels play more important roles in the feature extraction and the final classification process. In Fig. 4(b), we give the sum of the absolute value of the connection weights from each channel, and we pick out the first ten channels with the largest values and show the electrode physical location of these channels as red circles in Fig. 4(c). The larger the circle is, the greater the weight is, and thus the value of the weight can be used to evaluate the importance of the channel. The top ten channels are P1, CCP3h, FC3, PPO5h, CP3, C1, FFT9h, F1, PO9, and FCC1h.

Similarly, we can use the same method to analyze the connection weights of the attention gate in hard attention gates. As described before, we split the input channels to nearest-neighbor groups according to the physical locations in hard attention gate, and the channels in each group are fully connected with a separate node in attention gate. Thus, our analysis is relied on each group. We show the first four groups with the largest values of connection weights in Fig. 4(d).

As seen, the most important locations selected by soft attention gate and hard attention gate are not identical. But

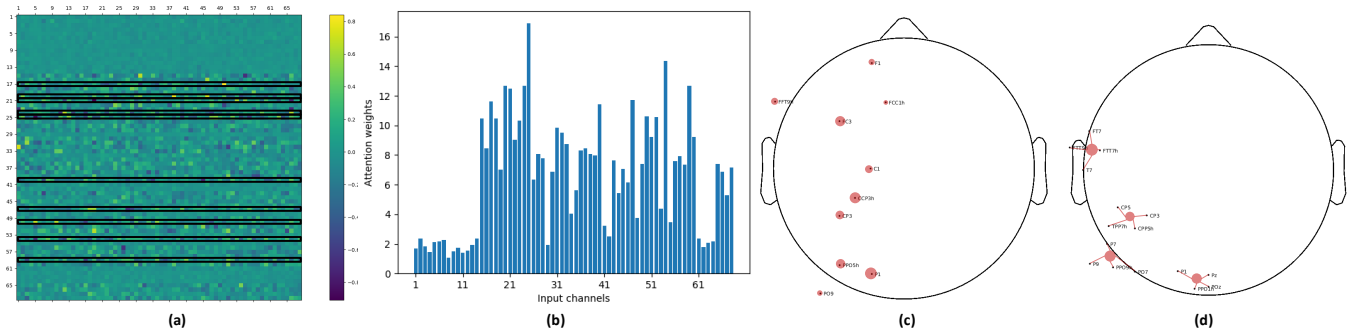

Figure 4: (a) Attention weights map. (b) The sum of the absolute value of the connection weights from each channel. (c) Electrodes physical location of the top ten channels. (d) Electrodes physical location of the top four regions.

we can still find that some channels are selected out by both of them. For example, the absolute value of the weight in channel CP3 is large, thus it is popped out from soft attention gate, and it is also included in the four most important region groups obtained by using the hard attention gate. The similar situation happened in P1. The object recognition is a complex task and involves several different areas of the brain. In principle, the frontal lobe is involved in memory encoding during incidental learning and then later maintaining and retrieving semantic memories [38]. In the soft attention gate, three channels are selected in this region. According to the timecourse analysis of electrophysiological correlates of object recognition [19], there are two distinct types of components in the event-related potential recorded during the categorization of natural images, examples of which indicate that the first peak is a frontal positivity, and the second peak is a central positivity, and this second peak could be some kind of top-down mechanism in object recognition. From the selected channels of the soft attention gate and the region groups of the hard attention gate, we can find that central electrodes are selected out. These theoretically prove that, as a machine learning structure, attention gate is capable of grasping important information of human brains in the process of object recognition.

After obtaining the importance of different channels based on the visualization results of the attention gate, we seek to explore the influence of different number of channels upon the classification accuracy. The results are shown in Fig. 5. Our original setting is referred to as RA-BiLSTM. The setting with the ten most important channels is referred to as RA-BiLSTM-10, the setting with the 34 most important channels is referred to as RA-BiLSTM-34, whilst the setting with the four most important regions is referred to as RA-BiLSTM-4R. As seen in Fig. 5, the precision rate accomplished by RA-BiLSTM, RA-BiLSTM-10, RA-BiLSTM-34, and RA-BiLSTM-4R are 98.4%, 94.9%, 97.2%, and 93.2%, respectively. These results tell us that the proposed attention gate has a good potential for feature selection. Even if we remove some unimportant channels based on weights, it will not have a big impact on the final results.

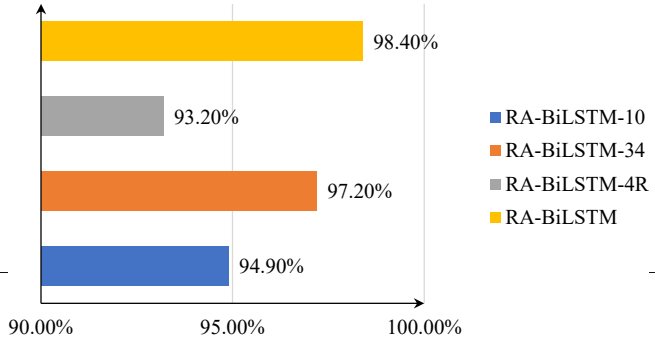

Figure 5: Comparative assessment of the proposed framework upon different configurations of the attention gate.

4.4 Case Study of The Neural Activations and Emotions

Unlike most of the existing EEG datasets that only incorporate less than 10 classes, ImageNet-EEG [31] contains 40 classes, and most of them are regular objects or animals. Thus, for this dataset, we are not satisfied with just introducing the novel deep learning framework with good classification results. Further, we attempt to examine the connections between the neural activations and the specific emotional states. The EEG data in ImageNet-EEG just contains the Beta and Gamma bands. From the existing work, the emotional processing enhanced Gamma band powers at frontal area as compared to processing neutral pictures [27], and the signal from Gamma band is suitable for EEG-based emotion classification [25]. Thus, all experiments provided here are focused on the signals from Gamma band.

This dataset includes 40 categories, including "dog", "cat", "butterfly", "sorrel", "capuchin", "elephant", "panda", "fish", "airliner", "broom", "canoe", "phone", "mug", "convertible", "computer", "watch", "guitar", "locomotive", "espresso", "chair", "golf", "piano", "iron", "jack", "mailbag", "missile",

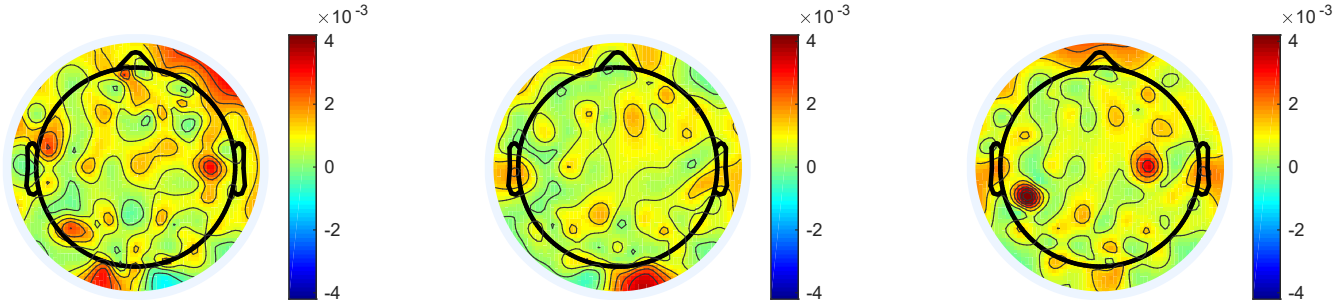

Figure 6: Scalp distribution of the average energy at Gamma frequency sub-band for all participants and sessions of the three categories: "gun", "phone", and "panda".

"mitten", "bike", "tent", "pajama", "parachute", "pool", "radio", "camera", "gun", "shoe", "banana", "pizza", "daisy", and "bolete" (fungus). Each category contains 50 images with 300 EEG signals for the six subjects. In the classes of the dataset, the class "gun" is a category that could cause negative emotions. Most of other classes are thought as typically neutral, such as "phone", "watch", "bike", "shoe". Fig. 6 demonstrates the average energy distribution in Gamma band of the "gun", "phone", and "panda" categories. From this figure we can see that an increase of the average relative energy in the prefrontal area during the period of the images from the category "gun" is observed as compared to that from the categories of "phone" and "panda". These results are consistent with the discoveries reported by [10, 16], which show that neural signatures associated with positive, neutral and negative emotions do exist.

5 CONCLUSIONS

Following recent efforts via directly using multimedia to stimulate brain activities towards brain image classification, we propose in this paper a region-dependent and attention-driven bi-directional LSTM deep learning approach for EEG-based classification of brain activities evoked by natural images. Our proposed framework provides an improved solution for the problem that, given an image used to stimulate brain activities, we should be able to identify which class the stimuli image comes from by analyzing the EEG signals. The region-level information is extracted to preserve and emphasize the hemispheric lateralization for neural functions or cognitive processes of human brains. In addition, a channel-level attention mechanism is integrated into our new framework to measure and seize the importance of different EEG channels, and a RA-BiLSTM is used to capture the dynamic correlations hidden in the EEG sequences. Extensive experiments on ImageNet-EEG, the most challenging EEG dataset for brain activity classifications, validate that our framework outperforms the existing state-of-the-arts under various contexts and experimental set ups. Further, our research has produced substantial evidences to support that data estimated straightforwardly from human minds could enable machine learning models to make better and more human-like understandings.

Finally, further research can be identified as: (i) applying our deep learning framework for other EEG-based content understanding or pattern analysis tasks; (ii) reconstructing the multimedia content information through the proposed EEG representations.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61620106008), the Natural Science Foundation of Guangdong Province (No. 2016A030310053), the Shenzhen Emerging Industries of the Strategic Basic Research Project under Grant (No. JCYJ20160226191842793), the Shenzhen high-level overseas talents program, and the Inlife-Handnet Open Fund.

REFERENCES

[1] Nawfal Al-Hadithi, Ahmed Al-Imam, Manolia Irfan, Mohammed Khalaf, and Sara Al-Khafaji. 2016. The relation between cerebral dominance and visual analytic skills in Iraqi medical students, a cross sectional analysis. *Asian Journal of Medical Sciences* 7, 6 (Oct. 2016), 47–52. https://doi.org/10.3126/ajms.v7i6.15205

[2] Andreas Antoniades, Loukianos Spyrou, David Martin-Lopez, Antonio Valentin, Gonzalo Alarcon, Saeid Sanei, and Clive Cheong Took. 2017. Detection of Interictal Discharges with Convolutional Neural Networks Using Discrete Ordered Multichannel Intracranial EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 12 (2017), 2285–2294.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014). arXiv:1409.0473 http://arxiv.org/abs/1409.0473

[4] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. 2015. Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448* (2015).

[5] Mark F Bear, Barry W Connors, and Michael A Paradiso. 2007. *Neuroscience: Exploring the brain (3rd ed.)*. Lippincott Williams and Wilkins. 377–379 pages.

[6] C Bolduc, A. M. Daoust, E Limoges, C. M. Braun, and R Godbout. 2003. Hemispheric lateralization of the EEG during wakefulness and REM sleep in young healthy adults. *Brain and Cognition* 53, 2 (2003), 193.

[7] S Cabral, R. A. Resende, A. C. Clansey, K. J. Deluzio, W. S. Selbie, and A. P. Veloso. 2016. A Global Gait Asymmetry Index. *Journal of Applied Biomechanics* 32, 2 (2016), 171–177.

[8] J. H. Cole, Poudel Rpk, D Tsagkrasoulis, Caan Mwa, C Steves, T. D. Spector, and G Montana. 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage* 163 (2017), 115.

[9] Koel Das, Barry Giesbrecht, and Miguel P Eckstein. 2010. Predicting variations of perceptual performance across individuals from neural activity using pattern classifiers. *Neuroimage* 51, 4 (2010), 1425–1437.

[10] Richard J. Davidson and Nathan A. Fox. 1982. Asymmetrical Brain Activity Discriminates between Positive and Negative Affective Stimuli in Human Infants. *Science* 218, 4578 (1982), 1235.

[11] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2018. Universal Transformers. *CoRR* abs/1807.03819 (2018). arXiv:1807.03819 http://arxiv.org/abs/1807.03819

[12] Hao Dong, Akara Supratak, Wei Pan, Chao Wu, Paul M Matthews, and Yike Guo. 2018. Mixed neural network approach for temporal sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 2 (2018), 324–333.

[13] Zhen Gao and Shangfei Wang. 2015. Emotion Recognition from EEG Signals byźLeveraging Stimulus Videos. In *Proceedings, Part II, of the 16th Pacific-Rim Conference on Advances in Multimedia Information Processing – PCM 2015 - Volume 9315*. Springer-Verlag, Berlin, Heidelberg, 118–127.

[14] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, and S. Zuo. 2019. EEG-Based Spatio-Temporal Convolutional Neural Network for Driver Fatigue Evaluation. *IEEE Transactions on Neural Networks and Learning Systems* (2019), 1–9. https://doi.org/10.1109/TNNLS.2018.2886414

[15] Anupriya Gogna, Angshul Majumdar, and Rabab Ward. 2017. Semi-supervised Stacked Label Consistent Autoencoder for Reconstruction and Analysis of Biomedical Signals. *IEEE Transactions on Biomedical Engineering* 64, 9 (2017), 2196–2205.

[16] Stelios K. Hadjidimitriou and Leontios J. Hadjileontiadis. 2012. Toward an EEG-Based Recognition of Music Liking Using Time-Frequency Analysis. *IEEE Trans Biomed Eng* 59, 12 (2012), 3498–3510.

[17] Y. Jia, C. Zhou, and M. Motani. 2017. Spatio-temporal autoencoder for feature learning in patient data with missing observations. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 886–890. https://doi.org/10.1109/BIBM.2017.8217773

[18] Zhicheng Jiao, Xinbo Gao, Ying Wang, Jie Li, and Haojun Xu. 2018. Deep Convolutional Neural Networks for mental load classification based on EEG data. *Pattern Recognition* 76 (2018), 582 – 595. https://doi.org/10.1016/j.patcog.2017.12.002

[19] Jeffrey S Johnson and Bruno A Olshausen. 2003. Timecourse of neural signatures of object recognition. *Journal of Vision* 3, 7 (2003), 4. https://doi.org/10.1167/3.7.4 arXiv:/data/journals/jov/932827/jov-3-7-4.pdf

[20] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M Norcia, and Patrick Suppes. 2015. A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *Plos one* 10, 8 (2015), e0135697.

[21] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. 2017. Brain2Image: Converting Brain Signals into Images. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1809–1817.

[22] M. K. Kim, M Kim, E Oh, and S. P. Kim. 2013. A review on the computational methods for emotional state estimation from the human EEG. *Computational and Mathematical Methods in Medicine,2013,(2013-3-24)* 2013, 2 (2013), 573734.

[23] Stefan Knecht, Michael Deppe, Bianca Dräger, Bobe, Hubertus Lohmann, Ringelstein, and Henningsen. 2000. Language lateralization in healthy right-handers. *Brain* 123, 1 (2000), 74–81.

[24] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of neural engineering* 15, 5 (October 2018), 056013. https://doi.org/10.1088/1741-2552/aace8c

[25] M. Li and B. Lu. 2009. Emotion classification based on gamma-band EEG. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 1223–1226. https://doi.org/10.1109/IEMBS.2009.5334139

[26] Xiaowei Li, Rong La, Ying Wang, Junhong Niu, Shuai Zeng, Shuting Sun, and Jing Zhu. 2019. EEG-based mild depression recognition using convolutional neural network. *Medical & Biological Engineering & Computing* (19 Feb 2019). https://doi.org/10.1007/s11517-019-01959-2

[27] Matthias M. Mller, Andreas Keil, Thomas Gruber, and Thomas Elbert. 1999. Processing of affective pictures modulates right-hemispheric gamma band EEG activity. *Clinical Neurophysiology* 110, 11 (1999), 1913–1920.

[28] Jinyoung Moon, Yongjin Kwon, Kyuchang Kang, Changseok Bae, and Wan Chul Yoon. 2015. Recognition of Meaningful Human Actions for Video Annotation Using EEG Based User Responses. In *International Conference on Multimedia Modeling*. Springer, 447–457.

[29] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. 2018. Decoding Brain Representations by Multimodal Learning of Neural Activity and Visual Features. *CoRR* abs/1810.10974 (2018). arXiv:1810.10974 http://arxiv.org/abs/1810.10974

[30] R Righart and Gelder B De. 2008. Rapid influence of emotional scenes on encoding of facial expressions: an ERP study. *Social Cognitive and Affective Neuroscience* 3, 3 (2008), 270.

[31] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. 2017. Deep Learning Human Mind for Automated Visual Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4503–4511. https://doi.org/10.1109/CVPR.2017.479

[32] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-To-End Memory Networks. In *NIPS*. 2440–2448. http://papers.nips.cc/paper/5846-end-to-end-memory-networks

[33] Chuanqi Tan, Fuchun Sun, Wenchang Zhang, Jianhua Chen, and Chunfang Liu. 2017. Multimodal Classification with Deep Convolutional-Recurrent Neural Networks for Electroencephalography. In *Neural Information Processing*, Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy (Eds.). Springer International Publishing, Cham, 767–776.

[34] B. O. Turner, N Marinsek, E Ryhal, and M. B. Miller. 2015. Hemispheric lateralization in reasoning. *Ann N Y Acad Sci* 1359, 1 (2015), 47–64.

[35] Oriol Vinyals, Lukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey E. Hinton. 2015. Grammar as a Foreign Language. In *NIPS*. 2773–2781. http://papers.nips.cc/paper/5635-grammar-as-a-foreign-language

[36] Jun Wang, Eric Pohlmeyer, Barbara Hanna, Yu-Gang Jiang, Paul Sajda, and Shih-Fu Chang. 2009. Brain state decoding for rapid image retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*. ACM, 945–954.

[37] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang. 2018. LSTM-Based EEG Classification in Motor Imagery Tasks. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26, 11 (Nov 2018), 2086–2095. https://doi.org/10.1109/TNSRE.2018.2876129

[38] Jamie Ward. 2006. *The student's guide to cognitive neuroscience*. Psychology Press. 548 pages.

[39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*. 2048–2057. http://proceedings.mlr.press/v37/xuc15.html

[40] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *CVPR*. 21–29. https://doi.org/10.1109/CVPR.2016.10

[41] Y. Yuan, G. Xun, K. Jia, and A. Zhang. 2019. A Multi-View Deep Learning Framework for EEG Seizure Detection. *IEEE Journal of Biomedical and Health Informatics* 23, 1 (Jan 2019), 83–94. https://doi.org/10.1109/JBHI.2018.2871678

[42] Dalin Zhang, Lina Yao, Xiang Zhang, Sen Wang, Weitong Chen, Robert Boots, and Boualem Benatallah. 2018. Cascade and Parallel Convolutional Recurrent Neural Networks on EEG-based Intention Recognition for Brain Computer Interface. In *AAAI Conference on Artificial Intelligence*. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16107