University of São Paulo

Institute of Mathematical Sciences and Computation

Research project

Explainability in Graph Neural Networks for Autism Assessment Using fMRI Analysis

Candidate: Matheo Angelo Pereira Dantas

Advisor: Dr. André Carlos Ponce de Leon Ferreira de Carvalho

Concentration area: Artificial Intelligence, Interpretability, Autism Spectrum Disorder, Functional Magnetic Resonance Imaging

São Carlos - June, 2024

Abstract

Autism diagnosis is made only from clinical analysis, which is often inaccessible. The use of AI could help doctors in improving the diagnosis with biological data, such as the Functional Magnetic Resonance Imaging (fMRI)[1]. However, such complex data usually relies on the use of similarly complex black-box models, so although the predictions can be precise, the knowledge learned by the network gets lost. Therefore, it is important to use the techniques of Explainable AI to uncover the rationales behind the predictions of the models. Within this context, we propose to conduct an investigation on such techniques. The main goal of the project is to study explainability methods applicable to Graph Neural Networks, which are machine learning architectures designed to operate on graphs, in order to observe the neurological patterns of autistic brains. For that purpose, we propose to test the currently available explainability algorithms for GNNs on our data and make a comparative analysis, using objective measures of explainability.

1 Introduction

1.1 Motivation

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by social and communication impairments and rigid and repetitive patterns of behaviour and interests. Autism is a very common disability, as recent studies reported that ASD was prevalent among 1 in 44 children in the US[2]. It is very important to diagnose autistic individuals early, in order to prevent psychic suffering[3] and enable them to access their disability rights. However, ASD has no discriminative biomarkers, so it can only be diagnosed from the observation and analysis of the individual's behavior by doctors and psychologists. Therefore, to make the diagnostic process more accessible, it is important to investigate the biological basis of autism.

Towards that goal, we propose the use of Graph Neural Networks (GNNs)[4]. GNNs are deep learning models designed to operate on graph-structured inputs and learn sophisticated patterns from them, taking into consideration it's topology. They have been extensively applied in aiding scientific discovery, being capable, for example, of identifying new antibiotics[5]. Therefore, they are promising tools to aid in the task of biological investigation of autism, as they have the potential to detect patterns that qualitative scientific investigation hasn't discovered yet.

However, neural networks are generally not interpretable, as they perform very large and complex calculations, so we don't have immediate access to the knowledge discovered by the network. The area of Explainable AI was created to address the problem of interpretability, by providing post-hoc techniques that extract semantic insight hidden in black-box machine learning models after they are trained. Explainability is very important in modern medical AI research, as it is one of the most frequent domains in which the 'explainability' keyword appears in scientific literature [6].

It is also worth noting that the use of Explainable AI for Graph Neural Networks is still in early development. Explainability research usually focuses on other domains, such as text and images, while graphs are less explored[7]. Graph Neural Networks are also more challenging to explain, as they process important discrete topological information from the input instead of solely numerical data. For those reasons, it is especially important to conduct more careful research on explainability when using Graph Neural Networks instead of other Deep Learning architectures.

In conclusion, it is important to develop research on AI-assisted evaluation of autism diagnosis, and a key aspect of those technologies is interpretability. Furthermore, within that context, the algorithm employed to make predictions in our research demands special attention to that aspect. The next section will present the main concepts about explainable GNNs necessary to follow the text.

1.2 Explainability in Graph Neural Networks

The goal of explainability methods for neural networks is to identify what patterns of the input of the network are important for it's predictions, and present them in the form of explanations, in a way that is humanly intelligi-

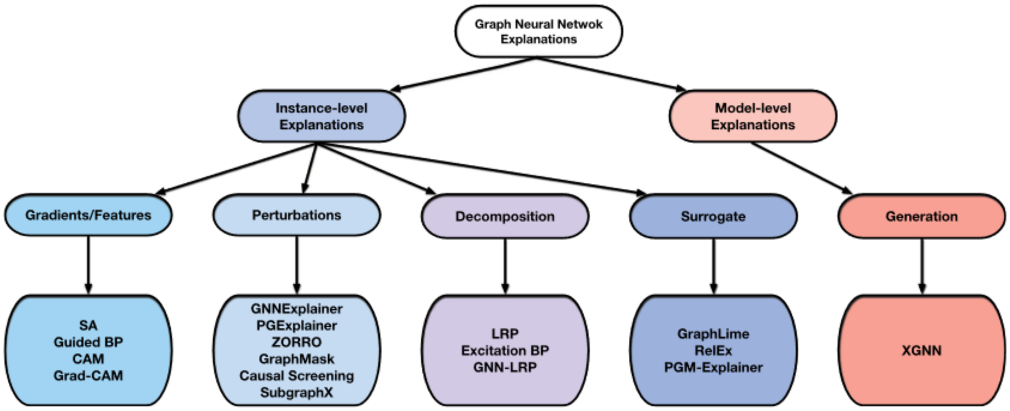

Figure 1: Taxonomy of existing explainability algorithms for Graph Neural Networks. Source: [7].

ble. There is a variety of ways to create and evaluate these explanations in the context of Graph Neural Networks, as will be explained in detail in the sections below.

1.2.1 Explainability Methods

Here, we follow the taxonomic survey proposed in 1. In Graph Neural Networks, there are two main approaches to explainability: Model-level and Instance-level. Instance-level algorithms explain the results of a particular input, highlighting parts of the input that were decisive for it's prediction. On the other hand, model-level algorithms explain the trained model as a whole, usually by detecting important patterns for the prediction of each class.

There are not many model-level algorithms for GNNs in the literature, because it is difficult to explain the entire GNN model due to it's discrete nature. The only relatively popular model-level algorithm for GNNs is XGNN[8], while other lesser known techniques have been developed later, namely GNNInterpreter[9] and PAGE[10]. XGNN and GNNInterpreter are based on generation: they generate input graphs that maximize the probabilities of each class. PAGE is a prototype-based technique, as it looks for "prototypes" (graph motifs [11]) that are common in each class.

As for instance-level algorithms, there is a much larger variety of methods available. They can be split into four different strategies: Gradients/Features, Perturbation, Decomposition, and Surrogate. The first consists of creating a heatmap of importance on the nodes, either from the magnitude of the gradient of the node features with respect to the output of the GNN[12], or from a projection of the last hidden feature maps into the input, in the specific case of Graph Convolutional Networks[13]; Perturbation-based algorithms [14][15][16][17][18][19] alter or remove features (including nodes and edges) to search for perturbations that keep the prediction unchanged, obtaining a mask of important features in the graph. Decomposition-based methods [12][20][21][13] spread the output of the GNN through the input features, giving a bigger fraction of the result to the most important features. Lastly, Surrogate-based methods are used to train an interpretable model, such as a Decision Tree, to fit neighboring data, which could consist of similar graphs in the case of graph classification[22], or the neighboring nodes of the explained node, in the case of node classification[23][24].

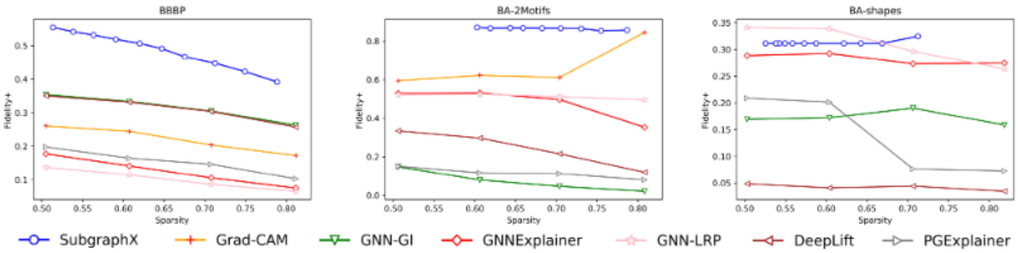

Figure 2: An experiment with different explainability methods for GNNs in a graph of fidelty in function of sparsity. Source: [7].

1.2.2 Measures to assess Explainability

It is important to evaluate the explanations using objective measures, as they provide a ground-truth view of the performance of the algorithm and make it possible to benchmark the algorithm for big amounts of input samples.

The two most frequently used measures are Fidelity and Sparsity[7]. Fidelity measures how well the explanation represents the behavior of the GNN by creating graph masks from the explanations and measuring how much the whole inputs and the masks alone have matching outputs in the model. Sparsity averages the fraction of nodes of the graph that are used for each explanation mask, following the intuition that good explanations should be sparse, showing only the most meaningful graph features.

The fidelity can be artificially increased by selecting bigger masks, and the sparsity can also be increased just by creating emptier masks, all of which can depend on the thresholding of the mask. A way to compare explainability algorithms more fairly, then, is to plot the fidelity in function of sparsity, as in 2.

Other measures include: Contrastivity [13], which is as much high as the explanations for each class are different (explanations should be discriminative for classification); Stability[25], which measures how unaltered the explanations are after small perturbations in the graph; and consistency[25], which measures how similar the explanations are for different trained models on the same task.

Model-level algorithms require different methods for evaluation. Generation-based methods[8][9] evaluate the results using the predicted class probability of the generated graphs. It is also possible to create new ad-hoc measures[10], which would have the downside of being less objective, as we would have no experiments to compare our algorithms to.

1.3 Investigated Application

As previously mentioned, this research project aims to investigate GNN Explanations applied to the classification of ASD diagnosis from fMRI data.

Our dataset consists of graphs generated from fMRI (functional Magnetic Resonance Imaging) exams. The exam records a three-dimensional video, where each voxel (3D pixel) is a BOLD (blood-oxygen-level-dependent) signal, a time series representing the amount of brain activity. Based on the recordings of the BOLD signal for different regions of the brain, we will form a graph where nodes are regions of the brain and edges represent connectivity between those regions.

For those graphs, we will perform two different classification tasks using

GNNs. The first task consists on a binary classification between autistic and neurotypical brain, that is, the prediction of autism diagnosis. For the second task, we classify the intensity of multiple phenotypic manifestations of autism, with the intent of identifying what are the specific support needs of the patient in different areas, such as communication skills and executive functions. The labels for that classification are obtained by discretizing the scores of the individuals in surveys correspondent to each phenotype.

After the models are trained, we aim to provide explanations for the results of our model, by highlighting regions of the brain that have differentiated activation in different manifestations of autism, or in autistic people generally. We would like to provide those explanations for individual exams, mapping brain activity that highly suggests autism or signs of autism in that person, and provide examples of exams where similar patterns were key to identify the predicted class.

With that in mind, this project should use explainability algorithms and performance measures that are compatible with the task of graph classification and that identify important regions of the brain, which corresponds to important nodes in the graph input. Those can be both instance-level explanations, that justify the predictions for a given individual, and model-level explanations, that help us understand the neural functioning of autistic people. Additionally, we should also be aware of which methods are better suited for binary classification or for multi-class classification.

2 Objective and Activities

The primary objective of this research project is to identify possible biomarkers of ASD present in fMRI exams, with the use of Graph Neural Networks and Explainable AI.

This involves following the specific objectives below:

- Study the use of Graph Neural Networks in classification of autism from fMRI analysis;
- Perform experiments on the data using explainability methods;
- Select the methods with the best scores of explainability;
- Identify the regions of the brain that are more active in autistic individuals, according to the selected GNN explanations;
- Compare the output of the experiments with neuroscience literature, to better understand the manifestation of autism in the brain.

3 Methodology

3.1 Sourced data

The data to be used in this project will be obtained from the ABIDE II dataset[1], which is a public dataset containing 1114 samples of fMRI exams, with labels of autistic and neurotypical for each sampled individual and their scores on the phenotypic surveys. It is a public dataset, approved by an ethics committee.

3.1.1 Pre-processing

As previously mentioned, we will construct a graph for each individual based on their fMRI exam. To do this, we group the voxels into different Regions Of Interest (ROI), which are important regions of the brain, and the BOLD series of each ROI is the average of the series for all voxels contained in that ROI. After that, we calculate the Pearson Correlation of the BOLD Series for each pair of ROI, and then construct the graph, where the nodes represent ROI and the edges are inserted for the 10% pairs of ROI with the highest correlation.

3.2 Experiments

In both of our binary and multiclass explanation tasks, the main goal is to highlight important regions of the brain for the chosen class. Therefore, we

[1] https://fcon$_1$000.$projects.nitrc.org/indi/abide/abide_I I.html$

must use algorithms that explain graph classification, and do so by creating node masks, or giving importance scores to the nodes. Among the methods mentioned in the previous sections, the best candidates are all of the gradient-based and feature-based methods (SA[12], Guided BP[12], CAM[13], and GradCAM[13]), the perturbation-based algorithms that create node masks (Zorro[16], SubgraphX[17], and GNNExplainer[15]), and the decomposition-based algorithms that target nodes (LRP[20][12] and Excitation BP[13]). We will also study how to adapt other methods to produce node importance representations.

We would like to have class-specific explanations, i.e. what patterns in the brain are important for the identification of autism. Initially, it would seem intuitive to use a classic model-level explanation method like GNNInterpreter[9], as it generates coherent input graphs by maximizing the similarity between the explanations and the graphs in the dataset, rather than demanding the use of specific rules of graph generation[8], which are unknown in our neurological application. However, model-level methods create optimal inputs rather that highlighting important input features, and adaptations would be needed in order to obtain node importance explanations from the input optimization. Instead, we will observe the ROIs that are most commonly selected in instance-level graph masks of each class, as in [26].

After the explanations are generated, they will be evaluated with the Fidelity-Sparsity graph. In the binary classification task, we will also use the Contrastivity metric, which wouldn't be suited for multiclass classification because it is possible that the input has features that are important to distinguish a group of classes, and Contrastivity would then penalize redundancy within that group of classes. Additionally, we will investigate ad-hoc measures that can be suited for our specific application.

Additionally, it is important to note that the phenotypic intensity classes are obtained by discretizing a continuous value in the survey score to enable the use of classification and compatible techniques, which are more common in explainability methods for GNNs[7]. We may also study how to adapt the explanations to the case of regression of the survey scores.

The source code for this project will be made available in GitHub, or similar, repository, primarily written in the Python language.

3.3 Expected results

We expect to find explainability methods that achieve satisfying scores on the objective metrics, which are our only tool to visualize explainability power, as our application lacks a known ground-truth scientific basis.

After we obtain the most reliable explanations, we will compare the produced outputs with the qualitative findings of modern neuroscience literature, to identify which parts of the brain are more active in autistic individuals, and how brain activity is related to the manifestation of autism. The study in [26] exemplifies this well, by printing a word cloud obtained from a meta-analysis research of the brain regions identified as more active and as less active in autistic brains in comparison with TD (Typical Development) brains.

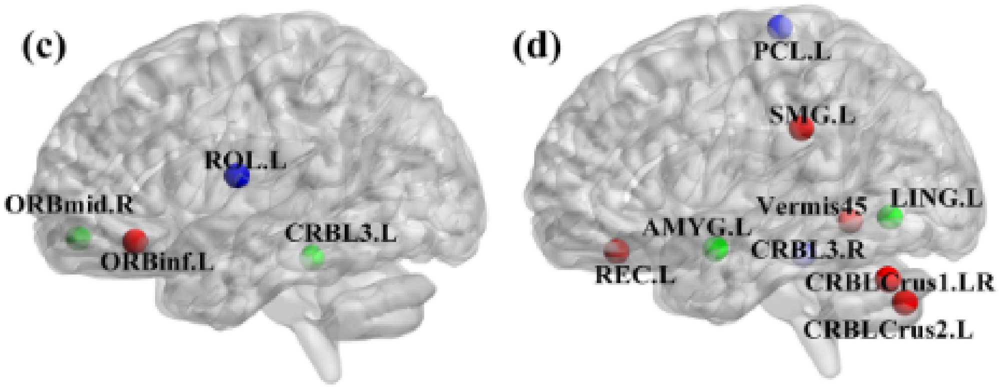

Figure 3: Regions of the brain that were identified as important to distinguish ASD and neurotypical brains in [26]. (c) Includes regions that are more active, and (d) includes regions that are less active in autistic brains.

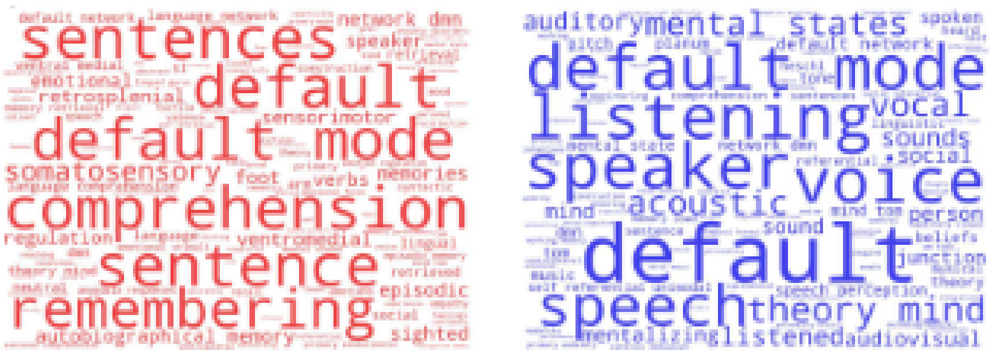

Figure 4: Word cloud created in [26] from meta-analysis of important brain regions identified by explainability algorithms for GNNs. The red cloud was samples from search of active regions in ASD brains, and the blue cloud was sampled from search of inactive regions.

4 Schedule

Activities	1-2	3-4	5-6	7-8	9-10	11-12
1. Autism Meta-Analysis.	▬					
2. Technical Meta-Analysis.	▬	▬				
3. Implementation of Explainability Methods.		▬	▬			
4. Evaluation with Objective Metrics.			▬			
5. Midterm Report.				▬		
6. Improvement Identification.				▬		
7. Experimentation with the Improvements.				▬	▬	
8. Assess Explainability Metrics.					▬	
9. Analysis of Results.						▬

Table 1: Activity schedule by month

5 Infrastructure

The proposed project will use the computing resources of the Analytics laboratory at ICMC-USP, which has several networked computers and workstations, as well as 1 cluster with 5 nodes with Intel Xeon CPU X5690 @ 3.47GHz processors, 4 of them with 2 processors and 1 of them with only 1 processor. Of the 5 nodes, 2 have 80GB of RAM, 2 have 64GB of RAM and 1 has 32GB of RAM. The laboratory also has 1 GPU Server with 2 Intel Xeon E5-2620V2 Processors, 128 GB Memory, 1 Seagate 3.5" 2TB SATA HDD, and 2 NVIDIA K20M passive cooling GPUs. The laboratory has the technical support of the ICMC-USP IT department to carry out the project.

This project will use the bibliographic resources from the Prof. Achille Bassi library at ICMC-USP [2], which has a large collection in the areas of Mathematics, Computing, Statistics, and related sciences. The collection contains around 25,000 volumes of books and 700 titles of periodicals, as well as technical reports, conference proceedings, microfiche, CDs, and DVDs. The library is part of the COMUT Program, in which it is integrated as a Base Library.

[2] http://www.icmc.usp.br/ biblio

References

[1] S. Zhang and R. L. Chiang-shan, “Functional connectivity mapping of the human precuneus by resting state fmri,” *Neuroimage*, vol. 59, no. 4, pp. 3548–3562, 2012.

[2] M. J. Maenner, “Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2018,” *MMWR. Surveillance Summaries*, vol. 70, 2021.

[3] M. Hosozawa, A. Sacker, and N. Cable, “Timing of diagnosis, depression and self-harm in adolescents with autism spectrum disorder,” *Autism*, vol. 25, no. 1, pp. 70–78, 2021.

[4] G. Corso, H. Stark, S. Jegelka, T. Jaakkola, and R. Barzilay, “Graph neural networks,” *Nature Reviews Methods Primers*, vol. 4, no. 1, p. 17, 2024.

[5] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, *et al.*, “A deep learning approach to antibiotic discovery,” *Cell*, vol. 180, no. 4, pp. 688–702, 2020.

[6] G. Vilone and L. Longo, “Explainable Artificial Intelligence: A Systematic Review,”

[7] H. Yuan, H. Yu, S. Gui, and S. Ji, “Explainability in graph neural networks: A taxonomic survey,” *CoRR*, vol. abs/2012.15445, 2020.

[8] H. Yuan, J. Tang, X. Hu, and S. Ji, "Xgnn: Towards model-level explanations of graph neural networks," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 430–438, 2020.

[9] X. Wang and H.-W. Shen, "Gnninterpreter: A probabilistic generative model-level explanation for graph neural networks," *arXiv preprint arXiv:2209.07924*, 2022.

[10] Y.-M. Shin, S.-W. Kim, and W.-Y. Shin, "Page: prototype-based model-level explanations for graph neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[12] F. Baldassarre and H. Azizpour, "Explainability techniques for graph convolutional networks," *arXiv preprint arXiv:1905.13686*, 2019.

[13] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10772–10781, 2019.

[14] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," *Advances in neural information processing systems*, vol. 33, pp. 19620–19631, 2020.

[15] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[16] T. Funke, M. Khosla, and A. Anand, "Hard masking for explaining graph neural networks," 2020.

[17] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of graph neural networks via subgraph explorations," in *International conference on machine learning*, pp. 12241–12252, PMLR, 2021.

[18] M. S. Schlichtkrull, N. De Cao, and I. Titov, "Interpreting graph neural networks for nlp with differentiable edge masking," *arXiv preprint arXiv:2010.00577*, 2020.

[19] X. Wang, Y. Wu, A. Zhang, X. He, and T.-s. Chua, "Causal screening to interpret graph neural networks," 2020.

[20] R. Schwarzenberg, M. Hübner, D. Harbecke, C. Alt, and L. Hennig, "Layerwise relevance visualization in convolutional text graph classifiers," *arXiv preprint arXiv:1909.10911*, 2019.

[21] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "Higher-order explanations of graph neural networks via relevant walks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7581–7596, 2022.

[22] M. Vu and M. T. Thai, "Pgm-explainer: Probabilistic graphical model explanations for graph neural networks," *Advances in neural information processing systems*, vol. 33, pp. 12225–12235, 2020.

[23] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "Graphlime: Local interpretable model explanations for graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[24] Y. Zhang, D. Defazio, and A. Ramesh, "Relex: A model-agnostic relational model explainer," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 1042–1049, 2021.

[25] B. Sanchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. Qian, K. McCloskey, L. Colwell, and A. Wiltschko, "Evaluating attribution for graph neural networks," *Advances in neural information processing systems*, vol. 33, pp. 5898–5910, 2020.

[26] Y. Chen, J. Yan, M. Jiang, T. Zhang, Z. Zhao, W. Zhao, J. Zheng, D. Yao, R. Zhang, K. M. Kendrick, and X. Jiang, "Adversarial learning based node-edge graph attention networks for autism spectrum disorder identification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2022.