

Pooling Regularized Graph Neural Network for fMRI Biomarker Analysis

Xiaoxiao Li^{*}, Yuan Zhou[†], Nicha C. Dvornek^{†*}, Muhan Zhang[◇], Juntang Zhuang^{*}, Pamela Ventola[‡] and James S. Duncan^{**†}

^{*} Biomedical Engineering, Yale University, New Haven, CT, USA

^{*} Electrical Engineering, Yale University, New Haven, CT, USA

[†] Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA

[‡] Child Study Center, Yale School of Medicine, New Haven, CT, USA

[◇] Facebook AI Research

Abstract. Understanding how certain brain regions relate to a specific neurological disorder has been an important area of neuroimaging research. A promising approach to identify the salient regions is using Graph Neural Networks (GNNs), which can be used to analyze graph structured data, e.g. brain networks constructed by functional magnetic resonance imaging (fMRI). We propose an interpretable GNN framework with a novel salient region selection mechanism to determine neurological brain biomarkers associated with disorders. Specifically, we design novel regularized pooling layers that highlight salient regions of interests (ROIs) so that we can infer which ROIs are important to identify a certain disease based on the node pooling scores calculated by the pooling layers. Our proposed framework, Pooling Regularized-GNN (PR-GNN), encourages reasonable ROI-selection and provides flexibility to preserve either individual- or group-level patterns. We apply the PR-GNN framework on a Biopoint Autism Spectral Disorder (ASD) fMRI dataset. We investigate different choices of the hyperparameters and show that PR-GNN outperforms baseline methods in terms of classification accuracy. The salient ROI detection results show high correspondence with the previous neuroimaging-derived biomarkers for ASD.

Keywords: fMRI Biomarker · Graph Neural Network · Autism.

1 Introduction

Explaining the underlying roots of neurological disorders (i.e., what brain regions are associated with the disorder) has been a main goal in the field of neuroscience and medicine [1,2,3,4]. Functional magnetic resonance imaging (fMRI), a non-invasive neuroimaging technique that measures neural activation, has been paramount in advancing our understanding of the functional organization of the brain [5,6,7]. The functional network of the brain can be modeled as a graph in which each node is a brain region and the edges represent the strength of the connection between those regions.

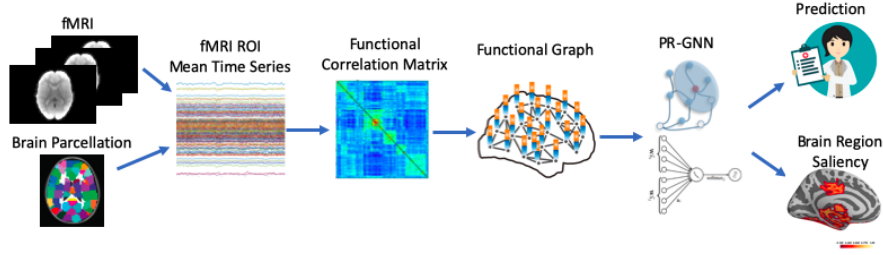


Fig. 1: The overview of the pipeline. fMRI images are parcellated by atlas and transferred to graphs. Then, the graphs are sent to our proposed PR-GNN, which gives the prediction of specific tasks and jointly selects salient brain regions that are informative to the prediction task.

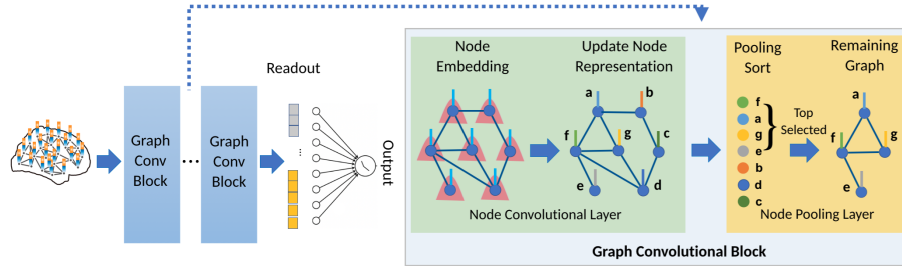


Fig. 2: PR-GNN for brain graph classification and the details of its key component - Graph Convolutional Block. Each Graph Convolutional Block contains a node convolutional layer followed by a node pooling layer.

The past few years have seen the growing prevalence of using graph neural networks (GNN) for graph classification [8]. Like pooling layers in convolutional neural networks (CNNs) [9,10], the pooling layer in GNNs is an important design to compress a large graph to a smaller one for lower dimensional feature extraction. Many node pooling strategies have been studied and can be divided into the following categories: 1) clustering-based pooling, which clusters nodes to a super node based on graph topology [11,12,13] and 2) ranking-based pooling, which assigns each node a score and keeps the top ranked nodes [14,15]. Clustering-based pooling methods do not preserve node assignment mapping in the input graph domain, hence they are not inherently interpretable at the node level. For our purpose of interpreting node importance, we focus on ranking-based pooling methods. Currently, existing methods of this type [14,15] have the following key limitations when applying them to salient brain ROI analysis: 1) ranking scores for the discarded nodes and the remaining nodes may not be significantly distinguishable, which is not suitable for identifying salient and representative regional biomarkers, and 2) the nodes in different graphs in the same group may be ranked totally differently (usually caused by overfitting),

which is problematic when our objective is to find group-level biomarkers. To reach group-level analysis, such approaches typically require additional steps to summarize statistics (such as averaging). For these two-stage methods, if the results from the first stage are not reliable, significant errors can be induced in the second stage.

To utilize GNN for fMRI learning and meet the need of group-level biomarker finding, we propose a pooling regularized GNN framework (PR-GNN) for classifying neurodisorder patients vs healthy control subjects and discovering disorder related biomarkers jointly. The overview of our methods is depicted in Fig. 1. Our key contributions are:

- We formulate an end-to-end framework for fMRI prediction and biomarker (salient brain ROIs) interpretation.
- We propose novel regularization terms for ranking-based pooling methods to encourage more reasonable node selection and provide flexibility between individual-level and group-level interpretation in GNN.

2 Graph Neural Network for Brain Network Analysis

The architecture of our PR-GNN is shown in Fig. 2. Below, we introduce the notation and the layers in PR-GNN. For simplicity, we focus on Graph Attention Convolution (GATConv) [16,17] as the node convolutional layer. For node pooling layers, we test two existing ranking based pooling methods: TopK pooling [14] and SAGE pooling [15].

2.1 Notation and Problem Definition

We first parcellate the brain into N ROIs based on its T1 structural MRI. We define ROIs as graph nodes $\mathcal{V} = \{v_1, \dots, v_N\}$. We define an undirected weighted graph as $G = (\mathcal{V}, \mathcal{E})$, where \mathcal{E} is the edge set, i.e., a collection of (v_i, v_j) linking vertices v_i and v_j . G has an associated node feature matrix $H = [\mathbf{h}_1, \dots, \mathbf{h}_N]^\top$, where \mathbf{h}_i is the feature vector associated with node v_i . For every edge connecting two nodes, $(v_i, v_j) \in \mathcal{E}$, we have its strength $e_{ij} \in \mathbb{R}$. We also define $e_{ij} = 0$ for $(v_i, v_j) \notin \mathcal{E}$ and therefore the adjacency matrix $E = [e_{ij}] \in \mathbb{R}^{N \times N}$ is well defined.

2.2 Graph Convolutional Block

Node Convolutional Layer To improve GATConv [8], we incorporate edge features in the brain graph as suggested by Gong & Cheng [18] and Yang et. al [17]. We define $\mathbf{h}_i^{(l)} \in \mathbb{R}^{d^{(l)}}$ as the feature for the i^{th} node in the l^{th} layer and $H^{(l)} = [\mathbf{h}_1^{(l)}, \dots, \mathbf{h}_{N^{(l)}}^{(l)}]^\top$, where $N^{(l)}$ is the number of nodes at the l^{th} layer (the same for $E^{(l)}$). The propagation model for the forward-pass update of node representation is calculated as:

$$\mathbf{h}_i^{(l+1)} = \phi_i^{\Theta^{(l)}}(H^{(l)}, E^{(l)}) = \alpha_{i,i} \Theta^{(l)} \mathbf{h}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j} \Theta^{(l)} \mathbf{h}_j^{(l)}, \quad (1)$$

where the attention coefficients α_{ij} are computed as

$$\alpha_{i,j} = \frac{\exp(\hat{\alpha}_{i,j})}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp(\hat{\alpha}_{i,k})}, \quad \hat{\alpha}_{i,j} = e_{i,j}^{(l)} \text{ReLU}((\mathbf{a}^{(l)})^\top [\Theta^{(l)} \mathbf{h}_i^{(l)} \parallel \Theta^{(l)} \mathbf{h}_j^{(l)}]), \quad (2)$$

where $\mathcal{N}(i)$ denotes the set of indices of neighboring nodes of v_i , \parallel denotes concatenation, $\Theta^{(l)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ and $\mathbf{a}^{(l)} \in \mathbb{R}^{2d^{(l+1)}}$ are model parameters.

Node Pooling Layer The choices of keeping which nodes in TopK pooling and SAGE pooling are determined based on the node importance score $\mathbf{s}^{(l)} = [s_1^{(l)}, \dots, s_{N^{(l)}}^{(l)}]^\top$, which is calculated in two ways as follows:

$$s_i^{(l)} = \begin{cases} \text{sigmoid}((\mathbf{h}_i^{(l)})^\top \mathbf{w}^{(l)} / \|\mathbf{w}^{(l)}\|), & \text{TopK pooling} \\ \text{sigmoid}(\phi_i^{\theta^{(l)}}(H^{(l)}, E^{(l)})), & \text{SAGE pooling} \end{cases} \quad (3)$$

where ϕ_i^{θ} is calculated in Eq. (1) and $\mathbf{w}^{(l)} \in \mathbb{R}^{d^{(l)}}$ and $\theta^{(l)} \in \mathbb{R}^{1 \times d^{(l)}}$ are model parameters. Note that $\theta^{(l)}$ is different from $\Theta^{(l)}$ in Eq. (1) such that the output of ϕ_i^{θ} is a scalar.

Given $\mathbf{s}^{(l)}$ the following equation roughly describes the pooling procedure:

$$\text{idx} = \text{top}(\mathbf{s}^{(l)}, k^{(l)}), \quad E^{(l+1)} = E_{\text{idx}, \text{idx}}^{(l)}. \quad (4)$$

The notation above finds the indices corresponding to the largest $k^{(l)}$ elements in score vector $\mathbf{s}^{(l)}$, and $(\cdot)_{\mathbf{i}, \mathbf{j}}$ is an indexing operation which takes elements at row indices specified by \mathbf{i} and column indices specified by \mathbf{j} . The nodes receiving lower scores will experience less feature retention.

Lastly, we seek a flattening operation to translate graph information to a vector. Suppose the last layer is L , we use $\mathbf{z} = \text{mean}\{\mathbf{h}_i^{(L)} : i = 1, \dots, N^{(L)}\}$, where mean operates elementwisely. Then \mathbf{z} is sent to a multilayer perceptron (MLP) to give the final prediction.

3 Proposed Regularizations

3.1 Distance Loss

To overcome the limitation of existing methods that ranking scores for the discarded nodes and the remaining nodes may not be distinguishable, we propose two distance losses to encourage the difference. Before introducing them, we first rank the elements of the m^{th} instance scores, $\mathbf{s}_m^{(l)}$, in a descending order, denote it as $\hat{\mathbf{s}}_m^{(l)} = [\hat{s}_{m,1}^{(l)}, \dots, \hat{s}_{m,N^{(l)}}^{(l)}]^\top$, and denote its top $k^{(l)}$ elements as $a_{m,i}^{(l)} = \hat{s}_{m,i}^{(l)}, i = 1, \dots, k^{(l)}$, and the remaining elements as $b_{m,j}^{(l)} = \hat{s}_{m,j+k^{(l)}}^{(l)}, j = 1, \dots, N^{(l)} - k^{(l)}$. We apply two types of constraint to all the M training instances.

MMD Loss Maximum mean discrepancy (MMD) loss [19,20] was originally proposed in Generative adversarial nets (GANs) to quantify the difference of the scores between real and generated samples. In our application, we define MMD loss for the pooling layer as:

$$L_{MMD}^{(l)} = -\frac{1}{M} \sum_{m=1}^M \left[\frac{1}{(k^{(l)})^2} \sum_{i,j=1}^{k^{(l)}} \kappa(a_{m,i}^{(l)}, a_{m,j}^{(l)}) + \frac{1}{(N^{(l)} - k^{(l)})^2} \sum_{i,j=1}^{N^{(l)} - k^{(l)}} \kappa(b_{m,i}^{(l)}, b_{m,j}^{(l)}) - \frac{2}{k^{(l)}(N^{(l)} - k^{(l)})} \sum_{i=1}^{k^{(l)}} \sum_{j=1}^{N^{(l)} - k^{(l)}} \kappa(a_{m,i}^{(l)}, b_{m,j}^{(l)}) \right],$$

where $\kappa(a, b) = \exp(-\|a - b\|^2 / \sigma)$ is a Gaussian kernel and σ is a scaling factor.

BCE Loss Ideally, the scores for the selected nodes should be close to 1 and the scores for the unselected nodes should be close to 0. Binary cross entropy (BCE) loss is calculated as:

$$L_{BCE}^{(l)} = -\frac{1}{M} \sum_{m=1}^M \frac{1}{N^{(l)}} \left[\sum_{i=1}^{k^{(l)}} \log(a_{m,i}^{(l)}) + \sum_{i=1}^{N^{(l)} - k^{(l)}} \log(1 - b_{m,i}^{(l)}) \right]. \quad (5)$$

The effect of this constraint will be shown in Section 4.3.

3.2 Group-level Consistency Loss

Note that $\mathbf{s}^{(l)}$ in Eq. (4) is computed from the input $H^{(l)}$. Therefore, for $H^{(l)}$ from different instances, the ranking of the entries of $\mathbf{s}^{(l)}$ can be very different. For our application, we want to find the common patterns/biomarkers for a certain neuro-prediction task. Thus, we add regularization to force the $\mathbf{s}^{(l)}$ vectors to be similar for different input instances in the first pooling layer, where the group-level biomarkers are extracted. We call the novel regularization group-level consistency (GLC) and only apply it to the first pooling layer, as the nodes in the following layers from different instances might be different. Suppose there are M_c instances for class c in a batch, where $c \in \{1, \dots, C\}$ and C is the number of classes. We form the scoring matrix $S_c^{(1)} = [\mathbf{s}_{1,c}^{(1)}, \dots, \mathbf{s}_{M_c,c}^{(1)}]^\top \in \mathbb{R}^{M_c \times N}$. The GLC loss can be expressed as:

$$L_{GLC}^c = \frac{1}{M_c^2} \sum_{i=1}^{M_c} \sum_{j=1}^{M_c} \|\mathbf{s}_{i,c}^{(1)} - \mathbf{s}_{j,c}^{(1)}\|_2 = 2\text{Tr}((S_c^{(1)})^\top L S_c^{(1)}), \quad (6)$$

where $L_c = D_c - W_c$, W_c is a $M_c \times M_c$ matrix with all 1s, D_c is a $M_c \times M_c$ diagonal matrix with M_c as diagonal elements. We propose to use Euclidean distance for $\mathbf{s}_{i,c}$ and $\mathbf{s}_{j,c}$ due to the benefits of convexity and computational efficiency.

前k个重要性尽可能相同，后面的重要性尽可能相同，它们之间的重要性尽可能有区别



最小化同一社团内的节点的重要性距离

Cross entropy loss L_{ce} is used for the final prediction. Then, the final loss function is formed as:

$$L_{total} = L_{ce} + \lambda_1 \sum_{l=1}^L L_{Dist}^{(l)} + \lambda_2 \sum_c^C L_{GLC}^c, \quad (7)$$

where λ 's are tunable hyper-parameters, l indicates the l^{th} GNN block and L is the total number of GNN blocks, $Dist$ is either MMD or BCE.

4 Experiments and Results

4.1 Data and Preprocessing

We collected fMRI data from a group of 75 ASD children and 43 age and IQ-matched healthy controls (HC), acquired under the "biopoint" task [21]. The fMRI data was preprocessed following the pipeline in Yang et al. [22]. The Desikan-Killiany [23] atlas was used to parcellate brain images into 84 ROIs. The mean time series for each node was extracted from a random 1/3 of voxels in the ROI by bootstrapping. In this way, we augmented the data 10 times. Edges were defined by top 10% positive partial correlations to achieve sparse connections. If this led to isolated nodes, we added back the largest edge to each of them. For node attributes, we used Pearson correlation coefficient to node 1 – 84. Pearson correlation and partial correlation are different measures of fMRI connectivity. We aggregate them by using one to build edge connections and the other to build node features.

4.2 Implementation Details

The model architecture was implemented with 2 conv layers and 2 pooling layers as shown in Fig. 2, with parameter $d^{(0)} = 84, d^{(1)} = 16, d^{(2)} = 16$. We designed a 3-layer MLP (with 16, 8 and 2 neurons in each layer) that takes the flattened graph $\mathbf{z} \in \mathbb{R}^{16}$ as input and predicts ASD vs. HC. The pooling layer kept the top 50% important nodes ($k^{(l)} = 0.5N^{(l)}$). We will discuss the variation of λ_1 and λ_2 in Section 4.3. We randomly split the data into five folds based on subjects, which means that the graphs from a single subject can only appear in either the training or test set. Four folds were used as training data, and the left-out fold was used for testing. Adam was used as the optimizer. We trained the model for 100 epochs with an initial learning rate of 0.001, annealed to half every 20 epochs. We set $\sigma = 5$ in the MMD loss to match the same scale as BCE loss.

4.3 Hyperparameter Discussion and Ablation Study

We tuned the parameters λ_1 and λ_2 in the loss function Eq. (7) and showed the results in Table 1. λ_1 encouraged more separable node importance scores for selected and unselected nodes after pooling. λ_2 controlled the similarity of the

Table 1: Model variations and hyperparameter (λ_1 - λ_2) discussion.

Loss	Pool	0-0	0.1-0	0.1-0.1	0.1-0.5	0.1-1
MMD	TopK	0.753(0.042)	0.784(0.062)	0.781(0.038)	0.780(0.059)	0.744(0.060)
	SAGE	0.751(0.022)	0.770(0.039)	0.771(0.051)	0.773(0.047)	0.751(0.050)
BCE	TopK	0.750(0.046)	0.779(0.053)	0.797(0.051)	0.789(0.066)	0.762(0.044)
	SAGE	0.755(0.041)	0.767(0.033)	0.773(0.047)	0.764(0.050)	0.755(0.041)

Table 2: Comparison with different baseline models.

Model	SVM	Random Forest	MLP	BrainNetCNN [24]	Li et al. [25]	PR-GNN*
Acc	0.686(0.111)	0.723(0.020)	0.727(0.047)	0.781(0.044)	0.753(0.033)	0.797(0.051)
#Par	3k	3k	137k	1438k	16k	6k

Acc: Accuracy; #Par: The number of trainable parameters; PR-GNN*: TopK+BCE.

selected nodes for instances within the same class. A larger λ_2 moves toward group-level interpretation of biomarkers. We first performed an ablation study by comparing setting (0-0) and (0.1-0). Mean accuracies increased at least 3% in TopK (1-2% in SAGE) with MMD or BCE loss. To demonstrate the effectiveness of L_{Dist} , we showed the distribution of node pooling scores of the two pooling layers in Fig. 3 over epochs for different combination of pooling functions and distance losses, with $\lambda_1 = 0.1$ and $\lambda_2 = 0$. In the early epochs, the scores centered around 0.5. Then the scores of the top 50% important nodes moved to 1 and scores of unimportant nodes moved to 0 (less obvious for the second pooling layer using SAGE, which may explain why SAGE got lower accuracies than TopK). Hence, significantly higher scores were attributed to the selected important nodes in the pooling layer. Then, we investigated the effects of λ_2 on the accuracy by varying it from 0 to 1, with λ_1 fixed at 0.1. Without L_{GLC} , the model was easier to overfit to the training set, while larger L_{GLC} may result in underfitting to the training set. As the results in Table 1 show, the accuracy increased when λ_2 increased from 0 to 0.1 and the accuracy dropped if we increased λ_2 to 1 (except for TopK+MMD). For the following baseline comparison experiments, we set λ_1 - λ_2 to be 0.1-0.1.

4.4 Comparison with Existing Models

We compared our method with several brain connectome-based methods, including Random Forest (1000 trees), SVM (RBF kernel), and MLP (one 20 nodes hidden layer), a state-of-the-art CNN-based method, BrainNetCNN [24] and a recent GNN method on fMRI [25], in terms of accuracy and number of parameters. We used the parameter settings indicated in the original paper [24]. The inputs and the architecture parameter setting (node conv, pooling and MLP layers) of the alternative GNN method were the same as PR-GNN. The inputs of BrainNetCNN were Pearson correlation matrices. The inputs of the other alternative methods were the flattened up-triangle of Pearson correlation matrices. Note that the inputs of GNN models contained both Pearson and partial

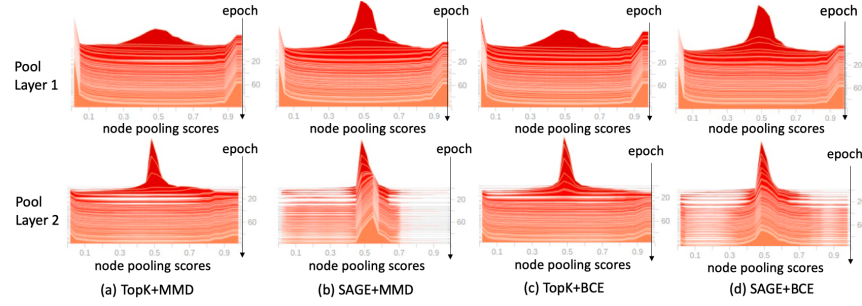


Fig. 3: Distributions of node pooling scores over epochs (offset from far to near).

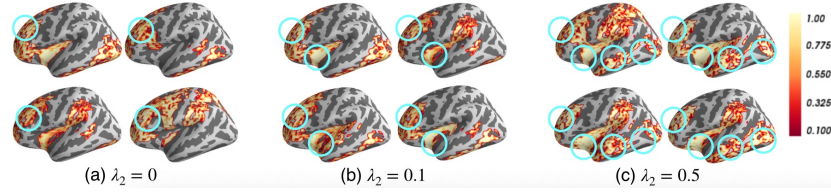


Fig. 4: Selected salient ROIs (importance score indicated by yellow-red color) of four randomly selected ASD individuals with different weights λ_2 on GLC. The commonly detected salient ROIs across different individuals are circled in green.

correlations. For a fair comparison with the non-GNN models, we used Pearson correlations (node features) as their inputs, because Pearson correlations were the embedded features, while partial correlations (edge weights) only served as message passing filters in GNN models. The results are shown in Table 2. Our PR-GNN outperformed alternative models. With regularization terms on the pooling function, PR-GNN achieved better accuracy than the recent GNN [25]. Also, PR-GNN needs only 5% parameters compared to the MLP and less than 1% parameters compared to BrainNetCNN.

4.5 Biomarker Interpretation

Without losing generalizability, we investigated the selected salient ROIs using the model TopK+BCE ($\lambda_1 = 0.1$) with different levels of interpretation by tuning λ_2 . As we discussed in Section 3.2, large λ_2 led to group-level interpretation and small λ_2 led to individual-level interpretation. We varied λ_2 from 0-0.5. Without losing generalizability, we show the salient ROI detection results of four randomly selected ASD instances in Fig. 4. We show the remaining 21 ROIs after the 2nd pooling layer (with pooling ratio = 0.5, 25% nodes left) and corresponding node pooling scores. As shown in Fig. 4(a), when $\lambda_2 = 0$, we could rarely find any overlapped area among the instances. In Fig. 4(b-c), we circled the large overlapped areas across the instances. By visually examining the salient ROIs, we found two overlapped areas in Fig. 4(b) and four overlapped

areas in Fig. 4(c). By averaging the node importance scores (1st pooling layer) over all the instances, dorsal striatum, thalamus and frontal gyrus were the most salient ROIs associated with identifying ASD. These ROIs are related to the neurological functions of social communication, perception and execution [26,27,28,29], which are clearly deficient in ASD.

5 Conclusion

In this paper, we propose PR-GNN, an interpretable graph neural network for fMRI analysis. PR-GNN takes graphs built from fMRI as inputs, then outputs prediction results together with interpretation results. With the built-in interpretability, PR-GNN not only performs better on classification than alternative methods, but also detects salient brain regions for classification. The novel loss term gives us the flexibility to use this same method for individual-level biomarker analysis (small λ_2) and group-level biomarker analysis (large λ_2). We believe that this is the first work using a single model in fMRI study that fills the critical interpretation gap between individual- and group-level analysis. Our interpretation results reveal the salient ROIs to identify autistic disorders from healthy controls. Our method has a potential for understanding neurological disorders, and ultimately benefiting neuroimaging research. We will extend and validate our methods on larger benchmark datasets in future work.

Acknowledgements

This research was supported in part by NIH grants [R01NS035193, R01MH100028].

References

1. M. D. Kaiser, C. M. Hudac, S. Shultz, S. M. Lee, C. Cheung, A. M. Berken, B. Deen, N. B. Pitskel, D. R. Sugrue, A. C. Voos, *et al.*, “Neural signatures of autism,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 49, pp. 21223–21228, 2010.
2. A. A. Goldani, S. R. Downs, F. Widjaja, B. Lawton, and R. L. Hendren, “Biomarkers in autism,” *Frontiers in psychiatry*, vol. 5, p. 100, 2014.
3. J. T. Baker, A. J. Holmes, G. A. Masters, B. T. Yeo, F. Krienen, R. L. Buckner, and D. Öngür, “Disruption of cortical association networks in schizophrenia and psychotic bipolar disorder,” *JAMA psychiatry*, vol. 71, no. 2, pp. 109–118, 2014.
4. E. McDade, G. Wang, B. A. Gordon, J. Hassenstab, T. L. Benzinger, V. Buckles, A. M. Fagan, D. M. Holtzman, N. J. Cairns, A. M. Goate, *et al.*, “Longitudinal cognitive and biomarker changes in dominantly inherited alzheimer disease,” *Neurology*, vol. 91, no. 14, pp. e1295–e1306, 2018.
5. K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. Duncan, F. Morales, and A. Evans, “A general statistical analysis for fmri data,” *Neuroimage*, vol. 15, no. 1, pp. 1–15, 2002.

6. R. A. Poldrack, Y. O. Halchenko, and S. J. Hanson, "Decoding the large-scale structure of brain function by classifying mental states across individuals," *Psychological science*, vol. 20, no. 11, pp. 1364–1372, 2009.
7. X. Wang, X. Liang, Z. Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, H. Wang, Y. Li, Y. Zhu, F. Wu, *et al.*, "Decoding and mapping task states of the human brain via deep learning," *Human Brain Mapping*, 2019.
8. W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, pp. 1024–1034, 2017.
9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
10. M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," in *Advances in neural information processing systems*, pp. 136–144, 2016.
11. M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, pp. 3844–3852, 2016.
12. I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 11, pp. 1944–1957, 2007.
13. Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Advances in neural information processing systems*, pp. 4800–4810, 2018.
14. H. Gao and S. Ji, "Graph u-nets," *arXiv preprint arXiv:1905.05178*, 2019.
15. J. Lee, I. Lee, and J. Kang, "Self-attention graph pooling," *arXiv preprint arXiv:1904.08082*, 2019.
16. P. Veličković *et al.*, "Graph attention networks," in *ICLR*, 2018.
17. H. Yang, X. Li, Y. Wu, S. Li, S. Lu, J. S. Duncan, J. C. Gee, and S. Gu, "Interpretable multimodality embedding of cerebral cortex using attention graph network for identifying bipolar disorder," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 799–807, Springer, 2019.
18. L. Gong and Q. Cheng, "Exploiting edge features for graph neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9211–9219, 2019.
19. A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.
20. C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, "Mmd gan: Towards deeper understanding of moment matching network," in *Advances in Neural Information Processing Systems*, pp. 2203–2213, 2017.
21. M. D. Kaiser *et al.*, "Neural signatures of autism," *PNAS*, 2010.
22. D. Yang *et al.*, "Brain responses to biological motion predict treatment outcome in young children with autism," *Translational psychiatry*, vol. 6, no. 11, p. e948, 2016.
23. R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, *et al.*, "An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest," *Neuroimage*, vol. 31, no. 3, pp. 968–980, 2006.
24. J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh, "Brainnetcnn: Convolutional neural networks

- for brain networks; towards predicting neurodevelopment,” *NeuroImage*, vol. 146, pp. 1038–1049, 2017.
25. X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, “Graph neural network for interpreting task-fmri biomarkers,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 485–493, Springer, 2019.
 26. M. Schuetze, M. T. M. Park, I. Y. Cho, F. P. MacMaster, M. M. Chakravarty, and S. L. Bray, “Morphological alterations in the thalamus, striatum, and pallidum in autism spectrum disorder,” *Neuropsychopharmacology*, vol. 41, no. 11, pp. 2627–2637, 2016.
 27. A. Y. Hardan, R. R. Girgis, J. Adams, A. R. Gilbert, M. S. Keshavan, and N. J. Minshew, “Abnormal brain size effect on the thalamus in autism,” *Psychiatry Research: Neuroimaging*, vol. 147, no. 2-3, pp. 145–151, 2006.
 28. J. P. Bhanji and M. R. Delgado, “The social brain and reward: social information processing in the human striatum,” *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 1, pp. 61–73, 2014.
 29. C. Press, N. Weiskopf, and J. M. Kilner, “Dissociable roles of human inferior frontal gyrus during action execution and observation,” *Neuroimage*, vol. 60, no. 3, pp. 1671–1677, 2012.