Biomedical Signal Processing and Control 69 (2021) 102743

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Leveraging spatial-temporal convolutional features for EEG-based emotion recognition

Yi An, Ning Xu, Zhen Qu

School of Information Science, Tibet University, Lhasa, China

ARTICLE INFO

Keywords:
EEG
Emotion recognition
DCNN
Attention
ConvLSTM

ABSTRACT

The electroencephalogram (EEG) signal is a medium to realize a brain–computer interface (BCI) system due to its zero clinical risk and portable acquisition devices. As deep learning technology has been considered to obtain a great success towards solving various vision-based research problems such as affective computing. Therefore, in the present paper, a novel framework for EEG-based emotion recognition is proposed. The framework consists of two modules. The first module is deep convolutional neural network (DCNN) architecture, which can represent the inter-channel correlation among physically adjacent EEG signals by converting the chain-like EEG sequence into 2D frame sequences. The second module is ConvLSTM, which can represent the sequence information of the EEG data samples. After that, the features of DCNN and ConvLSTM are concatenated and represented by attention mechanism for final emotion recognition. Extensive experiments conducted on the DEAP database demonstrate that: (1) The proposed framework effectively improves the accuracies of both emotion classification, with arousal dimension up to 87.69%, which is higher than the most of the state-of-the-art methods. (2) The dimension of valence also obtains comparable emotion recognition performance with the accuracy of 87.84%, which surpass the most of the state-of-the-art methods.

1. Introduction

Recently, researchers from machine learning and affective computing field have focused on emotional expression analysis based on audiovisual and physiological signals. From the perspective of emotional classification, emotion recognition can be categorized into two groups: the first one is non-verbal behaviors, such as facial expression [1], posture [2], physiological signals (i.e., functional magnetic resonance imaging (FMRI) [3], magnetoencephalography (MEG) [4], electroencephalogram (EEG) [5], electrocardiogram (ECG)). In [6], the authors found that facial images from videos are vital for representing the expressions. However, the physiological signals can objectively represent the states of human emotion. In some cases, the EEG signals can obviously reflect certain important human emotional changes. Hence, many researchers focus on studying emotion recognition based on EEG signals.

In recent years, researchers often utilize traditional methods for EEG-based emotion recognition. In [7], the authors first pre-process the EEG signals, and extract the EEG parameters by discrete wavelet transform. Then an input matrix is generated from the above feature extraction method on 63 biosensors for emotion classification. Finally, they adopt Fuzzy C-Means (FCM) and Fuzzy K-Means (FKM) clustering methods for emotion classification. In [8], the authors adopt mutual information to select the reduced signals for emotion recognition. For the reduced signals, which include the highest mutual information for extracting the features. Experiments are conducted on DEAP [9] and MAHNOBHCI datasets [10], and show some improvements for emotion recognition task.

Advances in deep learning technology have also give rise to many applications by training deep models based on neural networks in affective computing study [11], [12]. In [13], the authors provide a comprehensive review in adopting EEG signals for emotion recognition, and suggest that the most studies adopt deep learning efficiently for feature extraction and classification. In [14], the authors extract time, frequency and location features of EEG by convolutional neural networks (CNN), and also used Stacked Auto Encoders (SAE) to improve the performance of emotion classification. In [15], a cascade and parallel convolutional recurrent neural network is proposed to recognize movement intention. In [16], the author propose to adopt a pre-processing approach to transform the EEG signal into 2D format data and combine CNN and recurrent neural network (RNN) to predict the emotional state of the trial samples. The authors adopt Power Spectral Density of different EEG channels and transform it into two-dimensional plane to generate the EEG multidimensional feature image (MFI), then CNN was used to learn temporary image information from EEG MFI sequences, while LSTM was used to classify human emotions [16].

However, most of the methods still use hand-crafted features, and then adopt some classification approaches (i.e., support vector machine

https://doi.org/10.1016/j.bspc.2021.102743
Received 17 May 2020; Received in revised form 4 May 2021; Accepted 7 May 2021
Available online 29 June 2021

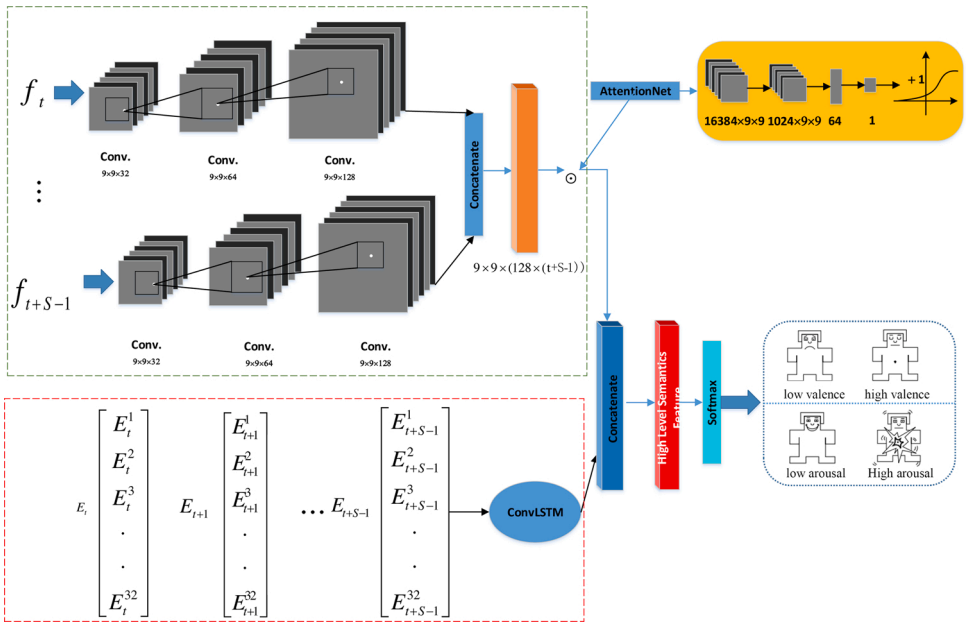

Fig. 1. The illustration of the proposed framework. The red rectangle is ConvLSTM channel, which can model the temporal patterns from the EEG data. The green rectangle is DCNN channel, which can represent the spatial patterns from the 2D like frame sequences.

(SVM) etc.) for predicting the emotional state. For example, in [14], [16], [17], these works have not adopted the feature extraction capability of DCNN. Specifically, the advantage of DCNN have not fully mined in the EEG signals. Additionally, the importance of the baseline (EEG signals without stimulation) is not fully considered by the most of EEG-based emotion recognition studies.

Meanwhile, salient objects are recognized rapidly in a cluttered visual scene [18] by human, in other words, we can directly recognize region of interest in the scene. In [19], the authors propose a novel framework, named Transformer, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention. In EEG-based emotion recognition task, few of studies used attention mechanism to learn discriminative representations. Also, sequence information is important for representing human states [16]. However, there still exist few studies to model the characteristic of sequence for EEG-based emotion recognition.

From the literatures, to address the above-mentioned issues, we propose a novel deep framework, namely ENet, to take the baseline signals into account and transform the raw 1D chain-like EEG signals into 2D frame-like sequences.

The main contributions can be summarized as follows:

1 We propose an automated emotion recognition framework, namely ENet, which can effectively learn the discriminative characteristic information for assessing the human state.
2 To extract the spatial features, DCNN technology is adopted. Meanwhile, to mine the important information in the deep learned features of DCNN, soft attention mechanism is used.
3 To further extract the temporal patterns of EEG signals, ConvLSTM technology is adopted in our task.
4 Extensive experiments on the DEAP database demonstrate that the superior performance of the proposed framework when compared with most of the-state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 briefly discuss related works on EEG-based emotion recognition. Section 3 show the proposed hybrid framework. Section 4 introduce the adopted database and experimental results. Conclusion and future works are given in Section 5.

2. Related works

For EEG-based emotion recognition studies, a great number of researchers focused on design feature extraction methods. Especially in recent years, many researchers attempt to design different machine learning methods to model the human emotion state. In the following, we first briefly describe the methods for assessing the emotional state.

In [20], the authors propose a Bayesian network for EEG-based emotion recognition. The goal of the research is focused on representing uncertainty emotional states. Experiments conducted on the DEAP dataset demonstrated the effectiveness of the proposed method with the average classification rate 86.8% for arousal and 85.9% for valence, respectively.

In [21], the authors propose a novel method, named graph regularized sparse linear discriminant analysis (GraphSLDA), to address the emotion recognition problem. Based on the conventional linear discriminant analysis (LDA) method, GraphSLDA is proposed by imposing a graph regularization and a sparse regularization on the transform matrix of LDA. To learn the discriminative representations, some features are extracted to train the GraphSLDA model and also adopt it as classifier to test EEG signals. The raw EEG signals are divided into five frequency bands, i.e., δ, θ, α, β, and γ. The experimental performance demonstrate that the proposed method outperforms the most of the state-of-the-art methods on the SJTU emotion EEG dataset (SEED) database.

In [22], the authors propose an improved empirical mode decomposition (EMD) applying singular value decomposition (SVD)-based feature extraction method to extract the features coefficients of expansion based on all IMFs. Extensive experiments are conducted on four EEG datasets for assessing the severity of depression. The proposed method achieved promising performance, when compared with the pre-proposed EMD-based feature extraction method.

In [23], a multichannel EEG emotion recognition method based on a novel dynamical graph convolutional neural networks (DGCNN) is proposed. The DGCNN is to adopt a graph to model the multichannel EEG features. To valid the effectiveness of the proposed method, they conduct the experiments on the SEED and DREAMER dataset. The experimental results show that the proposed method obtains better recognition performance than the most of the state-of-the-art methods.

Wang et al. [5] compare three different kinds EEG features for

Table 1
DEAP dataset representation for each subject.

Array name	Array shape	Array contents
Data	40 × 40 × 8064	Video/trail × Channel × Data
Labels	40 × 4	video/trial × label(valence, arousal, dominance, liking)

emotion classification, adopt a feature smoothing method to remove the unrelated features of emotion task, and propose a novel method to tracking the trajectory of emotion changes by manifold learning. To valid the effectiveness of the proposed method, they develop a movie stimulation experiment to generate real emotional state. Based on the experimental results, they obtain the following conclusions: (1) power spectrum feature obtained superior performance to other two ones; (2) the proposed feature smoothing method can obviously improve the accuracy of emotion classification.

Li et al. [24] propose a novel method, named R2G-STNN, which adopt spatial-temporal neural network to learn discriminative representation with regional to global hierarchical procedure. In their work, they adopt a bidirectional long short term memory (BiLSTM) network to learn the spatial features of EEG electrodes. Meanwhile, R2G-STNN model equipped with a region-attention layer to learn a series of weights to strengthen or weaken the contributions of brain regions. Furthermore, to learn both regional and global spatial-temporal features, BiLSTM is also used. In order to valid the performance of the proposed approach, subject-dependent and subject-independent experiment are performed on the SEED database, and the experimental results demonstrate that the proposed approach obtains state-of-the-art performance.

In [25], the researchers adopt deep canonical correlation analysis (DCCA) for multimodal emotion recognition. To transform each modality separately and coordinate different modalities into a hyperspace, canonical correlation analysis constraints is used. They evaluate the performance of DCCA on five databases, and the experimental results show that DCCA obtains the-state-of-the-art performance.

Most of the above-mentioned approaches are not adopt attention mechanism to improve the emotion recognition accuracy. Furthermore, few of works fuse spatial and temporal information to analysis the emotional state. Therefore, in the present paper, we will introduce a novel framework, which can better mine discriminative information of the EEG-signals for emotion recognition. In other words, we propose a novel end-to-end framework based on DCNN and ConvLSTM with attention mechanism, for EEG-based emotion recognition.

3. Our method

Fig. 1 is outlined the proposed framework for EEG-based emotion recognition. The raw EEG signals are first converted into 2D EEG frame sequences. Then the 2D EEG frame sequences are feed into the proposed hybrid framework ENet. In the framework, DCNN and ConvLSTM along with attention mechanism can represent discriminative characteristic for emotion recognition. In the following, we present each procedure in detail.

3.1. Dataset

The DEAP dataset consists of two parts, the first is that the ratings from an online self-assessment where 120 one-minute extracts of music videos were each rated by 14–16 volunteers based on arousal, valence and dominance. Second, 32 volunteers watched a subset of 40 of the music videos, and participant ratings, physiological recordings and facial images are recorded in the experiment. 22 participants frontal facial videos are also recorded. As illustrated in Table 1, each participant has two arrays. The dataset contains 32 channel EEG signals and 8 channel peripheral physiological signals. EEG signals are only adopted in our study. In our experiment, the EEG signals are sampled at 512 Hz and then down-sampled at 128 Hz. EOG signals are not used in our work. A bandpass frequency filter from 4.0 to 45.0 is applied. 60 s trail data and 3 s baseline data are recorded of EEG. The length of emotional music videos is 40 min, and participants were asked to rate the levels of arousal, valence, liking and dominance for each video. For a detailed introduction of DEAP, please refer to [9]. In addition, to compare the performance of another work (i.e., [26], [27], [28], etc.), we use 5 as threshold to split the trials into two classes based on the rated levels of valence and arousal. Therefore, our study can be considered as two binary classification problems (i.e., low arousal and high arousal, low valence and high valence).

3.2. Data pre-processing

For raw EEG data, it has some noise during collecting the data procedure [9]. To further improve the recognition accuracy, we compute the differences between baseline signals and stimulation signals. Firstly, pre-trial signals are taken out from the all C channels and split into N segments (the length is Q). N matrices with the size of $C \times Q$ are generated. Secondly, we perform element-wise addition operation for all of these matrices, and compute the mean value of the matrices. Formally, the computational paradigm can be written as:

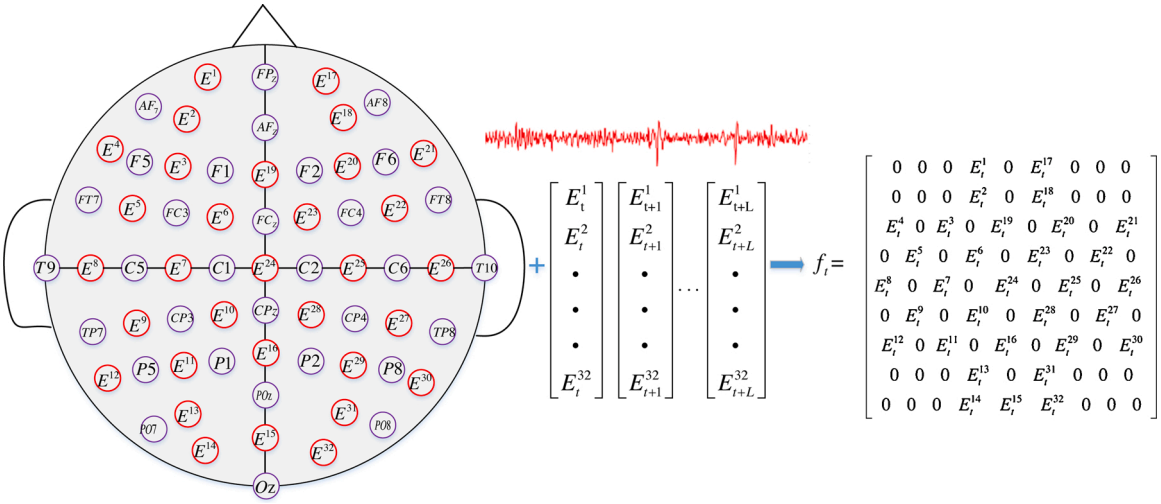

Fig. 2. The illustration of 1D data to 2D.

$$Mean = \frac{\sum_{i=1}^{N} M_i}{N} \tag{1}$$

where $M_i \in \mathbb{R}^{C \times Q}$ represents the ith matrix.

As mentioned above, the $C \times Q$ matrix can reflect subjects' original emotional state which is not stimulated by any material. Thirdly, we split the raw EEG signals into $P(C \times Q)$ matrices named REEG and subtract the *Mean* for every matrix. To simplify matter, the data that generated by subtraction are denoted as EEG_{sub}, it is can be written as:

$$\mathrm{EEG}_{\mathrm{subj}} = \mathrm{REEG}_j - Mean \tag{2}$$

Lastly, a final matrix is generated by concatenating all the EEG_{sub} matrices. The size of the final matrix is the same as the raw EEG signals. For a clear illustration, we give a detailed description of the three steps as shown in Fig. 1.

3.3. 1D EEG data to 2D frame sequences

As illustrated in Fig. 2, the 1D EEG data are converted into 2D data for emotion recognition. For DEAP database, when wearable device is used to record some physiological signals (i.e., EEG, ECG, etc.) based on international 10–20 system, it contains multiple electrodes. The 10–20 system is an international description and application of the identified method of the scalp electrode and the underlying region of the cerebral cortex, which based on the relationship between the location of an electrode and the underlying area of the brain, specifically the cerebral cortex.

In the present paper, the EEG data of DEAP is defined as follows. Let us assume that $E_t = [E_t^1, E_t^2, E_t^3, \ldots, E_t^n]^T$ is a 1D data vector at time t. Where n represents the number of channels and E_t^n denotes a data of the jth channel. In the current study, n is 32. The 10–20 System is exhibit at the left side of Fig. 2, it can be seen that the EEG electrodes are plot in red of the DEAP database. To further represent the discriminative characteristic information of EEG data, we convert the original data into a matrix with the size of $(h \times w)$, where h denotes the maximum point number of the vertical test points and w represents the maximum point number of the horizontal test points. h and w are set to 9 in our study. From the EEG electrode map, each electrode is physically adjacent to a plurality of electrodes in which the EEG signal is recorded in the brain, and the elements of the chained 1D EEG data are limited to two neighbors. To obtain a spatial characteristic information among multiple adjacent channels, 1D EEG data are converted into 2D EEG frames by the electrode map. Formally, the generated 2D data frame f_t can be written as follows:

$$\mathrm{f_t} = \begin{bmatrix} 0 & 0 & 0 & E_t^1 & 0 & E_t^{17} & 0 & 0 & 0 \\ 0 & 0 & 0 & E_t^2 & 0 & E_t^{18} & 0 & 0 & 0 \\ E_t^4 & 0 & E_t^3 & 0 & E_t^{19} & 0 & E_t^{20} & 0 & E_t^{21} \\ 0 & E_t^5 & 0 & E_t^6 & 0 & E_t^{23} & 0 & E_t^{22} & 0 \\ E_t^8 & 0 & E_t^7 & 0 & E_t^{24} & 0 & E_t^{25} & 0 & E_t^{26} \\ 0 & E_t^9 & 0 & E_t^{10} & 0 & E_t^{28} & 0 & E_t^{27} & \\ E_t^{12} & 0 & E_t^{11} & 0 & E_t^{16} & 0 & E_t^{29} & 0 & E_t^{30} \\ 0 & 0 & 0 & E_t^{13} & 0 & E_t^{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & E_t^{14} & E_t^{15} & E_t^{32} & 0 & 0 & 0 \end{bmatrix} \tag{3}$$

In Eq. (3), zero is the unused electrodes of DEAP. In our work, it has no effect on the performance of emotion recognition. After the above operation, he 1D data vector sequences $[E_t, E_{t+1}, \ldots, E_{t+L}]$ is converted into 2D frame sequences $[f_t, f_{t+1}, \ldots, f_{t+L}]$. To maintain the data consistency and improve the performance of emotion recognition, Z-score is adopted to compute $[f_t, f_{t+1}, \ldots, f_{t+L}]$ at same scale. Z-score can be represented as follows:

$$Z = \frac{x - \mu}{\sigma} \tag{4}$$

where x denotes a non-zero element at the frame, μ is the mean of the elements, and σ represents the stand deviation of the elements.

Finally, sliding window is used to segment the 2D frames into some sub-frames to capture the state of human emotion. The procedure is exhibited at the right side of Fig. 2. Every sub-frame is a sequence of 2D frames which has a inflexible length without any overlap among adjacent data frames. The procedure can be represented as:

$$\mathrm{Sub}_j = [f_t, f_{t+1}, \ldots, f_{t+\mathrm{Sub}-1}] \tag{5}$$

where Sub_j represents the jth segment of the data. In the present paper, we aim to develop a discriminative approach for recognizing the dimension of emotion (i.e. arousal, valence) from each data segment Sub_j.

3.4. Feature extraction

As shown in Fig. 1, DCNN and ConvLSTM technology as well as attention mechanism are used to model discriminate characteristic information for assessing the emotional state of human. In the following we introduce each step in detail.

Due to the size of small samples of DEAP dataset, convolutional operation of DCNN is used to model the features of EEG for emotion recognition. For the pipeline of DCNN part, 2D convolutional layers with kernel size of 4×4 is used to extract the spatial features. The reason why we use this kernel as the size of 4×4 is because for the valuable patterns among the multi-channels can be captured. For each convolutional layer, we use zero-padding to maintain the valuable information of the data frames. The feature map size of the first convolutional layer is 32. The size of the followed convolutional layers is 64, 128, respectively. After each convolutional layer, a batch normalization (BN) operation is adopted to accelerate deep network training by reducing internal covariate shift.

In addition, to model the sequence information of EEG, ConvLSTM is used in our work. The reason for using ConvLSTM is that its convolutional structures include input-to-state and state-to-state transitions, which can model spatialtemporal characteristic information quite well. Formally, the inputs, the cell states, the hidden states and the gates of ConvLSTM are 4D tensors whose first dimension denotes the time step, the second and the third are spatial dimensions (height, width), and the last dimension is the feature map. The computation of the hidden value h_t of a ConvLSTM cell is updated at every time t.

Let assume that " $\odot$", " $*$" and "σ" are the Hadamard product, convolutional operation and the sigmoid function. Formally, the ConvLSTM can be written as follows:

$$\mathrm{i}_t = \sigma_i(W_{\mathrm{ai}} * \mathrm{x}_t + W_{\mathrm{hi}} * \mathrm{h}_{t-1} + W_{\mathrm{ci}} \odot \mathrm{c}_{t-1} + \mathrm{b}_i) \tag{6}$$

$$\mathrm{f}_t = \sigma_f(W_{\mathrm{af}} * \mathrm{x}_t + W_{\mathrm{hf}} * \mathrm{h}_{t-1} + W_{\mathrm{cf}} \odot \mathrm{c}_{t-1} + \mathrm{b}_f) \tag{7}$$

$$\mathrm{c}_t = \mathrm{f}_t \odot \mathrm{c}_{t-1} + i_t \odot tanh(W_{\mathrm{ac}} * \mathrm{x}_t + W_{\mathrm{hc}} * \mathrm{h}_{t-1} + \mathrm{b}_c) \tag{8}$$

$$\mathrm{o}_t = \sigma_o(W_{\mathrm{ao}} * \mathrm{x}_t + W_{\mathrm{ho}} * \mathrm{h}_{t-1} + W_{\mathrm{co}} \odot \mathrm{c}_{t-1} + \mathrm{b}_o) \tag{9}$$

$$\mathrm{h}_t = \mathrm{o}_t \odot tanh(\mathrm{c}_t) \tag{10}$$

where $\mathbf{i}$,$\mathbf{f}$,$\mathbf{o}$ and $\mathbf{ch}_t$ represent the input gate, forget gate, output gate and cell activation 4D tensors, all of them have the same size as tensor h_t. W_{ai},W_{hi},W_{af},W_{hf}, W_{cf},W_{ac}, W_{hc}, W_{ao}, W_{ho} and W_{co} represent the weight matrices, with subscripts representing the relationships. For example, W_{ai} is the input–input gate matrix, W_{hi} represents the hidden-input gate matrix. However, $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_c$ and $\mathbf{b}_o$ are bias vectors.

To capture the valuable information of the data frame, attention mechanism is used in our task. After the concatenate operation, the data

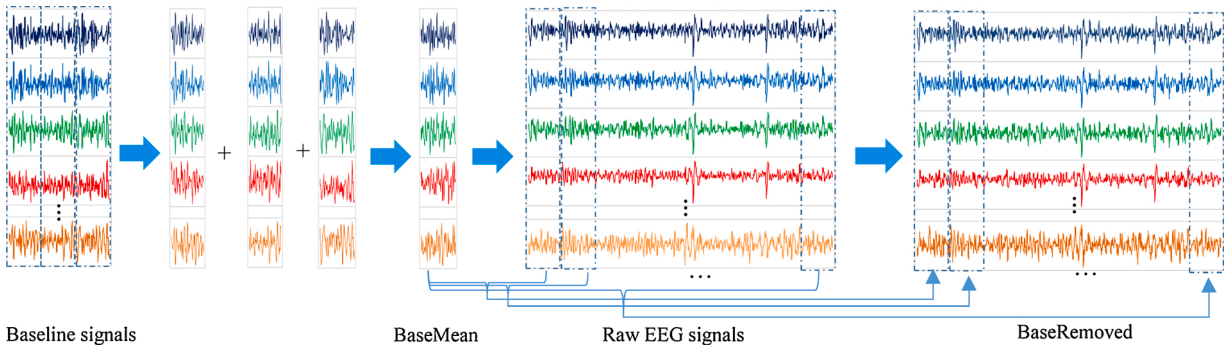

Fig. 3. The illustration of the pre-processing procedure.

with a size of $9 \times 9 \times (128 \times (t+S-1))$. We embed the AttentionNet into the framework to automatically learn the valuable emotional patterns. The detailed architecture of the AttentionNet is shown in the right orange dashed rectangle in Fig. 1. As can be seen in the rectangle, the AttentionNet is designed as follows. Two convolution operation, two inner productions and a sigmoid activation are included in the AttentionNet. The range of the sigmoid activation is between 0 and 1. 64 1×1 convolution kernels are used to compress the cube features with the size of $64 \times 9 \times 9$. Finally, the cube features are flattened into a spatial feature vector with the size of $\varphi \in \mathbb{R}^{5184}$. As a convolution operation, it can compute the features at different scale for reducing the dimension of feature maps.

In the present paper, we first pre-process the input 1D EEG signals into 2D data frame sequences. Let us assume that $Sub_j = [f_t f_{t+1}, \ldots, f_{t+\mathrm{Sub}-1}] \in \mathbb{R}^{S \times h \times w}$ is the jth input segment of DCNN, and each data frame denotes $f_k (k = t, t+1, \ldots, t+Sub-1)$. Every segment is feed into 2D-CNN and transformed into a spatial feature vector φ:

$$\varphi_j = \mathrm{Conv}(\mathrm{Sub}_j), \varphi_j \in \mathbb{R}^{5184} \tag{11}$$

For ConvLSTM part, we use two stacked ConvLSTM layers. The hidden state of the ConvLSTM unit in the first layer at time step t represents as h_t, and h_{t-1} is the hidden state of the previous time step $t-1$. The information from the previous time step that is conveyed to the current time step, and affects the final output. In this work, the hidden state of the ConvLSTM unit is adopted as its output. Hence, the second ConvLSTM layer is the hidden state sequence of the first ConvLSTM layer $[\mathbf{h}_t, \mathbf{h}_{t+1}, \mathbf{h}_{t+S-1}]$. As we concentrate on segment-level emotion recognition, and only utilize the last output h'_{t+S-1}, is input into the fully connected layer. For the data of ConvLSTM, we adopt 1D EEG data vectors as the input data. The jth inputted windowed segment of the ConvLSTM is:

$$\mathrm{ConvL}_j = [v_t, v_{t+1}, \ldots, v_{t+S-1}] \tag{12}$$

where v_t represents the vector at time step t, and S is the window size.

The hidden state at last time step $t+S-1$ in one segment is:

$$h'_{t+S-1} = \mathrm{ConvLSTM}(\mathrm{ConvL}_j), \quad h'_{t+S-1} \in \mathbb{R}^m \tag{13}$$

where m represents the hidden size of the ConvLSTM unit.

The concatenation operation of DCNN and ConvLSTM is performed after feature extraction. A fully connected layer is adopted before ConvLSTM layers to model the temporal representation ability.

Finally, the spatial and temporal features are concatenated into a joint spatialtemporal feature vector. Then a softmax layer is used to assess the human emotion states:

$$\chi_j = \mathrm{Softmax}([\varphi_j, h'_{t+S-1}]), \chi_j \in \mathbb{R}^n \tag{14}$$

where n represents the number of classes. In order to avoid over-fitting, we apply dropout operation as a form of regularization after fully connected layers in ConvLSTM part. Additionally, L2 regularization term is also used to cost function to improve the generalization capability of the training modal.

4. Experimental results

This section describes the experimental evaluation of the proposed framework for emotion recognition. In Section 4.1, we briefly introduce the experimental setup and evaluation measures. The experimental results are shown in Section 4.2.

4.1. Experimental setup

For EEG-based emotion recognition task, the length of window size is significant for assessing the emotion recognition. In our study, we consider the extensive works and follow the work [5] to adopt 1s window length for recognizing the human emotion state. In [5], the authors conduct extensive experiments and consider 1s to predict the emotion state. Therefore, 1s window is considered in the present paper. The sample rate of EEG data is 128 Hz, the pre-trial baseline signals are split into three matrices with the size of 32×128 to calculate the baseMean matrix. As illustrate in Fig. 3, we pre-process the raw EEG data to make the data more robust for emotion recognition. Above the pre-preprocessing procedure, for each trial, 32 channels of EEG signals are obtained. After using the 1s window, we get 60 segments as a trial. And the window size of S is 128. For every subject, we get a total of 2400 samples for each subject. Therefore, as shown in Fig. 2, the 2D data frames are transformed into the format with the size of 9×9. To evaluate the performance of the proposed framework, 5-fold and 10-fold cross-validation method are used, respectively. Average pooling method is also used to assess the overall performance on the 5-fold and 10-fold cross-validation. For the selection of the optimal number of cross folds, on the one hand, we follow the experience from other works. On the other hand, we try different numbers of cross validation to compare the performance of emotion recognition.

All experiments are conducted using Titan-X GPU with 12 GB memory. The networks are trained with stochastic gradient using TensorFlow deep learning framework with a batch size 100. The dropout is set to 0.5. The selection of parameters is based on our experience in training deep networks, without any special fine-tuning for the given dataset. The learning rate is set to 0.0001. To obtain the fast convergence, if the emotional accuracy exceeds 80% but less than 85%, the learning rate is set to 5×10^{-5}. As the accuracy exceed 85%, the learning rate is set to 5×10^{-6}. The number of iterations is set to 50.

4.2. Results and analysis

To evaluate the performance of the proposed framework, we conduct extensive experiments on the DEAP database. An ablation study is first performed to demonstrate the effectiveness of each component of the

Table 2
Performance (%) of each subject on valence and arousal using 5-fold validation.

Sub.	With attention		Without attention		Sub.	With attention		Without attention	
	Valence	Arousal	Valence	Arousal		Valence	Arousal	Valence	Arousal
S01	45.44	46.54	36.61	45.72	S17	39.25	40.69	35.40	36.03
S02	41.50	42.32	35.91	36.50	S18	44.69	44.97	39.27	41.52
S03	45.83	46.25	42.77	43.05	S19	44.43	45.03	**43.66**	38.32
S04	40.68	41.01	37.42	36.14	S20	46.27	46.50	38.99	41.43
S05	44.08	44.33	41.83	41.78	S21	43.39	44.34	36.42	40.19
S06	43.65	44.67	37.65	43.72	S22	44.68	44.87	42.90	42.86
S07	44.95	46.23	40.32	44.74	S23	45.93	46.03	37.56	41.18
S08	45.29	46.29	37.65	44.52	S24	44.82	44.84	41.09	41.42
S09	43.47	42.34	39.79	38.20	S25	43.76	44.68	37.73	40.71
S10	45.86	46.12	40.57	43.98	S26	43.41	42.93	38.71	40.65
S11	42.50	41.88	36.22	34.33	S27	**46.63**	**47.07**	43.23	43.18
S12	44.53	45.30	33.67	41.30	S28	44.10	44.81	38.67	42.16
S13	42.61	44.04	31.89	42.47	S29	46.51	45.27	40.78	43.28
S14	42.21	44.09	32.99	41.39	S30	45.77	46.34	43.72	44.72
S15	45.62	45.30	39.16	44.55	S31	43.92	44.69	38.83	44.45
S16	45.90	46.15	41.57	**46.12**	S32	44.77	44.22	36.81	40.05
Average accuracy results (meanstd)						44.361.76	44.621.62	38.743.05	41.582.95

Table 3
Performance (%) of each subject on valence and arousal using 10-fold validation.

Sub.	With attention		Without attention		Sub.	With attention		Without attention	
	Valence	Arousal	Valence	Arousal		Valence	Arousal	Valence	Arousal
S01	**93.08**	91.38	**90.48**	86.41	S17	77.77	79.58	66.75	64.48
S02	82.90	82.27	69.61	68.71	S18	87.63	88.18	75.97	78.02
S03	91.32	88.63	84.53	80.53	S19	89.33	86.65	82.63	69.16
S04	79.55	79.88	67.12	68.08	S20	92.02	89.46	89.32	78.85
S05	87.43	87.5	77	84.02	S21	89.25	84.4	82.58	79.90
S06	85.33	88.27	74.33	80.05	S22	87.55	88.83	81.53	82.63
S07	91.37	91.43	80.22	85.82	S23	90.73	89.08	76	76.70
S08	90.07	91.42	85.17	84.90	S24	89	88.27	83.30	82.60
S09	86.62	84.05	79.02	72.60	S25	86.72	87.77	80.85	76.67
S10	91.30	89.97	86.05	83.4	S26	85.17	84.63	70.25	72.67
S11	81.73	80.25	64.02	62.72	S27	91.37	92.48	80.48	79.40
S12	88.88	86.02	85.58	82.3	S28	85.88	88.80	70.93	83.35
S13	84.83	86.62	68.22	85.25	S29	92.37	91.38	83.55	83.50
S14	83.87	87.03	77.41	76	S30	89.73	**93.40**	70.68	86.70
S15	90.33	89.92	85.15	86.22	S31	87.92	89.73	81.47	85.87
S16	90.95	92.17	82.73	**89.83**	S32	89.11	86.78	85.27	69.85
Average Accuracy Results(meanstd)						87.843.73	87.693.61	78.697.07	78.977.06

proposed framework. The proposed model is further compared with the-state-of-the-art methods to show its superior emotion recognition performance.

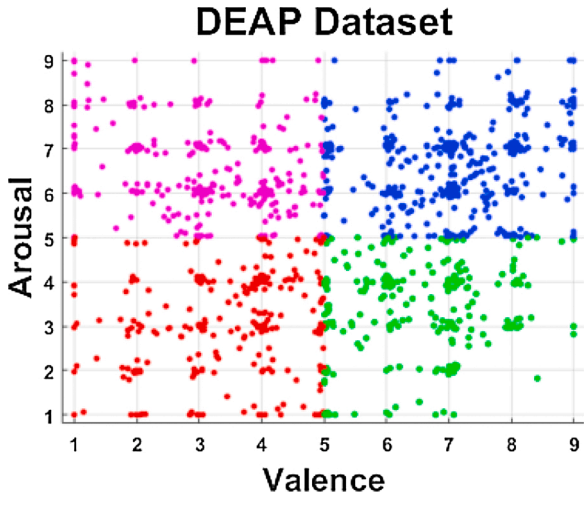

Fig. 4. Distribution of emotion classes in the DEAP dataset.

4.2.1. *Performance of individual subjects*

The results of emotion recognition accuracy on DEAP dataset are shown in Tables 2 and 3 , respectively. Firstly, we explore the performance using individual subjects with 5-fold and 10-fold cross validation experiment, respectively. From Table 2, one can notice that, when 5-fold cross validation is conducted for each subject, the average recognition accuracy with attention method on valence and arousal is 44.361.76% and 44.621.62%, respectively. While the performance of without attention on valence and arousal is 38.743.05% and 41.582.95%, respectively. This observation indicates that the attention mechanism is important for emotion analysis, and the DCNN-ConvLSTM with attention mechanism can characterize the emotion state well. From Table 3, when 10-fold cross validation experiment is conducted for each subject, one can see that, the average recognition accuracy with attention method on valence and arousal is 87.843.73% and 87.693.61%, respectively. While the results of DCNN-ConvLSTM method on arousal and valence is 78.697.07% and 78.977.06%, respectively. The results are shown in the last row in Tables 2 and 3, for 5-fold and 10-fold cross validation, respectively. Moreover, for the DCNN-ConvLSTM-attention and DCNN-ConvLSTM method, no matter what the method is adopted, the subject S01 obtain the best performance on arousal of 93.08% and 90.48, respectively. While for the arousal dimension, the subject S30 obtain the best performance of the DCNN-ConvLSTM-attention method with 93.40. The subject S16 get the best results of 89.83. These results

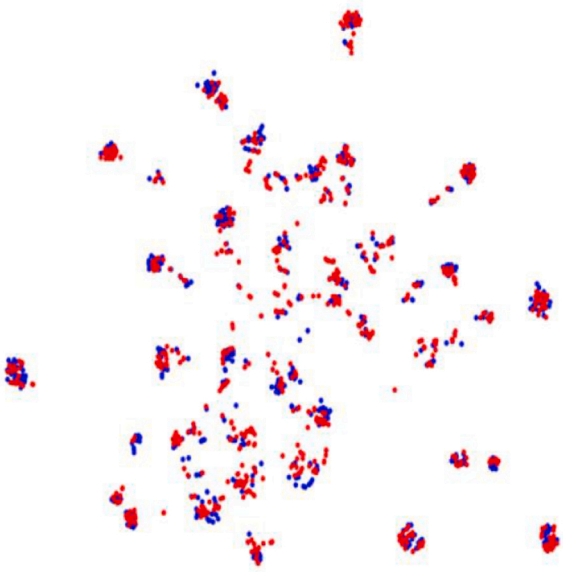

Fig. 5. t-SNE on two valence classes (high-valence in red and low-valence in blue).

Table 4
Emotion recognition results comparison to other method.

Methods	Arousal	Valence
Zhang et al. [27]/deep belief network (DBN)	64.2	58.4
John et al. [29]/minimum-redundancy-maximum-relevance (mRMR)	73.06	73.14
Zhong et al. [30]/Transfer recursive feature elimination (T-RFE)	78.67	78.75
Hao et al. [26]/deep neural networks (DNN)	83.23	83.82
Wei et al. [31]/Bimodal Deep AutoEncoder (BDAE)	80.5	85.2
Luo et al. [32]/spiking neural networks (SNN)	74	78
Yin et al. [33]/LSTM	84	85
Du et al. [34]/LSTM	72	69
Ours Approach (10-fold)	**87.69**	**87.84**

also demonstrate the effectiveness of proposed framework, which is capable of learning the discriminative information for emotion recognition from 2D EEG data frames. Furthermore, the comparison with subject-dependent also shows the effectiveness of the proposed approach. Especially, one can notice that, the accuracy of emotion can improve when attention mechanism is used. The reason is that self attention mechanism can represent the important patterns from the features. The attention mechanism can assign different weights for the features. Specifically, the high weights are corresponding to the important features, and vice versa. Moreover, one can notice that the number of cross folds for validation from 5 to 10 can improve the accuracy for emotion recognition. The main reason is that the number of cross folds 10 can represent more discriminative patterns than the number of cross folds 5. Furthermore, different individuals have respective representations when recording the EEG data samples. Also, different internal representations are also roused when face the stimuli videos or audios.

In addition, Fig. 4 shows the distribution of self-reported valence and arousal for the DEAP dataset. One can notice that the DEAP database has a higher concentration of trials closer to neutral emotion, i.e., near the center of the graph.

Fig. 5

Furthermore, to visualize class-separability when using the proposed framework, in Fig. 4, one can notice that, a better classification degree is obtained.

4.2.2. Comparison with previous methods

Finally, we compare our approach with other methods on DEAP dataset. For a fair comparison, we show the results that are only considering the binary classification issue for emotion recognition in Table 4. From the table, one can notice that, our approach achieves better performance than the listed methods on DEAP dataset. This further shows the effectiveness of our proposed approach for depression recognition.

5. Conclusion

In this paper, an novel framework ENet is proposed for emotion recognition. ENet consists of two branches. The first branch is DCNN. ENet utilizes DCNN and attention mechanism to model the spatial patterns of EEG signals. In addition, to represent the temporal features, ConvLSTM with attention mechanism are adopted. The second branch is ConvLSTM, which can model the temporal information from original EEG signals. After that, the features of DCNN and ConvLSTM are concatenated for final emotion recognition. Extensive experimental results have demonstrated that the effectiveness of the proposed framework, when compared with the most of the state of the art emotion recognition methods based on EEG data.

In the future, we will explore various hand-crafted and deep-learned features for EEG-based emotion recognition. In addition, we will study multimodal (i.e., audio, video, text, etc) cues for emotion analysis.

Declaration of Competing Interest

The authors report no declarations of interest.

References

[1] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikä inen, Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition, IEEE Trans. Affect. Comput. 10 (1) (2017) 32–47.

[2] J. Yan, W. Zheng, M. Xin, J. Yan, Integrating facial expression and body gesture in videos for emotion recognition, IEICE Trans. Inform. Syst. 97 (3) (2014) 610–613.

[3] E. Juárez-Castillo, H.G. Acosta-Mesa, J. Fernandez-Ruiz, N. Cruz-Ramirez, A feature selection method based on a neighborhood approach for contending with functional and anatomical variability in FMRI group analysis of cognitive states, Intel. Data Anal. 21 (3) (2017) 661–677.

[4] Y. Guo, H. Nejati, N.-M. Cheung, Deep neural networks on graph signals for brain imaging analysis, in: 2017 IEEE International Conference on Image Processing (ICIP), IEEE (2017) 3295–3299.

[5] X.-W. Wang, D. Nie, B.-L. Lu, Emotional state classification from EEG data using machine learning approach, Neurocomputing 129 (2014) 94–106.

[6] S. Poria, E. Cambria, A. Hussain, G.-B. Huang, Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.

[7] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, I. Zunaidi, D. Hazry, Eeg feature extraction for classifying emotions using fcm and fkm, Int. J. Comput. Commun. 1 (2) (2007) 21–25.

[8] L. Piho, T. Tjahjadi, A mutual information based adaptive windowing of informative EEG for emotion recognition, IEEE Trans. Affect. Comput.

[9] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis;using physiological signals, IEEE Trans. Affect. Comput. 3 (1) 18–31.

[10] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2011) 42–55.

[11] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International conference on multimodal interaction, ACM (2013) 543–550.

[12] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM (2015) 435–442.

[13] S. M. Alarcao, M. J. Fonseca, Emotions recognition using eeg signals: A survey, IEEE Trans. Affect. Comput.

[14] Y.R. Tabar, U. Halici, A novel deep learning approach for classification of eeg motor imagery signals, J. Neural Eng. 14 (1) (2016) 016003.

[15] D. Zhang, L. Yao, X. Zhang, S. Wang, W. Chen, R. Boots, Eeg-based intention recognition from spatio-temporal representations via cascade and parallel convolutional recurrent neural networks, arXiv preprint arXiv:1708.06578.

[16] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, 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) 352–359.
[17] Y. Li, J. Huang, H. Zhou, N. Zhong, Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks, Appl. Sci. 7 (10) (2017) 1060.
[18] L. Itti, C. Koch, Computational modelling of visual attention, Nat. Rev. Neurosci. 2 (3) (2001) 194.
[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems (2017) 5998–6008.
[20] X. Zhang, D. Cao, P. Moore, J. Chen, L. Zhou, Y. Zhou, X. Ma, A bayesian network (bn) based probabilistic solution to enhance emotional ontology, in: in: Human Centric Technology and Service in Smart Space, Springer, 2012, pp. 181–190.
[21] Y. Li, W. Zheng, Z. Cui, X. Zhou, A novel graph regularized sparse linear discriminant analysis model for eeg emotion recognition, in: in: International Conference on Neural Information Processing, Springer, 2016, pp. 175–182.
[22] J. Shen, X. Zhang, B. Hu, G. Wang, Z. Ding, An improved empirical mode decomposition of electroencephalogram signals for depression detection, IEEE Trans. Affect. Comput.
[23] T. Song, W. Zheng, P. Song, Z. Cui, Eeg emotion recognition using dynamical graph convolutional neural networks, IEEE Trans. Affect. Comput.
[24] Y. Li, W. Zheng, L. Wang, Y. Zong, Z. Cui, From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition, IEEE Trans. Affect. Comput.
[25] W. Liu, J.-L. Qiu, W.-L. Zheng, B.-L. Lu, Multimodal emotion recognition using deep canonical correlation analysis, arXiv preprint arXiv:1908.05349.
[26] H. Tang, W. Liu, W.-L. Zheng, B.-L. Lu, Multimodal emotion recognition using deep neural networks, in: in: International Conference on Neural Information Processing, Springer, 2017, pp. 811–819.
[27] Z. Peng, L. Xiang, Y. Hou, G. Yu, D. Song, B. Hu, Eeg based emotion identification using unsupervised deep feature learning.
[28] Z.W.-L. Wei, Liu, L. Bao-Liang, Emotion recognition using multimodal deep learning, in: in: International Conference on Neural Information Processing, Springer, 2016.
[29] J. Atkinson, D. Campos, Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers, Expert Syst. Appl. 47 (2016) 35–41.
[30] Z. Yin, Y. Wang, L. Liu, W. Zhang, J. Zhang, Cross-subject eeg feature selection for emotion recognition using transfer recursive feature elimination, Front. Neurorobot. 11 (2017) 19.
[31] W. Liu, W.-L. Zheng, B.-L. Lu, Emotion recognition using multimodal deep learning, in: in: International conference on neural information processing, Springer, 2016, pp. 521–529.
[32] Y. Luo, Q. Fu, J. Xie, Y. Qin, G. Wu, J. Liu, F. Jiang, Y. Cao, X. Ding, Eeg-based emotion classification using spiking neural networks, IEEE Access 8 (2020) 46007–46016.
[33] Y. Yin, X. Zheng, B. Hu, Y. Zhang, X. Cui, Eeg emotion recognition using fusion model of graph convolutional neural networks and lstm, Appl. Soft Comput. 100 (2021) 106954.
[34] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, H. Wang, An efficient lstm network for emotion recognition from multichannel eeg signals, IEEE Trans. Affect. Comput.