Cognitive Neurodynamics (2022) 16:805–818
https://doi.org/10.1007/s11571-021-09751-5

RESEARCH ARTICLE

4D attention-based neural network for EEG emotion recognition

Guowen Xiao[1] · Meng Shi[1] · Mengwen Ye[2] · Bowen Xu[1] · Zhendi Chen[1] · Quansheng Ren[1]

Received: 1 February 2021 / Revised: 11 October 2021 / Accepted: 29 October 2021 / Published online: 3 January 2022

Abstract

Electroencephalograph (EEG) emotion recognition is a significant task in the brain-computer interface field. Although many deep learning methods are proposed recently, it is still challenging to make full use of the information contained in different domains of EEG signals. In this paper, we present a novel method, called four-dimensional attention-based neural network (4D-aNN) for EEG emotion recognition. First, raw EEG signals are transformed into 4D spatial-spectral-temporal representations. Then, the proposed 4D-aNN adopts spectral and spatial attention mechanisms to adaptively assign the weights of different brain regions and frequency bands, and a convolutional neural network (CNN) is utilized to deal with the spectral and spatial information of the 4D representations. Moreover, a temporal attention mechanism is integrated into a bidirectional Long Short-Term Memory (LSTM) to explore temporal dependencies of the 4D representations. Our model achieves state-of-the-art performances on both DEAP, SEED and SEED-IV datasets under intra-subject splitting. The experimental results have shown the effectiveness of the attention mechanisms in different domains for EEG emotion recognition.

Keywords EEG · Emotion recognition · Attention mechanism · Convolutional recurrent neural network

Introduction

Emotion plays an important role in daily life and is closely related to human behavior and cognition (Dolan 2002). As one of the most significant research topics of affective computing, emotion recognition has received increasing attention in recent years for its applications of disease detection (Bamdad et al. 2015; Figueiredo et al. 2019), human–computer interaction (Fiorinia et al. 2020; Katsigiannis and Ramzan 2017), and workload estimation (Blankertz et al. 2016). In general, emotion recognition methods can be divided into two categories (Mühl et al. 2014). One is based on external emotion responses including facial expressions and gestures(Yan et al. 2016), and the other is based on internal emotion responses including electroencephalograph (EEG) and electrocardiography (ECG) (Zheng et al. 2017). Neuroscientific researches have shown that some major brain cortex regions are closely related to emotions, making it possible to decode emotions based on EEG (Brittona et al. 2006; Lotfia and Akbarzadeh-T 2014). EEG is non-invasive, portable, and inexpensive so that it has been widely used in the field of brain-computer interfaces (BCIs) (Pfurtscheller et al. 2010). Besides, EEG signals contain various spatial, spectral, and temporal information about emotions evoked by specific stimulation patterns. Therefore, more and more researchers concentrate on EEG emotion recognition recently (Alhagry et al. 2017; Li and Lu 2009).

Traditional EEG emotion recognition methods usually extract hand-crafted features from EEG signals first and then adopt shallow models to classify the emotion features. EEG emotion features can be extracted from the time domain, frequency domain, and time–frequency domain. Jenke et al. conduct a comprehensive survey on EEG feature extraction methods by using machine learning techniques on a self-recorded dataset (Jenke et al. 2014). For classifying the extracted emotion features, many researchers have adopted machine learning methods over the past few years (Kim et al. 2013). Li et al. apply a linear support vector machine (SVM) to classify emotion features extracted from the gamma frequency band (Li and Lu

✉ Quansheng Ren
qsren@pku.edu.cn

[1] Department of Electronics, Peking University, Beijing, China

[2] School of Electrical Engineering, Beijing Jiaotong University, Beijing, China

2009). Duan et al. extract differential entropy (DE) features, which are superior to representing emotion states in EEG signals (Shi et al. 2013), from multichannel EEG data and combine a k-Nearest Neighbor (KNN) with SVM to classify the DE features (Duan et al. 2013). However, shallow models require lots of expert knowledge to design and select emotion features, limiting their performance on EEG emotion classification.

Deep learning methods have been demonstrated to outperform traditional machine learning methods in many fields such as computer vision, natural language processing, and biomedical signal processing (Craik et al. 2019; Goh et al. 2018) for the ability to learn high-level features from data automatically (Krizhevsky et al. 2012). Recently, some researchers have applied deep learning to EEG emotion recognition. Zheng et al. introduce a deep belief network (DBN) to investigate the critical frequency bands and EEG signal channels for EEG emotion recognition (Zheng and Lu 2015). Yang et al. propose a hierarchical network to classify the DE features extracted from different frequency bands (Yang et al. 2018c). Song et al. use a graph convolutional neural network to classify the DE features (Song et al. 2020). Ma et al. propose a multimodal residual Long Short-Term Memory model (MMResLSTM) for emotion recognition, which shares temporal weights across the multiple modalities (Jiaxin Ma et al. 2019). To learn the bi-hemispheric discrepancy for EEG emotion recognition, Yang et al. propose a novel bi-hemispheric discrepancy model (BiHDM) (Li et al. 2020). Although all those deep learning methods outperform the shallow models, it is still challenging to fuse more important information on different domains and capture discriminative local patterns in EEG signals.

Inspired by the convolutional recurrent neural network and attention mechanisms which have been introduced to process EEG signals (Shen et al. 2020; Tao et al. 2020; Jia et al. 2020), first, we transform raw EEG signals into 4D spatial-spectral-temporal representations which consist of several temporal slices and contain information on different domains. Then, we use a CNN to extract spatial and spectral information of 4D representations, and a bidirectional LSTM to extract temporal information of 4D representations. Generally, the critical brain regions and frequency bands in the brain activities could vary with different subjects, time, and emotions. However, traditional CNNs and LSTMs ignore the importance of different spatial positions, frequency bands, and temporal slices of 4D representations. For instance, traditional LSTMs only use the output of the last hidden unit for subsequent processing, but there exist complex temporal dependencies between outputs of all hidden units in fact, which might be helpful for emotion recognition. To adaptively capture discriminative information on different domains, we employ attention mechanisms on both the CNN and the bidirectional LSTM. For the CNN, attention mechanisms are applied to the spatial and spectral dimensions of each temporal slice after the convolutional layers, so that different weights are assigned to different brain regions and frequency bands. For the bidirectional LSTM, we use self-attention mechanism (Vaswani et al. 2017) to utilize long-range temporal dependencies of different temporal slices. Because of these two factors, information that is helpful to recognize different emotions is strengthened, while the opposite is weakened by attention mechanisms. Therefore, more discriminative information on different domains can be automatically obtained, instead of artificially selected.

The primary contribution of this paper are summarized as follows: a) We propose a four-dimensional attention-based neural network (4D-aNN), which fuses information on different domains and captures discriminative patterns in EEG signals based on the 4D spatial-spectral-temporal representation. b) We conduct experiments on DEAP and SEED datasets, and the experimental results indicate average emotion recognition accuracies of 96.90% and 97.39% in the valence and arousal classification tasks of DEAP dataset. Besides, 4D-aNN achieves a mean accuracy of 96.25% in the classification task of SEED dataset.

The remainder of this paper is organized as follows. We describe our proposed method in the *Method* section. Dataset, experiment settings, results, ablation studies, and discussion are presented in the *Experiment* section. Finally, conclusions are given in the *Conclusion* section.

Method

The overall structure of 4D-aNN for EEG emotion recognition consists of the 4D spatial-spectral-temporal representation, the attention-based CNN, the attention-based bidirectional LSTM, and the classifier. We will describe the details of each part in sequence.

4D spatial-spectral-temporal representation

To generate 4D spatial-spectral-temporal representations from original multi-channel EEG signals, we first split original EEG signals into T seconds long segments without overlapping (for an nT seconds long data, we cut it into n segments of T seconds long.) as previous works do (Shen et al. 2020; Yang et al. 2018a). Each segment is assigned with the same label as the original EEG signals. Then, we decompose each segment into several frequency bands (i.e. $\delta[1 \sim 4$ Hz], $\theta[4 \sim 8$ Hz], $\alpha[8 \sim 14$ HZ], $\beta[14 \sim 31$ Hz], and $\gamma[31 \sim 51$ Hz]) with Butterworth bandpass filters. The Differential Entropy (DE) features of all EEG channels, which have been proven to be effective for emotion recognition (Zheng et al. 2017), are extracted from different frequency bands respectively with a 0.5 s window for each segment.

DE feature is capable of discriminating EEG patterns between low and high frequency energy, which is defined as

$$h_D(X) = -\int_X f(x)\log(f(x))dx \quad (1)$$

where x is formally a random variable and in this context, the signal acquired from a certain frequency band on a certain EEG channel, $f(x)$ is the probability density function of x. If x obeys the Gaussian distribution $N(\mu, \sigma^2)$, DE feature can simply be calculated by the following formulation:

$$h_D(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\sigma^2}}\exp\frac{(x-\mu)^2}{2\sigma^2}\log\frac{1}{\sqrt{2\sigma^2}}\exp\frac{(x-\mu)^2}{2\sigma^2}dx$$
$$= \frac{1}{2}\log 2e\sigma^2 \quad (2)$$

where e and σ are Euler's constant and standard deviation of x, respectively.

Thus, We extract a 3D feature tensor $F_n \in R^{cf2T}, n = 1, 2, \ldots, N$ from each segment, where N is the number of total segments, c is the number of EEG channels, f represents the number of frequency bands, and 2 T is derived by the 0.5 s window without overlapping, as depicted in Fig. 1.

Then, to utilize the spatial information of electrodes, we organize all the c channels as a 2D map so that the 3D feature tensor F_n is transformed into a 4D representation $X_n \in R^{hwf2T}$, where h and w are the height and width of the 2D map, respectively. The 2D map of all the c channels with zero-padding is shown in Fig. 2, which preserves the topology of different electrodes.

Attention-based CNN

For a 4D spatial-spectral-temporal representation X_n, we extract the spatial and spectral information from each temporal slice $S_i \in R^{hw2f}, i = 1, 2, ..., 2T$ with a CNN, explore the discriminative local patterns in spatial and spectral domains with a convolutional attention module, and finally get its spatial and spectral representation. The attention module here is similar to what Woo et al. propose (Woo et al. 2018), which is originally used to improve the representation power of CNN networks.

The structure of the attention-based CNN is shown in Fig. 3. It contains four convolutional blocks and one fully-connected layer. As shown in Fig. 4, Each convolutional block consists of a ResBlock (He et al. 2016) and a convolutional attention module. The convolutional attention module is used in each convolutional block to utilize the spatial and spectral attention mechanisms, and the details will be given later. The four convolutional blocks have 16, 32, 64, and 32 feature maps with the filter size of 5×5, respectively. Finally, outputs of the last convolutional block are flattened and fed to the fully-connected layer with 150 units. Thus, for each temporal slice S_i, we take the final output $P_i \in R^{150}$ as its spatial and spectral representation.

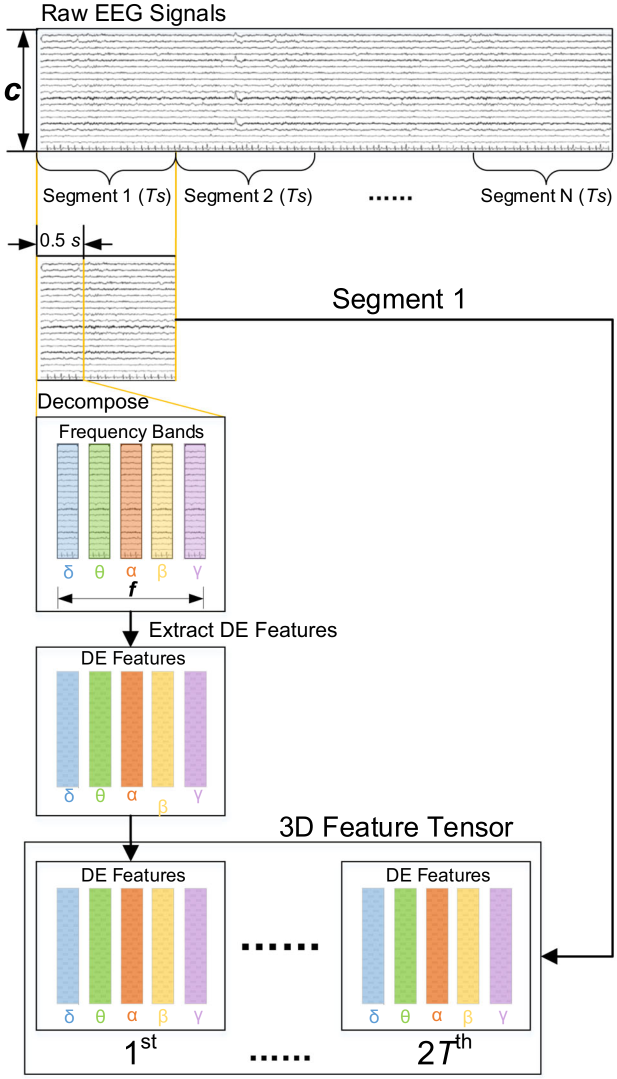
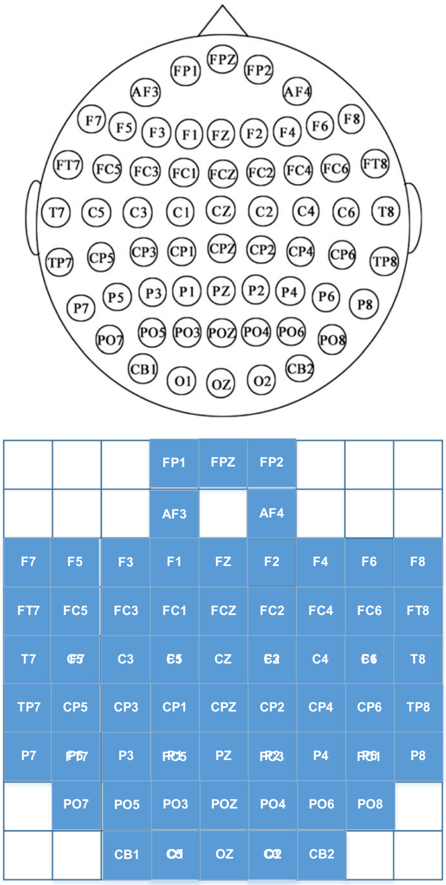


Fig. 1 The generation of 3D feature tensors. For each T seconds EEG signal segment, we extract DE features from c channels and f frequency bands with a 0.5 s window, generating the 3D feature tensors $F_n \in R^{cf2T}, n = 1, 2, \ldots, N$, where N is the number of total segments


FPZ
FP1 FP2
AF3 AF4
F7 F5 F3 F1 FZ F2 F4 F6 F8
FT7 FC5 FC3 FC1 FCZ FC2 FC4 FC6 FT8
T7 C5 C3 C1 CZ C2 C4 C6 T8
TP7 CP5 CP3 CP1 CPZ CP2 CP4 CP6 TP8
P7 P5 P3 P1 PZ P2 P4 P6 P8
PO7 PO5 PO3 POZ PO4 PO6 PO8
CB1 O1 OZ O2 CB2

			FP1	FPZ	FP2			
			AF3		AF4			
F7	F5	F3	F1	FZ	F2	F4	F6	F8
FT7	FC5	FC3	FC1	FCZ	FC2	FC4	FC6	FT8
T7	C5	C3	C1	CZ	C2	C4	C6	T8
TP7	CP5	CP3	CP1	CPZ	CP2	CP4	CP6	TP8
P7	P5	P3	P1	PZ	P2	P4	P6	P8
	PO7	PO5	PO3	POZ	PO4	PO6	PO8	
		CB1	O1	OZ	O2	CB2		

Fig. 2 The 2D map with zero-padding of 62 channels. The purpose of the organization is to preserve the positional relationships among different electrodes

Convolutional attention module

The convolutional attention module is applied after each convolutional layer to adaptively capture important brain regions and frequency bands. The structure of the convolutional attention module is shown in Fig. 5. It consists of two sub-modules, i.e. the spatial attention module and the spectral attention module.

For each convolutional layer above, its output is a 3D feature tensor $V \in R^{h_v \times w_v \times c_v}$, where h_v, w_v, and c_v are the height, the width, and the number of the 2D feature maps of V, respectively. We take V as the input of the convolutional attention module.

The spectral attention module is applied to identify valuable frequency bands for emotion recognition. The average pooling has been widely used to aggregate spatial information and the maximum pooling has been commonly adopted to gather distinctive features. Therefore, we shrink the spatial dimension of V by a spatial-wise average pooling and a spatial-wise maximum pooling, which are defined as:

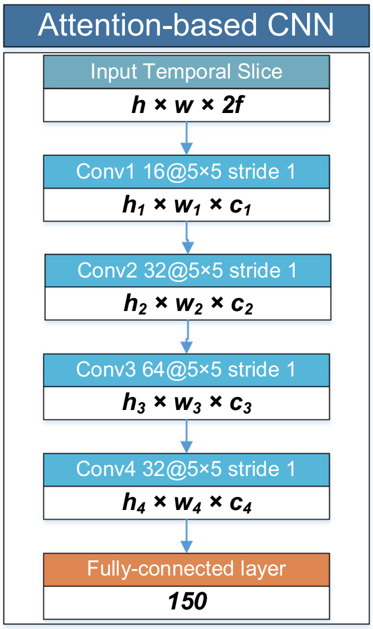


Fig. 3 The structure of the attention-based CNN. The upper half of the blocks in the figure is the type of layers and the lower denotes the shape of its output tensors

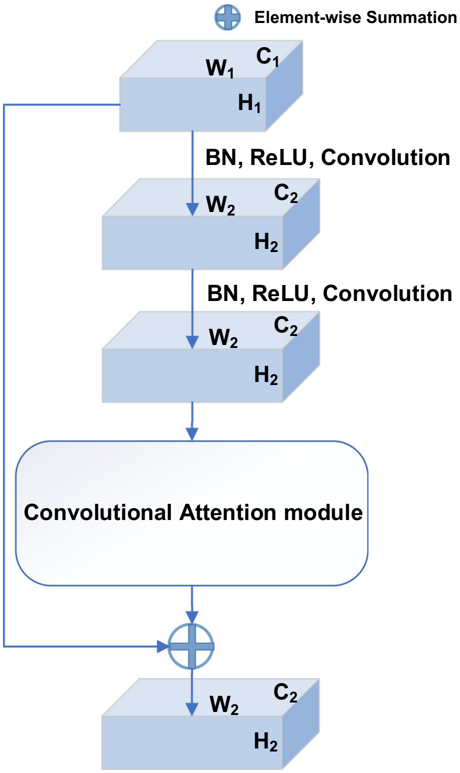


Fig. 4 The structure of the convolutional block which consists of a ResBlock and a convolutional attention module. H1, W1, and C1 represent the height, width, and channels of the input feature map, respectively. H2, W2, and C2 represent the height, width, and channels of the output feature map, respectively. BN denotes the batch normalization and ReLU represents the activation function

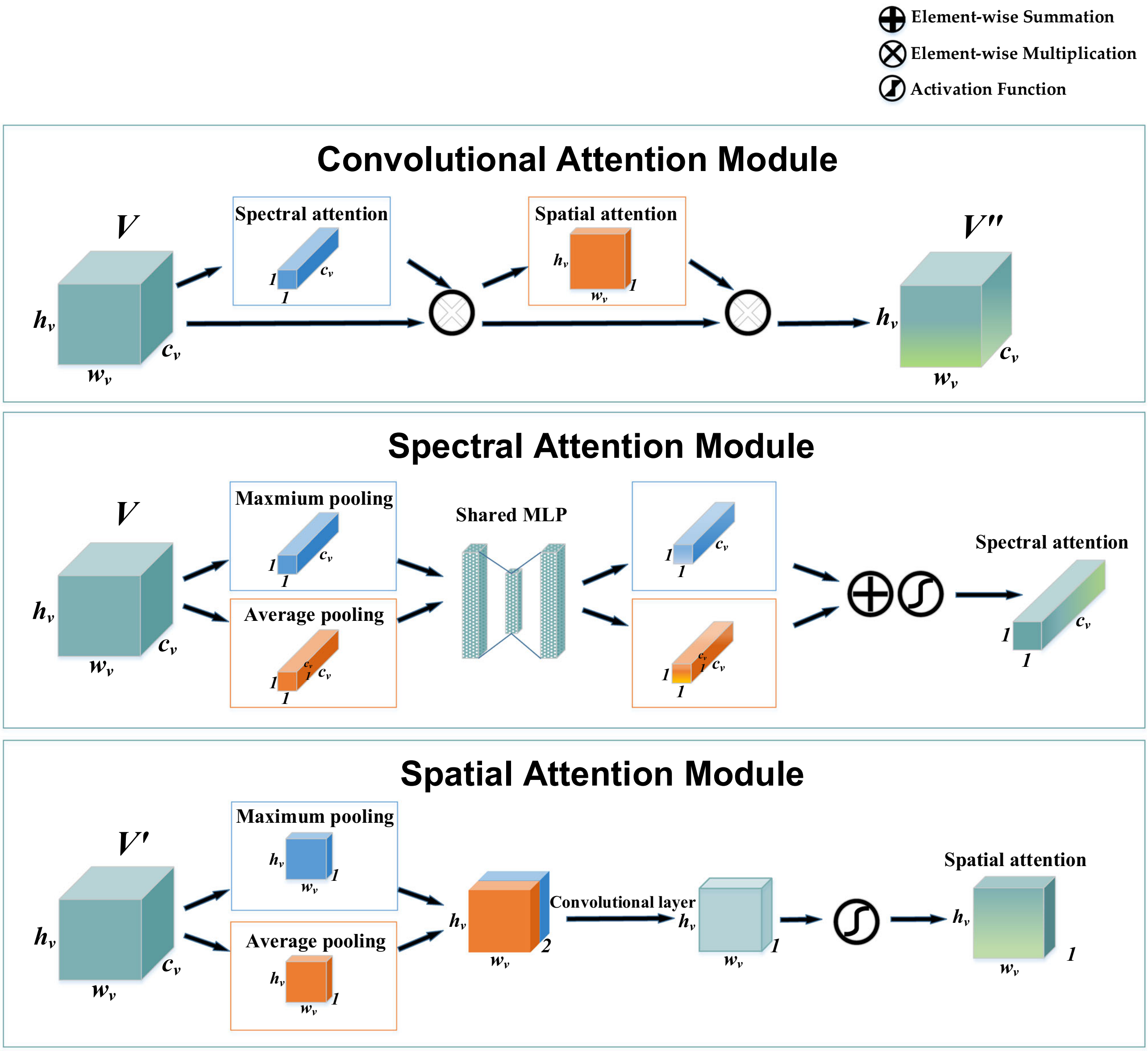


Fig. 5 The top block is the overall structure of the convolutional attention block, it consists of the spectral attention module and the spatial attention module. The middle block represents the generation of spectral attention. The bottom block denotes the generation of spatial attention

$$C_{avg,i} = \frac{1}{h_v \times w_v}\sum_{h=1}^{h_v}\sum_{w=1}^{w_v} V_i(h,w), i = 1,2,\ldots,c_v \tag{3}$$

$$C_{\max,i} = \max(V_i), \quad i = 1,2,\ldots,c_v \tag{4}$$

where $V_i \in R^{h_v \times w_v}$ denotes the 2D feature map in the *i-th* channel of V, $C_{avg,i}$ represents the element in the *i*th channel of the spatial average representation $C_{avg} \in R^{c_v}$, $max(V_i)$ returns the largest element in V_i, and $C_{max,i}$ is the element in the *i-th* channel of the spatial maximum representation $C_{max} \in R^{c_v}$. Subsequently, we implement the spectral attention by two fully-connected layers, a *Relu* activation function and a *sigmoid* activation function, which is defined as:

$$A_{spectral,avg} = W_2^S(\mathrm{Relu}(W_1^S C_{avg}) \tag{5}$$

$$A_{spectral,\max} = W_2^S(Relu(W_1^S C_{\max}) \tag{6}$$

$$A_{spectral} = sigmoid(A_{spectral,avg}A_{spectral,\max}) \tag{7}$$

$$A_{spatial} = Sigmoid(Conv(SPA)) \tag{8}$$

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{9}$$

where W_1^S and W_2^S are learnable parameters, $\oplus$ denotes the element-wise addition, and $A_{spectral} \in R^{1\times 1\times c_v}$ is the spectral attention. The elements of $A_{spectral}$ represent the importance of the corresponding 2D feature maps of the spectral domain. After generating the spectral attention $A_{spectral}$, the output of the spectral attention module can be defined as:

$$V' = A_{spectral}\ V \tag{10}$$

where V' denotes the refined 3D feature tensor, and $\otimes$ represents the element-wise multiplication.

The spatial attention module is applied to identify valuable brain regions for emotion recognition. Firstly, we shrink the spectral dimension of V' by spectral-wise average pooling and spectral-wise maximum pooling, which is defined as:

$$\begin{aligned} SPA_{avg,(h,w)} &= \frac{1}{c_v}\sum_{c=1}^{c_v} S'_{h,w}(c), h = 1, 2, \ldots, h_v; \\ &w = 1, 2, \ldots, w_v \end{aligned} \tag{11}$$

$$\begin{aligned} SPA_{\max,(h,w)} &= \max\left(S'_{h,w}\right), h = 1, 2, \ldots, h_v; \\ &w = 1, 2, \ldots, w_v \end{aligned} \tag{12}$$

where $S'_{h,w} \in R^{c_v}$ denotes the channel in the *h-th* row and *w-th* column of V', $SPA_{avg,(h,w)}$ represents the element in the *h-th* row and *w-th* column of the spectral average representation $SPA_{avg} \in R^{h_v \times w_v \times 1}$ and $SPA_{max,(h,w)}$ is the element in the *h-th* row and *w-th* column of the spectral maximum representation $SPA_{max} \in R^{h_v \times w_v \times 1}$. In the following, we implement the spatial attention with a convolutional layer and a *sigmoid* activation function, which is defined as:

$$SPA = Cat(SPA_{avg}, SPA_{\max}) \tag{13}$$

$$A_{spatial} = Sigmoid(Conv(SPA)) \tag{14}$$

where $Cat(SPA_{avg}, SPA_{max})$ denotes the concatenation of SPA_{avg} and SPA_{max} along the spectral dimension, $Conv(SPA)$ represents the convolutional layer for SPA, and $A_{spatial} \in R^{h_v \times w_v \times 1}$ is the spatial attention. The elements of $A_{spatial}$ represent the importance of the corresponding regions of the spatial domain. Subsequently, the output of the spatial attention module can be defined as:

$$V'' = A_{spatial} V' \tag{15}$$

where $V'' \in R^{h_v \times w_v \times c_v}$ denotes the final output 3D feature tensor of the convolutional attention module.

Attention-based bidirectional LSTM

For each temporal slice $S_i \in R^{hw2f}$, $i = 1, 2, \ldots, 2T$, the final output of the attention-based CNN is $P_i \in R^{150}$. Since the variation between different temporal slices contains temporal information for emotion recognition, we utilize an attention-based bidirectional LSTM to explore the importance of different slices, as shown in Fig. 6.

A bidirectional LSTM connects two unidirectional LSTMs with opposite directions to the same output. Comparing with a unidirectional LSTM, a bidirectional LSTM preserves information from both past and future, making it understand the context better. In this paper, the bidirectional LSTM comprises two unidirectional LSTMs with 36 memory cells. The unidirectional LSTM for positive time direction, $\mathrm{LSTM_P}$ takes the output sequence of the attention-based CNN $P^P = (P_1, P_2, \ldots, P_{2T})$ as the input sequence, while the other for negative time direction, $\mathrm{LSTM_N}$ takes the reverse sequence $P^N = (P_{2T}, P_{2T-1}, \ldots, P_1)$ as the input sequence. The outputs of the *i-th* node of the unidirectional LSTMs are $Y_i^P \in R^{36}$ and $Y_i^N \in R^{36}$, $i = 1, 2, \ldots, 2T$, respectively. Then, we concatenate Y_i^P and Y_{2T+1-i}^N as the output of the *i-th* node of the bidirectional LSTM $Y_i \in R^{72}$. Different from traditional ways that only use the output of the last node of an LSTM for classification or other applications, we take the outputs

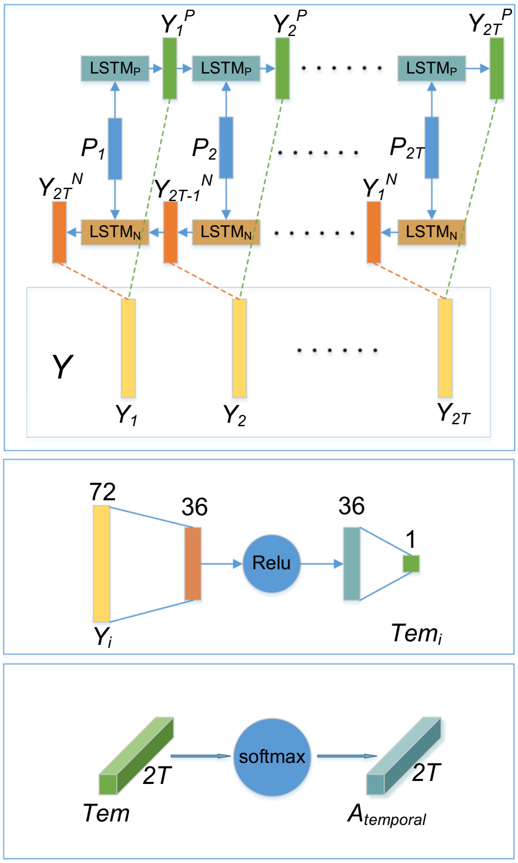


Fig. 6 The top block is the structure of the bidirectional LSTM. We concatenate the outputs of $\mathrm{LSTM_P}$ and $\mathrm{LSTM_P}$ as the output of the bidirectional LSTM, $Y \in R^{2T \times 72}$. The middle block represents the projection of the outputs of the bidirectional LSTM. The bottom block denotes the generation of temporal attention

of all the bidirectional LSTM nodes $Y \in R^{2T \times 72}$ into consideration and explore the importance of different temporal slices by the temporal attention mechanism.

The temporal attention mechanism is implemented with two fully-connected layers, a *Relu* activation function, and a *softmax* activation function, which is defined as:

$$Tem_i = W_2^T\left(Relu\left(W_1^T Y_i + b_1^T\right)\right) + b_2^T \tag{16}$$

$$A_{temporal} = softmax(Tem) \tag{17}$$

$$softmax(x) = \frac{\exp(x)}{\sum \exp(x)} \tag{18}$$

where W_1^T, W_2^T, b_1^T and b_2^T are learnable parameters, Tem_i represents the *i-th* element of $Tem \in R^{2T \times 1}$ which projects $Y \in R^{2T \times 72}$ to a lower dimension, and $A_{temporal} \in R^{2T \times 1}$ is the temporal attention. The elements of $A_{temporal}$ represent the importance of the corresponding temporal slices. Subsequently, the high-level representation of the 4D sample X_n can be defined as:

$$L_n(e) = \sum A_{temporal} Y_e, \; e = 1, 2, \ldots, 72 \tag{19}$$

where $Y_e \in R^{2T \times 1}$ denotes the *e-th* column of $Y \in R^{2T \times 72}$ and $L_n(e)$ is the *e-th* element of the high-level representation $L_n \in R^{72}$, which integrates spatial, spectral, and temporal information of X_n.

Classifier

Based on the high-level representation L_n of EEG signals, we apply a fully-connected layer and a *softmax* activation function to predict the label of the 4D sample X_n, which can be defined as follows:

$$Pre = softmax(W^p L_n + b^p) \tag{20}$$

where W^p, b^p are learnable parameters and $Pre \in R^C$ denotes the probability of X_n belonging to all the C classes. Specifically, the class of the largest probability is the predicted label of 4D-aNN.

Experiment

In this section, we firstly introduce two widely used datasets. Then, the experiment settings are described. Finally, the results on the datasets are reported and discussed.

DEAP dataset

DEAP dataset (Koelstra et al. 2011) includes 32-channel EEG signals of 32 participants recorded while they watch 40 video clips. For each clip, participants record their level of valence and arousal from 1 to 9 as labels, and we set the threshold to divide labels into two classes as 5. EEG signals of each clip contain 3 s baseline signals and 60 s trial signals. Then, the EEG signals are down-sampled to 128 Hz and passed to a bandpass filter between 4 and 45 Hz to remove noise. We split the 3 s baseline signals into 6 segments of 0.5 s, and extract DE features of each segment for different frequency bands. Then, we calculate the baseline DE features by averaging DE features of those 6 segments for each frequency band. Finally, we subtract baseline DE features from DE features of trial signals before recognition.

SEED dataset

SEED dataset (Zheng and Lu 2015) contains 3 different categories of emotion data: positive, neutral, and negative. For each kind of emotion, 5 film clips that are about 4 min long and can elicit the desired target emotion are selected. 15 healthy subjects (7 males and 8 females, with age (23.27 ± 2.37)) take part in the EEG signals collection. 3 groups of experiments are conducted for each subject, and each experiment consists of 15 clips viewing processes. Each clip viewing process can be divided into four stages, including a 5 s hint of start, a 4 min clip period, a 45 s self-assessment, and a 15 s rest period. The order of the 15 clips is arranged so that two clips eliciting the same emotion are not shown consecutively. The EEG signals in the experiments are recorded by a 62-channel's ESI NeuroScan system and down-sampled to 200 Hz. Besides, the EEG signals seriously contaminated by electromyography (EMG) and electrooculography (EOG) are removed manually. Then, a bandpass filter between 0.3 to 50 Hz is applied to filter the noise.

SEED-IV dataset

SEED-IV dataset (Zheng et al. 2018) contains 4 different categories of emotion data: neutral, sad, fear, and happy. Similar to the previous one, this dataset also involves three groups of experiments and 15 subjects (8 females, aged between 20 and 24). All participants watch 24 videos (6 videos per emotion type) in every single experiment while their EEG signals and eye movements are collected by the 62-channel ESI NeuroScan System and SMI eye-tracking glasses, respectively. There is a 5 s preparation stage before the video and a 45 s self -assessment stage after the video. The original EEG signals are filtered by a bandpass filter between 1 and 75 Hz after the down-sample operation.

Settings

The proposed 4D-aNN takes a 4D segment $X_n \in R^{hwf2T}$ as the input. In this paper, we adopt the 2D compact map with $h = 9$ and $w = 9$ to maintain the positional relationship of electrodes. As shown in previous works, the combination of all the frequency bands can contribute to better results so that we set $f = 4$ for DEAP dataset (θ[4 ~ 8 Hz], α[8 ~ 14 HZ], β[14 ~ 31 Hz], and γ[31 ~ 51 Hz]) and $f = 5$ for SEED dataset and SEED-IV dataset (δ[1 ~ 4 Hz], θ[4 ~ 8 Hz], α[8 ~ 14 HZ], β[14 ~ 31 Hz], and γ[31 ~ 51 Hz]). We set the length of segments T as 3, obtaining about 800 samples per subject in the DEAP dataset, about 1130 samples in the SEED dataset and 1100 samples in the SEED-IV dataset per experiment for each subject. For DEAP dataset, we conduct a fivefold cross-validation on each subject: First, following the experimental settings in previous works (Shen et al. 2020; Tao et al. 2020; Yang et al. 2018b), we shuffle all the samples of each subject and divide them into 5 groups randomly. Second, each group is used as the test set while the other 4 groups are used as the train set. Finally, we calculate the average classification accuracy (ACC) and standard deviation (STD) of each subject. For SEED dataset and SEED-IV dataset, we randomly shuffle the samples as Tao (Tao et al. 2020) and Shen (Shen et al. 2020) do, then conduct a fivefold cross-validation on each experiment and calculate the average ACC and STD of 3 experiments of each subject. Moreover, we implement a cross-video classification on SEED-IV dataset by protocols of Zheng, Li and Zhong (Li et al. 2020; Zheng et al. 2018; Zhong et al. 2020).The average ACC and STD of all subjects are taken as the final performances of our method. We train the 4D-aNN on an NVIDIA GTX 1080 GPU. The Adam optimization is applied to minimize the loss function. We set the learning rate as 0.0001, the maximum of epochs as 100 and the batch size as 12 (16 for SEED-IV).

Baseline models

- PCRNN (Yang et al. 2018b): It concatenates the spatial features obtained by CNN and the temporal features extracted by LSTM to finish emotion recognition.
- ACRNN (Tao et al. 2020): It uses a convolutional recurrent neural network to learn spatial and temporal features, and extracts more important features with attention mechanisms.
- 4D-CRNN (Shen et al. 2020): It builds DE features extracted from EEG signals into 4D feature structures and uses a convolutional recurrent neural network to extract spatial features, spectral features, and temporal features for EEG emotion recognition.
- SST-EmotionNet (Jia et al. 2020): It uses a two-stream network to extract spatial, spectral, and temporal features. Besides, SST-EmotionNet adopts the attention mechanisms to improve its performance on EEG emotion recognition.
- EmotionMeter (Zheng et al. 2018): It develops a multimodal framework using bimodal deep auto-encoder to utilize both eye movement and EEG signals for promotion.
- BiHDM (Li et al. 2020): It considers the asymmetric differences between two hemispheres for EEG emotion recognition and applies four directed RNNs to obtain the deep representation of all the EEG electrodes' signals.
- RGNN (Zhong et al. 2020): It takes the biological topology among different brain regions into consideration to capture both global and local relations among different EEG channels.

Results

We compare our model with several baseline models on DEAP dataset, SEED dataset and SEED-IV dataset. Table 1 presents the average ACC and STD of these models for EEG emotion recognition. On DEAP dataset, the valence classification accuracy of 4D-aNN is 96.90%, exceeding PCRNN, ACRNN and 4D-CRNN by 6.64%, 3.18% and 2.68%, respectively. The arousal classification accuracy of 4D-aNN is 97.39%, which outperforms PCRNN, ACRNN and 4D-CRNN by 6.41%, 4.01% and 2.81%, respectively. On SEED dataset, the classification accuracy of 4D-aNN is 96.25%, beating 4D-CRNN, and SST-EmotionNet by 1.51% and 0.23% (for a 3 s sample, the data size of SST-EmotionNet is more than 100 times bigger than 4D-aNN), separately. On SEED-IV dataset, the four-category classification accuracy of 4D-aNN reaches 86.77%, which shows a promotion of 1.85% over SST-EmotionNet.

Comparing with the baseline models, the proposed 4D-aNN achieves the state-of-the-art performance on DEAP, SEED and SEED-IV datasets under intra-subject splitting. As shown in Figs. 7, 8 and 9, we demonstrate the performances of 4D-aNN on each subject on DEAP, SEED and SEED-IV datasets. For valence and arousal classification on DEAP dataset, 4D-aNN performs well on most subjects (except #5 and #22). The valence accuracy of subject #5 and arousal accuracy of subject #22 are 92.0% and 90.63%, respectively, which are obviously lower than the average accuracies. The possible reason is that the subjective feelings about music videos of subjects #5 and #22 were not recorded precisely. For classification on SEED dataset,

Table 1 The performance (average ACC and STD (%)) of the compared models

Model	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)	SEED-IV ACC ± STD (%)
PCRNN Yang et al. (2018b)	90.26 ± 2.88	90.98 ± 3.09	–	–
ACRNN Tao et al. (2020)	93.72 ± 3.21	93.38 ± 3.73	–	–
4D-CRNN Shen et al. (2020)	94.22 ± 2.61	94.58 ± 3.69	94.74 ± 2.32	–
SST-EmotionNet Jia et al. (2020)	–	–	96.02 ± 2.17	84.92 ± 6.66
4D-aNN	96.90 ± 1.65	97.39 ± 1.75	96.25 ± 1.86	86.77 ± 7.29

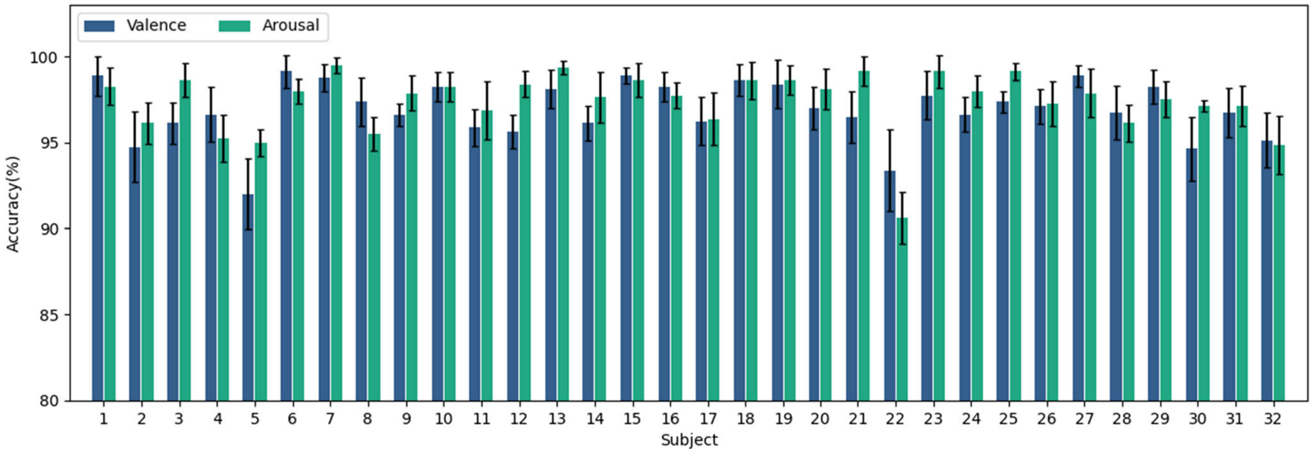


Fig. 7 The performance of 4D-aNN on DEAP dataset. We conduct a fivefold cross-validation for each subject for valence and arousal classification, respectively

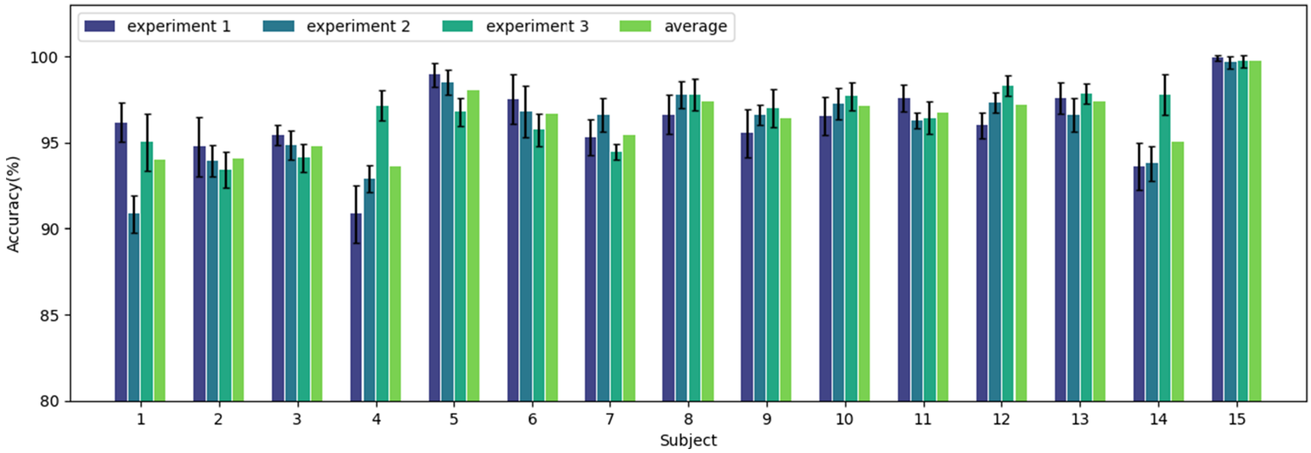


Fig. 8 The performance of 4D-aNN on SEED dataset. 3 experiments are conducted for each subject, we conduct a fivefold cross-validation for each experiment, and also present the average classification accuracy for each subject

there are 9 subjects (#5, #6, #8, #9, #10, #11, #12, #13, and #15) whose performances are better than the mean accuracy. Likewise, the performances of 8 subjects (#2, #4, #6, #7, #9, #13, #14, and #15) on SEED-IV dataset are above average.

The experimental results indicate that the proposed 4D-aNN that integrates attention mechanisms on different domains to capture discriminative information could achieve superior performances on DEAP, SEED and SEED-IV datasets.

In addition, to further illustrate the effectiveness of 4D-aNN, we implement a cross-video classification on SEED-IV dataset following the protocols of Zheng, Li and Zhong (Li et al. 2020; Zheng et al. 2018; Zhong et al. 2020). More

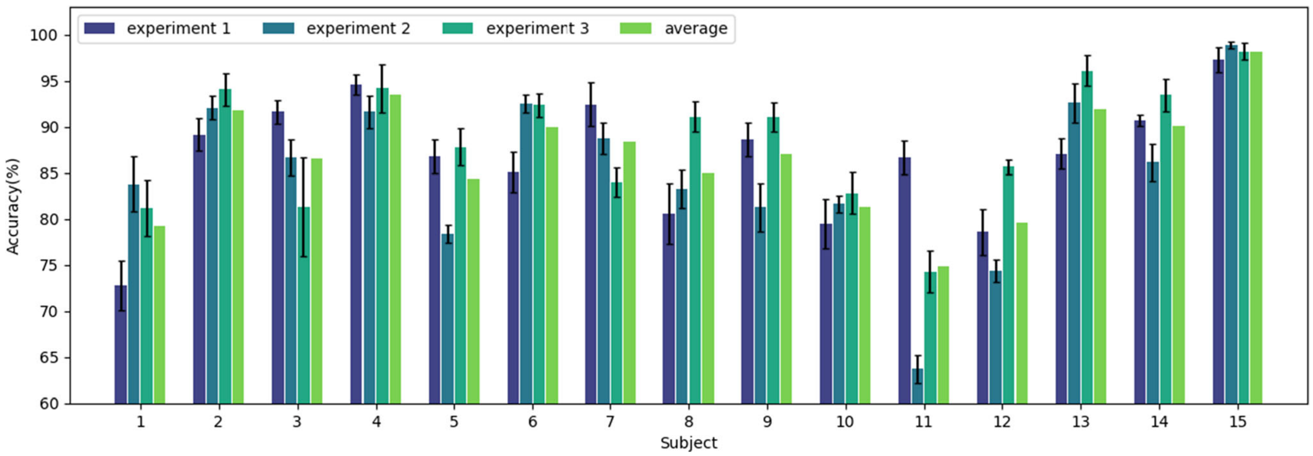


Fig. 9 The performance of 4D-aNN on SEED-IV dataset. 3 experiments are conducted for each subject, we conduct a fivefold cross-validation for each experiment, and also present the average classification accuracy for each subject

exactly, we use samples of 16 videos as the train set and the remaining samples of 8 videos (2 videos per emotion type) as the test set for every single experiment. Evaluation is conducted among 3 experiments of all subjects. The results are shown in Table 2. As can be seen, the proposed 4D-aNN also achieves a better accuracy of 79.80% in the cross-video classification, compared with other models. Therefore, 4D-aNN can also perform well under different conditions.

To demonstrate the importance of the attention mechanisms in our model, we conduct an additional experiment for ablation studies on 3 datasets. We evaluate the performances of 4D-aNN when all the attention mechanisms are ablated. As shown in Fig. 10, after any attention mechanism is ablated, the classification accuracies decrease obviously, especially the temporal attention mechanism. For valence and arousal classification on DEAP dataset, the accuracies of 4D-aNN without all the attention mechanisms decrease by 1.93% and 1.76%, respectively. For classification on SEED dataset and SEED-IV dataset, the accuracies of 4D-aNN without all the attention mechanisms decreases by 2.33% and 3.21%, separately. The results of the ablation experiment indicate that the attention mechanisms make contributions to EEG emotion recognition for the ability to capture the discriminative local patterns in spatial, spectral, and temporal domains.

Table 2 The performance (average ACC and STD (%)) of the compared models on SEED-IV (cross-video)

Model	SEED-IV ACC ± STD (%)
BiDANN Li et al. (2018)	70.29 ± 12.03
Emotion Meter Zheng et al. (2018)	70.58 ± 17.01
BiHDM Li et al. (2020)	74.35 ± 14.09
RGNN Zhong et al. (2020)	79.37 ± 10.54
4D-aNN	79.80 ± 10.31

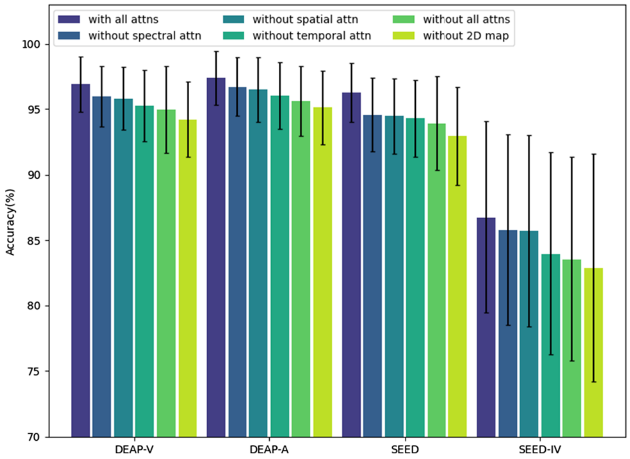


Fig. 10 Ablation studies on the attention mechanisms of 4D-aNN on SEED and DEAP dataset. "DEAP-V" and "DEAP-A" represent valence and arousal classification, respectively

Besides, we conduct experiments of ablation on 2D map: removing the spatial relationship of the signal and discarding padding operation. The figure shows that the evaluation metrics fall significantly: four accuracy indicators decrease to 94.21%, 95.12%, 92.93% and 82.9%, which means all reductions are more than 2%. As described above, the superiorities of preserving the spatial relationship of the signal outweigh the inferiorities of introducing some noise by 2D map.

In particular, to explore the critical brain regions for different emotions, we separately depict the electrode activity heatmaps in Fig. 11. We draw the heatmaps using *Grad-CAM* + + (Chattopadhay et al. 2018), based on the

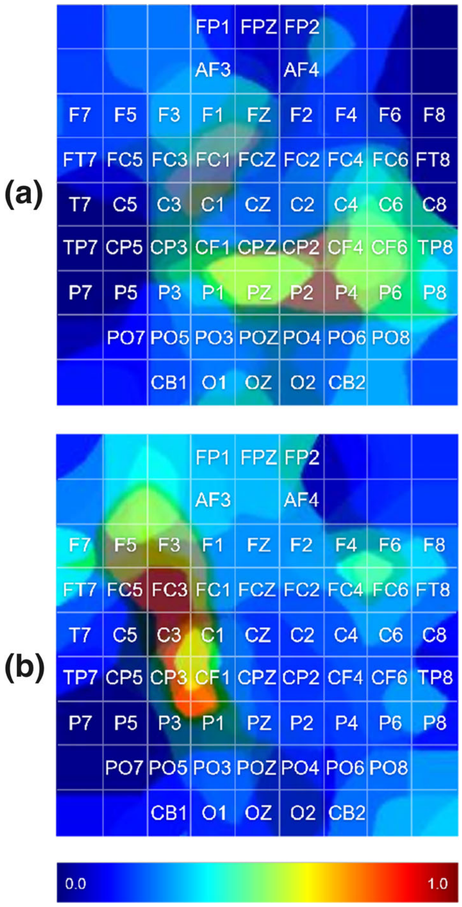


Fig. 11 The electrode activity heatmaps based on the arousal classification results of subject #7 on DEAP dataset. Parts (a), (b) correspond to the rated level of arousal less and more than 5, respectively. Dark red regions denote more significant contributions to the recognition of the corresponding emotions

arousal classification results of subject #7. *Grad-CAM* + + uses the last convolutional layer feature maps and the class scores of the classifier to generate heatmaps. The heatmaps are able to explain which input regions are important for predictions. In this work, the size of each heatmap is 9 × 9, which is the same as the 2D compact map. The elements in the heatmaps represent the contributions of the corresponding brain regions to the recognition of the target emotions. From Fig. 11, We can observe the distinct distributions of important brain regions with regard to different emotions: When the rated level of arousal is less than 5, channels P2, P4, and C1 are more active than others. In contrast, when the rated level of arousal is more than 5, channels P1, P3, and FC3 are obviously more active than others.

As for the temporal domain, we obtain and normalize the temporal attention weights from the model, based on 4 subjects' different types of classifications on SEED-IV dataset. Then we similarly visualize the weights similarly into Fig. 12. We can notice that the model does pay attention to some certain time points usefully while inferring.

Some researchers have studied the relationship between emotion and brain activity from a physiological point of view. Damasio et al. indicate that the activation patterns in different regions of brain including insular lobe, pons, cingulate cortex, etc. are also different (Damasio et al. 2000). Besides, Koven's experiments show that the brain responds faster to pleasant stimuli than to unpleasant stimuli (Koven et al. 2003). Moreover, Leon-Carrion et al. find that the brain relies on the frontal and posterior cortex to complete the temporal response to emotional stimuli (Leon-Carrion et al. 2006).

Distinctly, the brain activities could vary with different subjects, time, and emotions so that the attention mechanisms that enable 4D-aNN to adaptively capture discriminative patterns make sense for EEG emotion recognition.

Discussion

We conduct several experiments to investigate the use of 4D-aNN which fuses the spatial-spectral-temporal information and the effectiveness of the attention mechanisms on different domains for EEG emotion classification. In this section, we discuss three noteworthy points.

First, to deal with the spatial-spectral information, we apply an attention-based CNN which consists of a CNN network, a spectral attention module, and a spatial attention module. The CNN network extracts the spatial-spectral representation from inputs first. Then, the spectral attention mechanism is applied to each spectral feature to explore the importance of different frequency bands. Besides, the spatial attention mechanism is applied to each 2D feature map to adaptively capture the critical brain regions. The critical brain regions and frequency bands could vary with different individuals, emotions, and time. Therefore, the ability to capture discriminative information of the attention modules improves the performance of 4D-aNN despite not exploring the connectivity patterns.

Second, to explore the temporal dependencies in 4D spatial-spectral-temporal representations, we utilize an attention-based bidirectional LSTM. The bidirectional LSTM extracts high-level representations from the outputs of the attention-based CNN. Different from traditional ways that only use the output of the last node of an LSTM for classifications or other applications, we consider outputs of all the nodes with the temporal attention mechanism. The temporal attention mechanism adaptively assigns weights of different temporal slices so that the dynamic content of emotions in 4D representations could be captured better.

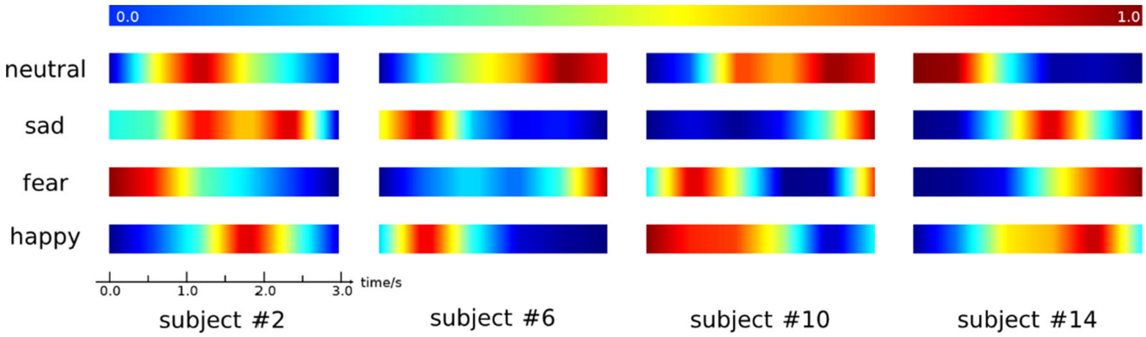


Fig. 12 The visualized heatmaps of temporal attention weights based on 4 subjects' different types of classifications on SEED-IV dataset. Dark red regions denote more significant contributions to the recognition of the corresponding emotions

Third, to address the importance of the attention mechanisms and 2D map, we conduct ablation studies on 4D-aNN. For valence and arousal classification on DEAP dataset, accuracies of 4D-aNN without all the attention mechanisms decrease by 1.93% and 1.76%, respectively. For classification on SEED dataset and SEED-IV, accuracies of 4D-aNN decreases by 2.33% and 3.21%, respectively. For experiments of ablation on 2D map, four accuracies indicators decrease by 2.69%, 2.27%, 3.32% and 3.87%, separately. The experimental results demonstrate the effectiveness of the attention mechanisms and 2D map to adaptively capture discriminative patterns.

Conclusion

In this paper, we propose the 4D-aNN model for EEG emotion recognition. The 4D-aNN takes 4D spatial-spectral-temporal representations containing spatial, spectral, and temporal information of EEG signals as inputs. We integrate the attention mechanisms into the CNN module and the bidirectional LSTM module. The CNN module deals with the spatial and spectral information of EEG signals while the spatial and spectral attention mechanisms capture critical brain regions and frequency bands adaptively. The bidirectional LSTM module extracts temporal dependencies on the outputs of the CNN module while the temporal attention mechanism explores the importance of different temporal slices. The experiments on DEAP, SEED and SEED-IV datasets demonstrate better performance than all baselines. In particular, the ablation studies show the effectiveness of the attention mechanisms and 2D map in our model for EEG emotion recognition.

Acknowledgements Sponsored by Zhejiang Lab (No. 2021RD0AB01), and supported by High-performance Computing Platform of Peking University.

Funding The work was sponsored by Zhejiang Lab (No. 2021RD0AB01).

Availability of data and material The dataset used is available on https://www.eecs.qmul.ac.uk/mmv/datasets/deap/ and http://bcmi.sjtu.edu.cn/home/~seed.

Code availability The code of our model will be available after acceptance.

Declarations

Conflict of interest Not applicable.

References

Alhagry S, Fahmy AA, El-Khoribi RA (2017) Emotion recognition based on EEG using LSTM recurrent neural network. Int J Adv Comput Sci Appl 8:335–358. https://doi.org/10.14569/IJACSA.2017.081046

Bamdad M, Zarshenas H, Auais MA (2015) Application of BCI systems in neurorehabilitation: a scoping review disability and rehabilitation. Assist Technol 10:355–364. https://doi.org/10.3109/17483107.2014.961569

Blankertz B et al (2016) The Berlin brain-computer interface: progress beyond communication and control. Front Neurosci 10:530. https://doi.org/10.3389/fnins.2016.00530

Brittona JC, Phan KL, Taylor SF, Welsh RC, Berridge KC, Liberzon I (2006) Neural correlates of social and nonsocial emotions: an fMRI study. Neuroimage 31:397–409. https://doi.org/10.1016/j.neuroimage.2005.11.027

Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. Paper presented at the 2018 IEEE winter conference on applications of computer vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018. doi:https://doi.org/10.1109/WACV.2018.00097

Craik A, He Y, Contreras-Vidal JL (2019) Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 16:031001. https://doi.org/10.1088/1741-2552/ab0ab5

Damasio AR, Grabowski TJ, Bechara A, Damasio H, Ponto LL, Parvizi J, Hichwa RD (2000) Subcortical and cortical brain activity during the feeling of self-generated emotions. Nat Neurosci 3:1049. https://doi.org/10.1038/79871

Dolan RJ (2002) Emotion, cognition, and behavior. Science 298:1191–1194. https://doi.org/10.1126/science.1076358

Duan R-N, Zhu J-Y, Lu B-L (2013) Differential entropy feature for eeg-based emotion classification. Paper presented at the 2013 6th international IEEE/EMBS conference on neural engineering (NER), San Diego, CA, USA. doi:https://doi.org/10.1109/NER.2013.6695876

Figueiredo GR, Ripka WL, Romaneli EFR, Ulbricht L (2019) Attentional bias for emotional faces in depressed and nondepressed individuals: an eye-tracking study. Paper presented at the 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), Berlin, Germany, 23–27 July 2019. doi:https://doi.org/10.1109/EMBC.2019.8857878

Fiorinia L, Mancioppi G, Semeraro F, Fujita H, Cavallo F (2020) Unsupervised emotional state classification through physiological parameters for social robotics applications. Knowledge-Based Syst. https://doi.org/10.1016/j.knosys.2019.105217

Goh SK, Abbass HA, Tan KC, Al-Mamun A, Thakor N, Bezerianos A, Li J (2018) Spatio-spectral representation learning for electroencephalographic gait-pattern classification. IEEE Trans Neural Syst Rehabil Eng 26:1858–1867. https://doi.org/10.1109/TNSRE.2018.2864119

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. pp 770–778. doi:https://doi.org/10.1109/CVPR.2016.90

Jenke R, Peer A, Buss M (2014) Feature extraction and selection for emotion recognition from eeg. IEEE Trans Affect Comput 5:327–339. https://doi.org/10.1109/TAFFC.2014.2339834

Jia Z, Lin Y, Cai X, Chen H, Gou H, Wang J (2020) SST-EmotionNet: spatial-spectral-temporal based attention 3D Dense Network for EEG emotion recognition. In: Proceedings of the 28th ACM international conference on multimedia, Seattle, WA, USA, 2020. Association for Computing Machinery, pp 2909–2917. doi:https://doi.org/10.1145/3394171.3413724

Katsigiannis S, Ramzan N (2017) Dreamer: a database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inf 22:98–107. https://doi.org/10.1109/JBHI.2017.2688239

Kim M-K, Kim M, Oh E, Kim S-P (2013) A review on the computational methods for emotional state estimation from the human eeg. Comput Math Methods Med. https://doi.org/10.1155/2013/573734

Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Patras I (2011) Deap: a dataset for emotion analysis using physiological signals. IEEE Trans Affect Comput 3:18–31. https://doi.org/10.1109/T-AFFC.2011.15

Koven NS, Heller W, Banich MT, Miller GA (2003) Relationships of distinct affective dimensions to performance on an emotional stroop task. Cogn Ther Res 27:671–680. https://doi.org/10.1023/A:1026303828675

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 2012. Curran Associates, Inc., pp 1097–1105. doi:https://doi.org/10.1145/3065386

Leon-Carrion J, Mcmanis MH, Castillo EM, Papanicolaou AC (2006) Time-locked brain activity associated with emotion: a pilot MEG study. Brain Injury: BI 20:857–865. https://doi.org/10.1080/02699050600832304

Li Y, Zheng W, Zong Y, Cui Z, Zhang T, Zhou X (2018) A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2018.2885474

Li M, Lu B-L (2009) Emotion classification based on gamma-band EEG. Paper presented at the 2009 annual international conference of the IEEE engineering in medicine and biology society, Minneapolis, MN, USA. doi:https://doi.org/10.1109/IEMBS.2009.5334139

Li Y et al (2020) A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Trans Cogn Dev Syst. https://doi.org/10.1109/TCDS.2020.2999337

Lotfia E, Akbarzadeh-T M-R (2014) Practical emotional neural networks. Neural Netw 59:61–72. https://doi.org/10.1016/j.neunet.2014.06.012

Jiaxin Ma, Tang H, Zheng W-L, Lu B-L (2019) Emotion recognition using multimodal residual LSTM network. In: Proceedings of the 27th ACM international conference on multimedia, Nice, France. Association for computing machinery, New York, USA, pp 176–183. doi:https://doi.org/10.1145/3343031.3350871

Mühl C, Allison B, Nijholt A, Chanel G (2014) A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. Brain-Comput Interfaces 1:66–84. https://doi.org/10.1080/2326263X.2014.912881

Pfurtscheller G et al (2010) The Hybrid BCI. Front Neurosci 4:3. https://doi.org/10.3389/fnpro.2010.00003

Shen F, Dai G, Lin G, Zhang J, Kong W, Zeng H (2020) EEG-based emotion recognition using 4D convolutional recurrent neural network. Cogn Neurodyn 14:815–828. https://doi.org/10.1007/s11571-020-09634-1

Shi L-C, Jiao Y-Y, Lu B-L (2013) Differential entropy feature for eeg-based vigilance estimation. Paper presented at the 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), Osaka, Japan, 3–7 July 2013. doi:https://doi.org/10.1109/EMBC.2013.6611075

Song T, Zheng W, Song P, Cui Z (2020) EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans Affect Comput 11:532–541. https://doi.org/10.1109/TAFFC.2018.2817622

Tao W, Li C, Song R, Cheng J, Liu Y, Wan F, Chen X (2020) EEG-based emotion recognition via channel-wise attention and self attention. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.3025777

Vaswani A, et al (2017) Attention is all you need. Paper presented at the Advances in Neural Information Processing Systems. doi:https://doi.org/10.5555/3295222.3295349

Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. Computer Vision—ECCV 2018. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-01234-2_1

Yan J, Zheng W, Xu Q, Lu G, Li H, Wang B (2016) Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech. IEEE Trans Multimed 18:1319–1329. https://doi.org/10.1109/TMM.2016.2557721

Yang Y, Wu QMJ, Zheng W-L, Lu B-L (2018c) EEG-based emotion recognition using hierarchical network with subnetwork nodes. IEEE Trans Cogn Dev Syst 10:408–419. https://doi.org/10.1109/TCDS.2017.2685338

Yang Y, Wu Q, Fu Y, Chen X (2018a) Continuous convolutional neural network with 3D input for EEG-based emotion recognition. In: Cheng L, Leung ACS, Ozawa S (eds) Neural information processing. Springer International Publishing, pp 433–433. doi:https://doi.org/10.1007/978-3-030-04239-4_39

Yang Y, Wu Q, Qiu M, Wang Y, Chen X (2018b) Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In: 2018 International joint conference on neural networks (IJCNN). IEEE, pp 1–7. doi:https://doi.org/10.1109/IJCNN.2018.8489331

Zheng W-L, Lu B-L (2015) Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans Auton Ment Dev 7:162–175. https://doi.org/10.1109/TAMD.2015.2431497

Zheng W-L, Zhu J-Y, Lu B-L (2017) Identifying stable patterns over time for emotion recognition from EEG. IEEE Trans Affect

Comput 10:417–429. https://doi.org/10.1109/TAFFC.2017.2712143

Zheng W-L, Liu W, Lu Y, Lu B-L, Cichocki A (2018) Emotionmeter: A multimodal framework for recognizing human emotions. IEEE Trans Cybern 49:1110–1122. https://doi.org/10.1109/TCYB.2018.2797176

Zhong P, Wang D, Miao C (2020) EEG-based emotion recognition using regularized graph neural networks. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.2994159

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.