EasyChair Preprint

№ 13754

Detection of Depression in EEG Signals Based on Convolutional Transformer and Adaptive Transfer Learning

Qianqian Tan and Minmin Miao

July 2, 2024

Detection of depression in EEG signals based on convolutional transformer and adaptive transfer learning

Qianqian Tan✉ and Minmin Miao

School of Information Engineering, Huzhou University, Huzhou 313000, China
1002537620@qq.com
02746@zjhu.edu.cn

Abstract. Electroencephalography (**EEG**) signals provide an objective reflection of the inner workings of the brain, making them a promising tool for the diagnosis of depression. However, the classification of EEG signals for depression is severely affected by individual differences among subjects, complex intrinsic properties, and low Signal-to-Noise Ratio (**SNR**), which limits the classification accuracy. Additionally, traditional convolutional neural networks extract local features but fail to capture long-term dependencies in EEG decoding. To address the aforementioned issues, we introduce an adaptive transfer learning method based on a convolutional transformer model for depression detection. The experimental results demonstrate the effectiveness of the proposed model on the public MODMA dataset and EDRA dataset. The results indicate that the MODMA and EDRA datasets exhibit optimal accuracies of 100% and 98.61%, respectively, outperforming some state-of-the-art depression identification methods. Our findings provide new perspectives on the recognition of depression, which could be used as an assisted diagnostic tool in the future.

Keywords: Transfer learning · EEG · Depression detection · Convolutional transformer.

1 Introduction

Depression is a common mental disorder that causes persistent sadness, feelings of hopelessness, and low self-esteem [4]. Unfortunately, clinical diagnostic methods for depression have some limitations, which can delay effective treatment for patients. That is why we need to develop an objective assessment method to detect depression accurately and quickly. EEG is widely used in the medical field as a rapid, noninvasive tool for monitoring brain activity and is effective in diagnosing a variety of disorders, including depression [19], epilepsy [6],

This work was supported in part by the National Natural Science Foundation of China under Grants 62101189 and U20A20228, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LTGC23F010001.

and psychiatric disorders [3]. In recent years, there has been a growing interest among researchers in utilizing deep learning models to analyze depression EEG signals. Due to the limitation of the kernel size of Convolutional Neural Networks (**CNNs**), they can only capture features in the local receptive field, thus making it difficult to capture long-term dependencies in a time series. For this reason, researchers have further introduced Recurrent Neural Networks (**RNN**) and Long Short-Term Memory Networks (**LSTM**) to capture temporal features in EEG classification tasks [12, 17]. However, such models are not suitable for training in parallel and are prone to lose their hidden states rapidly as the time step increases. Given the significance of global dependency, transformer models based on the attention mechanism have emerged in the field of EEG decoding and have demonstrated promising performance by capitalizing on long-term temporal relationships [14, 10]. Moreover, the effect of individual variability on EEG signal data, including differences in individual brain structure and neural response levels, results in significant differences in the distribution of EEG signal data among subjects. Consequently, cross-subject classification tasks frequently exhibit lower accuracy.

In order to address the aforementioned challenges and shortcomings, we introduce an adaptive transfer learning method based on a convolutional transformer model for depression detection. The paper's main contributions include:

- An end-to-end adaptive transfer learning method based on convolutional transformer model for decoding EEG signals with depression. In addition, the necessity of adaptive transfer learning is demonstrated, and the generalization ability of the model is enhanced by a fine-tuning approach.
- An extensive number of experiments using the publicly available datasets MODMA and EDRA are conducted to validate the performance of our model. The results of these experiments provide strong evidence that the proposed model achieves state-of-the-art performance. This suggests that the transfer learning approach is an effective method for addressing the challenges of data distribution differences and limited samples.

2 Related Work

In recent years, the application of deep learning models to analyze depression EEG signals has attracted the attention of many researchers. Qayyum et al. [9] integrated a shallow network of convolutional neural networks and gated recurrent units (**GRUs**) for depression diagnosis. Zhang et al. [21] proposed a 2DCNN-LSTM classification model that fully utilizes spatial and temporal information, achieving an accuracy of 95.1% in depression diagnosis. Zhang et al. [20] introduced LSTM to extract temporal-domain features and a two-dimensional convolutional neural network to extract spatial-domain features. The detection of major depressive disorder (**MDD**) through the temporal and spatial feature fusion approach achieved an accuracy of 96.33%. Wan et al. [16] proposed

a transformer-based EEG analysis model, EEGformer, which employs a one-dimensional convolutional neural network to automatically extract EEG channel features and combines multiple transformer modules in order to uniquely capture multiple EEG features.

3 Methodology

This study introduces an adaptive transfer learning method based on a convolutional transformer model for EEG signal decoding in depression. The model employs convolution to learn local features, and then employs self-attention to encapsulate global features, enhancing the model's generalization ability through different methods of fine-tuning the scheme. The architecture of this network is shown in Fig. 1.

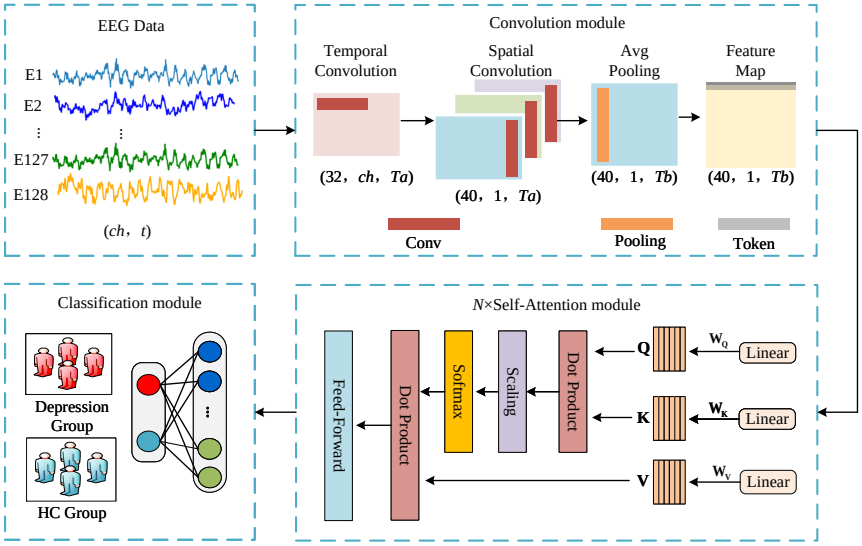

Fig. 1. The convolutional transformer model consists of four modules: data input, convolution, self-attention, and classification, where t, Ta, and Tb represent the sampling points, ch represents the number of channels, and $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent matrices.

3.1 Convolution Module

First, two 1D convolutional layers are employed to design the convolutional module for temporal and spatial dimensions, respectively, with the objective of extracting local features effectively. Subsequently, the batch normalization technique is introduced to simplify the training process and alleviate the overfitting

problem. Meanwhile, the exponential linear unit (**ELU**) nonlinear activation function is chosen. In the subsequent stage of feature extraction, we employ the average pooling operation along the time dimension. Finally, the feature maps of the convolution module are rearranged, and each sample point is input as a token to the self-attention module.

3.2 Self-Attention Module

In this module, we use self-attention mechanisms to capture the global time dependence of EEG features in depression. The process starts by creating query vectors ($\mathbf{Q}$), key vectors ($\mathbf{K}$), and value vectors ($\mathbf{V}$) using three linear layers. The dot product of the $\mathbf{Q}$ and $\mathbf{K}$ vectors is used to compute the pairwise similarities between each query and all the keys. To prevent the gradient from vanishing, we normalize these similarities by dividing by the scaling factor. Next, we apply the softmax function to obtain the weight matrix. Finally, the weight matrix is multiplied with the values ($\mathbf{V}$) by a dot product operation[15]. This entire attention computation is repeated N times in the self-attention module. The process can be expressed as follows:

$$\text{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}}\right)\mathbf{V} \tag{1}$$

In order to enable the model to perceive the global dependence of depressed EEG signals from different locations, we used a multi-head strategy. This requires dividing the feature map token of the previous module into H segments and then merging the output of each head to form the final output of the module. This process can be described as follows:

$$\begin{aligned} \text{MultiHead}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) &= \text{Concat}\left(head_0; \cdots; head_{H-1}\right) \in \mathbf{R}^{N \times d} \\ head_h &= \text{Attention}\left(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h\right) \in \mathbf{R}^{N \times \frac{d}{H}} \end{aligned} \tag{2}$$

where the dimension of the input feature map is $N \times d$, $head_h$ denotes the output of the hth head after the attention mechanism, Concat denotes the operation used to splice all $head_h$.

3.3 Classification Module

In this module, the output of the self-attention module is employed as the input of the fully connected layer. Subsequently, the category with the greatest probability value is regarded as the final classification result through the softmax function. The loss function of the entire framework is based on cross entropy, which is calculated as follows:

$$Loss = -\frac{1}{N_c}\sum_{i=1}^{N_c}\sum_{j=1}^{n} y_{ij}\log\left(\hat{y}_{ij}\right) \tag{3}$$

where N_c denotes the number of samples in the current batch and n denotes the number of categories. In addition, y_{ij} and $\hat{y}_{ij}$ are the true and predicted labels, respectively.

3.4 Transfer Learning Strategy

CNN-based classification algorithms generally require a significant amount of training data, which can lead to longer computation times. To address this issue, transfer learning is commonly used to obtain pre-trained models. Due to the differences in data distribution between the source and target domains, the application of transfer learning to the EEG classification task for depression faces certain challenges. In this study, the source domain consists of data from non target subject, while the target domain consists of data from target subject. To achieve domain adaptation, this study employs a fine-tuning strategy so that the model trained using the source domain can be better adapted to the properties of the target domain. Hence, the adaptation scheme focuses on fine-tuning the model parameters using a part of target subject data to optimize the model's performance in the EEG depression classification task in the target subject. This study considered two different classification strategies: subject-independent and subject-adaptive. In the subject-independent classification, all data except for the target subject were used for training. For each target subject, we performed a 2-fold cross validation on the data from remaining subjects for model selection. Since the model never observes any data from the target subjects during training, this is prone to inter-subject variation. Therefore, in the subject-adaptive classification method, different proportions of target subject data (see Section 4.2 for dataset partitioning) are used to fine-tune the pre-trained model in order to investigate the effect of different degrees of adaptive α on classifier accuracy. In this study, three different adaptation schemes AS-1 , AS-2 and AS-3 were employed. Each scheme fine-tunes different parts of the pre-trained model to improve the classification performance on the target subject. In the first adaptation scheme AS-1, the fully connected (**FC**) layer has been optimized while the rest of the network parameters remain unchanged. In the second adaptation scheme AS-2, the transformer module and the fully connected layer have been retrained using adaptation data, while the rest of the network parameters remain unchanged. In the third adaptation scheme AS-3, the entire network model has been retrained using adaptation data. The framework of the scheme is illustrated in Fig. 2.

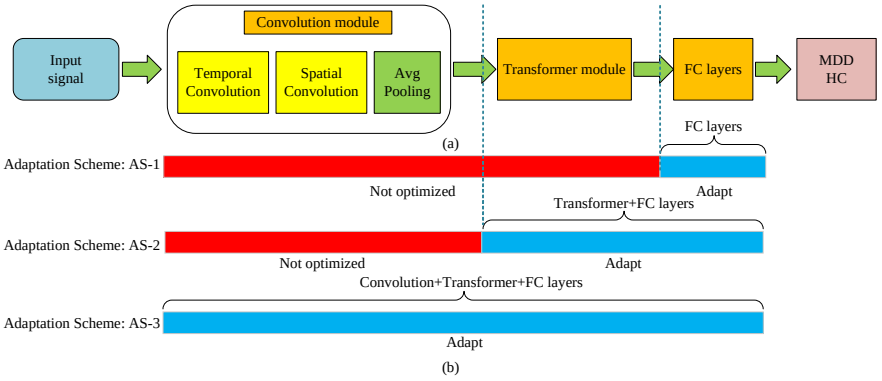

Fig. 2. Illustrations of (a) Network architecture and (b) three different adaptation schemes AS-1, AS-2 and AS-3.

4 Experiments and Results

This section presents the dataset preprocessing, experimental setup, and results for various schemes, which are then compared with state-of-the-art algorithms to validate the superiority of our approach.

4.1 Dataset Preprocessing

In this paper, we use the MODMA dataset [1] and the EDRA dataset [18]. For the MODMA dataset, the dataset consists of 53 subjects, of which 24 are diagnosed depression patients and 29 are healthy individuals. The device sampling frequency was 250 Hz and the reference electrode was a Cz electrode. The data recording time for each subject was approximately 5 minutes. The raw EEG data signals from all 128 channels were filtered using a band-pass filter from 0.5 Hz to 50 Hz, and then re-referenced using the common average reference (**CAR**) method. To reduce computational resources, we chose to study 64 of the 128 electrodes in this paper [20]. For uniform processing, we divided each participant's data into 150 segments of 2 seconds each, for a total of 7950 segments. For the EDRA dataset, which consisted of 50 subjects, 26 were categorized as being at high risk for depression, while 24 were considered to be at low risk. The EEG signal acquisition device was an EEG cap with 64 electrodes, in accordance with international standards. Following the final screening, 62 valid electrodes were selected for further experimental studies. The reference electrode for the online recordings was the FCz electrode, while the offline recordings used a global average reference. The EEG signals were filtered using a 0.05-100 Hz online bandpass filter. The sampling frequency was 500 Hz. Independent component analysis (**ICA**) was used to correct for ocular artifacts. Each subject

intercepted a 1-minute continuous EEG recording and divided it into 60 equal-length segments of one second in length. Finally, a total of 3000 samples were obtained.

4.2 The setting of Experiment

Subject-independent and subject-adaptive experiments were conducted separately in the current study. (1) In subject-independent classification, all data except for the target subjects were used for training. For each target subject, we performed a two-fold cross-validation of the remaining subjects' data for model selection. (2) In subject-adaptive classification, the model with minimum validation loss across cross-validation folds in the subject-independent classification is used as the base model. We used only the first 60% of the target subjects' data to fine-tune the network, ranging from 10% to 100% in steps of 30%. The model was implemented in the PyTorch framework and deployed on a GeForce RTX 3090. The Adam optimizer [5] is employed to minimize the loss function, with a learning rate of 0.00008, and the batch size to 50. We set the number of filters in the convolution module to 32, the number of self-attention executions N to 6, and the number of heads h to 10.

4.3 Experimental Results and Analysis

This section presents a discussion of the accuracy of subject-independent and subject-adaptive classification models. In subject-independent experiments, the average accuracy among all subjects on the MODMA dataset was 51.36%. The experimental results reflect that the generalization ability of the model varies due to differences in the distribution of the collected EEG signal data among subjects. In order to reduce the individual differences among subjects and enhance the model's generalization performance, the method of fine-tuning was employed. Table 1 provides the average results for all subjects in the subject adaptive experiment for the MODMA dataset. For the MODMA dataset, the accuracy of scheme AS-1 can exceed 97% at an adaptation rate of 10%, and can reach 100% at an adaptation rate of 40%. The remaining schemes achieve 100% accuracy. To further validate the effectiveness of the model, we conducted further experiments on the EDRA dataset according to scheme AS-3, which showed an average accuracy of 98.61% among all subjects, which is better than some of the state-of-the-art depression recognition methods. Experimental results show that the model has a certain generalization ability. A comparison was made between our method and other state-of-the-art depression identification methods to demonstrate the superiority of our work. Table 2 and Table 3 list the state-of-the-art studies and their corresponding performances in recent years for the MODMA dataset and EDRA dataset, respectively. The following methods have achieved impressive results on the MODMA dataset or EDRA dataset. For example, the study [2] fully considered all channel information in EEG-based major depressive disorder recognition and designed a random search algorithm to select the best discriminative features describing each channel. The study [8]

transformed EEG signals into brain maps containing temporal, frequency and spatial information, and utilized CNNs and GRUs to achieve classification of depressed and healthy individuals. The study [21] presents a hybrid neural network based on CNN and LSTM for the automatic detection of depression. In study [13], a novel Twin Pascal's Triangles Lattice Pattern (**TPTLP**) was employed to extract local texture features from raw depression EEG signals and subbands. The study [22] proposed a new model for learning depression detection from EEG signals by adaptive channel optimization (**MGCL-ACO**) multi-view contrast. Research [18] proposed a graph-based adaptive least absolute shrinkage and selection operator model (**GA-LASSO**) to learn the discriminant features of FC matrices.

Table 1. The average accuracy (%) of the fine-tuned pre-trained model among all subjects on the MODMA dataset, where AS denotes the adaptation scheme and α denotes the proportion of adapted data.

Scheme	Adaptation Rate			
	α=10%	α=40%	α=70%	α=100%
AS-1	97.70	100	100	100
AS-2	100	100	100	100
AS-3	100	100	100	100

Table 2. Results of our model compared to state-of-the-art models of depression diagnosis using the MODMA dataset.

Method	Accuracy (%)
BrainMap + CNN + GRU [8]	89.63
CNN-LSTM [21]	95.10
GA-LASSO [18]	97.43
MGCL-ACO [22]	99.19
BLDA-RSSA [2]	99.32
TPTLP+Greedy algorithm [13]	**100**
Proposed model	**100**

* GRU: Gate Recurrent Unit; LSTM: Long Short-Term Memory; GA-LASSO: Graph-based Adaptive Least Absolute Shrinkage and Selection Operator; MGCL-ACO: Multi-view Graph Contrastive Learning via Adaptive Channel Optimization; TPTLP: Twin Pascal's Triangles Lattice Pattern.

Table 3. Results of our model compared to state-of-the-art models of depression diagnosis using the EDRA dataset.

Method	Accuracy (%)
ShallowConvNet [11]	84.69
EEGNet [7]	94.67
DeepConvNet [11]	95.94
GA-LASSO [18]	97.33
MGCL-ACO [22]	98.38
Proposed model	**98.61**

5 Conclusion

In this study, we introduce an adaptive transfer learning method based on a convolutional transformer model for depression detection. By studying the effects of different adaptation ratios on different adaptation schemes, we determine the optimal adaptation strategy. Experiments on the MODMA dataset and EDRA dataset show that when the AS-3 scheme is fine-tuned using an adaptive rate of 10%, the accuracies are up to 100% and 98.61%, respectively, which suggests that the model still has a strong generalization ability with a small number of data samples.

References

1. Cai, H., Yuan, Z., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., Li, J., Yang, Z., Li, X., et al.: A multi-modal open dataset for mental-disorder analysis. Scientific Data **9**(1), 178 (2022)
2. Chang, H., Zong, Y., Zheng, W., Xiao, Y., Wang, X., Zhu, J., Shi, M., Lu, C., Yang, H.: Eeg-based major depressive disorder recognition by selecting discriminative features via stochastic search. Journal of Neural Engineering **20**(2), 026021 (2023)
3. Chen, H., Lei, Y., Li, R., Xia, X., Cui, N., Chen, X., Liu, J., Tang, H., Zhou, J., Huang, Y., et al.: Resting-state eeg dynamic functional connectivity distinguishes non-psychotic major depression, psychotic major depression and schizophrenia. Molecular Psychiatry pp. 1–11 (2024)
4. Kennis, M., Gerritsen, L., van Dalen, M., Williams, A., Cuijpers, P., Bockting, C.: Prospective biomarkers of major depressive disorder: a systematic review and meta-analysis. Molecular psychiatry **25**(2), 321–338 (2020)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Kunekar, P., Gupta, M.K., Gaur, P.: Detection of epileptic seizure in eeg signals using machine learning and deep learning techniques. Journal of Engineering and Applied Science **71**(1), 21 (2024)
7. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering **15**(5), 056013 (2018)

8. Liu, W., Jia, K., Wang, Z., Ma, Z.: A depression prediction algorithm based on spatiotemporal feature of eeg signal. Brain Sciences **12**(5), 630 (2022)
9. Qayyum, A., Razzak, I., Mumtaz, W.: Hybrid deep shallow network for assessment of depression using electroencephalogram signals. In: Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part III 27. pp. 245–257. Springer (2020)
10. Qayyum, A., Razzak, I., Tanveer, M., Mazher, M., Alhaqbani, B.: High-density electroencephalography and speech signal based deep framework for clinical depression diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2023)
11. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for eeg decoding and visualization. Human brain mapping **38**(11), 5391–5420 (2017)
12. Song, X., Yan, D., Zhao, L., Yang, L.: Lsdd-eegnet: An efficient end-to-end framework for eeg-based depression detection. Biomedical Signal Processing and Control **75**, 103612 (2022)
13. Tasci, G., Loh, H.W., Barua, P.D., Baygin, M., Tasci, B., Dogan, S., Tuncer, T., Palmer, E.E., Tan, R.S., Acharya, U.R.: Automated accurate detection of depression using twin pascal's triangles lattice pattern with eeg signals. Knowledge-Based Systems **260**, 110190 (2023)
14. Tigga, N.P., Garg, S.: Efficacy of novel attention-based gated recurrent units transformer for depression detection using electroencephalogram signals. Health Information Science and Systems **11**(1), 1 (2022)
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
16. Wan, Z., Li, M., Liu, S., Huang, J., Tan, H., Duan, W.: Eegformer: A transformer–based brain activity classification method using eeg signal. Frontiers in Neuroscience **17**, 1148855 (2023)
17. Wu, X., Yang, J.: The superiority verification of morphological features in the eeg-based assessment of depression. Journal of Neuroscience Methods **381**, 109690 (2022)
18. Yang, L., Wei, X., Liu, F., Zhu, X., Zhou, F.: Automatic feature learning model combining functional connectivity network and graph regularization for depression detection. Biomedical Signal Processing and Control **82**, 104520 (2023)
19. Ying, M., Shao, X., Zhu, J., Zhao, Q., Li, X., Hu, B.: Edt: An eeg-based attention model for feature learning and depression recognition. Biomedical Signal Processing and Control **93**, 106182 (2024)
20. Zhang, B., Wei, D., Yan, G., Li, X., Su, Y., Cai, H.: Spatial–temporal eeg fusion based on neural network for major depressive disorder detection. Interdisciplinary Sciences: Computational Life Sciences **15**(4), 542–559 (2023)
21. Zhang, J., Xu, B., Yin, H.: Depression screening using hybrid neural network. Multimedia Tools and Applications **82**(17), 26955–26970 (2023)
22. Zhang, S., Wang, H., Zheng, Z., Liu, T., Li, W., Zhang, Z., Sun, Y.: Multi-view graph contrastive learning via adaptive channel optimization for depression detection in eeg signals. International Journal of Neural Systems **33**(11), 2350055–2350055 (2023)