Introspection of deep learning models trained on brain fMRI data

IntroDL Project proposal

Pavel Popov, Ufuk Seçilmiş, and Selim Süleymanoğlu

Georgia State University

1 Description

In recent years deep learning methods have been extensively used in brain image analysis. Deep learning (DL) models, while being inspired by the organic neural mechanics, in return proved to be highly effective in establishing causal connections between brain regions and showed decent results on the classification of brain disorders from brain imaging data. However, despite this classification success, the complexity of DL models makes them effectively 'black boxes' when it comes to the problem of pinpointing the abnormalities in the data which led to a particular classification decision. Solving this problem would be useful for understanding brain processes and disorders. For this project, we want to analyze the ability of a few general and brain-focused DL models, trained for classification on some synthetic and brain imaging datasets, to produce sound saliency maps.

2 Outline

Datasets We will use ABIDE [1] (Autism spectrum disorder/healthy controls) and COBRE [2] (Schizophrenia/healthy controls) fMRI datasets. Both datasets are processed to remove noise and transform 3D brain images into 1D vectors of ICA components. This way, for each sample from these datasets we will have a time series of signals extracted from ICA regions.

We are also planning to craft a few synthetic datasets with manually introduced abnormalities in the data. Using these datasets will allow us to perform sanity checks of the saliency maps produced by the studied models.

Models We will use the general LSTM and Vanilla Transformer DL models adapted to the data we are working with. We will also use a brain-focused DICE model [3], which shows decent classification results on the time-series fMRI data.

Saliency maps We will use the Captum package to generate saliency maps of the trained models. We will use the baseline Saliency, Integrated Gradients, and Noise Tunnel methods.

3 Preliminary report

We will describe the techniques we used for preparing the synthetic datasets. We expect to obtain the saliency maps of LSTM and Transformer models trained on synthetic datasets, and assess the models‘ ability to detect the manually introduced abnormalities.

4 Final report

We will describe the methods used for producing saliency maps. In addition to the results of the preliminary report, we expect to obtain the saliency maps of all of the studied models trained on synthetic and brain imaging datasets. We will analyze the obtained result and make conclusions on the models‘ ability to detect abnormalities in the fMRI data, and discuss the abnormalities themselves.

References

[1] A Di Martino, C-G Yan, et al. “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism”. en. In: *Mol Psychiatry* 19.6 (June 2013), pp. 659–667.

[2] Mustafa S Çetin, Fletcher Christensen, et al. “Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia”. en. In: *Neuroimage* 97 (Apr. 2014), pp. 117–126.

[3] Usman Mahmood, Zening Fu, et al. “Through the looking glass: Deep interpretable dynamic directed connectivity in resting fMRI”. In: *NeuroImage* 264 (2022), p. 119737.