Multi-Modal Graph Neural Network with Transformer-Guided Adaptive Diffusion for Preclinical Alzheimer Classification

Jaeyoon Sim[1] Minjae Lee[1] Guorong Wu[2] Won Hwa Kim[1]

[1]Pohang University of Science and Technology, Pohang, South Korea
[2]University of North Carolina at Chapel Hill, Chapel Hill, USA

Introduction

- **Key Idea: Guiding diffusion process at each node by a downstream transformer via diffusion-kernel and multi-head attention.**
- **Problem: Limitations in interpreting the brain networks in a scenario with multiple imaging biomarkers.**
  - Convolutional approaches ineffectively aggregate information from distant neighborhoods, while attention-based methods exhibit deficiencies in capturing node- centric information, particularly in retaining critical properties from pivotal nodes.
  - These shortcomings reveal challenges for identifying disease-specific variation from diverse features from different modalities.
- **Contribution:**
  1. We propose a novel framework to aggregate both short- and long- range properties for better prediction of graph labels.
  2. We demonstrate superior performance on graph classification in comparisons to the state-of-the-art methods.
  3. We show interpretability on the brain networks in a scenario with multiple imaging biomarkers.

Preliminary: Graph Kernel Convolution

- An undirected graph $G = \{V, E\}$ with N nodes comprises a node set V and an edge set E. A symmetric adjacency matrix A and a diagonal degree matrix D can be computed from E. A graph Laplacian is defined as $L = D - A$. It has a complete set of orthonormal eigenvectors $U = [u_1|u_2|...|u_N]$ and corresponding real and non-negative eigenvalues $0 = \lambda_1 \leq ... \leq \lambda_N$, so does the normalized Laplacian $\hat{L} = D^{-1/2}LD^{-1/2}$.
- From Spectral Graph Theory, the choice of a kernel function determines specific graph characteristics. A prominent heat-kernel between nodes p and q is spanned by U as

$$h_s(p,q) = \sum_{i=1}^{N} e^{-s\lambda_i} u_i(p) u_i(q) \quad (1)$$

where u_i is the i-th eigenvector. The kernel $e^{-s\lambda_i}$ captures smooth transition between nodes within the scale s as a low-pass filter. Graph Fourier transform, i.e., $\hat{x} = U^T x$, defines the graph convolution $*$ of a signal $x(p)$ with a filter h_s as

$$h_s * x(p) = \sum_{i=1}^{N} e^{-s\lambda_i} \hat{x}(i) u_i(p) \quad (2)$$

whose band-width is controlled by the scale s.

GNN with Transformer-guided Adaptive Diffusion (GTAD)

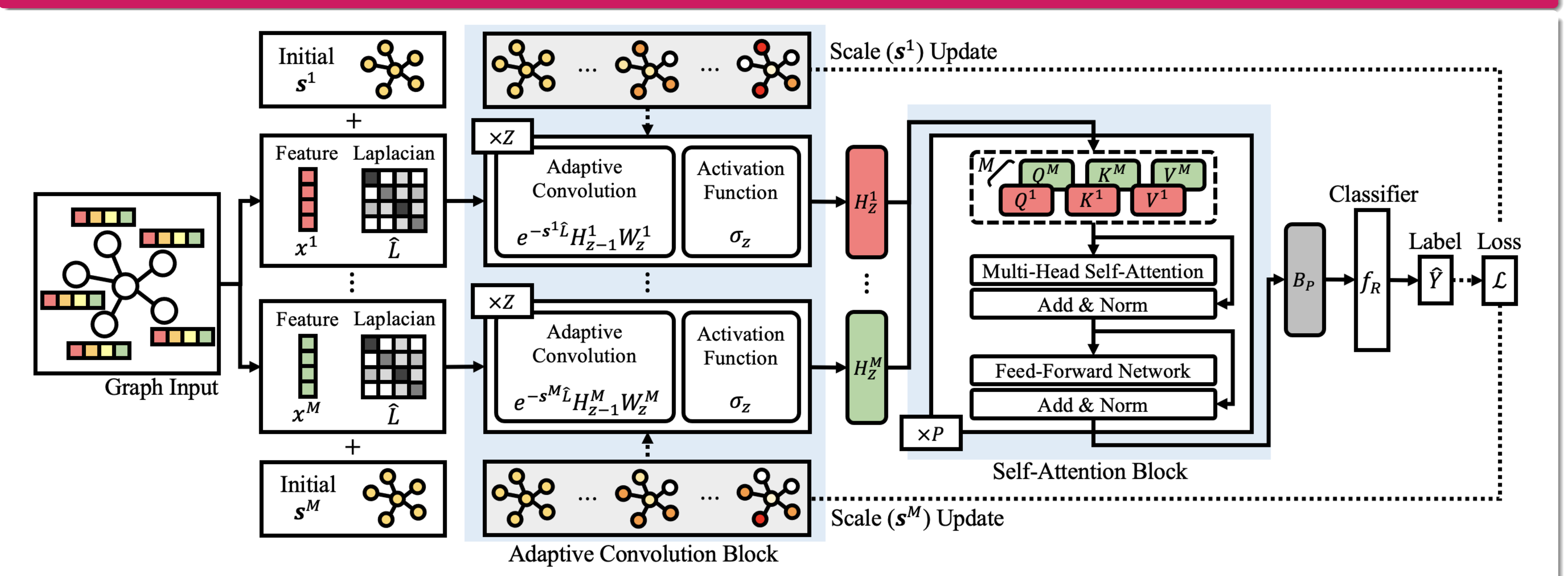

Figure: Illustration of our framework (GTAD).

We introduce a novel end-to-end framework GTAD that learns node-centric parameters of a diffusion kernel which are governed by a transformer.

- **Modality-wise Adaptive Convolution Block.** Consider G given as $\hat{L} \in \mathbb{R}^{N\times N}$, a set of features (i.e., imaging measures) $X = \{\boldsymbol{x}^m\}_{m=1}^M$ defined on N nodes from M modalities, a set of trainable multi-variate scales $\{\boldsymbol{s}^m\}_{m=1}^M$ where $\boldsymbol{s}^m \in \mathbb{R}^N$ and a graph label Y. Each encoder consists of multiple graph convolution layers that adaptively aggregate features for each node with a non-linear activation function σ_z as

$$H_z^m = \sigma_z(e^{-\boldsymbol{s}^m \hat{L}} H_{z-1}^m W_z^m). \quad (3)$$

- **Modality-wise Self-Attention Block.** The obtained embeddings $\{H_Z^m\}_{m=1}^M$ are inputted to an attention block to compute node-wise attention scores. Using the self-attention scores, a self-attention value is computed as

$$\phi(Q^m, K^m, V^m) = \sigma(\frac{Q^m K^{mT}}{\sqrt{C}})V^m. \quad (4)$$

- **Transformer-Guided Scale Update.** To update a scale s_n^m at the n-th node for the m-th encoder, the objective function is defined by cross-entropy between the true value Y_{tj} and the prediction $\hat{Y}_{tj}$.

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} Y_{tj} \ln \hat{Y}_{tj} + \alpha \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} \mathbb{1}_{s<0}|s_n^m|. \quad (5)$$

Update of the modality-specific scales is performed as $s \leftarrow s - \beta\frac{\partial \mathcal{L}}{\partial s}$ via gradient-descent with a learning rate β.

Alzheimer's Disease Neuroimaging Initiative (ADNI)

- On the same parcellation, region-wise imaging features such as Standard Uptake Value Ratio (SUVR) of metabolic intensity from FDG-PET, β-Amyloid protein from Amyloid-PET and cortical thickness from MRI were measured.
- Diagnostic labels: Control (CN), Significant Memory Concern (SMC), Early Mild Cognitive Impairment (EMCI)

Table: Demographics of the preclinical ADNI dataset.

Category	CN	SMC	EMCI
# of subjects	333	172	414
Gender (Male / Female)	156 / 177	62 / 110	240 / 174
Age (Mean±Std)	73.0 ± 5.9	71.7 ± 5.2	71.0 ± 7.7

Classification Result

Table: Preclinical AD classification performance (CN/SMC/EMCI) on ADNI data.

Modalities	Cortical Thickness & β-Amyloid			Cortical Thickness & FDG		
Methods	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
GCN	0.861±0.04	0.772±0.06	0.780±0.06	0.873±0.02	0.802±0.02	0.813±0.03
GAT	0.896±0.01	0.827±0.03	0.839±0.02	0.882±0.02	0.811±0.03	0.844±0.03
GraphHeat	0.868±0.02	0.777±0.05	0.797±0.04	0.887±0.03	0.821±0.04	0.834±0.03
GDC	0.858±0.02	0.767±0.03	0.786±0.04	0.842±0.01	0.743±0.02	0.765±0.03
ADC	0.906±0.02	0.835±0.03	0.861±0.04	0.896±0.01	0.831±0.01	0.847±0.02
LSAP	0.911±0.01	0.847±0.03	0.872±0.02	0.934±0.02	0.899±0.05	0.904±0.03
NodeFormer	0.916±0.02	0.856±0.04	0.865±0.02	0.944±0.01	0.913±0.03	0.921±0.02
DIFFormer	0.930±0.01	0.877±0.03	0.900±0.02	0.954±0.01	0.923±0.02	0.944±0.01
SGFormer	0.941±0.01	0.894±0.03	0.911±0.02	0.959±0.01	0.931±0.01	0.945±0.01
GTAD (Ours)	**0.945±0.02**	**0.901±0.03**	**0.919±0.02**	**0.963±0.01**	**0.935±0.02**	**0.948±0.01**
Modalities	**β-Amyloid & FDG**			**All Imaging Features**		
Methods	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
GCN	0.880±0.01	0.806±0.02	0.813±0.02	0.888±0.02	0.816±0.02	0.826±0.02
GAT	0.877±0.02	0.815±0.03	0.814±0.04	0.912±0.01	0.858±0.02	0.864±0.02
GraphHeat	0.880±0.02	0.804±0.05	0.824±0.03	0.893±0.02	0.824±0.03	0.839±0.03
GDC	0.866±0.02	0.787±0.03	0.790±0.03	0.867±0.02	0.779±0.03	0.799±0.02
ADC	0.910±0.01	0.865±0.02	0.856±0.02	0.904±0.02	0.855±0.04	0.858±0.02
LSAP	0.922±0.02	0.862±0.05	0.893±0.03	0.912±0.01	0.844±0.04	0.879±0.02
NodeFormer	0.931±0.01	0.887±0.03	0.893±0.03	0.938±0.02	0.900±0.03	0.902±0.03
DIFFormer	0.951±0.01	0.919±0.03	0.933±0.02	0.953±0.01	0.920±0.02	0.936±0.02
SGFormer	0.954±0.01	0.923±0.03	0.936±0.02	0.951±0.01	0.911±0.02	0.933±0.02
GTAD (Ours)	**0.962±0.01**	**0.935±0.02**	**0.946±0.02**	**0.963±0.01**	**0.943±0.01**	**0.941±0.02**

Interpretation of the Trained GTAD

- **Discussion on the Scales**
  - The trained model yields node-wise optimized scales, where each node corresponds to a specific ROI in the brain.

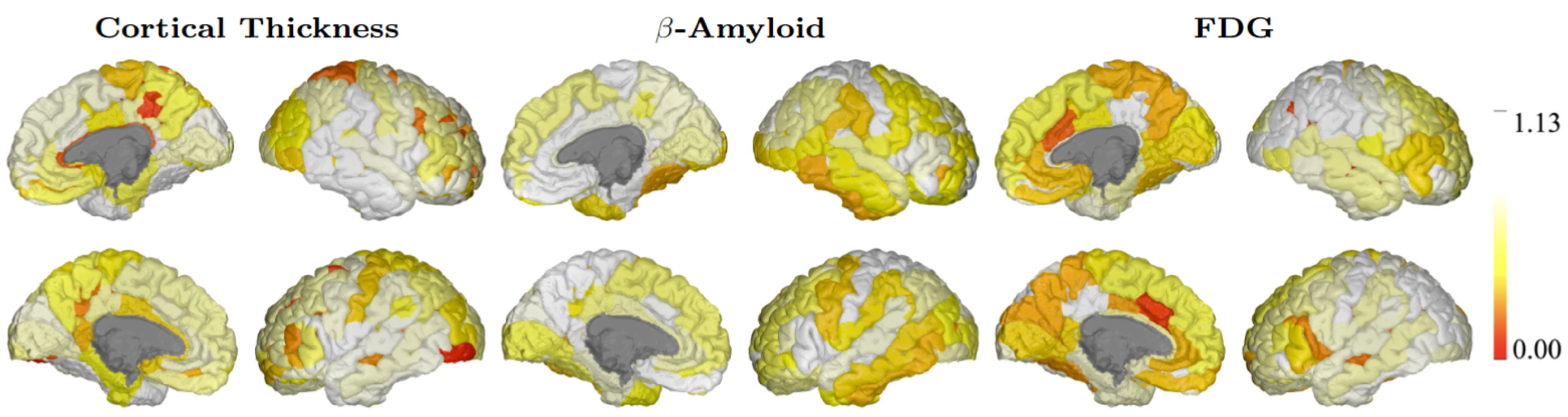

Figure: Visualization of learned scales on the cortical regions of left (top) and right (bottom) hemispheres.

- **Pre-clinical AD via ROI Attention**
  - From the attention block, each ROI gains long-range characteristics from other ROIs by modality-wise attention mechanism.
  - Most relevant ROIs in Preclinical AD prediction can be detected by total attention scores that represent the intensity of attention at each ROI in the brain.

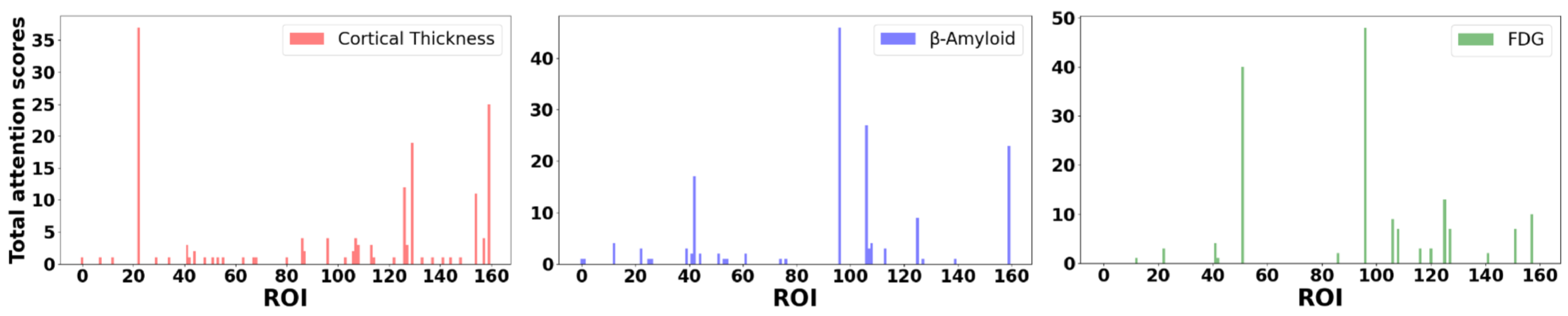

Figure: Distribution of attention scores across all brain regions with cortical thickness (left), β-Amyloid (center) and FDG (right).

- **Ablation Study on the Blocks**
  - To explore the effect of each block, ablation study on convolution types and attention types for preclinical AD classification is given.

Table: Performance comparisons of different blocks. For attention block, our multi-modal (MM) attention and existing position-wise attention are compared.

Convolution Block	MM Attention	Accuracy	Precision	Recall
Multi-Layer Perceptron	✗	0.939±0.03	0.893±0.05	0.913±0.04
	✓	0.947±0.02	0.906±0.04	0.933±0.02
Graph Convolution Layer	✗	0.899±0.01	0.835±0.03	0.849±0.03
	✓	0.900±0.01	0.834±0.03	0.852±0.02
Adaptive Convolution Layer (Ours)	✗	0.945±0.03	0.903±0.05	0.922±0.04
	✓	**0.963±0.01**	**0.943±0.01**	**0.941±0.02**