A Multi-Scale Fusion Convolutional Neural Network Based on Attention Mechanism for the Visualization Analysis of EEG Signals Decoding

Donglin Li, Jiacan Xu, Jianhui Wang, Xiaoke Fang, and Ying Ji

***Abstract*— Brain-computer interface (BCI) based on motor imagery (MI) electroencephalogram (EEG) decoding helps motor-disabled patients to communicate with external devices directly, which can achieve the purpose of human-computer interaction and assisted living. MI EEG decoding has a core problem which is extracting as many multiple types of features as possible from the multi-channel time series of EEG to understand brain activity accurately. Recently, deep learning technology has been widely used in EEG decoding. However, the variability of the simple network framework is insufficient to satisfy the complex task of EEG decoding. A multi-scale fusion convolutional neural network based on the attention mechanism (MS-AMF) is proposed in this paper. The network extracts spatio temporal multi-scale features from multi-brain regions representation signals and is supplemented by a dense fusion strategy to retain the maximum information flow. The attention mechanism we added to the network has improved the sensitivity of the network. The experimental results show that the network has a better classification effect compared with the baseline method in the BCI Competition IV-2a dataset. We conducted visualization analysis in multiple parts of the network, and the results show that the attention mechanism is also convenient for analyzing the underlying information flow of EEG decoding, which verifies the effectiveness of the MS-AMF method.**

***Index Terms*— BCI, MI, EEG, CNN, multi-brain regions, spatio temporal multi-scale features, dense fusion strategy, attention mechanism, visualization analysis.**

I. Introduction

BRAIN-COMPUTER interface promotes direct communication between the human body and external devices by analyzing the activity of neuronal electrical signals in the brain [1] to achieve the purpose of understanding human intentions. With the development of technology, BCI technology has gradually played an essential role in human-computer interaction [2] and assisted living [3]. Non-invasive BCI based on electroencephalogram compared with other non-invasive methods such as fMRI, MEG, PET, which have many characteristics, looks like portability, easy operation, and low cost, etc. BCI detects and analyzes different brain activity patterns in EEG signals [4] to understand the user's intentions for communication. Among them, motor imagery is a pattern widely used in brain-computer interface system. When the subjects imagine moving a certain part of the body, the neurons in the sensorimotor cortex area of the brain appear some oscillating activities corresponding to the specific imagination [5]. Machine learning technology and deep learning technology have been used to decode this oscillating activity [6], [7]. People can control external devices by decoding motor imagery EEG signals, such as autonomous driving [8], motor rehabilitation [9], and helping healthy people control devices to perform delicate tasks [10]. Therefore, improving the method of decoding EEG signals is very important for the future development of motor imagery BCI.

The EEG signals are essentially a multi-channel non-linear time series. BCI decoding neuron activity has a core problem which is extracting as many recognizable features as possible from the EEG time series to understand brain activity accurately [11], [12], and how to extract and select the appropriate features determines the decoding performance of BCI. Appropriate denoising processing of EEG signals is beneficial to decoding, multivariate noise normalization [13] can make use of the information contained in the multi-channel signal data. Initially, BCI based on ERD / ERS [14] can classify EEG signals by recognizing the energy changes of brain signals during motor imagery [15]. Later, to solve the individual difference problem of the ERD/ERS [16], many modern machine learning methods were introduced and widely applied to MI-BCI [17], [18]. One of the commonly used methods for extracting features is [19] common spatial pattern (CSP) algorithm [20], which attempts to find the optimal spatial filter to enhance the sensitivity of the difference between the two types of EEG signals. To solve complex and diverse problems, researchers proposed a series of variant methods based on CSP. e.g., the filter bank common spatial pattern (FBCSP) [21] and discriminant filter bank public space model (DFBCSP) [22] have solved the problem that CSP is highly dependent on the frequency band. Regulative CSP (RCSP) [23] and common

Manuscript received May 14, 2020; revised July 25, 2020 and October 14, 2020; accepted November 5, 2020. Date of publication November 11, 2020; date of current version January 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61733003. *(Corresponding author: Donglin Li.)*

The authors are with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: 598855925@qq.com).

Digital Object Identifier 10.1109/TNSRE.2020.3037326

spatio-spectral patterns (CSSP) [24] have solved the problem of regularization. Although the methods based on CSP show good performance in the MI-BCI classification field, most of them only considered part of the information in the time domain, space, frequency and static energy of the EEG signals, while ignoring the dynamic nature of signals in the MI process. Therefore, valuable MI information will be lost in this process. If these features are combined with the CSP framework, the MI-BCI framework can still improve the classification performance.

The combination of the machine learning method and the above algorithm has been widely used in the field of EEG, but its performance and accuracy are not enough. As an important branch of machine learning, deep learning technology has been used for analyzing EEG signals and attracted much attention in recent years. Among the researches on decoding MI EEG signals, convolutional neural network (CNN) is the most widely used in deep learning models, which can extract features of higher discrimination and robustness [25], while other models have the same outstanding effect on EEG signals of nonlinear time series. e.g., Sakhavi *et al.* [26] designed different types of convolutional neural network (CNN) structures such as temporal convolution, channel convolution and two-dimensional convolution to classify signal. Tabar *et al.* [27] proposed a deep network that used SAE to classify the features extracted from CNN network. Although the features extracted by the deep learning method are better than the manual extraction, there are still many complex problems in the process of decoding the MI EEG signals using the deep learning model, However, the simple network [28] just makes changes in the field of data input and receptive field, which cannot satisfy the requirement of the complex EEG signals. Therefore, this framework has a weak variability to solve the problem.

EEG signal is essentially a complex data object due to the characteristic of non-linear, dynamic and multi-channel correlation. Many complex network frameworks were proposed to adapt to the multi-faceted information of EEG signals, which considered the characteristics of EEG signals from multiple angles. To overcome the limitations of multiple channels, Molinari [29] *et al.* proposed a novel cascaded recurrent neural network (RNN) architecture based on long term memory (LSTM) blocks to analyze EEG signals. Wang *et al.* [30] proposed an AX-LSTM network combined with the channel sizing technique, which considered the time-varying characteristics of the network and the problem of network overfitting. To ensure the integrity of features among the different convolutional layers in the network model, Li *et al.* [31] proposed a feature fusion network with dense connections. Li *et al.* [32] proposed a channel-projection mixed-scale convolutional neural network (CP-MixedNet) to capture the multi-scale temporal feature of EEG signals. Therefore, to combine the diversity of EEG signals characteristics, the analysis and research of complex network frameworks are also necessary.

Before applying the deep learning method in the MI EEG classification task, the EEG data should be processed and expressed in a form that satisfied the needs. The EEG reflects the electrical signals activity of the brain, and the communication between each neuron arises ongoing brain cognitive or motor tasks. The classification task of MI EEG signals aims to analysis the movement-related brain regions signals, the signals of different brain regions also correspond to different functions. The motor cortex [33], [34] is responsible for planning and controlling movement. However, during the transmission of electrical signals from neurons through brain tissues, the mixed activity of neurons due to the volume conduction effect was recorded by electrodes [35], [36], each electrode directly affects each other, the EEG signals collected are not only corresponding to the electrical signals in the functional brain region.

Different methods of EEG signals representation have been put forward to solve the above problems. Short-time Fourier transform (STFT) [29] method and wavelet transform [37] etc. were used to convert EEG time series into 2D images. To consider the spatial feature of multi-channel signals, Bashivan *et al.* [38] transform EEG activities into a sequence of topology preserving multi-spectral images. These methods ignore the relationship between different EEG regions and may lose the spatial information of EEG signals [39]. However, each node reflects different brain functions and influences each other, so the relationship between different brain regions is essential for EEG analysis.

In this paper, we propose a multi-scale fusion convolutional neural network based on the attention mechanism for the visualization analysis of EEG signals decoding. The modular structure of the network increases the variability of the network, makes the model adapt to complex EEG signals, and preserves the integrity of EEG features to the greatest extent. In terms of data representation, we propose a method for EEG representing based on multiple brain regions, which not only preserves the temporal features of EEG but also increases the spatial resolution. MS-AMF network is specifically composed of the spatial multi-scale module, attention module, temporal multi-scale module and dense fusion module. The temporal and spatial multi-scale modules are responsible for extracting the features within and between brain regions. We embedded temporal multi-scale module in the dense fusion module, such a combined module considered the dynamic temporal features of EEG signals, and the dense connection of fusion features also ensures the maximum information flow among the various layers of the network to avoid losing important information. The attention module improves the expression ability of the network model by accurately modelling the functional relationship between the feature-map of each channel in the convolution layer, enabling the network to selectively amplify valuable feature channels and suppress useless feature channels based on global information. Experimental results illustrate that this method has a good effect on the dataset of BCI Competition IV-2a. We conducted a visualization analysis in multiple parts of the network, and the attention mechanism is also convenient for analyzing the underlying information flow of EEG decoding.

The rest of this paper is organized as follows: Section II introduces the data representation and model of MS-AMF. Experiments and results are illustrated in Section III. In Section IV, the role of different modules in the network

is discussed through visualization analysis methods from different views; the paper is finally concluded in Section V.

II. Method

In the following section, we will first describe the EEG representation based on multiple brain regions divide. Then, we introduce the basic modules of MS-AMF, including spatial multi-scale module, attention module (AM), temporal multi-scale module (TMSM) and dense fusion module. Finally, we introduce the parameters and experimental process of the network framework.

A. EEG Representation Based on Multiple-Brain Regions Divide

To increase the ability of learning local details in the brain region, we proposed EEG representation based on multi-brain region: firstly, we divided the original EEG data into multi-brain regions, and then extracted the spatial domain information from the divided signals, which makes the receptive field of the neuron cover a designated brain region, and the extracted features include the local details of the brain region. We provide a new input method for subsequent networks.

1) Multi-Brain Regions Division: In the most BCI experiments, the electrodes used to collect EEG signals were placed following the international 10-20 system, as shown in Fig. 1 The electrode placement area can be divided into five brain regions, e.g. F(frontal), T(temporal), C(central), P(parietal) and O(occipital). Existing paper [33] points out that the central region and the parietal region are related to motor function.

In the field of EEG signals recognition, with the addition of deep learning methods, EEG representation is usually used to extract primary surface information. However, in most studies, two-dimensional signals tensors of the whole brain are represented as input and put into the deep learning framework to reduce the dimension to one-dimensional tensors. Predicting reaching targets from Human EEG [40], different brain regions work together in motor imagery and they are activated differently. Therefore, the direct combination method rarely considers the correlation between the features in the brain regions and between the electrodes in the brain regions, we proposed the representation based on multi-brain regions divide, which can extract detailed information. We achieve the partition input method by replacing one convolution kernel with multiple convolution kernels, which can add nonlinear activation layers to increase the discriminative ability. As shown in Fig. 1(b), in the division of the regions we selected the original EEG data of the central region and the parietal regions to represent as $C_i, P_j, (i = 1, 2, \ldots 13; j = 1, 2, \ldots 9)$. Due to the lack of frontal and occipital electrodes in the data set, we divided them into adjacent brain regions, defined as C and P. As shown in Fig. 1(a), region C mainly contains information of the central region, and region P mainly contains information of the parietal region. This method not only preserves the temporal characteristics of EEG signals but also preserves the spatial characteristics of brain regions to some extent. In this way, the influence of electrode volume conduction can be alleviated and the spatial resolution can be improved, which provides a novel angle for the follow-up analysis of EEG signals in deep learning analysis.

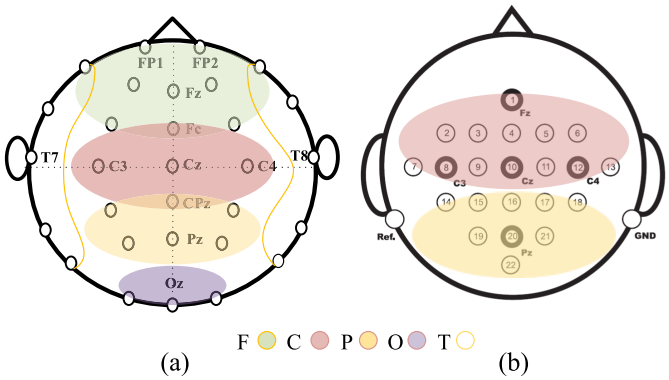

Fig. 1. The distribution of the brain regions divided by the positions of the electrodes. (a) represents the traditional way of dividing brain regions, (b) represents the EEG representation method based on brain regions division.

2) EEG Representation: At present, the combination of the CSP algorithm and one-versus-rest strategy [20] can well implement the multi-classification of EEG signals. CSP algorithm is a spatial feature extraction algorithm, which can effectively distinguish two kinds of EEG data by maximizing the variance of one kind and minimizing the variance of another. The mathematical expression of CSP optimization function is as follows:

$$J(w) = \frac{w^T X^T X w}{w^T Y^T Y w} = \frac{w^T S_h w}{w^T S_f w} \tag{1}$$

where w is the spatial filter learned by the eigenvectors extracted. S_i is the covariance matrix of the two categories, we use CSP algorithm to extract a coefficient vector of a spatial filter W_{csp}, and chose the first and last $N = 2$ rows of CSP projection matrix W_{csp}. However, the CSP algorithm is primarily used to distinguish two categories of data, so in the case of multiple categories, the one-versus-rest strategy needs to be chosen. The multi-EEG signals are projected as $Z_C = W_c^T C, Z_P = W_p^T P$.

The EEG representation based on multi-brain regions was only obtained the primary spatial features through shallow learning after separating the brain regions, without extracting the temporal features. We put two sets of spatial feature matrices into the MS-AMF model in parallel, which can implement in-depth extraction of spatial and temporal features of EEG signals.

B. A Multi-Scale Fusion Network With Attention-Module

In this section, we propose a multi-scale feature fusion CNN (MS-AMF) based on attention mechanism to implement the decoding of MI multi-channel EEG signals, which includes spatial multi-scale module, attention module, and dense fusion module. Fig. 2 is a sample framework for MS-AMF CNN presented in this article. In the following sections, we will describe the principles and configuration of each module in detail.

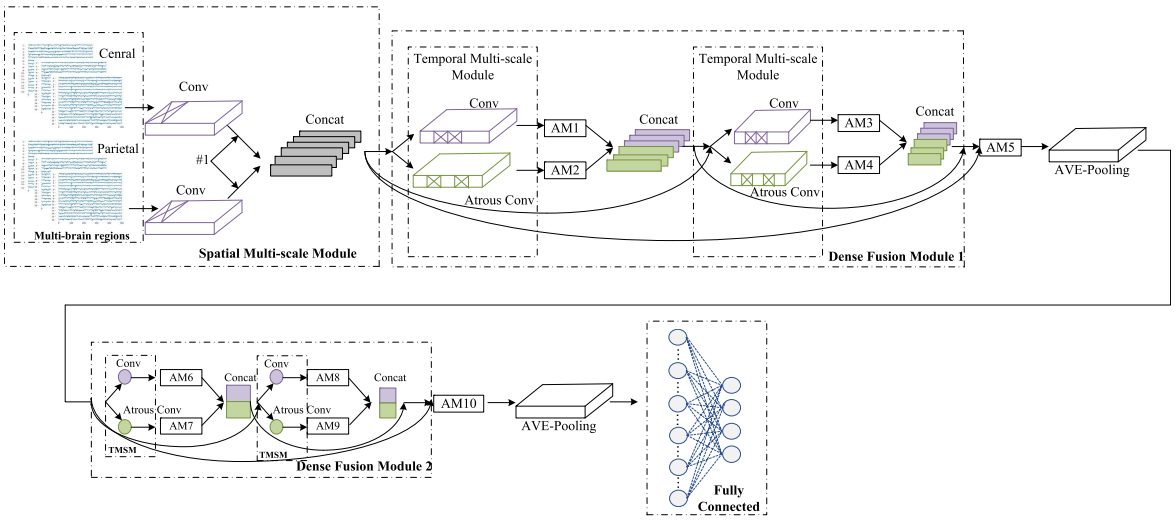

Fig. 2. A schematic illustration of the MS-AMF architecture, where the purple cuboid represents the normal convolution kernel, the green cuboid represents the Atrous convolution kernel, and the rectangles corresponding to the colors are stacked together to represent the feature map. (The #1 flag in the Spatial Multi-scale Module is introduced in the discussion section).

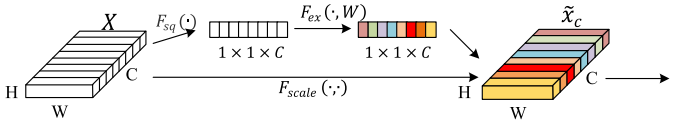

Fig. 3. The specific composition structure of attention module, where different colors indicate different sensitivity.

1) Spatial Multi-Scale Module: The spatial multi-scale module is a simple multi-input convolution structure, which can extract information within and between brain regions. To reduce the loss of spatial information when the dimension of EEG signals channel drops to 1, we adopt the multi-input mode and put the represented data of Z_C, Z_P in the parallel convolution module at the same time. Finally, the corresponding features of the brain regions are F_C, F_P, respectively. We spliced the third dimension of the features extracted from the two brain regions $X_0 = Concat(F_C, F_P)$, and extract the features between them. As shown in Fig. 2, the feature maps in different brain regions are fused and put into a dense connection module, which avoids the loss of information caused by feature extraction in advance or the conversion of EEG signals into images.

2) Attention Module: The attention module is composed of Squeeze-and-Excitation [49], which can improve the expression ability of the framework and increase the sensitivity of the model to the information characteristics. To meet this expectation, we added intra-block AM and inter-block AM into the dense fusion module, enabling the network to selectively amplify valuable feature channels and suppress useless feature channels based on global information. We define the algorithm of attention module as AM $(\cdot)$, and the structure is shown in Fig. 3. We introduce its composition as follow:

a) Squeeze module: Assume that $X \in \mathbb{R}^{H \times W \times C}$ is the feature-map output extracted from the previous layer of the module. The formula for the Squeeze module is as follows:

$$z_c = F_{sq}(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i, j) \qquad (2)$$

where $X_c \in \mathbb{R}^{H \times W}, c \in [1, 2, \ldots\ldots, C]$, represents the feature-map of the c channel in X, $x_c(i, j)$ represents the data point of the position (i, j) in X_c. $F_{sq}(\cdot)$ represents feature compression along spatial dimensions, which compresses each two-dimensional feature X_c into a real number z_c, which has a global perceptive field to some extent, and the dimension output matches the number of feature channels input.

b) Excitation module: We use Squeeze operation to aggregate information, and excitation each channel activator factor S, the formula is as follows:

$$S = F_{ex}(Z, W) = \sigma(g(Z, W)) = \sigma(W_2 \delta(W_1 Z)) \qquad (3)$$

where δ represents the sigmoid function operation, and two sets of parameters, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, are learned to explicitly model the correlation between channels. To limit model complexity and aid generalization, the module forming a bottleneck with two fully-connected (FC) layers to control the parameters. Between the FC, there is a dimensionality-reduction layer with a reduction ratio r, a Sigmoid, and then a dimensionality- increasing layer returning to the channel dimension of the conversion input X. Finally, the output of the AM module is obtained by adjusting X by activator factor S. The formula is as follows:

$$\tilde{x}_c = F_{scale}(x_c, s_c) = s_c x_c \qquad (4)$$

where $\tilde{X} = \left[\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_c\right]$ and $F_{scale}(x_c, s_c)$ refers to channel-wise multiplication between the scalar s_c and the feature map $x_c \in \mathbb{R}^{H \times W}$.

3) Dense Fusion Module: The network MS-AMF proposed in this paper retains the fusion function involved in the Densenets architecture [41] we proposed earlier, which is a simple connection mode. The network directly connects the output characteristic graph of each layer to ensure the

maximum information flow between each layer. Each layer of the network receives additional input from the previous layer and passes its feature-map to the latter layer, so that the network maintains the feed-forward characteristics. The process of feature fusion is to splice the feature-map according to a certain dimension. By embedding the parallel multi-scale temporal feature extraction method into the dense model, the network can implement considerable accuracy.

As shown in Fig. 2, our proposed MS-AMF framework contains two dense fusion modules. The dense fusion model has a total of 2 layers, in which there is a composite algorithm of nonlinear transformation between each layer, defined as $H(\cdot)$: $BN - ReLU - [Conv(\cdot) + AM(\cdot)] \&[Atrous\ Conv(\cdot) + AM(\cdot)] - Concat(\cdot) - AM(\cdot)$, $Concat(\cdot)$ indicates that the features of different time scales are fused and stitched. $Atrous\ Conv(\cdot)$ convolution layer is to inject holes in the standard convolution map, in this way, the reception field is increased and the temporal signal size is reduced so that each convolution output contains a large range of information. At the same time, the network would effectively pay attention to the EEG information in the topical temporal and not allow each temporal signal interact with the surrounding signal blocks to affect the independence of time domain information. The parallel operations of normal convolution and Atrous convolution constitute the temporal multi-scale module and extract the temporal feature from the receptive fields of different sizes.

TABLE I
STRUCTURE OF MS-AM FUSION NETWORK MODEL

layers	output	MS-AM Fusion Network	
Multi-Brain Region	1*500 @512	F_C: $conv(16,1),(16,1),256$	F_P: $conv(16,1),(16,1),256$
Dense Block1	1*500 @768	$conv(1,25),(1,1),128$	$Atrous\ conv(1,3),(1,1),2{,}128$
		$AM_1(\cdot)$	$AM_2(\cdot)$
		Concat$(\cdot)$	
	1*500 @1024	$conv(1,25),(1,1),128$	$Atrous\ conv(1,3),(1,1),2{,}128$
		$AM_3(\cdot)$	$AM_4(\cdot)$
		Concat$(\cdot)$	
AM Module		$AM_5(\cdot)$	
Ave-Pooling	1*20 @1024	$pooling(1*25),(1*25)$	
Dense Block2	1*20 @1280	$conv(1,3),(1,1),128$	$Atrous\ conv(1,3),(1,1),2{,}128$
		$AM_6(\cdot)$	$AM_7(\cdot)$
		Concat$(\cdot)$	
	1*20 @1536	$conv(1,3),(1,1),128$	$Atrous\ conv(1,3),(1,1),2{,}128$
		$AM_8(\cdot)$	$AM_9(\cdot)$
		Concat$(\cdot)$	
AM Module		$AM_{10}(\cdot)$	
Ave-Pooling	1*4 @1536	$pooling(1*5),(1*5)$	
Fully Connected	4	4, *softmax*	

C. Parameter Selection of MS-AMF Network

Although individual differences will affect the optimal parameters of the network, to prove that the network is universal, we fixed all the important network framework parameters of the subjects. Firstly, we referred to the main parameters of the dense connection module in the previous study [31], and then we adjusted the parameters of the spatial scale module and the attention module under a relatively stable framework. Due to the limitation of time and calculation, we did not choose a complex optimization algorithm for parameter optimization and finally used the cross-validation method for parameter selection.

The specific parameters of each part of MS-AMF network are shown in Table I, where the format of parameters is $conv\,[size, stride, num]$, $atrous[size, stride, ratio, num]$, $pooling\,[size, stride]$, the RELU function was selected as the activation function of the convolutional layer and the fully connected layer. The size and number of convolution kernels were selected according to the features of the input data in the multi-brain regions and the requirement that the convolution output is an integer. In dense block1, we chose the filter size of the ordinary convolution to be 1×25; this is because the MI experiment EEG signals sampling frequency is 250hz, and the non-linear EEG data select 0.1s temporal signal will enhance the receiving domain decoding effect. Too much receptive field lead to inaccurate morphological learning, while too few receptive field leads to too much local information learning. Therefore, in the dense fusion module, we combined temporal multi-scale convolution with normal convolution in parallel, which reduces the unicity of a single convolution scale in the ordinary frame. We had installed AM modules within and between the dense fusion modules, which can first help normalize the output of the network, then calibrate the feature map to improve the expression of the network and finally facilitate subsequent visualization analysis to explore the learning process of the network.

The training configuration of the network as follows:

1) Adam algorithm was selected as the optimizer algorithm of the network, and the parameter was set as the initial value.

2) the loss function algorithm chose the cross-entropy.

3) The learning rate was set as 10e-5.

4) In addition, we insert a batch normalization layer between the activation layer and the convolution layer.

For experiments, the data preprocessing and signal representation that we carried out in Matlab 2017b environment, and we used the 16GB RAM CPU of Intel(R) Core i7-7700hq 2.80Ghz. For deep learning, we used the GeForce GTX 1080 GPU with 8GB of RAM; the MS-AMF network was implemented using the TensorFlow deep learning framework.

III. EXPERIMENTS AND RESULTS

A. Dataset

In this paper, we used the 2008 BCI competition IV-2a EEG data set, which had been widely used in the study of motor imagery for comparative analysis. The dataset consists of four different motor image tasks, including thinking to move hands, feet, and tongue, and the data set labels are divided into the left hand (class 1), right hand (class 2), feet (class 3), and tongue (class 4). Motor image data set recorded from 22 Ag/AgCl electrodes with a 250-Hz sampling rate and band-pass filtered between 0.5 and 100 Hz from 9 subjects. The timing scheme

consists of a fixed 2 seconds, a reminder time of 1.25 seconds, followed by a period of a MI of 4 s. The data set consists of two sessions of data, training data and testing data. Each session has 288 trials for each of the training and evaluation data (72 trials per class).

Data preprocessing refers to our previous research. [31] To restore the original appearance of data as much as possible and ensure that it will not be excessively interfered with by noise, null value and other problems, the data preprocessing in this paper are as follows.

1) we used half a period of MI (0∼2 s). According to [37], we eliminated the data from wrong experiments. For the null value problem, we chose linear interpolation to fill in the missing value.

2) filter bank of bandpass filter (8-30hz) was used to filter the EEG signals of all recorded channels.

B. Experimental Results and Comparison of Baseline

We preprocessed the original BCI data and constructed the input data set by multi-brain region signal characterization. Finally, the MS-AMF network is used to train and test the input data set. We selected several mainstream algorithms in the field of MI EEG signals recognition as the baseline as follow:

1) FBCSP, NCSP and other EEG feature extraction algorithms [21], [42], [26] have a good decoding effect. However, they need a lot of computing power to find a suitable filter acting on the original EEG signals, to make the variance of multi-type data bigger and easier to classify. This kind of method only changes the difference between the data but could not reflect the underlying information of the data characteristics.

2) In the field of EEG signals recognition, the typical deep learning algorithm [26] and [43] generally only adopts the common CNN architecture, which can only roughly consider the spatial and temporal characteristics of EEG signals, and the extracted features are relatively weak. Although the effect is slightly improved compared with the traditional method, the variability of the network framework does not meet the requirements of extracting the characteristics from complex EEG signals.

3) In the field of EEG signals recognition, the complex deep learning network structure [30], [44], [45] generally adopts the methods of RNN, complex CNN, the combination of CNN and RNN and the change of network input to realize the flexible change of network structure to solve specific problems. This algorithm further improves the classification effect of EEG on the original basis.

It can be seen from the Table II that the average classification accuracies of the nine subjects with the MS-AMF algorithm mentioned in this paper are only slightly lower than that of the algorithm [43], with a difference of 0.13%. Moreover, MS-AMF algorithm has obtained the optimal classification rate on S1, S3, S8, and S9. For the subjects S5 and S6, the classification results obtained by the MS-AMF algorithm are significantly different from the optimal method [43], respectively, 21.73% and 7%, which may be due to the phenomenon of overfitting as the network depth and parameters increase. Compared with [43], MS-AMF algorithm uses a more concise preprocessing strategy. Compared with [45] in the complex network framework, the multi-scale temporal processing method is also adopted, MS-AMF network fuses the multi-stage temporal features and obtains remarkable results. The algorithm in Baseline [26], [33] and [39] has significantly adjusted the corresponding network parameters for each subject, so the classification effect of these algorithms are better than other algorithms in baseline. The algorithm proposed in this paper does not change the main network parameters for individual differences, but only adjusts the depth of the network according to the amount of dataset, the convergence degree of loss in the training process and the decoding accuracy in the testing process. We change the depth of the network by increasing or decreasing the number of the TMSM in the dense block (num = 2, 3) and finally get the optimal decoding accuracy.

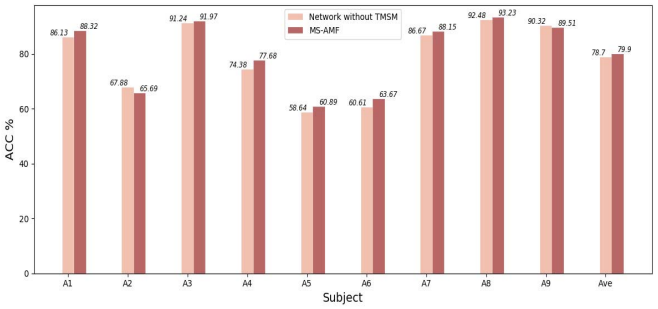

Fig. 4. The performance comparison between accuracy across subjects for MS-AMF with and without TMSM.

C. The Influence of Modules on MS-AMF

The MS-AMF network consists of two important modules, temporal multi-scale module (TMSM), and the attention module. The TMSM captures the temporal dynamic information by extracting the temporal features of EEG in parallel. The AM can assist the network model in calibrating the features so that the network selectively amplifies the valuable feature channels and inhibits the useless feature channels to increase the sensitivity of the network. To study the effects of the two modules on the classification effect of motor imagery signals, we used the method of contrast test and confounding matrix analysis to compare the classification effect.

1) Analysis of TMSM: As shown in Fig. 4, we compared the classification effect of the MS-AMF network with or without TMSM(a branch with normal Conv and the same parameters), and the network can only extract information within one scale without the TMSM.TMSM had varying degrees of influence on the classification effect for all subjects, among which the classification effect of 7 subjects had been improved, the max increase of which is 3.3% for S4, the classification rate of two subjects decreased, and the max decrease of which is 2.19% for S2. This shows that although there are individual differences among subjects, the temporal multi-scale module is beneficial to EEG decoding.

TABLE II
STRUCTURE OF MS-AM FUSION NETWORK MODEL

	FBCSP		n-CSP	Normal Deep-learning		Complex architecture			
	FBCSP[21]	SVM+FBCSP [26]	NCSP[42]	C2CM [26]	Modular Network [43]	AX-LSTM [30]	FDBN[44]	3D CNN [45]	MS-AMF
Subject 1	76.00	82.29	79.16	87.5	84.91	75.12	71.08	77.39	88.32
Subject 2	56.50	60.42	52.08	65.28	66.38	71.38	55.56	60.14	65.69
Subject 3	81.25	82.99	83.33	90.28	84.74	72.24	76.87	82.92	91.97
Subject 4	61.00	72.57	62.15	66.67	81.36	72.92	65.62	72.28	77.68
Subject 5	55.00	60.67	54.51	62.50	79.22	82.62	69.08	75.83	60.89
Subject 6	45.25	44.10	39.24	45.49	70.67	69.64	64.98	68.98	63.67
Subject 7	82.75	86.11	83.33	89.58	86.12	88.98	71.68	76.03	88.15
Subject 8	81.25	77.08	82.64	83.33	83.81	80.28	92.37	76.85	93.23
Subject 9	70.75	75.00	66.67	79.51	83.04	75.07	82.38	84.66	89.51
AVE	67.75	71.18	67.01	74.46	80.03	76.47	72.18	75.0	79.90

TABLE III
THE PERFORMANCE COMPARISON BETWEEN ACCURACY ACROSS SUBJECTS FOR MS-AMF WITH AND WITHOUT AM

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	AVE
Without AM	76.64	55.47	83.94	62.81	47.36	52.06	80.00	87.21	83.87	69.93
MS-AMF	88.32	65.69	91.97	77.68	60.89	63.67	88.15	93.23	89.51	79.90

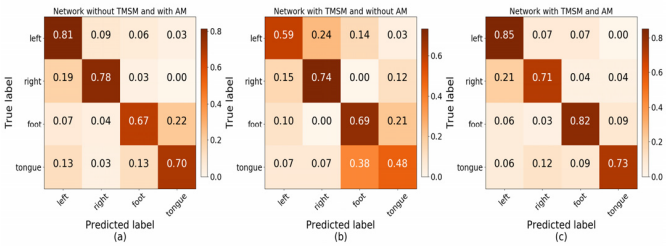

Fig. 5. Confusion matrixes of the subject 4. (a) network without TMSM and with AM. (b) network with TMSM and without AM. (c) network with TMSM and AM.

2) Analysis of AM: Table III shows the EEG decoding effect of the proposed ME-AMF network with and without AM. The addition of the attention module greatly improved the classification effect of all subjects' networks, especially subject 4, which increased by 14.87%. The ninth participant increased only 5.64%; The classification rate of subjects 3 and 8 further exceeded 90% under the condition of good basic classification effect, indicating that AM is beneficial to EEG decoding function.

3) Analysis of Confusion Matrixes: According to the above two parts, we could see that subject4 was the most sensitive to these two modules, subject 8 had the best classification result, and their classification results had different representativeness, so we chose subject4 and subject8 to observe their classification result respectively.

First, we analyzed the confusion matrix of subject4. Fig. 5 (a) and (c) showed the confusion matrix of subject4 with or without TMSM. It could be seen that TMSM improved the classification effect between foot and tongue significantly, but it also affected the classification result between right hand and tongue. Fig. 5(b) and (c) illustrated the confusion matrix of subject4 with or without AM; we could discover that the AM increased the classification rate between three classes MI except for the right hand by comparison.

Second, we could see that the classification effect of subject 8 was more prominent than that of other subjects; it had the most discriminating information, so we analyzed the confusion matrix of subject8. We could compare (a) and (d), in which (a) did not add TMSM and AM, and found that the classification effect of the four categories had a significant improvement, it might indicate that the two modules were beneficial to the motor imaginary EEG decoding. As shown in Fig. 6(a) and (b), we found that the presence of TMSM affected the classification effect between the feet and hands, and the sensitivity between left hand and foot, tongue and right hand was enhanced, but the classification effect between foot and right hand was sacrificed, which might further explain that the influence of TMSM on subjects was different due to individual differences.

The paper [46] shows that the upper limb (left hand and right hand) task produces a more obvious pattern classification effect than the lower limb (left foot and right foot) task, as Fig. 5, 6 show that the addition of the AM module in the special training improves the decoding effect of the lower limbs and tongue, and reduces the difference in the classification effects of the upper and lower limbs. Fig. 6(a) and (c), (b) and (d) showed that the influence of attention module is focused on the classification effect of feet. Compared with the results in subject4 above, other categories of classification effect have a larger baseline, so it has a little effect, and the sensitivity between feet and other parts is greatly improved.

D. The Influence of Different Ratios in AM and Atrous Conv

The reduction ratio r introduced in paper is a hyperparameter which allows us to vary the capacity and computational cost of the SE blocks in the network. To investigate the trade-off between performance and computational cost mediated by this hyperparameter, we conduct experiments with MS-AMF

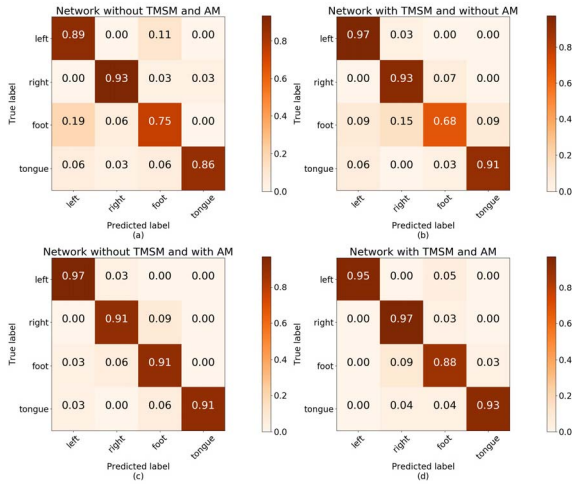

Fig. 6. Confusion matrixes of the subject 8. (a) network without TMSM and AM. (b) network with TMSM and without AM. (c) network without TMSM and with AM. (d)network with TMSM and AM.

TABLE IV
DECODING ACCURACY AND PARAMETER SIZES FOR MS-AMF MODEL AT DIFFERENT REDUCTION RATIOS

r	Acc(%)	Params
2	84.96	13.5Mb
4	93.23	6.75Mb
8	91.72	3.87Mb
16	88.72	1.94Mb
32	89.47	0.97Mb

model of typical subject 8 for a range of different r values. The comparison in Table IV. shows that performance is fluctuant to a range of reduction ratios. Increased complexity does not improve performance monotonically, while a smaller ratio dramatically increases the parameter size of the model. Finally, we set r = 4 which can achieve a good balance between accuracy and complexity.

The kernel length of Atrous convolution also affects the decoding accuracy of EEG signals. We choose a typical subject 8 for analysis. As can be seen from the Fig. 7, firstly we fix the kernel length of the Atrous convolution in the second dense layer to 3 and increase the kernel length of the Atrous convolution in the first dense layer. The kernel length was increased from 3 to 11 then obtain 1-dense change decoding accuracy bar chart. The classification accuracy first decreases, and the accuracy increases when the length is 11. The other group of histograms represents the decoding accuracy obtained by simultaneously increasing the kernel length of the Atrous convolution in the module, and the fluctuation of the decoding accuracy is similar. It indicates that large kernel length of Atrous convolution has little improvement on the EEG signals decoding in this model. By weighing the decoding performance and the number of parameters, we determined that the kernel length was 3. We use parallel computing of common convolution and Atrous convolution to extract multi-scale features. Therefore, our MS-AMF model can obtain

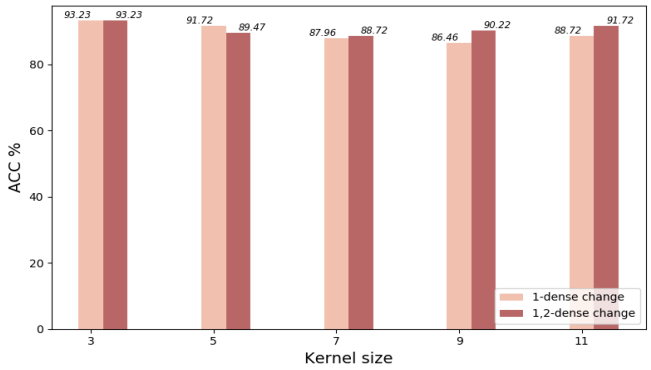

Fig. 7. Kernel sizes for Atrous convolution at different reduction ratios.

more discriminative temporal information at different scales by using the multi-scale temporal module with the optimal kernel length.

IV. DISCUSSION AND ANALYSIS OF NETWORK VISUALIZATION

Visualization is an important method to analyze the network bottom layer and explain network behavior [47]. Firstly, to prove the effectiveness of the proposed multi-brain input, we visualized the part of the network. In the previous part, we got the conclusion that AM has a great influence on the network classification effect. Therefore, in this section, we explored the underlying reason why AM is beneficial to EEG decoding by visualizing it. Finally, due to the time-domain dynamics of EEG signals, different time periods have different responses to external excitation, which we explored through Grad-CAM [48] visualization.

A. Brain Regions Activation Analysis

In this paper, a network input method for parallel processing of multi-brain regions was proposed, in which two brain regions were used, namely the central region and the top region. To analyze the importance of the two brain regions, we added an AM into the spatial multi-scale module to facilitate visualization (the location on #1 flag). We marked and explained the newly added modules in Fig. 2. As shown in Fig. 2, we added two AM modules to the marked location and fine-tuned the network. In order to reduce the gap between the fine-tuned network and the original network, we also reduced the number of iterations, learning rate and decay rate during the training process. The average decoding accuracy of the fine-tuned model will fluctuate within 3% compared to the original model.

We select a group of left-hand test data of subject8 for the test. which obtained the activation values of the two brain regions' respective feature maps from attention modules (see Formula 3). We sort the corresponding feature-maps in descending order according to the activation value; Then in different range we observe the proportion of feature-maps that each brain region corresponding to. The results are shown in the Fig. 8. The y-axis represents the number of the feature-map of the brain region in the current range;

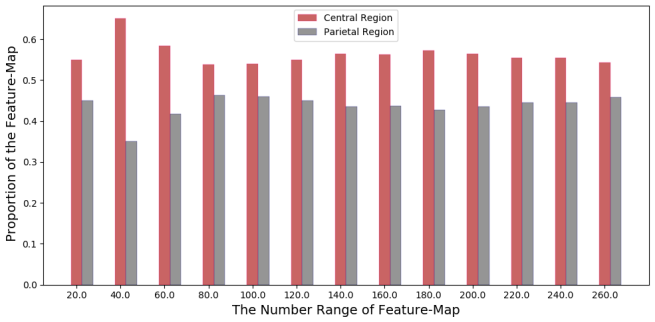

Fig. 8. The proportion of the activation values after sorted in the two brain regions.

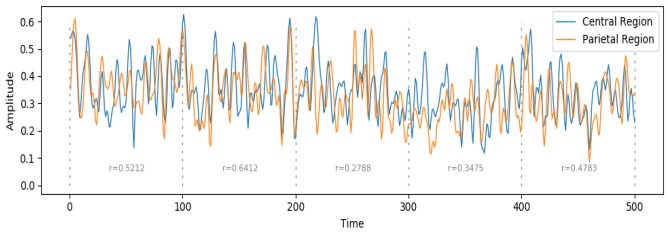

Fig. 9. The average value of the activated feature map by the AM module.

the x-axis represents the number range of important feature-maps after sorting. It can be found that in the range of 1-60 after the weight ranking, the proportion and gap of the two brain regions fluctuate considerably, with the maximum difference of 0.3. The proportion of the central region is always much more significant than that of the parietal region. In the range 1-80 and 1-260, the proportion and the gap between two brain regions less volatile, the most significant gap is 0.14, the central brain regions, where the proportion is still higher than the parietal region. We can get the conclusion: the central brain regions dominated by information in the decoding process, the central regions for motor imagine to stimulate more sensitive, this conclusion consistent with the literature research.

Besides, we compared the average values of the feature maps obtained after the activation of two brain regions by AM, and the results are shown in Fig. 9. In the first 200 time-domain data, the correlation degree of the two groups of brain regions was relatively large, especially in the range of 100-200, and the correlation coefficient reached 0.6412. Since the 0-200 time-domain data we selected included part of the cueing process and the beginning of the motor imagery trial process, we can speculate that the reason for the significant correlation coefficient between the top area and the central area in this time period is that the effect of the top area on the decoding of motor imagine may be related to attention and additional stimulus.

B. Activator Analysis of Attention Module

Attention module has a significant influence on the network, which realizes the weight mapping of the input feature-map through the activity factor. The activity factor represents the sensitivity of the information characteristics contained in the corresponding feature-map. To further explore the role of AM in the network, in this section, we respectively study the distribution of activity factors within and between dense blocks of MS-AMF model.

TABLE V
THE DEGREE OF DISPERSION OF ACTIVATOR FACTOR DISTRIBUTION

AM model of normal convolution	Std(%)	AM model of Atrous convolution	Std(%)
AM1	1.210	AM2	0.706
AM3	1.295	AM4	0.717
AM6	2.059	AM7	1.576
AM8	2.245	AM9	1.638

Due to the inter-class differences, we explored the data of left hand, right hand, foot and tongue. For each category, we respectively extracted a group of samples from the test set for testing, respectively calculated the excitation factors of the channels from am1-am10, and plotted their distribution. The results are shown in Fig. 10, 11. Due to the large number of activity factors of AM between dense blocks, only a part of them is selected for demonstration in order to facilitate observation and analysis.

First, we explored the activity factor in the dense block. In dense block1, we observed AM1∼AM4 and found that the activation-distribution between different classes in the shallow layer of the network is similar and inseparable. Besides, it could be seen that different convolution has different sensitivities to features; we observe the difference in the distribution of activator factor in the AM module after common convolution and Atrous convolution. The distribution of the activator factor among the AM1\3\6\8 modules after common convolution shows more distinct preferences than the AM2\4\7\9 after Atrous convolution. We measure the degree of dispersion for four types from each AM model by calculating the average standard deviation of the activator values in the same channel. $diff = \frac{1}{128}\sum_{i}^{128} std\,(L_i, R_i, F_i, T_i)$. As shown in the Table V, the degree of dispersion of activator factor distribution after common convolution is greater than that after Atrous convolution, which indicates that the discriminability of the decoded information obtained by the same input after common convolution is outstanding.

In dense block2, we observed AM6∼AM9 and found that with the increase of block depth, the specificity of activity factors among different types also increased. We observed AM in order 2-4-7-9 and 1-3-6-8. With the increase of depth, four types of MI data show different preferences for feature sensitivity. The activation-distribution of each channel became more specific. As the Table V show that, the standard deviation of AM8 and AM9 is higher than that of AM3 and AM4. It is indicating that the activator values of deep features have greater dispersion degree and specificity.

We computed the correlation of the activity distribution in the AM8 and AM9 between the four types. As shown in the Table VI, the correlation coefficients of the foot types are smaller than the other types. The specificity of the foot is higher than that of other classes, which also proves that

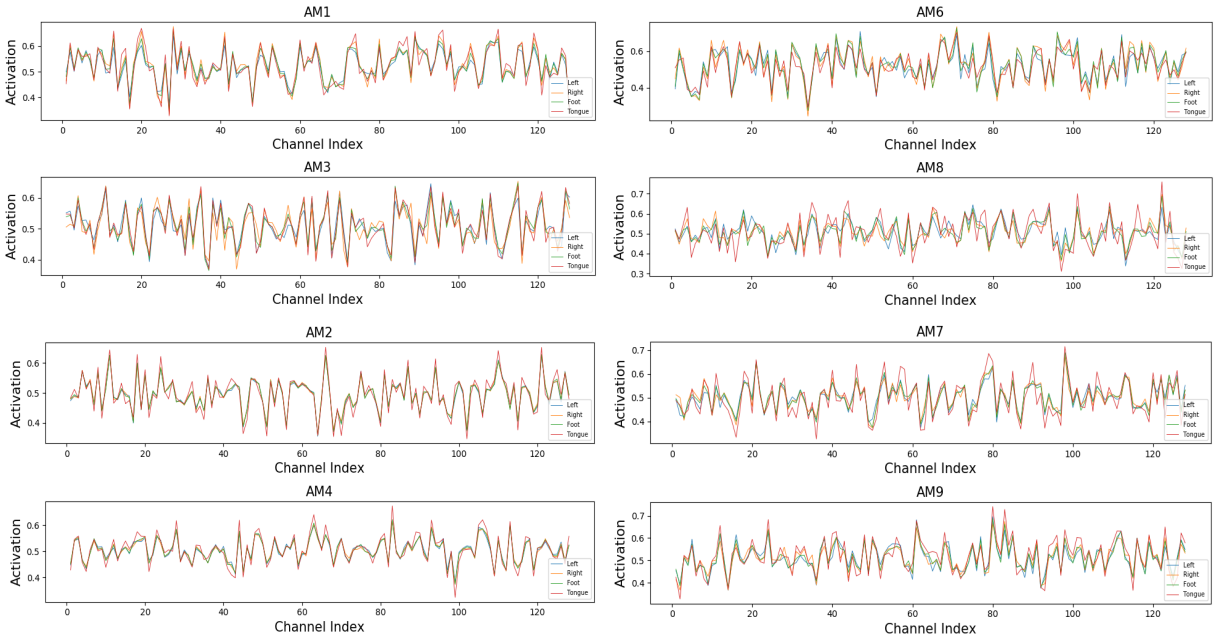

Fig. 10. The activity factors value distribution of AM modules within dense blocks.

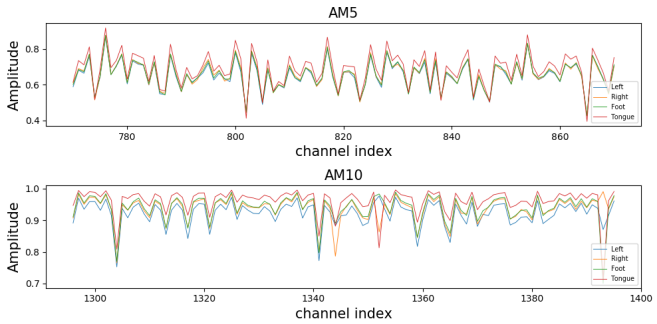

Fig. 11. The activity factors value distribution of AM modules between dense blocks.

TABLE VI
THE CORRELATION OF THE ACTIVITY DISTRIBUTION BETWEEN THE FOUR TYPES

	AM8		AM9
L vs R	0.7703	L vs R	0.8593
F vs R	**0.81**	**F** vs L	**0.9191**
F vs L	**0.8275**	**F** vs T	**0.9244**
F vs T	**0.8418**	**F** vs R	**0.9377**
R vs T	0.8931	L vs T	0.9494
L vs T	0.9253	R vs T	0.9755

confounding matrix for the AM of analysis in the previous section has a greater influence on feet.

Second, we explore the activity factors between the dense block. It shows that the separability of AM5 is far higher than the AM in dense block1, but there is a problem of distribution aliasing among classes. In contrast, the distribution of AM10 gradually shows a similar trend. This shows that as the depth of the network increases, the AM module has a limited effect in increasing the specificity of different types of data.

C. Temporal Attention Analysis of Different Classes

The Grad-cam method was used to visualize the signals to explore the temporal attention interval of four kinds of EEG signals by the network model. It has been pointed out that the deeper convolutional layer in the network can capture more advanced signals features, and some spatial information can be retained compared with the fully connected layer. Therefore, the last convolutional layer in the network can balance the high-level semantic features and detailed spatial information. In grad-cam, the gradients of the score (before the soft-max) for four classes were computed with respect to the feature map of the last convolutional layer; then the gradients of each feature map were global averaged to obtain the importance weight of the corresponding feature map, and got the preliminary coarse heat map; finally, we conducted up-sampling on the heat map to get the final map which could show the temporal attention of the original signal.

Fig. 12 shows the four classes Grad-CAM result of MS-AMF. MS-AMF focused on different ranges of EEG signals for MI classification; the network has a high degree of attention during the first 50 time points of the four MI classes temporal signal. The selected data of 0s-2s contained the temporal point of 0.25s cue, the subjects received external visual stimuli during this period, so their attention was relatively concentrated, and the MI EEG signals' feature was prominent. After the completion of the curing process, the attention level of the three MI EEG signals is shallow, and the attention level increases around the 200th temporal point and lasts for a while. Since the original data we selected is only the temporal signal of the first half trial, we speculated that the temporal attention level of MI EEG signals has a certain relationship with the

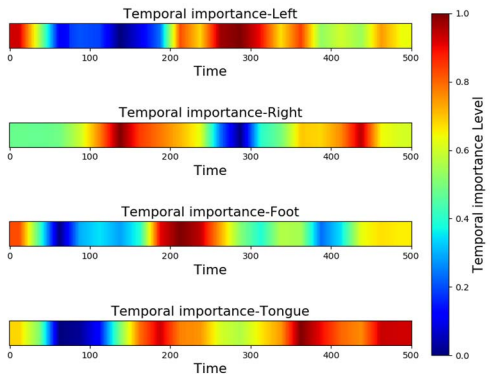

Fig. 12. Grad-Cam heat map for MI data.

attention degree of the subjects. Interestingly, we observed a set of opposite trends in the temporal attention of the right and left hands at certain time intervals, which we will do more research in the future.

V. Conclusion

In summary, we proposed a multi-scale fusion convolutional neural network based on the attention mechanism for EEG signals decoding and visualization analysis. The multi-brain region representation method and the spatial multi-scale module can preserve the temporal features of the EEG and the spatial features of the brain region, improve the resolution and analyze the EEG signals from different views. The temporal multi-scale module and the dense fusion module ensure the maximum information flow among each layer of the network and then extract the multi-scale features of the electrodes between brain regions. Finally, the attention module improves the expression ability of the network model. The results on the public motor imagery EEG datasets (BCI IV-2a) reached 79.90%, and our proposed method achieved higher classification accuracy than the baseline method. The five modules of the network also provide different perspectives for the analysis, which is convenient to analyze the feature extraction process of the motor imagery EEG signals from the model framework. However, with the increase of precision and analysis module, the network parameters also increase gradually. Besides, the application of visualization technology in decoding EEG signals can help analyze deep learning underlying decoding process and the importance localization of time-frequency domain and brain regions. e.g. Activate visual images of brain regions in fMRI research to identify brain state; locate regions where brain diseases such as epilepsy occur. We can explore the architecture and functional network of the brain through the methods as weight visualization and regional visualization. The proposed MI classification framework has a certain universality, which can be extended to other areas of analysis related to EEG. The influence of the number and position of attention modules on decoding and the generalization of the model need to be further studied in the future. We can simplify the deep learning model and apply it to the data under different conditions. The visualization analysis also requires further validation with rich data and different environments.

References

[1] L. J. Greenfield, J. Geyer, and P. Carney, *Reading EEGs: A Practical Approach*. Philadelphia, PA, USA: Lippincott Williams & Wilkins, 2012.

[2] L. Chen *et al.*, "Three-layer weighted fuzzy support vector regression for emotional intention understanding in human–robot interaction," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2524–2538, Oct. 2018.

[3] J. Rafferty, C. D. Nugent, J. Liu, and L. Chen, "From activity recognition to intention recognition for assisted living within smart homes," *IEEE Trans. Human-Machine Syst.*, vol. 47, no. 3, pp. 368–379, Jun. 2017.

[4] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," *NeuroImage*, vol. 31, no. 1, pp. 153–159, May 2006.

[5] G. Pfurtscheller and F. H. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: Basic principles," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1842–1857, Nov. 1999.

[6] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, no. 2, pp. pp. R32–R57, 2007.

[7] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008.

[8] L. Tonin, T. Carlson, R. Leeb, and J. del R Millan, "Brain-controlled telepresence robot by motor-disabled people," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 4227–4230.

[9] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, 2008.

[10] T. Carlson and J. del R. Millan, "Brain-controlled wheelchairs: A robotic architecture," *IEEE Robot. Autom. Mag.*, vol. 20, no. 1, pp. 65–73, Mar. 2013.

[11] H.-J. Hwang, S. Kim, S. Choi, and C.-H. Im, "EEG-based brain-computer interfaces: A thorough literature survey," *Int. J. Human-Comput. Interact.*, vol. 29, no. 12, pp. 814–826, Dec. 2013.

[12] Z. Gao *et al.*, "An adaptive optimal-kernel time-frequency representation-based complex network method for characterizing fatigued behavior using the SSVEP-based BCI system," *Knowl.-Based Syst.*, vol. 152, pp. 163–171, Jul. 2018.

[13] M. Guggenmos, P. Sterzer, and R. M. Cichy, "Multivariate pattern analysis for MEG: A comparison of dissimilarity measures," *NeuroImage*, vol. 173, pp. 434–447, Jun. 2018.

[14] F. Pichiorri *et al.*, "Brain-computer interface boosts motor imagery practice during stroke recovery: BCI and motor imagery," *Ann. Neurol.*, vol. 77, no. 5, pp. 851–865, May 2015.

[15] N. F. Ince, F. Goksu, A. H. Tewfik, and S. Arica, "Adapting subject specific motor imagery EEG patterns in space–time–frequency for a brain computer interface," *Biomed. Signal Process. Control*, vol. 4, no. 3, pp. 236–246, Jul. 2009.

[16] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimination between imagination of right and left hand movement," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, pp. 642–651, Dec. 1997.

[17] M. Sharma, A. Dhere, R. B. Pachori, and U. R. Acharya, "An automatic detection of focal EEG signals using new class of time–frequency localized orthogonal wavelet filter banks," *Knowl.-Based Syst.*, vol. 118, pp. 217–227, Feb. 2017.

[18] H. Yuan and B. He, "Brain–computer interfaces using sensorimotor rhythms: Current state and future perspectives," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1425–1435, May 2014.

[19] Z. J. Koles, J. C. Lind, and A. C. K. Soong, "Spatio-temporal decomposition of the EEG: A general approach to the isolation and localization of sources," *Electroencephalogr. Clin. Neurophysiol.*, vol. 95, pp. 219–230, Nov. 1995.

[20] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[21] S.-H. Park, D. Lee, and S.-G. Lee, "Filter bank regularized common spatial pattern ensemble for small sample motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 498–505, Feb. 2018.

[22] H. Higashi and T. Tanaka, "Simultaneous design of FIR filter banks and spatial patterns for EEG signal classification," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1100–1110, Apr. 2013.

[23] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve BCI designs: Unified theory and new algorithms," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 355–362, Feb. 2011.
[24] S. Lemm, B. Blankertz, G. Curio, and K.-R. Muller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, Sep. 2005.
[25] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[26] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain-computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.
[27] Y. R. Tabar and U. Halici, "A novel deep learning approach for classification of EEG motor imagery signals," *J. Neural Eng.*, vol. 14, no. 1, Feb. 2017, Art. no. 016003.
[28] B. Ghazi, R. Panigrahy, and J. R. Wang, "Recursive sketches for modular deep learning," in *Proc. 36th Int. Conf. Mach. Learn.*, Long Beach, CA, USA: PMLR, vol. 97, 2019, pp. 1–52.
[29] N. Michielli, U. R. Acharya, and F. Molinari, "Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals," *Comput. Biol. Med.*, vol. 106, pp. 71–81, Mar. 2019.
[30] P. Wang, A. Jiang, X. Liu, J. Shang, and L. Zhang, "LSTM-based EEG classification in motor imagery tasks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 11, pp. 2086–2095, Nov. 2018.
[31] D. Li, J. Wang, J. Xu, and X. Fang, "Densely feature fusion based on convolutional neural networks for motor imagery EEG classification," *IEEE Access*, vol. 7, pp. 132720–132730, 2019.
[32] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019.
[33] S. Kantak, J. Stinear, E. Buch, and L. Cohen, "Rewiring the brain: Potential role of the premotor cortex in motor control, learning, and recovery of function following brain injury," *Neurorehabilitation Neural Repair*, vol. 26, pp. 282–292, Sep. 2011.
[34] D. Hermes *et al.*, "Functional MRI-based identification of brain areas involved in motor imagery for implantable brain–computer interfaces," *J. Neural Eng.*, vol. 8, no. 2, Apr. 2011, Art. no. 025007.
[35] J. C. Henry, "Electroencephalography: Basic principles, clinical applications, and related fields," *Neurology*, vol. 67, no. 11, p. 2092, 2006. [Online]. Available: https://n.neurology.org/content/67/11/2092.2
[36] P. Hammon, S. Makeig, H. Poizner, E. Todorov, and V. De Sa, "Predicting reaching targets from human EEG," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 69–77, Dec. 2008.
[37] Y. Zhang, X. Zhang, H. Sun, Z. Fan, and X. Zhong, "Portable brain-computer interface based on novel convolutional neural network," *Comput. Biol. Med.*, vol. 107, pp. 248–256, Apr. 2019.
[38] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," Nov. 2015, *arXiv:1511.06448*. [Online]. Available: https://arxiv.org/abs/1511.06448
[39] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.
[40] D. Hermes *et al.*, "Functional MRI-based identification of brain areas involved in motor imagery for implantable brain–computer interfaces," *J. Neural Eng.*, vol. 8, no. 2, Mar. 2011, Art. no. 025007.
[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[42] H. Yang, S. Sakhavi, K. Keng Ang, and C. Guan, "On the use of convolutional neural networks and augmented CSP features for multi-class motor imagery of EEG signals classification," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 2620–2623.
[43] B. E. Olivas-Padilla and M. I. Chacon-Murguia, "Classification of multiple motor imagery using deep convolutional neural networks and spatial filters," *Appl. Soft Comput.*, vol. 75, pp. 461–472, Feb. 2019.
[44] N. Lu, T. Li, X. Ren, and H. Miao, "A deep learning scheme for motor imagery classification based on restricted Boltzmann machines," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 6, pp. 566–576, Jun. 2017.
[45] X. Zhao, H. Zhang, G. Zhu, F. You, S. Kuang, and L. Sun, "A multi-branch 3D convolutional neural network for EEG-based motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2164–2177, Oct. 2019.
[46] A. M. Batula, J. A. Mark, Y. E. Kim, and H. Ayaz, "Comparison of brain activation during motor imagery and motor movement using fNIRS," *Comput. Intell. Neurosci.*, vol. 2019, p. 12, May 2017.
[47] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional neural networks," in *Proc. ECCV*, Nov. 2013, p. 8689.
[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.