Joint Optimization of Algorithmic Suites for EEG Analysis

Eder Santana, Austin J. Brockmeier, and Jose C. Principe

Abstract— **Electroencephalogram (EEG) data analysis algorithms consist of multiple processing steps each with a number of free parameters. A joint optimization methodology can be used as a wrapper to fine-tune these parameters for the patient or application. This approach is inspired by deep learning neural network models, but differs because the processing layers for EEG are heterogeneous with different approaches used for processing space and time. Nonetheless, we treat the processing stages as a neural network and apply backpropagation to jointly optimize the parameters. This approach outperforms previous results on the BCI Competition II - dataset IV; additionally, it outperforms the common spatial patterns (CSP) algorithm on the BCI Competition III dataset IV. In addition, the optimized parameters in the architecture are still interpretable.**

I. INTRODUCTION

In the analysis of functional brain activities, the electroencephalogram (EEG) plays an important role because it is a non-invasive, and non-obtrusive screening method with a long history of application. Several practical methods for reliably decoding brain activity have been developed, and these methods have enabled brain computer interfaces (BCI) systems [1] [2].

Over the years, neurophysiology knowledge has guided the development of new EEG analysis methods. For example, during movement tasks one can observe event-related synchronization and desynchronization (ERS and ERD respectively) on mu (8-13 Hz) and beta (15-30 Hz) rhythms in electrodes near the motor cortex area. Given this information, the most successful methods for motor imagery decoding filter EEG signals at those specific bands [3] [4].

The topology of neural processing in the cortex means that the spatial patterns of the neural activity are indicative of the type of brain processing. The common spatial pattern (CSP) [4] method and its extensions exploit differences in the spatial patterns between conditions to decode the condition type. Furthermore, spatiotemporal information is exploited by first band-pass filtering the signals before applying CSP. Finally, the power of different spatiotemporal projections are treated as features for general classifiers such as linear discriminant analysis [3] or support vector machines (SVM).

Algorithms that optimize the temporally filtering, spatial projection, and classification have additional free parameters that need to be selected, and often depend upon the subject. Furthermore, we note that any processing stages manually chosen by neuroscience knowledge may not provide the best classification performance. The maturing field of deep learning uses adaptively trained deep neural networks (DNN) [5] as a substitute for manually chosen processing. DNNs are usually large hierarchical networks composed of homogeneous multilayer perceptrons with more than one hidden layer in several different connection configurations, such as convolutive, recurrent, or fully connected. Given their generality, DNNs are powerful methods, but their practical performance is limited by the potential to overfit the training data, performing poorly on novel test data. Thus, DNNs only became useful when combined with pre-training and elaborated regularization schemes [5].

The main application of DNNs has been problems for which large amounts of data are available: speech recognition, image classification [7], and natural language processing [5]. They have not been applied to small-scale problem, such as EEG where the number of subjects and trials is limited. On these small-scale problems, the learned parameters are still interpretable and are useful for the investigator, whereas the hierarchical, homogenous processing layers in deep learning are not readily interpretable.

Given these problems, we propose a strategy for "interpretable" deep learning for EEG classification. The strategy consists of framing the existing EEG processing pipelines as a specific deep learning network, using specific algorithms (e.g. CSP) to provide a pre-trained initialization, and then optimizing the overall algorithmic suite for each patient and condition. Using a fixed cost function on the output of the neural network, backpropagation of the cost function's gradient enables the joint optimization of all the algorithmic stages.

The remainder of this paper is organized as follows: initially, we introduce the neural network architecture and comment on its advantages and disadvantages, then we demonstrate the algorithm's record breaking performance on the BCI Competition II dataset IV and show its performance on 5 other human EEG datasets provided by the BCI competition III dataset IV-a, and finally, we propose future research directions as well as recommendations for the practitioner using the proposed method.

II. PROPOSED FRAMEWORK

Based on existing EEG motor imagery literature, mainly inspired by the CSP method [4], we propose an analysis framework consisting of three steps. The first layer is a temporal filter, which can be implemented as a convolutional layer for each channel. In neural networks, convolutional layers can be interpreted as regular fully connected layers with shared parameters [6]. The second step consists of a spatiotemporal projection of the filtered signal. This can be

Research supported by UF Opportunity Fund, Grant #00098693, 2014.

E. Santana, A. J. Brockmeier and J. C. Principe are with the University of Florida, Department of Electrical and Computer Engineering, Gainesville, FL, 32611 USA. (e-mail: principe@cnel.ufl.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

implemented as a single fully connected feedforward layer or two divided layers, one for spatial projection and another for temporal projection. The third step is the classification layer, which assigns labels to the input based on the features learned by the previous layers. We can implement this layer as a simple logistic regression layer, as a multilayer perceptron (MLP), or as a SVM.

Given this architecture, all the layers can be optimized together using backprogation. All the parameters are tuned together for the final goal of classification. This is the main advantage of the proposed framework versus methods that define each layer independently based on heuristic knowledge.

Another method that also proposes a unified framework for EEG analysis is the Second-order Bilinear Discriminant Analysis (SOBDA) [7], where the spatial filtering and spatiotemporal projections are all well defined under a single and elegant equation. Nevertheless, this method is not as flexible as the one proposed here as it is limited by only linear operations. Also, SOBDA does not allow us to exploit heuristics for training neural networks that we discuss in Section IV.

The main drawback of our method is also related to its flexibility. If we try to learn an arbitrarily large DNN using small datasets, the method will most certainly overfit the training data. Thus, in order to verify the validity of the present framework, we focus on strategies that can deal with limited number of samples. For instance, the BCI competition II contains only 316 training examples, each example composed of 50 samples for 28 channels, which means that if we try to train a neural network with a single neuron and 5028 connections we would already have more parameters than examples. Instead, what we propose for this case is an architecture based on and initialized by the regular CSP pipeline [3]. We represent this architecture in Fig. 1.

Mathematically, the proposed architecture can be described as follows. Given a multichannel EEG matrix $\mathbf{X}_{channel \times time}$, spatial projection $\mathbf{w}$, temporal projection $\mathbf{v}$, filter $\mathbf{h}$, and the classifier weights $\mathbf{u}$, we have

$$y = \mathbf{u}^T f\left(\left[\mathbf{w}^T(\mathbf{X} \star \mathbf{h})\right]^2 \mathbf{v}\right) \tag{1}$$

where is the value that will be plugged into an activation function, where $f(\cdot)$ is the logistic function, and then feed forward to a cost function. Here we used the cross-entropy cost function, which is standard in logistic regression [5]. Also, the same filter is used on every row of the matrix. The following gradients can be plugged in the backpropagration chain rule for adaptation [7]:

$$\frac{\partial y}{\partial \mathbf{w}} = \mathbf{u}^T \frac{2}{\left[\mathbf{w}^T(\mathbf{X} \star \mathbf{h})\right]^2 \mathbf{v}} (\mathbf{X} \star \mathbf{h})\mathbf{v} \tag{2}$$

$$\frac{\partial y}{\partial \mathbf{v}} = \mathbf{u}^T \frac{1}{\left[\mathbf{w}^T(\mathbf{X} \star \mathbf{h})\right]^2 \mathbf{v}} \left[(\mathbf{X} \star \mathbf{h})^T \mathbf{w}\right]^2 \tag{3}$$

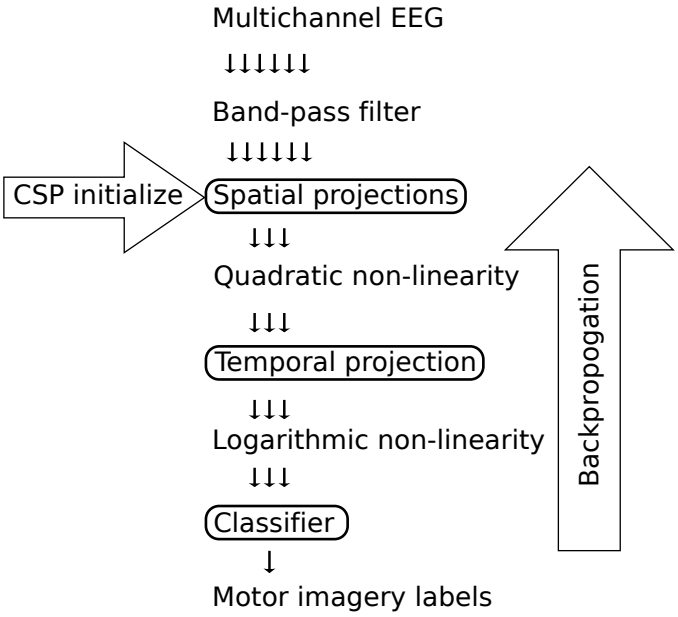

Fig. 1. Diagram of the deep neural network architecture for classifying motor imagery from EEG. The adapted layers are circled.

III. EXPERIMENTS WITH BCI COMPETITION DATASETS

Here we apply the proposed framework to human EEG datasets.

A. BCI competition II, dataset IV

The first dataset, proposed in BCI competition II [8], was recorded from a normal subject during a no-feedback session, while the subject pressed keyboard keys using either hand. Signals were recorded at 1000 Hz and down-sampled to 100 Hz with a band-pass filter between 0.05 and 200 Hz. Each trial contains 50 samples per channel. The goal is identify which hand the subject was using for each press. The recorded signal is divided into 416 epochs: 316 are for training and the remaining 100 are for testing.

To classify this data, we used the general strategy proposed in Fig. 1. We separated part of the training set for cross-validation. We fixed the learning rate η, and we used early stopping [9] to determine how many epochs M, the network needs to converge. Given those values, we trained the network using the full training set for M epochs.

Our best results were obtained by using a fixed IIR filter with pass-band between 8 and 30 Hz; initializing the spatial projections with 5 CSP projections; and setting the temporal projection to a single vector of ones. Note, as a rule of thumb, we choose the CSP projections as those that correspond to eigenvalues that are one standard deviation larger than the eigenvalues mean.

The classification layer was randomly initialized from a Gaussian distribution. We trained the network using $\eta = 0.01$ and $M = 1700$. For each training epoch, we use 4 random mini-batches of 1/4 of the training data and update the weights using the stochastic gradient of each mini-batch. The resulting classification rate is shown in Table I along

with comparative results from other methods applied to the same dataset, including a multilayer perceptron (MLP) with a fully connected hidden layer.

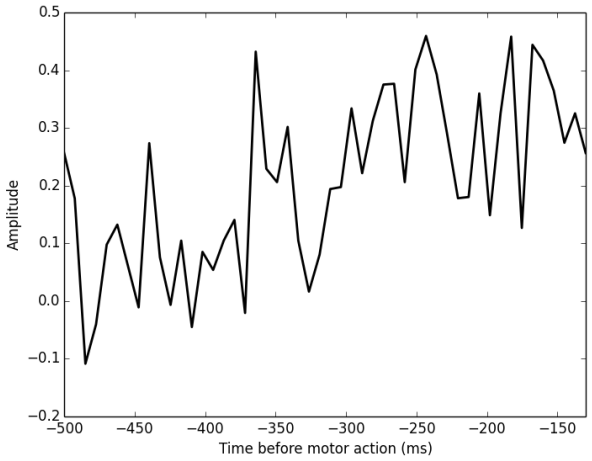

Fig. 2. Temporal projection learned through backpropagation. The samples nearer the motor action have higher magnitude weights.

Note that our method was initialized with the solution that won the actual competition and reduced its final error from 16% to 10%. This result is also 3 percentage points better than the results obtained with SOBDA [7]. The code to reproduce our results is freely available online[1].

TABLE I
CLASSIFICATION ACCURACY (% CORRECT) OF SEVERAL METHODS APPLIED TO BCI COMPETITION II DATASET IV.

DNN[a]	CSP [3]	SOBDA [7]	MLP[b]
90%	84%	87%	65%

[a]Method proposed in the present paper.
[b]Multilayer perceptron trained with a fully connected hidden layer.

To understand what the network learned, we investigated its coefficients after adaptation. In Fig. 2, we show the vector learned for the temporal projection emphasizes the time points appearing closer to the motor action time. To verify that these were important for classification, we used only the last 30 samples in the original CSP framework and found that this improves its performance by 2 percentage points. However, we could only achieve the reported performance when we adapt the layers together.

B. BCI competition III, dataset IV-a

This dataset consists of EEGs of 5 healthy human subjects (aa, al, av, aw, ay). The number of training/testing sets for each subject were 168/112, 224/56, 84/196, 56/224, and 28/252, respectively. Neural network based approaches are not expected to perform well on such small datasets, but, again, we initialized our results with CSP and fine-tuned the parameters using backpropagation. We used the same approach and parameters settings, $\eta = 0.01$ and $M = 1700$, as the last dataset. Results are in Table II. The average error was reduced from 27% to 17%, but the error reduction was more modest than the previous dataset due to the smaller training samples. Throughout, the DNN performed better than or equal to CSP

[1]https://github.com/EderSantana/DeepEEG

TABLE II
CLASSIFICATION ACCURACY (% CORRECT) FOR 5 SUBJECTS OF BCI COMPETITION III DATASET IV-A.

	aa	al	av	aw	ay
DNN	**80.4**%	**100**%	**69.4**%	**93.7**%	**71**%
CSP	63.4%	**100**%	**69.4**%	91.5%	67%

IV. DISCUSSIONS

In this section we comment on the machine learning aspects that should be of most interest for those further developing the present framework.

A. Data augmentation

Here, we dealt with the problem of the small dataset by pre-training our neural network with a previously defined method. On the other hand, future research should consider data augmentation, a technique that uses added variations of training samples to increase robustness. Data augmentation is explored in image processing where possible augmentations consider the possible translations and rotations of the main object of interest in the scene [10]. For EEG, this can be done by breaking each trial into multiple temporal windows and classifying the windows independently. Since the exact timing may vary among trials, augmenting the data in time could possibly make the network more robust to temporal shifts.

B. Regularization techniques

Training large networks requires parameter regularization to avoid overfitting. Our reported results used early-stopping, which can also be interpreted as an L2-norm regularization [9]. An L1-norm norm could be considered for automatic channel selection or for selecting the number of CSP projections.

C. Parameter initialization

The recent success of neural networks for challenging problems is based on better initialization techniques that first try to model the distribution of the data [10] using unsupervised learning. Here, we initialized our network with parameters proposed by CSP, which is based on a discriminative statistical test. Using discriminative pre-traning has not been previously explored by the deep learning literature.

D. Training band-pass filtering layer

Here our best results were obtained using a fixed band-pass filter. The filter was a zero-phase IIR. Thus, its implementation goes beyond what can be done with the convolution layer, which defines a FIR. Such filter can only be implemented using recurrent connections, but as training recurrent systems is beyond the scope of the present paper, we leave this for future research.

V. CONCLUSIONS

In this paper we proposed a strategy for the joint optimization of multiple signal processing stages involved EEG motor imagery classification. Basically, we reinterpreted the main steps of the conventional EEG analysis pipeline as layers in a neural network. This new approach allowed us to wrap the previously ad hoc stages in EEG processing using neural networks adapted with backpropagation. We investigated this framework for datasets with limited number of samples. To do so, we initialized our network with the architecture and parameters provided by CSP and obtained results that were better or equal to those obtained with CSP alone. Additionaly, we showed that the deep neural network could still have interpretable components.

REFERENCES

[1] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, Dec. 1988.

[2] J. C. Principe, "The Cortical Mouse: A Piece of Forgotten History in Noninvasive Brain Computer Interfaces," *Pulse, IEEE*, vol. 4, no. 4, pp. 26–29, 2013.

[3] Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao, and F. Yang, "BCI competition 2003-data set IV:An algorithm based on CSSD and FDA for classifying single-trial EEG," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1081–1086, Jun. 2004.

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

[5] Y. Bengio, "Learning Deep Architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[6] 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.

[7] C. Christoforou, R. Haralick, P. Sajda, and L. C. Parra, "Second-Order Bilinear Discriminant Analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 665–685, Mar. 2010.

[8] B. Blankertz, K.-R. R. Müller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer, "The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials," *Biomedical Engineering, IEEE Transactions on*, vol. 51, no. 6, pp. 1044–1051, Jun. 2004.

[9] J. C. Principe, N. R. Euliano, and W. C. Lefebvre, *Neural and Adaptive Systems: Fundamentals Through Simulations*. John Wiley and Sons, 2000.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," Jun. 2012.