Concept-Based Explainability for an EEG Transformer Model

Anders Gjølbye Madsen[*†] William Theodor Lehn-Schiøler[*†]
Áshildur Jónsdóttir[*] Bergís Arnardóttir[*] Lars Kai Hansen[*]

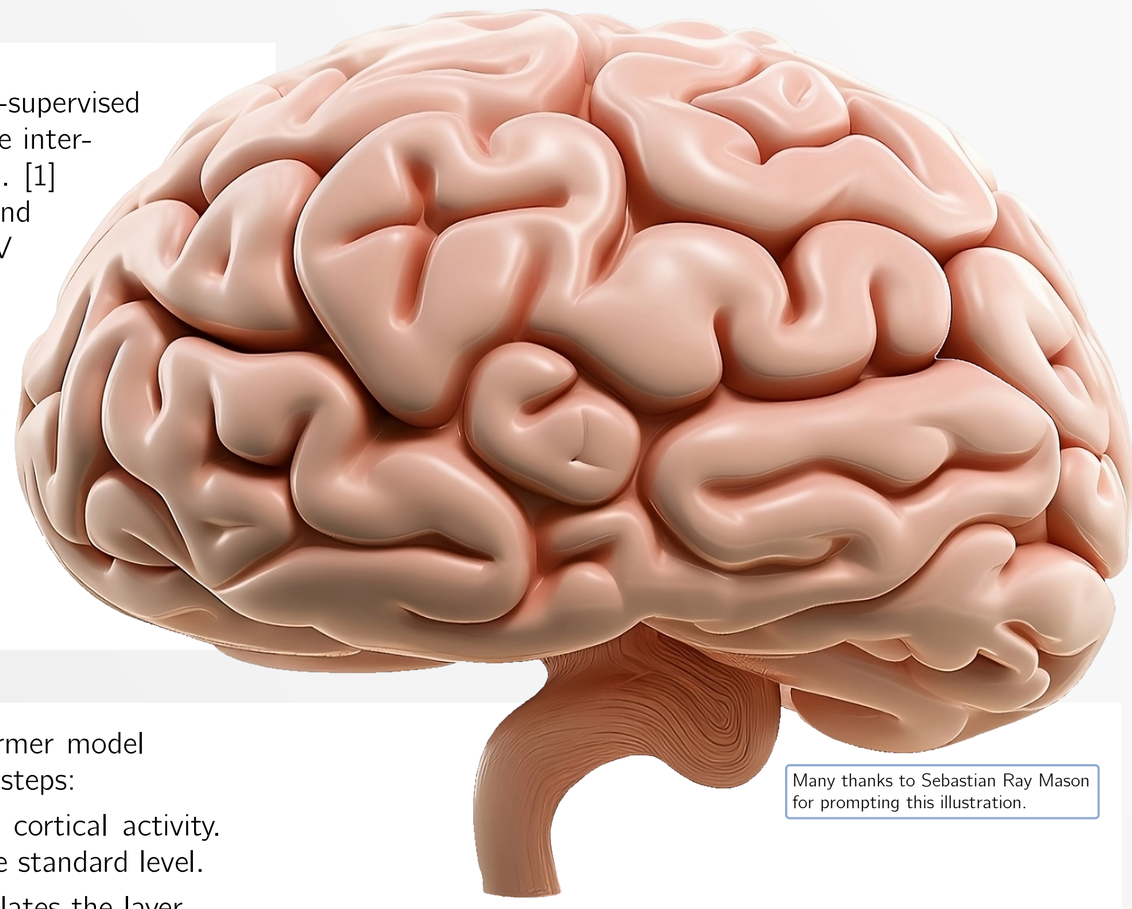
[*]Technical University of Denmark [†]BrainCapture

INTRODUCTION

This work investigates representations of electroencephalogram (EEG) data obtained by self-supervised learning methods, motivated by the lack of labeling in large-scale EEG datasets. We apply the interpretability method of Testing Concept Activation Vectors (TCAV) approach from Kim et al. [1] to BENDR-based models from Kostas et al. [2], to provide insights into their structure and decision-making processes. A better understanding of EEG transformer models using TCAV could support the use of these models as diagnostic support tools for identifying EEG abnormalities, such as epileptic discharges [3]. To address this, we present the following scientific contributions:

- TCAV workflows for EEG data, proposing concepts based on human-annotated data as well as concepts defined by cortical areas and frequency bands.
- Sanity checks for TCAV to ensure valid explanations in simple EEG settings.
- Two practical applications: seizure prediction and brain-computer interfacing (BCI).

All code used in this research has been made publicly accessible for validation and replication.

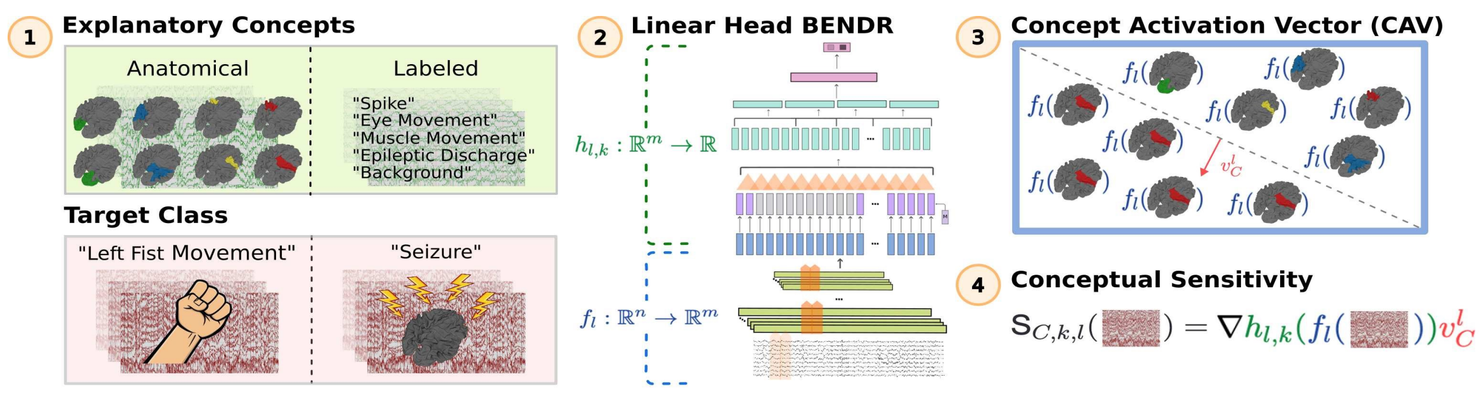


Many thanks to Sebastian Ray Mason for prompting this illustration.

METHODOLOGY

The experiment aims to evaluate if explanatory concepts can provide insights into a transformer model designed for EEG data [1, 2, 5]. The approach is novel for EEG data and consists of four key steps:

1) **Explanatory concepts** are defined as either event-based EEG labels or frequency-based cortical activity. The latter involves assessing the difference in alpha activity between a cortical area and the standard level.
2) **The BENDR model**, inspired by language modeling [4], processes these concepts and isolates the layer activations. BENDR enhances EEG-based BCIs with self-supervised learned EEG data representations.
3) **Concept Activation Vectors (CAV)** are computed as the normal vector to the hyperplane that separates layer activation for explanatory and random concepts.
4) **The TCAV Score** measures the explanatory concept's sensitivity at the bottleneck, using directional derivatives relative to the respective CAV. TCAV scores close to 0 or 1 signify negative or positive alignment with the activation layer, respectively.



Testing with Concept Activation Vectors (TCAV) is a technique used to quantify the degree to which layers of neural networks align with human-defined concept with are separated from *random* concepts using linear discriminants. Directional derivatives are used to measure how sensitive the network is to changes in the input data,

$$S_{C,k,l}(\boldsymbol{x}) = \nabla h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l,$$

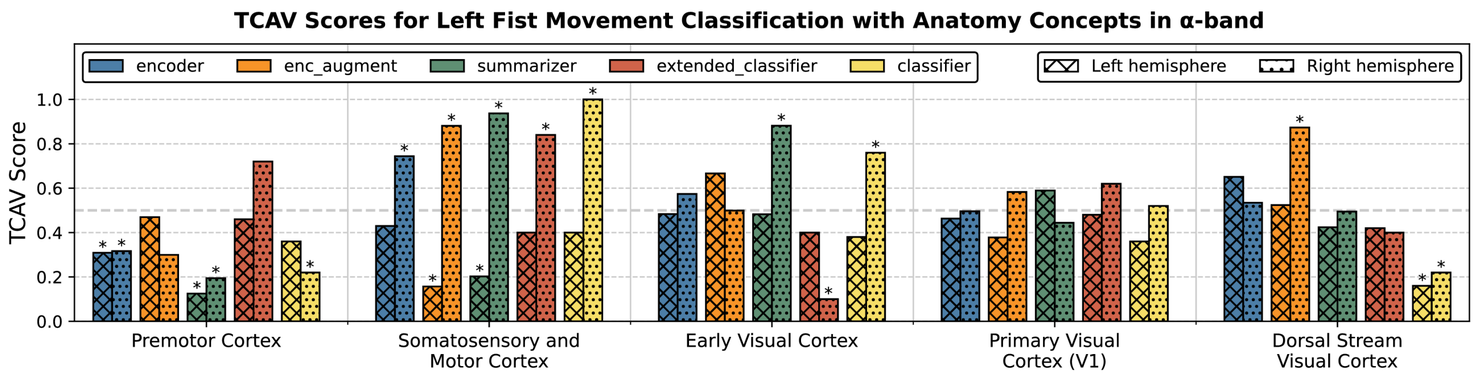
i.e., where ***l*** is the layer, ***k*** is the class, and ***C*** is the concept. A TCAV score is defined as the ratio of examples that have positive sensitivity, i.e.,

$$\mathrm{TCAV}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}.$$

Concepts ***C*** for which a null hypothesis that the average sensitivity lies around zero is rejected thus relate to the target class prediction ***k***, and may bring positive or negative evidence for the given target class. The null hypothesis of the test is that half of the examples have positive sensitivity and the other half have negative or zero sensitivity, i.e.,

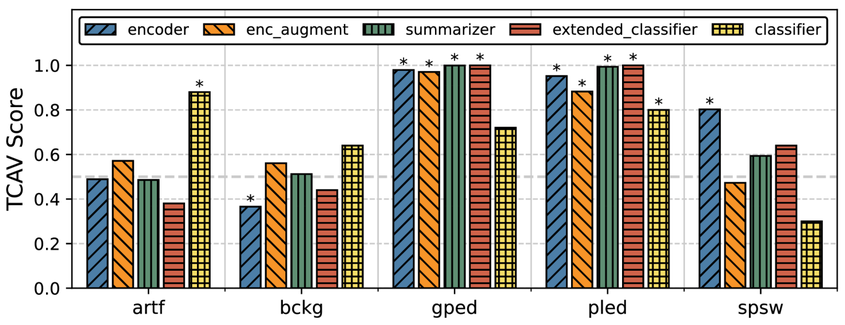
$$H_0 : \mathrm{TCAV}_{C,k,l} = 0.5.$$

RESULTS



The TCAV scores analyze the alignment between α-band activity in cortical regions and the *Left Fist Movement* classification task in the model at five bottlenecks. The analysis reveals significant lateralization in the *Motor Cortex* across all five bottlenecks, indicating its positive significance in the task. These results strongly suggest that the model's internal representation incorporates lateralization, reflecting the fact that one hemisphere exhibits more electrical activity than the other. It is noteworthy that lateralization is most significant in the *Encoding Augment* and *Summarizer* bottlenecks, indicating that it is captured early in the network. While no apparent lateralization is present in the *Premotor Cortex*, this part of the cortex is negatively significant in the *Encoder* and *Summarizer* bottlenecks for both the left and right hemispheres. This suggests that the EEG data only captures the performing and not the planning of the movement.

The TCAV scores assess the alignment of event-based EEG labels with the internal representation of the Seizure classification task in the model at five bottlenecks. From the right, the concepts are defined as (1) technical artifacts, (2) background, (3) generalized periodic epileptic discharge, (4) periodic lateralized epileptic discharge, and (5) spike and short wave. We observe a significant alignment in (3) and (5) for most bottlenecks and observe that (3) is only significant on the first layer which likely corresponds to it being filtered out due to it being irrelevant for seizure classification.



CONCLUSION

In this work, we present two new workflows for concept-based explainability within the TCAV framework for EEG data. Specifically, we provide

- Concepts derived from labeled data, i.e., the TUH EEG database.
- A workflow based on source location of resting-state EEG data.
- Concepts derived from anatomical cortical areas and for specific frequency bands.
- A case study involving seizure prediction, where TCAV score reveals the role of fundamental spike patterns.
- A brain-computer interface case, hinting at how the TCAV method offer valuable insights into classifier design for EEG data.

REFERENCES

[1] Been Kim, Martin Wattenberg, et al., *"Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),"* 2018.

[2] Demetres Kostas, Stèphane Aroca-Ouellette, et al., *"Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data,"* Frontiers in Human Neuroscience, vol. 15, 2021.

[3] Dragoljub Gajic, Zeljko Djurovic, et al., *"Detection of epileptiform activity in eeg signals based on time-frequency and nonlinear analysis,"* Frontiers in computational neuroscience, ol. 9, pp. 38, 2015.

[4] Alexei Baevski, Henry Zhou, et al., *"wav2vec 2.0: A framework for self-supervised learning of speech representations,"* CoRR, vol. abs/2006.11477, 2020.

[5] Ashish Vaswani, Noam Shazeer, et al., *"Attention is all you need,"* 2017.

This work is supported by The Pioneer Centre for AI, DNRF grant number P1, The Novo Nordisk Foundation grant NNF22OC0076907 "Cognitive spaces - Next generation explainability".