Multimodal Region-Based Transformer for the Classification and Prediction of Alzheimer's Disease

Kevin Mueller, Gordon Erlebacher, Anke Meyer-Baese

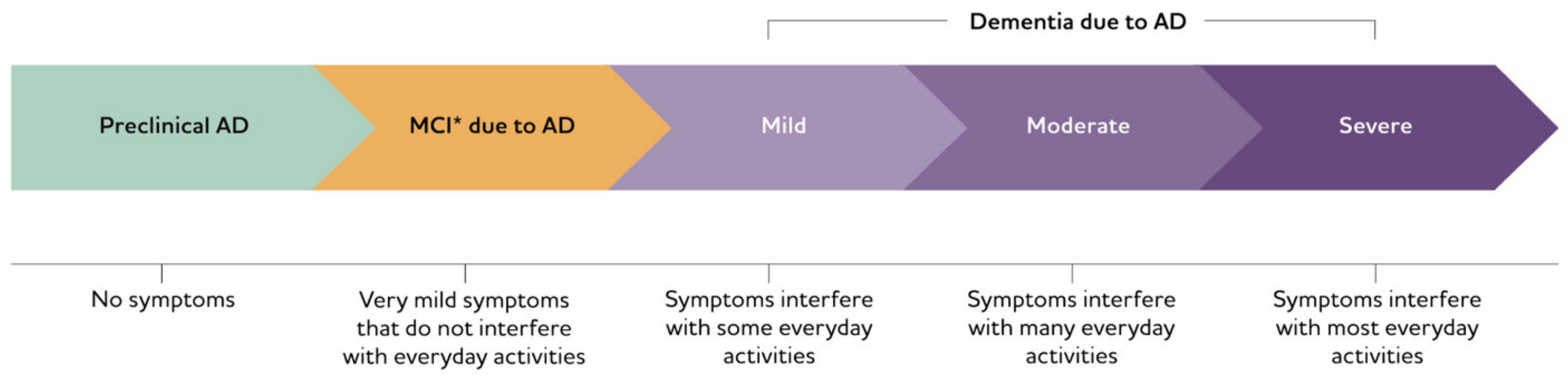
Department of Scientific Computing, Florida State University

Abstract

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 applying deep learning approaches to region-based features that are commonly attained from atlas partitions of known regions of interest (ROI). In this work, we propose a self-attention mechanism to jointly learn a graph of connectivity's 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.

Introduction

Alzheimer's disease is a progressive and irreversible disease that causes significant memory loss, thinking, and eventually even the ability to carry out simple tasks [2]. As such, there is a strong need to develop better models for accurate prediction of the severity and progression of the disease. Modeling the progression of AD as a graph has many inherent advantages such as interpretability and enabling scientific discovery, but most graph connectivity measures are defined for an entire population and unable to work at the level of individual subjects. This work seeks to explore alternative connectivity measures defined by neural networks as a way to bridge this gap.


Dementia due to AD

Preclinical AD — No symptoms

MCI* due to AD — Very mild symptoms that do not interfere with everyday activities

Mild — Symptoms interfere with some everyday activities

Moderate — Symptoms interfere with many everyday activities

Severe — Symptoms interfere with most everyday activities

Preprocessing

For this work, we use the baseline PET and T1 image scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) as input for our model. While it is possible to train a classifier for an assortment of clinical conditions with the ADNI dataset, we only consider the control vs. AD diagnosis case for 473 (260 CN / 212 AD) total subjects. We then apply the following preprocessing:

- Segmenting an anatonmical T1 image into WM, GM, and CSF and creating a warp field for nonlinear registration.
- Coregistering PET with thresholded GM and WM images.
- Nonlinear registration of T1 into MNI standard space.
- Nonlinear registration of coregistered PET into MNI standard space.
- Intensity normalization of each PET volume such that the sum of all voxels within the subjects brain mask are equal to one.
- Calculating mean activity of the 116 regions of the normalized GM and PET images provided from the AAL atlas.

Model

Given a dataset, $\{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^{R \times d}$ consists of the PET and GM mean voxel values for all R regions, and y_i is the subject's label at time of baseline scan, our goal is to train a classifier, $f : X_i \longrightarrow \hat{y}_i$, to predict the correct label. Our model consists of two main parts:

- A **multi-headed self-attention (MHA) layer** to extract inter-region features. For this layer, we use the standard formulation [1], where $H^{(l)} \in \mathbb{R}^{R \times d_h}$ is a matrix representation of all the node features $h_j^{(l)}$, and all nodes are updated at subsequent layer $l+1$ as

$$H^{(l+1)} = \sigma\left(\Psi W^U\right) \tag{1}$$

$$\text{where } \Psi = \left(\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)\right) V. \tag{2}$$

  $Q \in \mathbb{R}^{R \times d_q}$, $K \in \mathbb{R}^{R \times d_k}$ and $V \in \mathbb{V}^{R \times d_v}$ are the query, key, and value hidden representations of each node defined as:

$$Q = H^{(l)} W^Q, \quad K = H^{(l)} W^K, \quad V = H^{(l)} W^V \tag{3}$$

  where W^Q, W^K, W^V are their corresponding transformation matrices. In general, $\sigma(\cdot)$ can be any nonlinearity, but for our simple 1-layer model we use a linear activation function. The multi-headed version of the above formulation is then

$$H^{(l+1)} = \mathcal{C}(\Psi, \ldots, \Psi_n) W^U, \tag{4}$$

  which represents the concatenation of n attention heads $\Psi \in \mathbf{R}^{R \times d_v}$, each computed by applying equation 2. Each head contains its own attention matrix and increases the models capacity when modeling interactions between nodes.

- A **readout layer** that learns a mapping from the node embeddings at the final layer $H^{(L)}$ to the class label $\hat{y}_i$. In order to preserve the most information in node embeddings, we project the embeddings to a lower dimensional space and then concatenate all nodes as input for a multilayer perceptron (MLP), such that

$$\hat{y}_i = \text{MLP}(h_G) \tag{5}$$

$$\text{where } h_G = \mathcal{C}(h'_1, h'_j, \ldots, h'_R) \tag{6}$$

$$\text{and } h'_j = W^G(h_j^L). \tag{7}$$

We then optimize the model applying the standard cross entropy loss between the predicted and true output using stochastic optimization.

Results

We use Monte Carlo cross-validation for 20 trials with a 20% hold-out data set to validate our model and compare it with established baselines.

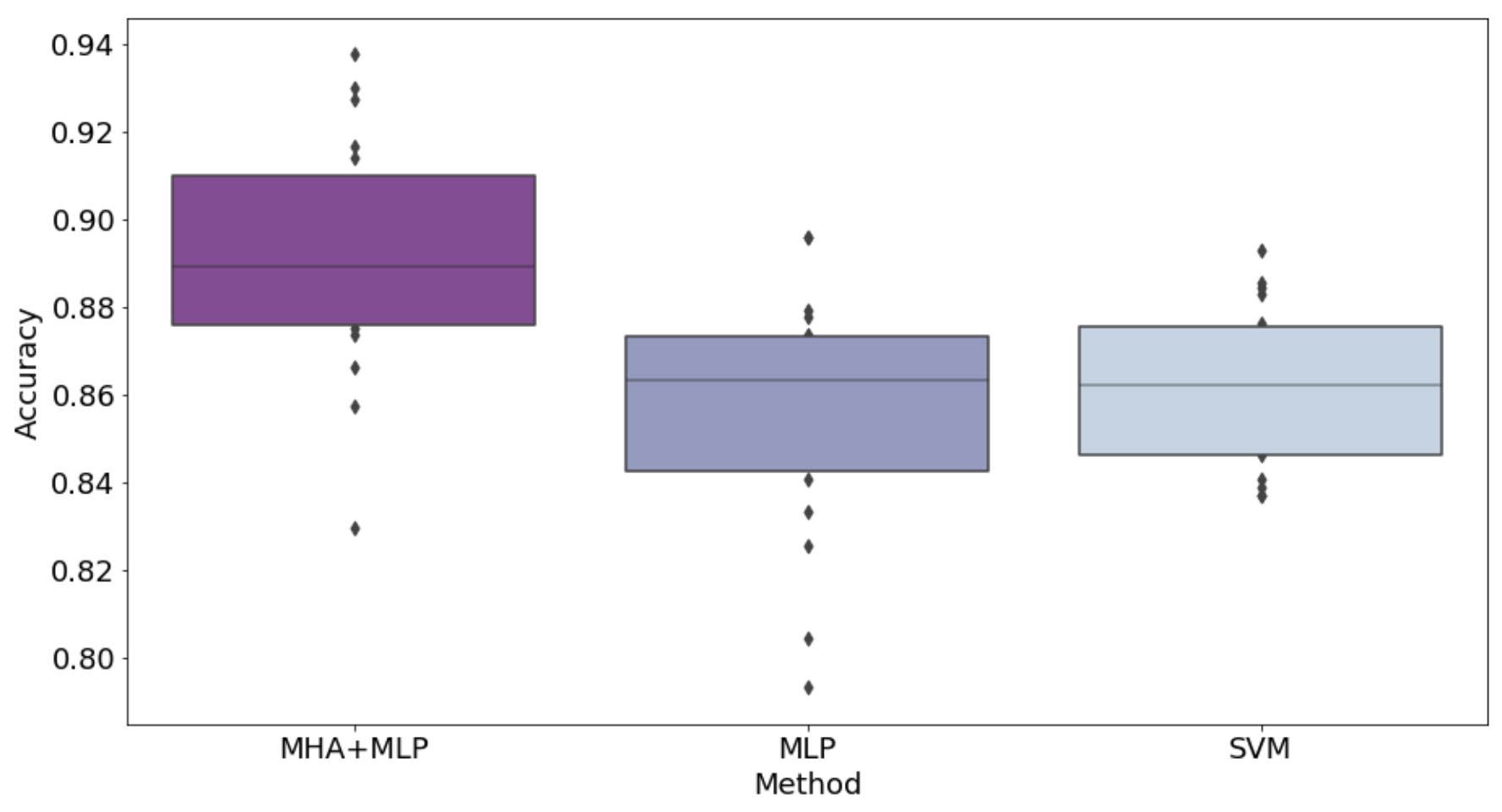


Figure 2: *Distribution of the accuracy for the AD vs. CN classification task using different region-based classifiers (multihead attention + multilayer perceptron, multilayer perceptron, support vector machine)*

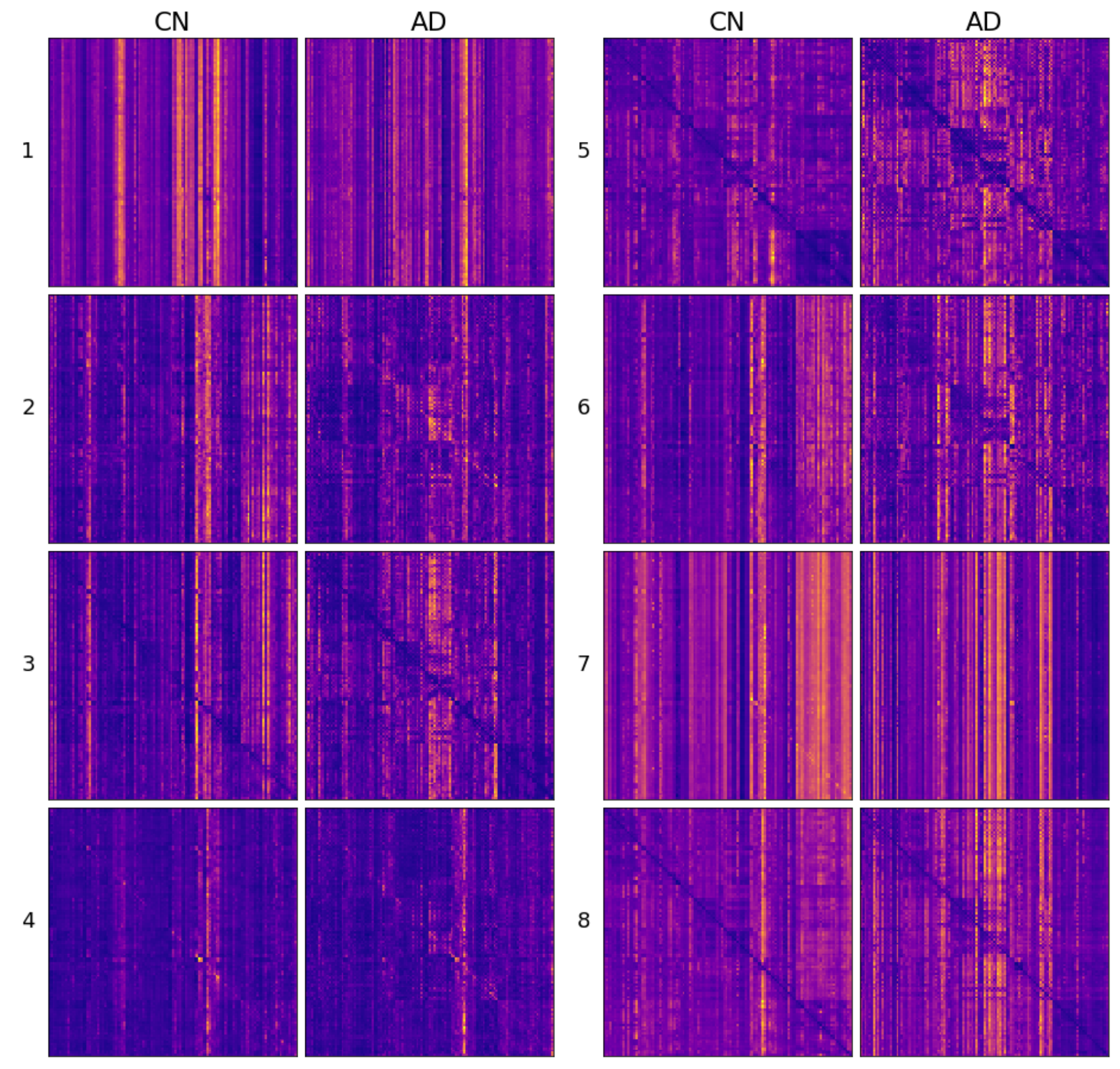


Figure 3: *Comparison of 8 attention heads between CN and AD subjects.*

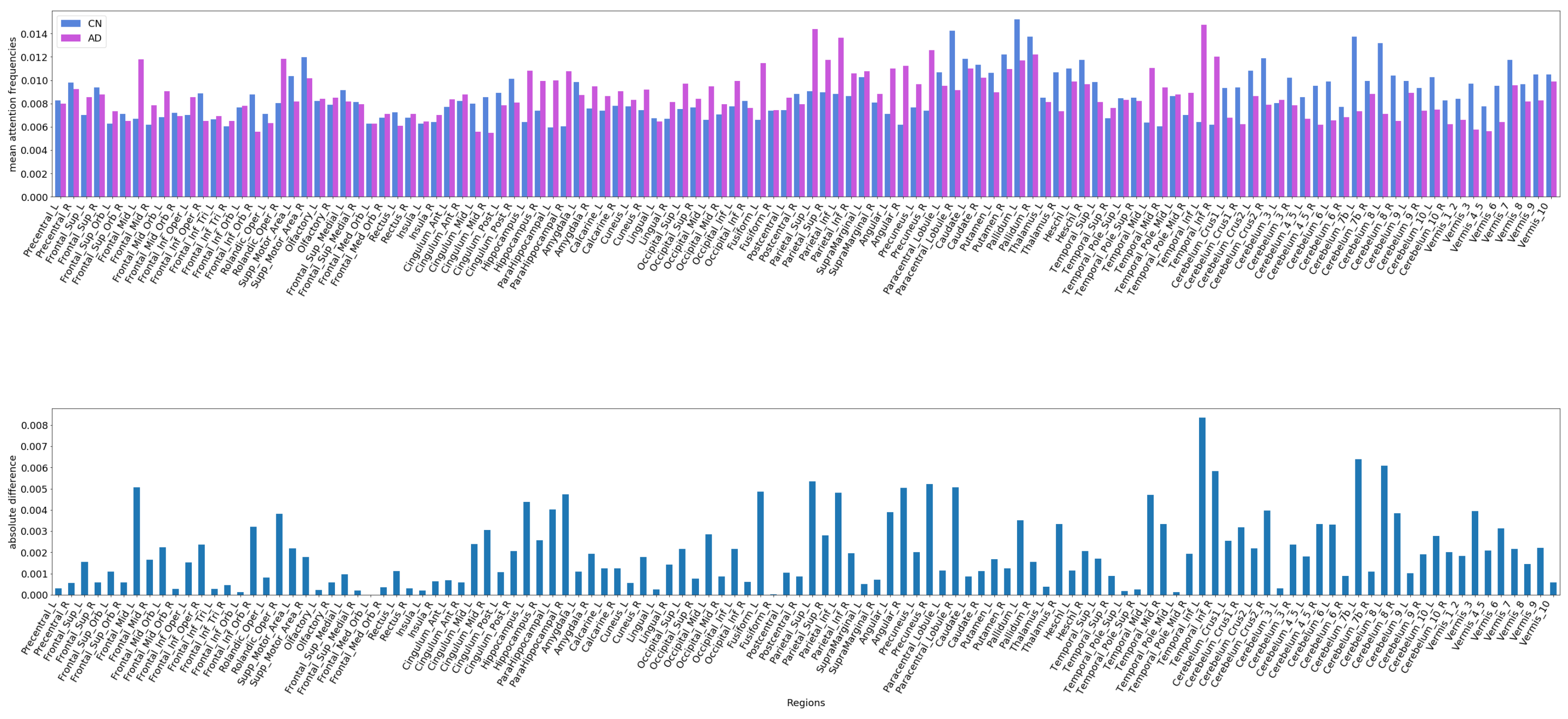


Figure 4: *Marginal distribution of the projected attention columns for the mean of the CN and AD subjects across all heads (top). The absolute difference between each region (bottom).*

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

[1] Vaswani et al. Attention is all you need. arXiv:1706.03762 (2017).

[2] https://www.alz.org/