PARALLEL CONVOLUTIONAL-LINEAR NEURAL NETWORK FOR MOTOR IMAGERY CLASSIFICATION

Siavash Sakhavi[*†], *Cuntai Guan* [*], *Shuicheng Yan*[†]

[*] A*STAR
Institute for Infocomm Research (I^2R)
Brain-Computer Interface Lab
Singapore

[†] National University of Singapore
Department of Electrical and Computer Engineering
Learning and Vision Lab
Singapore

ABSTRACT

Deep learning, recently, has been successfully applied to image classification, object recognition and speech recognition. However, the benefits of deep learning and accompanying architectures have been largely unknown for BCI applications. In motor imagery-based BCI, an energy-based feature, typically after spatial filtering, is commonly used for classification. Although this feature corresponds to the estimate of event-related synchronization/desynchronization in the brain, it neglects energy dynamics which may contain valuable discriminative information. Because traditional classification methods, such as SVM, cannot handle this dynamical property, we proposed an architecture that inputs a dynamic energy representation of EEG data and utilizes convolutional neural networks for classification. By combining this network with a static energy network, we saw a significant increase in performance. We evaluated the proposed method and compared with SVM on a multi-class motor imagery dataset (BCI competition dataset IV-2a). Our method outperforms SVM with static energy features significantly ($p < 0.01$).

Index Terms— Convolutional Neural Network, Deep Learning, Motor Imagery, Brain-Computer Interface, EEG

1. INTRODUCTION

Machine learning algorithms for EEG are not as old as algorithms used for speech, image and text. They are mainly known for their application in brain-computer interfaces (BCI) such as motor imagery (MI), steady state evoked potentials (SS-EP) and event-related potentials (ERP) [1]. When using machine learning for BCI, important properties of EEG which discriminates it from other data must be taken into consideration (i.e. non-stationarity, low signal-to-noise ratio, channel correlation). In most BCI settings, the main goal in BCI is to discriminate brain states in a single recording or trial or in the most limited number of trials possible. This raises the issue of how to expose information that can be seen in average ensembles but not in single trials. Furthermore, because task recording is time-consuming, the number of recorded samples are limited in each session. Solutions to these key challenges discriminates machine learning for BCI from its application for other data.

Proposed algorithms for MI-BCI classification mainly include common spatial patterns (CSP) [2], which combines spatial filtering and static energy feature extraction, and in some cases, multi-band temporal filtering [3] for frequency band selection. As a result, static energy is the widely used representation of EEG for many BCI problems such as domain adaptation [4], channel selection [5], dealing with session-to-session non-stationarity [6] and performance estimation [7]. Consequently, a common negation seen in these algorithms is neglect for the dynamics of the signal during the trial; when static energy features are extracted, the energy dynamics is compacted into a single number and hence, temporal information is discarded. However, Pfurtscheller [8] showed that tongue and foot motor imagery do not have a evident drop in energy, such as the left or right motor imagery, but rather a energy pattern in certain channels and frequencies. Therefore, a methodology that can exploit information temporally can be used in collaboration with energy features to build a classifier which can then handle a wide variety of tasks that have discrimination in time, energy or both.

In recent years, Deep Learning (DL) has become a growing trend in the area of machine learning and artificial intelligence. Changes in activation functions [9], regularization methods to escape overfitting [10], incorporating data and model parallelism [11], new architecture modifications [12], optimized libraries and software for DL research [13] and large amounts of data have all contributed to the success of DL and deep architectures in recent years. In our case, we searched for architectures and DL methods which were suitable for classifying dynamic energy based features and have chosen the convolutional neural network (CNN) [14] as the methodology. Convolution, by nature, slides on input dimensions to detect a pattern based on a learned kernel and can be used a methodology to detect events in signal processing.

CNNs have been used for EEG processing and classification [15, 16] but not in the context of MI-BCI and four-class

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

classification. The main contribution of this paper is design and analysis of a parallel convolutional-linear neural network for 4-class motor imagery classification. In the following sections, we describe how to represent a EEG signal in which can capture energy dynamicity. A parallel CNN and linear architecture is then designed to input both static and dynamic energy features. The overall architecture is evaluated and classification results for the well-known BCI competition IV-2a dataset [17] are shown.

2. METHODOLOGY

2.1. Representing EEG Time Series

Event-related synchronization/desynchronization (ERS/ERD) in the motor cortex are the activity associated with motor-imagery. ERD is defined as the relative difference of energy before and after cue and equivalent to the subject modulating the amplitude of the recorded EEG signal during a motor-imagery task. A sample ERD/ERS map can be seen in Figure 1, showing the consistency of activity during a four-class motor imagery class for selected channels. In some tasks, such as left/right motor imagery, a time-consistent pattern of energy drop can bee seen whereas in feet/tongue imagery, a dynamic pattern of energy can be seen (see [8] for a detailed analysis). This dynamicity of EEG energy can be interpreted as change in the energy envelope of the signal and therefore, it is rational to use this envelope as the representation of the signal. Here, we propose to use the energy of the analytic signal computed by the Hilbert transform. In static energy feature algorithms, rather than fitting the absolute energy of each trial, the relative energy of each channel is considered as the feature. In our work, we also follow the practice by dividing the energy signals by the by sum energy of all channels, $\hat{x}_c / \sum_{c=1}^{N_C} \hat{x}_{ce}$, where $\hat{x}_c$ is a single channel energy envelope, $\hat{x}_{ce}$ is the channel energy and N_C is the number of channels. Eventually, the logarithm of the energy is taken to obtain a normal distribution before feeding into the network. From here on, we use the term dynamic energy for this representation and static energy for trial segment energy throughout the paper.

2.2. Designing the Architectures

Generally, when designing an architecture, the input format must be taken into account. Static energy features have a format of $x_e \in \mathbb{R}^{N_{sC} \times C}$, where C is the number of classes and N_{sC} is the selected channels. For dynamic energy, a format of $x_t \in \mathbb{R}^{T \times N_{sc} \times C}$ is considered, where T is the number of time samples in the representation. There is an option of concatenating the data on the class dimension, C, and then feeding into the network, which we have chosen to consider for static energy features.

In terms of architecture, a three-layered multilayer perceptron (MLP), is used for static energy features. Our simulations showed that a linear support vector machine (SVM) results in comparable performance relative to [3], hence, deepening the classifier would not have had significant improvement. For dynamic energy features, a convolutional neural network is designed with the following configuration:

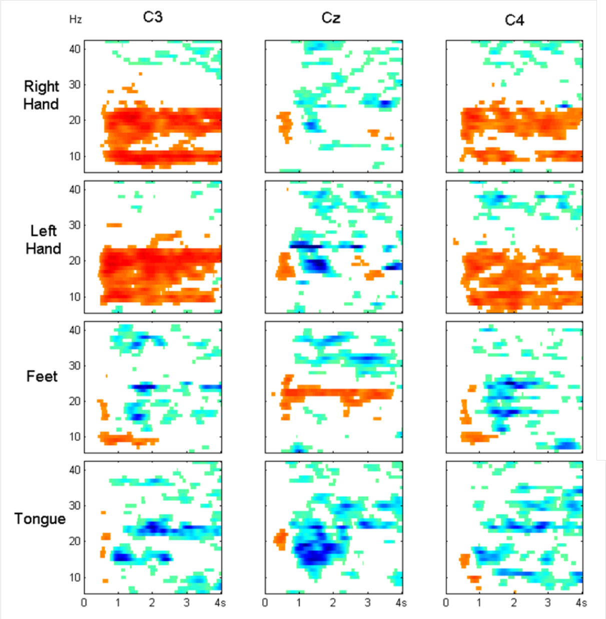

Fig. 1: Sample ERD/ERS plot for a four-class task in selected EEG channels. In color version, red indicates desynchronization and blue indicates synchronization. It is evident that patterns for left/right hand MI differs from feet/tongue MI, supporting the fact that a static energy is not sufficient for some tasks.

- *Parallel convolutional layers with a one dimensional kernel in time*. This kernel size means that the output of each activation function will save information for each channel.
- *Average pooling*. In CNN architectures for images, max pooling is used because of its spatial invariance-inducing property. But training the EEG data with max pooling resulted in poor results. Therefore, average pooling is used.
- *Convolutional layer with a one dimensional kernel in time*. In this stage the kernel is chosen as the same size as input and can be viewed as template matching.
- *MLP before and after concatenation*. MLP after the convolutional layers is an embedding that can be viewed as a sort of modified energy feature. After concatenating data from the convolutional and MLP layers above, the data is fed into another MLP for classification.

For both static and dynamic energy architectures, dropout regularization is used. Dropout, proposed by [10], is a semi-ensemble learning regularization that helps network to not overfit. The activation functions chosen for the two architectures is a Rectified Linear Unit ($ReLU$). Each architecture is trained individually using Stochastic Gradient Descent

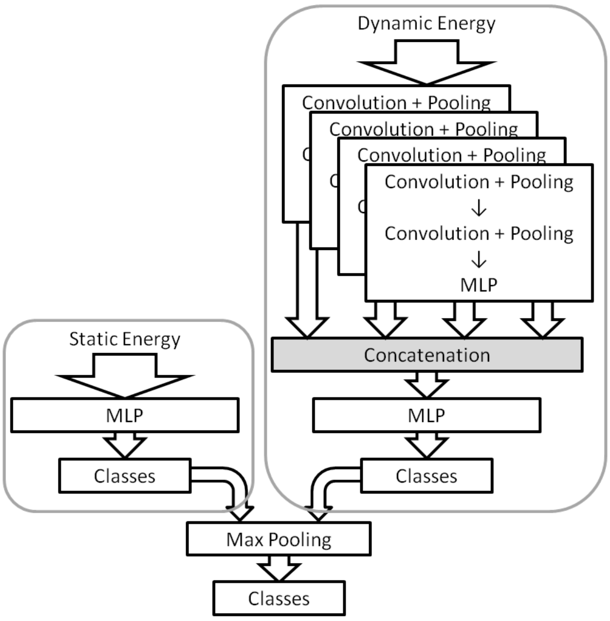

Fig. 2: Proposed Architecture. Grey boxes indicate the modules inside is trained independently.

(SGD) and a Negative Log-Liklihood (NLL) criterion. Final decision making is done on the maximum value of each class from each independent architecture. A depiction of the described architecture can be seen in Fig.2. Torch7 [13] is used to implement the architectures and training.

3. PREPROCESSING AND DATASET

Preprocessing, especially spatial filtering, is a crucial stage in EEG analysis. Here, we adopt the Filter-Bank CSP (FBCSP) [3] preprocessing stages for our data. Simplified, the FBCSP process is *a*) applying a bank of 9 filters from 4 to 40 Hz with a width of 4 Hz and *b*) Computing CSP for each frequency band using equation 1:

$$\max_{w \in \mathbb{R}^{N_C}} \frac{w^T C_1 w}{w^T (C_1 + C_2) w} \tag{1}$$

where w is the spatial filter, N_C is the number of channels and C_1 and C_2 are the channel covariance matrix of two distinct tasks (classes). The static energy features are calculated using $var(w^T X)$, with var being the variance operator on time and $X \in \mathbb{R}^{N_C \times T}$. Eventually, *c*) a feature selection algorithm, typically mutual information, is performed. Feature selection is carried out to select the most discriminative spatial filters and frequency bands.

BCI competition IV-2a [17] data is used as the dataset to evaluate the proposed architecture. This 9-subject dataset consists of 4-class ($C = 4$) MI data (Right hand, Left hand, Feet, Tongue) with each class having 72 samples for train and the test data having the same amount of data. We decide to follow the competition nature, therefore, test data is not used for training. The CSP algorithm is performed on a 0.5 to $2.5\,\mathrm{s}$ segment after cue and 4 pairs of spatial filters are picked for each frequency band. 8 pairs of features are selected from the 36 features as the input of the MLP network ($N_{sC} = 8$). This choice of spatial filters and features is based on [3]. The same spatial filters for each of the selected static energy features are applied on a segment of 0 to $3\,\mathrm{s}$ after cue and the log-energy representation described in section 2.1 is extracted. Preprocessing stages are implemented using MATLAB, version R2012a.

4. RESULTS

A support vector machine (SVM), applied on the static energy features, is used as the benchmark and implemented using LIBSVM [18]. To test significance, we choose the Wilcoxon signed-rank test. This non-parametric hypothesis test is used for cases where, due to limited samples, values do not follow a normal distribution. In our case, because of the limited number of subjects, this test is ideal. It should be noted that the reported p-value is not accurate due to the limited number of samples.

Accuracies are achieved by averaging over 10 models, Table 1. Results of independently trained static and dynamic energy networks and their combination are presented so these three results can be used to interpret which feature has more contribution to the accuracy of the combined network in each individual. Overall, based on the mean accuracy, there is an increase in classification performance and p-value of the Wilcoxon signed-rank test falls into the $p < 0.01$ range and therefore, shows the accuracy increase is significant. An in-

	SVM	MLP	CNN	CNN$\Vert$MLP
Sub1	79.16	75.69	78.82	**80.55**
Sub 2	52.08	48.96	53.47	**53.82**
Sub 3	83.33	75.35	82.64	**84.72**
Sub 4	62.15	**64.93**	60.76	64.58
Sub 5	54.51	52.08	**59.03**	**59.03**
Sub 6	39.24	39.93	43.75	**44.1**
Sub 7	83.33	82.99	82.64	**84.03**
Sub 8	82.64	84.72	83.68	**86.8**
Sub 9	66.67	67.36	81.25	**77.77**
mean	67.01	65.78	69.56	**70.60**
p-value	-	0.4065	0.2127	0.0091

Table 1: Classification results over BCI competition IV-2a test data. *SVM* column is the benchmark static energy features using an SVM classifier. *MLP* and *CNN* columns show results of linear and convolutional neural network using static and dynamic energy features respectively. *CNN*$\Vert$*MLP* shows the combined network using both-features. The p-value for the Wilcoxin rank-signed test can be in row *p-value*. Results show a significant increase in classification accuracy when a static and dynamic energy is used together.

teresting observation is the amount of increase in the accuracy value in some subjects, specifically subjects 5, 6 and 9. When checking the table for the reason of increase, it is evident that the convolution network and therefore, dynamic energy features are boosting performance in these subjects.

With the improved performance, we sought to look into in what aspects the CNN is contributing to the overall results. For this, we derive the individual classification accuracies which is defined as the number of correct classifications for one class over the number of samples in that class and also, the average confusion matrix over all subjects. These values can be seen in Table 2a and 2b. In 2a, we see class accuracies for the SVM given static energy feature and the CNN given dynamic energy features. It shows that the main strength of the CNN network is in classifying the feet and tongue class without compromising the classification of right and left classes much and therefore, balancing the accuracies and overally increasing performance. Table 2b also supports this by showing the overall reducion of confusion between tongue/feet and left/right classes.

5. DISCUSSION AND FUTURE WORK

We have proposed a parallel linear-convolutional architecture for the analysis of motor imagery data. In terms of static energy classification, our network does not outperform the benchmark methodology and although the average accuracy of dynamic energy features is higher, the increase is not consistent. Combining these two methods, however, results in consistent increases in almost all subjects. This confirms that energy dynamics contains discriminative information which cannot be seen in energy features alone and some subjects benefit from using energy dynamics as a feature.

Our algorithm is not perfect in several aspects and can be improved: heavy pre-processing of data, choice of architecture and network parameters. For these items, we propose the following research directions and future work:

- Network-based implementation of pre-processing stages
- Hyper-parameter optimization for parameter selection
- Modified architecture based on combination of static and dynamic energy

Furthermore, deep architectures, due to there high learning capacity, have gained their success by being trained on large amounts of data. Unfortunately, limitation on gathering data for individual subjects is a barrier in EEG research. If more subject-specific data can be collected or the current data augmented in a way that can capture the non-stationary nature of EEG data, better classification results can be achieved.

Classification aside, there is one other element in BCI research that is valuable: interpretability of learned algorithms. The value of these interpretations is clear when these systems are used for medical diagnosis and monitoring such as changes in spatial patterns in stroke patients [19] or ADHD [20]. In these applications, machine learning techniques and their interpretation are not only needed but vital. We hope to have a deeper investigation into the meaning of learned network parameters.

REFERENCES

[1] Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of neural engineering*, vol. 4, 2007.

[2] Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *Rehab. Eng., IEEE Trans. on*, vol. 8, no. 4, pp. 441–446, 2000.

[3] Kai Keng Ang, Zheng Yang Chin, Chuanchu Wang, Cuntai Guan, and Haihong Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Front. in neurosci.*, vol. 6, 2012.

[4] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek, "EEG data space adaptation to reduce intersession nonstationarity in brain-computer interface.," *Neural comp.*, vol. 25, no. 8, pp. 2146–71, Aug. 2013.

[5] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Chai Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI.," *IEEE transactions on bio-medical engineering*, vol. 58, no. 6, pp. 1865–73, June 2011.

[6] W. Samek, M. Kawanabe, and K.-R. Muller, "Divergence-based framework for common spatial patterns algorithms," *Biomedical Engineering, IEEE Reviews in*, vol. 7, pp. 50–72, 2014.

[7] Atieh Bamdadian, Cuntai Guan, Kai Keng Ang, and Jianxin Xu, "The predictive role of pre-cue EEG rhythms on MI-based BCI classification performance.," *Journal of neuroscience methods*, vol. 235, pp. 138–44, Sept. 2014.

[8] 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. 153–9, 2006.

[9] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted Boltzmann machines," in *International Conference on Machine Learning (ICML)*, 2010.

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[11] Alex Krizhevsky, "One weird trick for parallelizing

	Energy SVM					CNN				
	L	R	F	T	Ovr	L	R	F	T	Ovr
Sub 1	**0.92**	**0.79**	0.57	0.85	**0.79**	0.85	**0.79**	**0.64**	**0.88**	0.79
Sub 2	0.54	**0.38**	0.64	**0.43**	0.52	**0.64**	0.31	**0.81**	0.39	**0.53**
Sub 3	**0.90**	0.96	0.76	**0.78**	**0.83**	0.75	**0.97**	**0.81**	**0.78**	0.82
Sub 4	**0.46**	0.69	**0.76**	0.61	**0.62**	0.39	**0.75**	0.53	**0.76**	0.61
Sub 5	**0.81**	0.67	0.11	0.50	0.54	0.75	**0.88**	**0.17**	**0.57**	**0.59**
Sub 6	**0.50**	**0.58**	0.07	0.38	0.39	0.46	0.40	**0.28**	**0.61**	**0.44**
Sub 7	0.86	**0.94**	**0.63**	0.83	**0.83**	**0.88**	0.93	0.56	**0.94**	0.83
Sub 8	**0.93**	**0.82**	0.85	0.68	0.83	**0.93**	0.74	**0.86**	**0.82**	**0.84**
Sub 9	0.75	**0.89**	0.67	0.39	0.67	**0.78**	0.82	**0.85**	**0.81**	**0.81**
mean	**0.74**	**0.75**	0.56	0.60	0.67	0.71	0.73	**0.61**	**0.73**	0.70

(a) Class accuracy for each subject

SVM	L	R	F	T
L	**74.07**	15.28	5.86	4.78
R	16.05	**74.69**	4.48	4.78
F	10.19	15.28	56.17	18.36
T	13.12	15.12	11.27	60.49

CNN	L	R	F	T
L	71.30	13.12	5.71	9.88
R	11.11	73.15	7.25	8.49
F	6.33	8.80	**60.96**	23.92
T	8.49	8.33	10.34	**72.84**

(b) Average confusion matrix of all subjects for four classes

Table 2: Class accuracy and average confusion matrix for BCI competition IV-2a data. *SVM* is the benchmark FBCSP features using an SVM classifier. *CNN* show results of linear and convolutional neural network. *L, R, F* and *T* correspond to Left, Right, Feet and Tongue classes in the MI task. *Ovr* is the overall accuracy. It can be seen from *(a)* that the convolutional neural network has caused an increase of classification accuracy in Feet and Tongue classes for almost all subjects. In *(b)*, it can be seen that the network is reducing misclassification between classes.

convolutional neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[12] Min Lin, Qiang Chen, and Shuicheng Yan, "Network In Network," *arXiv preprint arXiv:1312.4400*, 2013.

[13] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011, number EPFL-CONF-192376.

[14] Y LeCun, L Bottou, Y Bengio, and P Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[15] Sebastian Stober, Daniel J. Cameron, and Jessica A. Grahn, "Using Convolutional Neural Networks to Recognize Rhythm Stimuli from Electroencephalography Recordings," in *Advances in Neural Information Processing Systems*, 2014, pp. 1449–1457.

[16] Hubert Cecotti, "A time–frequency convolutional neural network for the offline classification of steady-state visual evoked potential responses," *Pattern Recognition Letters*, vol. 32, no. 8, pp. 1145–1153, 2011.

[17] Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *Journal of Neural engineering*, vol. 4, no. 2, pp. R32, 2007.

[18] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[19] Kai Keng Ang, Karen Sui Geok Chua, Kok Soon Phua, Chuanchu Wang, Zheng Yang Chin, Christopher Wee Keong Kuah, Wilson Low, and Cuntai Guan, "A Randomized Controlled Trial of EEG-Based Motor Imagery Brain-Computer Interface Robotic Rehabilitation for Stroke.," *Clinical EEG and neuroscience*, pp. 1550059414522229–, Apr. 2014.

[20] Choon Guan Lim, Tih-Shih Lee, Cuntai Guan, D Sheng Fung, Yin Bun Cheung, S Teng, Haihong Zhang, and K Krishnan, "Effectiveness of a brain-computer interface based programme for the treatment of ADHD: a pilot study," *Psychopharmacol Bull*, vol. 43, no. 1, pp. 73–82, 2010.