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

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

[1] Pohang University of Science and Technology, Pohang, South Korea
{simjy98, lalswo010, wonhwa}@postech.ac.kr
[2] University of North Carolina at Chapel Hill, Chapel Hill, USA

Abstract. The graphical representation of the brain offers critical insights into diagnosing and prognosing neurodegenerative disease via relationships between regions of interest (ROIs). Despite recent emergence of various Graph Neural Networks (GNNs) to effectively capture the relational information, there remain inherent limitations in interpreting the brain networks. Specifically, convolutional approaches ineffectively aggregate information from distant neighborhoods, while attention-based methods exhibit deficiencies in capturing node-centric information, particularly in retaining critical characteristics from pivotal nodes. These shortcomings reveal challenges for identifying disease-specific variation from diverse features from different modalities. In this regard, we propose an integrated framework guiding diffusion process at each node by a downstream transformer where both short- and long-range properties of graphs are aggregated via diffusion-kernel and multi-head attention respectively. We demonstrate the superiority of our model by improving performance of pre-clinical Alzheimer's disease (AD) classification with various modalities. Also, our model adeptly identifies key ROIs that are closely associated with the preclinical stages of AD, marking a significant potential for early diagnosis and prevision of the disease.

1 Introduction

Amyloid deposition and neurofibrillary tangles disrupt neural connections, indicating the potential of using brain connectomes in neuroimaging to identify early signs of brain disorders such as Alzheimer's Disease (AD) [4,19,11,15]. The white-matter connectome establishes structural interconnections between distinct anatomical regions of interest (ROI) within the brain, and it constructs a brain network per subject. As the brain network guides pathological variation on the ROIs [17], it is critical to incorporate the connection information in addition to regional measures from other images, e.g., magnetic resonance image (MRI) and positron emission tomography (PET) scans with various tracers, to characterize preclinical/early symptoms of AD.

A typical representation of a brain network involves a graph, mathematically formulated by its nodes and edges. The nodes correspond to each ROI, and connectome features, e.g., number of tracts between ROIs and fractional anisotropy

(FA), determine the edges with strength (i.e., edge weight). The graph representations of brain networks, together with image-derived measurements at each ROI, naturally lift the utilization of a graph neural network (GNN) for disease classification and characterization.

Traditionally well-known GNNs [12,22] incorporate the structure of graphs via graph convolution, and later methods use kernel convolutions based on diffusion process to obtain better representations [26,7,27,18]. These conventional methods rely on the homophily condition that node features locally connected by the edges behave similarly, overlooking the relationships between nodes far apart. Graph Transformers use global attention to capture far-distance influence beyond neighboring nodes within the graph [24,25,23], however, these methods often disregard sufficient expressive power of the central nodes, lacking interpretation of the result. The problem becomes more challenging when using multiple biomarker magnitudes as nodal features, as the interaction among multiple biomarkers and their diverse characteristics introduce heterogeneity, further complicating the analysis.

Therefore, it is necessary to develop an interpretable multi-modal method to capture both local characteristic and global graph-level information. The architecture we propose, i.e., **G**NN with **T**ransformer-guided **A**daptive **D**iffusion (GTAD), addresses the issues above by learning node-centric parameters of a diffusion kernel which are governed by a transformer. The encoder part of GTAD first obtains locally-effective representation of each node per imaging modality with a heat-kernel, which is later mixed by multi-head attention in the transformer to achieve globally-effective representation for classification. The node-wise kernel parameter as well as the attention scores let us interpret the local and global graph characteristics learnt by the model, especially when each node corresponds to anatomical ROI in the brain network.

The key contributions of our work are **1)** proposing a novel framework to aggregate both short- and long-range properties for better prediction of graph labels, **2)** demonstrating superior performance on graph classification in comparison to the state-of-the-art methods, and **3)** showing interpretability on the brain networks in a scenario with multiple imaging biomarkers. Experiments on structural brain networks from Diffusion Tensor Imaging (DTI) and ROI measures from functional imaging from Alzheimer's Disease Neuroimaging Initiative (ADNI) study show that the developed framework yields practical results for pre-clincal AD classification and interpretation to facilitate early diagnosis and prevention of AD.

2 Method

Prelim: 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, whose elements encode connectivity among its nodes and the volume of each node respectively. A graph Laplacian is defined as $L = D - A$, which is real and

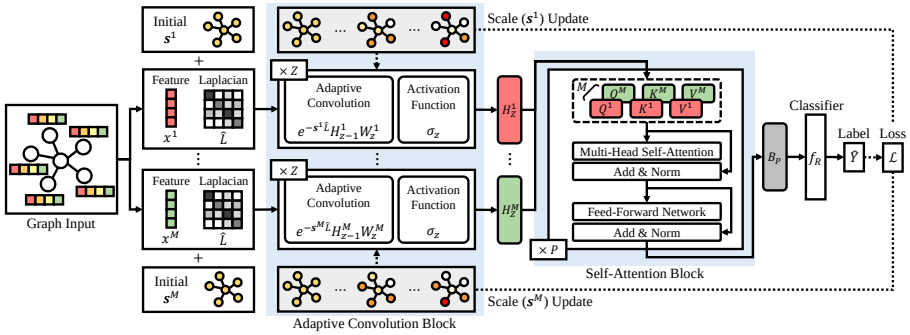


Fig. 1. Illustration of GTAD. A graph (as $\hat{L}$) and node feature $\boldsymbol{x}^m$ are inputted to m-th encoder at the adaptive convolution block. Then, all outputs $\{H_Z^m\}_{m=1}^M$ from this block are inputted to the self-attention block, producing an output B_P. Finally, the B_P is entered into a classifier f_R which yields a prediction $\hat{Y}$. To adaptively adjust the node-wise scales for each modality, the loss $\mathcal{L}$ from $\hat{Y}$ is backpropagated to update m-th encoder with scales $\boldsymbol{s}^m$.

positive semi-definite. It has a complete set of orthonormal eigenvectors $U = [u_1|u_2|...|u_N]$ and corresponding real and non-negative eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$, so does the normalized Laplacian $\hat{L} = D^{-1/2}LD^{-1/2}$.

From Spectral Graph Theory [2], the choice of a kernel function determines specific graph characteristics. For example, a prominent heat-kernel between nodes p and q is spanned by U as

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

where u_i is the i-th eigenvector. The kernel $e^{-s\lambda_i}$ captures smooth transition between the nodes within the scale s as a low-pass filter. Using convolutional theorem [14], 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

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

whose band-width is controlled by the scale s.

Modality-wise Adaptive Convolution Block. Consider a graph G given as a normalized Laplacian $\hat{L} \in \mathbb{R}^{N\times N}$, 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. To obtain representations from individual modality, encoders from our model take $\hat{L}$ and $\boldsymbol{x}^m$ for $m \in \{1, \ldots, M\}$ as inputs and perform convolution with heat kernel in Eq. (2) across all nodes and modalities respectively. Each encoder consists of multiple graph

convolution layers that adaptively aggregate features for each node with a non-linear activation function σ_z as

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

where H_z^m is an output from z-th convolution layer for m-th modality with $H_0^m = \boldsymbol{x}^m$, and W_z^m is a weight matrix. Within our framework, $\boldsymbol{s}^m$ is made trainable to capture local characteristic of individual node for different modalities. Since the Eq. (3) is an operation in a single convolution layer, better representation for the original feature $\boldsymbol{x}^m$ can be achieved by stacking Z of them.

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. Here, the multi-head self-attention module is inherited from the transformer layer [21]. Unlike typical use of transformer [5,6], each head is assigned to an individual modality to integrate *long-range* information from other nodes, which is not captured in the convolution block.

The input of attention module consists of query $Q^m \in \mathbb{R}^{N\times C}$, key $K^m \in \mathbb{R}^{N\times C}$, and value $V^m \in \mathbb{R}^{N\times C}$ from modality-wise embedding H_Z^m, and C is the dimension for hidden units. The self-attention scores are computed as $Q^m {K^m}^T/\sqrt{C}$, and softmax σ is applied to obtain weights on the values. Using the self-attention scores, a self-attention value is computed as

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

As a single attention head is assigned to a single modality, the global characteristics for all modalities is averaged with a multi attention function as $\Phi(Q, K, V) = [h^1|h^2|\ldots|h^m]W^\Phi$ where $h^m = \phi(Q^m W^{Q^m}, K^m W^{K^m}, V^m W^{V^m})$. Here, W^{Q^m}, W^{K^m}, W^{V^m} and W^Φ are weight matrices for Q^m, V^m, K^m and $\Phi(\cdot)$ respectively. Thus, multi-modal self-attention enables the model to jointly attend to information from different modalities across various ROIs in long-range.

The attention block contains a fully connected feed-forward module $\Psi(\cdot)$, which consists of multiple linear transformations with a non-linear activation function in between. To stabilize learning process and improve generalization, residual connections [8] are used, followed by layer normalization $f_L[\cdot]$ [1]. Therefore, a comprehensive context across all nodes is captured as

$$B_p = f_L[f_L[B_{p-1} + \Phi(B_{p-1})] + \Psi(f_L[B_{p-1} + \Phi(B_{p-1})])] \tag{5}$$

where B_p is an output from p-th attention layer, and multi-modal representations $\{H_Z^m\}_{m=1}^M$ are used as Q, K and V for B_0. To capture complex dependencies in the input modalities, multiple attention layers, e.g., P-layers, can be stacked.

Transformer-Guided Scale Update. Consider a set of graphs $\{G_t\}_{t=1}^T$ with corresponding labels $\{Y_t\}_{t=1}^T$, and learning a classification model finds a function $f(G_t) = Y_t$. For this, a downstream classifier $f_R(\cdot)$ takes the B_P from Transformer as an input and returns a prediction $\hat{Y}_{tj}$ at the j-th class for the t-th

sample, which is computed via $Softmax$ as

$$\hat{Y}_{tj} = \frac{f_R(B_P)_{tj}}{\sum_{j' \in J} f_R(B_P)_{tj'}} \tag{6}$$

where J is a class size. 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 $\hat{Y}_{tj}$. With an ℓ_1 norm regularization on s_n^m to impose positive scale for the heat-kernel, the overall objective function $\mathcal{L}$ is defined as

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{J} Y_{tj} \ln \hat{Y}_{tj} + \alpha \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{N} \mathbb{1}_{s<0}|s_n^m| \tag{7}$$

where α is a user-parameter and $\mathbb{1}$ is an indicator function. 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.

3 Experiments

Dataset. Neuroimages of T=919 preclinical AD subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study were used for the experiment. Each brain was partitioned into 148 cortical regions and 12 sub-cortical regions with Destrieux atlas [3] with MRI, and Tractography on diffusion weighted imaging (DWI) was applied to calculate the number of white matter fibers connecting the 160 brain regions to construct brain network. On the same parcellation, region-wise imaging features such as Standard Uptake Value Ratio (SUVR) [20] of metabolic intensity from FDG-PET, β-Amyloid protein from Amyloid-PET and cortical thickness from MRI were measured. Each subject was assigned to Control (CN, T=333), Significant Memory Concern (SMC, T=172) and Early Mild Cognitive Impairment (EMCI, T=414) for group comparisons.

Setup. We designed various 3-way classifications to classify the pre-clinical groups using various combinations of biomarkers. 5-fold cross validation was used to obtain unbiased results, and accuracy, precision and recall in their mean were computed for evaluation. As the baselines, we categorized GNNs into three groups and adopted them; 1) Convolution-based GNNs such as GCN [12] and GAT [22], 2) GNNs with graph diffusion such as GraphHeat [26], GDC [7], ADC [27] and LSAP [18], and 3) Graph transformers such as NodeFormer [24], DIFFormer [23] and SGFormer [25]. More details are given in the supplementary.

Classification Result. The performance comparisons between our model and nine baselines across four experiments are reported in Table 1. As shown in Table 1, aggregating both local (i.e., short-range) features by adaptively learned modality-wise scales and global (i.e., long-range) information by global attentions performed the best in all experimental cases, and accuracy from most experiments showed over 96% except for the case using cortical thickness and β-Amyloid. Notably, GTAD outperformed the outstanding transformers in preclinical AD prediction, indicating that our model is more suitable on the brain

Table 1. Preclinical AD classification performance (CN/SMC/EMCI) on ADNI data.

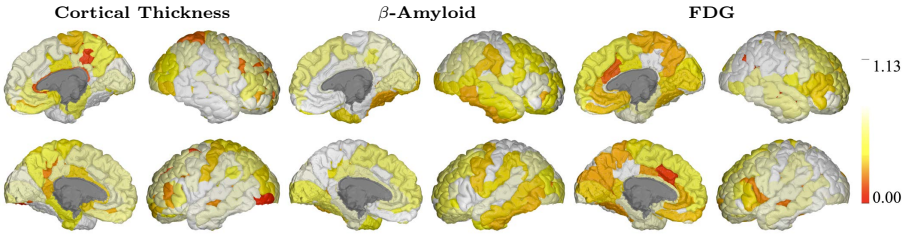
Modalities	**Cortical Thickness & β-Amyloid**			**Cortical Thickness & FDG**		
Methods	**Accuracy**	**Precision**	**Recall**	**Accuracy**	**Precision**	**Recall**
GCN [12]	0.861±0.04	0.772±0.06	0.780±0.06	0.873±0.02	0.802±0.02	0.813±0.03
GAT [22]	0.896±0.01	0.827±0.03	0.839±0.02	0.882±0.02	0.811±0.03	0.844±0.03
GraphHeat [26]	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 [7]	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 [27]	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 [18]	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 [24]	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 [23]	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 [25]	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 [12]	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 [22]	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 [26]	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 [7]	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 [27]	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 [18]	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 [24]	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 [23]	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 [25]	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**

network even under difficult conditions (i.e., prediction for early stages in AD given multiple imaging scans). Also, the stability of our model can be explained by low standard deviations for all evaluations within 5-folds.

4 Interpretation of the Trained GTAD

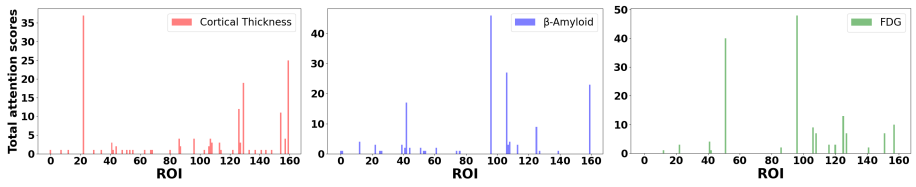
Discussion on the Scales. In the pre-clinical AD classification, the trained model yields node-wise optimized scales, where each node corresponds to a specific ROI in the brain. As the trained scales denote the optimal ranges of ROI-wise neighborhood for each modality, they represent modality-wise characteristics across all ROIs providing the interpretability of GTAD. Therefore, while ROIs with small trained scales require information from neighboring ROIs on the classification, ROIs with large scales need distant features as they are less effective individually. The learned scales on brain regions per modality are visualized in Fig. 2. Even for the same region in the brain, the local ranges are set differently depending on the modalities, which provides multi-dimensional understandings of subnetwork for AD progression.

In addition to the visualization of localized scales, 8 ROIs with the smallest scales for each modality are listed in the bottom of Fig. 2. Using the ROI-wise optimized scales with all biomarkers, GTAD selected most independent ROIs in the subcortical regions (i.e., *thalamus, putamen and globus pallidus*), temporal regions (i.e., *inferior, superior and occipito temporal regions*), frontal regions (i.e., *middle, superior and orbital regions*) and other important regions that are closely linked to AD. Based on these results, the ROIs with small scales are significantly important in interpreting the classification results depending on the characteristics that each imaging modality captures.



Cortical Thickness		β-Amyloid		FDG	
ROI	**Scale**	**ROI**	**Scale**	**ROI**	**Scale**
(R) S.cingul.Marginalis	0.0089	(R) Lat.Fis.ant.Horizont	0.0662	(L) sub.putamen	0.0187
(L) G&S.occipital.inf	0.0153	(L) S.oc.middle&Lunatus	0.0771	(L) G&S.cingul.Mid.Ant	0.0231
(L) S.front.sup	0.0252	(R) S.calcarine	0.0868	(R) S.temproal.sup	0.0254
(R) S.suborbital	0.0254	(R) sub.thalamus	0.1124	(R) sub.globus.pallidus	0.0255
(L) S.oc.temp.med&Lingual	0.0387	(R) G.oc.temp.lat.fusifor	0.1159	(L) sub.globus.pallidus	0.0303
(R) S.pericallosal	0.0387	(L) S.calcarine	0.1166	(R) G&S.cingul.Mid.Ant	0.0451
(R) S.front.middle	0.0420	(R) G.temporal.inf	0.1208	(L) S.temporal.sup	0.0609
(R) G.parietal.sup	0.0500	(R) S.orbital.lateral	0.1224	(L) S.calcarine	0.0662

Fig. 2. Top: Visualization of learned scales on the cortical regions of left (top) and right (bottom) hemispheres. Bottom: 8 Localized ROIs with the smallest trained scales for classification. (L) and (R) denote left and right hemisphere, respectively.



Cortical Thickness		β-Amyloid		FDG	
ROI	**IR**	**ROI**	**IR**	**ROI**	**IR**
(L) G.oc.temp.med.Lingual	23.13 %	(R) G.oc.temp.med.Lingual	28.75 %	(R) G.oc.temp.med.Lingual	30.00 %
(R) sub.putamen	15.63 %	(R) G.subcallosal	16.88 %	(L) S.collat.transv.post	25.00 %
(R) S.interm.prim.Jensen	11.88 %	(R) sub.putamen	14.38 %	(R) S.collat.transv.post	08.13 %
(R) S.front.inf	07.50 %	(L) Pole.occipital	10.63 %	(R) sub.hippocampus	06.25 %
(L) sub.globus.pallidus	06.88 %	(R) S.collat.transv.post	05.63 %	(R) G.subcallsoal	05.63 %

Fig. 3. Top: Distribution of attention scores across all brain regions with cortical thickness (left), β-Amyloid (center) and FDG (right). Bottom: Corresponding ROIs with the 5 highest attention scores for classification. Importance Rate (IR) indicates how many ROIs pay attention. (L) and (R) denote left and right hemisphere, respectively.

Pre-clinical AD via ROI Attention. From the attention block, each ROI gains long-range characteristics from other ROIs by modality-wise attention mechanism. In this regard, 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. Here, the total attention score is defined as the result of calculating how many ROIs give the highest attention score to the corresponding ROI. In Fig. 3, distributions of these scores per ROI show which ROIs are making long-range influences. Since the distributions of total attention score vary across all modalities, we can explain which ROI is most important from a specific modality in making predictions.

Table 2. 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**

Top 5 ROIs with the highest importance rate, i.e., the ratio of total attention scores, are listed in the bottom of Fig. 3. Notably, *Lingual gyrus* was detected with the highest importance rate from all modalities in common. *Lingual gyrus*, which is especially related to processing logical order of events and encoding visual memories, is belong to temporal regions and highly linked to AD [10,13]. In particular, *hippocampus* showed a high importance rate in FDG, and *putamen* also simultaneously exhibited a high score in Cortical Thickness and β-Amyloid. These regions are one of the first areas to be affected in AD, indicating that they are closely associated with pre-clinical AD [16,9]. From these results, we can observe the key regions in distinguishing the progressions of neurodegenerative brain diseases through modality-wise attentions.

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 in Table 2. For the convolution block, Multi-Layer Perceptron (MLP), Graph Convolution and Adaptive Graph Convolution are compared with a choice of multi-head or position-wise attention which was obtained by inputting concatenated features into a single encoder [21]. The flexible capture of local properties for each node using adaptive graph convolution exhibits better expressive power with 94.5% accuracy. This metric was boosted up to 96.3% by the multi-modal attention, demonstrating capturing local and global features with separate blocks but training them jointly if highly effective. As the MLP connects all ROIs globally and the Graph Convolution is not adaptively guided by the transformer, the effect of the multi-modal attention was very marginal.

5 Conclusion

In this work, we proposed a novel end-to-end framework GTAD to dynamically define node-centric ranges per imaging modality via diffusion kernel, guided by a subsequent transformer. Our framework captures local characteristics on graphs by flexibly optimizing node-wise scales separately on imaging modalities, and obtains a global representation by employing multi-modal self-attention, which guides the model to better prediction. Leveraging multiple imaging measures, GTAD demonstrates superiority as evidenced by improved performance in preclinical AD classification, and the results identifies disease-specific variation through AD-specific key ROIs in the brain.

Acknowledgments. This research was supported by NRF-2022R1A2C2092336 (50%), RS-2022-II2202290 (20%), RS-2019-II191906 (AI Graduate Program at POSTECH, 10%) funded by MSIT, RS-2022-KH127855 (10%), RS-2022-KH128705 (10%) funded by MOHW from South Korea.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. Advances in Neural Information Processing Systems (2016)
2. Chung, F.R.: Spectral graph theory, vol. 92. American Mathematical Soc. (1997)
3. Destrieux, C., Fischl, B., Dale, A., Halgren, E.: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage **53**(1), 1–15 (2010)
4. DeTure, M.A., Dickson, D.W.: The neuropathological diagnosis of Alzheimer's disease. Molecular neurodegeneration **14**(1), 1–18 (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. Annual Meeting of the Association for Computational Linguistics (2019)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021)
7. Gasteiger, J., Weißenberger, S., Günnemann, S.: Diffusion improves graph learning. Advances in Neural Information Processing Systems **32** (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
9. de Jong, L.W., van der Hiele, K., Veer, I.M., Houwing, J., Westendorp, R., Bollen, E., de Bruin, P.W., Middelkoop, H., van Buchem, M.A., van der Grond, J.: Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. Brain **131**(12), 3277–3285 (2008)
10. Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A.: Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. Clinical Neurophysiology **126**(11), 2132–2141 (2015)
11. Kim, W.H., Adluru, N., Chung, M.K., Okonkwo, O.C., Johnson, S.C., Bendlin, B.B., Singh, V.: Multi-resolution statistical analysis of brain connectivity graphs in preclinical Alzheimer's disease. NeuroImage **118**, 103–117 (2015)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (2017)
13. Liu, X., Chen, W., Hou, H., Chen, X., Zhang, J., Liu, J., Guo, Z., Bai, G.: Decreased functional connectivity between the dorsal anterior cingulate cortex and lingual gyrus in Alzheimer's disease patients with depression. Behavioural brain research **326**, 132–138 (2017)
14. Oppenheim, A.V., Willsky, A.S., Nawab, S.H., Ding, J.J.: Signals and systems, vol. 2. Prentice hall Upper Saddle River, NJ (1997)

15. Park, J., Hwang, Y., Kim, M., Chung, M.K., Wu, G., Kim, W.H.: Convolving directed graph edges via Hodge Laplacian for brain network analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 789–799. Springer (2023)
16. Rao, Y.L., Ganaraja, B., Murlimanju, B., Joy, T., Krishnamurthy, A., Agrawal, A.: Hippocampus and its involvement in Alzheimer's disease: a review. 3 Biotech **12**(2), 55 (2022)
17. Ryyppö, E., Glerean, E., Brattico, E., Saramäki, J., Korhonen, O.: Regions of interest as nodes of dynamic functional brain networks. Network Neuroscience **2**(4), 513–535 (2018)
18. Sim, J., Jeon, S., Choi, I., Wu, G., Kim, W.H.: Learning to approximate adaptive kernel convolution on graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4882–4890 (2024)
19. Stam, C.J., De Haan, W., Daffertshofer, A., Jones, B., Manshanden, I., van Cappellen van Walsum, A.M., Montez, T., Verbunt, J., De Munck, J.C., Van Dijk, B.W., et al.: Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer's disease. Brain **132**(1), 213–224 (2009)
20. Thie, J.A.: Understanding the standardized uptake value, its methods, and implications for usage. Journal of Nuclear Medicine **45**(9), 1431–1434 (2004)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
22. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. International Conference on Learning Representations (2018)
23. Wu, Q., Yang, C., Zhao, W., He, Y., Wipf, D., Yan, J.: Difformer: Scalable (graph) transformers induced by energy constrained diffusion. International Conference on Learning Representations (2023)
24. Wu, Q., Zhao, W., Li, Z., Wipf, D.P., Yan, J.: Nodeformer: A scalable graph structure learning transformer for node classification. Advances in Neural Information Processing Systems **35**, 27387–27401 (2022)
25. Wu, Q., Zhao, W., Yang, C., Zhang, H., Nie, F., Jiang, H., Bian, Y., Yan, J.: Simplifying and empowering transformers for large-graph representations. Advances in Neural Information Processing Systems **36** (2024)
26. Xu, B., Shen, H., Cao, Q., Cen, K., Cheng, X.: Graph convolutional networks using heat kernel for semi-supervised learning. International Joint Conference on Artificial Intelligence (2019)
27. Zhao, J., Dong, Y., Ding, M., Kharlamov, E., Tang, J.: Adaptive diffusion in graph neural networks. Advances in Neural Information Processing Systems **34**, 23321–23333 (2021)