Visual cues can bias EEG Deep Learning models

David Trocellier, Bernard N'Kaoua, Fabien Lotte

Visual cues can bias EEG Deep Learning models

[David Trocellier[1], Bernard N'Kaoua[2], Fabien Lotte[1]]
[[1]Affiliation-A Inria center at the university of Bordeaux/ LaBRI, Talence, France]
[[2]Univ. Bordeaux, Bordeaux, France]

Synopsis

The use of Deep Learning (DL) for classifying motor imagery-based brain-computer interfaces (MI-BCIs) has seen significant growth over the past years, promising to enhance EEG classification accuracies. However, the black-box nature of DL may lead to accurate but biased and/or irrelevant DL models. Here, we study the influence of using visual cue EEG (which is commonly done) in the DL input window on both the features learned and the classification performance of a state-of-the-art DL model, DeepConvNet. The classifier was tested on a large MI-BCI dataset with two time windows post visual cue: 0-4s (with the cue EEG) and 0.5-4.5s (without). Performance-wise, the first condition significantly outperformed the second (86.82% vs. 76.11%, $p<0.001$). However, saliency maps analyses demonstrated that the inclusion of the visual cue EEG leads to the extraction of cue-related evoked potentials, which are distinct from the MI features used by the model trained without visual cues EEG.

Background

DL excels in learning complex, non-linear features, leading to often superior performance in classifying EEG in MI-BCIs, notably in cross-subject paradigms (Lawhern et al, 2018). Here, models are trained on a large dataset from numerous users and evaluated on data from new users not included in the training set. Among the many parameters to tune for DL model training, the input window start and end times for the signal segments used for classification— can significantly impact performance. Usually, the more time points used as input, the higher the classification accuracy. Thus, EEG DL models tend to use the longest possible window, from the MI beginning to its end (see, e.g., the BEETL competition (Wei et al, 2022)). However, in controlled experiments, the MI start often coincides with the presentation of visual cues indicating which task to perform, potentially biasing the classification of the MI signal.In this article, we stress the importance of examining both the features learned by DL models and their classification performance, to avoid biased and/or irrelevant DL EEG models that could not be used in practice. We highlight the effect of incorporating visual cues EEG in the DL input window and analyze its impact on both model performance and the characteristics of the learned features.

Methods

We investigated the influence of including visual cue EEG in the DL input window on the features learned by the DeepConvNet model (Schirrmeister et al., 2017) by utilizing a large, open-source dataset (Dreyer et al, 2023) comprising 87 participants. These participants engaged in a single-session left and right-hand MI-BCI task, which included six runs—two for calibration and four for online tasks with feedback. Each run consisted of 40 trials (20 per class),

during which participants were shown a red arrow pointing left or right (the visual instruction cue), indicating the hand with which they should perform MI. Participants engaged in MI for 5 seconds and received feedback via a continuously updated horizontal gauge.

The model was trained using a leave-one-subject-out, cross-subject paradigm, where one subject was used as the test set, and the remaining subjects were divided between the training (80%) and validation (20%) sets. For the training set, all trials were included, whereas only the online feedback runs were utilized for the test and validation sets. All data were normalized based on the mean and standard deviation of the training dataset. The training followed the standard parameters proposed by the original DeepConvNet paper (Schirrmeister et al., 2017). Modifications were made to the batch size and the number of epochs (batch size: 256, epochs: 150), with the best model saved based on validation set performance and subsequently evaluated on the test set.

The model was trained with two different input windows, both incorporating 4 seconds of signal as input: the first starting simultaneously with the visual cue (0-4s) and the second starting 0.5 seconds after the cue (0.5-4.5s). Both input windows having the same size: 27 channels and 2048 time points. All other parameters remained unchanged, ensuring a controlled comparison of the impact of including the cue on the model's performance and learned feature representation.

We used saliency maps (Montavon et al, 2018) to identify the features utilized for discrimination between the two classes. This method is based on gradient backpropagation conducted on test data. To enhance the interpretability, we transform the gradient matrix (27 x 2048) into two more comprehensible forms: a temporal vector (2048) and a spatial vector (27). This transformation is achieved by averaging the gradients across one dimension while preserving the other one.

Results

Classification performances significantly differ with (86.82%) and without (76.11%) the visual cue EEG, ($p<0.001$ with a paired t-test). Saliency maps show distinct feature utilization across conditions. Temporally, the most discriminative features appear before 0.5s for the 0-4s window and around 1.5s post-cue for the 0.5-4.5s window. Spatially, electrodes P3 and P4 emerge as most salient in the 0-4s window, accounting for 16.6% and 12.8% of sensitivity, respectively. In contrast, for the 0.5-4.5s window, electrodes C3 and C4 emerge as the most salient, with 9.4% and 10.4% of sensitivity explained.

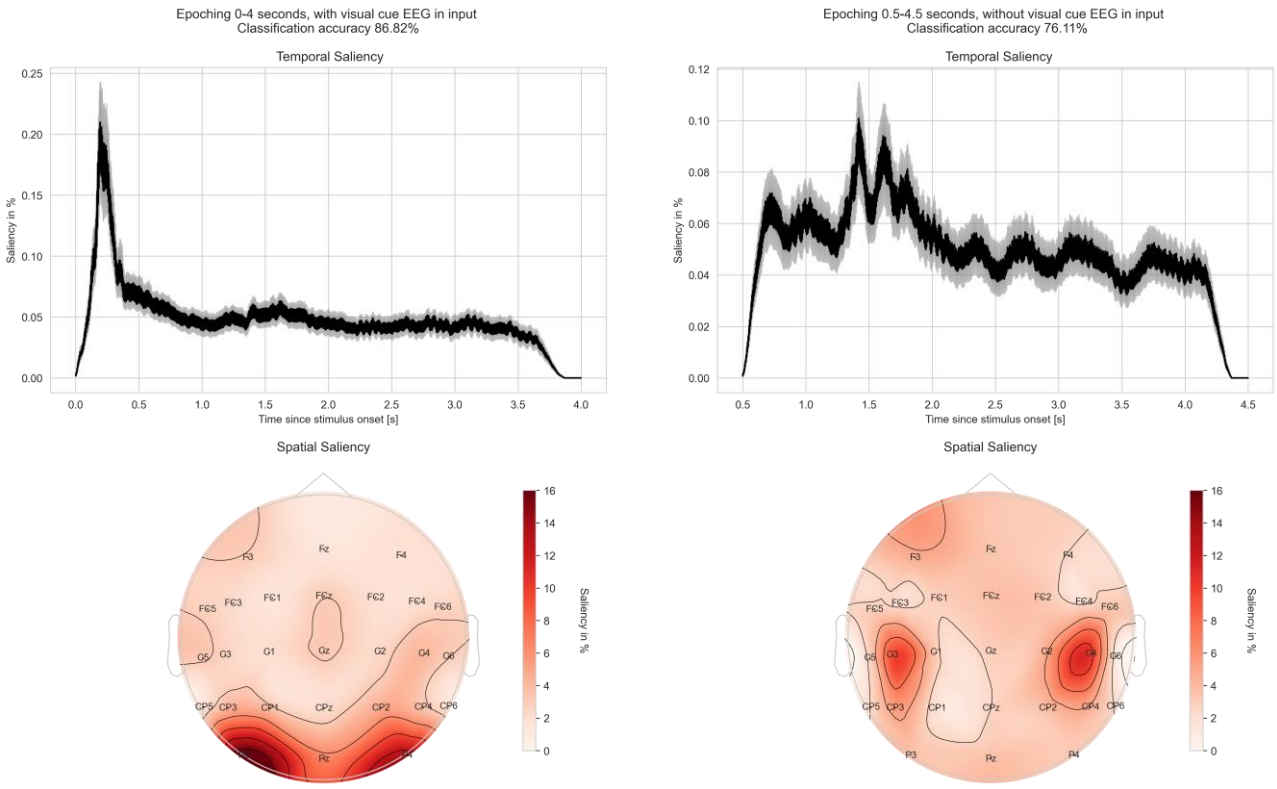

Figure 1: The temporal and spatial saliency maps for the DeepConvNet model including the visual cue EEG or not (right). The gradient features are extracted for each subject and their corresponding model, resulting in 87 distinct gradient matrices. The shown saliency maps are thus the average maps across subjects. **Temporal Saliency:** The black line represents the mean percentage of saliency at each time point, while the grey shading indicates its standard deviation. **Spatial Saliency:** The topography shows the percentage of saliency of each electrode.

Discussion

Including the visual cue EEG in the DL input affects both classification performance and the features learned by the model. It leads to an improvement of classification while extracting features seemingly corresponding to evoked potentials processed in the visual cortex. Many BCI DL papers do use this visual cue EEG. However, this improvement is not practically relevant. Indeed, such visual cues are only present in controlled lab experiments, but not anymore in any actual MI-BCI use in which users choose their own commands. Features learned by a MI-BCI should thus concentrate around C3 and C4 as it is the case when the visual cue is not included in the DL input window.These findings emphasize the importance of utilizing saliency maps to ensure that high classification accuracy is indeed due to the proper extraction of MI features, rather than the exploitation of bias.

Disclosures

This work was supported by the ANR projects ”Doctoral contracts in Artificial Intelligence” of Univ. Bordeaux/Inria (Grant No. ANR-20-THIA-0008-01) & PROTEUS (Grant ANR-22-CE33-0015-01).

References

(Al-Saegh et al., 2021) Al-Saegh, A., Dawwd, S. A., & Abdul-Jabbar, J. M. (2021). Deep learning for motor imagery EEG-based classification : A review. *Biomedical Signal Processing and Control*, *63*, 102172. https://doi.org/10.1016/j.bspc.2020.102172

(Dreyer et al., 2023) Dreyer, P., Roc, A., Pillette, L., Rimbert, S., & Lotte, F. (2023). A large EEG database with users' profile information for motor imagery brain-computer interface research. *Scientific Data*, *10*(1), Article 1. https://doi.org/10.1038/s41597-023-02445-z

(Lawhern et al, 2018) Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet : A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

(Montavon et al, 2018) Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing*, *73*, 1-15. https://doi.org/10.1016/j.dsp.2017.10.011

(Schirrmeister et al. , 2017) Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, *38*(11), 5391-5420. https://doi.org/10.1002/hbm.23730

(Wei et al, 2022) Wei, X., Faisal, A. A., Grosse-Wentrup, M., Gramfort, A., Chevallier, S., Jayaram, V., ... & Tempczyk, P. (2022, July). 2021 BEETL competition: Advancing transfer learning for subject independence and heterogenous EEG data sets. In *NeurIPS 2021 Competitions and Demonstrations Track* (pp. 205-219). PMLR.