Numerous deep learning approaches have been proposed to automatically classify Alzheimer’s disease (AD) from medical images. However, common approaches such as convolutional neural networks (CNNs) lack interpretability and are prone to overfitting when trained on small datasets. As an alternative, significantly less work has explored deep learning approaches to region-based features that are commonly attained from atlas partitions of known regions of interest (ROIs). In this work, we propose a self-attention mechanism to jointly learn a graph of connectivity between ROIs as a prior for learning meaningful features for AD prediction. We apply our method to a standard benchmark classification task using the ADNI dataset and systematically compare its performance to other ML approaches for ROI-based methods. Finally, we perform exploratory analysis and analyze the interpretability properties of the learned attention graphs for AD prediction.

Discussion and Future Work

• This work shows that utilizing self-attention to extract inter-region connectivity is a powerful inductive bias for classifying Alzheimer’s disease.
• For future work, we would like to further inspect the properties of the attention graphs for individual subjects over time with a focus on differentiating MCI caused from AD and MCI from other causes.
• A straightforward extension to the above model is to use more complicated architectures for the readout layer, such as graph neural networks (GNNs), to map the final node embeddings to the classification target.
• Another straightforward extension is to build a more complicated input embedding to summarize the region information by working with the volume partitions directly.
• A less straightforward extension would be to learn the regions and embeddings directly from the image volumes of multiple modalities.

References