BrainNNExplainer: An Interpretable Graph Neural Network Framework for Brain Network based Disease Analysis

Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, Carl Yang (corresponding: j.carlyang@emory.edu)

Emory University, Atlanta, GA 30322, USA

Problem and preliminaries

Brain network. Complex graphs with anatomic regions as nodes and connectivities between the regions as links.
GNN explanation. Current GNN explanation models usually produce a unique explanation subgraph for each graph sample *(e.g. GNNExplainer)*, or through the model-level explanation *(e.g. GAT)*, without considering the unique properties of brain networks *(i.e. fixed number and order of nodes under a given atlas)* and the characteristics of disease analysis *(i.e. subjects with the same disease may share similar connection patterns)*.

Motivation.
1) Unleash the prediction power of GNNs in brain analysis
2) Provide disease-specific explanation by a shared mask

Problem definition. Given a weighted brain network $G = (\mathcal{V}, \mathcal{E}, \boldsymbol{W})$, where $\mathcal{V} = \{v_i\}_{i=1}^n$ is the regions of interest node set (ROIs), $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ is the edge set, and $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ is the weighted adjacency matrix describing connection strengths, the model outputs a disease prediction y.
The *interpretability* is provided by learning a shared edge mask $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ to highlight the disease-specific prominent ROI connections.
Neural system mapping. Partition the ROIs of brain networks into eight neural systems based on structural and functional roles under a specific atlas (e.g. AAL90 and Brodmann82).

The backbone BrainNN

Node features construction. Common node features such as degree, binning degree, node2vec and degree profiles (LDP)

$$\boldsymbol{x}_i = [\deg(v_i); \min(\mathcal{D}_i); \max(\mathcal{D}_i); \text{mean}(\mathcal{D}_i); \text{std}(\mathcal{D}_i)] \tag{1}$$

Edge-weight-aware message passing. To adopt valuable edge weights, we construct the message vector $\boldsymbol{m}_{ij} \in \mathbb{R}^D$ by concatenating node embeddings of i, j, and edge weight w_{ij}

$$\boldsymbol{m}_{ij}^{(l)} = MLP_{\boldsymbol{\Theta}}\left(\left[\boldsymbol{h}_i^{(l)}; \boldsymbol{h}_j^{(l)}; w_{ij}\right]\right) \tag{2}$$

Then aggregate messages from all neighbors followed by a non-linear transformation

$$\boldsymbol{h}_i^{(l)} = \sigma\left(\sum_{j \in \mathcal{N}_i \cup \{i\}} \boldsymbol{m}_{ij}^{(l-1)}\right) \tag{3}$$

The graph-level embeddings can be obtained by summarizing all node embeddings with residual connections. The training objective for BrainNN is a supervised cross-entropy loss (denoted as $\mathcal{L}_p$) towards disease predictions.

The explanation generator

Shared edge mask as the explanation. Train the shared mask $\boldsymbol{M}$ by maximizing the mutual information between the BrainNN predictions $\hat{y}$ on the original graph G and $\hat{y}'$ on the masked graph G', where $\boldsymbol{W}' = \boldsymbol{W} \odot \sigma(\boldsymbol{M})$.

$$\mathcal{L}_m = -\sum_{c=1}^{C} \mathbb{1}[y = c] \log P_\Phi\left(y' = y \mid G = \boldsymbol{W}'\right) \tag{4}$$

Further apply a sparsity loss $\mathcal{L}_s$ to improve compactness and an element wise entropy loss $\mathcal{L}_e$ to encourage discreteness in mask weight values. Final training objective:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_e \tag{5}$$

Three-step training strategy. BrainNN-Explanation Generator-BrainNN

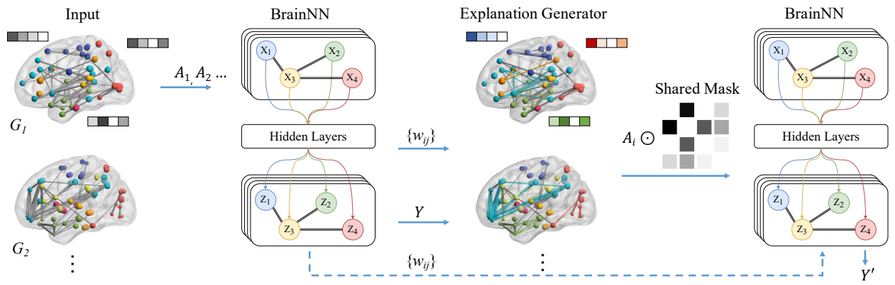

Figure: The proposed BrainNNExplainer trained in three-steps: the initial training of BrainNN on the original data, the explanation generation based on trained BrainNN, and the further adjustment of BrainNN based on the explanation masked graph.

Performance Comparison

Datasets. We use two real-world datasets, Human Immunodeficiency Virus Infection (HIV) and Bipolar Disorder (BP).

Method	HIV		BP	
	Accuracy	AUC	Accuracy	AUC
M2E	50.61	51.53	57.78	53.63
MIC	55.63	56.61	51.21	50.12
MPCA	67.24	66.92	56.92	56.86
MK-SVM	65.71	68.89	60.12	56.78
GAT	68.58	67.31	61.31	59.93
GCN	70.16	69.94	64.44	64.24
DiffPool	71.42	71.08	62.22	62.54
BrainNN	74.29	71.67	71.11	64.71
BrainNNExplainer	**77.14**	**75.00**	**75.56**	**69.88**

Table: Performance of different models on HIV and BP datasets. Our methods are colored in gray background and the highest performance is highlighted in boldface.

Baselines. The compared baselines include both shallow (i.e. M2E, MIC, MPCA, MK-SVM) and deep models (GAT, GCN, DiffPool).
Results. Our backbone BrainNN outperforms all SOTA baselines by up to 11%. The three-step training with globally shared mask achieves a further performance improvements of 5%.

Interpretability Analysis

Visualization. Use a threshold to obtain a explanation subgraph G_s' by removing low-weight edges from G'

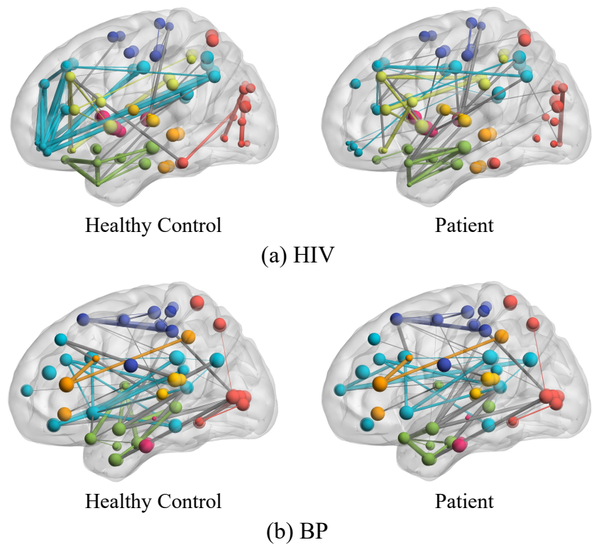

Figure: Comparison of explanation graph connectomes in brain networks of a healthy control and a patient on HIV and BP datasets. The colors of 8 neural systems are described as: VN, AN, BLN, DMN, SMN, SN, MN, CCN.

In the *HIV* dataset, the explanation subgraph of patients excludes many interactions within the Default Mode Network (DMN). For the *BP* patients, the connections within Bilateral Limbic Network (BLN) are much more sparse.
Interpretation of important brain systems. Observing the most manifest nodes with different comparative measures

Dataset	Type	Comparative Measures		
		Degree	Strength	Cluster Coefficient
HIV	Normal	DMN BLN CCN	DMN BLN CCN	DMN CCN BLN
	Patient	BLN CCN AN	BLN CCN AN	BLN
BP	Normal	BLN SMN DMN	BLN DMN SMN	SMN VN DMN
	Patient	BLN DMN SMN	BLN DMN SMN	SMN VN

Table: Top ranked neural systems of the explanation subgraph on HIV and BP for both Healthy Control (Normal) and Patient.

For *HIV* dataset, both healthy control and patients' explanation subgraphs reveal the importance of BLN, while DMN is missing from all three metrics in the patient group. Regarding *BP* dataset, BLN, SMN (Somato-Motor Network), and DMN are prominent in both patient and healthy controls.
Community structure and modularity. Compare the modularity of our explanation graph G' against the original graph G. The explained graph achieves about 5.10%-7.21% improvement over the original graph based on multiple metrics.