EEG-based Emotion Recognition via Channel-wise Attention and Self Attention

Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan and Xun Chen

Abstract—Emotion recognition based on electroencephalography (EEG) is a significant task in the brain-computer interface field. Recently, many deep learning-based emotion recognition methods are demonstrated to outperform traditional methods. However, it remains challenging to extract discriminative features for EEG emotion recognition, and most methods ignore useful information in channel and time. This paper proposes an attention-based convolutional recurrent neural network (ACRNN) to extract more discriminative features from EEG signals and improve the accuracy of emotion recognition. First, the proposed ACRNN adopts a channel-wise attention mechanism to adaptively assign the weights of different channels, and a CNN is employed to extract the spatial information of encoded EEG signals. Then, to explore the temporal information of EEG signals, extended self-attention is integrated into an RNN to recode the importance based on intrinsic similarity in EEG signals. We conducted extensive experiments on the DEAP and DREAMER databases. The experimental results demonstrate that the proposed ACRNN outperforms state-of-the-art methods.

Index Terms—Electroencephalogram (EEG), emotion recognition, channel-wise attention, self-attention.

1 Introduction

EMOTION analysis is important in daily life, particularly in the human-computer interaction field [1]–[5]. Emotion analysis can help increase the quality of human-computer communication and improve the intelligence of computer. In addition, emotion analysis plays an important role in health care to understand the behavioral and cognitive functioning of patients [6], [7], and physiological signals are generally used to measure the emotional state, including galvanic skin response, electromyography, heart rate, respiration rate and electroencephalography (EEG) [8].

In the past decade, the relationship between emotion and EEG signals has been studied extensively [9]–[11]. EEG signals, which can be obtained easily, measure voltage fluctuations resulting from ionic current flows in the neurons of the brain [12]. EEG is noninvasive, fast, and inexpensive, thereby making it a preferred method to brain responses to emotional stimuli. In addition, EEG signals are widely used for emotion analysis because EEG can be explored various information about emotions from frequency band, electrode position, and temporal information [13]–[15].

Generally, most EEG emotion recognition methods first design features from EEG signals and adopt classifiers to classify the emotion features. For example, Li *et al.* extracted features from the gamma frequency band and used a linear support vector machine (SVM) to classify the extracted features [13]. Patil *et al.* adopted higher-order crossings as features, which are better than other statistical features to classify emotions [16]. Shi *et al.* first proposed differential entropy (DE) features from five frequency bands and validated that DE features are superior for representing EEG signals [17]. In addition, Duan *et al.* extracted DE features from multichannel EEG data and combined an SVM and k-Nearest Neighbor (KNN) to classify the extracted features [18].

Recently, deep learning has been demonstrated to outperform traditional machine learning in many fields, e.g., computer vision [19], natural language processing [20] and biomedical signal processing [21]–[23]. In addition, many deep learning-based methods have been widely used for EEG-based emotion recognition. On one hand, deep learning methods can be considered as classifiers after feature extraction. For example, Yang *et al.* combined the DE of multiple bands as EEG features and employed a continuous convolutional neural network as a classifier [24]. Song *et al.* designed DE features according to the electrode position relationship and adopted a graph convolutional neural network as a classifier [25]. On the other hand, many deep learning methods are data-driven and function in an end-to-end manner, which does not require handcrafted features

- *This work is supported by the National Key Research and Development Program of China (Grant 2019YFA0706203), the National Natural Science Foundation of China (Grants 61922075, 41901350 and 61701160), the Provincial Natural Science Foundation of Anhui (Grants 2008085QF285 and 1808085QF186), the Key Research and Development Program of Anhui Province (1804h08020244) and the Fundamental Research Funds for the Central Universities (Grants JZ2019HGBZ0151 and JZ2020HGPA0111).*

- *W. Tao is with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China, and also with the Department of Electrical and Computer Engineering, University of Macau, Macau, China. E-mail: taovi1996@mail.hfut.edu.cn.*
- *C. Li, R. Song, J. Cheng and Y. Liu are with the Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China. E-mail: $\{changli, rcsong, chengjuan, yuliu\}$@hfut.edu.cn.*
- *F. Wan is with the Department of Electrical and Computer Engineering, University of Macau, Macau, China. E-mail: fwan@um.edu.mo.*
- *X. Chen is with the Department of Neurosurgery, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, and also with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, Anhui, 230001, China. E-mail: xunchen@ustc.edu.cn.*
- *Corresponding author: Chang Li.*

from EEG signals. For example, Alhagry *et al.* proposed an end-to-end deep learning neural network to recognize emotion from raw EEG signals, which used an LSTM-RNN to learn features from EEG signals and used the dense layer for classification [15]. Yang *et al.* proposed a parallel convolutional recurrent neural network for EEG emotion recognition and achieved good performance [7]. However, it still remains challenging to extract more discriminative features for EEG emotion recognition. Therefore, it is important to design an effective deep learning framework that can extract features and perform classification directly from raw EEG signals.

Inspired by the cascade convolutional recurrent network (CRNN), which combines CNN and RNN to extract spatial and temporal features from EEG signals [26], we use a CNN to extract the spatial information of EEG signals. Then, we employ two long short-term memory (LSTM) layers to extract temporal information, which is better at storing and accessing information than a standard RNN [27]. Different from a traditional CRNN, we employ a framework to extract more discriminative spatiotemporal information using two attention mechanisms, *i.e.,* a channel-wise attention mechanism [28] and an extended self-attention mechanism [29]. Generally, CNNs are used to extract the spatial features of EEG signals [7], however, this ignores the importance of features among different channels. To extract more discriminative features from the spatial information, some methods adopt channel selection to choose more relevant channels [30]. Different from traditional methods that need first select the relevant channels artificially [31], in this study, we first adopt an adaptive channel-wise mechanism, that transforms channels to a probability distribution as weights and recodes the EEG signals based on the transformed weights. Then CNNs are employed to extract the discriminative spatial features of recoded signals. In addition, an RNN is employed to explore the time information of EEG signals, however, this also ignores the importance of different EEG samples. Note that extended self-attention can be applied to LSTM to utilize long-range dependencies [32]. We integrate the extended self-attention mechanism into the RNN to explore the importance of different EEG samples, because this mechanism can update the weight according to the similarity of EEG signals. As a result, more discriminative temporal and spatial characteristics of EEG signals can be obtained by integrating the two attention mechanisms in our framework.

In this paper, we propose the attention-based convolutional recurrent neural network (ACRNN) to deal with EEG-based emotion recognition. Raw EEG signals can contain spatial information by the intrinsic relationship among different channels and time dependence among temporal slices, thus, the proposed ACRNN can learn the spatial features of multichannel EEG in the convolutional layer and explore the temporal features of different temporal slices using LSTM networks. In addition, the channel-wise attention and extended self-attention mechanisms can extract more discriminative spatial and temporal features, respectively. The proposed model was evaluated on two publicly available databases, *i.e.,* DEAP [2] and DREAMER [3], and the proposed method demonstrated superior performance relative to recognition accuracy in two databases. Our primary contributions are summarized as follows.

1) We have developed a data-driven ACRNN framework for EEG-based emotion recognition. This framework integrates the channel-wise attention mechanism into a CNN to explore spatial information, which can take the importance of different channels by channel-wise attention and the spatial information of multichannel EEG signals by a CNN into consideration. Besides, ACRNN integrates extended self-attention mechanism into RNN to explore temporal information of EEG signals, which can take the different temporal information by LSTM and the intrinsic similarity of each EEG sample by extended self-attention into consideration.

2) We conducted experiments on the DEAP and DREAMER databases, and the experimental results indicate average emotion recognition accuracies of 92.74% and 93.14% in the valence and arousal classification tasks of the DEAP database, respectively. In addition, the proposed method achieved mean accuracies of 97.79%, 97.98% and 97.67% in the valence, arousal and dominance classification tasks of the DREAMER database, respectively.

The remainder of this paper is organized as follows. Section II introduces related work, and Section III presents the proposed method. Section IV discusses extensive experiments conducted to demonstrate the effectiveness of the proposed ACRNN. Finally, a discussion is given in Section V, and the paper is concluded in Section VI.

2 Related Work

Here, we introduce the general flow of the traditional EEG emotion recognition framework. We then introduce the channel-wise attention and self-attention mechanisms.

2.1 General Flow of EEG Emotion Recognition

Recently, emotion recognition from EEG signals has received significant attention. The general flow of the EEG emotion recognition framework is summarized as follows (Fig. 1).

(**i**) Test protocol: First, the type of stimulus used, trial duration, the number of subjects, their gender, and the emotions to be recognized are recorded. Then, the subjects are exposed to the stimulus, e.g., music or a movie [2], [3].

(**ii**) EEG recordings: The number of electrodes and test duration are recorded, and then EEG signals are recorded by electrodes. The subjects then assess their emotional state by labeling the EEG recording after each trial [2], [3].

(**iii**) Preprocessing: To avoid artifacts in the EEG signals, e.g., eye blinks, the EEG signals should be preprocessed using artifact removal methods, e.g., blind source separation and independent component analysis [33].

(**iv**) Feature extraction: To extract relevant emotion features from EEG signals, information about the EEG signals is explored, e.g., the EEG characteristics in the time, frequency, and spatial domains [9].

(**v**) Various classifiers can be used to classify the extracted features, e.g., Bayesian, support vector machines, decision trees, and deep learning classifiers [34]. Depending on whether the classifier was trained on user-dependent data, EEG emotion recognition can be also divided into user-dependent and user-independent tasks.

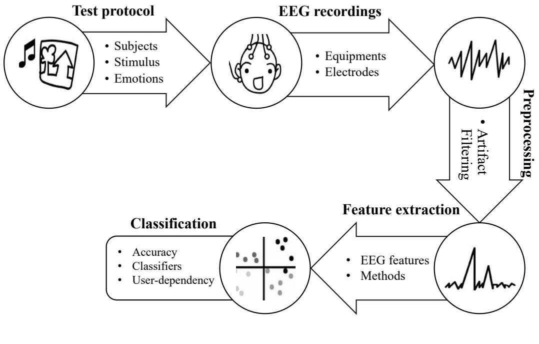


Fig. 1: The general flow of EEG emotion recognition.

2.2 Channel-wise Attention

Attention plays an important role in human perception [35], [36]. For example, humans can exploit a sequence of partial glimpses and selectively focus on salient parts to better capture a visual structure [37]. Inspired by the human attention mechanism, spatial attention mechanisms have been proposed for various vision tasks, e.g., semantic attention [38], multi-layer attention [39] and channel-wise attention [28]. Channel-wise attention demonstrates superior performance because it can change the weight of different channels to explore the information of a feature map; thus, it can extract more important information about channels. Therefore, the channel-wise attention mechanism has been used to exploit interdependencies among feature channels. For example, Hu *et al*. introduced a compact module to exploit the inter-channel relationship of a feature map [40], and Chen *et al*. combined spatial and channel-wise attention for image captioning [28].

Generally, channel-wise attention can squeeze the global spatial information and generate channel-wise statistics [28]. In addition, it is trainable with CNNs, thus, it can be integrated into CNN architectures [41]. Considering that multichannel EEG signals contains the spatial information via channels, channel-wise attention can be integrated into a CNN to explore the importance between the channels of EEG signals, and more discriminative spatial information can be extracted by a CNN.

2.3 Self Attention

Self-attention is an intra-attention mechanism that relates different positions of a single sequence to encode sequence data based on an importance score [20]. In addition, the self-attention mechanism is popular because it can improve long-range dependency modeling [42]. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. Self-attention has been demonstrated to perform well on simple-language question answering and language modeling tasks. For example, Vaswani *et al.* proposed an attention-based architecture for machine translation [32], and Shen *et al*. proposed the directional self-attention network to focus on the attention between elements in an input sequence [29]. In EEG recognition tasks, to augment the amount of training samples, one EEG trial is often segmented into several input samples. However, many methods ignore the importance of different EEG samples. Inspired by self-attention, we adopt this technique to further explore the time dependence between EEG samples.

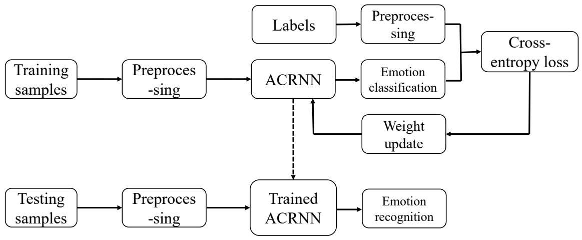


Fig. 2: Overview of the attention based convolutional recurrent neural network on EEG-based emotion recognition.

3 Proposed Methods

In this section, we first introduce the proposed EEG emotion recognition framework, and then we introduce our raw EEG signal preprocessing technique. Finally, we describe the construction of the proposed ACRNN in detail.

3.1 Framework of Proposed ACRNN

Generally, most EEG-based emotion recognition studies have focused on first extracting relevant features, and then the extracted features are used to classify the subjects emotional state [8], [14], [25]. In practice, raw EEG signals contain rich spatial and temporal information, which can be extracted to recognize a subjects emotional state. The proposed ACRNN is a data-driven method that integrates channel-wise [28] and extended self-attention [29] mechanisms into a CNN-RNN simultaneously. In addition, ACRNN can extract spatial and temporal information as emotion features, and classifies the extracted features using $\mathrm{softmax}$ function. Consequently, this end-to-end technique improves the accuracy of EEG based emotion recognition (Fig. 2). First, we divide the EEG samples into training and testing samples. Then, the training and testing samples are preprocessed by removing baseline signals, respectively. In addition, labels are preprocessed using the slicing window technique. Next, we use the training samples to train the proposed ACRNN model, compute the cross-entropy loss and update network parameters using the Adam optimizer [43]. Finally, the trained model is used to identify the emotional state of the testing samples, and classification accuracy are considered as the final recognition results.

3.2 Preprocessing of Proposed ACRNN

In the proposed ACRNN, preprocessing involves removing baseline signals and sliding windows. Generally, recorded EEG signals contain baseline and trial signals [2], [3]. Yang *et al.* proposed that baseline removal preprocessing can improve EEG emotion recognition on the DEAP database [7]. Here, let $\mathbf{X_R} = [\mathbf{X_B}, \mathbf{X_T}] \in \mathbb{R}^{M \times N}$ be the recorded EEG signals with H Hz sampling frequency and duration T_1, where M is the number of EEG electrode nodes, N is

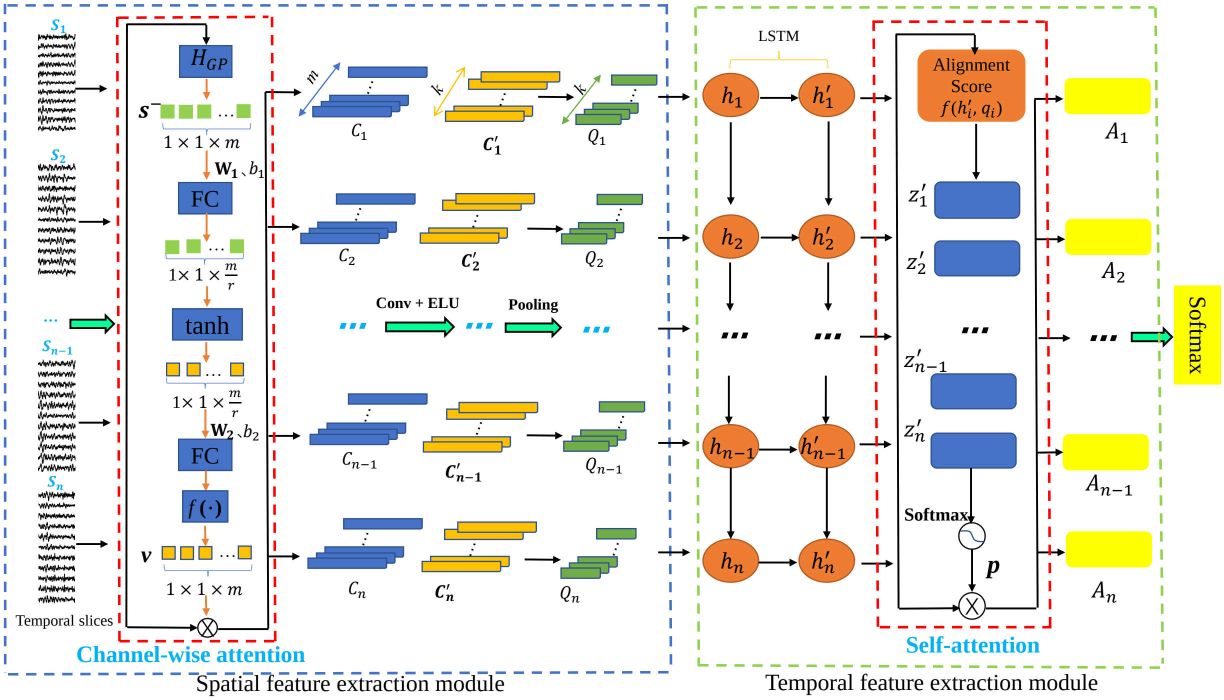


Fig. 3: The structure diagram of the attention based convolutional recurrent neural network.

the number of sampling points. In addition, $\mathbf{X_B} \in \mathbb{R}^{M\times L}$ denotes the baseline signal with duration T_2, L denotes the number of sampling points, X_i $(i = 1, 2, \ldots, T_2) \in \mathbb{R}^{M\times H}$ denotes the i-th second baseline signals. Thus, the mean value of baseline signal per-second can be formulated as follows:

$$\overline{\mathbf{X}}_\mathbf{B} = \frac{\sum_{i=1}^{T_2} \mathrm{X}_i}{T_2}, \tag{1}$$

where $\overline{\mathrm{X}}_\mathbf{B} \in \mathbb{R}^{M\times H}$ denotes the mean value of baseline signal per-second. Let $\mathbf{X_T} \in \mathbb{R}^{M\times J}$ represents the trial EEG signals with duration T_3, where J is the number of sampling points. To remove the baseline of trial EEG signals, $\mathbf{X_T}$ is segmented into several slices X_j $(j = 1, 2, \ldots, T_3) \in \mathbb{R}^{M\times H}$ with a one-second nonoverlapping sliding window, and the per-second baseline-removed signals can be formulated as follows:

$$\mathrm{X}'_j = \mathrm{X}_j - \overline{\mathrm{X}}_\mathbf{B}. \tag{2}$$

Finally, these baseline signals removed temporal slices X'_j are concatenated into a new matrix $\mathbf{X_T}' \in \mathbb{R}^{M\times J}$.

Generally, to augment the amount of training data, one EEG trial $\mathbf{X_T}' \in \mathbb{R}^{M\times J}$ is often segmented into several temporal slices $S = \{S_1, S_2, \ldots, S_n\}$ by sliding window. Here, S_i $(i = 1, 2, ..., n) \in \mathbb{R}^{M\times T}$ represents the i-th EEG sample, and T denotes the number of sampling points in each sliding window. Generally, a human emotional state lasts from 1 s to 12 s, and previous studies have shown that a 3-s sliding window can achieve good classification accuracy [44], thus, we employ a 3-s sliding window, *i.e.*, $T/H = 3$.

3.3 The Construction of Proposed ACRNN

The proposed ACRNN comprises the channel-wise attention mechanism, a CNN, an RNN and the extended self-attention mechanism. The structure of the proposed ACRNN is shown in Fig. 3. The left side of diagram shows the spatial feature extraction module. First, to explore the importance among the different channels of multichannel EEG signals, we employ the attention mechanism in a channel-wise manner into the EEG signals (Fig. 3). In actual EEG signal acquisition, different EEG channels in multichannel devices often contain redundant or less relevant information. To enhance emotion recognition accuracy, some methods adopt channel selection to choose more relevant channels [30]. Different from traditional methods that need select relevant channels artificially [31], we adopt the adaptive channel-wise mechanism, which can consider the information of all channels and assign weights to different channels based on importance. In our framework, $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ represents EEG samples after preprocessing, and $S_i = [s_1, s_2, \ldots, s_m]$ $(i = 1, 2, ..., m)$ is the i-th EEG sample, where s_j $(j = 1, 2, ..., m)$ represents the j-th channel of EEG sample S_i, and m is the total number of channels of each sample. In this model, we first apply mean pooling for each channel of EEG sample to obtain channel-wise statistics as follows:

$$\boldsymbol{s}^- = \left[s_1^-, s_2^-, \ldots, s_m^-\right], \tag{3}$$

where s_j^- $(j = 1, 2, ..., m)$ is the mean of the j-th channel. To reduce model complexity and improve generalizability, the channel-wise attention mechanism adopts two fully-connected (FC) layers around the non-linearity, *i.e.* a dimensionality-reduction layer with parameter $\mathbf{W_1}$ and bias terms b_1 with reduction ratio r and $\tanh$ function as the activation function, and a dimensionality increasing layer with parameter $\mathbf{W_2}$ and bias terms b_2. Thus, the gating mechanism of channel-wise attention is expressed as follows:

$$\boldsymbol{v} = \mathrm{softmax}\left(\mathbf{W_2} \cdot \left(\tanh\left(\mathbf{W_1} \cdot \boldsymbol{s}^- + b_1\right) + b_2\right)\right), \tag{4}$$

where the $\mathrm{softmax}$ function transforms the importance of channels to probability distribution $\boldsymbol{v} = [v_1, v_2, ..., v_m]$, which represents the importance of different channels. Finally, we consider probability as the weight to recode the information of the EEG sample $S_i = [s_1, s_2, ..., s_m]$ in each channel. Thus, the j-th (j = 1, 2, ..., m) attentive channel feature extracted via channel-wise attention can be represented as follows:

$$c_j = v_j \cdot s_j, \tag{5}$$

therefore, $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ represents the extracted channel-wise attentive features, the i-th extracted feature $C_i = [c_1, c_2, ..., c_m]$ can be obtained by channel-wise multiplication between each channel of $S_i = [s_1, s_2, ..., s_m]$ and each element of $\boldsymbol{v} = [v_1, v_2, ..., v_m]$.

Then, we use the CNN to further extract spatial information of EEG signals, where the number of convolution kernels is k, the kernel height is the same as the number of electrodes. Here, the kernel width is also designed to explore temporal information of the EEG signals. In addition, we use the exponential linear unit (ELU) function as the activation function in the convolution operations, which is better than the commonly used rectified linear unit (ReLU) function [45]. Thus, the i-th feature C_i' $(i = 1, 2, ..., n)$ can be obtained from the i-th channel attentive feature C_i after convolution and activation operations.

After that, we adopt a pooling layer to reduce the number of parameters and further extract features. Here, the i-th encoded representation after pooling is $\{Q_i | Q_i = \mathrm{MaxPool}\left(C_i'\right), i = 1 \ldots n\}$.

The right side of the structure diagram shows the temporal feature extraction module (Fig. 3), which comprises a two-layer LSTM and extended self-attention mechanism. The LSTM network can learn the context information of the sequence because it is based on a recurrent structure [27]. The LSTM network has been successfully used for EEG emotion recognition because it can learn features from EEG data based on temporal dependence [15]. As shown in Fig. 4, an LSTM cell receives three inputs, *i.e.*, input Q_i at the current time i, output c_{i-1} of previous time $i-1$, and h_{i-1} representing the hidden state of the previous time $i-1$. Then, the LSTM cell exports two outputs, *i.e.*, output c_i at the current time i and hidden state h_i represented as the i-th temporal feature extracted from LSTM. The LSTM cell contains three gates, *i.e.*, the input, forget and output gates to control the data flow by the sigmoid and tanh activation functions. Although here training samples and testing samples are different and not consecutive in time, the encoded samples contain spatial information after the spatial feature extraction module. Meanwhile, the input gate weight is organized to learn spatial information while the forget weight map organizes to learn more temporal information, the input gate and forget gate compete with each other to input new information into the cell or keep the current temporal information, respectively [46]. Thus, the LSTM network can extract the spatiotemporal features.

In this study, the number of LSTM units in each layer is the same as the number of EEG samples, and the output in each time step can be considered as the temporal information extracted from each sample. Generally, the LSTM network adopts two stacked layers to remember and encode all scanned spatial and temporal areas [47], thus, we set the number of LSTM layers to two. Therefore, the i-th output of the LSTM network is the hidden states of the second recurrent layer $\{h_i' | h_i' = \mathrm{lstm}\left(Q_i\right), i = 1 \ldots n\}$.

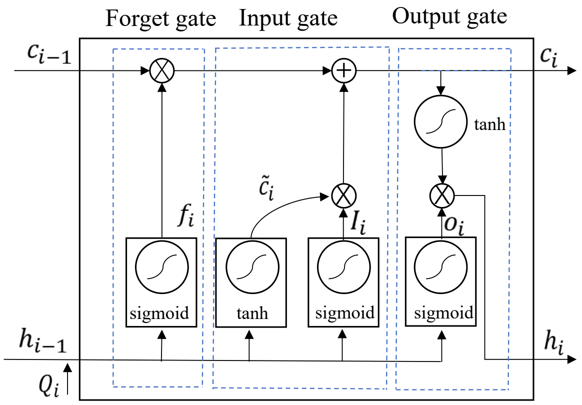


Fig. 4: LSTM unit architecture.

To extract more discriminative temporal information, we adopt the extended self-attention mechanism [29] to assign weights to each EEG signal sample by exploring the intrinsic importance of each sample. The structure of self-attention is shown in Fig. 5. Different from the traditional self-attention mechanism, which is used to assign importance to each recurrent encoded slice and aggregate this information to form a final representation [48], the extended attention is a natural extension of additive attention at the multi-dimensional feature level. It can better describe the specific meaning by computing the similarity within each sample from different points, and the obtained z_i' can be considered as a feature-wise score vector from the i-th sample h_i'. In addition, the extended self-attention mechanism adds two bias terms to the inside and outside of the activation function, and the i-th feature-wise score vector z_i' can be expressed as follows:

$$z_i' = f\left(h_i', q_i\right) = W^T \sigma\left(W_1 h_i' + W_2 q_i + b_1\right) + b, \tag{6}$$

where $f\left(h_i', q_i\right)$ represents the intrinsic similarity of the i-th encoded EEG sample, and q_i is the aligned pattern vector generated based on the feature vector h_i' by linear transformation, where the dimension is the same as the feature vector. Here, the activation functions $\sigma(\cdot)$ is an exponential linear unit (ELU), W and b are the weight and bias terms of σ function, respectively, W_1, W_2 are weight parameters, and b_1 is the bias terms. Then, $\boldsymbol{p} = \{p_1, p_2, ..., p_n\}$ denotes the probabilities of all samples, and the probability of the i-th EEG sample can be expressed as follows:

$$p_i = \frac{\exp\left(z_i'^T \cdot h_i'\right)}{\sum_{i=1}^{n} \exp\left(z_i'^T \cdot h_i'\right)}. \tag{7}$$

Lastly, $A = \{A_1, A_2, \ldots, A_n\}$ denotes the features extracted by the extended self-attention mechanism, and the i-th attentive feature extracted by the extended self-attention mechanism can be computed as follows:

$$A_i = p_i \cdot h_i'. \tag{8}$$

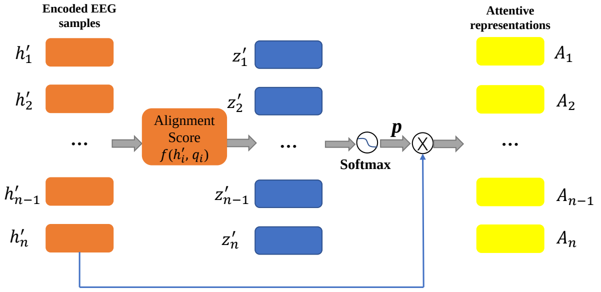


Fig. 5: Extended self-attention architecture.

In the last part of the proposed ACRNN, we employ the softmax layer as the classifier. The extracted spatiotemporal attentive features are $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$, and the softmax classifier receives these extracted features as input to recognize emotion as follows:

$$\boldsymbol{P} = \mathrm{softmax}(W\boldsymbol{A} + b), \tag{9}$$

where $\boldsymbol{P} = \{P_1, P_2, \ldots, P_n\}$, P_i $(i = 1, 2, ..., n)$ represents the predicted probability of the i-th EEG sample, and W and b are the weight and bias terms of the softmax function, respectively. Then, the cross-entropy error over all labeled samples is evaluated:

$$\mathcal{L} = -\sum_{i=1}^{n} \hat{Y}_i \log\left(P_i\right), \tag{10}$$

where $\hat{Y}_i$ is the label of the i-th EEG sample, and the lower cross-entropy error $\mathcal{L}$ indicates higher emotion recognition accuracy.

In summary, we have designed a framework to extract features and classify emotion from raw EEG signals. We adopt a channel-wise attention mechanism to adaptively assign the weights of different channels, which can extract the intrinsic information among channels, then employ a CNN to extract the spatial information of encoded EEG signals. In addition, we adopt two-layer LSTM to explore the temporal information of different EEG samples, and we integrate the extended self-attention mechanism to assign weight to EEG samples based on the importance of each sample. Finally, spatiotemporal attentive features can be obtained for EEG emotion recognition.

4 Experiments

Here, we introduce two widely used databases. Then, we introduce six deep learning methods and two traditional methods for comparison. We then demonstrate the model implementation in our experiments. Finally, we present and compare the experimental results obtained by the proposed method and compared methods.

4.1 Data Materials

To validate the performance of the proposed ACRNN, we conducted experiments on two widely used databases, *i.e.*, the database for emotion analysis using physiological signals (DEAP) [2] and database for emotion recognition through EEG and ECG signals (DREAMER) [3]. The DEAP database includes the EEG and peripheral physiological signals of 32 participants recorded while they subjects watched 40 pieces of music videos. The database contains 32-channel EEG signals and eight-channel peripheral physiological signals, where the EEG signals are used for emotional recognition, and the peripheral physiological signals are abnegated. In this experiment, the EEG signals were sampled at 512 Hz and then downsampled to 128 Hz. In addition, electrooculography (EOG) artifacts were removed using the blind source separation technique. The preprocessed EEG data of each trial contain 60-s trial data and 3-s baseline data. The emotional music videos include 40 one-minute clips, and the participants were asked to record their levels of arousal, valence, liking, and dominance for each video from 1 to 9. In our experiment, we selected valence and arousal as the emotional evaluation criteria, and the threshold to divide trials into two classes according to the rated levels of arousal and valence was set to five. Each subject file contained two arrays, and the data format of these files is detailed in Table 1.

TABLE 1
DEAP DATABASE

Array name	Array shape	Array centents
Data	40 ×40×8064	video/trial×channel×data
Labels	40×4	video/trial×label

The second database is DREAMER (Table 2), which is a multimodal database of EEG and ECG signals recorded during affect elicitation by means of audio visual stimuli. Signals from 23 participants (14 males and 9 females) were recorded, and participants were asked to record the levels of arousal, valence, and dominance after each stimuli. The EEG signals were recorded at a sampling rate of 128 Hz using an emotive EPOC system [49]. Each film clip is 65 to 393 s, which is sufficient to elicit single emotions. Moreover, the recorded EEG signals contain baseline signals and typically last 4 s before each film clip. In addition, most ocular artifacts (eye blinking, eye movement, cardiac interferences, etc.) have been removed with linear phase FIR filters. Furthermore, to avoid contaminating the data with multiple emotions, the recordings captured during the last 180 s of each clip were used for further analysis. The threshold of rating values is placed in the middle, where values less than or equal to 3 represent low valence, arousal, and dominance, and values greater than 3 represent high valence, arousal, and dominance.

4.2 Model Implementation

After the preprocessing stage, we obtained a total of EEG 800 samples for each subject in DEAP database, where each sample is S_i $(i = 1, 2, ..., 800) \in \mathbb{R}^{32\times384}$. For DREAMER, we obtained a total of EEG 1250 samples for each subject, where each sample is S_i $(i = 1, 2, ..., 1250) \in \mathbb{R}^{14\times384}$. We shuffled all samples from different trials for each subject. Then, we used 10-fold cross-validation to evaluate the performance of the proposed and baseline methods. The average performance of the 10-fold validation process was taken as the final experimental results. The model was implemented with the TensorFlow framework and trained on an NVIDIA TITAN Xp pascal GPU. In addition, the Adam optimizer was employed to minimize the cross-entropy loss function, and the network parameters were optimized with a learning rate of 10^{-4}, and the dropout regularization was set to 0.5. Batch normalization was adopted to achieve better performance during training. The size of the convolution kernel was $a \times b$, the height was set a = 32 for DEAP and a = 14 for DREAMER, and the width was b = 40. The number of kernels was k = 40 and the pooling size was 1 × 75 with a stride of 10. In addition, we set the dimension of the hidden state in LSTM to 64.

4.3 Results and Analysis

To validate the effectiveness of the proposed method, we conducted extensive experiments on two databases. To validate the performance of the attention mechanisms, we designed three models to demonstrate the influence of the channel-wise attention and extended self-attention mechanisms, including CNN + RNN (CNN-RNN), Channel-wise attention mechanism + CNN + RNN (A-CNN-RNN) and CNN + RNN + extended self-attention mechanism (CNN-RNN-A). Details for these models are shown in Table 3. The CNN-RNN model comprised a CNN and LSTM network, and was designed to validate the effectiveness of the baseline framework, which can extract emotional features from raw EEG signals using a cascade framework. The A-CNN-RNN model comprised the channel-wise attention mechanism, a CNN and LSTM network, and was designed to validate the effectiveness of the channel-wise attention mechanism for the baseline framework. The CNN-RNN-A model comprised a CNN, LSTM network and the extended self-attention, and was designed to validate the effectiveness of extended self-attention mechanisms for the baseline framework. In addition, we compared the proposed method with three recent deep learning methods: continuous convolutional neural network (Conti-CNN) [24], a graph convolutional neural network (GCNN) [25], and a convolutional recurrent attention model (CRAM) [48]. The Conti-CNN can combine the features of multiple bands to improve recognition accuracy [24], the GCNN can adopt different entropy (DE) feature as inputs, and use the spectral graph filtering to extract features and recognize emotion [25], and the CRAM can utilize a CNN to encode the high-level representation of EEG signals and a recurrent attention mechanism to explore the temporal dynamics [48]. In addition, we employed two traditional feature-based classifiers for comparison, including support vector machine (SVM) and decision tree (DT) [24]. All methods were processed by the same preprocessing as ACRNN, *i.e.*, baseline signal removal and sliding windows.

For the traditional classifiers, we used DE features as inputs [17], [18]. DE feature has the balance ability of discriminating EEG pattern between low and high frequency energy, which is typically used as frequency-domain features in EEG emotion recognition [14], [47], [50]. According to the literatures [18] and [51], the band-pass filter is applied to EEG signals to obtain the sub-band signals, which approximately follow a Gaussian distribution. Consequently, five sub-bands were defined: 1) delta (1-3 Hz); 2) theta (4-7 Hz); 3) alpha (8-13 Hz); 4) beta (14-30 Hz); and 5) gamma (31-50 Hz). Note that we extracted DE features from the laterer four sub-band signals because the higher-frequency band (approximately 30-100 Hz) is more suitable for EEG emotion recognition [13]. The final feature vector was a concatenation of features from all channels. For DEAP, the final feature vector was 4 × 32 = 128 dimensions, and each subject yielded 800 samples, where each sample $\mathbf{X}_i \in \mathbb{R}^{32\times128}$ (i = 1, 2, ..., 800). For DREAMER, the final feature vector was 4 × 14 = 56 dimensions, and each subject yielded 1250 samples, where each sample $\mathbf{X}_i \in \mathbb{R}^{14\times56}$ (i = 1, 2, ..., 1250).

TABLE 2
DREAMER DATABASE

Audio-visual stimuli	
Number of videos	18
Video content	Audio-Video
Video duration	65-393 s (M = 199 s)
Experiment information	
Number of participants	23
Number of males	14
Number of females	9
Age of participants	22-33
Rating scales	Arousal, Valence, Dominance
Rating values	1-5
Recorded signals	14-channel 128 Hz EEG

TABLE 3
BASELINE MODEL AND ATTENTION-BASED MODELS FOR EEG EMOTION RECOGNITION

Component Model	channel-wise attention	CNN	LSTM network	self-attention
CNN-RNN	×	✓	✓	×
A-CNN-RNN	✓	✓	✓	×
CNN-RNN-A	×	✓	✓	✓
ACRNN	✓	✓	✓	✓

In our work, we conducted the experiments on the same subject with the proposed ACRNN and compared methods for subject-dependent EEG emotion recognition. We divided the sample data into training sets and test sets, and then used 10-fold cross validation [52]. Typically, 10-fold cross validation divides data into 10 equal data subsets, and one subset is used as the test set, and the other nine subsets form the training set. This process was repeated 10 times. For the DEAP database, the number of training samples was 720, and the remaining 80 samples for each subject were used as test samples. For the DREAMER database, the numbers of training and test samples were 1125 and 125, respectively.

To further analyze the contribution of channel-wise attention, we performed experiments to compute the channel

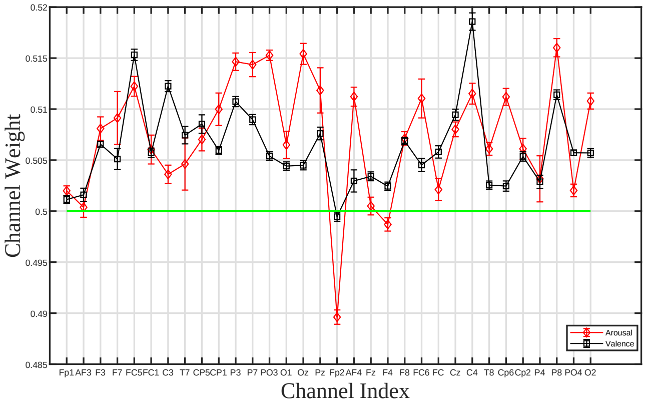


Fig. 6: The average channel weight on DEAP database.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3025777, IEEE Transactions on Affective Computing

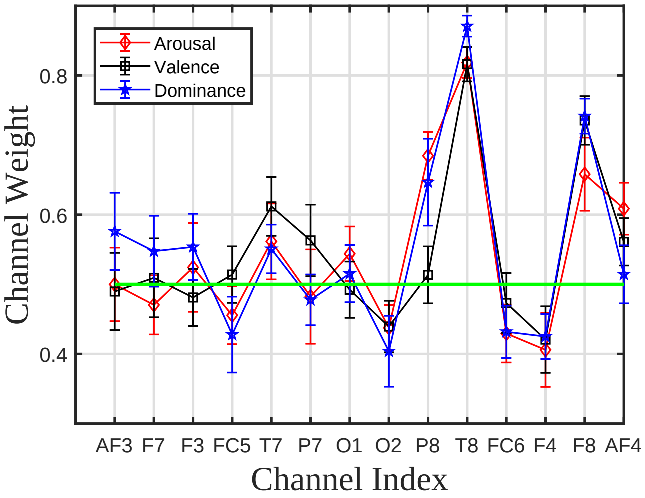


Fig. 7: The average channel weight on DREAMER database.

TABLE 4
The relationship between greater weight channels and brain region in DEAP and DREAMER

Database	Channel	Brain region
DEAP	F3	Frontal
	FC5	Frontal
	P3	Parietal
	P7	Parietal
	F8	Frontal
	Cz	Central
	C4	Central
	P8	Parietal
DREAMER	T8	Temporal
	F8	Frontal

weights of EEG signals by channel-wise attention. Figs. 6 and 7 show the average channel weights in DEAP and DREAMER, respectively. There are 32 channels in DEAP dataset and 14 channels in DREAMER dataset. As shown, the channel weights of EEG signals in DEAP and DREAMER are both different in channel-wise attention mechanism. Fig. 6 shows the channel weights of FC5, P3, C4, P8 are obviously greater than other channels on two dimensions in DEAP database. Fig. 7 shows the channel weights of T8 and F8 are obviously greater than other channels on three dimensions in DREMAER database. The electrodes are placed according to the international 10-20 system in two databases, the relationship between greater weight channels and brain region is as shown in Table 4. The results are also consistent with some studies, which has demonstrated the EEG signals related to emotions are mostly distributed in the frontal lobe, the temporal lobe and parietal lobe [30], [31], [53]. The greater channel weight also indicates the given channels are more related to emotions and more important in the EEG signals.

Table 5 shows the average recognition accuracies of the proposed and compared methods on valence and arousal in the DEAP database. As can be seen, the average recognition accuracy of A-CNN-RNN improved by approximately 30% and 25% compared to the baseline framework CNN-RNN on two dimensions because channel-wise attention focuses on the spatial features among different channels. In addition, the proposed ACRNN improved the average recognition accuracy by 0.6% and 0.5% compared to A-CNN-RNN because the ACRNN combines the channel-wise and extended self-attention mechanisms to extract the spatiotemporal attentive information of the EEG signals. In addition, we found that CNN-RNN-A improved recognition accuracy by approximately 27% on two dimensions compared to CNN-RNN because the extended self-attention mechanism of CNN-RNN-A extracts attentive information according to the importance of each sample. In addition, the proposed ACRNN improved recognition accuracy by approximately 3% on two dimensions compared to CNN-RNN-A because the proposed model exploits both attention mechanisms simul-taneously. Compared to the three deep learning methods (Conti-CNN, CRAM, and GCNN), the experimental results indicate that the proposed ACRNN improved the average recognition accuracy by 10%, 8% and 5%, respectively. Compared to traditional methods, the ACRNN achieved better recognition performance than the traditional methods, e.g., DT and the SVM.

Table 6 shows the average recognition accuracies of the compared methods obtained on the DREAMER database. As shown, emotion recognition accuracy improved significantly on the DREAMER database. For example, the proposed ACRNN improved the average recognition accuracy by 15%, 9% and 5% compared to Conti-CNN, GCNN and CRAM, respectively. Thus, the proposed ACRNN can achieve the best recognition accuracy among all compared methods, and the experimental results demonstrate the effectiveness of integrating the two attention mechanisms into the CNN-RNN.

TABLE 5
AVERAGE ACCURACIES (%) OF DIFFERENT METHODS ON THE VALENCE AND AROUSAL CLASSIFICATION TASKS OF DEAP DATABASE (MEAN ± STD. DEV.)

	Valence	Arousal
DT	75.95 ± 4.76	78.18 ± 5.45
SVM	89.33 ± 7.41	89.99 ± 6.74
Conti-CNN	82.77 ± 4.47	81.55 ± 6.55
CRAM	87.09 ± 7.49	84.46 ± 9.27
GCNN	88.24 ± 3.18	87.72 ± 3.32
CNN-RNN	62.75 ± 7.53	67.12 ± 9.13
A-CNN-RNN	91.48 ± 5.02	91.59 ± 5.42
CNN-RNN-A	89.15 ± 6.66	89.96 ± 5.93
ACRNN	**93.72 ± 3.21**	**93.38 ± 3.73**

TABLE 6
AVERAGE ACCURACIES (%) OF DIFFERENT METHODS ON THE VALENCE, AROUSAL AND DOMINANCE CLASSIFICATION TASKS OF DREAMER DATABASE (MEAN ± STD. DEVC.)

	Valence	Arousal	dominance
DT	68.81 ± 6.87	67.50 ± 7.28	67.43 ± 6.73
SVM	76.71 ± 5.89	77.54 ± 5.62	75.76 ± 5.63
Conti-CNN	81.72 ± 5.24	82.48 ± 5.11	82.58 ± 5.28
CRAM	92.27 ± 2.95	93.03 ± 1.87	93.34 ± 1.78
GCNN	88.87 ± 3.58	88.79 ± 3.86	88.54 ± 3.89
CNN-RNN	78.59 ± 13.87	77.66 ± 13.34	77.75 ± 14.22
A-CNN-RNN	97.47 ± 2.32	97.92 ± 1.60	98.15 ± 1.76
CNN-RNN-A	96.61 ± 3.42	97.36 ± 2.63	97.54 ± 2.16
ACRNN	**97.93± 1.73**	**97.98 ± 1.92**	**98.23 ± 1.42**

To demonstrate the performance of the proposed method and compared methods for each subject, we conducted experiments on each subject. Figs. 8, 9, 10, 11 and 12 show the average accuracy and standard deviation of each subject on each dimension. As can be seen, the traditional

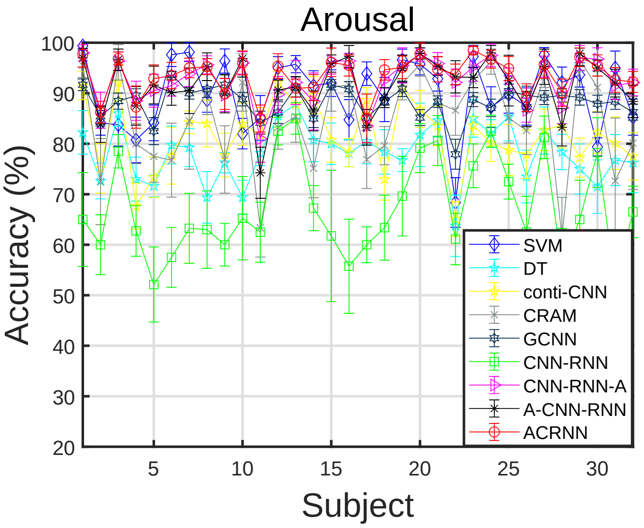


Fig. 8: Average accuracies (%) on each subject of different methods on arousal classification tasks on DEAP database.

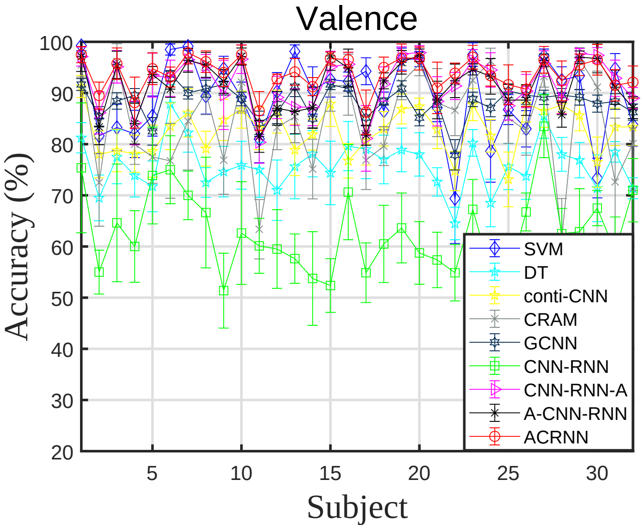


Fig. 9: Average accuracies (%) on each subject of different methods on valence classification tasks on DEAP database.

SVM and DT methods achieved good average recognition accuracy on some subjects, however, the standard deviations were very large. In addition, the compared methods performed worse on some subjects. However, we found that the three attention-based methods achieved better average recognition accuracy on each subject, and the standard deviations were less than those of the compared methods. Thus, the experimental results demonstrate that attention-based methods can work better than the compared methods for each subject. Furthermore, the results indicate that the proposed ACRNN combines channel-wise attention module and extended self-attention to exploit more discriminative information for EEG emotion recognition and can achieve superior recognition accuracy on two public databases.

5 Discussions

EEG-based emotion recognition is widely used to help computers better understand the current emotional state of

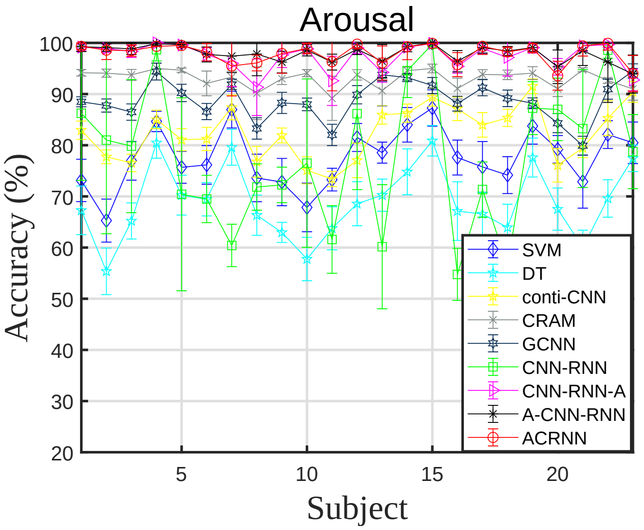


Fig. 10: Average accuracies (%) on each subject of different methods on arousal classification tasks on DREAMER database.

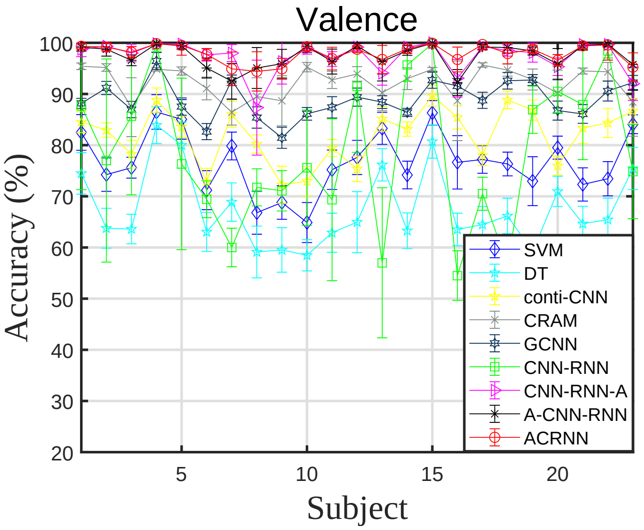


Fig. 11: Average accuracies (%) on each subject of different methods on valence classification tasks on DREAMER database.

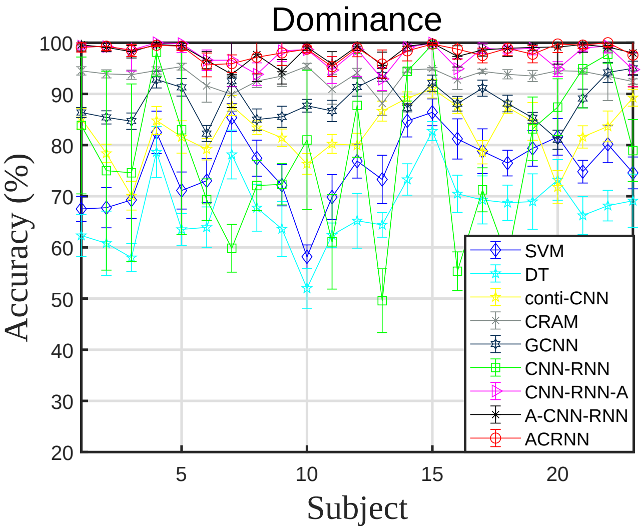


Fig. 12: Average accuracies (%) on each subject of different methods on dominance classification tasks on DREAMER database.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3025777, IEEE Transactions on Affective Computing

the user [21]. Traditional EEG recognition methods first design hand-crafted features from raw EEG signals, and then employ classifier to classify these features. Recently, deep learning methods have an employed end-to-end technique to recognize emotion from raw EEG signals. However, it still remains challenging to extract more discriminative features from raw signals for EEG emotion recognition. Thus, a discriminative framework to recognize the emotional state from the raw EEG signals is required. Considering that raw EEG signals contain spatial information by the intrinsic relationship between different channels and time dependencies among temporal slices, we have proposed ACRNN to extract spatial and temporal attentive information and classify the emotional state of a subject.

In this framework, a channel-wise attention mechanism extracts the difference among channels from the EEG signals by assigning the weights to different channels, and a CNN is designed to extract the feature map as spatial information by the convolution operation on all channels of the EEG signals. Different from some methods based on brain cognitive function, they need to focus on EEG channels and to design relevant features [54]–[56]. For example, Li et.al constructed emotion-related brain networks with phase locking value (PLV) and adopted a multiple feature fusion approach to combine the compensative activation and connection information for emotion recognition [55], Wang et.al combined brain directed connectivity (BDC) and DE features which are in different frequency bands among brain areas to extract discriminative information to improve recognition accuracy [56]. However, our proposed method adopt channel-wise attention to allocate weights in different channels. To further analyze the importance of different channels, we compute the average channel weights. From Figs. 6 and 7, we can find the channel weights of EEG signals in DEAP and DREAMER appear different according to the channel-wise attention mechanism. The channel weights of FC5, P3, C4, P8 are clearly greater than the other channels on two dimensions in DEAP database, and the channel weights of T8 and F8 are obviously greater than the other channels on three dimensions in DREMAER database. The result demonstrates that the EEG signals relevant to emotions are mostly distributed in the frontal lobe, the temporal lobe and the parietal lobe, which is consistent with the existing studies [30], [31], [53]. It can also be seen that the channels with computed greater weights are more related to emotions and thus more important in EEG-based emotion recognition. To demonstrate the effectiveness of the channel-wise attention mechanism, we integrate channel-wise attention into the baseline CNN-RNN framework, and the experimental results demonstrate that the channel-wise attention of the A-CNN-RNN can improve the average accuracy by approximately 30% compared to the CNN-RNN model on the DEAP and DREAMER databases because the channel-wise attention mechanism can transform channels to a probability distribution as weights and recode the EEG signals based on the transformed weights. In addition, the extended self-attention mechanism is designed to explore the importance of different EEG samples. To demonstrate the effectiveness of the extended self-attention mechanism, we integrated it into the baseline CNN-RNN framework, and the experimental results show that extended self-attention can improve average accuracy by 27% and 29% compared to the CNN-RNN model on these databases. The experimental results also demonstrate that the extended self-attention mechanism focuses on more important EEG samples by scoring the probability based on the similarities among samples.

In summary, the channel-wise attention and extended self-attention mechanisms improve the average accuracy greater than 25% on two both databases. This indicates that these attention mechanisms can improve EEG emotion recognition and achieve comparable recognition results. However, compared to extended self-attention, channel-wise attention improved the average recognition accuracy by approximately 2% and 1% on the DEAP and DREAMER databases, respectively. This indicates that channel-wise attention performs slightly better than extended self-attention. In summary, the proposed ACRNN is a cascade framework that integrates channel-wise attention and extended self-attention mechanisms. It can effectively extract spatiotemporal attentive features simultaneously. In addition, our all experimental results are obtained by 10-fold cross-validation, the high accuracy and low standard deviation have also demonstrated that the proposed ACRNN can achieve superior recognition accuracy.

6 Conclusion

In this paper, we have proposed an end-to-end deep learning method for EEG emotion recognition. The proposed ACRNN takes the spatial information, temporal information and attentive information of EEG signals into consideration. In addition, we integrate channel-wise attention into a CNN, which can extract spatial attentive features, and channel-wise attention can extract the attentive information among the channels. We also integrated extended self-attention into RNN, which can extract attentive information based on the importance of each sample. Finally, extensive experimental results have demonstrated that the proposed ACRNN achieved average accuracies of 93.72% and 93.38% on the valence and arousal classification tasks in the DEAP database, respectively. In addition, the proposed ACRNN achieved average accuracies of 97.93%, 97.78% and 98.23% on valence, arousal, and dominance classification tasks in the DREAMER database, respectively. Compared to existing methods, it is clear that the proposed ACRNN improved EEG emotion recognition accuracy in the DEAP and DREAMER databases. In future work, we will study the trial-based and inter-subjects EEG emotion recognition based on the attention mechanism.

References

[1] R. W. Picard, *Affective Computing*, 1997.

[2] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[3] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[4] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel eeg via deep forest," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[5] Y. Liu, Y. Ding, C. Li, J. Cheng, R. Song, F. Wan, and X. Chen, "Multi-channel eeg-based emotion recognition via a multi-level features guided capsule network," *Computers in Biology and Medicine*, vol. 123, p. 103927, 2020.

[6] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Eeg-based emotion recognition approach for e-healthcare applications," in *2016 eighth international conference on ubiquitous and future networks (ICUFN)*. IEEE, 2016, pp. 946–950.

[7] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[8] S. M. Alarco and M. J. Fonseca, "Emotions recognition using eeg signals: A survey," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[9] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 327–339, 2014.

[10] P. Ackermann, C. Kohlschein, J. Á. Bitsch, K. Wehrle, and S. Jeschke, "Eeg-based automatic emotion recognition: Feature extraction, selection and classification methods," in *2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom)*. IEEE, 2016, pp. 1–6.

[11] N. Sulthan, N. Mohan, K. A. Khan, S. Sofiya, and M. S. PP, "Emotion recognition using brain signals," in *2018 International Conference on Intelligent Circuits and Systems (ICICS)*. IEEE, 2018, pp. 315–319.

[12] C. Li, W. Tao, J. Cheng, Y. Liu, and X. Chen, "Robust multichannel eeg compressed sensing in the presence of mixed noise," *IEEE Sensors Journal*, vol. 19, no. 22, pp. 10 574–10 583, 2019.

[13] M. Li and B.-L. Lu, "Emotion classification based on gamma-band eeg," in *2009 Annual International Conference of the IEEE Engineering in medicine and biology society*. IEEE, 2009, pp. 1223–1226.

[14] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel eeg data through convolutional recurrent neural network," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2016, pp. 352–359.

[15] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on eeg using lstm recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.

[16] A. Patil, C. Deshmukh, and A. Panat, "Feature extraction of eeg for emotion recognition using hjorth features and higher order crossings," in *2016 Conference on Advances in Signal Processing (CASP)*. IEEE, 2016, pp. 429–434.

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

[18] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for eeg-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[19] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[20] D. Hu, "An introductory survey on attention mechanisms in nlp problems," in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2019, pp. 432–448.

[21] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (eeg) classification tasks: a review," *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.

[22] J. Li, Z. Struzik, L. Zhang, and A. Cichocki, "Feature learning from incomplete eeg with denoising autoencoder," *Neurocomputing*, vol. 165, pp. 23–31, 2015.

[23] S. K. Goh, H. A. Abbass, K. C. Tan, A. Al-Mamun, N. Thakor, A. Bezerianos, and J. Li, "Spatio–spectral representation learning for electroencephalographic gait-pattern classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 9, pp. 1858–1867, 2018.

[24] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3d input for eeg-based emotion recognition," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 433–443.

[25] T. Song, W. Zheng, S. Peng, and C. Zhen, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2018.

[26] 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," *arXiv preprint arXiv:1708.06578*, 2017.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[29] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for rnn/cnn-free language understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] M. S. Özerdem and H. Polat, "Emotion recognition based on eeg features in movie clips with channel selection," *Brain informatics*, vol. 4, no. 4, p. 241, 2017.

[31] L. Tong, J. Zhao, and W. Fu, "Emotion recognition and channel selection based on eeg signal," in *2018 11th International Conference on Intelligent Computation Technology and Automation (ICICTA)*. IEEE, 2018, pp. 101–105.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[33] X. Chen, X. Xu, A. Liu, S. Lee, X. Chena, X. Zhang, M. J. McKeown, and Z. J. Wang, "Removal of muscle artifacts from the eeg: a review and recommendations," *IEEE Sensors Journal*, 2019.

[34] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from eeg data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[35] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[36] R. A. Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[37] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *Advances in neural information processing systems*, 2010, pp. 1243–1251.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[39] P. H. Seo, Z. Lin, S. Cohen, X. Shen, and B. Han, "Hierarchical attention networks," *arXiv preprint arXiv:1606.02393*, vol. 2, 2016.

[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[41] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[42] M. Daniluk, T. Rocktäschel, J. Welbl, and S. Riedel, "Frustratingly short attention spans in neural language modeling," *arXiv preprint arXiv:1702.04521*, 2017.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] Y. Li, J. Huang, H. Zhou, and N. Zhong, "Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks," *Applied Sciences*, vol. 7, no. 10, p. 1060, 2017.

[45] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[46] S. Goldstein, Z. Hu, and M. Ding, "Decoding working memory load from eeg with lstm networks," *arXiv preprint arXiv:1910.05621*, 2019.

[47] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 839–847, 2018.

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

[49] N. A. Badcock, P. Mousikou, Y. Mahajan, P. De Lissa, J. Thie, and G. McArthur, "Validation of the emotiv epoc® eeg gaming system for measuring research quality auditory erps," *PeerJ*, vol. 1, p. e38, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3025777, IEEE Transactions on Affective Computing

[50] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[51] W.-L. Zheng, Y.-Q. Zhang, J.-Y. Zhu, and B.-L. Lu, "Transfer components between subjects for eeg-based emotion recognition," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 917–922.

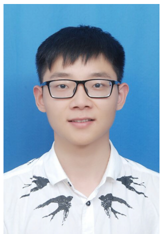
[52] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[53] Z.-M. Wang, S.-Y. Hu, and H. Song, "Channel selection method for eeg emotion recognition using normalized mutual information," *IEEE Access*, vol. 7, pp. 143 303–143 311, 2019.

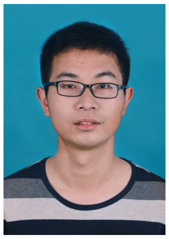
[54] X. Liu, T. Li, C. Tang, T. Xu, P. Chen, A. Bezerianos, and H. Wang, "Emotion recognition and dynamic functional connectivity analysis based on eeg," *IEEE Access*, vol. 7, pp. 143 293–143 302, 2019.

[55] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, and Y. a. Zhang, "Eeg based emotion recognition by combining functional connectivity network and local activations," *IEEE Transactions on Biomedical Engineering*, pp. 1–1, 2019.

[56] H. Wang, X. wu, and L. Yao, "Identifying cortical brain directed connectivity networks from high-density eeg for emotion recognition," *IEEE Transactions on Affective Computing*, vol. PP, pp. 1–1, 07 2020.



Wei Tao received the B.S. degree in Measurement, Control technology and Instrumentation from Hefei University of Technology, China, in 2017, and the M.E. degree in Biomedical instrument from Hefei University of Technology, China, in 2020. He is currently working towards his Ph.D. degree in the Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Macau. His research interests include EEG-based emotion recognition, compressed sensing, and machine learning.

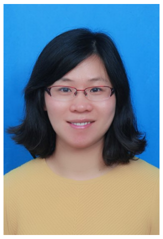


Chang Li received the B.S. degree in information and computing science from the Wuhan Institute of Technology, Wuhan, China, in 2012, and the Ph.D. degree in circuits and systems from the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, in 2018.

He is currently a Lecturer with the Department of Biomedical Engineering, Hefei University of Technology, Hefei, China. His research interests include biomedical signal processing, hyperspectral image analysis, computer vision, pattern recognition, and machine learning.

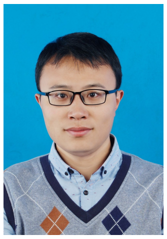
Rencheng Song received his B.S. degrees in Mathematics from Jilin University, Changchun, China, in 2005, and the Ph.D. degree from Zhejiang University, Hangzhou, China in 2010.

He is currently an associate professor with the Department of Biomedical Engineering at Hefei University of Technology. His research interests include biomedical signal processing, non-contact physiological parameter measurement, machine learning and electromagnetic imaging.



Juan Cheng received her B.S. degree and Ph. D degree from the Department of Electronic Science and Technology, University of Science and Technology of China (USTC) in 2008 and 2013, respectively.

She is currently an associated professor with the Department of Biomedical Engineering, Hefei University of Technology (HFUT), Hefei, China. Her research interests include biomedical signal/image processing, non-contact physiological parameter measurement, and machine learning.

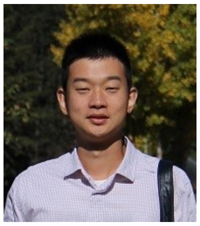


Yu Liu received his B.S. degree and Ph. D degree from the Department of Automation, University of Science and Technology of China in 2011 and 2016, respectively.

He is currently an assistant professor in the Department of Biomedical Engineering at Hefei University of Technology. His research interests include image/signal processing, computer vision, information fusion and machine learning.

Feng Wan received the PhD degree in Electrical and Electronic Engineering from the Hong Kong University of Science and Technology, Hong Kong.

He is currently an Associate Professor in the Department of Electrical and Computer Engineering, Faculty of Science and Technology, and also a Primary Faculty of the Centre for Cognitive and Brain Sciences, Institute of Collaborative Innovation, University of Macau, Macau. His research interests include braincomputer interfaces, biomedical signal processing, neuroimaging and neurofeedback training, deep and transfer learning, computational intelligence and intelligent control.



Xun Chen received the B.S. degree in the Department of Electronic Science and Technology at the University of Science and Technology of China (USTC) in 2009, and received the Ph.D degree in the Department of Electrical and Computer Engineering at the University of British Columbia (UBC) in 2014.

He is with the Department of Electronic Science and Technology at USTC as a professor. His research interests include the broad areas of statistical signal processing and machine learning in biomedical applications. He has published over 100 refereed scientific papers. He is serving as associate editors for Signal Processing-Image Communication, IEEE Signal Processing Letters, Frontiers in Neuroscience and IEEE Access.