Healthcare of the Future 2022
T. Bürkle et al. (Eds.)

doi:10.3233/SHTI220333

Classifying Numbers from EEG Data - Which Neural Network Architecture Performs Best?

Sugeelan SELVASINGHAM[a], Kerstin DENECKE[a,1]

[a] *Bern University of Applied Sciences, Bern, Switzerland*

Abstract. This paper presents a comparison of deep learning models for classifying P300 events, i.e., event-related potentials of the brain triggered during the human decision-making process. The evaluated models include CNN, (Bi | Deep | CNN-) LSTM, ConvLSTM, LSTM + Attention. The experiments were based on a large publicly available EEG dataset of school-age children conducting the "Guess the number"-experiment. Several hyperparameter choices were experimentally investigated resulting in 30 different models included in the comparison. Ten models with good performance on the validation data set were also automatically optimized with Grid Search. Monte Carlo Cross Validation was used to test all models on test data with 30 iterations. The best performing model was the Deep LSTM with an accuracy of 77.1% followed by the baseline (CNN) 76.1%. The significance test using a 5x2 cross validation paired t-test demonstrated that no model was significantly better than the baseline. We recommend experimenting with other architectures such as Inception, ResNet and Graph Convolutional Network.

Keywords. Convolutional neural networks, EEG, P300, Classification

1. Introduction

Brain-computer interfaces (BCI) enable communication without muscle activity based on brain signals measured with electroencephalography (EEG). The P300 is an event-related potential that is triggered during the decision-making process of a human. P300-based BCIs have gained attention in recent years and are considered one of the most important BCI categories [1]. Compared to other BCI paradigms, P300 BCIs are relatively fast, effective for most users, straightforward and require virtually no training of subjects. The challenge is to classify the P300 events with sufficient accuracy to enable a good communication. Deep learning and neural networks have been applied to this classification task. Vareka created a Convolutional Neural Network (CNN) model and trained it with EEG data [2]. CNN is an artificial intelligence method and is often used in image classification. The author was able to achieve an average classification accuracy of 62.18% using CNN. He also tested Recurrent Neural Networks (RNN) [3]. The RNN architecture is often used for classifying time series, i.e. also for EEG. Since the accuracy for guessing the number using P300 event streams with only three channels is still

[1] Corresponding Author, Kerstin Denecke, Bern University of Applied Sciences, Institute for Medical Informatics, Quellgasse 21, 2502 Biel, Switzerland; E-mail: kerstin.denecke@bfh.ch.

insufficient for reliable BCI, the objective of this work is to develop and test different deep learning architectures to achieve a better accuracy. A practical application of the algorithm could support individuals who are unable to communicate verbally in their interaction with computers. The main contribution of this work is a comprehensive analysis of multiple neural network architectures for the classification of numbers from EEG data (P300 event streams).

2. Material and methods

P300 dataset. We used the P300-dataset described by Mouvcek et al. [4] collected using the "guess the number" experiment. Participants are asked to pick a number between 1 and 9. During the following EEG measurement phase, the individual is stimulated with these numbers. He or she is silently counting the number of occurrences of the selected number. The target number is supposed to trigger the P300 response. After the experiment, this number is revealed and compared with the guess of the experimenters observing averages EEG waveforms [4]. The dataset used in this paper was collected in experiments with 250 participating school-age children which were carried out in elementary and secondary schools in the Czech Republic. Electroencephalographic data from three EEG channels (Fz, Cz, Pz) and stimuli markers were stored. Since we want to compare our results to the work published by Vareka [2], we used the dataset as prepared by the author without additional preprocessing. Before classification, the data were randomly split into training (75 %) and testing (25 %) sets. Training data was additionally split into 75% training and 25% validation set.
Experimental setup. We used two evaluation strategies to test the performance of the models: K-fold cross validation and Monte Carlo Cross Validation (MCCV [5]). In this validation method, the entire data set or the number of epochs is randomly divided into test and training data. Epochs are randomly divided into test and training data. With each iteration, the division of test and training data varies. Due to the random division, the same data components can serve as test data several times in some iterations. K-fold cross validation splits training data into K equal parts. To assess the quality of classification, we calculated the following metrics: AUC, Precision, Recall, Accuracy and Duration for training and classification. Results are tested for statistical significance using the 5x2 Paired t-Test. The following hyperparameter were fix throughout all experiments: Optimizer = Adam, Loss = categorical_crossentropy, epoch = 30, batchsize = 16. While adapting iteratively the model architecture or adjusting the hyperparameters of the model, the performance is tested with the validation data. Using the testing set, we computed results in each cross-validation iteration and averaged scores at the end of the processing. The experiments were run on a DGX station with Tesla V100-DGXS-32GB. Different deep learning models have been tested, among them CNN, (Bi | Stacked)LSTM, CNN-LSTM, ConvLSTM und LSTM + Attention. The models were derived from literature when they considered EEG data for classification or a similar classification problem: Models as tested by Vareka [2,3] have been rebuilt as baseline without any modification (labeled with “gtn_”). Other architectures were taken from Zhang et al. [7] who used the models to classify emotions from EEG signals (labeled with “emotion_”). Anguita et al. [8] applied different architectures to a smartphone dataset for activity recognition. Architectures based on this work are labeled with “mastery_”. Finally, the CNN-BiLSTM architecture suggested by Mansar for classifying sleep stages based on EEG data have been used [9]. All models were optimized (label

"optimized"): We identified good performing hyperparameters using grid search, resulting in optimized architectures.

3. Results

Best performing model was the Deep LSTM *emotion_deep_lstm* with an accuracy of 63.7% in the single trial (see Table 1) and 77.1% in the averaged trial (see Table 2). However, it was not significantly better than the baseline CNN (model *gtn_cnn* shown in the second line in Table 1). We define the model "emotion_deep_lstm" as a Deep LSTM Model with an LSTM layer being the first layer, followed by a Dense layer with 50 units. After that we define another LSTM layer with 6 units, a Dropout layer (dropout rate = 0.9) and an output layer with 2 units (softmax activation). The averaged trial used average values from six epochs as was done by Vareka [3]. Using the average trial, almost all models performed better, but also the standard deviation increased.

modelName	train_acc	dev_acc	test_acc	auc	precision	recall	trainable_param	architecture
emotion_deep_lstm	63.05%	62.71%	63.71%	68.28%	63.71%	63.71%	36428	Deep LSTM
gtn_cnn	**65.13%**	**61.96%**	**63.04%**	**67.43%**	**63.04%**	**63.04%**	89974	**CNN**
optimized_emotion_recog_deep_lstm_model	63.37%	62.31%	63.02%	67.47%	63.02%	63.02%	38450	Deep LSTM
mastery_lstm_attention_stackoverflow	59.79%	59.08%	60.56%	63.94%	60.56%	60.56%	53202	LSTM - Attention
optimized_gtn_cnn	69.69%	60.06%	60.49%	64.00%	60.49%	60.49%	268772	CNN
optimized_emotion_regoc_cnn_lstm_model	60.53%	58.83%	60.05%	63.64%	60.05%	60.05%	590	CNN - LSTM
emotion_cnn_lstm	60.60%	58.79%	59.97%	63.89%	59.97%	59.97%	590	CNN - LSTM
emotion_bidirectional_lstm_attention_stackoverflow	60.15%	58.89%	59.90%	63.34%	59.90%	59.90%	43330	BiLSTM - Attention
emotion_lstm_attention_stackoverflow	58.38%	57.64%	59.46%	62.44%	59.46%	59.46%	1460	LSTM - Attention
emotion_more_layer_cnn_v1	66.66%	58.83%	59.45%	63.18%	59.45%	59.45%	35717	CNN
optimized_gtn_lstm	55.59%	58.52%	59.30%	63.15%	59.30%	59.30%	122132	LSTM
mastery_lstm_cnn	58.71%	58.76%	59.10%	62.23%	59.10%	59.10%	2602622	CNN - LSTM
optimized_sleep_timedistributed_cnn_lstm_model	60.95%	58.16%	58.82%	62.17%	58.82%	58.82%	373202	CNN - BiLSTM
sleep_cnn_lstm	60.22%	57.82%	58.64%	61.94%	58.64%	58.64%	373202	CNN - BiLSTM
gtn_lstm	**56.24%**	**57.74%**	**58.45%**	**62.13%**	**58.45%**	**58.45%**	**120592**	**LSTM**

Table 1. Single trial: Results of the 15 best performing models. Model *gtn_cnn* and gtn_lstm are the baseline models from Vareka [2,3]. Accuracy values for training, validation and test set are shown, as well as values for AUC, precision, and recall.

4. Discussion and future work

Our results show that there are models that can perform slightly better on the considered classification task. Taking Vareka's LSTM (gtn_lstm) as a baseline [3] our best performing model differs from this baseline model in two ways: our model (emotion_deep_lstm) has as first layer an LSTM layer with 64 time steps followed by a dense layer with 50 nodes. In future work, it could be tested to add additional layers since it seems to have an impact on the accuracy. The optimization of the models was not as successful as expected. A reason might be that during Monte Carlo Cross Validation only one iteration was performed. Vareka optimized with 30 iterations. The averaged trial

experiments achieved good performance. Through averaging noise in the data is reduced and only relevant data is considered for classification.

Our work has several limitations: complex architectures were not implemented. Future work is needed to test the performance of such models. We used an existing EEG dataset that only recorded 3 channels. Other researchers use EEG data with 8 recorded channels which might result in more expressive data. Furthermore, we could not test all possible variations of hyperparameter. This also remains open for future work. New deep learning architectures such as Inception, Graph convolutional network or ResNet could be tested on the data.

modelName	train_acc	dev_acc	test_acc	auc	precision	recall	trainable_param	architecture
emotion_deep_lstm	63.10%	62.71%	77.11%	84.97%	77.11%	77.11%	36428	Deep LSTM
gtn_cnn	**65.13%**	**61.96%**	**76.15%**	**83.70%**	**76.15%**	**76.15%**	89974	**CNN**
optimized_emotion_recog_deep_lstm_model	63.41%	62.11%	74.51%	81.75%	74.51%	74.51%	38450	Deep LSTM
optimized_gtn_cnn	69.53%	60.16%	72.64%	79.53%	72.64%	72.64%	268772	CNN
mastery_lstm_attention_stackoverflow	59.79%	59.08%	69.75%	75.78%	69.75%	69.75%	53202	LSTM - Attention
mastery_lstm_cnn	58.23%	58.10%	69.56%	75.16%	69.56%	69.56%	2602622	CNN - LSTM
optimized_gtn_lstm	55.75%	58.68%	69.13%	76.16%	69.13%	69.13%	122132	LSTM
mastery_ConvLSTM	67.54%	57.68%	68.73%	74.85%	68.73%	68.73%	1319214	ConvLSTM
emotion_bidirectional_lstm_attention_stackoverflow	60.31%	58.96%	68.71%	75.12%	68.71%	68.71%	43330	BiLSTM - Attention
optimized_mastery_convLSTM_model	67.50%	57.95%	68.68%	74.73%	68.68%	68.68%	647470	ConvLSTM
optimized_emotion_regoc_cnn_lstm_model	60.44%	58.69%	67.96%	76.03%	67.96%	67.96%	590	CNN - LSTM
emotion_more_layer_cnn_v1	66.59%	58.97%	67.79%	74.52%	67.79%	67.79%	35717	CNN
emotion_cnn_lstm	60.46%	58.86%	67.18%	74.74%	67.18%	67.17%	590	CNN - LSTM
gtn_lstm	**56.25%**	**58.04%**	**67.13%**	**75.07%**	**67.13%**	**67.13%**	**120592**	**LSTM**
optimized_sleep_timedistributed_cnn_lstm_model	60.95%	58.16%	66.90%	72.77%	66.90%	66.90%	373202	CNN - BiLSTM

Table 2. Averaged trial: Results of the 15 best performing models. Model *gtn_cnn* is the baseline. Accuracy values for training, validation and test set are shown, as well as values for AUC, precision, and recall.

References

[1] R. Fazel-Rezai, B. Z. Allison, C. Gugeret al. P300 brain computer interface: current challenges and emerging trends, Frontiers in neuroengineering. 2012;5:14.

[2] L. Vareka, Evaluation of convolutional neural networks using a large multi-subject p300 dataset, Biomedical Signal Processing and Control. 2020;58:101837.

[3] L. Vareka, Comparison of convolutional and recurrent neural networks for the p300 detection, In: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC) 2021, Vol. 4 - BIOSIGNALS, pp. 186–191.

[4] R. Moucek, L. Vareka, T. Prokop, J. Stebetak, P. Brha, Event-related potential data from a guess the number brain-computer interface experiment on school children, Scientific data. 2017;4(1):1–11.

[5] Q.-S. Xu, Y.-Z. Liang, Monte carlo cross validation, Chemometrics and Intelligent Laboratory Systems . 2001; 56 (1):1–11.

[6] H. Shan, Y. Liu, T. P. Stefanov, A simple convolutional neural network for accurate p300 detection and character spelling in brain computer interface., in: IJCAI, 2018, pp. 1604–1610.

[7] Y. Zhang, J. Chen, J.H. Tan et al. An investigation of deep learning models for EEG-based emotion recognition. Frontiers in Neuroscience. 2020;14.

[8] D. Anguita, A. Ghio, L. Oneto et al. A public domain dataset for human activity recognition using smartphones. In: Esann 2013 Apr 24; 3, p. 3.

[9] Y. Mansar. Sleep Stage Classification from Single Channel EEG using Convolutional Neural Networks [Internet]. Medium. 2020 [last access 13.12.2021]. Accessible at: https://bit.ly/32CgZmA