IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DEEP LEARNING ALGORITHMS FOR INTERNET OF MEDICAL THINGS

Received August 11, 2019, accepted August 18, 2019, date of publication August 22, 2019, date of current version September 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936817

A Hierarchical Bidirectional GRU Model With Attention for EEG-Based Emotion Classification

J. X. CHEN[1,2], D. M. JIANG[1], AND Y. N. ZHANG[1]

[1]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2]Department of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China

Corresponding author: J. X. Chen (chenjx_sust@foxmail.com)

This work was supported by the Youth Science Foundation Projects of National Natural Science Foundation of China under Grant 61806118 (Project Name: Research on EEG Based Emotion Recognition With Semi-Supervised Deep Generative Adversarial Network, Project Principal: J. X. Chen).

ABSTRACT In this paper, we propose a hierarchical bidirectional Gated Recurrent Unit (GRU) network with attention for human emotion classification from continues electroencephalogram (EEG) signals. The structure of the model mirrors the hierarchical structure of EEG signals, and the attention mechanism is used at two levels of EEG samples and epochs. By paying different levels of attention to content with different importance, the model can learn more significant feature representation of EEG sequence which highlights the contribution of important samples and epochs to its emotional categories. We conduct the cross-subject emotion classification experiments on DEAP data set to evaluate the model performance. The experimental results show that in valence and arousal dimensions, our model on 1-s segmented EEG sequences outperforms the best deep baseline LSTM model by 4.2% and 4.6%, and outperforms the best shallow baseline model by 11.7% and 12% respectively. Moreover, with increase of the epoch's length of EEG sequences, our model shows more robust classification performance than baseline models, which demonstrates that the proposed model can effectively reduce the impact of long-term non-stationarity of EEG sequences and improve the accuracy and robustness of EEG-based emotion classification.

INDEX TERMS Hierarchical, bidirectional GRU, attention, EEG, emotion classification.

I. INTRODUCTION

With the development of Deep Learning and Artificial Intelligence technology, affective recognition has become a hot research topic in the field of human-computer interaction [1]. Numerous studies in neurophysiology and psychology have found that the production or activity of human emotions is closely related to the dynamics of the cerebral cortex [2]. Electroencephalogram (EEG) is a general reflection of the electrophysiological activity of brain neurons in the cerebral cortex or scalp surface, which contains a lot of physiological, psychological and pathological information. Different human motor, cognitive and emotional activities can induce different EEG signals. Effective feature extraction and classification on EEG signals make it possible to read people's mind and to achieve some control purposes [3]. Considering the strong objectivity, non-forgery and easy acquisition of EEG signals, they are gradually applied to the detection and classification of human emotions [4], fatigue [5], identity verification [6], and target images [7].

The associate editor coordinating the review of this article and approving it for publication was Wei Wei.

In this paper, we study the methods of feature learning and representation of large-scale EEG data, and then make emotion classification. Because EEG signals are highly time-dependent, non-stationary and susceptible to noise interference, it is difficult to tell and extract the key time points or segments with high emotional correlation in the EEG sequence. Therefore, it is a great challenge to design a model that can accurately predict the emotional categories of EEG. Some linear inverse algorithms, such as Dynamic Statistical Parametric Maps (DSPM), Minimum Norm Estimation (NME), Phase Shift and Vector Modulation [8], all assume that EEG signals are stationary. Sliding window method solved some non-stationary problems of EEG to a certain extent, but its statistical efficiency is low and incomplete. Zhang *et al.* [9] combined autoregressive model with wavelet decomposition to improve the performance of EEG sequence modeling. Chen *et al.* [10] proposed an EEG feature extraction method by combining data space adaptation (DSA) and common spatial patterns (CSP) algorithms, which alleviated the degradation of emotion classification performance caused by fluctuations and differences of cross-day EEG signals. Based on that, the researchers further proposed a method to

combine common space pattern with wavelet packet decomposition to extracted the cross-day robust emotion-related EEG features and further reduced the influence of non-stationarity of EEG signals on emotion classification [4].

Recently, deep learning models including Convolution Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short Time Memory Network (LSTM) networks have achieved great success in natural language processing, image classification, video tracking and speech sequence modeling because of their strong end-to-end self-learning ability for complex feature representation without tremendous feature engineering. Now, CNN, RNN and LSTM have been gradually applied to the modeling of EEG signals, but the research is still in its infant stage. Jirayucharoensak *et al.* [11] proposed a deep learning network based on Stacked Autoencoder (SAE) to model the energy spectral density features of EEG signals and applied a PCA-based principal component covariant adaptive transformation algorithm to make emotion classification in valence and arousal dimensions. Compared with traditional SVM classifier, their performance improved by 5.55% and 6.53% respectively. Stober *et al.* [12] discussed the application of deep convolution autoencoder in capturing the invariant features of EEG data between the same and different subjects and achieved good results. Alhagry *et al.* [13] proposed a deep LSTM network to make EEG-based within-subject emotion classification in dimensions of arousal, valence and liking on DEAP data set and achieved the accuracy of 85.65%, 85.45% and 87.99% respectively, which was much higher compared with traditional methods. Soleymani *et al.* [14] proposed a method to detect the emotional state of the subjects from their EEG signals and facial expressions in real time by using LSTM-RNN and continuous conditional random field algorithm and achieved better performance. Salama *et al.* [15] used a three-dimensional convolutional neural network to recognize emotional categories from with-subject multi-channel EEG data in DEAP dataset and obtained the accuracy of 87.44% and 88.49% respectively in arousal and valence dimensions. In our previous work [16], we also explored the deep CNN model to learn the high-level discriminative feature representation from time and frequency domain combination features of EEG on DEAP data set and acquired the emotional within-subject classification accuracy of 88.6% and 86.7% respectively in valence and arousal dimensions.

These methods have improved the performance of EEG emotional classification to a certain extent, but there are few models considering the different importance of EEG epochs and samples to the emotional category of the sequence. Moreover, most of the current studies focus on emotion classification from subject-dependent EEG data rather than subject-independent EEG data which instead is very important to develop a high-performance EEG based affective human-machine interaction application. Because we don't expect to train a specific model for each user to make emotion prediction. Instead, we hope to extract the common EEG features of the same category of emotions among users and fine-tune a general model to predict the specific user's emotion status online.

Therefore, we propose a hierarchical bidirectional Gated Recurrent Unit (GRU) network with attention mechanism which is called H-ATT-BGRU to model the time-varying EEG sequence recorded during external emotional stimulation. The inspiration of the model comes from the Hierarchical Attention Network (HAN) proposed in [17] for document classification based on the hierarchical structure of words, sentences and documents, which assumes that not all words are equally important for the sentence class, and not all sentences are equally important for the document classification. That is very similar to the non-stationarity of EEG sequence. It is observed that different samples and epochs in an EEG sequence are differentially informative, moreover, the importance of samples and epochs are highly context dependent, we introduce the attention mechanism in both layers to solve the sensitivity to that fact. By focusing on different weights of content of different importance, more significant EEG feature vector is constructed, which highlights the contribution of important epochs and samples to the emotion classification of an EEG sequence. To evaluate the performance of our model, we use the EEG data of the open DEAP data set to make cross-subject emotion classification experiments in valence and arousal dimensions. The experimental results are compared with those of baseline traditional classifiers and deep neural networks, which proves that our model has obvious advantages over previous baseline methods.

The content of the paper is organized as follows, we first introduce our hierarchical bidirectional attention-enhanced GRU model which consists of sample encoder, attention-based sample aggregation, epoch encoder, attention-based epoch aggregation and sequence classification. Next, we describe the EEG data set and data preprocessing, elaborate the EEG-based cross-subject emotion classification experiments on baseline models and our proposed model. Then, we present and analyze the experimental results, make comparison and discussion with the baseline shallow classifiers and deep networks. We also present the attention weight distribution of some testing EEG samples and visualize the attention weights of epochs and samples learned by our model. Conclusion is provided in the end.

II. HIERARCHICAL GRU MODEL WITH ATTENTION

The hierarchical structure of our proposed H-ATT-BGRU mode is shown in Fig.1. It includes an EEG sample encoder, a sample-level attention layer, an EEG epoch encoder and an epoch-level attention layer. Details of these components are described below.

A. BIDIRECTIONAL GRU NETWORK FOR ENCODER

In this paper, we applied a bidirectional GRU network [18] for EEG sequence encoding. GRU model is a famous variant of LSTM [19], which synthesizes the forgetting gate and input gate to a single update gate and also mixes cell state and

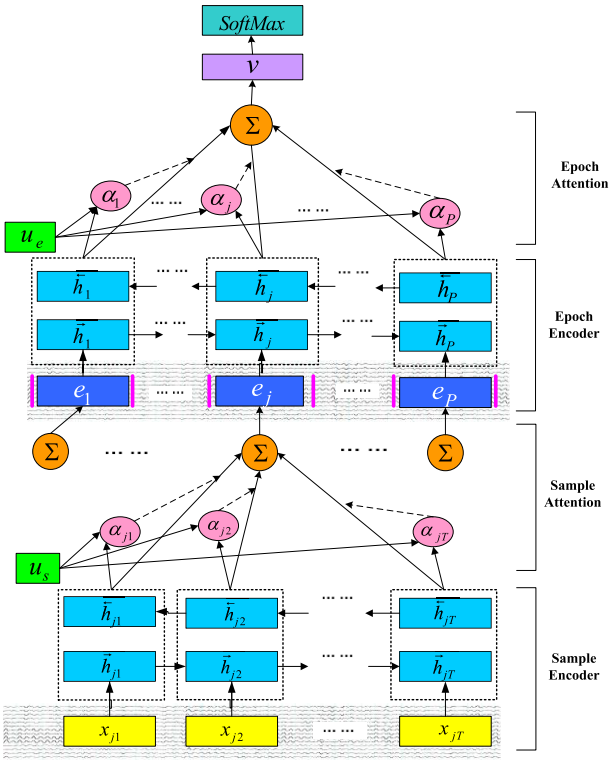

FIGURE 1. Hierarchical Bidirectional GRU model with Attention.

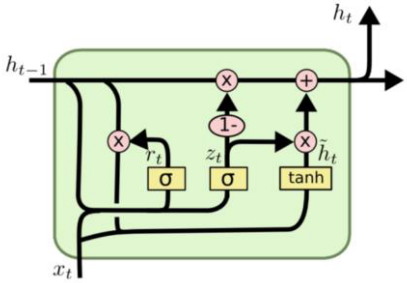

FIGURE 2. Internal computing structure of GRU.

hidden state. So, the final GRU model is simpler and faster than the standard LSTM model. Especially when training big data, it can save a lot of time with small performance difference from that of standard LSTM model. Both LSTM and GRU can retain important features through various Gates and ensure these features will not be lost in long-term transmission. The internal structure of GRU model is shown in Fig.2, where z_t represents update gate and r_t represents reset gate.

At time t, the GRU calculates the new state as:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{1}$$

This is to compute a linear interpolation between the previous state h_{t-1} and the current candidate state $\tilde{h}_t$ with the new sequence information. The update gate z_t decides to keep how much past information and to add how much new information. It controls the extent to which the information of the previous state is brought into the current state. The larger

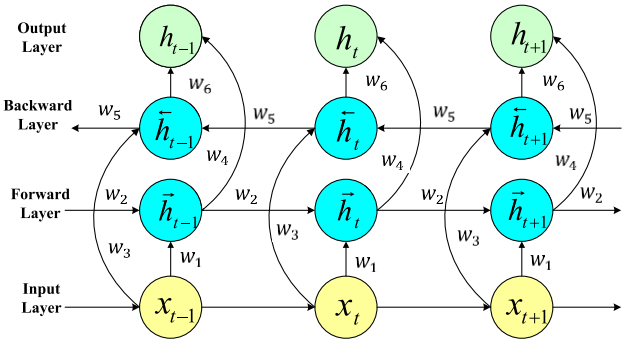

FIGURE 3. Bidirectional GRU network structure.

the value of z_t, the more information of the previous state is brought in. The state of z_t is updated as:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{2}$$

where x_t is the sample vector at time t, $\tilde{h}_t$ is the candidate state computed in the same way as the hidden layer of traditional RNN network:

$$\tilde{h}_t = \tanh\left(W_h x_t + r_t \odot (U_h h_{t-1}\right) + b_h \tag{3}$$

where r_t denotes a reset gate which controls how much the previous state contributes to the current candidate state $\tilde{h}_t$. The smaller the r_t value, the smaller the contribution from the previous state. If r_t=0, it will forget the previous state. The reset gate is updated as:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{4}$$

For many sequence modelling tasks, it is beneficial to have access to future as well as past context. However, standard GRU networks process sequences in temporal order and they ignore future context. Bidirectional GRU networks extend the unidirectional GRU networks by introducing a second layer, where the hidden to hidden connections flow in opposite temporal order. The model is therefore able to exploit information both from the past and the future, which architecture is shown in Fig.3.

B. HIERARCHICAL ATTENTION MODEL

In order to apply the H-ATT-BGRU model, the EEG sequence is divided into several segments of equal length, each of which is called an epoch and contains an equal number of time points which are called samples. The purpose of this model is to encode the raw EEG sequence into a vector representation, on which we construct a classifier to make emotion classification. Next, we will describe in detail how to use the hierarchical attention model to build the sequence level vector from sample vectors step by step.

1) SAMPLE ENCODER

Because the cognitive activity of human brain is time-varying, the occurrence time of the related EEG signal is also time-varying. The first layer of our H-ATT-BGRU model is

the sample encoder as shown in Fig.1, which is used to learn the local correlation among the samples in an EEG epoch.

The sample encoder uses a bidirectional GRU network which architecture is shown in Fig.3. Suppose an EEG sequence contains P epochs $e_j (j \in [1, P])$, each epoch contains T samples, and $x_{jt} (t \in [1, T])$ represents the T^{th} sample in the J^{th} epoch. x_{jt} is a vector with size of $C \times 1$, C is the number of channels. The number of hidden units is also T. The forward GRU reads the epoch e_j from x_{j1} to x_{jT} and a backward GRU reads the epoch e_j from x_{jT} to x_{j1}:

$$\vec{h}_{jt} = \vec{GRU}_f\left(x_{jt}\right), t \in [1,T]$$
$$\overleftarrow{h}_{jt} = \overleftarrow{GRU}_b\left(x_{jt}\right), t \in [1, T]$$

The annotation vector of each EEG sample is calculated by summarizing its hidden outputs from both directions as (5)-(7), so that the bidirectional contextual information of the epoch is incorporated into the annotation.

$$\vec{h}_t = f(w_1 x_t + w_2 \vec{h}_{t-1}) \tag{5}$$
$$\overleftarrow{h}_t = f(w_3 x_t + w_5 \overleftarrow{h}_{t+1}) \tag{6}$$
$$h_t = g(w_4 \vec{h}_t + w_6 \overleftarrow{h}_t) \tag{7}$$

2) SAMPLE ATTENTION

Not all samples have the same contribution to the emotion category of the EEG epoch. Some local samples may be more informative than others. Therefore, attention mechanism is introduced to extract samples that are important to the emotion of the epoch and aggregate the annotations of these informative samples into a vector representation of the EEG epoch.

Specifically, the annotation h_{jt} of each sample is input into a single layer Multiple-layer Perceptron (MLP), and its hidden representation u_{jt} is obtained by (8). Then the similarity between u_{jt} and the sample level context vector u_s is calculated and the normalized weight α_{jt} representing the importance of the sample x_{jt} is measured through a SoftMax function as (9). After that, we calculate the epoch vector e_j as a weighted sum of the sample annotations based on their weights as (10). The sample level context vector u_s can be regarded as a high-level representation of a fixed query ''what is the important sample'' over the samples in the epoch like the method used in [20], [21]. The vector of u_s is randomly initialized and fine-tuned through the training process. $w_s \in R^{l \times 1}$, $b_s \in R$ are the default weights and bias vectors of the single-layer MLP which are randomly initialized with normal distribution respectively, and l is the length of sample annotation vectors.

$$u_{jt} = \tanh(W_s h_{jt} + b_s) \tag{8}$$
$$\alpha_{jt} = \frac{\exp(u_{jt}^T u_s)}{\sum_t \exp(u_{jt}^T u_s)} \tag{9}$$
$$e_j = \sum_t \alpha_{jt} h_{jt} \tag{10}$$

3) EPOCH ENCODER

Given the epoch vector e_j, a vector representation of the EEG sequence can be obtained in the same way. As shown in the Epoch Encoder section of Fig.1, P is the number of epochs in each sequence, and the number of hidden units is also P. EEG epochs are still encoded by bidirectional GRU network as:

$$\vec{h}_j = \vec{GRU}_f\left(e_j\right), \quad j \in [1,P]$$
$$\overleftarrow{h}_j = \overleftarrow{GRU}_b\left(e_j\right), \quad t \in [1, P]$$

We then calculate the annotation vector h_j of the epoch e_j by summarizing $\vec{h}_j$ and $\overleftarrow{h}_j$ from both directions in the similar way as (7), so that the neighbor contextual information around the epoch j is incorporated into the annotation h_j.

4) EPOCH ATTENTION

In order to reward the clue epochs that contribute to the correct classification of the sequence, attention mechanism is used again, and the epoch level context vector u_e is introduced to measure the importance of the epochs. Similarly, the annotation h_j of each epoch is feed into a single-layer MLP, and its hidden representation u_j is calculated by (11). The similarity between u_j and the epoch level context vector u_e is computed and the normalized importance weight α_j is measured through a SoftMax function as (12). Then, we get the sequence vector v as a weighted sum of the epoch annotations based on their weights as (13), which fuses all the information of epochs in a sequence. The context vector u_e, the default weights w_e and bias vector b_e are randomly initialized and fine-tuned during the training process.

$$u_j = \tanh(W_e h_j + b_e) \tag{11}$$
$$\alpha_j = \frac{\exp(u_j^T u_e)}{\sum_t \exp(u_j^T u_e)} \tag{12}$$
$$v = \sum_j \alpha_j h_j \tag{13}$$

C. SEQUENCE CLASSIFICATION

Vector v is a high-level feature representation of EEG sequence, which is more discriminative and robust. Therefore, it can be classified by:

$$p = \text{softmax}(W_c v + b_c) \tag{14}$$

We use N category-balanced EEG sequences $\{(X_1, Y_1), (X_2, Y_2), \ldots (X_N, Y_N)\}$ to train the model, where $N = S \times L$ denoting that these N sequences come from S subjects and each subject watches L videos to induce L EEG sequences. Each sequence consists of P epochs, each epoch consists of T samples, each sample is a vector in size of $C \times 1$, and C is the number of channels. Y_j is the label of the sequence j. The model is trained by minimizing the cross-entropy between the predicted label and the real label. The loss function of the model is shown in formula as:

$$\text{E}\Theta = -\frac{1}{N} \sum_{n-1}^{N} Y_n \log\left(\tilde{Y}_n\left(X_n, \theta\right)\right) + \frac{\lambda}{2} |\,|\theta|\,|_2^2 \tag{15}$$

here, θ is a set of model parameters, λ is the regularization parameter. The model is trained with Adam optimizer.

III. EXPERIMENTS AND RESULTS

Firstly, we introduce the DEAP data set and EEG data preprocessing methods. Then, we evaluate the effectiveness of our model compared with other baseline shallow classifiers and other deep neural networks through cross-subject binary emotion classification experiments in valence and arousal dimensions.

A. EEG DATA SET

The experiments are performed on the DEAP data set developed by researchers at Queen Mary University of London, which is a large open source dataset containing multiple physiological signals with emotional evaluations [22]. The data set records EEG, ECG, EMG and other bioelectrical signals while 32 subjects were watching 40 1-minute music videos with different emotional tendencies. The subjects then evaluate the videos with emotion scale from 1 to 9 in dimensions of valence, arousal, liking, dominance and familiarity. The rating value from small to large indicates that each index is from negative to positive or from weak to strong. The 40 stimulus videos consist of 20 high valence/arousal ones and 20 low valence/arousal ones.

B. DATA PREPROCESSING

We extract the first 32-channel EEG signals from DEAP data set, down sample the data to 128 Hz, filter the data with a band-pass filter of 4-47 Hz to eliminate the background noise, make common average referencing and remove the ocular artifacts by blind source separation algorithm. The total length of the preprocessed EEG signals is 63-s, including 60-s of watching video and 3-s before watching. We then remove the 3-s baseline signals' average from the 60-s EEG signals (7680 readings) of watching video and normalize the data across channel to acquire the normalized stimulus related dynamics. Next, in the time domain, the 60-s EEG sequences are segmented into sixty 1-s epochs. So, the size of each subject's data set is 40(trials)×60(epochs)×28 (samples)×32(channels) and each trial corresponds to an EEG sequence. Based on the emotional rating value of each video in scale of 1 to 9 in arousal and valence domain, the median 5 is used as the threshold to divide the rating value into two categories: the value more than 5 is labeled 1 that means high arousal/valence, and the value less than or equal to 5 is labeled 0 that means low arousal/valence. Finally, the labels with size of 40×1 for each subject are obtained.

To evaluate the performance of our model on different length epochs, we also segment each EEG sequence into 120, 24 and 12 epochs with length of 0.5-s, 2.5-s and 5-s respectively. Therefore, each epoch contains 64, 320 and 640 samples, and each sample contains 32-channel data. We call these segmented EEG data in time domain as RAW feature of each subject. The detail of the RAW feature in different epoch length and their labels are shown in Table 1.

TABLE 1. Size of RAW feature in different epoch length.

Epoch length	Size of RAW features	Size of Labels
0.5-s	40(trials)×120(epochs)×64(samples)*32(channels)	40×1
1-s	40(trials)×60(epochs)×128(samples)*32(channels)	40×1
2.5-s	40(trials)×24(epochs)×320(samples)*32(channels)	40×1
5-s	40(trials)×12(epochs)×640(samples)*32(channels)	40×1

C. BASELINE MODELS

1) SHALLOW CLASSIFIERS

We use the shallow classifiers including Bagging Tree (BT) and Supported Vector Machine (SVM) which have shown good performance in EEG-based emotion recognition [23] as baselines. Because the shallow classifiers cannot self-learn the high-level discriminant EEG features, we extract the features of autoregression (AR) and power spectral density (PSD) respectively, make feature selection and then perform binary emotion classification in valence and arousal.

Bagging Tree (BT) is a kind of supervised classifier which combines a group of weak decision tree classifiers into a strong classifier through iteration. Given a group of samples and each sample has a set of attributes and a predetermined category. In each iteration, 70% samples are randomly selected to form the sub training set D_t to train the t^{th} weak classifier and then put these samples back to repeat the same iteration 100 times. Finally, the category voted most by these weak classifiers is chosen as the final classification result by voting. The core idea of the Support vector machine is to determine an optimal hyperplane to make the samples fall in either side of it and make the distance between the sample and the hyperplane as large as possible. The SVM classifier can get good classification performance on even small data set. Here, we apply the linear kernel function and sequential minimal optimization for SVM to make binary classification experiments [16].

Previous studies [2], [6] have proved that BT and SVM have better performance on PSD and AR features in EEG-based emotion classification. We use the Fast Fourier algorithm to extract 64 PSD features by sliding the 0.5-s Hamming window with 0.25-s step in 4-47Hz frequency domain of each 1-s EEG epoch and the size of the PSD feature of each subject is 2400(epochs)×64 (PSD samples)×32(channels). Next, we use the Sequential Floating Forward Selection (SFFS) method [24] to select the more discriminative PSD features, as well as reduce the feature dimension. Finally, we get the selected 16 PSD features called PSD + SFFS for training BT and SVM models.

A real and stable autoregression process is shown as:

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + e(n) \tag{16}$$

where, p is the order of the model, $x(n)$ is the signal sampled at time n, a_k is the k-order autocorrelation

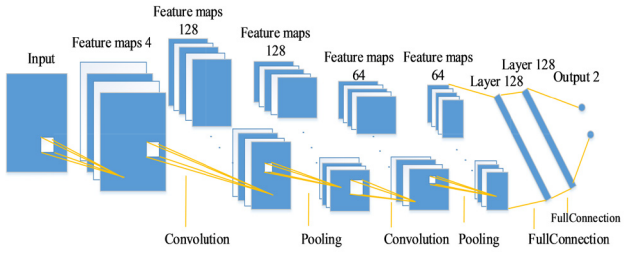

FIGURE 4. Structure of deep CNN model.

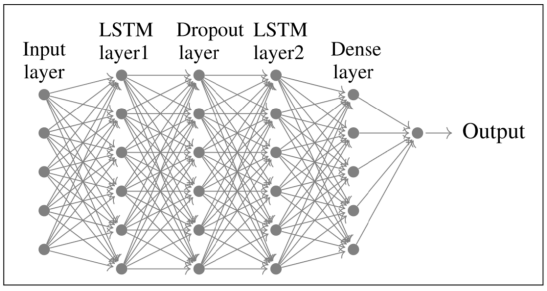

FIGURE 5. The structure of LSTM model.

coefficient, $e(n)$ is the error term independent of the past time samples. Assuming that a_k is a noise with zero mean and finite variance, the solution process of $x(n)$ is a regression over all the p sample points before time n. The value of a_k is estimated by the methods in [25] through limited samples $x(1), x(2), x(3), \ldots, x(N)$, and N is the number of samples. Finally, the first six maximal autocorrelation coefficients of each channel are taken as AR features of each epoch and the total AR features are size of 2400(epochs)×6(AR features)×32(channels). Then, we build 10-fold cross-subjects cross validation data sets on PSD+SFFS and AR features respectively to train and test the BT and SVM models in valence and arousal, which classification accuracies are taken as performance baselines.

2) DEEP NEURAL NETWORKS

We also explore two popular deep neural networks including CNN and LSTM to make cross-subject emotion classification on 1-s RAW features of DEAP data set as baselines. The CNN model regards each 1-s EEG epoch as an image of 128(samples)×32(channels) as the input. The total number of EEG epochs of 32 subjects is 76800 (32×40×60), so as the number of labels. We randomly select the epochs of 8 subjects (25% of total) as test data, and the epochs of the rest 24 subjects (75% of total) as training data. In this way, 10-fold cross-subject cross validation sets are constructed to train the CNN model.

The structure of the CNN model shown as Fig.4 is similar to that of CVCNN in [16], which consists of two convolution layers with four 3*3 convolutional kernels, two maximum pooling layers with kernel size of 2*2, one dropout layer with probability of 0.6 and two full-connection layers both with 128 hidden units. The first convolution layer has 128 feature maps and the second one has 64 feature maps. The convolution padding is 0 and the step is 1. The learning rate is 0.05. Finally, a SoftMax layer is appended to classify the output of the last full-connection layer into two emotional categories.

The LSTM model is shown as Fig.5, which also takes 1-s EEG epoch as an input sequence consisting of 128 samples. Each sample is represented as a 32-channel vector. The total number of EEG epochs and labels of 32 subjects is also 76800. Similarly, we construct 10-fold cross-subject cross validation sets to train the LSTM model. The LSTM model consists of two LSTM layers with 128 and 64 hidden units respectively, a dropout layer with probability of 0.3, a full-connection layer with 128 hidden units and a SoftMax classification layer. The *tanh* activation function is used for both LSTM layers. The learning rate is 0.03.

D. THE CONFIGURATION OF H-ATT-BGRU MODEL

We train our proposed H-ATT-BGRU model on 10-fold cross-subject cross validation data sets of EEG sequences with 1-s epochs, and test its binary emotion classification accuracy in the valence and arousal. The N EEG sequences fed into the model are similar to N documents. Each sequence contains 60 epochs, which is similar to each document contains 60 sentences. Each epoch contains 128 samples, which is similar to each sentence contains 128 words. Each sample is a 32-channel feature vector, which is similar to each word is represented with a 32-dimension word embedding vector. As shown in Fig.1, the high-level feature vector v of the input EEG sequence is generated through the process of sample encoder, attention-based samples aggregation, epoch encoder, and attention-based epochs aggregation. At last, the vector v is fed into the SoftMax layer to predict the emotional category. The model is trained by minimizing the cross-entropy between the predicted label and the real label.

The hyperparameters of the model are fine-tuned on the verification set which consists of 25% randomly selected sequences from the training set. In our model, we set the number of GRU hidden units to be 64, so the bidirectional GRU network can output 128-dimension sample/epoch annotation vectors after combining outputs of the forward and backward hidden units. The sample/epoch context vectors are also 128 dimensional initialized randomly. In training, mini-batch size is set to be 60 which means 60 sequences are organized to be a batch. It is also the number of epochs in each sequence. The model is trained with Adam optimizer. We pick the best learning rate of 0.05 using grid search on the validation set.

E. RESULTS AND ANALYSIS

The experimental results of different models on 1-s RAW features of DEAP data set in valence and arousal are shown in Table 2. In order to highlight the role of attention in our proposed H-ATT-BGRU model, we extend the hierarchical bidirectional GRU network to H-AVE-BGRU model and H-MAX-BGRU model. The H-AVE-BGRU model takes the average of sample/epoch annotations output by GRU as

TABLE 2. Emotion classification accuracy (%).

Method + Features	Valence	Arousal
BT+AR	55.7	54.4
BT+PSD+SFFS	56.2	53.5
SVM+AR	54.6	53.8
SVM+PSD+SFFS	55.3	54.5
CNN＋RAW	57.2	56.3
LSTM＋RAW	63.7	61.9
H-AVE-BGRU＋RAW	65.6	64.4
H-MAX-BGRU＋RAW	65.8	64.3
H-ATT-BGRU＋RAW	67.9	66.5

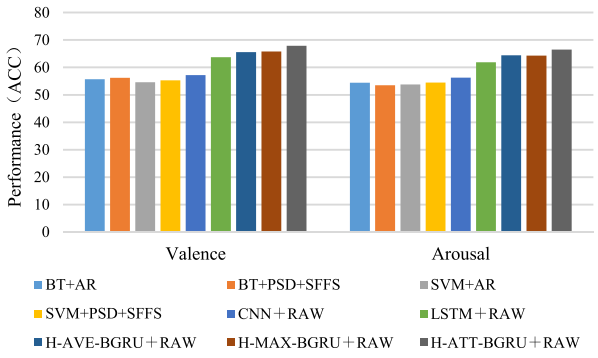

FIGURE 6. Comparison of performance of different models on 1-s epochs.

the aggregation feature vector of the epoch/sequence. The H-MAX-BGRU model takes the maximum of sample/epoch annotations output by GRU as the aggregation feature vector of the epoch/sequence. While our H-ATT-BGRU model aggregates the sample/epoch annotations with their context weight vectors to generate the epoch/sequence vector.

As shown in Table 2, our model gets the best classification performance of 67.9% in valence, which outperforms the best shallow baseline BT model by 11.7%, outperforms the best deep baseline LSTM network by 4.2%, and outperforms H-AVE-BGRU model and H-MAX-BGRU model by 2.3% and 2.1% respectively. This finding is consistent with that in arousal. Our H-ATT-BGRU model achieves the optimal classification accuracy of 66.5%, which is 12.0% higher than that of the best shallow baseline SVM model, 4.6% higher than that of the best deep LSTM network, 2.1% and 2.2% higher than those of H-AVE-BGRU model and H-MAX-BGRU model respectively.

Fig.6 show the comparison of performance of different models on 1-s segmented EEG sequence. We find that the performance of CNN model that do not explore hierarchical structure is not significantly better than that of the best traditional shallow classifier for large scale EEG sequence classification. For example, the best accuracy of CNN model in valence achieves 57.2%, which outperforms the best baseline BT model by 1%. The best accuracy of CNN model in arousal achieves 56.3%, which outperforms the best baseline SVM model by 1.8%. However, the performance of

TABLE 3. Classification accuracy on different length epochs (%).

Method + Features	0.5s	1s	2.5s	5s
LSTM+RAW	64.5	63.7	55.3	51.4
H-AVE-BGRU＋RAW	67.2	65.6	63.9	61.2
H-MAX-BGRU＋RAW	67.1	65.8	63.7	61.0
H-ATT-BGRU＋RAW	69.3	67.9	67.4	66.8

LSTM model is significantly better than that of CNN. For example, The LSTM model outperforms CNN by 6.5% and 5.6% in valence and arousal respectively, which shows that LSTM can learn the dependencies between samples in the whole sequence more efficiently than CNN.

When comparing the deep neural networks, we find that the H-AVE-BGRU model and H-MAX-BGRU model that explore hierarchical structure have obvious advantage over non-hierarchical CNN and LSTM models. For example, the classification accuracy of H-AVE-BGRU model is 8.4% and 8.1% higher than that of CNN, and 1.9% and 2.5% higher than that of LSTM in valence and arousal respectively. While our H-ATT-BGRU model that further utilizes attention method combined with hierarchical architecture outperforms the best baseline LSTM mode by 4.2% and 4.6% in valence and arousal respectively. What's more, in our experiments, H-AVE-BGRU is equivalent to initializing the sample/epoch context vectors with no information, such as they are all-one or all-zero vectors which makes the attention weights in (9) and (12) become uniform. In valence and arousal dimensions, our H-ATT-BGRU model improves over H-AVE-BGRU model by 2.3% and 2.1% respectively, which indicates that the proposed context sample/epoch importance-weighted vectors are effective for improving the classification performance of hierarchical attention networks.

To further explore the relationship between the performance of the model and the length of an epoch in an EEG sequence, we use four models to make valence classification experiments on EEG sequences segmented with different time-length. The experimental results are shown in Table 3.

From Table 3, we can see that the classification accuracies of four models are consistently improved with shortening the length of epochs. As a result, our H-ATT-BGRU model achieve its best cross-subject classification accuracy of 69.3% on 0.5-s segmented EEG sequences, which is 2.1% and 2.2% higher than that of H-AVE-BGRU and H-MAX-BGRU respectively, 4.8% higher than that of LSTM. When the EEG sequence is segmented in 0.5 seconds, one sequence is divided into 120 epochs and each epoch contains 64 samples. As the time-span of the epoch is short, the non-stationarity is not obvious. Taking the advantage of that, four models all present the optimal performance on 0.5-s segmented EEG sequences. Especially for our proposed model, its hierarchical structure and attention mechanism play significant roles in improving performance.

Observing each row in Table 3, we find that the accuracy of LSTM on 0.5-s segmented sequences is 0.8%, 9.2%

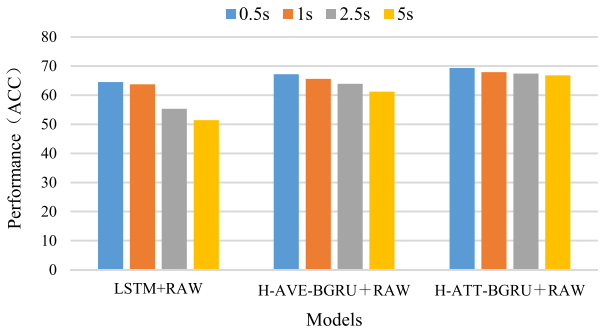

FIGURE 7. Comparison of performance on different length epochs.

and 13.1% higher than that on 1-s, 2.5-s and 5-s segmented sequences respectively. The accuracy of H-AVE-BGRU on 0.5-s segmented sequences is 1.6%, 3.3% and 6.0% higher than that on 1-s, 2.5-s and 5-s segmented sequences respectively. The accuracy of our H-ATT-BGRU model on 0.5-s segmented sequences is 1.4%, 1.9% and 2.5% higher than that on 1-s, 2.5-s and 5-s segmented sequences respectively. Fig.7 shows the comparison of classification performance of these three models on different time-length segmented sequences.

From Fig.7, we can see that there is no significant difference between the performance of three models on 0.5-s and 1-s segmented sequences. But when the time-length of epochs is 2.5-s and 5-s, the performance of LSTM decreases obviously, indicating that there may be significant non-stationarity in 2.5-s and 5-s EEG sequences and LSTM seemed to be inadequate to overcome this intrinsic non-stationarity in long time sequences. For the hierarchical H-AVE-BGRU model without attention mechanism, the performance difference is reduced, which indicates that the model is more capable of learning sample/epoch dependencies in long time span than LSTM. Our H-ATT-BGRU model further reduces the performance differences on different length segmented sequences, which demonstrates again that even for long time sequences, our hierarchical model with attention mechanism can learn more significant and dependent features according to the context importance weights at both sample and epoch levels to overcome the non-stationarity of EEG sequences and therefore improve the classification accuracy and robustness.

F. CONTEXT-DEPENDENT ATTENTION WEIGHTS

If each sample has no difference in importance for sequence classification, the model without attention mechanism may work well, because the model can automatically learn lower weights for irrelevant samples and higher weights for highly correlated samples. However, the importance of samples in EEG sequences is highly context-dependent. For example, in the EEG sequence with low valence, there may be a sample with a larger value, which may be due to the positive emotions of the subjects on part of the stimulus video. While in the EEG sequence with high valence, there may be a sample with a lower value, which may be due to the negative feeling on part

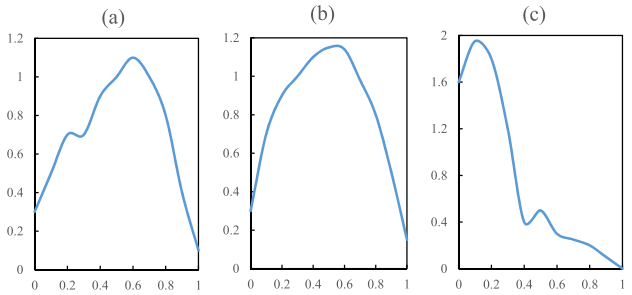

FIGURE 8. Attention weight distribution of the sample A with a larger value. (a) Joint distribution on the test set. (b) Distribution for EEG sequences with high valence. (c) Distribution for EEG sequences with low valence. We find that the weight distribution of sample A goes higher as the valence goes higher.

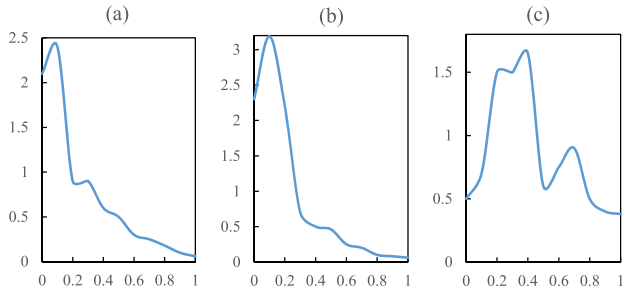

FIGURE 9. Attention weight distribution of the sample B with a smaller value. (a) Joint distribution on the test set. (b) Distribution for EEG sequences with high valence. (c) Distribution for EEG sequences with low valence. Unlike before, smaller-value samples are considered important for low valence, but less important for high valence.

of the stimulus video. However, the importance weights of these two samples in different categories of EEG sequences depend on their context.

To verify that our model can capture the context-related importance weights of EEG samples, we obtain the attention weight distribution of a sample A with a larger value and sample B with a smaller value from the test sequences of DEAP data set, as shown in Fig. 8 and Fig. 9. We can see that the range of attention weights assigned to each sample is set from 0 to 1. As can be seen from the graph, our model assigns context-dependent importance weights to the samples according to their different contexts.

G. ATTENTION VISUALIZATION

In order to verify that our model can select important epochs and samples with large amount of information from EEG sequences, we visualize the attention weights of epochs and samples learned by our hierarchical model of one test data set of DEAP. As shown in Fig.10, each row is an EEG epoch, pink to red on the left represents the weight of the epoch, and blue to purple on the right represents the weight of the sample. Because of the existence of hierarchical structure, we use epoch weights to normalize the sample weights to ensure that only the important samples in the important epochs are emphasized. We show the value of $\sqrt{\alpha}_e\alpha_s$ for visualization, $\sqrt{\alpha}_e$ denotes the weight of important samples in

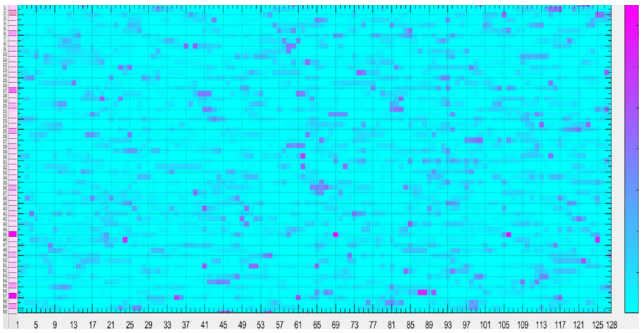

FIGURE 10. Visualization of the attention weights.

unimportant epochs to ensure that they are not completely invisible. However, we are not able to clearly explain the physical meaning of these highlighted important samples and epochs to the emotion category which need much comprehensive neurological and psychological professional knowledge. We will further study that in our future work.

IV. RELATED WORK

Bashivan *et al.* [26] proposed a deep recursive convolution neural network R-CNN for EEG-based cognitive and mental load classification. The network can learn robust features from EEG sequences that are insensitive to variations and distortions in time, space and frequency domains, and the classification error rate of SVM and random forest that they used before is reduced by more than 50%. Hefron *et al.* [27] used a LSTM-based deep recursive neural network to explain the temporal dependence of cognitive-related EEG signals, thus significantly improving the stationarity of the cross-day EEG features. Zhang *et al.* [28] proposed a deep convolution recurrent neural network model, which can accurately recognize human motion intentions by effectively learning the spatiotemporal correlation representation of the original EEG sequences, and obtained the subject-dependent classification accuracy of 98.3% on the MI-EEG data set, which is higher than other simple classifiers and other deep learning networks. Lawhern *et al.* [29] proposed an application of multilayer pure convolution neural network without complete connection layer and achieved the best performance in the paradigms of oddball recognition task based on P300, motor-related cortical potential recognition in finger motion tasks and sensory motor rhythm recognition in motion imagination tasks.

However, there is so far no hierarchical structured deep neural network applied for EEG-based classification. Hierarchical structure has been widely used for sequence generation and language modeling. Li et al. used a hierarchical LSTM auto-encoder to preserve and reconstruct multi-sentence paragraphs [30]. Lin et al. established a hierarchical RNN to capture the contextual information between sentences in a document, and integrated it into the word level RNN to predict the cross-sentence word sequence [31]. Bahdanau et al. firstly proposed the attention mechanism in machine translation and used an encoder decoder network with attention mechanism to select the candidate original words for foreign words before translation [18]. Vinyals et al. proposed an attention-based sequence-to-sequence model for syntactic constituency parsing [32]. Xu et al. introduced an attention-based model that automatically selected the relevant image regions and learn to generate words in the image caption [33]. Zhou et al. proposed an attention-based bidirectional LSTM network to capture the most important semantic information in a sentence for relation classification [34]. Hermann et al. developed a kind of deep neural networks with attention that learn to read real documents and answer questions with minimal prior knowledge of language structure [35]. Yang et al. presented a multiple-layer stacked attention networks that learn to answer natural language questions from images [36]. Unlike these works, we explore a hierarchical attention mechanism for EEG-based emotion recognition, and as far as we know this is the first such study instance.

V. CONCLUSION

In this paper, we propose a hierarchical bidirectional GRU model with attention mechanism (H-ATT-BGRU) for learning EEG sequence features and making cross-subject emotion classification on them. The first layer of the model encodes the local correlation among the samples in an epoch, and the second layer encodes the temporal correlation among the EEG epochs in a sequence. The model uses attention mechanism at both sample and epoch levels. By aggregating important EEG sample annotations into epoch vectors according to the sample's context weight, and then aggregating important epoch annotations into sequence vectors according to the epoch's context weight, the model highlights the contribution of important samples and epochs in the sequence to emotional categories.

We make cross-subject emotion classification experiments on large-scale EEG sequences of DEAP data set in valence and arousal to estimate the performance of our model. The experimental results show that our H-ATT-BGRU model achieves the best classification accuracy of 69.3% on the 0.5-s segmented EEG sequences, which is 2.1% and 2.2% higher than the hierarchical but non-attention H-AVE-BGRU model and H-MAX-BGRU model respectively, and 4.8% higher than that of the best baseline LSTM model. Our model also achieves the best classification accuracy of 67.9% and 66.5% on the 1-s segmented EEG sequences in valence and arousal, which is 4.2% and 4.6% higher than that of the optimal baseline LSTM model, and 11.7% and 12% higher than that of the best shallow baseline model respectively. Moreover, with increase of the epoch's length of EEG sequences, our model shows more robust classification performance than other baseline models, which demonstrates that the proposed model can effectively reduce the impact of long-term non-stationarity of EEG sequences and improve the accuracy and robustness of EEG-based emotion classification. The model will be applied to develop the applications of robust affective brain-computer interface in the future.

REFERENCES

[1] G. Lu, L. Yuan, W. Yang, J. Yan, and H. Li, "Speech emotion recognition based on long short-term memory and convolutional neural networks," *J. Nanjing Univ. Posts Telecommun.*, vol. 38, no. 5, pp. 63–69, Nov. 2018. doi: 10.14132/j.cnki.1673-5439.2018.05.009.

[2] J. Li, G. Z. Liu, and J. Gao, "Emotion classification based on EEG signal," *J. Beijing Univ. Inf. Sci. Technol.*, vol. 32, no. 2, pp. 34–39, May 2017.

[3] L. H. Zetterberg, "Estimation of parameters for a linear difference equation with application to EEG analysis," *Math. Biosci.*, vol. 5, nos. 3–4, pp. 227–275, 1969.

[4] J. Chen, D. Jiang, and Y. Zhang, "A common spatial pattern and wavelet packet decomposition combined method for EEG-based emotion recognition," *J. Adv. Comput. Intell. Intell. Inform.*, vol. 23, no. 2, pp. 274–281, 2019.

[5] R. Chai, G. Naik, T. N. Nguyen, S. Ling, Y. Tran, and A. Craig, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 3, pp. 715–724, May 2017.

[6] J. X. Chen, Z. J. Mao, W. X. Yao, and Y. F. Huang, "EEG-based biometric identification with convolutional neural network," *Multimedia Tools Appl.*, pp. 1–21, Feb. 2019. doi: 10.1007/s11042-019-7258-4.

[7] J. Chen, Z. Mao, R. Zheng, Y. Huang, and L. He, "Feature selection of deep learning models for EEG-based RSVP target detection," *IEICE Trans. Inf. Syst.*, vol. E102-D, no. 4, pp. 836–844, 2019.

[8] A. Gramfort, D. Strohmeier, J. Haueisen, and M. S. Hämäläinen, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," *NeuroImage*, vol. 70, pp. 410–422, Apr. 2013.

[9] Y. Zhang, B. Liu, X. Ji, and D. Huang, "Classification of EEG signals based on autoregressive model and wavelet packet decomposition," *Neural Process. Lett.*, vol. 45, no. 2, pp. 365–378, Apr. 2017.

[10] J. Chen, D. Jiang, and Y. Zhang, "An extended common spatial pattern framework for EEG-based emotion classification," in *Advances in Brain Inspired Cognitive Systems—BICS* (Lecture Notes in Computer Science), vol. 10989. Cham, Switzerland: Springer, 2018, pp. 215–223. doi: 10.1007/978-3-030-00563-4_27.

[11] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena, "EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation," *Sci. World J.*, vol. 2014, pp. 1–10, Sep. 2014.

[12] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn, "Deep feature learning for EEG recordings," 2015, *arXiv:1511.04306*. [Online]. Available: https://arxiv.org/pdf/1511.04306.pdf

[13] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 355–358, 2017.

[14] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affective Comput.*, vol. 7, no. 1, pp. 17–28, Jan./Mar. 2016.

[15] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. W. Shalaby, "EEG-based emotion recognition using 3D convolutional neural networks," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 329–337, 2018. doi: 10.14569/IJACSA.2018.090843.

[16] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44317–44328, 2019. doi: 10.1109/ACCESS.2019.2908285.

[17] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. HLT*, 2016, pp. 1480–1489.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2019, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[19] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 38, May 2013, pp. 6645–6649.

[20] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," 2015, *arXiv:1503.08895*. [Online]. Available: https://arxiv.org/abs/1503.08895

[21] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," 2015, *arXiv:1506.07285*. [Online]. Available: https://arxiv.org/abs/1506.07285?context=cs

[22] S. Koelstra, C. Muhl, and M. Soleymani, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[23] Y. Liu, O. Sourina, and M. K. Nguyen, "Real-time EEG-based emotion recognition and its applications," *Trans. Comput. Sci. XII*, vol. 6670, no. 12, pp. 256–277, 2011.

[24] D. Ververidis and C. Kotropoulos, "Information loss of the Mahalanobis distance in high dimensions: Application to feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2275–2281, Dec. 2009.

[25] R. Palaniappan, "Identifying individuality using mental task based brain computer interface," in *Proc. 3rd Int. Conf. Intell. Sens. Inf. Process.*, Dec. 2005, pp. 238–242.

[26] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," 2015, *arXiv:1511.06448*. [Online]. Available: https://arxiv.org/abs/1511.06448

[27] R. G. Hefron, B. J. Borghetti, J. C. Christensen, and C. M. S. Kabban, "Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation," *Pattern Recognit. Lett.*, vol. 94, pp. 96–104, Jul. 2017.

[28] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, and R. Boots, "Eeg-based intention recognition from spatio-temporal representations via cascade and parallel convolutional recurrent neural networks," 2017, *arXiv:1708.06578*. [Online]. Available: https://arxiv.org/abs/1708.06578

[29] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, 2018, Art. no. 056013.

[30] J. Li, M.-T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," 2015, *arXiv:1506.01057*. [Online]. Available: https://arxiv.org/abs/1506.01057

[31] R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, and S. Li, "Hierarchical recurrent neural network for document modeling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 899–907.

[32] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2773–2781.

[33] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[34] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. 54th Assoc. Comput. Linguistics*, 2016, pp. 207–212.

[35] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1693–1701.

[36] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 21–29.

J. X. CHEN was born in Shihezi, Xinjiang, China, in 1979. She received the B.S. and M.S. degrees from the Department of Electrical and Information Engineering, Shaanxi University of Science and Technology, China, in 2002 and 2005, respectively. She is currently pursuing the Ph.D. degree in computer science and technology with Northwestern Polytechnical University, Xi'an, China.

From 2006 to 2013, she was a Lecturer with the Shaanxi University of Science and Technology, where she has been an Associate Professor with the Department of Electrical and Information Engineering, since 2013. From 2016 to 2017, she researched on deep learning and electroencephalogram (EEG)-based event detection as a Visiting Scholar at The University of Texas at San Antonio, TX, USA. She is the author of one book, more than 20 articles, and two inventions. Her research interests include machine learning and pattern recognition, deep learning, EEG signal processing, and emotion recognition. Since 2016, she has been a member of the Chinese Computer Society.

D. M. JIANG was born in Zhengzhou, Henan, China, in 1973. She received the B.S., M.S., and Ph.D. degrees in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 1996, 1998, and 2000, respectively.

She researched at the Postdoctoral mobile station of control science and technology, Northwestern Polytechnical University (NPU), from December 2000 to November 2003. Since November 2003, she has been with the School of Computer Science, NPU, where she was promoted to be a Professor, in May 2010. From November 2001 to June 2002 and from June 2006 to October 2007, she worked on the International Cooperation Project between the Ministry of Science and Technology of China and the Belgian Government at the Free University of Brussels. Since 2005, she has been the Chinese Contact Person at the Joint Laboratory of Audiovisual Signal Processing, Northwestern Polytechnical University, and Brussels Free University. In April 2011, she was promoted to be the Supervisor of Ph.D. student with the School of Computer Science, NPU. She is the host of three national projects and three key provincial projects of Shaanxi Province and more than 50 articles. Her research interests include speech processing of audio-visual fusion, audiovisual emotion analysis and recognition, facial animation synthesis of expressive speakers, and emotion recognition of physiological signals. Since 2011, he has been the Contact Person of Northwestern Polytechnical University, China, and the Europe Liama Research Union. She is currently the Director of the Shaanxi Image and Graphics Society and the Shaanxi Signal Processing Society.

Dr. Jiang received the Second Prize for outstanding academic papers of Natural Science in Shaanxi Province and the First Prize for science and technology from Shaanxi University.

Y. N. ZHANG was born in Liquan, Shaanxi, China, in 1967. She received the B.S. degree in electronic engineering from the Dalian University of Technology, Dalian, China, in 1988, and the M.S. and Ph.D. degrees in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1992 and 1996, respectively. She finished her first-station Postdoctoral study at the Key Laboratory of Radar Defense Science and Technology, Xi'an University of Electronic Science and Technology, Xi'an, in December 1999. She finished her second-station Postdoctoral study at the School of Computer Engineering, Northwestern Polytechnical University, in March 2002.

From 2000 to 2010, she was the Director of the Shaanxi Key Laboratory of Speech and Image Information Processing, Northwestern Polytechnical University. In 2003, she was promoted to be a Professor at the School of Computer Science, NPU, where she was promoted to be the Supervisor of Ph.D. student, in 2005. From December 2012 to January 2018, she was the Dean of the School of Computer Science, Northwest Polytechnic University. Since December 2015, she has been the Assistant President of Northwest Polytechnic University. She has undertaken more than 40 national-level projects, such as 973 national defense projects, key projects of the National Natural Science Foundation of China, 863 national defense projects, and preliminary research of assembly projects. She is the author of three books and more than 100 papers that have been published in authoritative journals and important international conferences at home and abroad, such as the IEEE TPAMI, IEEE TIP, PR, IEEE TSMC-B, *Information Fusion*, CVPR, and ICCV. She is also the host more than 50 patents for invention. Her research interests include image processing, pattern recognition, computer vision, and remote sensing.

Dr. Zhang serves as the Standing Director of the Chinese Stereology Society, the Deputy Director and the Secretary-General of the Image Analysis Branch, the Director of the Chinese Image Graphics Society and the Chinese Artificial Intelligence Society, a member of the Signal Processing Branch, Chinese Electronics Society, and the Standing Vice-Chairman and the Secretary-General of the Shaanxi Signal Processing Society. She was rewarded as one of the Advanced Individuals of 863 Science and Technology Tackling of General Assembly, in 2011. She was selected as Young and Middle-aged Science and Technology Innovation Leading Talent in ''Talents Promotion Plan'' of the Ministry of Science and Technology of the CPC, in 2013. She was selected as one of the first batch of scientific and technological innovation leaders in the ''ten thousand people plan'' of the Organization Department of the Central Committee of the CPC, in 2014. She received one first prize for scientific and technological progress in Shaanxi Province and one first prize for national defense technological invention.