Investigating Deep Learning for fNIRS based BCI

Johannes Hennrich, Christian Herff, Dominic Heger and Tanja Schultz

Abstract— **Functional Near infrared Spectroscopy (fNIRS) is a relatively young modality for measuring brain activity which has recently shown promising results for building Brain Computer Interfaces (BCI). Due to its infancy, there are still no standard approaches for meaningful features and classifiers for single trial analysis of fNIRS. Most studies are limited to established classifiers from EEG-based BCIs and very simple features. The feasibility of more complex and powerful classification approaches like Deep Neural Networks has, to the best of our knowledge, not been investigated for fNIRS based BCI. These networks have recently become increasingly popular, as they outperformed conventional machine learning methods for a variety of tasks, due in part to advances in training methods for neural networks. In this paper, we show how Deep Neural Networks can be used to classify brain activation patterns measured by fNIRS and compare them with previously used methods.**

I. Introduction

In the last few years, functional Near Infrared Spectroscopy (fNIRS) has become an emerging technology for optical brain activity measurement that can be used in non-invasive Brain-Computer Interfaces (BCIs). However, there are still no common standards for feature extraction and classification in single trial fNIRS analysis. Often, rather simple methods are used for feature extraction, such as calculating the mean of the measured hemoglobin concentrations within a given time window [1]. Just as for feature extraction, classification methods used are comparably simple. Recently, Bauernfeind et al. [2] compared different classifiers for fNIRS BCIs and recommended using shrinkage LDA. However, their evaluation did only include classifiers that are well established in Brain Computer Interface research, but did not investigate Deep Neural Networks. Therefore, it is still unclear, whether more complex classification schemes, such as Deep Neural Networks (DNN), can be used to exploit additional information that may be hidden in the non linear dynamics of the hemodynamic responses that typically occur in fNIRS data.

In this paper, we investigate the suitability of deep learning, i.e. artificial neural networks with many layers, for fNIRS classification. Deep learning methods have lately regained popularity, as they have shown impressive results for many different classification problems. However, to the best of our knowledge, no studies have used deep neural networks for the classification of fNIRS data, before. We discuss on how to design deep neural networks for fNIRS and evaluate their classification performance in comparison to traditional approaches.

All authors are with the Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany johannes.hennrich@student.kit.edu

A. Functional Near-Infrared Spectroscopy

Functional Near-Infrared Spectroscopy (fNIRS) is a relatively new non-invasive method to capture hemodynamic responses to cortical activity. These responses can be measured with optical sensors which are cheap and portable and allow for high spatial resolution. The underlying metabolic effects fNIRS is based on are the same effects fMRI uses. fNIRS makes use of the fact that oxygenated hemoglobin absorbs near infrared light differently than deoxygenated hemoglobin. This is achieved by using two different wavelengths in the near infrared part of the electromagnetic spectrum to measure concentration levels of oxygenated and deoxygenated hemoglobin and derive the hemodynamic responses to cortical activity. The by far most prominent modality for non-invasive BCIs is electroencephalography (EEG), which derives cortical activity from electric signals measured on the scalp. In comparison to fNIRS, EEG has a higher temporal resolution because hemodynamic effects are inert and appear with some delay. On the other hand, fNIRS provides higher spatial resolution and does not require electrolyte gel, which makes fNIRS more comfortable to wear and faster to put on, especially on bald regions like the forehead.

Many fNIRS studies aim to discover patterns in the hemodynamic responses by averaging over many trials. For research on online BCIs a single-trial analysis of fNIRS activations is necessary. In [3] and [4] activations of a basic motor imagery task were classified online and on single-trial using brain activity produced by motor imagery on the motor cortex. In contrast to these BCIs for direct control, a passive investigation of memory load in fNIRS applying the n-back paradigm was conducted in [5].

B. Deep Learning

To classify the activity captured by fNIRS, a variety of machine learning approaches have been evaluated. In [4] both Support Vector Machines (SVM) and Hidden Markov Models (HMM) yielded significant classification accuracies for motor imagery. In [6] Support Vector Machines were used for a four-class classification task using the mean, median, range and slope of a trial as features. The slope was also used in [7], in combination with a Linear Discriminant Analysis (LDA). [8] and [5] also used LDA classifiers in combination with very simple features.

While all the mentioned classifiers have successfully been used for a long time, Deep Neural Networks have only recently gained popularity as highly efficient training procedures were developed. Training these deep architectures is complicated, as the process gets slower the more layers are

978-1-4244-9270-1/15/$31.00 ©2015 IEEE

added. To make training feasible, [9] presented a greedy, unsupervised algorithm which uses Restricted Boltzmann Machines (RBM) to pre-train the network and find good initial weights that speed up the actual training. The pre-training is done layer by layer which allows it to be fast even for very deep networks ([10], [11]). Deep neural networks pre-trained with RBMs were used for the classification of phones in automated speech recognition in [12] and [13] and outperformed all previous methods. With the same pre-training, a generative neural network yielded impressive results by learning and reconstructing pictures of handwritten images ([14]). In 2012 [15] showed how these handwritten images can be classified using a deep neural network and achieved a relative improvement of accuracy of 41% compared to conventional methods. Here, we investigate the performance of deep neural networks as classifiers for fNIRS based BCI.

II. Materials and Methods

A. Data Corpus

The data corpus we used for evaluation is from a BCI experiment conducted in [8]. The experiment consisted of three different mental tasks, namely mental arithmetics (MA), word generation (WG) and mental rotation (MR). In mental arithmetics, the subjects were asked to continuously subtract a given number from another given number, in word generation, they had to find words starting with the specified letter and in mental rotation, 3D objects were displayed which the participants had to imagine to rotate around an axis. Tasks were displayed for 10 seconds and followed by a rest interval of 15 seconds. After 30 randomly chosen rest intervals, 10 second long relax trials (R) were included.

To measure the brain activity in a subject's prefrontal cortex, we used a 8-channel fNIRS headset (Oxymon Mark III, Artinis) attached to the forehead.

In fNIRS, each channel produces one value for oxygenated and deoxygenated hemoglobin concentrations, the recorded signal is thus 16-dimensional. At a sampling rate of 10 Hz this resulted in 100 samples per trial for each of the 16 dimensions. A total of 30 trials for each of the three mental tasks and the relax task were recorded. Trials are extracted based on experiment timing and are labeled corresponding to the type of mental task.

B. Preprocessing

To remove common biological and technical artifacts from the measured signal and optimize it for neural networks, there are some important preprocessing steps. To remove slow trends, we subtracted the mean of the surrounding 240s window from each sample in every channel, this has been successfully applied to fNIRS signals in [5]. To attenuate spikes and high-frequency artifacts, in particular the subject's heartbeat, we lowpass filtered with a cutoff frequency of 0.5 Hz [5] using an elliptic IIR filter with filter order 6. After downsampling to 1 Hz to reduce the dimensionality of the data, the 16 channels were stacked so each trial formed a 160 dimensional vector. Since previous studies ([16], [12]) show that z-normalization reduce neural network training duration and increase robustness, we normalized the data to zero mean and unit variance.

C. Deep Neural Network

We used deep artificial neural networks to classify the data.

As deep neural networks have shown great potential at learning relevant features from raw data ([17], [11], [14]) we refrained from using a tailor-made feature extraction and trained the network directly on the preprocessed data.

Instead of initializing the network with random weights, we pre-trained the network in a layer wise, unsupervised manner using restricted boltzmann machines which can speed up training and lead to better generalization [14]. The pre-trained network was then fine tuned by minimizing the cross-entropy error with the method of conjugate gradients (CG). Using CG to train neural networks has been proposed in many publications and has several advantages over standard backpropagation, including a faster convergence and automated estimation of the learning rate ([18], [19]). For our particular problem, CG training was faster, more stable and resulted in higher classification accuracies than backpropagation.

The activation functions we used are linear functions for the input layer and softmax for the output. For the hidden layers, we decided to use logistic activation functions as they are the de facto standard for nonlinear neural networks and yield solid results for most purposes. Choosing the amount of hidden layers and units is a more difficult task. As the optimal network topology is highly dependent on problem type and the training data distribution, there is no universal procedure to derive these numbers. In our case, we only have 27 training samples per class if evaluated in a 10-fold cross-validation. Classic machine learning theory suggests using very small models with few parameters if there is such little training data available. However [20] showed that deep neural networks are able to learn models with much more parameters than available training samples. A common problem that arises when large models are trained with few training samples is overfitting. Overfitting is present if bad generalization reduces the test accuracy, however there are different solutions that address this problem ([21]). To settle on a network topology we ultimately ran a grid search which estimated the optimal size and amount of hidden layers. This way, we found that two hidden layers with 300 and 40 units, respectively, yielded the best results. The amount of units in the input- and classification layer is determined by the dimensionality of the input data and the amount of classes and therefore does not require tuning. To find out how the depth of the network affects its performance we employed two additional networks with one layer (40 units) and three hidden layers (300-100-100 units).

D. Evaluation

For the evaluation of its classification performance, we trained and tested the networks in a 10-fold cross-validation.

This procedure was repeated for all 10 subjects and classification tasks. The classes we tried to discriminate were the three mental tasks (MA, WG, MR) against relax (R) and the mental tasks against each other, which adds up to a total amount of 6 binary classification problems per subject. We compared our classification results with those from [8] which were achieved using a standard LDA classifier (LDA + Feat.Extr.). It is important to note that the LDA classifier was trained on custom-built features (slope on HbR/HbO) which require expert knowledge about fNIRS signal and the problem domain. The deep neural network on the other hand runs directly on the preprocessed data, thus providing a more generic solution.

For a more appropriate comparison we cross-validated a shrinkage-LDA on the same data as the deep neural network as a third method. It has been shown in [22] that regularized LDA classifiers perform well on classification tasks where only little training data is available for high dimensional feature spaces. An optimal shrinkage parameter was estimated using the analytic method proposed in [23]. This method uses the same feature space as the neural networks and does not require expert knowledge either.

III. Results

A. Classification Results

All classification accuracies achieved by the three deep neural networks, the LDA with feature extraction and the shrinkage-LDA are presented in Figure 1. Each bar represents the classification of one mental task against relax or another mental task using one of the five methods. All

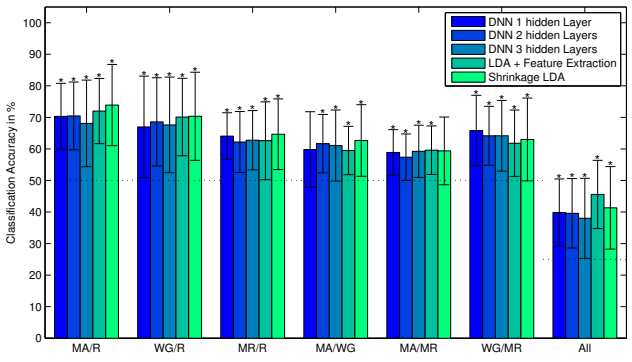

Fig. 1. Classification accuracies of different tasks averaged over the 10 subjects. Whiskers denote standard deviations. Results marked with $*$ are significantly better ($p < 0.01$) than naive classification (dotted lines).

classification results were found to be significantly better than chance level ($p < 0.05$).

The overall classification accuracies for the different methods averaged over subjects and tasks are shown in Table I. Paired t-tests showed no significant differences between classification accuracies of the methods for all classification conditions ($p > 0.05$), except for when comparing the shrinkage-LDA with the 1- and 3-hidden layer DNN ($p = 0.02$ and $p = 0.005$).

TABLE I

Overall accuracies for different classification methods.

DNN (1 hid.)	DNN (2 hid.)	DNN (3 hid.)	LDA Feat.Extr.	Shrinkage LDA
63.3%	64.1%	62.1%	64.3%	65.7%

To investigate the variances of the classification accuracies, we searched for systematic differences between the subjects. Using Pearson's correlation coefficient, we calculated the correlation of classification rates (averaged over 10 folds) for each two classification methods and each task across the 10 subjects (Table II) These high correlations (mean r: 0.61 for Neural Network vs. LDA + Feat. Extr., 0.89 for Neural Network vs. Shrinkage LDA, 0.70 for LDA + Feat. Extr. vs. Shrinkage LDA) suggest that there are significant variations in the data quality of the different subjects while all three classifiers run at a similar performance.

TABLE II

Pearson's correlation coefficients of classification accuracies over the subjects for each combination of classification methods and tasks.

	Neural Network LDA + Feat.Extr.	Neural Network Shrinkage LDA	LDA + Feat.Extr. Shrinkage LDA
MA/R	0.878	0.905	0.790
WG/R	0.869	0.966	0.861
MR/R	0.624	0.816	0.828
MA/WG	0.775	0.929	0.814
MA/MR	0.199	0.912	0.379
WG/MR	0.322	0.852	0.527

We also used the neural networks to classify all three mental tasks and relax against each other. In this 4-class task we achieved a classification accuracy of 40% using the 2-hidden layer DNN, which is significantly better ($p = 0.0012$) than naive classification (25%). With 45% for LDA with feature extraction and 41% for shrinkage LDA, the other methods yield higher, albeit not significantly better results than the neural network ($p = 0.24$ and $p = 0.75$). With the 1- and 3-hidden layer DNN we achieved accuracies of 39% and 38%.

Overall LDA-based methods outperformed all configurations of Deep Neural Networks.

B. Test for Overfitting

A simple but powerful tool to visualize the fine tuning training procedure and discover issues like overfitting is to test the classification accuracy of the network after each epoch of training. Testing is done twice after each epoch, once on the training set and on the test set, resulting in two time lines which can be plotted as a function of the amount of training epochs. Figure 2 shows classification accuracies depending on training epochs for subject 6 for mental arithmetics against relax. Comparable behavior was observed for all subjects. One can see that after several epochs the classification accuracy on the training set saturates on 100%, which denotes that all training samples are getting

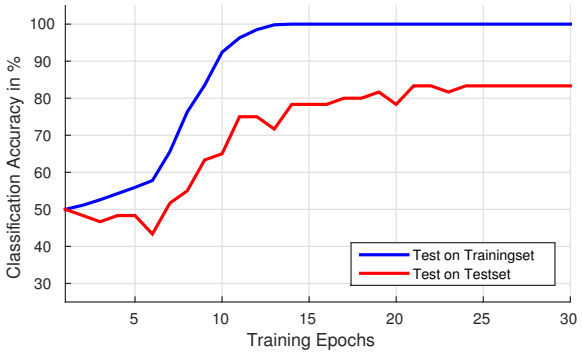

Fig. 2. Classification accuracies for subject 6 for mental arithmetics against relax when evaluated on training- and test set as a function of the amount of training epochs. Comparable behavior was observed for all subjects.

classified correctly. Similarly the classification accuracy on the test set rises and after several epochs saturates on a value of about 84%, which equals the final test accuracy of the network.

If after reaching a local maximum the test error would start to drop again, the plot would be a clear indicator for overfitting. In that case one should either reduce the complexity of the model by lowering the amount of hidden units or use a stopping criterion which monitors the training process and terminates it after the optimal amount of epochs as explained in [21]. Other possible scenario is that the training accuracy does not saturate, which can be caused by poor initial weights or too few training epochs. Further, it is possible that the training accuracy does saturate, but at a value lower than 100%. This is not necessarily a problem but can indicate that the model is too small and more hidden neurons might improve both training- and test accuracy.

The fact that we do not see any signs of overfitting, but a perfect classification of the training data suggests that far too few training samples are present.

IV. Conclusions

In this study, we showed how deep learning methods can be successfully used for building BCIs based on fNIRS. We achieved classification accuracies for the discrimination of different mental tasks that are comparable to those produced by conventional methods.

Even though the neural network did not yield higher classification rates, it is an promising approach as it does not require a tailor-made feature extraction and expert knowledge about the problem domain. Comparing the neural network with an optimally regularized LDA, also operating on the raw data, we found the shrinkage LDA to yield superior, albeit not significantly better results. By using networks with different amounts of hidden layers we showed that deeper networks do not perform better on this particular task. A possible reason for this is, presumably, the limited training data. Deep neural networks are known to require large training sets for successful learning. We thus recommend to use regularized classifiers if little training data is available.

References

[1] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous Recognition of Affective States by Functional Near Infrared Spectroscopy Signals," *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 832–837, Sept. 2013.

[2] G. Bauernfeind, D. Steyrl, C. Brunner, and G. R. Muller-Putz, "Single trial classification of fnirs-based brain-computer interface mental arithmetic data: A comparison between different classifiers," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014.

[3] S. M. Coyle, T. E. Ward, and C. M. Markham, "Brain–computer interface using a simplified functional near-infrared spectroscopy system," *Journal of neural engineering*, vol. 4, no. 3, p. 219, 2007.

[4] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer, "Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface," *NeuroImage*, vol. 34, no. 4, 2007.

[5] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS." *Frontiers in human neuroscience*, vol. 7, no. January, p. 935, Jan. 2014.

[6] A. M. Batula, H. Ayaz, and Y. E. Kim, "Evaluating a four-class motor-imagery-based optical brain-computer interface," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 2000–2003.

[7] S. D. Power, A. Kushki, and T. Chau, "Intersession consistency of single-trial classification of the prefrontal response to mental arithmetic and the no-control state by nirs," *PloS one*, vol. 7, no. 7, p. e37791, 2012.

[8] C. Herff, D. Heger, F. Putze, J. Hennrich, O. Fortmann, and T. Schultz, "Classification of mental tasks in the prefrontal cortex using fNIRS." *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 2160–3, 2013.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets." *Neural computation*, vol. 18, no. 7, pp. 1527–54, July 2006.

[10] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy Layer-Wise Training of Deep Networks," no. 1, 2007.

[11] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] A.-r. Mohamed, G. Dahl, and G. Hinton, "Deep Belief Networks for phone recognition," pp. 1–9, 2009.

[13] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," no. c, pp. 1–10, 2010.

[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks." *Science (New York, N.Y.)*, vol. 313, no. 5786, pp. 504–7, July 2006.

[15] D. Cirean, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," no. February, p. 20, Feb. 2012.

[16] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription," *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 24–29, Dec. 2011.

[17] Y. Bengio and N. Y. Georgios, "Learning Deep Physiological Models of Affect," no. April, pp. 20–33, 2013.

[18] J. L. Blue and P. J. Grother, "Training feed-forward neural networks using conjugate gradients," in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1992, pp. 179–190.

[19] J. Martens, "Deep learning via Hessian-free optimization," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 735–742.

[20] G. Hinton, "A Practical Guide to Training Restricted Boltzmann Machines," 2010.

[21] W. S. Sarle, "Stopped training and other remedies for overfitting," in *Proc. of the 27th symposium on the interface of computing science and statistics*, 1995, pp. 352–360.

[22] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components–a tutorial." *NeuroImage*, vol. 56, no. 2, pp. 814–25, May 2011.

[23] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, Feb. 2004.