Proceedings Paper:
Zhang, J and Li, K (2022) A Pruned Deep Learning Approach for Classification of Motor Imagery Electroencephalography Signals. In: Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)., 11-15 Jul 2022, Glasgow, Scotland, UK. IEEE , pp. 4072-4075. ISBN 978-1-7281-2783-5

https://doi.org/10.1109/EMBC48229.2022.9871078

A Pruned Deep Learning Approach for Classification of Motor Imagery Electroencephalography Signals

Jiayang Zhang[1] and Kang Li[1,*]

Abstract— **The Deep Learning (DL) approach has been gaining much popularity in recent years in the development of electroencephalogram (EEG) based Motor Imagery (MI) Brain-Computer Interface (BCI) systems, aiming to improve the performance of existing stroke rehabilitation strategies. A complex deep neural network structure has lots of neurons with thousands of parameters to optimize, and a great deal of data is often required to train the network and the training process can take an extremely long time. High training costs and high model complexity not only have negative impacts on the performance of the BCI system but also on its applicability to meet the real-time requirement to support the rehabilitation exercises of patients. To tackle the challenge, a contribution-based neuron selection method is proposed in this paper. A Convolutional Neural Network (CNN) based motor imagery classification framework is implemented, and a neuron pruning approach is developed and applied. The temporal and spatial features of EEG signals are captured by the CNN layers, and then the fast recursive algorithm (FRA) is applied to prune redundant parameters in the fully connected layers which reduces the computation cost of the CNN model without affecting its performance. The experimental results show that the proposed method can achieve up to 50% model size reduction and 67.09% computation savings.**

I. INTRODUCTION

Stroke is one of the major causes of long-term disability among adults and imposes significant socio-economic burden globally [3]. Stroke frequently results in long-term motor disabilities such as paralysis of the upper limbs [13]. Post-operative rehabilitation has a great impact on the normal life of patients. Therefore, there is a growing demand for better and efficient rehabilitative interventions. In recent years, the brain–computer interface (BCI)-based system shows promising results for post-stroke motor rehabilitation. BCI-based intervention can lead to swift functional recovery by building the relationship between motor intention and sensory feedback of motor movements. Electroencephalography (EEG), the measure of the electrical fields produced by the brain activities, is a brain mapping and neuroimaging technique widely used inside and outside the clinical domain [4]. Specifically, EEG picks up the electric potential differences, in the order of tens of μV, that reach the scalp when tiny excitatory postsynaptic potentials produced by pyramidal neurons in the cortical layers of the brain are aggregated. A BCI system enables real-time decoding of brain dynamics [5] using EEG signals generated solely by motor imagery (MI).

EEG signals often contain a vast amount of information about the functioning of the brain, and hence classification and evaluation of these signals are of paramount importance in the intention recognition of patients associated to various brain activities. Among various classification approaches, the deep learning (DL), as a family of machine learning tools for extracting high-level information embedded in data by applying deep-layered architectures, has been successfully applied to decode EEG signals in the last few years. In 2017, Schirrmeister et al created three ConvNets with different structures, where the number of convolutional layers ranges from 2 layers in a “shallow” ConvNet over a 5-layer deep ConvNet up to a 31-layer residual network (ResNet) [6]. Among them, the “Shallow” and “Deep” ConvNet structures have great influence on the performance of the resultant models. Lawhern in 2018 purposed a new structure namely “EEGnet” [7] which can achieve 70% classification accuracy on BCI competition IV dataset 2a [9]. In 2020, Amin et al proposed the FBCnet model and the binary classification accuracy on the Korean Dataset [1] reached 74%. With the development of new studies on using DL to decode EEG signals, complex networks with various layers combination have been proposed, such as the convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory (LSTM) and fully connected layers network. For example, Mammone et al [10] use three CNN layers and two fully connected layers to decode signals. This deep CNN model can achieve up to 90.3% accuracy rate on public dataset from the BNCI Horizon 2020 website. Amin et al further combine four models with different numbers of layers together to classify four-class tasks [11], achieving 6% higher accuracy than some previous studies. Zhang et al used a network with three CNN layers and LSTM layers respectively to achieve a satisfactory accuracy [12]. Although more complex networks can help improve the classification accuracy, high training cost and long training time issues cannot be ignored, especially considering that an useful BCI system needs to be used in real-time applications. Meanwhile, the nature of the EEG signals is non-linear, nonstationary and noisy. This means that a neural network with many redundant neurons and parameters may has the overfitting problem and the computational efficiency is also relatively low.

In this paper, to tackle these challenges, a new framework is proposed to prune neurons and save computations. The new framework first begins with the construction of a general

*Corresponding author

[1]Jiayang Zhang and Kang Li are with school of Electrical Engineering, University of Leeds, Woodhouse Lane Leeds, West Yorkshire, LS2 9JT, UK. Email:eljzh@leeds.ac.uk & k.li1@leeds.ac.uk

CNN model. In this model, one dimension CNN filter and depthwise CNN are used in the first two layers for learning the temporal and spatial features of EEG signals. Then an average pooling layer and two fully connected layers are used to combine features and produce final outputs. To reduce redundant trainable parameters, the fast recursive algorithm (FRA) [8] is used to calculate the contributions of each neuron in the last two layers. A threshold is used to ensure that the outputs after pruning are similar with the original ones, which implies that the resultant new model will take much less time to train and test while the model accuracy is not affected by large. Section III presents the details of the developed network and the FRA approach, and the experimental results are given in Section IV. Finally, Section V concludes the paper and future work is also discussed.

II. MATERIALS

We evaluate our method on a 54 subject MI dataset [1]. The data were collected from 62 Ag/AgCl electrodes with a sampling rate of 1,000 Hz. The EEG amplifier used in the experiment was a BrainAmp (Brain Products, Germany). The channels were nasion-referenced and grounded to electrode AFz. The dataset consists of binary data of left hand vs right hand MI. In this work, the data from the first session consists of 200 EEG trials. Every subject has 3s to prepare at the beginning of each trial. The center of the monitor showed a black fixation cross to remind subjects. Afterward, the subject would conduct an imagery task of grasping for 4s according to the right or left arrow that appeared as a visual cue. Each of the 4s in length were used (Fig. 1). According to [2], this report selected 20 channels in the motor region for the classification task including the channels FC1-6, Cz, C1-6, Cpz, and CP1-6. The data is down-sampled by a factor of 4 to the sampling frequency of 250 Hz.

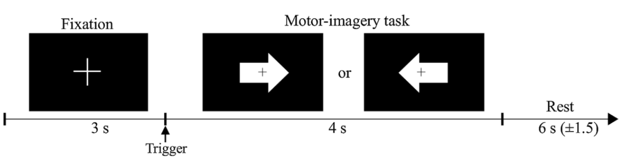

Fig. 1. Experimental design for binary MI-BCI testing

III. METHODS

A. Deep Neural Network Structure

The model is designed to extract the temporal and spatial information as signatures of MI. The modeling process has three main steps. 1) An 1-D CNN filter on each channel learns the temporal features of the EEG data; 2) The depthwise 2D convolution (Conv2D) layer extracts spatial feature based on all 20 channels; 3) The average pooling layer and final fully connected (FC) layer classify features from the CNN layer into given classes. The architecture is illustrated in Fig. 2. Exponential linear units (ELUs) are used as the activation function in the proposed network as they can accelerate learning and improve classification accuracy. To address the overfitting problem, dropout and batch normalization techniques are used.

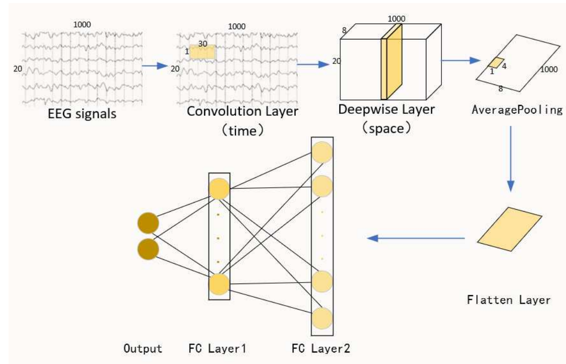

Fig. 2. The proposed deep neural network structure

The model developed in this paper has 2 convolutional layers, 1 average pooling layer, 1 flatten layer and 2 fully-connected layers with a total of 25,608 parameters. To reduce the model complexity and the number of parameters, a FRA-based network pruning method is used to remove redundant neurons according to the contributions of all neurons to the outputs. Because more than half of the total parameters of the CNN model come from the last two fully-connected layers, the proposed pruning method is therefore applied to these two layers. The redundant neurons and all the incoming and outgoing connections associated with these neurons are removed, leading to significantly reduced memory usage and computational complexity in online MI-BCI systems. To calculate the difference between the real output with the model predicted output after FRA pruning, the root mean square error is used:

$$rmse = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \quad (1)$$

when y_i is the output of the FC layer and $\hat{y}_i$ is the output of the new FC layer after pruning. The contribution of each neuron is calculated until rmse value < 0.01.

B. Pruning

The linear-in-the-parameter model is a model structure for approximating a large class of nonlinear systems. These models linearly combine a set of model terms that are non-linear functions of the system variables. Excessive number of candidate terms in these models may cause overfitting and high computational complexity, therefore, model selection algorithms have been proposed to generate parsimonious models with a much smaller number of model terms. The fast recursive algorithm (FRA) [8] was proposed to simultaneously select most significant model terms and estimate model parameters. Suppose a nonlinear dynamic system can be represented by a linear-in-the-parameters model, which is identified using N data samples$\{x(t), y(t)\}_{t=1}^{N}$, the linear-in-the-parameter model can be represented as:

$$y = \Psi\Theta + \Xi \quad (2)$$

where $y = [y(1), \ldots, y(N)]^T \in R^N$ denotes the system output, $\Psi = [\varphi_1, \ldots, \varphi_j, \ldots, \varphi_S] \in R^{N \times S}$ is the regression matrix that contains all candidate model terms, each term $\varphi_j \in R^{N \times 1}$, $\varphi_j = [\varphi_j(x(1)), \ldots, \varphi_j(x(N))]^T$ $(j = 1, \ldots, S)$ represents a nonlinear function of N input samples, $\Theta = [\theta_1, \ldots, \theta_S]^T$ are the unknown parameters to be identified, and $\Xi = [\xi_1, \ldots, \xi_N]^T$ is the model residual vector. Two recursive matrixes M_k and R_k, are predefined in FRA to fulfill the forward model selection procedure as:

$$M_k = \Psi_k^T \Psi_k \tag{3}$$

$$R_k = I - \Psi_k M_k^{-1} \Psi_k^T \tag{4}$$

where $\Psi_k \in R^{N \times K}$ includes the first k columns of the full regression matrix Ψ,$k = 1, \ldots, S$ and $R_0 = I$. Thus, when the first k columns in Ψ are selected, the estimation of parameters that minimizes the cost function and the associated minimal cost function can be formulated as:

$$\hat{\Theta}_k = M_k^{-1} \Psi_k^T y \tag{5}$$

$$E_k = y^T y - \hat{\Theta}_k^T \Psi_k^T y \tag{6}$$

When $\{\varphi j, j = 1, \ldots, S\}$ in Ψ are mutually linearly independent, the residual matrix R_k has the distinguished properties:

$$R_{k+1} = R_k - \frac{R_k \varphi_{k+1} \varphi_{k+1}^T R_k^T}{\varphi_{k+1}^T R_k \varphi_{k+1}}, k = 0, 1, \ldots, (S-1) \tag{7}$$

$$R_k^T = R_k, R_k R_k = R_k, k = 0, 1, \ldots, S \tag{8}$$

$$R_k R_j = R_j R_k = R_k, k \geqslant j; k, j = 0, 1, \ldots, S \tag{9}$$

$$R_{k\varphi j} = 0, j \in \{1, \ldots, k\} \tag{10}$$

Now, Equation (6) can be described as:

$$E_k = y^T R_k y \tag{11}$$

To simplify the formulas and decrease the computational complexity, three quantities are consequently defined as:

$$\begin{cases} \varphi_j^{(k)} \triangleq R_0 \varphi_j = \varphi_j \\ a_{k.j} \triangleq \left(\varphi_k^{(k-1)}\right)^T \varphi_j^{(k-1)}, a_1, j \triangleq \varphi_1^T \varphi_j \\ b_k \triangleq \left(\varphi_k^{(k-1)}\right)^T y, b_1 \triangleq \left(\varphi_1^{(0)}\right)^T y = \varphi_1^T y \end{cases} \tag{12}$$

where $j = 1, \ldots, S$ and $k = 0, 1, \ldots, S$. According to the properties of R_k, the net contribution of a new model term ϕ_{k+1} to the cost function can be explicitly calculated as:

$$\triangle E_{k+1} = -\frac{\left(y^T \varphi_{k+1}^{(k)}\right)^2}{\left(\left(\varphi_{k+1}^{(k)}\right)^T \varphi_{k+1}^{(k)}\right)} = \frac{\left(b_{k+1}^T\right)^2}{a_{k+1,k+1}}, k = 0, 1, \ldots, S-1 \tag{13}$$

By calculating the net contribution of each term, the model terms with maximum contributions will be selected one by one. Finally, after all the important model terms have been selected, the parameter for each selected term is calculated as:

$$\hat{\theta}_j = \frac{b_j - \sum_{i=j+1}^{k} \hat{\theta}_j a_{j,i}}{a_{j,j}}, j = k, k-1, \ldots, 1 \tag{14}$$

Equations (13) and (14) constitute the main steps of the FRA, which selects model terms one by one based on (13) and calculates the model parameters for the resultant model based on (14). Fig 3 illustrates the flowchart of the whole process. The main steps are:

- Step 1 - Calculate the contribution of each neuron in the FC1 layer;
- Step 2 - Rank the neurons based on their contributions, and then select the highest ranked neuron. The root mean square error between the new model output $\hat{Y}$ and the actual output Y over all the training samples is used for evaluate if the performance of the pruned model is not significantly different from the original model;
- Step 3 - Repeat Step 2 until the maximum number of neurons in the FC1 layer is reached or the rmse is smaller than the predefined error bound (0.01);
- Step 4 - For the FC2 layer, repeat steps 1-3 to reassign weights and bias of new neurons.

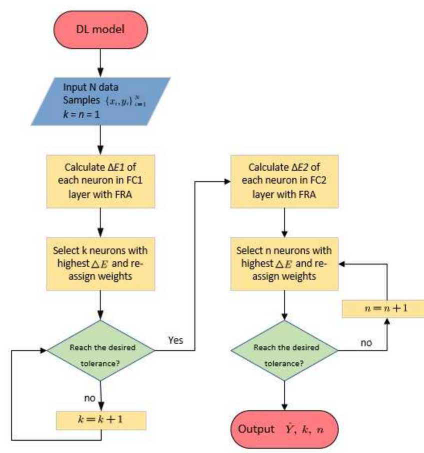

Fig. 3. Flowchart of the FRA-based neuron pruning process

IV. RESULTS

Although the goal of the paper is not to propose a new model to improve the classification accuracy, rather to demonstrate the effectiveness of the proposed network pruning approach, we compare the performance of other basic DL models such as EEGNET4-2, EEGNET8-2 [7], Shallownet [6] on the same dataset. The results are listed in table I.

TABLE I

COMPARISON WITH OTHER BASIC DL MODEL

	EEGNET4-2	Shallownet	EEGNET8-2	Purposed model
Accuracy	0.6444	0.6065	0.6585	0.627

The performance of the general DL model developed in this paper and other models on all 54 subjects is given in Fig 4. According to Fig 4, it is evident that the individual differences are huge. The lowest classification accuracy is only 44% while the highest can reach 95%. The number of tunable parameters in this general DL model is 25,608.

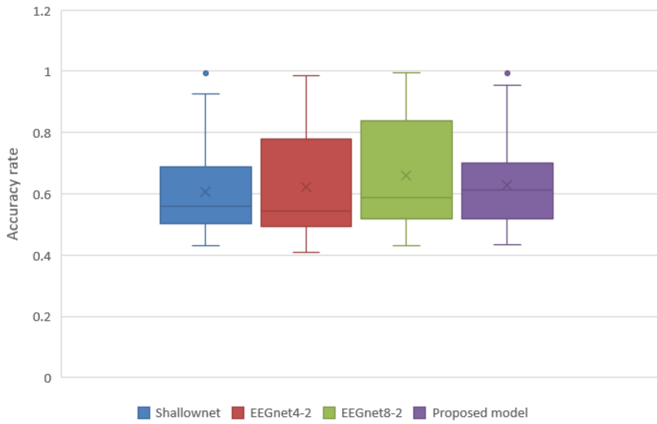

Fig. 4. The box plot based on results of different models

However, there are lots of redundant neurons in the last two fully connected layers. After network pruning using the FRA approach, 17,180 parameters were reduced on average. The average parameters reduction rate can reach 67.09% and the model size reduction rate can reach 50% (Table II). To show

TABLE II

THE MODEL SIZE, TOTAL PARAMETERS OF THE GENERAL DL MODEL BEFORE AND AFTER PRUNED

	General model	Pruned model	Pruned rate(%)
Model size (KB)	372	186	50
Parameter numbers	25608	8428	67.09

the relationship between the accuracy rate and the pruned rate on each subject, Fig 5 combines two results together. The coefficient rate p equals to -0.705 which shows the model with worse classification performance usually has more redundant parameters, leading to a higher pruned rate.

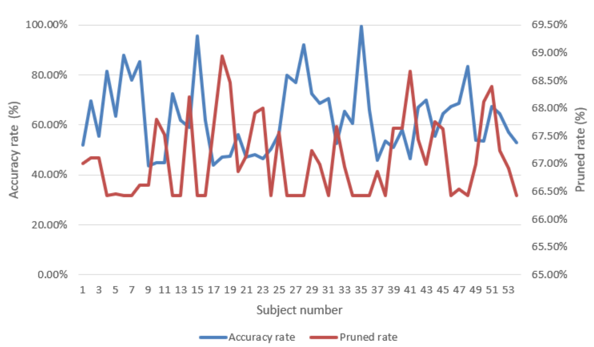

Fig. 5. The accuracy rate and pruned rate on all subjects

V. DISCUSSION AND CONCLUSION

In a BCI system, to better identify a stroke patient's intention, approaches for classifying EEG signals are important. Among these methods, deep learning as an end-to-end process has shown excellent performance in the field of natural language processing and computer vision. Hence, DL methods have been widely used MI BCI to achieve higher classification accuracy in recent years. Complex DL models structure with more layers can extract and learn more information from EEG signals. However, huge individual differences among patients and nonstationary nature of EEG signals usually cause the overfitting problem. Besides that, too much parameters will take hours even days to build a model which is not suitable in a real-time BCI system. Therefore, it is vital to develop approaches that can reduce many computation costs while having little influence on model output accuracy. In this paper, we have proposed a framework using the FRA method to evaluate the contributions of each neuron in the fully connected layers. Then eliminate the neurons that contribute the least while ensure that the difference between the final output of the model after pruning and the optimized output is less than 0.01. The results show that this framework can reduce 67.09% computation cost and 50% model size. The general model we used in this study is a generic CNN model among MI decoding models. Therefore, the proposed framework has a wide range of application fields. Future work will also investigate the pruning of models with deeper and wider neural network architectures.

REFERENCES

[1] Lee M H, Kwon O Y, Kim Y J, et al. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy[J]. GigaScience, 2019, 8(5): giz002.

[2] Mane R, Robinson N, Vinod A P, et al. A multi-view CNN with novel variance layer for motor imagery brain computer interface[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). IEEE, 2020: 2950-2953.

[3] Hankey G J. The global and regional burden of stroke[J]. The lancet global health, 2013, 1(5): e239-e240.

[4] Binnie C D, Prior P F. Electroencephalography[J]. Journal of Neurology, Neurosurgery Psychiatry, 1994, 57(11): 1308-1319.

[5] Wolpaw J R, Birbaumer N, McFarland D J, et al. Brain–computer interfaces for communication and control[J]. Clinical neurophysiology, 2002, 113(6): 767-791.

[6] Schirrmeister R T, Springenberg J T, Fiederer L D J, et al. Deep learning with convolutional neural networks for EEG decoding and visualization[J]. Human brain mapping, 2017, 38(11): 5391-5420.

[7] Amin S U, Alsulaiman M, Muhammad G, et al. Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion[J]. Future Generation computer systems, 2019, 101: 542-554.

[8] Wang, Y., Li, K., Gan, S., Cameron, C. Missing data imputation with ols-based autoencoder for intelligent manufacturing, IEEE Transactions on Industry Applications 55 (6) (2019) 7219-7229.

[9] Brunner C, Leeb R, Müller-Putz G, et al. BCI Competition 2008–Graz data set A[J]. Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology, 2008, 16: 1-6.

[10] Mammone N, Ieracitano C, Morabito F C. A deep CNN approach to decode motor preparation of upper limbs from time–frequency maps of EEG signals at source level[J]. Neural Networks, 2020, 124: 357-372.

[11] Amin S U, Alsulaiman M, Muhammad G, et al. Deep Learning for EEG motor imagery classification based on multi-layer CNNs feature fusion[J]. Future Generation computer systems, 2019, 101: 542-554.

[12] Zhang R, Zong Q, Dou L, et al. Hybrid deep neural network using transfer learning for EEG motor imagery decoding[J]. Biomedical Signal Processing and Control, 2021, 63: 102144.

[13] Kasadha B. Reducing the risk of stroke: looking beyond lifestyle changes[J]. British Journal of Healthcare Management, 2021, 27(9): 241-243.

[14] Dobkin B H. Strategies for stroke rehabilitation[J]. The Lancet Neurology, 2004, 3(9): 528-536.