Explaining the Shallow Network Trained to Classify EEG Signals by LRP and Sensitivity Analysis

Martyna Poziomska[1], Alan Drozd[2], Urszula Malinowska[1[0000-0001-8358-6024]], Jacek Rogala[2[0000-0001-5298-920X]], and Jarosław Żygierewicz[1[0000-0002-7536-0735]]

[1] Faculty of Physics, University of Warsaw, Pasteura 5, 02-093, Warsaw, Poland
[2] Nencki Institute of Experimental Biology, Polish Academy of Science, Pasteura 3, 02-093, Warsaw, Poland
m.poziomska@student.uw.edu.pl

Abstract. Explanation of EEG classification made by a neural network is still a challenging problem. Therefore, we test two approaches based on sensitivity analysis and Layer-wise relevance propagation. We show that the explanations obtained by these methods are consistent only to a certain degree, as each focuses on different aspects. Sensitivity analysis exposes which changes in the data would increase the probability of a given class, while LRP is more related to the values of specific features in the data. The additional difficulty is ensuring that the indicated features are indeed plausible for contrasting the conditions. Here we used the spectral power distribution across channels in the experimental conditions as a sanity check.

Keywords: Explainable Machine Learning, EEG Classification.

1 Introduction

Neural networks models are often treated as black-boxes, i.e., trained to solve a classification problem. They are effective at assigning class labels to input data. However, in many practical cases, it is important to understand the essential input features used by the model and whether they are reasonable in the given problem or the model uses some artifacts present in the training set. This contribution presents the attempts to assess the importance of the input EEG signal features using two methods.

2 Data

Data come from 87 participants who took part in a delayed match-to-sample experiment. There were two experimental conditions requiring (ATT) or not (CON) retention of an object in the visual memory. The signals used to train the network were 5s long fragments encompassing the delay phase, recorded with 19 electrodes from the 10-20 system, sampled at 400 Hz, and filtered in the 0.5-45 Hz frequency band. In total, there were 10937 CON and 8366 ATT trials.

3 Methods

We used the Shallow-ConvNet architecture as proposed in [1]. In a 3-fold cross-validation (CV) scheme, we trained it to classify trials as ATT or CON, using binary cross-entropy loss and AdamW optimizer. The importance of the signal features utilized by the model was estimated by two approaches.

First, we conducted a sensitivity analysis. It relied on evaluating the gradient of class ATT probability p over the perturbation parameters related to frequency-band power in a given EEG channel. The obtained gradients were averaged across 3 CV folds and 5 random initializations.

The second approach evaluated the relevance of the periods and filters in the first and second Conv layers using Layer-wise Relevance Propagation (LRP) [2]. Filters in the first Conv layer can be interpreted as FIR filters; thus, we can straightforwardly compute their transmittance $H_F(f)$ to obtain their frequency characteristics. We choose the most important filter F^* for each electrode e based on the mean absolute value of the relevance $R_{2,F,e}$ computed for each electrode and each filter F in the second layer of the model. The product of the absolute value of the transmittance $H_{F^*,e}(f)$, the spectral power of the signal (only for ATT) $S_e(f)$, the normalized[1] relative filter weight from the second layer $\hat{R}_{2,F^*,e}$ and the normalized relative filter weight from the firtst layer $\hat{R}_{1,e}$ yields the importance of a given frequency band ($f \pm 2$Hz) at a given electrode $I_e(f)$:

$$I_e(f) = \sum_{f_i=f-2Hz}^{f+2Hz} \left| H_{F^*,e}(f_i) \right| \cdot \mathrm{S}_e(f_i) \cdot \hat{R}_{2,F^*,e} \cdot \hat{R}_{1,e} \quad (1)$$

Additionally, we evaluated the average power spectral density in the two experimental conditions by periodogram. These spectra served as a reference for interpreting the results of features indicated by sensitivity analysis and LRP.

4 Results

The models trained in 3 CV and 5 random initializations achieved Matthews correlation coefficient of 0.22±0.02.

The average importance of frequency bands and channels according to sensitivity analysis is shown in Fig. 1a. Positive index values indicate that an increase of power in a given frequency band at a given electrode increases the probability of the input trial being of class ATT. We note a group of frontal (F7, Fz, F4) and temporal (T3, T4) electrodes showing that an increase of theta and alpha power in these regions suggests the trial being of ATT class. In contrast, an increase of theta and alpha power in parieto-temporal and occipital regions indicates the trial being of CON class.

The analysis of the most relevant filters indicated by the LRP method is shown in Fig. 1b. Here, we notice that the electrodes T6 and Pz are especially prominent at alpha and theta frequency bands. But in contrast to the sensitivity analysis, the frontal regions are not highly relevant.

[1] Divided by the max value

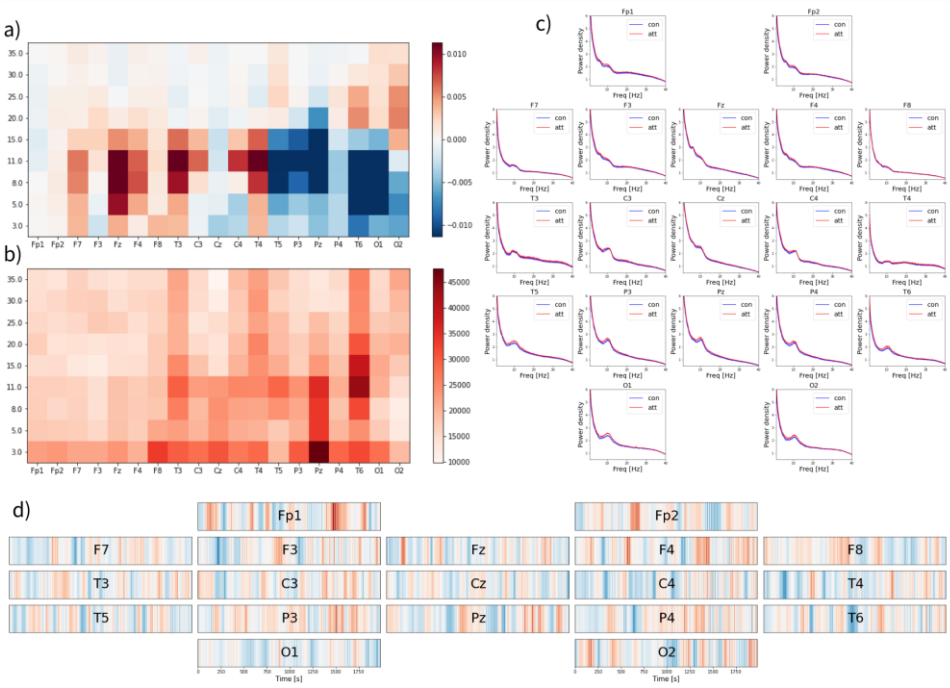

Fig. 1. a) Results of the sensitivity analysis; b) Importance of channels and frequency bands in the LRP analysis; c) Average spectral power, with standard error of measurement marked, for ACC and CON conditions; d) time course of relevance in the EEG channels.

We show average spectra for ATT and CON experimental conditions (Fig. 1c.) as a reference for the above-described results. Again, we observe an alpha rhythm peak, especially prominent in the central and parietal regions. Moreover, a closer look reveals that the power for ATT in the alpha peak is higher than for the CON class. Additionally, we notice a peak corresponding to theta rhythm in the frontal regions.

The relevance propagated to the input signals enables us to estimate the importance of various time periods at different channels. We visualize it in Fig. 1d.

5 Conclusion

The current study demonstrates that the sensitivity analysis and LRP based approaches are promising in explaining the neural network models trained to EEG classification. Still, in contrast to applying these methods to explaining models in the context of image classification, it is much more challenging to demonstrate the correctness of the indication of the critical features.

References

1. Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization. Hum. Brain Mapp. 38, 5391–5420 (2017) DOI: 10.1002/hbm.23730
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS ONE, vol. 10, no. 7, Art. no. e0130140 (2015)