Brain Network Transformer

Xuan Kan [1] **Wei Dai** [1] **Hejie Cui** [1] **Zilong Zhang** [2] **Ying Guo** [3] **Carl Yang** [1]

Abstract

Human brains are commonly modeled as networks of regions of interest (ROIs) and their connections for brain functions and mental disorders understanding. Recently, Transformer-based models have been studied over different types of data, including graphs, shown to bring performance gains widely. In this work, we study Transformer-based models for brain network analysis. Driven by the unique properties of data, we model brain networks as graphs with nodes of fixed size and order, which allows us to (1) use connection profiles as node features to provide natural and low-cost positional information and (2) learn pair-wise connection strengths among ROIs with efficient attention weights across individuals that are predictive towards downstream analysis tasks. Moreover, we propose an ORTHONORMAL CLUSTERING READOUT operation based on self-supervised soft clustering and orthonormal projection. This design accounts for the underlying functional modules that determine similar behaviors among groups of ROIs, leading to distinguishable cluster-aware node embeddings and informative graph embeddings. Finally, we re-standardize the evaluation pipeline on the only one publicly available large-scale brain network dataset of ABIDE, to enable meaningful comparison of different models. Experiment results show clear improvements of our proposed BRAIN NETWORK TRANSFORMER on both the public ABIDE and our restricted ABCD datasets. The implementation is available at https://anonymous.4open.science/r/BrainTransformer.

[*]Equal contribution [1]Department of Computer Science, Emory University, Atlanta, US [2]University of International Business and Economics, Beijing, China [3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, US. Correspondence to: Carl Yang <j.carlyang@emory.edu>.

Workshop of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, 2022.

1. Introduction

Brain network analysis has been an intriguing pursuit for neuroscientists to understand human brain organizations and predict clinical outcomes (Satterthwaite et al., 2015; Wang et al., 2016; Wang & Guo, 2019; Bullmore & Sporns, 2009; Deco et al., 2011). Among various neuroimaging modalities, functional Magnetic Resonance Imaging (fMRI) is one of the most commonly used for brain network construction, where the nodes are defined as Regions of Interest (ROIs) given an atlas, and the edges are calculated as pairwise correlations between the blood-oxygen-level-dependent (BOLD) signal series extracted from each region (Smith et al., 2011; Simpson et al., 2013; Wang et al., 2016). Researchers observe that some regions can co-activate or co-deactivate simultaneously when performing cognitive-related tasks such as action, language, and vision. Based on this pattern, brain regions can be classified into diverse functional modules to analyze diseases towards their diagnosis, progress understanding and treatment.

Nowadays Transformer-based models have led a tremendous success in various downstream tasks across fields including natural language processing (Vaswani et al., 2017; Dai et al., 2019) and computer vision (Dosovitskiy et al., 2021; Chu et al., 2021). Recent efforts have also emerged to apply Transformer-based designs to graph representation learning. GAT (Velickovic et al., 2018) firstly adapts the attention mechanism to graph neural networks (GNNs) but only considers the local structures of neighboring nodes. Graph Transformer (Dwivedi & Bresson, 2021) injects edge information into the attention mechanism and leverages the eigenvectors of each node as positional embeddings. SAN (Kreuzer et al., 2021) further enhances the positional embeddings by considering both eigenvalues and eigenvectors and improves the attention mechanism by extending the attention from local to global structures. Graphomer (Ying et al., 2021), which achieves the first place on the quantum prediction track of OGB Large-Scale Challenge (Hu et al., 2020a), designs unique mechanisms for molecule graphs such as centrality encoding to enhance node features and spatial/edge encoding to adapt attention scores.

However, brain networks have several unique traits that make directly applying existing graph Transformer models impractical. First, a brain network is a correlation matrix

defined on a complete graph. This impedes the designs like centrality, spatial, and edge encoding because each node in the brain network has the same degree and connects to every other node by a single hop. Second, in previous graph transformer models, eigenvalues and eigenvectors are commonly used as positional embeddings because they can provide identity and positional information for each node. Nevertheless, in brain networks, the connection profile, which is defined as each node's corresponding row in the brain network adjacency matrix, is recognized as the most effective node feature (Cui et al., 2022). This node feature naturally encodes both structural and positional information, making the aforementioned positional embedding design based on eigenvalues and eigenvectors redundant. The third challenge is scalability. Typically, the numbers of nodes and edges in molecule graphs are less than 50 and 2500, respectively. However, for brain networks, the node number is generally around 100 to 400, while the edge number can be up to 160,000. Therefore, operations like the generation of all edge features in existing graph transformer models can be time-consuming, if not infeasible.

In this work, we propose to develop BRAIN NETWORK TRANSFORMER (BRAINNETTF), which leverages the unique properties of brain network data to fully unleash the power of Transformer-based models for brain network analysis. Specifically, motivated by previous findings on effective GNN designs for brain networks (Cui et al., 2022), we propose to use the effective initial node features of connection profiles. Empirical analysis shows that connection profiles naturally provide positional features for Transformer-based models and avoid the costly computations of eigenvalues or eigenvectors. Moreover, recent work demonstrates that GNNs trained on learnable graph structures can achieve superior effectiveness and interpretability (Kan et al., 2022). Inspired by this insight, we propose to learn fully pairwise attention weights with Transformer-based models, which resembles the process of learning predictive brain network structures towards downstream tasks.

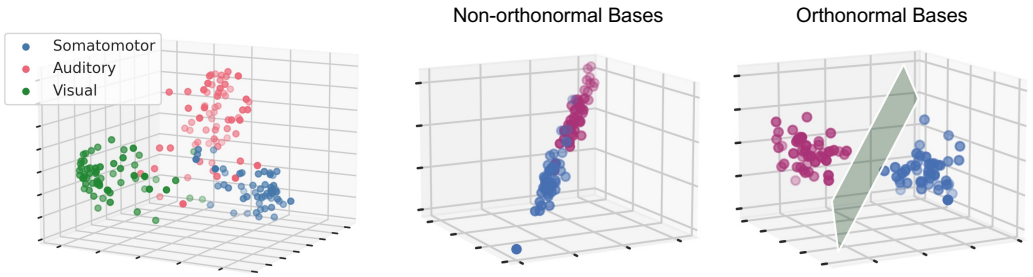
One step further, when GNNs are used for brain network analysis, a graph-level embedding needs to be generated through a readout function based on the learned node embeddings (Kawahara et al., 2017; Li et al., 2021; Cui et al., 2022). As is shown in Figure 1(a), a property of brain networks is that brain regions (nodes) belonging to the same functional modules often share similar behaviors regarding activations and deactivations in response to various stimulations (Caramazza & Coltheart, 2006). Unfortunately, the current labeling of functional modules is rather empirical and far from accurate. For example, (Akiki & Abdallah, 2019) provides more than 100 different functional module organizations based on hierarchical clustering. In order to leverage the natural functions of brain regions without the limitation of inaccurate functional module labels, we design a new global pooling operator, ORTHONORMAL CLUSTERING READOUT, where the graph-level embeddings are pooled from clusters of functionally similar nodes through soft clustering with orthonormal projection. Specifically, we first devise a self-supervised mechanism based on (Xie et al., 2016) to jointly assign soft clusters to brain regions while learning their individual embeddings. To further facilitate the learning of clusters and embeddings, we design an orthonormal projection and show its effectiveness in distinguishing embeddings across clusters, thus obtaining expressive graph-level embeddings after the global pooling, as illustrated in Figure 1(b).

Finally, the lack of open-access datasets has been a non-negligible challenge for brain network analysis. The strict access restrictions and complicated extraction/preprocessing of brain networks from fMRI data limit the development of machine learning models for brain network analysis. Specifically, among all the large-scale publicly available fMRI datasets in literature, ABIDE (Cameron et al., 2013) is the only one provided with extracted brain networks fully accessible without permission requirements. However, ABIDE is aggregated from 17 international sites with different scanners and acquisition parameters. This inter-site variability may conceal inter-group differences that are really meaningful, which is reflected in the unstable training performance and the significant gap between validation and testing performance in practice. To address these limitations, we propose to apply a stratified sampling method in the dataset splitting process and standardize a fair evaluation pipeline for meaningful model comparison on the ABIDE dataset. Our extensive experiments on this public ABIDE dataset and a restricted ABCD dataset (Casey et al., 2018) show significant improvements brought by our proposed BRAIN NETWORK TRANSFORMER.

2. Background and Related Work

2.1. GNNs for Brain Network Analysis

Recently, emerging attention has been devoted to the generalization of GNN-based models to fMRI-based brain network analysis (Li et al., 2019). GroupINN (Yan et al., 2019) utilizes a grouping-based layer to provide interpretability and reduce the model size. BrainGNN (Li et al., 2021) designs the ROI-aware GNNs to leverage the functional information in brain networks and uses a special pooling operator to select these crucial nodes. In addition, FBNetGen (Kan et al., 2022) considers the learnable generation of brain networks and explores the interpretability of the generated brain networks towards downstream tasks. Another benchmark paper (Cui et al., 2022) systematically studies the effectiveness of various GNN designs over brain network data. Different from other work focusing on static brain networks, STAGIN (Kim et al., 2021) utilizes GNNs with



(a) Node features projected to a 3D space with PCA. Colors indicate functional modules.

(b) Orthonormal bases can make indistinguishable nodes in non-orthonormal bases easily distinguishable.

Figure 1: Illustration of the motivations behind ORTHONORMAL CLUSTERING READOUT.

spatio-temporal attention to model dynamic brain networks extracted from fMRI data.

2.2. Graph Transformer

Graph Transformer raises many researchers' interest currently due to its outstanding performance in graph representation learning. Graph Transformer (Dwivedi & Bresson, 2021) firstly injects edge information into the attention mechanism and leverages the eigenvectors as positional embeddings. SAN (Kreuzer et al., 2021) enhances the positional embeddings and improves the attention mechanism by emphasizing neighbor nodes while incorporating the global information. Graphomer (Ying et al., 2021) designs unique mechanisms for molecule graphs and achieves the STOA performance. Besides, a fine-grained attention mechanism is developed for node classification (Zhao et al., 2021). Also, the Transformer is extended to larger-scale heterogeneous graphs with a particular sampling algorithm in HGT (Hu et al., 2020b). EGT (Hussain et al., 2021) further employs edge augmentation to assist global self-attention. In addition, LSPE (Dwivedi et al., 2022) leverages the learnable structural and positional encoding to improve GNNs' representation power, and GRPE (Park et al., 2022) enhances the design of encoding node relative position information in Transformer.

3. BRAIN NETWORK TRANSFORMER

3.1. Problem Definition

In brain network analysis, given a brain network $\boldsymbol{X} \in \mathbb{R}^{V\times V}$, where V is the number of nodes (ROIs), the model aims to make a prediction indicating gender, presence of a disease or other properties of the brain subject. The overall framework of our proposed BRAIN NETWORK TRANSFORMER is shown in Figure 2, which is mainly composed of two components, an L-layer attention module MHSA and a graph pooling operator OCREAD. Specifically, in the first component of MHSA, the model learns attention-enhanced node features $\boldsymbol{Z}^L$ through a non-linear mapping $\boldsymbol{X} \rightarrow \boldsymbol{Z}^L \in \mathbb{R}^{V\times V}$. Then the second component of OCREAD compresses the enhanced node embeddings $\boldsymbol{Z}^L$ to graph-level embeddings $\boldsymbol{Z}_G \in \mathbb{R}^{K\times V}$, where K is a hyperparameter representing the number of clusters. $\boldsymbol{Z}_G$ is then flattened and passed to a multi-layer perceptron for graph-level predictions. The whole training process is supervised with the cross-entropy loss.

3.2. Multi-Head Self-Attention Module (MHSA)

To develop a powerful Transformer-based model suitable for brain networks, two fundamental designs, the positional embedding and attention mechanism, need to be reconsidered to fit the natural properties of brain network data. In existing graph transformer models, the positional information is usually encoded via eigendecomposition, while the attention mechanism often combines node positions with existing edges to calculate the attention scores. However, for the dense (often fully connected) graphs of brain networks, eigendecomposition is rather costly, and the existence of edges is hardly informative.

ROI node features on brain networks naturally contain sufficient positional information, making the positional embeddings based on eigendecomposition redundant. Previous work on brain network analysis has shown that the connection profile $\boldsymbol{X}_{i\cdot}$ for node i, defined as the corresponding row for each node in the edge weight matrix $\boldsymbol{X}$, always achieves superior performance over others such as node identities, degrees or eigenvector-based embeddings (Li et al., 2021; Kan et al., 2022; Cui et al., 2022). With this node feature initialization, the self-connection weight $\boldsymbol{x}_{ii}$ on the diagonal is always equal to one, which encodes sufficient information to determine the position of each node in a fully connected graph based on the given brain atlas. To verify this insight, we also empirically compare the performance of the original Connection Profile with two variants concatenated with additional positional information in Table 1, *i.e.*, Connection Profile w/ Identity Feature and Connection Profile w/ Eigen

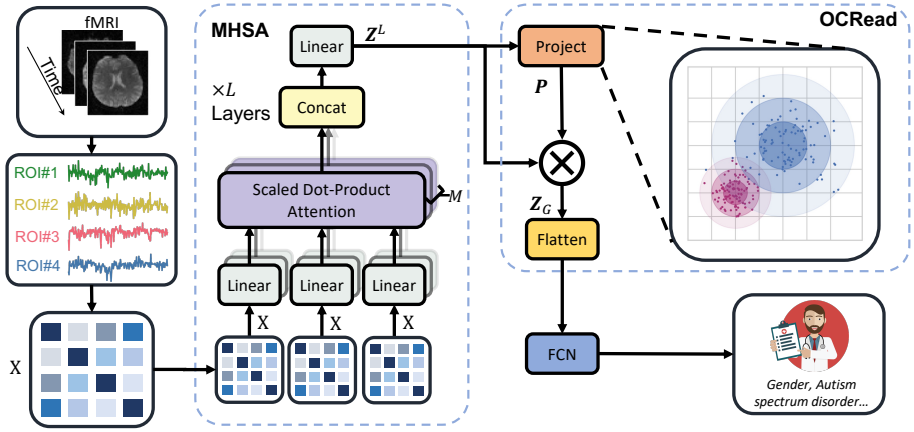


Figure 2: The overall framework of our proposed BRAIN NETWORK TRANSFORMER.

Feature. The results indeed show no benefit brought by the additional computations.

Node Feature	Dataset	
	ABIDE	ABCD
Connection Profile	76.4±1.2	94.3±0.7
Connection Profile w/ Identity Feature	75.4±1.9	94.5±0.6
Connection Profile w/ Eigen Feature	75.9±2.1	94.0±0.8

Table 1: The Performance (AUROC%) of Transformer with Different Node Features. **Connection profile** represents the corresponding row for each node in the adjacency matrix. **Identity feature** initializes a unique one-hot vector for each node. **Eigen feature** generates a k-dimensional feature vector for each node from the k eigenvectors based on the eigendecomposition on the adjacency matrix.

As for the attention mechanism, previous work (Cui et al., 2022) has empirically demonstrated that integrating edge weights into the attention score calculation can significantly degrade the effectiveness of attention on complete graphs, while the generation of edge-wise embedding can be unaffordable given a large number of edges in brain networks. On the other hand, the existence of edges provides no useful information for the computation of attention scores as well because all edges simply exist in complete graphs.

Based on the observations above, we design the basic BRAIN NETWORK TRANSFORMER by (1) adopting the Connection Profile as initial node features and eliminating any extra positional embeddings and (2) adopting the vanilla pair-wise attention mechanism without using edge weights or relative position information to learn a singular attention score for each edge in the complete graph.

Formally, we leverage a L-layer non-linear mapping module, namely Multi-Head Self-Attention (MHSA), to generate more expressive node features $\boldsymbol{Z}^L = \mathrm{MHSA}(\boldsymbol{X}) \in \mathbb{R}^{V \times V}$.

For each layer l, the output $\boldsymbol{Z}^l$ is obtained by

$$\begin{aligned} \boldsymbol{Z}^l &= (\|_{m=1}^{M} \boldsymbol{h}^{l,m}) \boldsymbol{W}_{\mathcal{O}}^l, \\ \boldsymbol{h}^{l,m} &= \mathrm{Softmax}\left(\frac{\boldsymbol{Q}^{l,m}(\boldsymbol{K}^{l,m})^\top}{\sqrt{d_{\mathcal{K}}^{l,m}}}\right)\boldsymbol{V}^{l,m}, \end{aligned} \tag{1}$$

where $\boldsymbol{Q}^{l,m} = \boldsymbol{W}_{\mathcal{Q}}^{l,m}\boldsymbol{Z}^{l-1}, \boldsymbol{K}^{l,m} = \boldsymbol{W}_{\mathcal{K}}^{l,m}\boldsymbol{Z}^{l-1}, \boldsymbol{V}^{l,m} = \boldsymbol{W}_{\mathcal{V}}^{l,m}\boldsymbol{Z}^{l-1}, \boldsymbol{Z}^0 = \boldsymbol{X}$, $\|$ is the concatenation operator, M is the number of heads, l is the layer index, $\boldsymbol{W}_{\mathcal{O}}^l, \boldsymbol{W}_{\mathcal{Q}}^{l,m}, \boldsymbol{W}_{\mathcal{K}}^{l,m}, \boldsymbol{W}_{\mathcal{V}}^{l,m}$ are learnable model parameters, and $d_{\mathcal{K}}^{l,m}$ is the first dimension of $\boldsymbol{W}_{\mathcal{K}}^{l,m}$.

3.3. ORTHONORMAL CLUSTERING READOUT (OCREAD)

The readout function is an essential component to learn the graph-level representations for brain network analysis (*e.g.*, classification), which maps a set of learned node-level embeddings to a graph-level embedding. $\mathrm{Mean}(\cdot), \mathrm{Sum}(\cdot)$ and $\mathrm{Max}(\cdot)$ are the most commonly used readout functions for GNNs. Xu et al. (Xu et al., 2019) show that GNNs equipped with $\mathrm{Sum}(\cdot)$ readout have the same discriminative power as the Weisfeiler-Lehman Test. Zhang et al. (Zhang et al., 2018) propose a sort pooling to generate the graph-level representation by sorting the final node representations. Ju et al. (Ju et al., 2022) present a layer-wise readout by extending the node information aggregated from the last layer of GNNs to all layers. However, none of the existing readout functions leverages the properties of brain networks that nodes in the same functional modules tend to have similar behaviors and clustered representations, as shown in Figure 1(a). To address this deficiency, we design a novel readout function to take advantage of the modular-level similarities between ROIs in brain networks, where nodes are assigned softly to well-chosen clusters with an unsupervised process.

Formally, given K cluster centers, each center has V dimensions, $\boldsymbol{E} \in \mathbb{R}^{K \times V}$, a Softmax projection operator is used as the function to calculate the probability $\boldsymbol{P}_{ik}$ of assigning node i to cluster k,

$$\boldsymbol{P}_{ik} = \frac{e^{\langle \boldsymbol{Z}^L_{i\cdot}, \boldsymbol{E}_{k\cdot} \rangle}}{\sum_{k'}^{K} e^{\langle \boldsymbol{Z}^L_{i\cdot}, \boldsymbol{E}_{k'\cdot} \rangle}}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\boldsymbol{Z}^L$ is the learned set of node embeddings from the last layer of MHSA module. With this computed soft assignment $\boldsymbol{P} \in \mathbb{R}^{V \times K}$, the original learned node representation $\boldsymbol{Z}^L$ can be aggregated under the guidance of the soft cluster information, where the graph-level embedding $\boldsymbol{Z}_G$ is obtained by $\boldsymbol{Z}_G = \boldsymbol{P}^\top \boldsymbol{Z}^L$.

However, jointly learning node embeddings and clusters without ground-truth cluster labels is difficult. To obtain representative soft assignment $\boldsymbol{P}$, the initialization of K cluster centers $\boldsymbol{E}$ is critical and should be designed delicately. To this end, we leverage the observation illustrated in Figure 1(b), where orthonormal embeddings can improve the clustering of nodes in brain networks *w.r.t.* the functional modules underlying brain regions.

Orthonormal Initialization. To initialize a group of orthonormal bases as cluster centers, we first adopt the Xavier uniform (Glorot & Bengio, 2010) to initialize K random centers and each center contains V dimensions $\boldsymbol{C} \in \mathbb{R}^{K \times V}$. Then, we apply the Gram-Schmidt process to obtain the orthonormal bases $\boldsymbol{E}$, where

$$\boldsymbol{u}_k = \boldsymbol{C}_{k\cdot} - \sum_{j=1}^{k-1} \frac{\langle \boldsymbol{u}_j, \boldsymbol{C}_{k\cdot} \rangle}{\langle \boldsymbol{u}_j, \boldsymbol{u}_j \rangle} \boldsymbol{u}_j, \quad \boldsymbol{E}_{k\cdot} = \frac{\boldsymbol{u}_k}{\|\boldsymbol{u}_k\|}. \tag{3}$$

3.4. Generalizing OCREAD to Other Graph Tasks and Domains

In this work, we tested the proposed OCREAD on functional connectivity (FC) based brain networks. Other popular modalities of brain networks include structural connectivities (SC), which describe the anatomical organization of the brain by measuring the fiber tracts between brain regions (Babaeeghazvini et al., 2021). In SC-based brain networks, ROIs that are positionally close to each other on the structural connectivity networks tend to share similar connection profiles. This means the idea of OCREAD is also naturally applicable to SC networks, where the orthonormal clustering is based on the physical distances instead of the functional modules on FC.

4. Experiments

This section evaluates the effectiveness of our proposed BRAIN NETWORK TRANSFORMER (BRAINNETTF) with extensive experiments. We aim to address the following research questions:

RQ1. How does BRAINNETTF perform compared with state-of-the-art models of various types?

RQ2. How does our proposed OCREAD module perform with different model choices?

RQ3. Does the learned model of BRAINNETTF exhibit consistency with existing neuroscience knowledge and suggest reasonable interpretability?

4.1. Experimental Settings

Datasets. We conduct experiments on two real-world fMRI datasets. (a) *Autism Brain Imaging Data Exchange (ABIDE)*: This dataset collects resting-state functional magnetic resonance imaging (Rs-fMRI) data from 17 international sites, and all data are anonymous (Cameron et al., 2013). The used dataset contains brain networks from 1009 subjects, with 516 (51.14%) being Autism spectrum disorder (ASD) patients (positives). The region definition is based on Craddock 200 atlas (Craddock et al., 2012). As the most convenient open-source large-scale dataset, it provides generated brain networks and can be downloaded directly without permission request. Despite the ease of acquisition, the heterogeneity of the data collection process hinders its use. Since multi-site data are collected from different scanners with different acquisition parameters, non-neural inter-site variability may mask inter-group differences. In practice, we find the training unstable, and there is a significant gap between validation and testing performances. However, we discover that most models can achieve a stable performance if we follow an appropriate stratified sampling strategy by considering collection sites during the training-validation-testing splitting process for ABIDE. Training curves in Figure 3 also show how different models achieve a stabler performance on our designed new splitting settings than the random splitting. Therefore, we use ABIDE as one of the benchmark datasets in this work, and we share our re-standardized data splitting to provide a fair evaluation pipeline for various future methods. (b) *Adolescent Brain Cognitive Development Study (ABCD)*: This is one of the largest publicly available fMRI datasets with restricted access (a strict data requesting process needs to be followed to obtain the data) (Casey et al., 2018). The data we use in the experiments are fully anonymized brain networks with only gender labels. After the quality control process, 7901 subjects are included in the analysis, with 3961 (50.1%) among them being female. The region definition is based on the HCP 360 ROI atlas (Glasser et al., 2013).

Metrics. The diagnosis of ASD is the prediction target on ABIDE, while gender prediction is used as the evaluation task for ABCD. Both prediction tasks are binary classification problems, and both datasets are balanced between classes. Hence, AUROC is the most proper performance metric adopted for a fair comparison, and accuracy is ap-

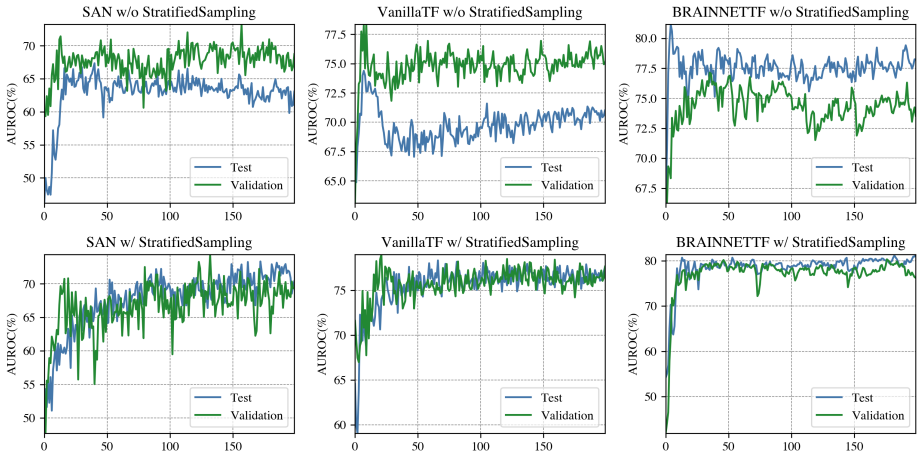


Figure 3: Training Curves of Different Models with or without StratifiedSampling. It is shown that (a) with stratified sampling, the performance gap between validation and test on ABIDE is much smaller than the one without stratified sampling; (b) stratified sampling can stabilize the training process on ABIDE, especially for VanillaTF and BRAINNETTF.

plied to reflect the prediction performance. All the reported performances are the average of 5 random runs on the test set with the standard deviation.

Implementation details. For experiments, we use a two-layer Multi-Head Self-Attention Module and set the number of heads M to 4 for each layer. We randomly split 70% of the datasets for training, 10% for validation, and the remaining are utilized as the test set. In the training process of BRAINNETTF, we use an Adam optimizer with an initial learning rate of 10^{-4} and a weight decay of 10^{-4}. The batch size is set as 64. All models are trained for 200 epochs, and the epoch with the highest AUROC performance on the validation set is used for performance comparison on the test set. The model is trained on an NVIDIA Quadro RTX 8000. Please refer to the supplementary material for the full implementation of BRAINNETTF.

Computation complexity. In BRAINNETTF, the computation complexity of Multi-Head Self-Attention Module and OCREAD are $\mathcal{O}(LMV^2)$ and $\mathcal{O}(KV)$ respectively, where L is the layer number of Multi-Head Self-Attention Module, V is the number of nodes, M is the number of heads, and K is the number of clusters in OCREAD. The overall computation complexity of BRAINNETTF is thus $\mathcal{O}(V^2)$, which is on the same scale as common GNNs on brain networks such as BrainGNN (Li et al., 2021) and BrainGB (Cui et al., 2022).

4.2. Performance Analysis (RQ1)

(a) BRAINNETTF vs. other graph transformers. We compare BRAINNETTF with two popular graph Transformers, SAN (Kreuzer et al., 2021) and Graphormer (Ying et al., 2021). In addition, we also include a basic version of BRAINNETTF without OCREAD, composed of a Transformer with a 2-layer Multi-Head Self-Attention and a CONCAT-based readout named VanillaTF. Our BRAINNETTF outperforms SAN and Graphormer by significant margins, with up to 6% absolute improvements on both datasets. VanillaTF also performs better than SAN and Graphormer. We believe this downgraded performance of existing graph transformers results from their design flaws facing the natures of brain networks. Specifically, both the preprocessing and the training stages of the Graphormer model accepts only discrete, categorical data. A bin operator has to be applied on the adjacency matrix, coarsening the node feature from connection profiles and dramatically hurting the performance. Furthermore, since brain networks are complete graphs, key designs like centrality encoding and spatial encoding of Graphormer cannot be appropriately applied. Similarly, for SAN, experiments in Appendix 1 show that adding eigen node features to connection profiles cannot improve the model's performance. Besides, the benchmark paper (Cui et al., 2022) reveals that injecting edge weights into the attention mechanism can significantly reduce the prediction power. **(b) BRAINNETTF vs. neural network models on fixed brain networks.** We further introduce another two neural network baselines on fixed brain networks. BrainGB (Cui et al., 2022) is a systematic study of how to design effective GNNs for brain network analysis. We adopt their best design as the BrainGB baseline. BrainnetCNN (Kawahara et al., 2017) represents state-of-the-art of specialized GNNs for brain network analysis, which models the adjacency matrix of a brain network similarly as a 2D image. As is shown in Table 2, BRAINNETTF consistently outperforms both BrainGB and BrainnetCNN. **(c) BRAINNETTF vs. neural network models on learnable brain networks.** Unlike classical GNNs, FBNETGEN (Kan et al., 2022) is

Table 2: Performance comparison with different baselines. The performance gains of BRAINNETTF over the baselines have passed the t-test with p-value$<$0.03.

Type	Method	Dataset: ABIDE		Dataset: ABCD	
		AUROC (%)	Accuracy (%)	AUROC (%)	Accuracy (%)
Graph Transformer	SAN	71.3±2.1	65.3±2.9	90.1±1.2	81.0±1.3
	Graphormer	63.5±3.7	60.8±2.7	89.0±1.4	80.2±1.3
	VanillaTF	76.4±1.2	65.2±1.2	94.3±0.7	85.9±1.4
Fixed Brain Network	BrainGB	69.7±3.3	63.6±1.9	91.9±0.3	83.1±0.5
	BrainnetCNN	77.4±2.4	70.4±2.7	93.5±0.3	85.7±0.8
Learnable Brain Network	FBNETGNN	77.6±1.2	70.0±1.4	94.5±0.7	87.2±1.2
Ours	BRAINNETTF	**80.2±1.0**	**71.0±1.2**	**96.2±0.3**	**88.4±0.4**

Table 3: Performance comparison AUROC (%) with different readout functions.

Readout	Dataset: ABIDE			Dataset: ABCD		
	SAN	Graphormer	VanillaTF	SAN	Graphormer	VanillaTF
MEAN	63.7±2.4	50.1±1.1	73.4±1.4	88.5±0.9	87.6±1.3	91.3±0.7
MAX	61.9±2.5	54.5±3.6	75.6±1.4	87.4±1.1	81.6±0.8	94.4±0.6
SUM	62.0±2.3	54.1±1.3	70.3±1.6	84.2±0.8	71.5±0.9	91.6±0.6
SortPooling	68.7±2.3	51.3±2.2	72.4±1.3	84.6±1.1	86.7±1.0	89.9±0.6
CONCAT	**71.3±2.1**	63.5±3.7	76.4±1.2	90.1±1.2	89.0±1.4	94.3±0.7
OCREAD	70.6±2.4	**64.9±2.7**	**80.2±1.0**	**91.2±0.7**	**90.2±0.7**	**96.2±0.4**

a GNN that aggerates node features on graphs with learnable weights based on learnable projections of the original BOLD signals, which achieves SOTA performance on the ABCD dataset for gender prediction. The learnable graphs can be seen as a type of attention score. Experiment results show that our proposed BRAINNETTF beats FBNETGEN on both datasets.

4.3. Ablation Studies on the OCREAD Module (RQ2)

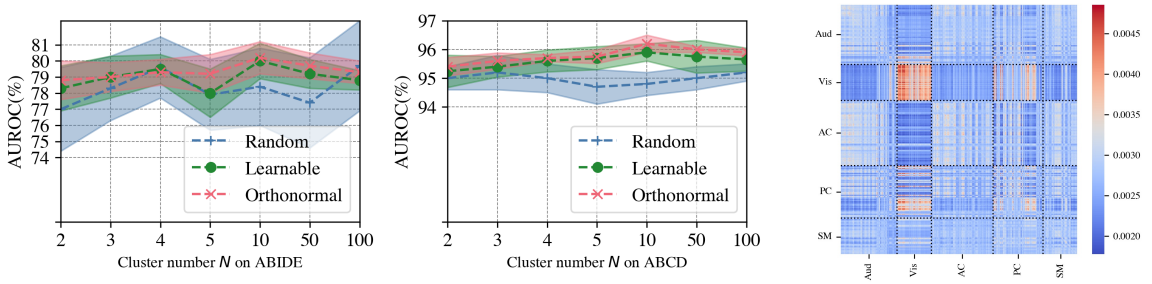
4.3.1. OCREAD WITH VARYING READOUT FUNCTIONS

We vary the readout function for various Transformer architectures, including SAN, Graphormer and VanillaTF, to observe the performance of each ablated model variant. The results shown in Table 3 demonstrate that our OCREAD is the most effective readout function for brain networks and improves the prediction power across various Transformer architectures.

4.3.2. OCREAD WITH VARYING CLUSTER INITIALIZATIONS

To further demonstrate how the design of OCREAD influences the performance of BRAINNETTF, we investigate two key model selections, the initialization method for cluster centers and the cluster number K. For the initialization, three different kinds of initialization procedures are compared, namely (a) **Random**: the Xavier uniform (Glorot & Bengio, 2010) is leveraged to randomly generate a group of centers, which are then normalized into unit vectors; (b) **Learnable**: the same initial process as Random, but the generated centers are further updated with gradient descent; (c) **Orthonormal**: our proposed process as described in Eq. (3).

Specifically, we test each initialization method with the cluster number K equals to 2, 3, 4, 5, 10, 50, 100. The results of adjusting these two hyper-parameters on ABIDE and ABCD datasets are shown in Figure 4(a). We observe that: (1) When cluster centers are orthonormal, the model's performance increases with the number of clusters ranging from 2 to 10, and then drops with the cluster number rising from 10 to 100, suggesting the optimal cluster number to be relatively small, which leads to less computation and is consistent with the fact that the typical number of functional modules are smaller than 25; (2) With a sufficiently large cluster number, all three initialization methods, Random, Learnable and Orthonormal, tend to reach similar performance, but orthonormal performs stably better when the number of clusters is smaller; (3) It is also notable that our OCREAD consistently achieves the best performance over other initialization methods regarding smaller standard deviations.



(a) Influence of two key hyper-parameters for model performance. (b) Attention heatmap on ABCD.

Figure 4: The hyper-parameter influence and the heatmap from self-attention.

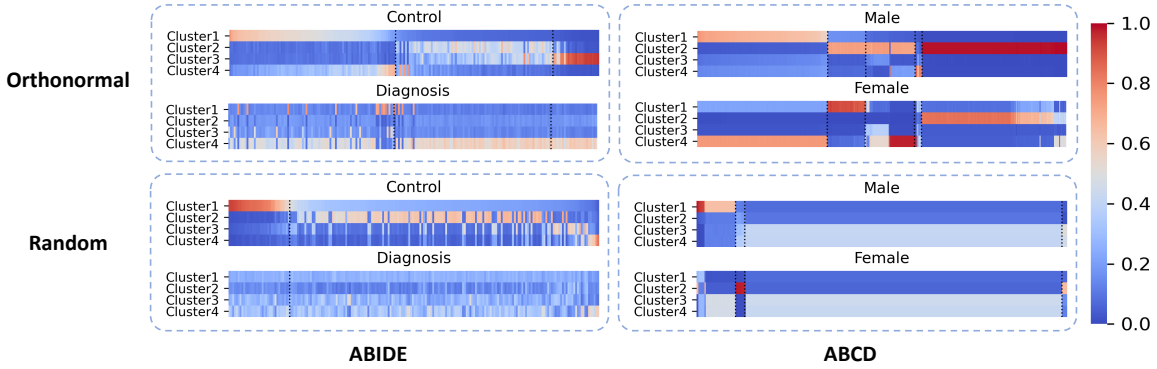


Figure 5: Visualization of cluster (module-level) embeddings learned with Orthonormal vs. Random cluster center initializations on two datasets. Each group in the dotted box contains two heatmaps (one for each prediction class) with the same node ordering on the x-axis.

4.4. In-depth Analysis of Attention Scores and Cluster Assignments (RQ4)

Figure 4(b) displays the self-attention score from the first layer of Multi-Head Self-Attention. The attention scores are the average across all subjects in the ABCD test set. This figure shows that the learned attention scores well match the divisions of functional modules based on available labels, demonstrating the effectiveness and interpretability of our Transformer model. Note that since there exists no available functional module labels for the atlas of the ABIDE dataset, we cannot visualize the correlations between attention scores and functional modules.

Figure 5 shows the cluster soft assignment results $\boldsymbol{P}$ on nodes in OCREAD with two initialization methods. The cluster number K is set to 4. The visualized numerical values are the average $\boldsymbol{P}$ of all subjects in each dataset's test set. From the visualization, we observe that (a) Orthonormal initialization produces more discriminative $\boldsymbol{P}$ between classes than random initialization; (b) Within each class, orthonormal initialization encourages the nodes to form groups. These observations demonstrate that our OCREAD with orthonormal initialization can leverage potential clusters underlying node embeddings, thus automatically grouping brain regions into potential functional modules.

5. Conclusion

This paper presents BRAIN NETWORK TRANSFORMER, a specialized graph Transformer model with ORTHONORMAL CLUSTERING READOUT for brain network analysis. Extensive experiments on two large-scale brain network datasets demonstrate that our BRAINNETTF achieves superior performance over SOTA baselines of various types. Specifically, to model the potential node feature similarity in brain networks, we design OCREAD and demonstrate its effectiveness empirically. Lastly, the re-standardized dataset split for ABIDE can provide a fair evaluation for new methods in the community. For future work, BRAINNETTF can be improved with explicit explanation modules and used as the backbone for further brain network analysis, such as digging essential neural circuits for mental disorders and understanding cognitive development in adolescents.

References

Akiki, T. J. and Abdallah, C. G. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific Reports*, 9: 1–15, 2019.

Babaeeghazvini, P., Rueda-Delgado, L. M., Gooijers, J., Swinnen, S. P., and Daffertshofer, A. Brain structural and functional connectivity: A review of combined works of diffusion magnetic resonance imaging and electroencephalography. *Frontiers in Human Neuroscience*, 15, 2021.

Bullmore, E. and Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10:186–198, 2009.

Cameron, C., Yassine, B., Carlton, C., Francois, C., Alan, E. A., András, J., Budhachandra, K., John, L., Qingyang, L., Michael, M., Chaogan, Y., and Pierre, B. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Frontiers in Neuroinformatics*, 2013.

Caramazza, A. and Coltheart, M. Cognitive neuropsychology twenty years on. *Cognitive Neuropsychology*, 2006.

Casey, B., Cannonier, T., and et al., M. I. C. The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018.

Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H., and Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, 2021.

Craddock, R. C., James, G. A., Holtzheimer III, P. E., Hu, X. P., and Mayberg, H. S. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Human Brain Mapping*, 33:1914–1928, 2012.

Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A. A. C., Lukemire, J., Zhan, L., He, L., Guo, Y., and Yang, C. Braingb: A benchmark for brain network analysis with graph neural networks, 2022.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

Deco, G., Jirsa, V. K., and McIntosh, A. R. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nature Reviews Neuroscience*, 12: 43–56, 2011.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Dwivedi, V. P. and Bresson, X. A generalization of transformer networks to graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021.

Dwivedi, V. P., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. Graph neural networks with learnable structural and positional representations. In *ICLR*, 2022.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80:105–124, 2013.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, 2020a.

Hu, Z., Dong, Y., Wang, K., and Sun, Y. Heterogeneous graph transformer. In *WWW*, pp. 2704–2710, 2020b.

Hussain, M. S., Zaki, M. J., and Subramanian, D. Edge-augmented graph transformers: Global self-attention is enough for graphs. *arXiv*, 2021.

Ju, M., Hou, S., Fan, Y., Zhao, J., Zhao, L., and Ye, Y. Adaptive kernel graph neural network. *AAAI*, 2022.

Kan, X., Cui, H., Lukemire, J., Guo, Y., and Yang, C. FBNETGEN: Task-aware GNN-based fMRI analysis via functional brain network generation. In *Medical Imaging with Deep Learning*, 2022.

Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., Zwicker, J. G., and Hamarneh, G. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

Kim, B.-H., Ye, J. C., and Kim, J.-J. Learning dynamic graph representation of brain connectome with spatio-temporal attention. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *NeurIPS*, 2021.

Kreuzer, D., Beaini, D., Hamilton, W. L., Létourneau, V., and Tossou, P. Rethinking graph transformers with spectral attention. In *NeurIPS*, 2021.

Li, X., Dvornek, N. C., Zhou, Y., Zhuang, J., Ventola, P., and Duncan, J. S. Graph neural network for interpreting task-fmri biomarkers. In *MICCAI*, 2019.

Li, X., Zhou, Y., Gao, S., Dvornek, N., Zhang, M., Zhuang, J., Gu, S., Scheinost, D., Staib, L., Ventola, P., et al. Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis*, 2021.

Park, W., Chang, W., Lee, D., Kim, J., and Hwang, S.-w. Grpe: Relative positional encoding for graph transformer, 2022.

Satterthwaite, T. D., Wolf, D. H., Roalf, D. R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E. D., Elliott, M. A., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R. E., and Gur, R. C. Linked Sex Differences in Cognition and Functional Connectivity in Youth. *Cerebral Cortex*, 25:2383–2394, 2015.

Simpson, S. L., Bowman, F. D., and Laurienti, P. J. Analyzing complex functional brain networks: fusing statistics and network science to understand the brain. *Statistics Surveys*, 7:1, 2013.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., and Woolrich, M. W. Network modelling methods for FMRI. *NeuroImage*, 54, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *NeurIPS*, 2017.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *ICLR*, 2018.

Wang, Y. and Guo, Y. A hierarchical independent component analysis model for longitudinal neuroimaging studies. *NeuroImage*, 189:380–400, 2019.

Wang, Y., Kang, J., Kemmer, P. B., and Guo, Y. An Efficient and Reliable Statistical Method for Estimating Functional Connectivity in Large Scale Brain Networks Using Partial Correlation. *Frontiers in Neuroscience*, 10:123, 2016.

Xie, J., Girshick, R., and Farhadi, A. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *ICLR*, 2019.

Yan, Y., Zhu, J., Duda, M., Solarz, E., Sripada, C., and Koutra, D. Groupinn: Grouping-based interpretable neural network-based classification of limited, noisy brain data. In *KDD*, 2019.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? In *NeurIPS*, 2021.

Zhang, M., Cui, Z., Neumann, M., and Chen, Y. An end-to-end deep learning architecture for graph classification. In *AAAI*, 2018.

Zhao, J., Li, C., Wen, Q., Wang, Y., Liu, Y., Sun, H., Xie, X., and Ye, Y. Gophormer: Ego-graph transformer for node classification, 2021.