Applied Intelligence
https://doi.org/10.1007/s10489-022-04228-2

Bi-hemisphere asymmetric attention network: recognizing emotion from EEG signals based on the transformer

Xinyue Zhong [1,2] · Yun Gu [2] · Yutong Luo [1,2] · Xiaomei Zeng [1,2] · Guangyuan Liu [1,2]

Accepted: 27 September 2022

Abstract
EEG-based emotion recognition is not only an important branch in the field of affective computing, but is also an indispensable task for harmonious human–computer interaction. Recently, many deep learning emotion recognition algorithms have achieved good results, but most of them have been based on convolutional and recurrent neural networks, resulting in complex model design, poor modeling of long-distance dependency, and the inability to parallelize computations. Here, we proposed a novel bi-hemispheric asymmetric attention network (Bi-AAN) combining a transformer structure with the asymmetric property of the brain's emotional response. In this way, we modeled the difference of bi-hemispheric attention, and mined the long-term dependency between EEG sequences, which exacts more discriminative emotional representations. First, the differential entropy (DE) features of each frequency band were calculated using the DE-embedding block, and the spatial information between the electrode positions was extracted using positional encoding. Then, a bi-headed attention mechanism was employed to capture the intra-attention of frequency bands in each hemisphere and the attentional differences between the bi-hemispheric frequency bands. After carring out experiments both in DEAP and DREAMER datasets, we found that the proposed Bi-AAN achieved superior recognition performance as compared to state-of-the-art EEG emotion recognition methods.

Keywords Cerebral hemispheric asymmetry · DEAP dataset · DREAMER dataset · EEG emotion recognition · Transformer

1 Introduction

EMOTION is a unique psychological phenomenon in humans that permeates all aspects of human life. With the development of artificial intelligence, a long-awaited dream of realizing harmonious human–computer interaction through computers' intelligent perception and understanding of human emotions has become a focus [1]. Consequently, emotion recognition has become an important branch of affective computing and has been widely studied by neuroscience and computer science research communities [2]. In addition, emotion recognition plays an important role in clinical healthcare and brain research.

Generally speaking, there are two methods of measuring emotional states: extrinsic and intrinsic. Compared to extrinsic responses, such as facial expressions [3], posture, and voice, physiological signals, such as those recorded on an electrocardiogram (ECG), electroencephalogram (EEG) and electromyogram (EMG), produced by humans in different situations reflect emotions more directly and objectively. Among these, EEG signals, which are triggered by voltage fluctuations caused by the flow of ionic currents in brain neurons [4], has become the preferred method for measuring emotional responses, because it is non-invasive, inexpensive, and easy to obtain [5]. In the last decade, research on emotional patterns based on EEG signals has been widely carried out, because the electrode position, temporal information, and frequency band of EEGs contain a large amount of emotional information [6], and is a reliable way to decode emotions.

Typically, the EEG emotion recognition method consists of two major steps: EEG feature extraction and their classification [5]. The extracted EEG features can be divided into three categories: The first type is the time domain feature, which captures temporal information, such as event-related potentials (ERPs) and high-order crossover features (HOCs). The second category is the frequency domain feature, which extracts EEG features from each frequency band, including differential entropy (DE) and power spectral density (PSD). The

✉ Guangyuan Liu
liugy@swu.edu.cn

[1] Institute of Affective Computing and Information Processing, Southwest University, Chongqing, China

[2] School of Electronic and Information Engineering, Southwest University, Chongqing, China

Published online: 14 November 2022

third category is time–frequency domain features, including the Hilbert-yellow spectrum (HHS) and amplitude-squared coherent estimation (MSCE). There are a large number of commonly used classifier families that are used to deal with the classification problem in the field of emotion recognition: Bayesian, support vector machines, decision trees, etc. [5, 7–10]. These classifiers are often followed by innovative ways of feature selection and preprocessing [9, 10]. Nonetheless, these traditional machine learning algorithms, whose input EEG features require experts to make subjective selection through experience and trial, can no longer meet the requirements of real-time emotion recognition, because the process remains time-consuming.

Under such circumstances, due to the rapid development of deep learning in recent years, EEG emotion recognition has achieved a breakthrough. Prior to this breakthrough, deep learning has achieved better performance in natural language processes (NLP), computer vision (CV), etc., which have attracted widespread attention in various fields. Consequently, deep learning networks have also been applied to EEG emotion recognition and have been proven to outperform traditional machine learning algorithms. Over the past few years, some scholars have applied classic CNN models, such as VGG and AlexNet, to EEG emotion recognition tasks [11–17]. In addition, a number of authors combined CNNs with long and short-term memory networks (LSTM) to construct a hybrid deep neural network for classifying EEG sequences [18–23]. In particular, Tao et al. [18] first adopted the attention mechanism for CNN and recurrent neural networks (RNN), named attention-based convolutional recurrent neural network (ACRNN). This enables the extraction of more discriminative spatiotemporal features of the EEG. Likewise, Liu et al. [24] also combined a 3D-CNN with an EEG channel attention-weighted learning module to extract the dynamic and spatial relations among EEG signals.

However, although these data-driven deep neural networks have achieved very high accuracy, some problems remain. First, most models are based on CNNs, or on CNNs combined with other networks, such as RNNs. Single-layer convolution cannot capture long-distance features [25, 26], and deepened multi-layer convolution requires large-scale superparameter adjustment, leading to a large number of calculations and slow training speed. At the same time, the pooling layer will cause a lack of positional information in the EEG signal, but is essential for the EEG signal. If the pooling layer is discarded to preserve positional information, it will increase the number of calculations. In addition, RNNs, another commonly used neural network, is essentially a Markov decision recursion. Owing to sequence dependency, long sequences cannot be processed in parallel and these models train slowly [27, 28].

Recently, the transformer [28] that rely entirely on the self-attention mechanism have been shown to be effective in drawing global dependencies on sequences. Compared with RNNs and CNNs, its maximum path-length is shorter [29], making remote dependencies easier to learn. Moreover, they have strong general modeling capabilities and better interpretability [30]. In a recent study [31], the transformer and BERT architectures used for language modelling were applied to learn from massive amounts of EEG data. It has been proven that the transformer can be fine-tuned to a variety of downstream brain–computer interfaces (BCI) and EEG classification tasks, outperforming previous approaches. Previously, revolutionary improvement of NLP and CV was promoted, implying that NLP and CV could be unified under a transformer structure, making the joint modeling of images and language easier [32]. However, emotion recognition, an indispensable part of human–computer interaction, has not been promoted to use this revolutionary deep learning framework. In particular, for real-time emotion recognition via physiological signals, no study has employed the overall structure of the transformer, which has resulted in slow progress in the joint modeling of physiological signals, facial expressions, and language in emotion recognition research. If the EEG-based emotion recognition task were to be extended to the transformer and could achieve good performance, the modeling and learning of physiological signal-based real-time emotion recognition, facial expression recognition, and language emotion recognition could be integrated, thereby accelerating the progress of these respective fields. A multi-modal recognition system of physiological signals and explicit emotional signals would be more sophisticated and would promote harmonious and synchronous human–computer interaction.

It is also worth noting that most deep learning emotional classifiers lack interpretability, relying entirely on blindly extracted features [33]. Hence, the utilization of the emotional asymmetry properties of the brain to improve emotion recognition performance has recently attracted considerable attention. From the perspective of neuroscience, the left and right hemispheres of the human brain are not completely symmetrical in terms of different emotions [34]. As early as half a century ago, Dimond et al. discovered that “emotional vision” differs between the right and left hemispheres of the brain [35]. Later, Davidson et al. proved that the EEG signals of the left frontal cortex are closely linked to positive emotions, while the right frontal cortex is more closely related to negative emotions [36, 37]. Consequently, the lateralization of the human brain has led scholars to combine the physiological significance of emotion recognition algorithms in an attempt to increase the accuracy of emotion recognition. Yang et al. [38] proposed a bi-hemispheric domain adversarial neural network, which includes a global domain and two local domain discriminators. They used the classifier to learn the distinctive emotional characteristics of each hemisphere. Similarly, to capture and amplify the different responses of the two hemispheres to emotional stimuli, Huang et al. [39] constructed three different EEG feature matrices and input them into a

three-layer CNN to extract spatial and temporal features. Nevertheless, the majority of deep learning methods currently used in EEG emotion recognition are black box systems lacking interpretability, are not integrated with the physiological mechanism of the human brain, ignore the fundamentals of brain neuroscience, and thus yield blind extracted features that are difficult to visualize.

Inspired by the transformer and the asymmetry properties of the brain hemispheres, we here propose a bi-hemispheric asymmetric attention network (Bi-AAN) that appropriately combined the transformer with the emotional asymmetry characteristic of the brain. No previous emotion recognition model has completely relied on the attention mechanism to calculate the original EEG signal and emotion representation, without using a sequence alignment RNN or convolutional layer. By allocating different attention weights to the left and right brains, the difference in emotional responses between the two hemispheres can be simulated. To achieve this goal, the transformer architecture was applied to emotion recognition tasks. First, the original EEG signal was converted to the frequency domain through a DE-embedding block. Second, we changed the multi-head attention mechanism to a bi-head, representing the left and right hemispheres; hence, we calculated the internal attention of each hemisphere's frequency bands and the attention difference between frequency bands of the two hemispheres. After capturing and fusing internal and differential relationships, more discriminative emotional representations were extracted. We evaluated the proposed model using the DEAP public dataset [1] and DREAMER dataset [40] deriving superior performance. Our primary contributions are summarized as follows:

1. We combined the transformer structure and brain lateralization property to EEG emotion recognition and proposed the Bi-AAN model depending on the attention mechanism, to extract the emotional bi-hemispheric asymmetry features globally for the first time.
2. Our Bi-AAN model employed bi-headed attention to model the emotional response discrepancy between the left and right brains, capturing the internal attention of each hemisphere's frequency bands and the attention difference between frequency bands of bi-hemispheres to extract more discriminative features from EEG signals.
3. We conducted subject-dependent and subject-independent experiments using the DEAP and DREAMERdatasets. The experimental results showed that the average accuracies can represent state-of-the-art performance.

The remainder of this paper is organized as follows: In Section 2, we provide a brief description of the preliminary work. In Section 3, we present the proposed bi-AAN model for EEG emotion recognition. The experiments are illustrated and discussed in Sections 4 and 5, respectively. Section 6 implys the future works and in Section 7, we conclude the paper.

2 Preliminaries

2.1 Differential entropy

As a commonly used EEG signal feature extraction method, DE has proven to be a powerful tool for measuring the complexity of continuous random variables in emotion recognition tasks, as continuous random variables can be discretized by DE [41–44]. For instance, the EEG signal sequence was divided into tiny sessions by Δx. According to the mean value theorem, a value x_i can always be found to establish (1):

$$\int_{i\Delta x}^{(i+1)\Delta x} p(x)dx = p(x_i)\Delta x \tag{1}$$

where $p(\Delta x)$ represents the probability density function of the discrete signals. Eq. 1 can be substituted into the discrete Shannon formula, shown as (2), if each point at i is assigned to x_i.

$$\begin{aligned} h(\Delta x) &= -\sum_{i=1}^{n} p(x_i)\Delta x \; ln[p(x_i)\Delta x] \\ &= -\sum_{i=1}^{n} p(x_i)\Delta x \; lnp(x_i) - \sum_{i=1}^{n} p(x_i)\Delta x ln\Delta x \end{aligned} \tag{2}$$

According to Lopida's law, the right side of (2) approximates 0 as Δx approaches 0. Therefore, the DE of a continuous signal can be defined as:

$$h(X) = -\int_X f(X)\log[f(x)]dx \tag{3}$$

where X is a random variable and $f(X)$ denotes the probability density function of X. Assuming that series X obeys the Gaussian distribution $N(\mu, \delta^2)$, the DE can be expressed as:

$$\begin{aligned} h(X) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\delta^2}} \log\left[\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{\delta^2}}\right] dx \\ &= \frac{1}{2}\log\left(2\pi e\sigma^2\right) \end{aligned} \tag{4}$$

For a specific frequency band, DE can be defined using the following formula:

$$h_i(X) = \frac{1}{2}\log\left(2\pi e{\sigma_i}^2\right) \tag{5}$$

where h_i and σ_i represent the differential entropy and signal variance in the corresponding frequency band i, (respectively)

2.2 Transformer

For a long time, the dominant sequence transduction models have mainly been based on complicated RNNs or CNNs to

form the encoder and decoder, while the one of the best-performing models relies on the attention mechanism to connect the encoder and decoder [45–47]. Inspired by this, the transformer was proposed by the Google team in 2018, eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output, where the self-attention relates to varying positions of a single sequence to calculate a representation of the sequence [28]. Moreover, the transformer based on the attention mechanism rather than on sequence-aligned recurrence allow significantly more parallelization, making it easier to learn dependencies between distant locations. To date, the transformer with powerful representation capabilities have made major breakthroughs in a number of artificial intelligence fields, such as NLP [48], CV [49, 50], and audio processing [51, 52]. Recently, an increasing number of studies have indicated the powerful potential of the transformer in various tasks [53, 54].

3 BI-AAN for EEG emotion recognition

In this section, the framework of the proposed Bi-AAN for EEG emotion recognition is introduced. Next, the preprocessing of the raw EEG signal is shown. Finally, the construction and algorithm of Bi-AAN are described in detail.

3.1 Framework of bi-AAN model

The core issue in emotion recognition is that subjects generate different subjective emotional states when facing the same stimulus. Several theories have been proposed by various scholars to describe emotions. Typical emotional models include discrete and dimensional models.

For discrete emotion description [55], emotions are classified into a set of discrete states, for example, there are six basic emotions (joy, sadness, surprise, fear, anger, and disgust). Contrary to the discrete model, the dimensional method maps emotions in a continuous axis, where emotions are often characterized by two dimensions (valence and arousal). The valence dimension aims to describe the degree of positivity or negativity, while the arousal dimension mainly represents the degree of excitement or apathy of the emotions. Based on the dimensional model of emotions, our research proposes an end-to-end data-driven EEG emotion recognition deep learning approach using Bi-AAN, as shown in Fig. 1. The proposed model consists of the following steps.

1. Data preprocessing: Initially, we removed the baseline EEG data spontaneously generated by the brain, to prevent the resting state data from interfering with the extraction of emotional features. Several researchers have found that baseline removal preprocessing can improve the accuracy of EEG emotion recognition [18, 41]. Therefore, the same operation was performed during our preprocessing. Subsequently, a sliding Hamming window is used to segment each trial of the EEG signal into a set of blocks, in which every block contains a 1-s EEG signal. Adjacent blocks overlap by half.
2. Data division: We scrambled the acquired samples and divided them into training and test sets.
3. Training: We constructed an EEG emotion recognition model that relied entirely on the attention mechanism to calculate the emotional features, without sequence alignment RNN or CNN (Bi-AAN), which is composed of a DE-embedding vector, bi-head attention mechanism, and two dense layers. The bi-headed attention mechanism assigned different attention weights to the left and right brain for different valence and arousal levels, by calculating the attention between the frequency bands of the two hemispheres. The extracted asymmetric features of the hemispheres are used to predict low/high arousal or negative/positive valence. Training samples were used to train the Bi-AAN model. Then, the cross-entropy loss was calculated, and the network parameters were updated using the Adam optimizer.
4. Test: Next, we used the trained model to predict the emotional state of the test sample to validate the final performance of the model.

3.2 Construction of bi-AAN model

The aim of the proposed Bi-AAN is to model the discrepancy in emotional responses between the two hemispheres by assigning different attention weights to the left and right hemispheres, and extracting the difference between the hemispheres. To achieve this goal, we referred to the architecture of the transformer by changing the multi-head attention mechanism to dual heads that represent the left and right hemispheres, respectively. Thus, we calculated the internal attention of each hemisphere's frequency bands and the attention difference between the frequency bands of the two hemispheres, which further enhanced the ability to discriminate EEG characteristics. Figure 2 illustrates the structure of the Bi-AAN model, which is composed of the following main parts: (1) DE-embedding block, (2) position-encoding module, (3) transformer encoder, and (4) two dense layers. The Bi-AAN model is specified on the DE-embedding block, position-encoding module, and transformer encoder.

3.2.1 DE-embedding block

The DE-embedding block is used to convert the raw EEG temporal slices to the frequency domain, thereby transforming

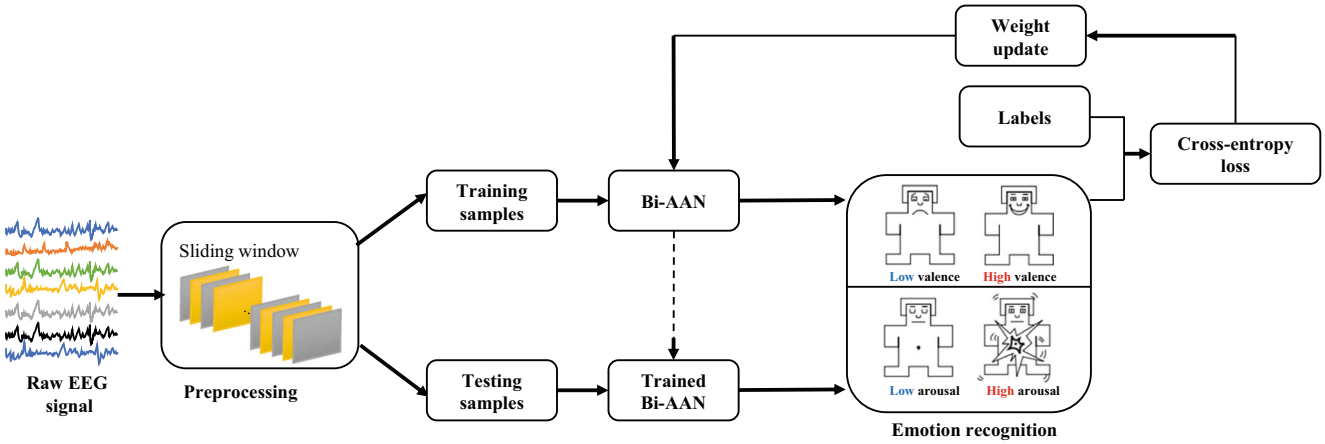

Fig. 1 Overview framework of proposed Bi-AAN for EEG based emotion recognition

the signal into DE-embedding vectors that are added to the subsequent position-encoding step. It has been confirmed in previous studies [38, 42, 56, 57] as the best feature extraction method for deep learning classification in the frequency domain.

Assuming that the recorded EEG signals with M Hz sampling frequency and duration T can be denoted as $X_R \in \mathbb{R}^{P \times Q}$, where P is the number of channels and Q represents the sampling points. The total number of EEG samples is n, and S = S_1, S_2, …, S_n represents the preprocessed EEG temporal slices. The input data of the DE-embedding block $S_i = [s_1, s_2, \ldots, s_P]$ (i = 1, 2, …, P) is the i-th EEG temporal slice, where s_j (j = 1, 2, …, P) represents the j-th channel of the sample S_i. Because the sampling rate is M Hz and the length of the slice window is T_s, the input sample S_i is denoted as $S_i \in \mathbb{R}^{P \times H * T_s}$. This module extracts the DE features for each frequency band on each channel. The calculation process is as follows:

$$H(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx = \frac{1}{2}\log\left(2\pi e\sigma^2\right) \tag{6}$$

where the time series EEG signal X obeys the Gauss distribution $N(\mu, \sigma^2)$, e and π are constants. DE-embedding was extracted from four frequency bands: θ band (4–7 Hz), α band (8–13 Hz), β band (14–30 Hz) and γ band (31–50 Hz).

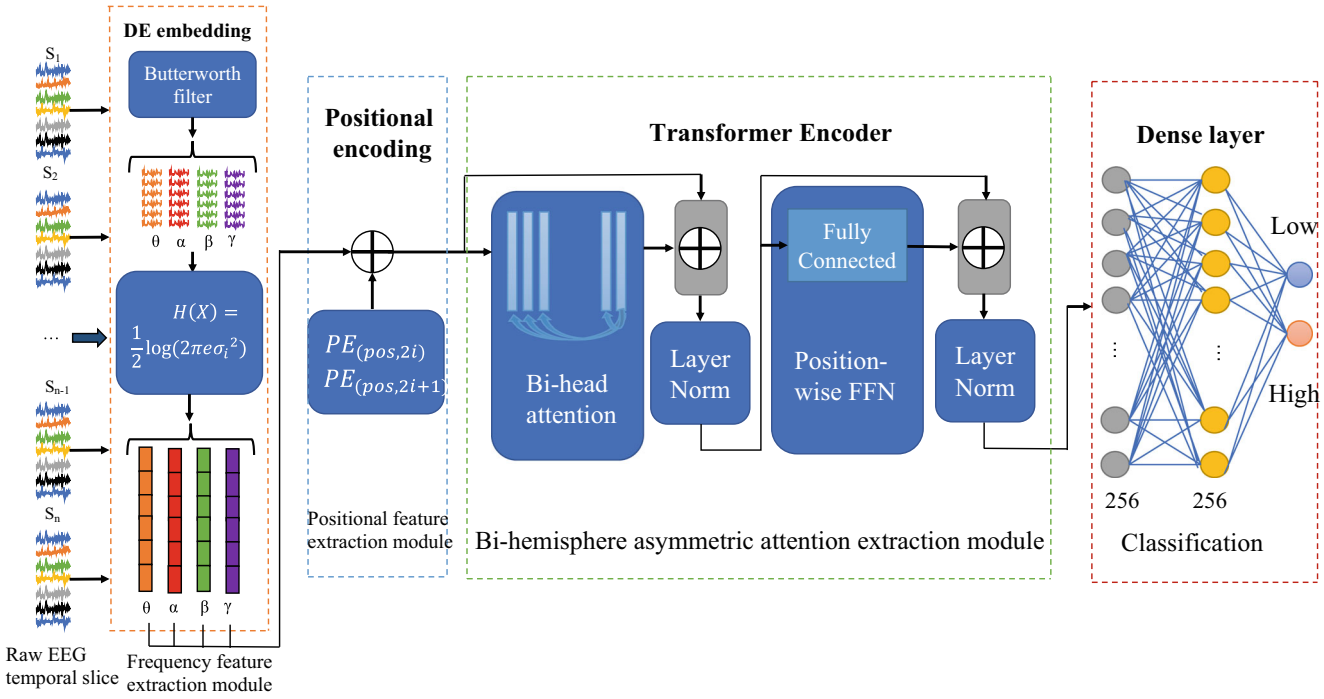

Fig. 2 Structure diagram of the proposed Bi-AAN

Thus, a DE-embedding vector, $H_{DE} \in \mathbb{R}^{bands \times P}$, with four bands, is generated. This step is inspired by previous studies [22, 38, 42–44] that used handcrafted DE features to alleviate redundant information of the original EEG signal before deep feature learning. Correspondingly, the DE feature extraction in our model was integrated with the input-embedding process of the transformer, resulting in a unique DE-embedding block that plays a role in feature extraction and dimensionality reduction of raw EEG data. Additionally, the integration of this feature extraction and deep network module makes the entire emotion recognition a thorough end-to-end data-driven process, which does not rely on the experience of experts, and also eliminates time-consuming manual feature extraction and selection.

3.2.2 Positional-encoding

Because the Bi-AAN model does not include a convolutional layer or a recursive layer, it lacks the ability to capture positional information, and thus the order of the EEG electrodes is missing. However, in the EEG electrode distribution map, each electrode is physically adjacent to multiple electrodes to record the EEG signal of a specific brain area; thus, the position of the EEG electrode contains information related to emotions [41]. In order to preserve the spatial information between adjacent channels, positional encoding, a unique part of the transformer framework, is added to the DE-embedding vector, which supplements the defect that the attention mechanism itself cannot capture positional information. Specifically, the position is encoded using the following equation of sine and cosine functions:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right) \tag{7}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right) \tag{8}$$

where pos represents the position of the channel in the EEG sequence, and i is the current dimension of the encoded position.

In this way, each element of the position-encoding process corresponds to a sinusoid, which makes the input vector levels of the encoder completely equal. The smaller the PE value, the longer is the wavelength. In addition, it allows the transformer model to learn to participate through relative positions and to extrapolate to longer sequence lengths during inference.

3.2.3 Transformer encoder

A transformer [28] is a sequence-to-sequence model that was originally used in the field of NLP for machine translation tasks. It consists of an encoder and a decoder module. In general, the encoder consists of N = 6 stacks of the same layer. Each layer has two sub-layers: a multi-head self-attention mechanism and a position-wise fully-connected feedforward network. To build a deeper model, a residual connection [58] is adopted around each module, followed by layer normalization [59]. To model the discrepancy in emotional response between hemispheres, the proposed Bi-AAN only employs an encoder session to extract the difference in attention.

First, after extracting the frequency and positional features, the EEG signal is entered into the bi-head attention mechanism shown in Fig. 3, which performs scaled dot-product attention on queries, keys, and values on each head. The input of scaled dot-product attention includes the d_k dimension of queries and keys and the d_v dimension of values. In brief, the dot products of the query with all keys are calculated initially, then divided by $\sqrt{d_k}$ for normalization, and a Softmax function is employed to obtain the weights of values. Practically, the attention function on a series of queries will be computed simultaneously; thus, they can be stacked into a matrix Q. Likewise, keys and values are stacked into K and V, respectively. Thus, the output matrix is expressed as

$$\mathrm{A} = \mathrm{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{9}$$

This calculation process is modified by dot-product attention by adding a scaling factor to enhance the gradient stability to improve training. Although this attention calculation method has the same theoretical complexity as additive attention, in practice, dot-product attention is faster and is more space-efficient, as it can be implemented using a highly optimized matrix multiplication code. Softmax converts scores into probabilities. Finally, each value vector is

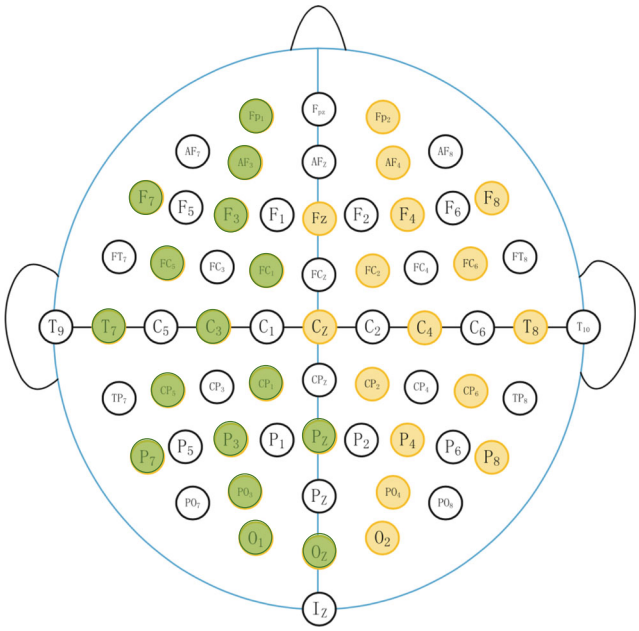

Fig. 3 The location of 32 electrodes in the DEAP dataset. The electrodes are placed according to the International 10–20 system

multiplied by the sum of the probabilities. Those vectors with greater probability will receive extra attention from the subsequent layers.

In the transformer, multi-head attention can be used to boost the performance of the self-attention layer. It permits the model to jointly attend to information from different representation subspaces at different positions. In contrast, the single-head self-attention layer limits the ability to focus on one or more specific positions. Specifically, the essence of multi-head attention is to map the query, key, and value to different subspaces of the original high-dimensional space to calculate the attention, and then merge the attention information in different subspaces in the last step with the total number of remaining parameters [30]. From the training perspective, it reduces the dimension of each vector when calculating the attention of each head and prevents overfitting to some extent. Because the distribution of attention in subspaces varies, multi-head attention actually finds the association between sequences from different aspects, and ultimately synthesizes the captured association in different subspaces.

Previous studies [38, 39] indicated that the left and right hemispheres of the human brain respond differently to various emotions, which can be incorporated into the emotion recognition model to boost the performance of the model. This inspired the bi-head attention mechanism in the Bi-AAN model. We set the number of heads of the multi-head attention mechanism to 2, such that each head represents a hemisphere for calculating the attention difference between the frequency bands of the bilateral hemispheres. Figure 3 shows the locations of 32 electrodes in the DEAP dataset, where the green and yellow electrodes represent the left and right brain, respectively. The bi-head attention divides the data into the original $H_{DE} \in \mathbb{R}^{bands \times P}$ into two datasets: $H_L \in \mathbb{R}^{bands \times \frac{P}{2}}$ and $H_R \in \mathbb{R}^{bands \times \frac{P}{2}}$ where L is the left brain and R is the right brain, according to the data distribution of the channel. As illustrated in Fig. 3, the 32-channel EEG data are divided into the first 16 channels in green and the last 16 channels in yellow. The dataset $H_L \in \mathbb{R}^{bands \times \frac{P}{2}}$ is placed in the subspace of the first head, and the subspace of the second head stores the $H_R \in \mathbb{R}^{bands \times \frac{P}{2}}$ dataset. Specifically, the first head of the bi-head attention only contains the first 16 electrode signals, whereas the data of the last 16 electrodes are placed in the second head, as shown in Fig. 4. Under this premise, two types of frequency band attention can be obtained: internal attention in each hemisphere and attention between the two hemispheres. Subsequently, the extracted attention is concatenated and expressed by the following mathematical formula, as the features are input into the next module.

$$\begin{aligned} &BiHead(Q,K,V) = Concat(head_1, head_2)W^o \\ &where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{10}$$

After the bi-head attention layer, a position-wise feed-forward network (FFN) is applied, which is composed of two linear transformation layers and a nonlinear activation function between them. It can be denoted as:

$$FFN(X) = ReLU\left(H^{'}W_1 + b_1\right)W_2 + b_2 \tag{11}$$

where X is the output of the previous layer. On the other hand, the position-wise FFN can be regarded as two convolutions with kernel size 1, because these parameters are shared across varying positions. The addition of FFN introduces nonlinearity (ReLU activation function), which transforms the space of the attention output, thereby increasing the performance ability of the model.

It is noteworthy that a residual connection followed by a normalization layer is inserted apart from the bi-head attention model and the fully connected feed-forward module. This operation strengthens the flow of information and standardizes the optimization space, thereby accelerating convergence for greater performance. After this, the output is expressed as

$$LayerNorm(H + Attention(H)) \tag{12}$$

where H represents the input of the bi-head attention layer. Hence, the transformer encoder maps low-level descriptors to a sequence of high-level representations. Finally, two dense layers map the emotion-related attention features into the class label space to predict whether the valence and arousal are high or low. In addition, each dense layer was followed by a batch normalization layer and a dropout layer to avoid overfitting.

3.3 Algorithm description of bi-AAN

The loss function is defined in formula (13):

$$Loss = crossentropy(p, l) + \alpha\|\Theta\| \tag{13}$$

where p and l denote the predicted lable of training data and the original lable, respectively. Θ represents all of parameters in Bi-AAN model, and α is the trade-off regularization weight. The cross-entropy function *crossentropy*(p, l) aims at measuring the difference between the original lables and the desired ones, while the regular term $\alpha \parallel \Theta \parallel$ is to avoid the over-fitting of the model parameters learning.

The following formula denotes the update rule of Bi-AAN:

$$\theta = \theta + \lambda \frac{\partial Loss}{\partial \theta} \tag{14}$$

Where θ and λ denote learnable parameters of Bi-AAN and the learning rate. Algorithm 1 summarizes the detailed procedures of training the Bi-AAN model in EEG emotion recognition.

Algorithm 1 The description of Bi-ANN.

Input: Sample collection $X \epsilon R^{n \times bands \times P}$, Data label set Y, learning rate λ, batch size k

Output: The desired parameters of Bi-AAN

1: Initialize model parameters;

2: **repeat**

3: Calculating the positional-encoding PE according to formula (7) and formula (8);

4: Calculating the bi-head attention matrix A according to formula (9);

5: Calculating the normalized residual connection according to formula (12);

6: Calculating the result of the fully connection layer FFN according to formula (11);

7: Calculating the loss function $Loss$ according to formula (13);

8: Updating the parameters of the model

$$\theta = \theta + \lambda \frac{\partial Loss}{\partial \theta};$$

10: **until** the iterations satisfy the predefined algorithm convergence condition.

4 Experiments

To validate the performance of the proposed Bi-ANN, we conducted experiments on two widely used benchmark databases: DEAP database [1] and DREAMER database [40] descriped in Section 4.1. To accurately evaluate the generalization of the Bi-AAN, we conducted both subject-dependent and subject-independent experiments. For the subject-dependent experiment, the model was trained and tested with the same subject's data, whereas the testing data and training data were obtained from different subjects in the subject-independent experiment. Sections 4.3 and 4.4 elaborate on the results on DEAP database and DREAMER database, respectively.

4.1 Data preparation

The DEAP database [1] is a large open-source dataset containing multimodal physiological signals for the analysis of emotion states. During the data collection experiments, bioelectric signals, such as evoked EEG, ECG, and EMG, were detected and recorded. Detailed descriptions are listed in Table 1. Each subject was asked to conduct a self-assessment after watching the video, to assess their emotional state on five rating scales. In this study, the valence–arousal level is used to carry out binary classification. Consequently, the labels of the EEG data were defined as low/high valence (LV/HV) and low/high arousal (LA/HA). In these two dimensions, 5 was the threshold; thus, 5 or more is considered high, and less than 5 is considered low.

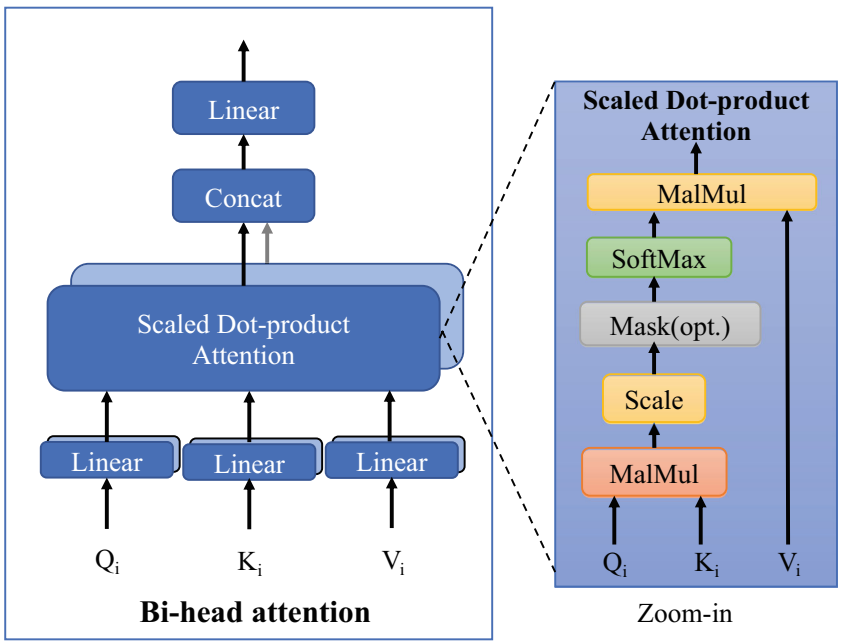

Fig. 4 Bi-head attention mechanism, where i∈ (L, R). L denotes the left brain, and R denotes the right brain. When i = L, the Q, K, and V will be computed as the scaled dot-product attention on the first head. If not, the calculation will be processed on the other

Table 1 The description of DEAP database

Attribute	Description
Participants	32 (16 males and 16 females)
Number of videos	40
Duration of video	60 s
Number of trials	40
Trial length	63 s (3 s baseline and 60 s trial)
Rating scales	Arousal, valence, likes, dominance and familiarity
Recorded EEG signals	32 electrode positions
EEG data	40×32×8064 (trial × channel × data)
EEG label	40×4 (trial × label)

In terms of preprocessing, the EEG signal of each channel was down-sampled to 128 Hz, and EOG artifacts were removed. Since the EEG signal from DEAP was filtered at 4.0–45.0 Hz, the δ band was already removed. Then, as mentioned in Section 3.1, we preprocessed the EEG data. Thus, the raw EEG signal of each trial was cut into 118 segments of 1-s length. Consequently, 4720 sliding time-window EEG samples were obtained for each subject.

The DREMAER database is a multimodal database of EEG and ECG signals recorded during emotion elicitation by audiovisual stimuli. It consists of 23 subjects' (14 males and 9 females) EEG data via 14 electrodes. To build this database, 18 movie clips ranging from 65 to 393 seconds were used to elicit 9 different emotions, and participants were asked to fill in the self-assessment manikins (SAM) after each stimulus. The valence-arousal level was also used for binary classification, with a threshold of 3.

EEG signals were recorded at a sampling rate of 128 Hz using an emotional EPOC system. And most eye artifacts (blinks, eye movements, cardiac disturbances, etc.) have been removed by FIR filters. Then, as with the DEAP dataset, we preprocessed the EEG data and obtained 3710 sliding time-window EEG samples per subject.

4.2 Model training

The proposed algorithm was implemented on the Pytorch 1.8 platform. The hardware experimental apparatus was a Windows Server 2008 R2 Standard operating system, 64G RAM, Intel(R) Xeon(R) CPU E5–2630, NVIDIA Titan V×4 GPU. The specific training details of the Bi-AAN model are as follows. To reduce the contingency caused by the single division of the training set and the test set, 10-fold cross validation was used to eschew the contingency and improve the generalization ability. The average performance of the 10-fold validation was regarded as the final result. The parameters of Bi-AAN are shown in Table 2.

Table 2 Parameters setting of Bi-AAN model

Bi-AAN model parameters	Values
The number of transformer encoder layers N	1
The number of heads h	2
The number of dense layers D	2
Dropout	0.5
Learning rate	0.0001

4.3 Experiment results obtained with the DEAP dataset

4.3.1 Subject-dependent experiments

In the subject-dependent evaluation, samples from the same subject were divided into separate training and test sets. For 10-fold cross validation, the test set had 472 samples, while the rest formed the training set for the same subject.

Based on the above parameters, the results of the proposed Bi-AAN model for each subject's emotion classification are presented in Table 3. The experimental results illustrate that the minimum, maximum, and average classification accuracies of Bi-AAN for 32 subjects were 99.08%, 93.04%, and 96.96%, respectively, for the valence binary classification. The minimum, maximum, and average classification accuracies for the arousal binary classification were 98.41%, 92.45%, and 96.64%, respectively. In addition, the average F1-score reached more than 96% in both the valence and arousal classifications.

Next, to verify the effectiveness of the proposed method, we compared the results of Bi-AAN with five comparator methods, including DT (decision tree), RF (random forest), SVM (support vector machine), ECLGCNN [22] and BiDCNN [39]. ECLGCNN constructs a feature cube input fusion graph CNN (GCNN) and LSTM. BiDCNN constructs three different EEG feature matrices that are input into a three-layer CNN to extract spatial and temporal features. To ensure relatively high persuasiveness and fairness, the experimental data of the ECLGCNN and BiDCNN directly used the results from the literature. Figures 5 and 6 display the average accuracies of the six methods for each subject in the valence and arousal dimensions, respectively. On the one hand, it is apparent from Fig. 5 that Bi-AAN has the highest classification accuracy and the smallest fluctuation range, whereas the accuracies of the DT and RF are below 85%, significantly lower than others and unstable. Moreover, we observed that the fluctuation trends of the DT and RF were similar. The reason for this phenomenon is that RF is an extended version of the DT. On the other hand, the performance of the two comparable deep learning algorithms on subject ID 11 and ID 17 was significantly lower than their average levels, while the oscillations of DT, RF, and SVM appeared in subjects ID 11 and

Table 3 Average accuracies and F1-score on each subject in DEAP dataset

Subject	Binary classification of valence		Binary classification of arousal		Subject	Binary classification of valence		Binary classification of arousal	
	Accuracy (%)	F1_score (%)	Accuracy (%)	F1_score (%)		Accuracy (%)	F1_score (%)	Accuracy (%)	F1_score (%)
1	97.81	98.01	97.29	96.05	17	96.04	96.36	95.85	96
2	95.33	95.1	94.37	94.83	18	97.44	97.96	97.6	96.85
3	97.47	95.37	96.61	97.62	19	97.5	97.29	96.3	96.76
4	96.56	95.54	96.35	95.26	20	97.86	96.54	97.6	97.12
5	93.59	92.45	94.29	94.07	21	97.63	97.38	95.83	94.93
6	97.5	97.87	98.38	97.95	22	93.04	93.29	92.73	91.63
7	97.52	96.33	98.02	96.54	23	98.22	97.7	98.02	97.99
8	96.69	97.73	96.51	97.02	24	96.79	94.09	94.66	94.6
9	97.16	97.01	97.65	98.41	25	98.41	97.8	97.23	96.18
10	96.84	95.98	96.97	97.09	26	95.8	97.09	96.04	95.18
11	96.79	97.11	96.45	97.36	27	97.73	97.5	98.09	97.41
12	98.67	97.24	96.22	96.03	28	97.63	97.55	96.95	96.72
13	99.08	97.52	98.43	98.84	29	96.87	95.98	97.23	96.57
14	96.14	94.55	95.57	96.01	30	95.83	95.87	96.11	94.71
15	97.03	97.32	97.78	97.6	31	97.94	98.4	97.1	97.86
16	97.7	98.41	97.86	97.94	32	96.22	94.24	96.35	96.61
Accuracy mean on valence			96.96		Fl-score mean on valence			96.52	
Accuracy mean on arousal			96.64		Fl- score mean on arousal			96.43	

ID 22. The proposed Bi-AAN had no similar large-amplitude oscillations, and the accuracy in almost all subjects was maintained above 95%. Therefore, these experimental results indicated that Bi-AAN can provide a more stable and robust

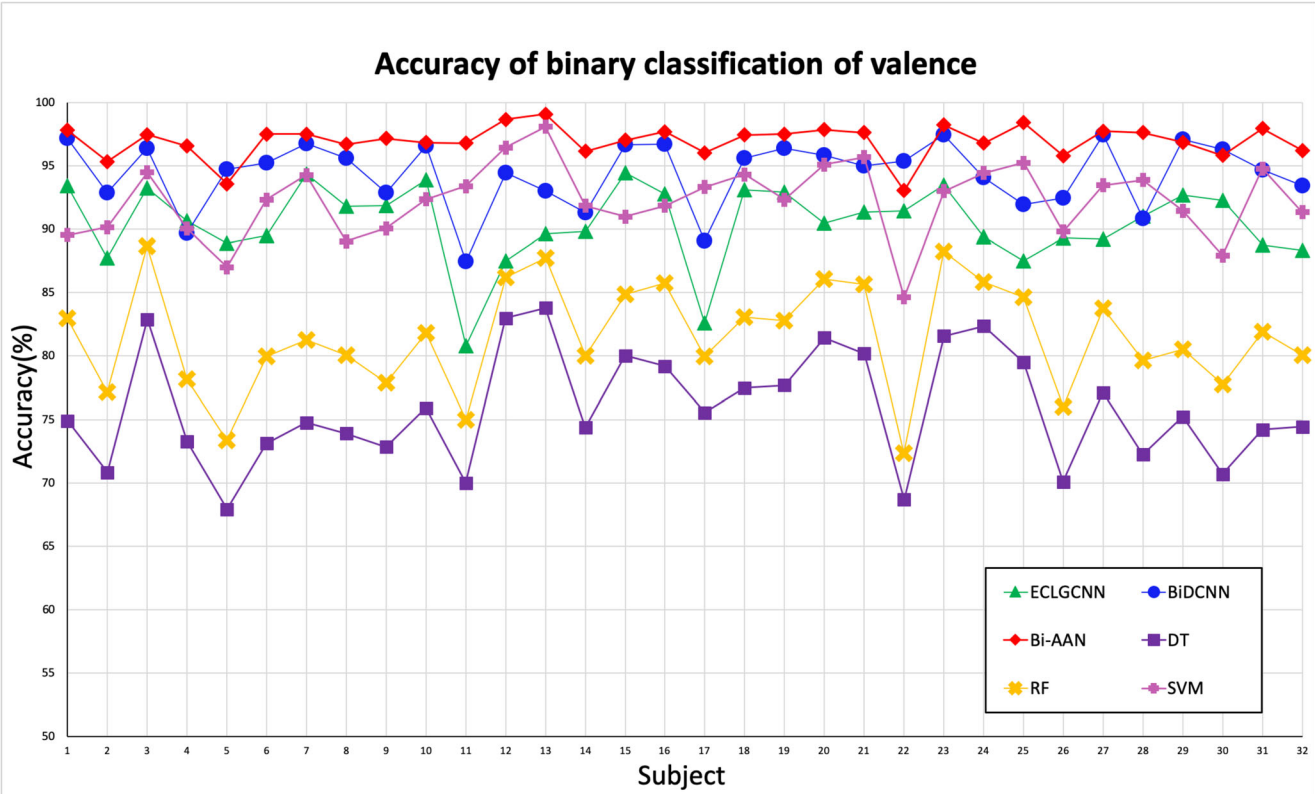

Fig. 5 Comparison of accuracies in each subject for Bi-AAN, SVM, DT, RF, ECLGCNN, and BiDCNN in binary classification of valence

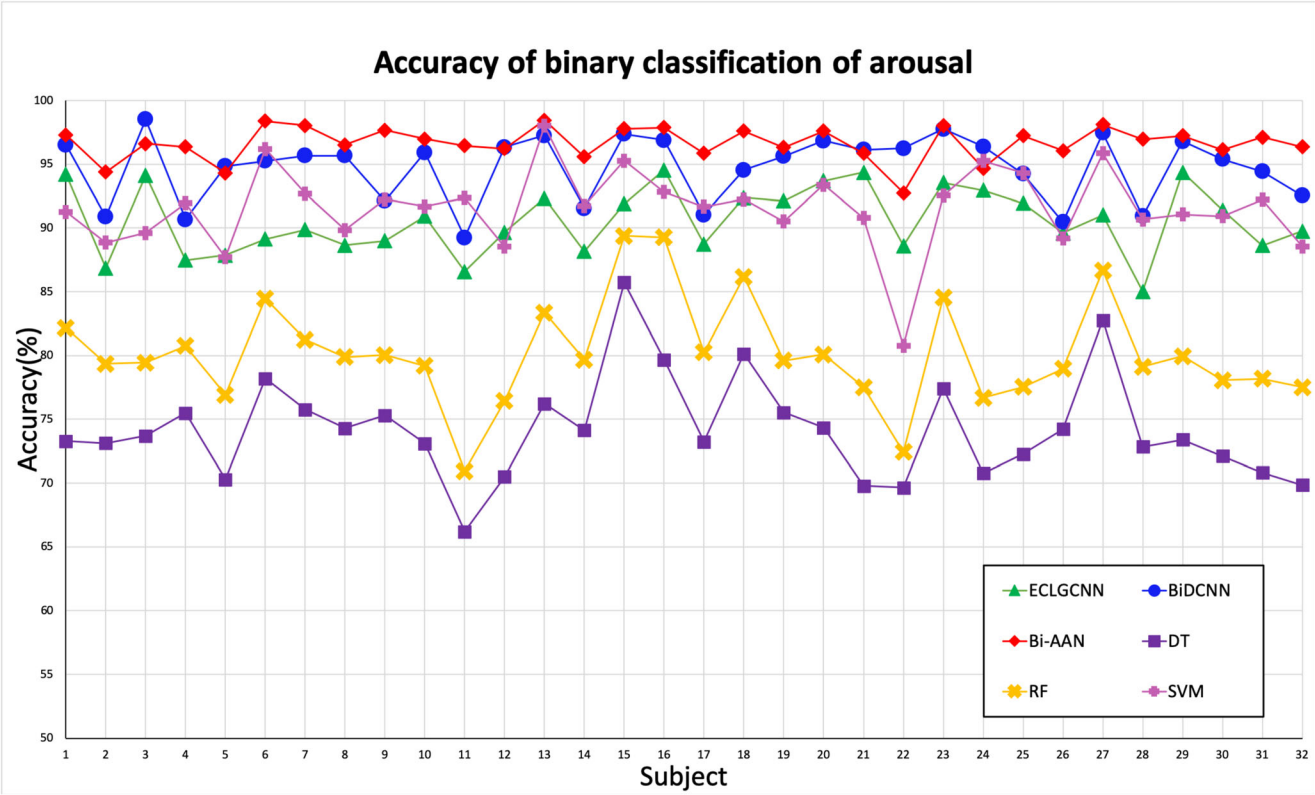

Fig. 6 Comparison of accuracies in each subject for Bi-AAN, SVM, DT, RF, ECLGCNN, and BiDCNN in binary classification of arousal

performance in the emotion classification task in each subject. Likewise, it can be seen from Fig. 6, Bi-AAN also achieved the highest accuracy rate with stable fluctuations in the arousal binary classification. As can be seen, the accuracy of BiDCNN on several specific subjects coincided with that of Bi-AAN, but BiDCNN demonstrated a larger fluctuation range, oscillating around 95%. In summary, Bi-AAN demonstrated the best overall performance, which also proved that Bi-AAN is effective in binary emotion classification, regardless of the dimensions.

For a comprehensive comparison, we also compared the average accuracy, standard deviation, and F1-score of the proposed Bi-AAN with nine other classifiers. In addition to the aforementioned DT, RF, SVM, ECLGCNN, and BiDCNN, four recent deep learning classifiers were also selected: 3DCNN [60], CNN + LSTM [20], CDCN [42], and ACRNN [18]. 3DCNN formulates 3D data representation from multichannel EEG signals as the data input. CNN + LSTM combines the CNN and RNN to form a hybrid neural network to learn the spatiotemporal representation of EEG signals effectively. The CDCN uses a 1D-convolutional layer to receive the contextual features from the time dimension with a 1D-dense structure to capture the electrode correlation. ACRNN applies the attention mechanism to CNN–RNN, which can extract more discriminative spatio-temporal features. As mentioned above, we directly used their results from the literature to ensure the fairness of comparison.

As shown in Table 4, the recognition accuracies of the deep learning methods in the binary emotion recognition of valence and arousal in the DEAP dataset were all lower than 95%, while most of the machine learning methods only reached approximately 80% accuracy, or even lower. For the deep learning classification classifiers, Bi-AAN produced at least 2.85% and 1.91% higher accuracy than did other classifiers in the binary classification of valence and arousal, respectively. Bi-AAN yielded a significantly reduced standard deviation (to 1.3%) as compared to

Table 4 The average recognition accuracies, standard deviation and F1-score of Bi-AAN and other methods on valence and arousal in the DEAP dataset

Methods	Valence	F1score	Arousal	F1score
DT	75.91/4.48	72.29	74.18/4.01	72.98
RF	81.54/4.26	76.10	80.18/4.08	78.06
SVM	92.36/2.89	91.32	91.57/3.15	91.17
3DCNN [60]	87.44/–	86.00	88.49/–	86.00
ECLGCNN [22]	90.45/3.09	91.08	90.60/2.63	90.94
CNN+LSTM [20]	90.80/3.08	–	91.03/2.99	–
CDCN [42]	92.24/–	–	92.92/–	–
ACRNN [18]	93.72/3.21	–	93.38/3.73	–
BiDCNN [39]	94.38/2.61	–	94.72/2.57	–
Bi-AAN	96.96/1.29	96.52	96.63/1.30	96.43

the other approaches, which implies that the Bi-AAN had greater stability. On the other hand, machine learning algorithms demonstrated greater instability (approximately 3–4%). At the same time, the highest F1-score further confirmed the better recognition performance of Bi-AAN. Overall, our method was superior to machine learning algorithms and state-of-the-art deep learning classifiers in subject-dependent EEG emotion recognition, proving the effectiveness and superiority of Bi-AAN.

4.3.2 Subject-independent experiments

In this section, we conducted subject-independent experiments to further evaluate the performance of Bi-AAN in scenarios where the training and test sets contain data from multiple different subjects. Consequently, data from 32 subjects were synthesized into one dataset containing 151,040 (32 × 4720) samples. A 10-fold cross validation with a random strategy was adopted to verify whether Bi-AAN can effectively reduce the discrepancy between subjects. To verify the superiority of Bi-AAN in this case, we also compared it with other methods, such as Emotionet [61], CNN–RNN [62], MEMD [63], DeepCNN [64], SAE + LSTM [65], PNN [66], and ECLGCNN [22]. Similar to the first experiment, we obtained their results directly from the literature. The comparison is presented in Table 5. The average accuracies of the proposed Bi-AAN model for the two emotion recognition tasks were 89.45% and 88.37%, respectively. Nevertheless, the performance of the seven comparison methods varied between 72% and 85%. The results showed that the performance of the Bi-AAN model on the DEAP dataset was at least 3.1% higher than the others. Among these eight methods, Bi-AAN presented excellent performance for subject-independent emotion recognition, and was considerably better than the other methods.

4.4 Experiment results obtained with the DREAMER dataset

In this section, we conduct experiments on the DREAMER database to evaluate the generalization of Bi-AAN on a variety of different databases. Through the division of 10-fold cross-validation, the test set of the subject-dependent has 371 samples, the rest of them was the train set.

The results of the proposed Bi-AAN model for each subject's emotion classification are shown in Table 6. For the valence, the results illustrate that the minimum, maximum, and average classification accuracies of Bi-AAN for 23 subjects were 97.00%, 86.44%, and 92.68%, respectively. The minimum, maximum, and average classification accuracies for the arousal binary classification were 98.25%, 86.72%, and 92.95%, respectively. In addition, the mean F1 score was over 90% in both valence and arousal categories.

For a comprehensive comparison, we compare the results of Bi-AAN with five methods, including DT (Decision Tree), RF (Random Forest), SVM (Support Vector Machine), Conti-CNN [41], and GCNN [56], which is shown in Table 7. The recognition accuracy of the machine learning method in both valence and arousal binary emotion recognition in the DEAP dataset is lower than 90%. Bi-AAN outperforms the other deep learning classifiers by at least 3.26% and 3.37% in binary classification of valence and arousal, respectively. And compared with other methods, Bi-AAN yielded a significantly lower standard deviation (around 3%), implying higher stability. The highest F1 score further confirms the better recognition performance of Bi-AAN, demonstrating the effectiveness and superiority of Bi-AAN.

Table 5 The comparison of Bi-AAN and other algorithm in subject-independent experiments

Methods	Valence (%)	Arousal (%)
Emotionet [61]	72.10	73.30
CNN-RNN [62]	72.06	74.12
MEMD [63]	72.87	75.00
DeepCNN [64]	82.41	73.3o
SAE+LSTM [65]	81.10	74.38
PNN [66]	81.21	81.76
ECLGCNN [22]	84.81	85.27
Bi-AAN	89.45	88.37

5 Discussion

The Bi-AAN proposed in this paper is a novel deep learning model that combines a transformer and bilateral hemispheric asymmetry properties, relying entirely on the attention mechanism for emotion recognition. To validate the performance and generalization of Bi-AAN, we conducted both subject-dependent and subject-independent experiments using the DEAP dataset. Experimental results showed that Bi-AAN had better emotion recognition performance than classical machine learning algorithms, such as DT, RF, SVM, and state-of-the-art deep learning methods, such as ECLGCNN and BiDCNN. In the binary classification of valence and arousal, the average accuracies for valence and arousal reached 96.96% and 96.63%, respectively, in the subject-dependence experiments. In the subject-independent experiments, the accuracies reached 89.45% and 88.37% for valence and arousal, respectively. In both cases, the Bi-AAN achieved superior performance. We found that machine learning algorithms performed unsatisfactorily, indicating that traditional machine learning cannot automatically extract feature representations, and that feature information is deficient [24]. In deep learning methods, such as BiDCNN, the difference between the two hemispheres was introduced, focusing on the preprocessing involving construction of different EEG feature matrices to

Table 6 Average accuracies and F1-score on each subject in DREAMER dataset

Subject	Binary classification of valence		Binary classification of arousal		Subject	Binary classification of valence		Binary classification of arousal	
	Accuracy (%)	F1_score (%)	Accuracy	F1_score (%)		Accuracy (%)	F1_score (%)	Accuracy (%)	F1_score (%)
1	91.66	89.40	93.91	93.37	13	95.34	96.80	95.75	95.01
2	92.09	88.08	93.47	91.59	14	94.56	93.20	93.09	90.40
3	90.63	90.74	90.34	87.18	15	96.66	93.57	97.31	94.01
4	97.00	96.85	98.25	100.00	16	95.19	97.38	94.94	94.13
5	92.13	91.59	92.25	91.02	17	93.72	92.35	90.78	93.05
6	89.41	87.34	86.72	87.74	18	94.75	95.19	95.84	91.68
7	95.91	94.22	94.06	92.12	19	96.78	94.30	94.44	94.23
8	86.44	75.57	90.34	92.22	20	90.28	91.55	93.72	90.21
9	91.09	88.15	87.13	85.29	21	89.84	91.25	92.84	96.08
10	89.16	85.47	89.25	80.03	22	94.41	92.01	95.56	95.45
11	86.81	81.70	91.09	87.66	23	95.10	90.80	95.94	95.46
12	92.75	93.41	90.75	89.16					
Accuracy mean on valence			92.68		F1-score mean on valence			90.91	
Accuracy mean on arousal			92.95		F1-score mean on arousal			91.61	

represent asymmetric features. In contrast, Bi-AAN does not require a complex matrix design before the classifier. The entire process of extracting asymmetric features is achieved by the network, which mainly uses the core bi-head attention mechanism to capture and fuse the attention differences, thereby also providing interpretable possibilities for the deep extraction of frequency features. From the perspective of the transformer, bi-head attention allows the model to attend to information from different hemispheres at different positions jointly, to enhance the performance of the self-attention layer. Its essence is to map the query, key, and value of both hemispheres to different subspaces of the original high-dimensional space to calculate the internal attention and attentional differences. Then, the attention information of the two hemispheres is merged. In this way, Bi-AAN combines neuroscientific knowledge with the transformer to co-train the emotion classification task that can capture and amplify the different emotional responses between the left and right hemispheres by extracting more powerful and discriminative attention features. This implies that emotional representations that include both intra-hemispheric and differential relationships play an important role in improving model performance. Therefore, the incorporation of neuroscience principles can guide construction of better emotion recognition models while avoiding blind searches in deep learning, and can help to extract more discriminative features [39].

Table 7 The average recognition accuracies, standard deviation and F1-score of Bi-AAN and other methods on valence and arousal in the DREAMER dataset

Methods	Valence	Fl score	Arousal	Fl score
DT	84.77/5.44	81.27	84.48/5.33	80.34
RJl	89.42/4.94	84.61	89.58/4.99	85.97
SVM	84.63/4.41	83.45	84.00/4.39	83.11
Conti-CNN [41]	81.72/5.24	–	82.48/5.11	–
GCNN [56]	88.87/3.58	–	88.79/3.86	–
Bi-AAN	92.68/3.03	90.91	92.95/2.96	91.61

Second, the architecture of the transformer makes few assumptions about the structural information of the data, making it a general and flexible architecture that allows extension of its advantages to sequence classification tasks other than NLP. Furthermore, a transformer can be viewed as a graph neural network [30], defined on a complete directed graph (with self-loops), where each input is a node in the graph. The standard attention mechanism can be viewed as a complete bipartite graph, where each query receives information from all memory nodes and updates its representation. This hypothesis provides a convincing basis for the outstanding performance of the transformer in EEG emotion recognition. Since the effectiveness of graph neural networks in extracting discriminative emotional features has been demonstrated early [22, 56, 67], it provides an efficient way to describe the intrinsic relationship between multiple EEG channels. These traditional graph neural networks usually need to be combined with CNN or RNN to extract comprehensive spatiotemporal features; thus, they are relatively complex, with heavy parameter calculation, and cannot mine long-range relationships of sequences. In contrast to the above methods, the proposed model can extract multi-domain features, including time, spatial, and frequency domain features, by using only the transformer structure. It effectively explores long-distance sequence information that is

difficult to capture by a CNN [26], which provides a more compact classification model for EEG emotion recognition.

In addition, the interpretability of the attention mechanism is a non-negligible advantage. Compared with other networks that blindly extract features with CNN and RNN as the main body, the attention extracted by Bi-AAN has great potential for visualization. This provides a potential method for determining the frequency bands or hemispheres that contribute most to EEG emotion recognition, which is beneficial for further combining emotion recognition tasks with neuroscientific conclusions, to improve the performance of real-time EEG emotion recognition.

6 Future works

Firstly, the positional encoding in Bi-AAN has certain limitations: the original 1D linear positional encoding is used in the model, while the electrode positions of EEG are in a 2D plane. Unlike words in a sentence that only have contextual relationships, EEG electrodes are not only multi-directional, but also have more complex distance relationships. Changing the position-encoding to a 2D form suitable for the distribution of EEG electrodes should be considered in future studies.

Secondly, integration of multimodal data is necessary to improve performance [30]. Because transformers have achieved great success in image, video, speech, and EEG, we have an opportunity to build a unified framework to converge explicit emotional signals and EEG signals, capture the intrinsic connection between multimodal data, and further explore the opportunities and challenges of emotion recognition.

Another topic that can be explored in future is the generalization of the data scale. A previous study has shown [30] that a transformer has a larger capacity than a CNN or RNN, and is more capable of handling large amounts of training data. This is because the transformer has few prior assumptions about the data structure, and is therefore more flexible. Additionally, the parallel processing of the attention mechanism makes the model training faster, particularly in the case of a large amount of data, such as subject-independent tasks. Generally speaking, due to the multiplication of the data in the subject-independent experiments, the parameters of subject-dependence need to be adjusted empirically to achieve optimal performance during subject-independent experiments [22]. In contrast, Bi-AAN has better generalizability on different data scales that can share the same parameters in both subject-independent and subject-dependent experiments, which is particularly important for real-time emotion recognition applications. However, the emotion recognition dataset used in this experiment was relatively small, which largely underutilized the capabilities of the deep learning network. Hence, one of the main tasks in future research is to apply larger-scale EEG datasets to this model to confirm the generalization ability of the model, allowing it to exert an advantage in large-scale EEG datasets, such as brain–computer interface tasks.

7 Conclusions

In this paper, we proposed a novel Bi-AAN model based on the transformer structure and asymmetric nature of the human brain to address EEG emotion recognition. Extensive experiments on the DEAP and DREAMER datasets demonstrated that the proposed Bi-AAN achieves superior performance as compared to other state-of-the-art classifiers. It benefits from the fact that Bi-AAN applies the attention mechanism to model attentional differences between the two hemispheres to extract more discriminative emotional representations. Therefore, this study highlights the applicability and superiority of the transformer model for EEG sequence data, extending the effectiveness of language model structures of sequences to EEG emotion recognition. In summary, the Bi-AAN model, an emotion recognition model composed entirely of attention mechanisms, provides a valuable basis for realizing a new generation of simpler and smarter human–computer interaction technology. Finally, we plan to explore the robustness and accuracy of optimization algorithms to accelerate the development and progress in the field of multimodal emotion recognition.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (No. 61872301 and No. 61472330).

References

1. Koelstra S (2012) DEAP: A Database for Emotion Analysis ;Using Physiological Signals. IEEE Trans Affect Comput *3*(1):18–31
2. Huang Y, Wu C, Wong AM, Lin B (2014) Novel active comb-shaped dry electrode for EEG measurement in hairy site. Biomed Engin IEEE Transact 62(1):256–263
3. Xiaohua H et al (2017) Discriminative Spatiotemporal Local Binary Pattern with Revisited Integral Projection for Spontaneous Facial Micro-Expression Recognition. *Affective Comput IEEE Transac 10*(1):32–47
4. C. Li, W. Tao, J. Cheng, Y. Liu, and X. Chen, "Robust Multichannel EEG Compressed Sensing in the Presence of Mixed Noise," *IEEE Sensors Journal,* vol. PP, no. 99, 2019
5. Alarcao SM, Fonseca MJ (2019) Emotions recognition using EEG signals: a survey. IEEE Trans Affect Comput 10(3):374–393. https://doi.org/10.1109/taffc.2017.2714671
6. Alhagry S, Aly A, Reda A (2017) Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *Int J Adv Comput Sci Appl* 8(10)
7. Subasi A, Tuncer T, Dogan S, Tanko D, Sakoglu U (2021) EEG-based emotion recognition using tunable Q wavelet transform and rotation forest ensemble classifier. Biomed Sig Proces Contrl 68: 102648. https://doi.org/10.1016/j.bspc.2021.102648
8. Tuncer T, Dogan S, Subasi A (2021) A new fractal pattern feature generation function based emotion recognition method using EEG.

Chaos, Solitons Fractals 144:110671. https://doi.org/10.1016/j.chaos.2021.110671
9. Tuncer T, Dogan S, Baygin M, Rajendra Acharya U (2022) Tetromino pattern based accurate EEG emotion classification model. Artif Intell Med 123:102210. https://doi.org/10.1016/j.artmed.2021.102210
10. Dogan A et al (2021) PrimePatNet87: prime pattern and tunable q-factor wavelet transform techniques for automated accurate EEG emotion recognition. Comput Biol Med 138:104867. https://doi.org/10.1016/j.compbiomed.2021.104867
11. M. A. Asghar, M. J. Khan, M. Rizwan, R. M. Mehmood, S-H. Kim, "An Innovative Multi-Model Neural Network Approach for Feature Selection in Emotion Recognition Using Deep Feature Clustering," *Sensors,* vol. 20, no. 13, 2020, Art no 3765, https://doi.org/10.3390/s20133765
12. R. Alhalaseh and S. Alasasfeh, "Machine-Learning-Based Emotion Recognition System Using EEG Signals," *Computers,* vol. 9, no. 4, 2020, Art no. 95, https://doi.org/10.3390/computers9040095
13. Moon S-E, Chen C-J, Hsieh C-J, Wang J-L, Lee J-S (2020) Emotional EEG classification using connectivity features and convolutional neural networks. Neural Netw 132:96–107. https://doi.org/10.1016/j.neunet.2020.08.009
14. Pandey P, Seeja KR (2021) Subject independent emotion recognition system for people with facial deformity: an EEG based approach. J Ambient Intell Humaniz Comput 12(2):2311–2320. https://doi.org/10.1007/s12652-020-02338-8
15. Senguer D, Siuly S (2020) Efficient approach for EEG-based emotion recognition. *Electron Lett* 56(25). https://doi.org/10.1049/el.2020.2685
16. Y. Cimtay and E. Ekmekcioglu, "Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition," *Sensors,* vol. 20, no. 7, 2020, Art no. 2034, https://doi.org/10.3390/s20072034
17. Demir F, Sobahi N, Siuly S, Sengur A (2021) Exploring deep learning features for automatic classification of human emotion using EEG rhythms. IEEE Sensors J:1–1. https://doi.org/10.1109/JSEN.2021.3070373
18. Tao W et al (2020) EEG-based emotion recognition via channel-wise attention and self attention. IEEE Trans Affect Comput:1–1. https://doi.org/10.1109/taffc.2020.3025777
19. Youjun L, Jiajin H, Haiyan Z, Ning Z (2017) Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. Appl Sci 7(10):1060
20. Yang Y, Wu Q, Qiu M, Wang Y, Chen X (2018) "Emotion recognition from Multi-Channel EEG through parallel convolutional recurrent neural network," presented at the 2018 International Joint Conference on Neural Networks (IJCNN)
21. Kim Y, Choi A EEG-Based Emotion Classification Using Long Short-Term Memory Network with Attention Mechanism. *Sensors (Basel, Switzerland)* 20(23):6727
22. Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl Soft Comput,* vol. 100, 2021, Art no 106954, https://doi.org/10.1016/j.asoc.2020.106954
23. Li J, Wu X, Zhang Y, Yang H, Wu X (2022) DRS-net: a spatial–temporal affective computing model based on multichannel EEG data. Biomed Sig Proces Contrl 76:103660. https://doi.org/10.1016/j.bspc.2022.103660
24. Liu S, et. al. (n.d.) "3DCANN: a Spatio-temporal convolution attention neural network for EEG emotion recognition," IEEE Journal of Biomedical and Health Informatics, https://doi.org/10.1109/JBHI.2021.3083525
25. X. Tan, K. L. Gao, B. Liu, Y. M. Fu, and L. Kang, "Deep global-local transformer network combined with extended morphological profiles for hyperspectral image classification," *J Appl Remote Sens,* vol. 15, no. 3, 2021, Art no. 038509, https://doi.org/10.1117/1.Jrs.15.038509
26. Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers Advance Multi-Modal Medical Image Classification," *Diagnostics,* vol. 11, no. 8, 2021, Art no. 1384, https://doi.org/10.3390/diagnostics11081384
27. Zhao B, Gong M, Li X (2022) Hierarchical multimodal transformer to summarize videos. Neurocomputing 468:360–369. https://doi.org/10.1016/j.neucom.2021.10.039
28. A. Vaswani *et al.* 2017 "attention is all you need," *arXiv*
29. J. Schmidhuber (2001) *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. Gradient flow in recurrent nets: The Difficulty of Learning LongTerm Dependencies
30. T. Lin, Y. Wang, X. Liu, and X. Qiu (2021) "A Survey of Transformers"
31. D. Kostas, S. Aroca-Ouellette, and F. Rudzicz (2021) "BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data"
32. Chen H, Jiang D, Sahli H (2020) Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transac Multimed* 99:1–1
33. Craik A, He Y, Contreras-Vidal JL (2019) Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 16(3):031001. https://doi.org/10.1088/1741-2552/ab0ab5
34. Zatorre RJ, Jonesgotman M, Evans AC, Meyer E (1992) Functional localization and lateralization of human olfactory cortex. Nature 360(6402):339–340. https://doi.org/10.1038/360339a0
35. Dimond SJ, Farrington L, Johnson P (1976) Differing emotional response from right and left hemispheres. Nature 261(5562):690–692. https://doi.org/10.1038/261690a0
36. Davidson RJ (1992) Emotion and affective style - hemispheric substrates. Psychol Sci 3(1):39–43. https://doi.org/10.1111/j.1467-9280.1992.tb00254.x
37. R. J. Davidson, C. D. Saron, J. A. Senulis, P. Ekman, and W. V. Friesen, "Approach withdrawal and cerebral asymmetry - emotional expression and brain physiology I," J Pers Soc Psychol, vol. 58, no. 2, pp. 330–341, Feb 1990, https://doi.org/10.1037/0022-3514.58.2.330
38. 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:1–1
39. Huang D, Chen S, Liu C, Zheng L, Jiang D (2021) Differences first in asymmetric brain: a bi-hemisphere discrepancy convolutional neural network for EEG emotion recognition. Neurocomputing 448
40. 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 Inform:1–1
41. Yang Y, Wu Q, Fu Y, Chen X (2018, ch. Chapter) Continuous Convolutional Neural Network with 3D Input for EEG-Based Emotion Recognition. *Neural Inform Proc, (Lecture Notes in Comp Science)* 39:433–443
42. Gao Z, Wang X, Yang Y, Li Y, Ma K, Chen G (2020) A channel-fused dense convolutional network for EEG-based emotion recognition. *IEEE Transac Cogn Develop Syst*:1–1. https://doi.org/10.1109/tcds.2020.2976112
43. Y. Zhu and Q. Zhong, "Differential entropy feature signal extraction based on activation mode and its recognition in convolutional gated recurrent unit network," *Front Phys,* vol. 8, 2021, Art no. 629620, https://doi.org/10.3389/fphy.2020.629620
44. Chao H, Dong L (2021) Emotion recognition using three-dimensional feature and convolutional neural network from multi-channel EEG signals. IEEE Sensors J 21(2):2024–2034. https://doi.org/10.1109/jsen.2020.3020828
45. Kalchbrenner N, Espeholt L, Simonyan K, Oord A, Graves A, Kavukcuoglu K (2016) "Neural Machine Translation in Linear Time"

46. Kaiser U, Sutskever I (2015) "Neural GPUs learn algorithms", *Computer Science*
47. Luong MT, Pham H, Manning CD (2015) "Effective approaches to attention-based neural machine translation," *Computer ence*
48. Devlin J, Chang MW, Lee K, Toutanova K (2018) "BERT: pre-training of deep bidirectional transformers for language understanding"
49. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Houlsby N (2020) "An image is worth 16x16 words: Transformers for Image Recognition at Scale"
50. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-End Object Detection with Transformers
51. Gulati A *et al.* (2020) "Conformer: Convolution-augmented Transformer for Speech Recognition"
52. Chen X, Wu Y, Wang Z, Liu S, Li J (2020) "Developing real-time streaming transformer transducer for speech recognition on large-scale dataset"
53. Rives A, Meier J, Sercu T, Goyal S, Fergus R (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 118(15): e2016239118
54. Schwaller P, Laino T, Gaudin T, Bolgar P, Bekas C, Lee AA (2018) "Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction"
55. Fox E (2008) Emotion Science*:* Cognitive and Neuroscientific Approaches to Understanding Human Emotions. Emotion science: cognitive and neuroscientific approaches to understanding human emotions
56. Song T, Zheng W, Song P, Cui Z (2020) EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Trans Affect Comput 11(3):532–541. https://doi.org/10.1109/taffc.2018.2817622
57. Liu S, Wang X, Zhao L, Zhao J, Xin Q, Wang S (2020) "Subject-independent Emotion Recognition of EEG Signals Based on Dynamic Empirical Convolutional Neural Network", *IEEE/ACM Trans Comput Biol Bioinform,* vol. PP, https://doi.org/10.1109/TCBB.2020.3018137
58. He K, Zhang X, Ren S, Sun J (2016) "Deep residual learning for image recognition," *IEEE*
59. Ba JL, Kiros JR, Hinton GE (2016) "Layer normalization"
60. Shawky E, El-Khoribi R, Shoman MAI, Wahby MA (2018) EEG-based emotion recognition using 3D convolutional neural networks. Int J Adv Comput Sci Appl 9(8):329
61. Wang Y, Huang Z, Mccane B, Neo P (2018) EmotioNet: a 3-D convolutional neural network for EEG-based emotion. Recognition:1–7
62. Xiang L, Song D, Peng Z, Yu G, and Hu B (2017) "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *IEEE International Conference on Bioinformatics & Biomedicine*
63. Mert A, Akan A (2016) Emotion recognition from EEG signals by using multivariate empirical mode decomposition. Pattern Anal Applic 21(1):81–89
64. Tripathi S, Acharya S, Sharma RD, Mittal S, Bhattacharya S (2017) "Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset"
65. Xing X, Li Z, Xu T, Shu L, Xu X (2019) "SAE+LSTM: A New Framework for Emotion Recognition From Multi-Channel EEG", *Frontiers in Neurorobotics,* vol. 13
66. Zhang J, Ming C, Hu S, Yu C, Kozma R (2016) "PNN for EEG-based emotion recognition," in *2016 IEEE International Conference on Systems,* Man, and Cybernetics (SMC)
67. Zhang G, Yu M, Liu YJ, Zhao G, Zhang D, Zheng W (2021) SparseDGCNN: recognizing emotion from multichannel EEG signals. IEEE Trans Affect Comput:1–1. https://doi.org/10.1109/TAFFC.2021.3051332

Xinyue Zhong received the bachelor's degree (with honours) in Electrical Engineering from the University of Tasmania and Southwest University in 2020. She currently enrolled in M.S. degree in Information and Communication Engineering, Southwest University. Her research fields are affective computing, artificial intelligence, deep learning, computational neuroscience and EEG-based emotion recognition.

Yun Gu received his B.S. degree in the Automation from Wuhan Polytechnic University in 2017. He is now pursuing his master's degree in the Graduate School of Electronic Information Engineering in Southwest University engaging in the research of machine learning and deep learning.

Yutong Luo received the M.S. degree in software engineering from Chongqing Normal University in 2019. He is currently a Ph.D. candidate in the School of Computer Science and Technology, Southwest University. His current research interests include affective computing, computer vision and machine learning.

Xiaomei Zeng received the B.E. degree in electronic information science and technology from Yangtze Normal University in 2020. She is currently a M.S. candidate in the School of Electronic Information Engineering, Southwest University. Her current research interests include affective computing, computational neuroscience and classification and recognition of emotion.

Guangyuan Liu received the PhD degree in circuit and system at the University of Electronic Science and Technology of China in 1999. He is currently a professor at Southwest University of China. His main research interests include computational intelligence and affective computing.