Study the Interpretability of Brain MR Image Deep Learning Modelling

Justin Kennedy\textsuperscript{1}, Siyuan Shi\textsuperscript{2}, Jiook Cha\textsuperscript{3}
\textsuperscript{1}Data Science, \textsuperscript{2}Data Science, \textsuperscript{3}Psychiatry, Columbia University

Overview

In this project, we looked to identify parts of the brain that are relevant towards prediction of specific levels of cognitive impairment, from cognitively normal to diagnosed Alzheimer’s disease patients. To do this, we applied several machine learning algorithms in aggregation on patient brain MR image sequences to generate region-specific heatmaps of each patient sample. In building off the resulting interpretations, this project has implications on forecasting an individual’s neurocognitive development, cognition, and behaviors.

Method

Layer-wise relevance propagation (LRP) works by creating a heatmap of per-pixel contributions to an input image’s output label by computing relevances of neurons in each layer in a backward pass. In Guided Backpropagation (GBP), the gradient is backpropagated and all negative gradients are set to 0 to identify the neurons that contribute positively to the predicted group. Grad-CAM(++) works by constructing a weighted sum of feature maps using gradients in the network.

Results

A subsample of 133 instances was chosen to test the MLI algorithms on. The model’s predictions on the sample are summarized as a confusion matrix in Figure 3.

Here we show the resulting MLI interpretations of one of the instances. The LRP and GBP algorithms give per-pixel details; the colored pixels are important for making the prediction. The Grad-CAM (Grad-CAM++) result focuses on the region-level details; the colored regions contain features that the model thinks belong to the predicted class.

Conclusion

Though difficult to assess the accuracy of the resulting MLI heatmaps, especially given their dependence on the accuracy of the model they’re based on, we can observe a consistency between separate MLI methods in certain brain regions being commonly specified as predictive of cognitive group. As the model’s accuracy improves, these results could help infer area-specific characteristics related to cognitive conditions.

Acknowledgments

We would like to thank Seungwook Han and Professor Jiook Cha for providing helpful feedback over the course of this project.

References

