EasyChair Preprint

№ 8406

EEG Classification Algorithm of Motor Imagery Based on CNN-Transformer Fusion Network

Haofeng Liu, Yuefeng Liu, Yue Wang, Bo Liu and Xiang Bao

July 6, 2022

EEG classification algorithm of motor imagery based on CNN-Transformer fusion network

line 1: Haofeng Liu
line 2: *School of Information Engineering*
line 3: *Inner Mongolia University of Science&Technology*
line 4: Baotou, China
line 5: 410244038@qq.com

line 1: Yuefeng Liu
line 2: *School of Information Engineering*
line 3: *Inner Mongolia University of Science&Technology*
line 4: Baotou, China
line 5: liuyuefeng@imust.edu.cn

line 1: Yue Wang
line 2: *School of Information Engineering*
line 3: *Inner Mongolia University of Science&Technology*
line 4: Baotou, China
line 5: wy18846794426@163.com

line 1: Bo Liu
line 2: *School of Information Engineering*
line 3: *Inner Mongolia University of Science&Technology*
line 4: Baotou, China
line 5: 15261865055@163.com

line 1: Xiang Bao
line 2: *School of Information Engineering*
line 3: *Inner Mongolia University of Science&Technology*
line 4: Baotou, China
line 5: 1643851893@qq.com

***Abstract*—In recent years, with the development of social economy and technology, the brain-computer interface based on motor imagery(MI-BCI) has gradually become the focus content of many re-searchers. However, the motor imagery EEG signal (MI-EEG) itself has the characteristics of non-linearity and low signal-to-noise ratio, and because the characteristics of different domains of MI-EEG cannot be effectively combined, the recognition rate of MI-EEG is unsatisfactory. To overcome the above problems, this paper proposes a Transformer-based one-dimensional convolutional neural network model(CNN-Transformer) for the classification and recognition of four types of motor imagery EEG signals. Firstly, the artifacts of the original EEG are removed and new time-space-frequency features are constructed by preprocessing such as bandpass filtering and PCA dimensionality reduction; then the local features in the time dimension are extracted through the convolution and pooling operations of 1D-CNN, while reducing the time The dimension of the feature; next, the Transformer based on the attention mechanism is used to extract more abstract and high-level temporal features from multiple perspectives; finally, the classification results are integrated and output through the fully connected layer. The performance of the CNN-Transformer model is evaluated using the competition dataset 2008 BCI-Competition 2A. The results show that the average accuracy and kappa value of the CNN-Transformer model are as high as 99.29%(±0.07%) and 98.43%(±0.21), respectively, which are 3.72% and 7.68% higher than the classical architecture (CNN-LSTM). This model provides a design idea for improving the accuracy of MI-EEG classification and recognition, and also lays a foundation for the wide application of MI-BCI.**

Keywords—motor imagery; brain-computer interface; deep learning; classification and recognition; convolutional neural network; transformer

I. INTRODUCTION

The 21st century is called "the age of brain science", and brain-computer interface (BCI) is one of the most popular research contents in the 21st century. BCI is a system that transmits information between the brain and a computer or other equipment, it does not directly rely on the utilization of external nerve and muscle tissue, but realizes the desire of human mind control by processing electroencephalographic signals (EEG) into control commands [1], such as exoskeleton machine [2], robotic arm [3], etc.

According to the different generation modes of EEG, BCI can be divided into spontaneous BCI and evoked BCI. Evoked BCI has a good effect due to its strong regularity and high stability of the evoked potential itself, and extensive research has been done in SSVEP, P300 and other aspects. Compared with the EEG modality of induced BCI, spontaneous MI-EEG has strong spontaneity and naturalness, so it is more suitable as an EEG control signal for the study of MI-BCI system [4]. But, MI-EEG itself has the characteristics of low signal-to-noise ratio and nonlinearity, which indirectly results in the poor recognition effect of spontaneous BCI classification. However, the implementation and application of BCI technology largely depends on the classification and recognition rate, so improving the accuracy of classification and recognition has become the main research content of researchers.

A. Related Work

According to the phenomenon of event-related desynchronization and event-related synchronization (ERS/ERD) [5], researchers have proposed many traditional machine learning methods for decoding MI-EEG, these methods mainly use wavelet packet transform (DWT), co-spatial mode (CSP) and other methods to manually extract time, frequency, space, or time-frequency features, and then use support vector machine (SVM), k-nearest neighbor classification, artificial neural network and other shallow model classification algorithms to obtain classification accuracy, such as Xiaojun Yu, Zhaohui Yuan[6] and others first proposed a data-adaptive empirical wavelet transform (EWT)-based signal decomposition method, which uses Welch power spectral density (PSD) and Hilbert transform (HT) to decompose the signal. Component extraction, and the use of least squares support vector machine (LS-SVM) to classify and identify motor imagery EEG signals with an accuracy of 94.6%. NiteshSingh Malan, Shiru Sharma [7] etc. used CSP algorithm to extract important features of MI-EEG, and used SVM to classify the features extracted by CSP, with an accuracy of 91.7%. Piyush Kant [8] et al. extracted mean, variance, wavelet energy, Shannon entropy, log energy

entropy, kurtosis and skewness features from EEG signals recorded at symmetrical electrode positions in the motor cortex, and used SVM classification with an accuracy of 86.4 %. Poonam Chaudhary [9] et al. used the CSP of the wavelet decomposition signal for feature extraction, and then used a decision tree classifier, which achieved a classification accuracy of 85.6%. Yimin Hou, Tao Chen [10], etc. proposed a new framework based on bi-spectrum, entropy and common space pattern (BECSP), specifically using the bi-spectrum, entropy and CSP methods in higher-order spectrum to extract MI- EEG signal features, and then select the feature with the largest contribution through a tree-based feature selection algorithm. By using the SVM algorithm based on the RBF kernel function for classification, the highest accuracy rate reaches 85%.To sum up, the traditional machine learning technology is very mature and the classification effect is also effective, but because the representative EEG features are extracted manually, the characteristics of the original EEG itself are ignored, and the process is more complicated. , the efficiency is also lower. And most classification algorithms are only for binary classification problems.

Deep learning methods have been proven to outperform traditional machine learning methods in many fields, such as: computer vision, natural language processing, biomedical signal processing, etc. Deep learning technology has gradually extended from these popular fields to the BCI field, and has become one of the latest and most important tools in the BCI field. The deep learning model can automatically extract the more representative deep abstract features of MI-EEG without complex preprocessing and feature extraction process. At the same time, the end-to-end features of deep learning well preserve the features of the original signal. In recent years, many researchers have carried out EEG signals decoding research based on deep learning methods, such as Hongli Li, Man Ding [11], etc. proposed a parallel CNN-LSTM hybrid neural network, in which CNN is responsible for extracting EEG space LSTM is responsible for extracting temporal features, and finally outputs the classification results through the fully connected layer, with an average accuracy of 87.68%. Chu Yaqi [12] and others proposed a neural network based on joint convolution of spatial-temporal features, which sequentially convolved temporal and spatial features to extract spatial-temporal features, and the average classification accuracy reached 80.09%. Jinzhen Liu, Fangfang Ye [13] and others designed a convolutional neural network and a cascaded network of gated recurrent units to learn time-frequency information from EEG data, and the classification and recognition accuracy reached 92.56%. Mouad Riyad, AbdellahAdib [14], etc. developed a ConvNet based on Inception and Xception architecture, which uses convolutional layers to extract temporal and spatial features, and employs separable convolution and depthwise convolution to achieve a faster and more efficient ConvNet. Then, a new block inspired by Inception is introduced to learn richer features to improve classification performance. Although the appeal decoding method has achieved good results, it is still a difficult problem to efficiently combine the features of different domains. Moreover, although CNN can better capture local features in different domains of EEG, local features between different layers cannot be well correlated. At the same time, with the increase of the amount of data and the number of network layers, the performance and adaptability of the model will decline. Although RNN is a natural time series model, its inherent sequential property hinders parallelization among training samples. For long sequences, memory constraints will hinder batch processing of training samples.

B. Central Idea

Transformer [15] completely abandoned the traditional CNN and RNN, and the entire network structure is completely composed of the attention mechanism, which is the latest and most popular deep learning model in recent years. Compared with CNN and RNN, Transformer can use the multi head attention mechanism to learn the relationship between different layers of features. It has better universality and strong comprehensive feature extraction ability. It can not only pay attention to the current information, but also expand from the current local information to the global information. Moreover, it outperforms other models in long-sequence feature correlation calculation and model visualization and interpretability. But it is also because Transformer abandons the structure of CNN and RNN, so its ability to capture and analyze local features is poor, and because of the small amount of data in this paper, it cannot make up for this defect through huge data. At the same time, although Transformer can calculate the attention between any two nodes through the self-attention mechanism, so that it has the ability to capture and analyze long time series data, for long time series, the analysis ability will still decline with the extension of the sequence.

Combining the above considerations, this paper proposes a Transformer-based one-dimensional convolutional neural network model CNN-Transformer, and applies the model to four types of MI tasks. The main contributions of this paper are as follows: (1) According to the original characteristics of MI-EEG, for the spatial domain, the original MI-EEG is decomposed according to different frequency bands, and then the decomposed feature sequences of different frequency bands are fused to construct new spatial features, and PCA is used to extract the main features of spatial dimensions; For the frequency domain, according to the filtered results, the differential entropy(DE) of each frequency band of each channel is calculated and transformed into one-dimensional characteristic sequence; For the time domain, a new time-frequency feature is constructed by combining the feature sequence processed in the frequency domain with the feature sequence of the time domain itself, and then the dimension of the time-frequency feature is reduced through the sliding window, which also solves the problem of small amount of data. (2) In order to solve the problems of poor analysis ability, large amount of calculation and weak ability to capture local features caused by long EEG time-frequency sequences, the CNN-Transformer model is designed. Through 1D-CNN convolution and pooling operations in time-frequency domain, the low-level time-frequency features are extracted and the time-frequency feature dimension is reduced for the Transformer, paving the way for subsequent Transformers to extract higher-level features. (3) Use optimization algorithms such as cross-validation to optimize the parameters and structure of the model CNN-Transformer. Finally, based on the same data set, the experimental results of the model proposed in this paper, the classic architecture CNN-LSTM

and other models are significantly compared and analyzed to verify the effectiveness and practicability of the model.

II. Materials and Methods

A. Data Description

The experimental data BCI Competition 2008 – Graz data set A is taken from the 2008 International Brain-Computer Interface Competition The dataset consists of MI-EEG of 9 subjects in 4 classes, namely left hand (class1), right hand (class2), foot (class3), and tongue (class4). The experimental procedures and contents were the same for each subject.

The content of the experiment is: each subject records 2 groups of experimental data at different time points, each group includes 6 groups of small experiments, and each group of small experiments contains 48 segments (i.e.: 4 types of actions, each type of action is randomly repeated 12 times, A total of 4 × 12 = 48 times), each group of experimental data contains a total of 288 sections (6 groups × 12 sections = 288 sections), and 2 groups of experimental data have a total of 576 sections (2 groups × 6 groups × 48 sections = 576 sections). Since one set of experiments does not contain the labels required for training, this paper only utilizes the set of experimental data with labels.

The experimental process is as follows: at the beginning of the test (t=0s), a fixed cross will appear on the black screen, in addition to a brief audible prompt tone, after two seconds (t=2s), one will point left and right as a prompt , down or up (corresponding to the four categories of left hand movement, right hand movement, foot movement and tongue movement) arrows will appear on the screen for about 1.25s, which prompts subjects to imagine the movement corresponding to the picture, each subject The subjects were asked to complete these imaginary tasks until the cross on the screen disappeared (t=6s), and then took a short rest until the screen turned black again, and this process was repeated 288 times in stages. Figure 1. shows the flow of the single-motion imagery experiment.

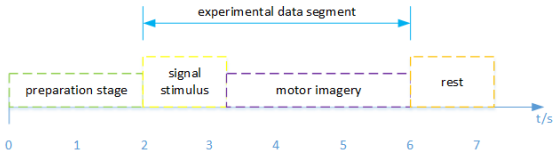

Figure 1. Single motor imagery sequence

For the accuracy of the experiment, we only intercepted the data 4s after the prompt as a single MI-EEG sample. Then, through the toolkit MNE-Python (MNE is an open source python toolkit mainly used for EEG/MEG analysis, processing and visualization), a MI-EEG data sample with a size of 22 × 1000 × 288 was constructed for each subject.

B. Data Preprocessing

Unlike most ways of preprocessing EEG using the Eeglab tool, all steps of preprocessing data in this paper are performed through the MNE-Python toolkit, and the preprocessing process is as follows: firstly, a sample data set with the structure [288,1000,22] is constructed for each subject according to the temporal axis of the single MI task, and to further improve the signal-to-noise ratio of MI-EEG In this paper, a 5th order Butterworth filter (i.e., δ[1-4Hz], θ[4-8Hz], α[8-13Hz], β[13-30Hz], γ[31-51Hz]) is used to filter and decompose each segment of the sample dataset into 5 new datasets, and then the datasets of these 5 bands are fused in spatial dimension and the data structure is reshaped as [288, 1000, 110], after which the spatial features are extracted and downscaled by PCA, and its downscaled data set structure is [288,1000,32].

We define Tn= (T1, T2,..., Tn) as an EEG signal sample containing 4S. According to the results of the initial filtering, the differential entropy of different frequency bands of all channels is calculated, which is transformed into a one-dimensional sequence and defined as D_m= (D_1, D_2,..., D_m). A new time-frequency characteristic sequence S_k=(S_1, S_2,..., S_k) is formed by splicing time series and differential entropy sequences. In order to make the model better learn features and solve the problem of small data volume, the data set matrix is divided into time steps through a sliding window with a step size of 60 and a window size of 510. Each part includes 510 sample sequences. Finally, a trainable set data with a structure of [2880,510,32] is formed for each subject. In order to carry out the following experiment smoothly, the EEG trainable data sets of 9 subjects were integrated, and the integrated data sets were divided into three parts according to 6:2:2, of which 60% were training sets for training models; 20% is the validation set, which is used to optimize the model parameters; The remaining 20% is the test set, which is used to evaluate the generalization and stability of the model.

C. Classical Architecture: CNN-LSTM

The classifier used by classical architectures is a serial combination of CNN and LSTM. First, the pre-extracted spatial features and time steps in the preprocessing process are used as the input of CNN-LSTM, the low-level features of the EEG data time-frequency dimension are extracted through CNN, and then the proposed features are input into LSTM to obtain a more abstract high-level Representative time-frequency features. Finally, all the features are integrated through the fully connected layer and the classification result is obtained. The overall structure of the model is shown in Figure 2.

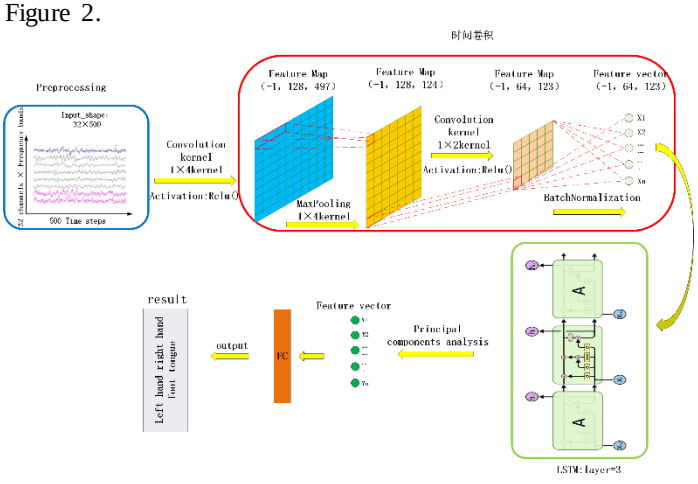

Figure 2. CNN-LSTM model structure

During data preprocessing, the data structure of each subject is [2880, 510, 32]. This means, (32, 510) as the input to the model. Among them, 32 is the spatial feature dimension after channel and frequency band fusion, and 510 sample points are temporal features. The kernel size and number of filters of the first convolutional layer are 4 and 128,

respectively. The activation function is selected as "ReLu". The maximum pool size is 4. The second convolutional layer has a kernel size of 2 and a number of filters of 64. Also select "ReLu" for the activation function. The batch normalized dimension is 64. The LSTM layers have a "tanh" activation function and a recurrent activation function with a dropout of 0.1, and the total number of units per LSTM layer is 25. The fully connected layer has a "softmax" activation function.

D. CNN-Transformer model

In the process of data preprocessing, we constructed a new time-frequency fusion feature sequence, in order to make the extracted EEG features better describe the time-frequency feature sequence characteristics of the signal, and solve the problem of long sequence characteristics of EEG signals, this paper proposes a Transformer-based one-dimensional convolutional neural network model (CNN-Transformer) for decoding MI-EEG.

First, the local temporal-frequency features are captured by down-sampling in the temporal-frequency dimension through 1D-CNN operations of convolution and pooling, and also the length of the time series is further reduced. In addition, this process serves to prevent model overfitting to some extent. The extracted short series temporal-frequency features are then fed into Transformer to further extract more abstract, high level time features. Finally, the high-level abstract features are integrated through the fully connected layer to output the classification results.

The process of extracting temporal features from Transformer is as follows: first, the output vector of 1D-CNN is used as the input feature vector X={X1,X2,...,Xn}, record the initial position information of the feature vector X by position-encoding (PE), and then use the PE feature vector as the input of the encoder. In the encoder, in order to be able to extract multi-angle temporal features in parallel as well as solve the gradient disappearance and gradient explosion problems, the value of each attention head is calculated in parallel through the multi-head attention mechanism, and the attention head values are input to “residual structure” and “layernorm” layer for processing, and then the new results are input to the feedforward neural network, and then the output results are input into the decoder. There are two inputs to decoder, one is the feature vector Z=={Z1,Z2,...,Zn} of encoder output and the other is the feature vector X1 of 1D-CNN output. In decoder, the output vector X of 1D-CNN is combined with the output vector X of 1D-CNN and the encoder output is decoded by the multi-head attention mechanism, the pre-feedback neural network and the residual structure. The CNN-Transformer model structure is shown in Figure 3.

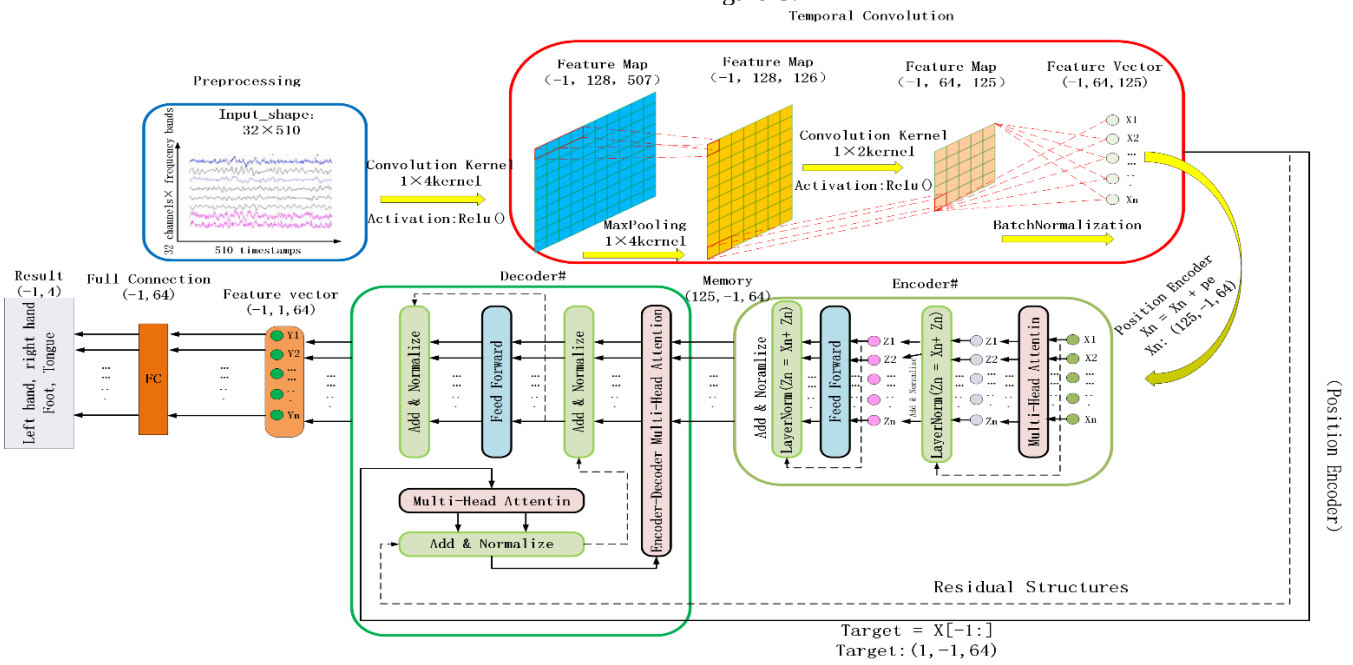

Figure 3. CNN-Transformer model structure

The overall model structure is divided into two parts. The first part is the temporal convolution, and the input shape and the parameter values involved in the whole convolution process are the same as the baseline model. The second part is the Transformer, and the parameters involved are described below. The overall structure of the Transformer consists of three parts: 1 layer of Position PE, 1 layer of encoder and 1 layer of decoder.PE layer: embedding dimension is 64; single time step is 125; dropout is 0.1; the value of pe is determined by the triangular functions sin and cos (pe: Position and order information of all feature vectors x); encoder layer: feature vector dimension is 64; multi-head attention (nhead) is 16; feedforward neural network dimension (dim_fre) is 16; dropout is 0.1; number of layers is 1. decoder layer: parameters are the same as those of encoder layer. In this paper, the decoder structure has no Mask structure. The fully connected layer has "softmax" activation function. The CNN-Transformer model is summarized in Table I.

E. Evaluation Indicators

The evaluation indicators involved in this paper include: ①Accuracy: The ratio of the number of correctly predicted categories to the total number of categories is used to measure the applicability and practicality of the model. ②Consistency (kappa): Similar to acc, it is also an indicator to measure the quality of the model. It is an indicator used for consistency

test and can also be used to measure the effect of classification. At the same time, it is also an indicator of penalizing the "bias" of the model. As shown in formula (1).

$$Kappa = \frac{Po-Pe}{1-Pe} \quad (1)$$

According to the calculation formula of kappa, the more unbalanced the confusion matrix is, the higher the Pe is, the lower the kappa value is, which is just enough to give a low score to a model with strong "bias". In the formula, Po is the recognition rate, and Pe is 0.25. ③ Confusion matrix: Confusion matrix, also known as error matrix, is a standard format for expressing accuracy evaluation and is used to observe the performance of the model in various categories. By calculating the confusion matrix composed of the recognition results of each category, it reflects the ratio of correct and wrongly divided motion MI-EEG for each category. Through the above evaluation indicators, while testing the quality of the models, the differences between the models can also be compared. ④Roc curve: Roc curve is a visual tool for evaluating classification models, which is used to describe the trade-off between classifier hit rate and false positive rate. The area value of the Roc curve and the abscissa is called Auc. Usually, the Auc value is used to compare the performance of different classifiers.

Table I. CNN-Transformer model overview

Model: CNN-Transformer		
Layer(type)	***Output Shape***	***Params***
conv1d_1	(-1, 128,507)	16512
max_pooling	(-1,128,126)	0
conv1d_2	(-1,64,125)	16448
batch_normalization	(-1,64,125)	256
encoder	(-1,64,125)	18432
decoder	(-1,64,125)	34816
fc	(-1,64)	256
Total Params: 87524 Trainable Params: 87396 Non-Trainable Params:128		

III. RESULTS AND ANALYSIS

A. *Experimental result and model parameter settings*

All experiments involved in this paper are carried out on the cloud server platform, and the GPU version is RTX3090. The results of model 5-fold cross-validation training are shown in Table II.

After 5-fold cross-validation, the averages of validation set accuracy and kappa were 0.9929 (±0.0007) and 0.9843 (±0.0021), respectively. After that, the CNN-Transformer model was tested with the test set data, and the average accuracy and kappa were as high as 99.68% and 98.67%, respectively. But the disadvantage of using cross-validation here is that the dataset for training and optimizing the model parameters and the dataset for testing the model are pre-partitioned. The accuracy rates of CNN-Transformer and the classic architecture training set and validation set are shown in Figures 4 and 5. The loss values are shown in Figures 6 and 7.

Table II. 5-fold cross-validation result

n-fold	***Acc(train)***	***Loss(train)***	***acc(validation)***	***loss(validation)***	***Kappa(validation)***
1	0.9998	0.0057	0.9937	0.2719	0.9822
2	0.9997	0.0068	0.9923	0.3443	0.9846
3	0.9996	0.0078	0.9914	0.3243	0.9827
4	0.9996	0.0082	0.9932	0.2819	0.9835
5	0.9998	0.0075	0.9942	0.2591	0.9887
average	0.9997(±0.0001)		0.9929(±0.0007)		0.9843(±0.0021)

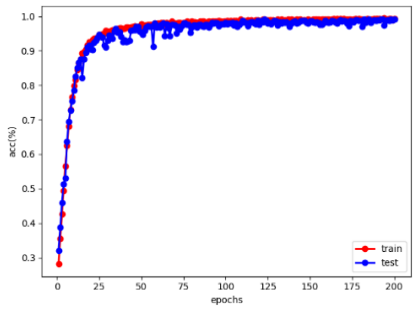

Figure 4. CNN-Transformer accuracy

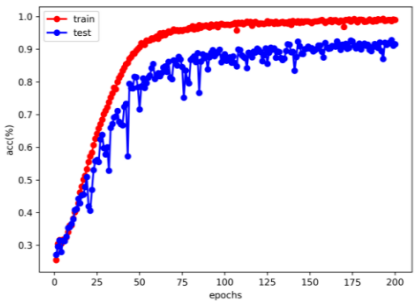

Figure 5. CNN-LSTM accuracy

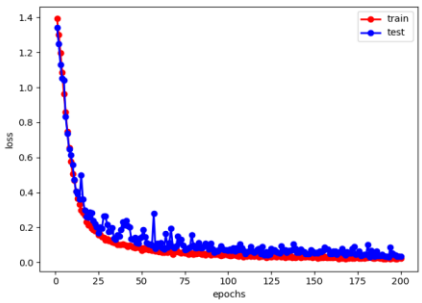

Figure 6. CNN-Transformer loss

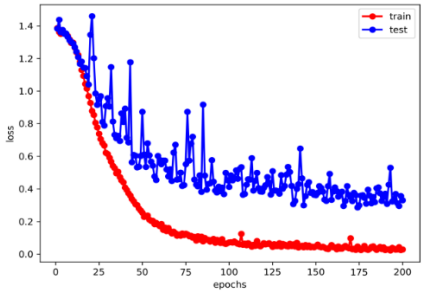

Figure 7. CNN-LSTM loss

As can be seen from the figure, when the CNN-Transformer model is trained for about 25 epochs, the model achieves a good effect, and the whole process has a high degree of fit and is more stable. The parameters of the CNN-Transformer model are optimized and adjusted through 5-fold cross-validation, and the parameters are shown in Table III.

Table III. CNN-Transformer parameters and values

Parameter	***Type or value***
Regularization	Dropout=0.1
Loss_function	Cross_entrop
Batch_size	128
Epoch	200
Learning_rate	0.0005
Nhead	16
Dim_fre	16

B. Comparative analysis of the result of two models

The accuracy and kappa values are as described in the previous section. After comparing acc and loss, we can see that CNN-Transformer can capture the features of EEG data faster and more accurately, and the efficiency and effect are better than the classic architecture. In order to further analyze the influence of the proposed method on MI-EEG recognition effect, this paper calculates the average confusion matrix of the two models. As shown in Figures 8 and 9.

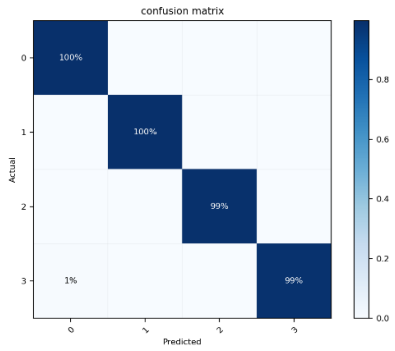

Figure 8. CNN-Transformer average confusion matrix

The horizontal axis is the predicted category, the vertical axis is the actual category (0, 1, 2, and 3 represent left-hand, right-hand footsteps, and tongue, respectively), and the diagonal line formed by the intersection of the horizontal axis and the vertical axis category is correctly divided , while the remaining intersections are the rate of mispredictions. It can be seen from Figure 9 that the average error rates at the intersection of the left and right hands are 5% and 6%, respectively. The average error rate at the junction of the foot and tongue was 7% and 3%, respectively. This could be the same spatial or temporal information when imagining different classes of tasks, or it could be that the features captured by the baseline model are inaccurate. As shown in Figure 8, Almost 0 error rate for all categories. Blank indicates 0 error rate. It further proves the effectiveness and stability of the CNN-Transformer model proposed in this paper.

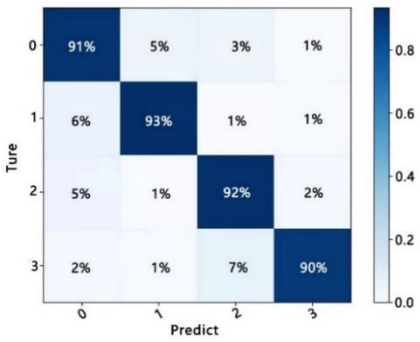

Figure 9. CNN-LSTM average confusion matrix

In order to more intuitively reflect the classification performance of CNN-Transformer, the roc curves of all categories of the two models are drawn at the same time, as shown in Figures 10 and 11. After comparison, it is found that both CNN-LSTM and CNN-Transformer have very good classification performance. However, for class1, the Auc value of CNN-Transformer is 0.5 higher than that of CNN-LSTM, and the other classes are 0.4 higher. Therefore, the classification performance of CNN-Transformer is better.

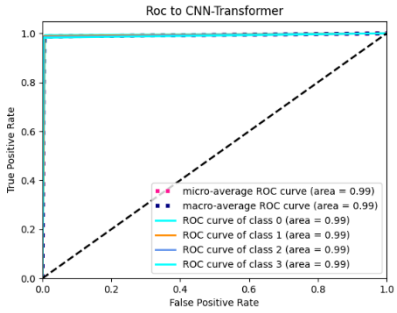

Figure 10. Roc curve of CNN-Transformer

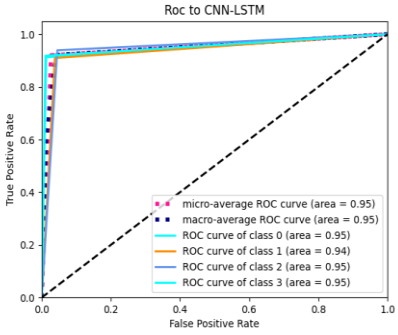

Figure 11. Roc curve of CNN-LSTM

Finally, this paper also summarizes other models that use the competition data set 2a as the experimental data, and compares them with the model in this paper. The comparison results are shown in Table IV.

Table IV. Comparison results with other models

Model	*accuracy*
CNN-Transformer	0.9929
CNN-LSTM	0.9231
TSCNN[12]	0.8009
PSD+CNN[16]	0.8797
WT+_HLMSFFCNN[17]	0.9295
TMCNN[18]	0.8085
CSP+CWT+CNN[19]	0.7790

IV. DISCUSSION AND CONCLUSIONS

The performance of the CNN transformer model proposed in this paper is evaluated by 5-fold cross validation and confusion matrix. The results show that the model has achieved good results. CNN transformer model mainly analyzes the data in the time dimension and extracts high-level features. Due to transformer, the model has good parallel computing ability and generalization ability of the model, and shows excellent results in both effect and efficiency. The robustness to different tasks can be improved by appropriate filtering and initial weight. In order to further prove the effectiveness of the model, compared with other models in the literature, it is better than other models on the premise of the same data. Because the model combines the advantages of CNN and transformer respectively, and also makes up for the respective defects of CNN, RNN and transformer. To sum up, the model CNN transformer proposed in this paper has good generalization ability and high practicability, provides a design idea for improving the accuracy of decoding MI-EEG, and lays a technical foundation for the implementation of MI-BCI. However, this study also has shortcomings and needs to be improved:

(1) KPCA is an improved version of PCA. For the nonlinear problems existing in the real world, principal component analysis PCA and linear discriminant analysis LDA are powerless, while KPCA has higher analysis and extraction ability for nonlinear data. However, due to the working principle of KPCA, although it can better reduce the dimension of nonlinear data, the cost is more it resource consumption. Therefore, due to the experimental equipment, PCA is still used to extract the principal components of the features of spatial dimensions.

(2) Transformer is independent of CNN and RNN. Although it has good feature extraction ability and parallel computing ability, it needs enough data to reflect its advantages. Experiments show that with the increase of encoder decoder structure, the classification effect of CNN transformer model is getting worse and worse. Therefore, for the amount of data in this paper, we only use the transformer with 1-layer encoder decoder structure, which does not give full play to the stackable advantage of transformer.

(3) According to the characteristics of MI-EEG, in the next step, we can try to build a fusion model of two-dimensional convolutional neural network and multi-layer transformer for time, space and frequency domain, so as to realize the joint learning of multi-source features and improve the recognition accuracy of MI multi classification tasks. At the same time, the visual structure of the output features of each model is studied to provide a practical basis for better explaining the feature structure of each field.

Identify applicable funding agency here. If none, delete this text box.

REFERENCES

[1] Wolpaw J R, Birbaumer N, Heetderks W J, et al. Brain-computer interface technology: areview of the first international meeting[C]. IEEE Transactionson Rehabilitation Engineering, 2000, 8(2):164.

[2] López-Larraz E, Trincado-Alonso F, Rajasekaran V, et al. Control of an ambulatory exoskeleton with a brain-machine interface for spinal cord injury gait rehabilitation. Front Neurosci, 2016, 10: 359.

[3] Meng J, Zhang S, Bekyo A, et al. Noninvasive electroence-phalogram based control of a robotic arm for reach and grasp tasks. Sci Rep, 2016, 6: 38565.

[4] Chaudhary U, Birbaumer N, Ramos-Murguialday A. Brain–computer interfaces for communication and rehabilitation. Nat Rev Neurol, 2016, 12(9): 513.

[5] Zhou Y, Wang T, Feng Huanqing et al. ERD/ERS analysis of motor imagery EEG[J]. Beijing Biomedical Engineering, 2004.

[6] M. T. Sadiq et al., "Motor Imagery EEG Signals Decoding by Multivariate Empirical Wavelet Transform-Based Framework for Robust Brain Computer Interfaces," in IEEE Access, vol. 7, pp. 171431-171451, 2019.

[7] NiteshSingh Malan，Shiru Sharma., "Time window and frequency band optimization using regularized neighbourhood component analysis for Multi-View Motor Imagery EEG classification," in Biomedical Signal Processing and Control, 2021.

[8] P. Kant, J. Hazarika and S. H. Laskar, "Wavelet transform based approach for EEG feature selection of motor imagery data for braincomputer interfaces," 2019 Third International Conference on Inventive Systems and Control (ICISC), 2019, pp. 101-105, doi: 10.1109/ICISC44355.2019.9036445.

[9] Poonam Chaudhary, Rashmi Agrawal, "Non-dyadic wavelet decomposition for sensory motor imagery EEG classification", Brain Computer Interfaces, 2020, pp. 11-21.

[10] Yimin Hou, Tao Chen, Xiangmin Lun, Fang Wang,A novel method for classification of multi-class motor imagery tasks based on feature fusion, Neuroscience Research, Volume 176,2022,Pages 40-48, ISSN 0168-0102.

[11] Hongli Li, Man Ding, Ronghua Zhang, Chunbo Xiu, Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network, Biomedical Signal Processing and Control, Volume 72, Part A, 2022, 103342, ISSN 1746-8094.

[12] Chu Yazi, Zhu Bo et al. A spatio-temporal feature learning convolutional neural network-based EEG decoding method for motor imagery[J]. Journal of Biomedical Engineering, 2021.

[13] Jinzhen Liu， Fangfang Ye, "Multi-class motor imagery EEG classification method with high accuracy and low individual differences based on hybrid neural network," Journal of Neural Engineering, 2021.

[14] Mouad Riyad, Mohammed Khalil, Abdellah Adib, MI-EEGNET: A novel convolutional neural network for motor imagery classification, Journal of Neuroscience Methods, Volume 353, 2021, 109037, ISSN 0165-0270.

[15] Ashish Vaswani, Noam Shazeer, "Attention is All you need", in Part of Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017.

[16] Pérez-Zapata A F, Cardona-Escobar A F, Jaramillo-Garzón J A, et al. Deep Convolutional Neural Networks and power spectral density features for Motor Imagery classification of EEG Signals// 2018 International Conference on Augmented Cognition. Cham: Springer, 2018: 158-169.

[17] Li, Ma., Han, Jf. & Yang, Jf. Automatic feature extraction and fusion recognition of motor imagery EEG using multilevel multiscale CNN. Med Biol Eng Comput 59, 2037–2050 (2021).

[18] Q. Zhou, P. F. Tian. A motion imagery EEG identification algorithm based on migration learning multilevel fusion [J/OL]. Journal of Electronic Measurement and Instrumentation.

[19] Prabhakar Agarwal, Sandeep Kumar, Electroencephalography based imagined alphabets classification using spatial and time-domain features, International Journal of Imaging Systems and Technology, 10.1002/ima.22655, 32, 1, (111-122), (2021).