EAI Endorsed Transactions
on e-Learning
Research Article EAI.EU

# EEG Emotion Recognition Based on Self-Distillation Convolutional Graph Attention Network

Hao Chao[1], Shuqi Feng[1,*]

School of Computer Science and Technology, Henan Polytechnic University, No. 2001, Century Road, Jiaozuo 454003, China

## Abstract

A convolution graph attention model based on self-distillation convolutional graph attention network (SDC-GAT) is proposed for multi-channel electroencephalograph (EEG) emotion recognition. Firstly, two-dimensional feature matrix based on EEG time-domain features are constructed, and the matrix is fed into the graph attention neural network to learn the internal connections between electrical brain channels located in different brain regions. Meanwhile, the three-dimensional feature matrix is constructed according to the relative positions of the electrode channels, and the self-distillation network is employed to extract local high-level abstract features containing electrode spatial position information from the three-dimensional feature matrix. Finally, outputs of the two networks are integrated to determine the emotional states. Experiments were performed on the DEAP dataset. The experimental results show that the spatial domain information of the electrode channel and the internal connection relationship between different channels are beneficial for emotion recognition. In addition, the proposed model can effectively fuse this information to improve the performance of multi-channel EEG emotion recognition.

**Keywords:** Multi-channel electroencephalogram signal, Emotion classification, Model compression, Self-distillation, Graph attention neural network, Convolutional neural networks, Attention

Received on 30 January 2024, accepted on 29 February 2024, published on 08 March 2024



doi: 10.4108/eetel.4974

*Corresponding author. Email: author@emailaddress.com

## 1. Introduction

Humans produce emotions as a physiological state in the face of external stimuli [1]. Emotion is a crucial factor in human life, which affects people's ability to work, spiritual state and judgment. Therefore, the recognition of emotions is essential for medical diagnosis, human-computer interaction, product design, and other fields [2]. Due to the ability of humans to self-camouflage, it is inaccurate to use non-physiological signals such as facial expressions to predict changes in human emotions. Electroencephalogram (EEG) signals are physiological signals, which can represent different emotional states and have the advantage of being difficult to camouflage. It manifests as waves of different frequencies, amplitudes, and shapes, which can accurately reflect fluctuations in emotional states in real time. As one of the most active research topics in affective computing, EEG emotion recognition has been widely concerned by the computer vision and pattern recognition research community [4].

Deep neural networks have shown good results in the field of EEG emotion recognition. Convolutional Neural Network (CNN) is an important deep learning model [5-8]. It can comprehensively mine and fuse the representation information of samples and is applied to EEG emotion recognition. However, a single network with fewer parameters cannot get more valuable information from training. When a network has a large number of parameters, it can be overfitted, which can affect performance. Knowledge distillation (KD) is a training method that can be used for the compression of ensemble models. The purpose of knowledge distillation is to use a network of teachers of high complexity to instruct a network of students of low complexity during training. However, the traditional knowledge distillation has two problems: teacher mode selection and knowledge transfer efficiency. Self-distillation (SD) [12,13] circumvents these problems. Self-distillation first appends several shallow classifiers based on concerns after the middle layers of the neural network at different

depths. In this paper, the self-distillation method is introduced into EEG emotion recognition, and the convolutional neural network based on self-distillation is used to extract the local spatial information of EEG signals. Self-distillation not only reduces training overhead, but also has higher recognition accuracy.

Due to the complex structure of the human brain, the arrangement of each channel is irregular during the EEG signal acquisition process. In the existing research, the continuous EEG signals are converted into a regular grid structure for signal sampling. But the assumption of this approach is that the electrodes are equidistant and ignores the functional neural connections between different parts of the brain. The EEG feature extraction process is oversimplified, so that the complex neural connectivity between different electrode locations cannot be explored. In order to solve this problem, some studies have used the connection relationship between electrode positions to construct the topological map structure of EEG signals. Then, the graph neural network (GNN) is used for EEG emotion recognition, and the graph neural network can update the state of vertices by exchanging neighborhood information periodically. Song et al. [14] proposed a multi-channel EEG-based Dynamic Graph Convolutional Neural Network (DGCNN) for emotion recognition. Yin et al. [15] proposed a fusion model of Graph Convolutional Network (GCN) and Long Short Term Memory (LSTM), and obtained better sentiment classification results on the DEAP dataset. Some studies [16][17] have used graph theory-based EEG network measurements or single-channel EEG complexity estimation for affective state studies.

Although deep learning has gotten good results in sentiment recognition, there are still some problems. Firstly, the location of EEG signal channels is complex, and it is worth exploring how to use the position relationship between different electrodes in the brain to improve the efficiency of emotion recognition. However, the approach of GCN needs to predetermine the weights between different connected nodes, which limits the flexibility and generalization ability of the network. Therefore, the use of graph attention network (GAT) to parameterize the weights between nodes is more helpful for sentiment recognition. Secondly, a single network with fewer parameters cannot get more valuable information from training. When a network has many parameters, it can be overfitted, which can affect performance. Therefore, self-distillation is introduced into emotion recognition to improve the performance of sentiment recognition while reducing network parameters. In order to solve the above problems, this paper proposes a self-distillation convolutional graph attention network (SDC-GAT). It uses a distillation network to learn the spatial position information of the electrode channels through a 3D feature matrix. Graph attention neural networks are used to obtain neural connections between different brain regions. The multi-head self-attention mechanism is used to adaptively adjust the adjacency matrix in the network, and the intrinsic relationship between EEG signals in different brain regions and different brain regions was fully utilized. Finally, the extracted high-level abstract features are fused and classified into the classification module.

The main contributions of this paper are as follows:

(i) The proposed SDC-GAT model can make full use of the intrinsic relationship between EEG signals in different brain regions and brain regions. Specifically, it can not only capture the global features of the brain through the connected edges on the undirected graph, so as to obtain the discriminant signals of different receptor domains of the EEG signal in the global learner, but also use the convolutional network based on distillation to extract the local features of each channel of the EEG signal.

(ii) In the graph attention neural network, the multi-head self-attention mechanism was used to adaptively adjust the adjacency matrix in the network, and the attention mechanism was used to parameterize the weights between nodes, which can improve the performance of emotion recognition.

(iii) Experiments were carried out on the DEAP dataset, and the experimental results show the superiority and rationality of the SDC-GAT model.

## 2. Related work

### 2.1. Emotion recognition model

Arousal and valence are the two main indicators of affective state, and two-dimensional affective models can be built based on these two dimensions. In this paper, the arousal-valence model proposed by Russell [18] is adopted, with arousal as the abscissa and valence as the ordinate. Arousal changes from inactive (e.g., uninterested, calm) to active (e.g., excited, alert), measuring activation of the sympathetic nervous system. Valence ranges from negative (e.g., nervous, sad) to positive (e.g., happy), measuring subjective attitudes. In the arousal dimension, if the score is less than 5, the definition label is low arousal, and if the score is higher than 5 or equal to 5, the definition is high arousal. Similarly, the labels on the valence dimension can be defined as low potency and high potency, respectively.

### 2.2. Emotional feature extraction

The time-domain features mainly capture the temporal statistics of EEG signals, and the common time-domain features include Hjorth features [19], higher-order crossover features [20], and event-related potentials [21]. In addition, statistical features such as mean, power, median, standard deviation, skewness, relative band energy, kurtosis, etc., are also used for emotion recognition.

### 2.3. Knowledge distillation

As one of the most widely used techniques in deep learning, the methods, applications, and principles of knowledge

distillation have attracted more and more attention [22-24]. The idea of using larger models to guide the training of smaller models was first proposed by Bucilua et al. for the compression of ensemble models [25]. Hinton et al. then extended this idea to neural networks, proposing the concept of "distillation" for the first time [26]. Then, based on the feature diagram [27], attention [28], solution process flow [29] and figure [30], an effective distillation method was proposed to transfer the knowledge of the teacher model to the student model. In addition to the compression and acceleration of neural networks, knowledge distillation has applications in other contexts. BAN [31] improves the accuracy of multiple student models by sequentially training them. Bagherinezhad et al. used knowledge distillation to refine the quality of labels and achieved significant accuracy improvements in classification [32]. Liu et al. [33] applied knowledge distillation to visual tasks such as object detection, segmentation, and depth prediction. Gupta et al. [34] proposed cross-modal knowledge distillation, which guides neural networks to train on unlabeled depth prediction and optical flow images. In addition, knowledge distillation is also used for neural network architecture search [35], semi-supervised learning, and distributed neural network training. Knowledge distillation can greatly reduce the number of parameters, thereby reducing the demand for resources such as CPU, memory, and energy consumption.

## 3. EEG emotion recognition based on SDC-GAT model

### 3.1. SDC-GAT Emotion Recognition Framework

In order to take advantage of the local dependency of EEG signal channels and the global spatial domain information, an SDC-GAT emotion recognition framework was proposed, as shown in Figure 1. The framework includes EEG signal feature processing and construction, neural network for high-level abstract feature extraction, and sentiment classification. Specifically, six time-domain features were extracted from 32 EEG signals, and then two-dimensional feature matrices and three-dimensional feature matrices were constructed according to the extracted time-domain features. A matrix of 2D features is fed into the GAT module to extract high-level abstract features that contain intrinsic connections between EEG channels located in different brain regions. The Convolutional Distillation Module is used to receive a 3D feature matrix and generate high-level abstract features that represent the spatial position of the electrodes. Finally, the classifier module is used to fuse two high-level abstract features and judge the emotional state.

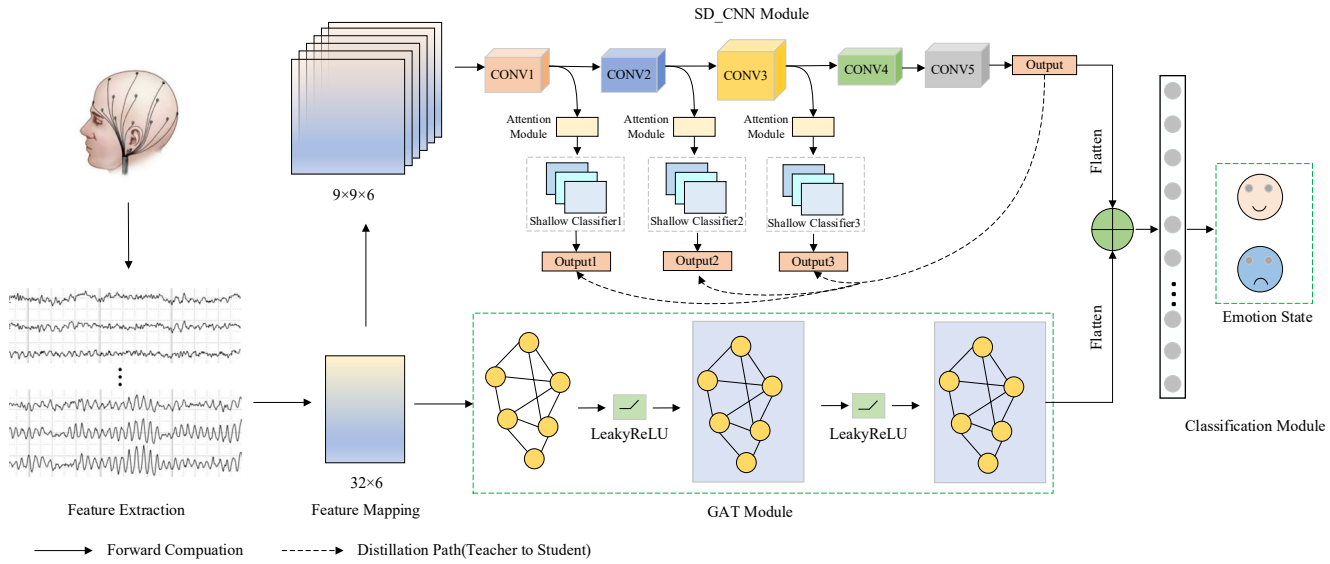


**Figure 1.** SDC-GAT Emotion Recognition Framework

Figure 1 shows the details of the self-distillation proposed on a convolutional neural network model. On the basis of self-distillation without changing the structure of the backbone layer, multiple early exit branches are added after the middle layer of the convolutional neural network. Each early exit branch consists of an attention module and a shallow classifier. In the training phase, all classifiers are trained using the self-distillation method proposed in this paper, which uses the deep classifier as the teacher model and the shallow classifier as the student model. During inference, all additional interest modules and shallow classifiers are discarded, so the deployed model has no additional parameters or computational loss. As shown in Figure 1, the backbone convolutional neural network acts as a deep classifier. According to the structure of the construction, it is divided into five parts, and three student models are constructed by adding an attention module and a shallow classifier in turn to the first three parts.

## 3.2. SDC-GAT implementation principle

Ref. [36] shows the floor plan of the international 10/20 system and its mapping matrix. The performance of emotion recognition can be improved by using the global information and spatial features of EEG channels, and a feature matrix can be constructed according to the position of the electrodes on the brain. The time-domain features extracted from different EEG channels are placed into the corresponding positions in the matrix according to their relative position coordinates. In order to maintain the integrity of the spatial information, 0 is used to represent the unused channels, and the mapped 2D matrix is shown in Figure 2. For each sample, a 9×9×1 matrix is constructed according to the mapping rules of Ref. [36]. The two-dimensional feature maps of the six features are superimposed to obtain a three-dimensional feature matrix of 9×9×6. This matrix not only contains the unique characteristics of each channel of the EEG, but also preserves the interaction and correlation information between the channels.

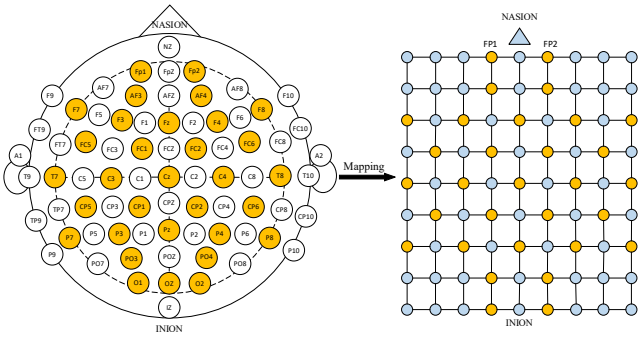


**Figure 2.** Mapped two-dimensional matrices

The first part of the convolutional layer is univariate convolution, which consists of 1×1 convolution kernels to form convolutional 1, with the purpose of extracting features from each EEG signal channel. The second part is a multi-scale convolutional layer, which uses 3×3, 5×5, 7×7 convolution kernels to form convolution 2, and uses three different convolution kernels to convolute the feature map. The convolutions are then merged in the same dimension. Subsequently, the conventional convolution operation is performed by convolution 3 to extract the spatial position information of adjacent electrodes in the EEG signal related to the affective state, and to explore the correlation between different brain regions. In the self-distillation convolutional neural network, the self-distillation technique shown in Figure 1 is used. The SD-CNN model adopts the following ideas to construct a self-distillation framework: firstly, the target convolutional neural network is divided into three shallow layers according to its depth and original structure. Secondly, a classifier is set after each shallow segment, which consists of an attention module and a shallow classifier. These two layers are only used for training and can be removed in inference. In the training phase, all the shallow sections with the corresponding classifiers are trained as student models by distilling the deepest sections, which are conceptually considered teacher models.

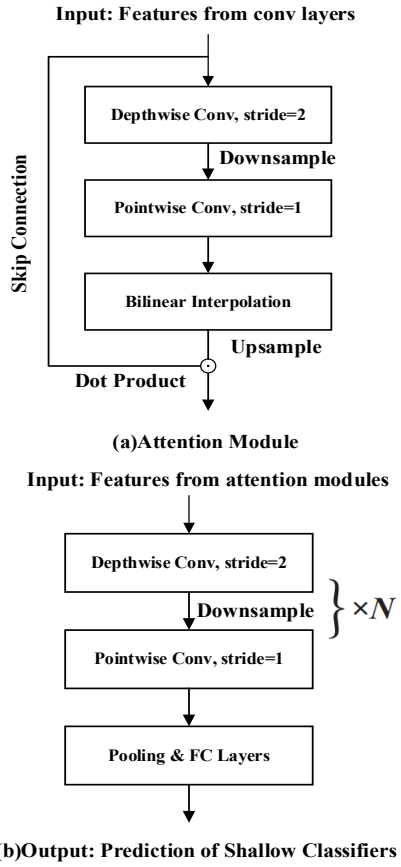


**Figure 3.** Pay attention to the structure of modules and shallow classifiers

Figure 3 shows the structure of the attention module and the shallow classifier. The attention module in Figure 3 consists of a downsampled convolutional layer and an upsampled bilinear interpolation layer. The attention mask learned by these two layers is used to enhance the original features through dot product operation. The shallow classifier consists of several pairs of depth layers and point-oriented layers in order to downsample features with fewer parameters and computations. N in Figure 3 is determined by the depth of the shallow classifier.

When the input 3D sparse matrix passes through the self-distillation convolution network, convolution operation, self-distillation and batch normalization operations are performed first. Then, the nonlinear interaction between the feature channels is learned by activating the layer, and the specific self-distillation process is as follows.

N samples $X = \{x_i\}_{i=1}^{N}$, in a given $M$ class, we denote the corresponding set of tags as $Y = \{y_i\}_{i=1}^{M}, y_i \in \{1,2,3,...,M\}$. The classifiers in a neural network (the proposed self-

distillation has multiple classifiers across the network) are represented as $\{\theta_{i/C}\}_{i=1}^{C}$, where C is the number of classifiers in the convolutional neural network.

$$q_i^c = \frac{\exp(z_i^c / T)}{\sum_j^c \exp(z_i^c / T)} \quad (1)$$

Here z is the output after the layer is fully connected, $q_i^c \in R^M$ is the i-class probability of the classifier $\theta_{c/C}$.T is usually set to 1 to indicate the distillation temperature.

To improve the performance of the student model, two types of losses are introduced during training:

lossCE: Cross-entropy loss from the label to the deepest classifier, and cross-entropy loss for all shallow classifiers. It is used to train the dataset with labels and a softmax layer for each classifier. In this way, the hidden knowledge in the dataset is ingested directly from the label into all classifiers.

In self-distillation, there are two sources of supervision $\theta_{i/C}$ for each classifier except for the deepest classifier. Balance them with hyperparameter $\alpha$.

$$loss_{CE} = (1-\alpha) \cdot CrossEntropy(q^i, y) \quad (2)$$

The first source is the cross-entropy loss calculated with $q^i$ and Y labels. Note that $q^i$ represents the output of the Softmax layer of the classifier $\theta_{i/C}$.

lossKL: Teacher-led KL (Kullback-Leibler) divergence loss. The KL divergence is calculated using the softmax output between students and teachers and introduced into the softmax layer of each shallow classifier. By introducing KL divergence, the self-distillation framework influences the deepest teacher network to each shallow classifier.

$$loss_{KL} = \alpha \cdot KL(q^i, q^C) \quad (3)$$

The goal is to approximate the shallow classifier to the deep classifier, which indicates the supervision of distillation. $q^C$ represents the output of the softmax layer of the deepest classifier.

In summary, the loss function of the 3D feature matrix through the SD-CNN neural network is composed of the loss function of each classifier, which can be written as:

$$Loss = loss_{CE} + loss_{KL} \quad (4)$$

The output of the final EEG signal through the self-distillation convolutional network layer is represented as $Z^S$.

The 2D feature matrix of 32×6 is input into the GAT layer. In this model, two layers of GAT are used to process the spatial information of EEG signals. Specifically, the electrode channels of the EEG signals are used as the nodes of the graph, the connections between the electrodes are used as the edges of the corresponding graph, and the weights of all edges (representing the functional relationship between the electrodes) constitute the adjacency matrix of the graph. Once constructed, GAT can learn the intrinsic relationships between different EEG electrodes. The flow of GAT processing EEG signal features is shown in Figure 1. After data collection, preprocessing and feature extraction, the correlation matrix is used to calculate the spatial correlation, and the index size indicates the closeness of the relationship between EEG signal channels to complete the construction of the input map.

First, the correlation matrix of node feature $H = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\}$, $\vec{h}_i \in R^F$, time t is input into the GAT network. N is the number of electrode channels, and F is the number of features of each node. The attention mechanism of a node determines the weight of the features of adjacent nodes during feature update. Here, the dimension of the input features is transformed according to a learnable weight matrix $W \in R^{F' \times F}$, where $F'$ represents the dimension of the output node.

Firstly, the weight matrix is initialized during the model training process, and it is assumed that each electrode channel has an intrinsic relationship with the remaining 31 electrode channels. W is initialized as a diagonal matrix with a major diagonal of 0 and other values of 1. The optimal weight matrix is obtained through iterative training. Then, the degree of influence of nodes $i$ and j is calculated by $\vec{h}_i$ and $\vec{h}_j$.

$$e_{ij} = a(W\vec{h}_i, W\vec{h}_j) \quad (5)$$

Among them, the feedforward neural network a(·) represents the self-attention mechanism, which can stitch together the result vectors to complete the feature mapping. $e_{ij}$ indicates the importance of the features of node $j$ to $i$, the proposed model only calculates the first-order neighbors of each node.

Then, the attention coefficients of all nodes of node $i$ are calculated, and the normalization of the attention weights is completed by using softmax to obtain the final attention coefficients. As shown in equation (6).

$$a_{ij} = \frac{\exp\left(LeakyReLU\left(\vec{a}^T\left[W\vec{h}_i \| W\vec{h}_j\right]\right)\right)}{\sum_{k \in N_i} \exp\left(LeakyReLU\left(\vec{a}^T\left[W\vec{h}_i \| W\vec{h}_k\right]\right)\right)} \quad (6)$$

Where, $\|$ is a connection operator.

$LeakyReLU(\cdot)$ as a nonlinear activation function, can enhance the generalization ability of the model. Finally, the multi-head attention mechanism is used to learn the attention weights of node features to enhance the learning ability of the model. After being processed by the GAT attention layer, the features of node i can be expressed as equation (7).

$$\vec{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \in N_i} a_{ij}^k W^k \vec{h}_j\right) \quad (7)$$

The aggregation process of the multi-head attention mechanism on the node is shown in equation (6). The above is a complete graph convolution process, and the EEG signal features will be output after multi-layer graph convolution. Furthermore, the transportation is fused and classified with the fully connected layer and the extracted high-level abstract spatial features, and $Z^G$ is obtained by batch normalization before full connection. The K in Equation (7) denotes K independent attention mechanisms. In the experiment, K is 2.

Finally, feature fusion is performed. The deep features extracted from the self-distillation convolutional network

and the graph neural network are flattened and spliced, as shown in equation (8).

$$Output(Z^S, Z^G) = Concat(flatten(Z^S), flatten(Z^G)) \quad (8)$$

Finally, the emotional state is output using the softmax function.

## 4. Experiments and analysis of results

### 4.1. Dataset and Feature Extraction

The experiment selects the DEAP data set for sentiment analysis. In the DEAP data set, 32 subjects (including 16 males and 16 females, aged 19 to 37, mean 26.9 years old) recorded peripheral physiological and EEG signals while watching 40 music videos as stimuli. The EEG signals recorded in each video were 60 seconds long. And each video was chosen to stimulate a relevant emotional state. Six time-domain features of the 32-channel EEG signals of the samples were extracted, including mean, median, peak, average of the first absolute value of the difference, average of the second absolute value of the difference, and approximate entropy. Approximate entropy is a nonlinear dynamic characteristic, which is used to quantify the regularity and unpredictability of time series fluctuations. And it also represents the complexity of the time series. Therefore, the approximate entropy function can be used to reflect the complexity of EEG signals.

### 4.2. Data preprocessing

SDC-GAT was validated on the DEAP dataset. There are 1280 (32×40) EEG signals in the dataset, and deep learning needs a large amount of data to get better results, so the time segmentation method is used to increase the number of samples. Firstly, remove the first 3 sec baseline in each segment of EEG signal. Then, each EEG signal is divided into 10 fragments without overlap, each fragment contains 6 s of EEG signal, and each fragment again constitutes a sample and inherits the original label. Finally, the number of samples obtained is 12800, and the time-domain features are extracted from the EEG signals of multiple channels in the sample to form the input of the network. SDC-GAT contains two different neural networks, so the corresponding feature matrices are constructed for different network models. Each sample is mapped into a 3D feature matrix of 9×9×6 according to the position of the electrode on the scalp, and the sample of this shape is used as the input of the SDC-GAT network. Each sample is constructed into a 32×6 two-dimensional matrix and used as input to the GAT network.

### 4.3. Experimental setup

All experiments were implemented on GPU devices using the PyTorch framework. For the 12,800 samples extracted, the experimental results were verified by the ten-fold cross-validation technique. In training, the order of the samples was shuffled and then divided into 10 subsets. Eight of these subsets were selected as the training set and the remaining 2 subsets were used as the validation set, and this was done 10 times until all subsets were tested. To avoid overfitting the model, a dropout function is added to each fully connected layer. In the self-distillation network, the recommended value for the hyperparameter α is 0.5 and the distillation temperature is set to 1. In addition, the batch size is 64, the learning rate of the network is defined as 0.001, the maximum number of learning iterations is 400, and the network is optimized using the Adam optimizer. The accuracy and F1 score were used as evaluation indicators.

### 4.4. SDC-GAT model performance analysis

The emotion recognition model in the experiment consists of a convolutional self-distillation framework and a graph attention network. Among them, the convolutional self-distillation framework can be divided into three parts: the backbone, the attention module and the shallow classifier. The main part contains five convolutional layers, the first four layers use two convolution kernels of 1×1, 3×3, 5×5, and 7×7 respectively, and the fifth layer uses 1×1, 3×3 convolution kernels. In order to better learn EEG features and improve the fitting ability of the network, an activation function is added after the output of all convolutional layers. The specific structure of the attention module and the shallow classifier is shown in Figure 3. Figure 4 shows the training results of the proposed network on the dataset, which includes two dimensions: arousal and valence. The binary sentiment recognition results of the proposed network are shown in Figure 4.

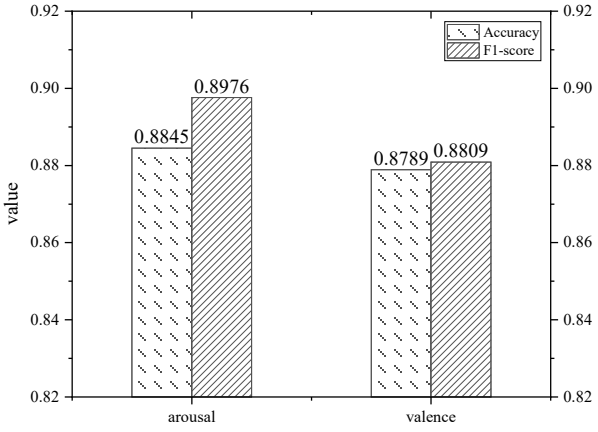


**Figure 4.** The result of binary classification of sentiment recognition

The SDC-GAT model achieved an accuracy of 0.8845 and an F1 score of 0.8976 in the arousal dimension. At the same time, in the valence dimension, the recognition accuracy

and F1 score were 0.8789 and 0.8809, respectively. The recognition results in Figure 4 demonstrate the effectiveness of the proposed model.

## 4.5. Comparison the ensemble model with the monolithic model

In order to prove that the ensemble model can effectively fuse the local and global information of EEG signals, the independent self-distillation convolution network and graph attention network are verified in experiments. In addition, a variety of CNN networks, graph convolutional neural networks (GCNs) and hybrid networks composed of these networks are constructed. Experiments were performed using processed EEG data and cross-validation with ten-fold cross-validation, and the experimental setup was the same as that of the SDC-GAT model. The CNN constructed in this subsection focuses on extracting the spatial information of EEG channels, including AlexNet and VGGNet. GCN is a natural generalization of CNN on graph structure. It is widely used in network analysis, traffic prediction, computer vision and other fields because it is suitable for extracting the structural features of graphs and has reliable performance in mining effective topological information and extracting key complex features from data. A graph convolutional network based on spectral domain is constructed in SDC-GAT, which consists of two convolutional layers. The parameter settings of the self-distilled convolutional network and the graph attention network are consistent with those in the integrated network, and the high-level abstract features extracted by the network are flattened and input into the fully connected layer for classification. The sample of subjects is consistent with the ensemble model and cross-validated using a 10-fold.

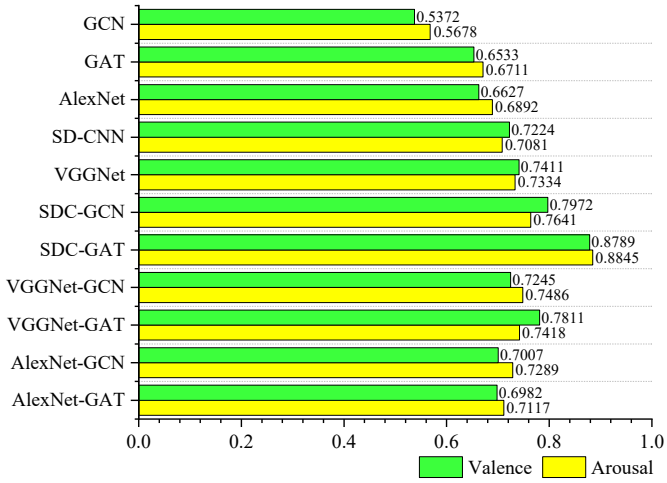


**Figure 5.** Comparison of the recognition results of multiple ensemble models and a single model

The EEG emotion recognition results of a single model and its ensemble model are shown in Figure 5. The results show that the performance of the integrated network is better than that of the single network, because each network in the integrated network can extract different information. In the arousal dimension, the Accuracy and F1 scores of SDC-GAT increased by 21.34% and 20.70% compared with GAT, and increased by 17.64% and 18.51% compared with SD-CNN, respectively. In terms of valence, the Accuracy and F1 scores of SDC-GAT increased by 22.56% and 19.88% compared with GAT, and increased by 15.65% and 14.54% compared with SD-CNN, respectively. This proves that each network in the ensemble model can extract different information, and the local information of the electrode channel and the global information of the EEG signal. VGG Net is the highest classification accuracy in a single network, with classification accuracy of 0.7334 and 0.7411 in the wake-up and valence dimensions, respectively. The SDC-GAT classification accuracy is the highest in the ensemble model, with classification accuracy of 0.8845 and 0.8789 in the wakefulness and titer dimensions, respectively. According to the classification results, GAT has a better ability to capture emotional information than GCN in EEG emotion recognition, and when combined with other models, it shows better emotion recognition performance.

## 4.6. Performance analysis of the distillation network

In order to explore the best performance of the distillation network, how to rationally use the distillation temperature to make the hidden knowledge better volatilize and condense, different distillation temperatures were set up for experiments. The distillation temperature is the hyperparameter T in equation (1). Except for the

distillation temperature change, the other parameters of the comparison experiment are the same as those of the SDC-GAT model. Table 1 shows the emotion recognition results.

Table 1. Effect of distillation temperature on distillation network performance

| Emotion dimension | Recognition results | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | T=1 | | T=2 | | T=3 | | T=4 | | T=5 | |
| | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score | Accuracy | F1-score |
| Arousal | **0.8845** | **0.8976** | 0.8789 | 0.8943 | 0.8226 | 0.8344 | 0.8757 | 0.8906 | 0.8257 | 0.8424 |
| Valence | **0.8789** | **0.8809** | 0.8679 | 0.8776 | 0.8734 | 0.8867 | 0.7992 | 0.8112 | 0.8703 | 0.8842 |

The classification results at different distillation temperatures are shown in Table 1. As can be seen from Table 1, distillation temperature affects the accuracy of sentiment classification. With the increase of distillation temperature, the classification accuracy generally decreased. In the arousal dimension, the classification results of T=2 and T=4 are similar, and the classification results of T=3 and T=5 are similar. In the valence dimension, the classification accuracy is the lowest when T=4, and the classification results are similar at other distillation temperatures. Distillation requires heating, and heating causes an increase in entropy. Increasing the temperature coefficient will lead to an increase in the information entropy of the output distribution, which will affect the results of sentiment classification. In the dimensions of arousal and valence, the distillation temperature T=1 has the highest classification accuracy and the best classification effect. Therefore, in the SDC-GAT model, the distillation temperature is set to 1.

## 4.7. Comparison with existing studies

In order to further validate the effectiveness of the SDC-GAT model, the model was compared with existing studies. These studies are based on the DAEP dataset. Gu et al. [37] proposed a frame-level distillation neural network to learn distillation features from the correlation of different frames. Joshi et al. [38] proposed a feature extractor based on Differential Entropy Linear (LF-DfE). Wang et al. [39] proposed a new emotion recognition model based on STFFNN, a hybrid spatiotemporal feature fusion neural network. Pandey et al. [40] used variational mode decomposition (VMD) as a feature extraction technique. Xefteris et al. [41] proposed a graph theory based on EEG functional connectivity patterns, which improved the performance of emotion recognition. Gao et al. [42] proposed EEG-GCN. In the dimensions of arousal and valence, the recognition accuracy of the binary classification task is shown in Figure 6.

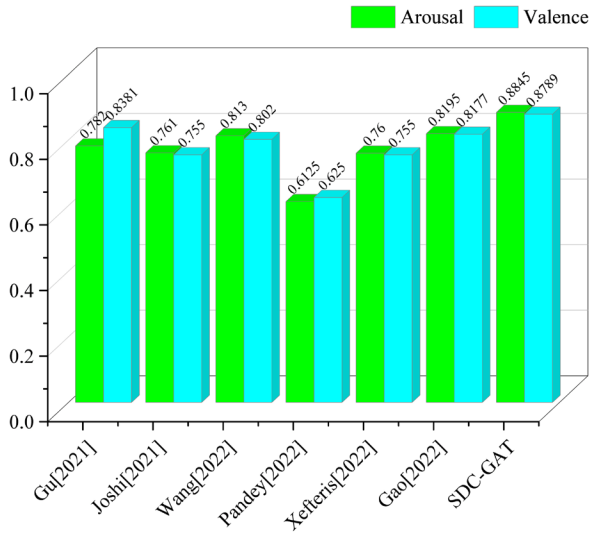


**Figure 6.** Comparison of SDC-GAT with existing studies

## 5. Conclusion

In this paper, a self-distillation convolutional graph attention network (SDC-GAT) EEG emotion recognition model is proposed. The model can excavate the sentiment information of EEG signals from the three-dimensional feature matrix and the two-dimensional feature matrix. The distillation convolutional network is used to excavate the local emotional features, and the GAT network is used to mine the global features. And through the ensemble model, the emotional feature fusion is effectively carried out.
Experimental results show that the SDC-GAT model can fuse the extracted local features with the global features. It uses the fused high-level abstract features to judge the emotional state, thereby improving the accuracy of emotion recognition. In addition, the sentiment recognition

performance of the model is compared with some existing models, which shows its superiority and verifies the feasibility and effectiveness of the model. In the next work, it is planned to combine EEG and facial video for multimodal continuous emotion recognition. The convolutional network using spatiotemporal attention mechanism is used to classify the sentiment of EEG signals. The decision-level fusion algorithm is used to iteratively learn and fuse the classification results of the two modalities, so as to further improve the performance of EEG emotion recognition.

## Acknowledgements.

This work is partially supported by the Program of National Natural Science Foundation of China (61872126), the Science and Technology Research Project of Henan Province (222102210078), the Natural Science Foundation of Henan Province (222300420445).

## References

[1] Liu S, Wang X, Zhao L, et al. Subject-independent emotion recognition of EEG signals based on dynamic empirical convolutional neural network[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020, 18(5): 1710-1721.

[2] Han J, Zhang Z, Pantic M, et al. Internet of emotional people: Towards continual affective computing cross cultures via audiovisual signals[J]. Future Generation Computer Systems, 2021, 114: 294-306.

[3] Liu S, Wang X, Zhao L, et al. 3DCANN: A spatio-temporal convolution attention neural network for EEG emotion recognition[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 26(11): 5321-5331.

[4] Hu W, Zhang Z, Zhao H, et al. EEG microstate correlates of emotion dynamics and stimulation content during video watching[J]. Cerebral Cortex, 2023, 33(3): 523-542.

[5] Dang W D, Lv D M, Li R M, et al. Multilayer network-based CNN model for emotion recognition[J]. International Journal of Bifurcation and Chaos, 2022, 32(01): 2250011.

[6] Iyer A, Das S S, Teotia R, et al. CNN and LSTM based ensemble learning for human emotion recognition using EEG recordings[J]. Multimedia Tools and Applications, 2023, 82(4): 4883-4896.

[7] Xin R, Miao F, Cong P, et al. Multiview Feature Fusion Attention Convolutional Recurrent Neural Networks for EEG-Based Emotion Recognition[J]. Journal of Sensors, 2023, 2023.

[8] Li Z, Zhang G, Wang L, et al. Emotion recognition using spatial-temporal EEG features through convolutional graph attention network[J]. Journal of Neural Engineering, 2023, 20(1): 016046.

[9] Zhang L, Bao C, Ma K. Self-distillation: Towards efficient and compact neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4388-4403.

[10] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129: 1789-1819.

[11] Wang X, Li Y. Harmonized dense knowledge distillation training for multi-exit architectures[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(11): 10218-10226.

[12] Zhang L, Bao C, Ma K. Self-distillation: Towards efficient and compact neural networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(8): 4388-4403.

[13] Yang Z, Li Z, Shao M, et al. Masked generative distillation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 53-69.

[14] Song T, Zheng W, Song P, et al. EEG emotion recognition using dynamical graph convolutional neural networks[J]. IEEE Transactions on Affective Computing, 2018, 11(3): 532-541.

[15] Yin Y, Zheng X, Hu B, et al. EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM[J]. Applied Soft Computing, 2021, 100: 106954.

[16] Ghandeharioun A, McDuff D, Czerwinski M, et al. Emma: An emotion-aware wellbeing chatbot[C]//2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, 2019: 1-7.

[17] Liu Y, Fu G. Emotion recognition by deeply learned multi-channel textual and EEG features[J]. Future Generation Computer Systems, 2021, 119: 1-6.

[18] Haj-Ali H, Anderson A K, Kron A. Comparing three models of arousal in the human brain[J]. Social Cognitive and Affective Neuroscience, 2020, 15(1): 1-11.

[19] Houssein E H, Hammad A, Ali A A. Human emotion recognition from EEG-based brain-computer interface using machine learning: a comprehensive review[J]. Neural Computing and Applications, 2022, 34(15): 12527-12557.

[20] Hjorth B. EEG analysis based on time domain properties[J]. Electroencephalography and clinical neurophysiology, 1970, 29(3): 306-310.

[21] Frantzidis C A, Bratsas C, Papadelis C L, et al. Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli[J]. IEEE transactions on Information Technology in Biomedicine,2010, 14(3): 589-597.

[22] Cai Y, Yao Z, Dong Z, et al. Zeroq: A novel zero shot quantization framework[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13169-13178.

[23] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[24] Liu Y, Cao J, Li B, et al. Knowledge distillation via instance relationship graph[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7096-7104.

[25] Cheng Y, Wang D, Zhou P, et al. Model compression and acceleration for deep neural networks: The principles, progress, and challenges[J]. IEEE Signal Processing Magazine, 2018, 35(1): 126-136.

[26] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[27] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv preprint arXiv:1412.6550, 2014.

[28] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. arXiv preprint arXiv:1612.03928, 2016.

[29] Yim J, Joo D, Bae J, et al. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning[C]//Proceedings of the IEEE conference

on computer vision and pattern recognition. 2017: 4133-4141.
[30] Lee S, Song B C. Graph-based knowledge distillation by multi-head attention network[J]. arXiv preprint arXiv:1907.02226, 2019.
[31] Furlanello T, Lipton Z, Tschannen M, et al. Born again neural networks[C]//International Conference on Machine Learning. PMLR, 2018: 1607-1616.
[32] Bagherinezhad H, Horton M, Rastegari M, et al. Label refinery: Improving imagenet classification through label progression[J]. arXiv preprint arXiv:1805.02641, 2018.
[33] Liu Y, Chen K, Liu C, et al. Structured knowledge distillation for semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2604-2613.
[34] Gupta S, Hoffman J, Malik J. Cross modal distillation for supervision transfer[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2827-2836.
[35] Kang M, Mun J, Han B. Towards oracle knowledge distillation with neural architecture search[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 4404-4411.
[36] Chao H and Dong L. Emotion Recognition Using Three-Dimensional Feature and Convolutional Neural Network from Multichannel EEG Signals[J]. IEEE SENSORS JOURNAL, 2021, 21(2): 2024-2034.
[37] Wang Z, Gu T, Zhu Y et al. FLDNet: Frame-level distilling neural network for EEG emotion recognition[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(7): 2533-2544.
[38] Joshi V `M, Ghongade R B. EEG based emotion detection using fourth order spectral moment and deep learning[J]. Biomedical Signal Processing and Control, 2021, 68: 102755.
[39] Wang Z, Wang Y, Zhang J, et al. Spatial-temporal feature fusion neural network for EEG-based emotion recognition[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.
[40] Pandey P, Seeja K R. Subject independent emotion recognition from EEG using VMD and deep learning[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(5): 1730-1738.
[41] Xefteris, V, TSANOUSA A, GEORGAKOPOULOU N, et al. Graph Theoretical Analysis of EEG Functional Connectivity Patterns and Fusion with Physiological Signals for Emotion Recognition. Sensors, 2022. 22(21): 8198.
[42] Gao Y, FU X L, OUYANG T X, et al. EEG-GCN: Spatio-Temporal and Self-Adaptive Graph Convolutional Networks for Single and Multi-View EEG-Based Emotion Recognition. IEEE SIGNAL PROCESSING LETTERS, 2022. 29: 1574-1578.