Universidade de São Paulo

Institute of Science Mathematics and Computation

Relatório final

Explicabilidade de Redes Neurais de Grafos Para a Avaliação de Autismo com Análise de fMRI

Aluno: Matheo Angelo Pereira Dantas

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

Concentration area: Artificial Intelligence, Machine Learning, Decision Making

Resumo

O diagnóstico de autismo atualmente só pode ser feito a partir da análise comportamental clínica, que é um procedimento pouco acessível. O uso de Redes Neurais Profundas pode contribuir para aprimorar o diagnóstico com dados biológicos mais complexos, mas os modelos de Aprendizado Profundo não são interpretáveis, então as previsões podem não ser confiáveis e o conhecimento adquirido pelo modelo não pode ser facilmente transposto para os seus usuários. Por essa razão, é importante é importante mitigar esse obstáculo de interpretabilidade com ferramentas de IA Explicável. Nesse contexto, propomos uma investigação dessas técnicas. O objetivo do projeto é estudar métodos de explicabilidade de Redes Neurais de Grafos, que são redes neurais projetadas para fazer previsões a partir de dados com conectividade complexa, e usá-los para investigar os padrões neurológicos presentes em dados cerebrais de pessoas autistas. Para isso, iremos aplicar diferentes técnicas de explicabilidade presentes na literatura científica, avaliá-las com métricas objetivas de interpretabilidade, e visualizar e interpretar os resultados.

Conteúdo

1 Introdução

1.1 Transtorno do Espectro Autista

De acordo com o catálogo de maior referência para transtornos mentais, o DSM-V[1], o TEA (Transtorno do Espectro Autista) é um transtorno do neurodesenvolvimento caracterizado por déficits na comunicação social e padrões repetitivos de comportamento, em diferentes contextos, se manifestando desde a infância. Essas características podem se apresentar com intensidades variadas entre os autistas.

Estima-se que em torno de 1% da população mundial tenha TEA[2]. Pessoas com TEA têm chances maiores de desenvolver depressão e outras condições de sofrimento psíquico [1]. Além disso, o TEA é legalmente reconhecido como uma deficiência no Brasil[3], o que garante o direito ao usufruto de políticas de acessibilidade e ações afirmativas. Assim, a identificação do autismo é uma questão de grande relevância social, sendo necessária para assegurar o devido suporte a esse público.

Entretanto, a condição não possui nenhum marcador biológico conhecido pela ciência, sendo o diagnóstico possível apenas a partir da análise clínica, por meio de testes, entrevistas e observação do comportamento por profissionais de psicologia. Esse processo pode ser inacessível e pouco eficiente em larga escala, motivando a comunidade acadêmica a buscar marcadores biológicos do TEA.

Uma das ferramentas mais utilizadas para esse objetivo é o Aprendizado de Máquina, onde computadores são "treinados"por meio de cálculos esta-

tísticos para identificar padrões em conjuntos de dados. Isso possibilita a detecção de padrões difíceis de serem percebidos pela análise humana manual, e assim, a descoberta de novos conhecimentos na ciência[4]. O aprendizado de máquina já foi empregado em diversos tipos de dados de autistas, como expressões faciais[5] e exames cerebrais[6].

Entre essas modalidades de dados, destaca-se o fMRI (Functional Magnetic Resonance Imaging[7]), um exame que filma o cérebro do paciente usando ressonância magnética. O exame grava a intensidade do sinal BOLD (Blood-Oxygen-Level-Dependent) em cada região do cérebro, visando analisar a atividade cerebral a partir da circulação de sangue no cérebro. Os exames de fMRI podem ser subdivididos entre task-fMRI, que observa a resposta do cérebro a uma tarefa específica desempenhada pelo paciente, e o R-fMRI (Resting-State fMRI), que apenas observa o cérebro normalmente, em estado de repouso.

1.2 Redes Neurais de Grafos

Uma das técnicas de aprendizado de máquina que têm sido empregadas para tentar extrair padrões biológicos do TEA, principalmente em exames de fMRI, são as Redes Neurais de Grafos, também chamadas de GNNs ("Graph Neural Networks"). As Redes Neurais Profundas, de modo geral, têm tido sucesso em tarefas que usam dados complexos, como imagens e textos, pois podem aproximar funções complexas a partir da composição de funções mais simples, as camadas da rede neural, e assim representar dados em diferentes níveis de abstração. As Redes Neurais de Grafos, em particular, operam em

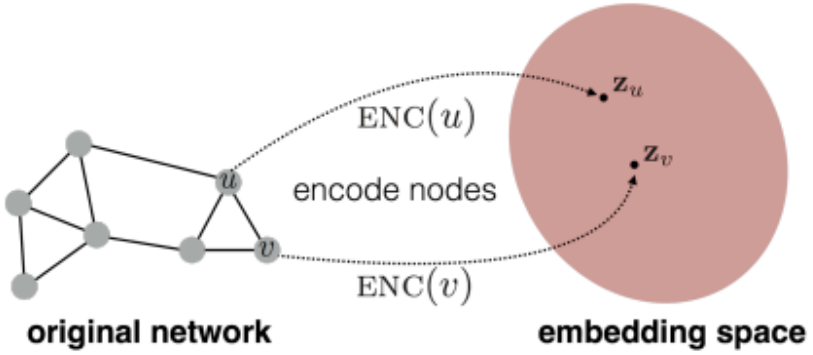

Figura 1: Ilustração do processo de aprendizado de representação em um grafo. Cada nó é associado a um vetor no espaço de embedding, que representa a sua função original no grafo. Fonte: [8].

dados na forma de grafos, onde um grafo consiste em um conjunto de pontos (nós) e um conjunto de conexões entre pares de nós (arestas).

1.2.1 GNNs como métodos de aprendizado de representação de grafos

Como ilustrado na Figura 1, as GNNs são treinadas para obter representações dos nós de um grafo no $\mathbb{R}^n$ (também chamadas de "embeddings") que codifiquem propriedades importantes do grafo. Essas representações podem ser tomadas como os atributos do nó e usadas para classificar o nó, ou podem ser agregadas para representar a estrutura do grafo todo, o que permite a tarefa de classificação de grafo.

1.2.2 Passagem de mensagem

Nas GNNs, essas representações são obtidas por meio do mecanismo de passagem de mensagem. Inicialmente, cada nó possui um vetor (que pode ser um escalar) e, a cada camada da rede neural, os nós atualizam seu valor em função dos valores da vizinhança. Quando esse processo é concluído, o resultado são as representações numéricas dos nós do grafo. Para obtermos uma representação do grafo inteiro, as representações dos nós são agregadas, o que geralmente é chamado de *pooling* ou *readout*. O pooling geralmente é uma operação invariante a permutação, como uma média simples, porque a maioria das aplicações considera que a representação do grafo não deve ser influenciada pela ordem de numeração dos nós[9].

De maneira mais formal, sendo $h_l(u)$ o vetor armazenado no nó u do grafo após a l-ésima camada da rede neural (onde $h_0(u) = x_u$, o vetor de features inicial do nó u), uma camada de passagem de mensagem obtém $h_{l+1}(u)$ a partir do seguinte algoritmo[8]:

- **Mensagem:** todo nó u do grafo recebe uma mensagem $MSG_{l+1}(u, v, e_{uv})$ de cada um dos seus vizinhos $v \in N(u)$, onde e_{uv} é a aresta ligando os nós u e v.

- **Agregação:** as mensagens $MSG_{l+1}(u, v, e_{uv})$ recebidas por u são agregadas com a operação $m_{l+1}(u) = AGG_{l+1}(\{MSG_{l+1}(u, v, e_{uv}), v \in N(u)\})$ invariante a permutação, como uma média, produzindo o vetor $m_{l+1}(u)$.

- **Atualização:** o vetor $h_{l+1}(u)$ é obtido a partir de uma função de atualização da forma $UPD_{l+1}(h_l(u), m_{l+1}(u))$.

1.3 Principais arquiteturas de GNN

A seguir, temos as principais arquiteturas de Redes Neurais de Grafos usadas atualmente:

1.3.1 Graph Convolutional Network

As Graph Convolutional Networks (GCNs)[10] aproximam as convoluções espectrais em grafos, de forma semelhante à ChebNet, mas limitam cada filtro de convolução a considerar apenas os sinais da vizinhança de cada nó, funcionando como uma camada de passagem de mensagem. Assim, uma camada de convolução da GCN é da forma:

$$H^{(l+1)} = \sigma\big(D^{-\frac{1}{2}}\tilde{A}D^{-\frac{1}{2}}H^{(l)}W^{(l)}\big)$$

Onde D é uma matriz diagonal tal que d_{ii} é o grau (número de vizinhos) do vértice i.

1.3.2 GraphSAGE

Uma camada de passagem de mensagem da GraphSAGE[11] segue a equação abaixo:

$$h_{l+1}(u) = \sigma(W^a_{l+1} \cdot h_l(u) + W^b_{l+1} \cdot AGG(\{h_l(v), v \in N(u)\}))$$

Onde $AGG(\cdot)$ é uma função de agregação pré-selecionada, como média simples ou valor máximo de cada coordenada dos vetores; W^a_{l+1} e W^b_{l+1} são

matrizes de pesos a serem otimizadas no treinamento da rede neural; e $\sigma(\cdot)$ é uma função de ativação não-linear.

1.3.3 Graph Attention Network

As Graph Attention Networks (GAT)[12] visam aplicar o mecanismo de atenção popularizado pelos Transformers[13] nas GNN, e calcula escores de atenção entre apenas entre nós adjacentes, ao invés de calcular a atenção para todos os pares de Tokens, como nos Transformers. Uma camada GAT padrão é da forma:

$$h(u) = \sigma\big(\sum_{v \in N_u} \alpha_{uv} W h(v)\big)$$

Onde consideramos que todo nó é adjacente a si mesmo, e α_{uv} é a atenção do nó u para o nó v, dada por:

$$\alpha_{uv} = \frac{\exp(e_{uv})}{\sum_{v \in N_u} \exp(e_{uv})} = \frac{\exp(\text{LeakyRELU}\big(a^T [Wh(u) \| Wh(v)]\big))}{\sum_{v \in N_u} \exp(\text{LeakyRELU}\big(a^T [Wh(u) \| Wh(v)]\big))}$$

Onde $\|$ indica concatenação de vetores e LeakyRELU[...].

1.3.4 Graph Isomorphism Network

A Graph Isomorphism Network (GIN)[14] utiliza o conceito de poder de representação, que é a capacidade de uma rede neural de grafos de gerar saídas diferentes para grafos não-isomórficos entre si. GNNs baseadas em passagem

de mensagem têm poder de representação no máximo tão alto quanto o Teste de Isomorfismo de Weisfeiler-Lehman[15], e a GIN é arquitetada para ser capaz de atingir esse poder de representação, possuindo a seguinte fórmula de passagem de mensagem:

$$h_u^{(k)} = \mathrm{MLP}^{(k)}\big((1+\epsilon^{(k)})h_u^{(k-1)} + \sum\nolimits_{v\in N(u)} h_v^{(k-1)}\big)$$

Onde $\epsilon^{(k)}$ é um número que pode ser fixado como um hiperparâmetro ou pode ser um parâmetro treinável da rede neural, e MLP é uma função Multi-Layer Perceptron[16] com pesos treináveis. Além disso, a camada de pooling global usa uma soma e não uma média.

1.4 Explicabilidade de Redes Neurais de Grafos

Um dos principais impecilhos para a implementação de redes neurais na sociedade é a falta de interpretabilidade. Como as redes neurais realizam cálculos muito complexos, o conhecimento adquirido durante o processo de treinamento não é humanamente inteligível. Com isso, não é possível entender o processo de tomada de decisão das redes neurais, o que dificulta a confiabilidade das previsões e a análise das suas implicações éticas. Para mitigar esse problema, surgiu a área da Inteligência Artificial Explicável (XAI), que visa explicar modelos treinados de Aprendizado de Máquina que têm baixa interpretabilidade.

Os grafos apresentam um desafio adicional para a interpretabilidade, pois ao contrário da maioria das técnicas de aprendizado profundo, os grafos po-

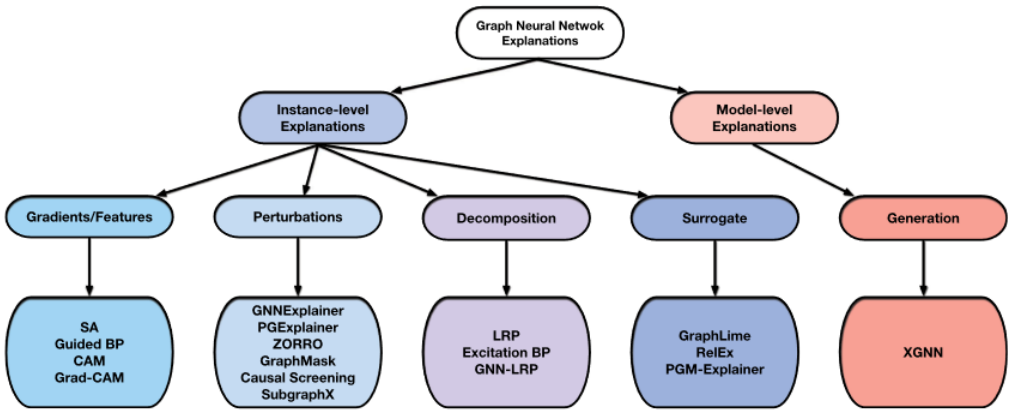

Figura 2: Taxonomia dos algoritmos de Explicabilidade de Redes Neurais de Grafos. Fonte: [19].

dem apresentar diferentes padrões de conectividade em uma mesma aplicação, o que deve ser interpretado em conjunto com os sinais numéricos dos dados.

Há uma série de aspectos das GNNs que podem ser objeto de investigação dos algoritmos de explicabilidade, e eles podem ser subdivididos de acordo com uma taxonomia de algoritmos[17][18], ilustrada na Figura 2.

1.4.1 Métodos de Saliência

Os métodos de saliência, adaptados das Redes Neurais Convolucionais, apresentam mapas de calor que medem o grau de influência de cada parte do grafo em uma determinada previsão. Isso pode ser medido através do gradiente das features de cada nó em relação à saída da rede neural, como é no caso do SA e do Guided BP[20], ou através dos mapas de ativação de classe

(CAM e GradCAM[21]), que salientam nós com valores altos nas features que têm influência positiva na classe predita, usando as features dos embeddings finais dos nós.

1.4.2 Métodos de Perturbação

O grupo mais numeroso de métodos de interpretabilidade de GNNs é o dos métodos de perturbação. Esses métodos consistem em perturbar as features (nós, arestas, ou features de nós) do grafo e analisar quais perturbações pequenas podem ter um efeito significativo na previsão, geralmente com o objetivo de gerar uma "máscara"com as partes mais importantes do grafo. Dependendo do tipo de perturbação, essas máscaras podem ser máscaras de nós (Zorro[22], SubgraphX[23]), arestas (GNNExplainer[24], PGExplainer[25], GraphMask[26], CausalScreening[27]), ou features de nós (GNNExplainer[24], Zorro[22]).

Além disso, podemos classificar esses métodos entre geradores de máscaras suaves (GNNExplainer[24]), que aceitam inserção parcial de features na máscara; discretas (Zorro[22], Causal Screening[27], SubgraphX[23], Gem[28]), onde cada feature está totalmente dentro ou fora da máscara; e aproximadamente discretas (PGExplainer[25], GraphMask[26]), que também aceitam inserção parcial, mas são otimizadas para evitar valores intermediários de inserção, aproximando máscaras discretas. Máscaras suaves podem enfrentar o problema da evidência introduzida[29], onde a remoção parcial de uma feature pode alterar o significado dela para a rede neural, ao invés de mascarar a contribuição da feature.

Além da geração de máscaras, outra abordagem possível é a geração de contra-factuais (CF-GNNExplainer[30], RCExplainer[31]), isto é, grafos parecidos com o grafo de entrada (no caso da classificação de nó, o grafo computacional do nó) mas que gerem resultados diferentes.

1.4.3 Métodos de Decomposição

Nos métodos de decomposição, a saída da rede neural é decomposta sucessivamente nas camadas da rede neural, aplicando regras de decomposição baseadas em Backpropagation[16] da última camada até a entrada da GNN, onde podemos visualizar a parcela de contribuição de cada feature. Essa decomposição pode ser feita entre os nós (LRP[32] e Excitation BP[21]), ou entre caminhos (GNN-LRP[33]) de tamanho k, onde k é o número de camadas de passagem de mensagem da GNN, de forma a exibir a circulação de informação no grafo por meio das passagens de mensagem da GNN.

1.4.4 Métodos de Substituição

Métodos de Substituição empregam modelos interpretáveis de aprendizado de máquina clássico (como Regressão Linear) que se ajustem à fronteira de decisão local de uma determinada previsão, substituindo a GNN. Os modelos de substituição para grafos atualmente existentes são GraphLIME[34], RelEx[35] e PGMExplainer[36]. GraphLIME e RelEx explicam apenas classificações de nó, enquanto o PGMExplainer explica tanto classificação de nó quanto classificação de grafo.

1.4.5 Explicações a Nível de Modelo

Todos os métodos mencionados nas seções acima explicam as redes neurais a nível de instância, isto é, servem para justificar uma previsão específica de uma GNN. Nas explicações a nível de modelo, o objetivo é explicar o processo de tomada de decisão da rede neural de modo geral. Para esse fim, é possível extrair as máscaras mais frequentes nas explicações a nível de instância, ou usar métodos especializados em explicações a nível de modelo.

Na literatura, há apenas dois métodos de interpretabilidade de GNNs de propósito geral a nível de modelo amplamente estabelecidos: o XGNN[37], que gera grafos de input para a GNN maximizando a probabilidade predita para cada classe, e o GCFExplainer[38], que gera para cada classe um conjunto pequeno de protótipos que sejam contra-factuais de muitos exemplos da classe.

1.4.6 GNNs auto-explicáveis

Outra abordagem possível é, ao invés de explicar modelos treinados com os algoritmos mencionados anteriormente, treinar GNNs voltadas para a explicabilidade [17]. As GNNs auto-explicáveis fazem as previsões através de mecanismos interpretáveis e usam esses mecanismos para gerar explicações em conjunto com as previsões. Os modelos que seguem esse propósito atualmente são o SE-GNN[39], que faz classificação de nó a partir da comparação com nós rotulados similares, e o ProtGNN[40], uma rede neural gerativa de protótipos de classes (como a XGNN[37]), onde a classe predita para um grafo novo é a classe com os protótipos mais similares ao grafo novo. Além

disso, o GIB[41], que é treinado para acertar as previsões usando a menor quantidade de informação possível (formalizada através de funções de informação mútua) do grafo de entrada, foi criado para mitigar o problema dos ataques adversariais, mas a representação sucinta gerada a partir do "gargalo de informação"pode ser explorada pela explicabilidade.

Também podemos incluir entre as GNNs auto-explicáveis modelos baseados em causalidade (DIR[42], DisC[43], CIGA[44]). GNNs causais são treinadas para separar correlações causais de correlações espúrias nos dados, e ao fazer previsões, primeiro identificam features causais (e.g. sub-grafos com evidência favorável a uma determinada classe) e usam apenas essas features para fazer previsões. Além do ganho de desempenho e confiabilidade nas previsões, essa abordagem também favorece a interpretabilidade, que é possível por meio da análise das relações de causalidade.

1.4.7 Métricas objetivas de explicabilidade de GNNs

As principais métricas objetivas de explicabilidade de GNNs são:

- **Fidelidade:** A Fidelidade[19] avalia se a remoção de máscaras selecionadas pela explicação tem um alto impacto no resultado da rede neural. Para métodos que não geram máscaras discretas, é obtida uma máscara a partir de todas as features cuja importância seja maior que um determinado limiar. Assim, há diferentes formas de se calcular a Fidelidade:
  - $Fidelity{+}_{acc} = \frac{1}{N}\sum_{i=1}^{N}\big(1(\hat{y}_i = y_i) - 1(\hat{y}_i{}^{1-m_i} = y_i)\big)$

- $Fidelity{-}_{acc} = \frac{1}{N}\sum_{i=1}^{N}\big(1(\hat{y}_i = y_i) - 1(\hat{y_i}^{m_i} = y_i)\big)$
- $Fidelity{+}_{prob} = \frac{1}{N}\sum_{i=1}^{N}\big(f(G_i)_{y_i} - f(G_i^{1-m_i})_{y_i}\big)$
- $Fidelity{-}_{prob} = \frac{1}{N}\sum_{i=1}^{N}\big(f(G_i)_{y_i} - f(G_i^{m_i})_{y_i}\big)$

Onde $\hat{y}_i$ é a classe predita para o exemplo i na rede neural, y_i é a classe verdadeira de i, $\hat{y_i}^{m_i}$ é a previsão da rede neural para i mascarando todas as features fora da máscara m_i, $1-m_i$ é a máscara complementar a m_i, $1(\hat{y_i}^{m_i} = y_i)$ é uma função que retorna 1 se $\hat{y_i}^{m_i} = y_i$ e 0 caso contrário, e $f(G_i)_{y_i}$ é a saída de probabilidade da rede neural para o grafo G_i na classe de G_i, y_i.

- **Esparsidade:** Uma boa explicação precisa ser sucinta e interpretável. A esparsidade[19] de uma previsão pode ser dada por $1 - \frac{|m|}{|M|}$, onde $|m|$ é o número de features selecionadas pela máscara e $|M|$ é o número total de features da modalidade analisada (e.g. o número de nós do grafo). A esparsidade de um método a nível de modelo é a média das esparsidades a nível de instância nas amostras testadas.

- **Acurácia:** Em casos onde são conhecidas as features determinantes da classe de um exemplo, é possível medir a acurácia[45] das explicações em identificar essas features.

Consideramos que existe um trade-off entre Fidelidade e Esparsidade, uma vez que, quanto mais completa é uma máscara de features, mais o comportamento da máscara se aproxima do conjunto completo das features.

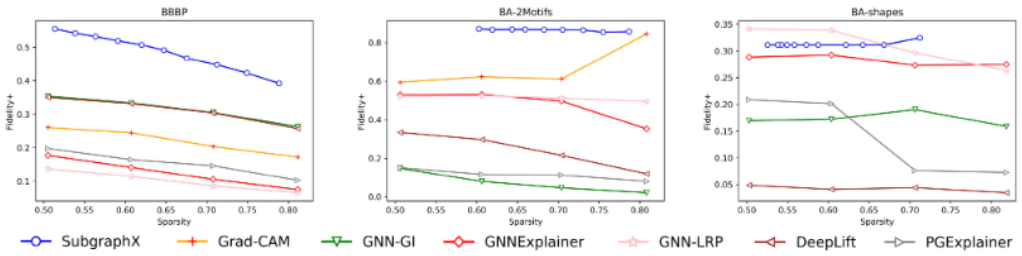

Figura 3: Gráficos comparativos da fidelidade e da esparsidade. Cada gráfico corresponde a um conjunto de dados diferente, e cada linha do gráfico corresponde a escores obtidos em um método de explicabilidade. Fonte: [19].

Assim, visando uma comparação mais justa, é comum visualizar a fidelidade da explicação em função da esparsidade, como na Figura 3.

Além da fidelidade, esparsidade e acurácia, que são as principais métricas, há alternativas:

- **Contrastividade:** Em uma explicação coerente, desejamos que evidência a favor de uma classe não seja também evidência a favor da classe oposta. Assim, para problemas de classificação binária, podemos avaliar a contrastividade[21] das explicações como $\frac{d_H(m_0,m_1)}{m_0 \vee m_1}$, onde m_0 e m_1 são respectivamente a máscara da classe negativa e da classe positiva para uma mesma previsão, $d_H(m_0, m_1)$ é a distância de Hamming (quantidade de features que estão em exatamente uma das máscaras dentre m_0 e m_1), e $m_0 \vee m_1$ é o tamanho da união entre as máscaras m_0 e m_1.

- **Estabilidade:** Métodos de interpretabilidade estáveis identificam features importantes para o modelo independentemente de variações pe-

quenas na implementação, de forma que as explicações não mudam com a ampla replicação dos resultados do modelo. Em situações onde isso é desejável, é avaliada a estabilidade[45].

- **Consistência:** Se há features específicas que determinam a classe de um dado exemplo, espera-se que modelos de alta performance sejam sempre consistentemente mais sensíveis a essas features. Isso pode ser avaliado por meio da métrica de Consistência[45].

1.5 GNNs aplicadas ao estudo do TEA

Dada a sua capacidade de extrair padrões complexos em dados conectados, as Redes Neurais de Grafos já foram empregadas na previsão do diagnóstico de TEA diversas vezes na literatura acadêmica[6], com muitas dessas usando os dados públicos de fMRI da ABIDE I[1].

1.5.1 Modelos de GNN para diagnóstico de TEA

Na literatura, há duas principais formas de representar dados de fMRI como grafos. A primeira delas é o grafo populacional, onde cada paciente é representado como um nó no grafo e arestas são inseridas entre pacientes similares, sendo a previsão do diagnóstico uma tarefa de classificação de nó. A segunda, que é o foco do estudo presente, é o grafo individual ("subject graph"). Nessa representação, o cérebro do paciente é modelado como um grafo, onde cada nó corresponde a uma região do cérebro, e arestas são inseridas entre pa-

[1] http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html

res de regiões que são altamente correlacionados na série BOLD do fMRI. Assim, as GNNs aplicadas em grafos individuais prevêem o diagnóstico de TEA por meio da tarefa de classificação de grafo. Além disso, dados fenotípicos são concatenados às representações geradas pelas GNNs para auxiliar na previsão.

[...]

1.5.2 Uso da explicabilidade para extração de biomarcadores do TEA

A interpretabilidade é bastante importante para a aplicação de redes neurais na área médica, não apenas para facilitar o uso responsável das previsões, mas também para identificar possíveis marcadores biológicos das condições investigadas, e por meio de estudos mais rigorosos, consolidá-los como conhecimento científico. No caso do diagnóstico a partir da análise de fMRI, um dos principais focos da interpretabilidade é identificar regiões do cérebro que são mais ativas em pessoas que possuem aquele diagnóstico.

No caso das GNNs, são usados algoritmos de interpretabilidade que detectem nós importantes do grafo, que corresponderiam às ROI do cérebro. São usados os algoritmos baseados em máscaras de nós, ou são criadas máscaras de nós a partir da discretização de mapas de saliência. Esses algoritmos originalmente funcionam a nível de instância, e para obter interpretações a nível de classe, são selecionadas as regiões do cérebro mais frequentes nas máscaras de nós.

Nos experimentos já registrados na literatura, foram utilizados tanto mé-

todos de uso geral para interpretabilidade de GNNs, quanto métodos *ad-hoc*, voltados apenas para a aplicação estudada.

[...]

2 Dados e Pré-Processamento

2.1 Obtenção dos dados

Os dados usados no projeto são provenientes da base de dados ABIDE I, mencionada anteriormente no texto. A base contém dados provenientes de 1112 pessoas, sendo 539 pessoas com TEA (Transtorno do Espectro Autista) e 573 pessoas sem TEA (Controle Típico). Esses dados foram obtidos de 16 diferentes pontos de coleta. Cada ponto de coleta tem um conjunto de arquivos para download, e além de dados dos equipamentos usados nos exames, há dois arquivos principais: o de dados fenotípicos e os dados dos exames.

Os dados fenotípicos, que podem estar ou no arquivo do seu respectivo ponto de coleta, ou no arquivo que tem o conjunto completo de dados fenotípicos, consistem em tabelas no formato *.csv*, onde cada linha representa um paciente e cada coluna representa diferentes características dos pacientes, como sexo, idade, presença ou não de TEA, e o diagnóstico do paciente a partir dos critérios do DSM-IV-TR[46].

Já a base de dados de exames possui dois tipos de exame: o anatômico, e o R-FMRI (Resting State Functional Magnetic Resonance Imaging), que é o dado de interesse para o treinamento das redes neurais. Os exames de cada ponto de coleta devem ser baixados separadamente, e vêm com várias pastas identificadas pelo número de identificação do paciente correspondente. Os exames estão no formato *.nii.gz*, que pode ser manipulado usando as biblioteca *nibabel* e *nilearn* da linguagem Python.

2.2 Visão Geral

2.2.1 Dados Fenotípicos

No total, há 74 colunas na tabela da base de dados fenotípicos, sendo uma delas o ponto de coleta, outra o número de identificação do paciente na base de dados (correspondente a uma pasta de mesmo número na base de dados de exames), e as demais os dados fenotípicos em si. A lista completa de atributos da tabela com suas respectivas legendas está no site da ABIDE I[2]. Dentre esses atributos, podemos destacar a presença de atributos demográficos (sexo, idade, diagnóstico do DSM-IV, IMC), e as pontuações de testes neuropsicológicos (QI, ADI-R, ADOS, SRS, SCQ, Quociente de Autismo Total, Vineland, WISC-IV), que descrevem de forma mais detalhada o perfil neuropsicológico do paciente.

2.2.2 Dados dos Exames

Um exame de fMRI da base de dados ABIDE I vem como um vídeo composto de imagens tridimensionais de resolução 64 x 28 x 28, coletadas em um total de 240 quadros ao longo do tempo.

2.3 Pré-processamento

Para construir um grafo a partir do exame de fMRI, primeiramente precisamos parcelar agrupar os sinais de cada região de interesse (ROI) do cérebro.

[2]https://fcon_1000.projects.nitrc.org/indi/abide/ABIDE_LEGEND_V1.02.pdf

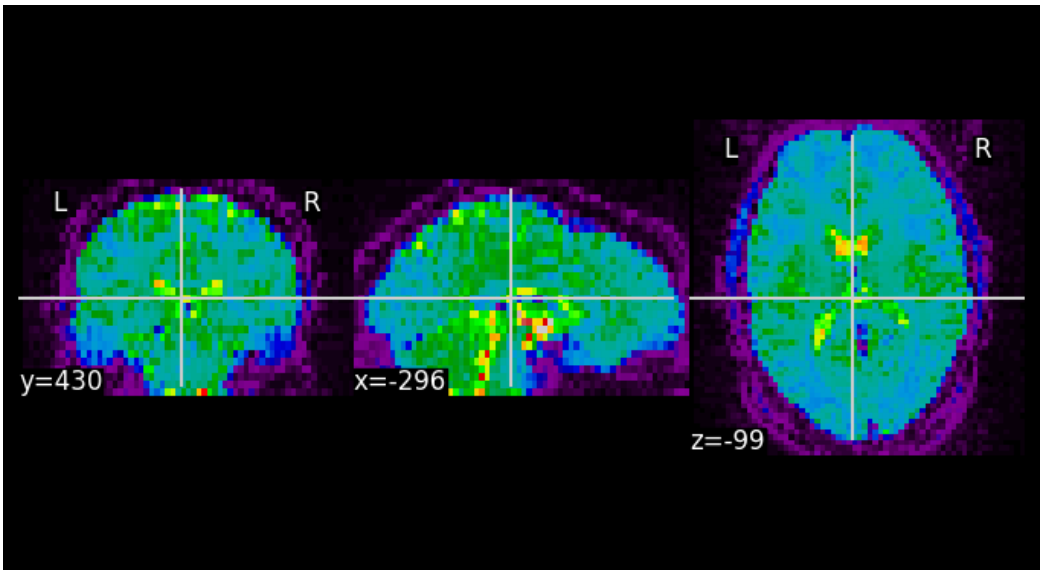

Figura 4: Exemplo de um quadro de um exame de fMRI da ABIDE I. Cada imagem mostra uma seção transversal do exame em 3D, com a legenda indicando a coordenada no sistema (x,y,z).

Isso pode ser feito usando um Atlas cerebral, que mapeia voxels de uma imagem cerebral para suas regiões correspondentes no cérebro. Usaremos o Atlas Harvard-Oxford, que é bastante usado em aplicações de Aprendizado de Máquina usando os dados da ABIDE I. Esses atlas são representados como arrays tridimensionais, onde o valor contido em (i, j, k) indica a região a ser atribuída ao voxel (i, j, k) no exame de fMRI. Cada região ROI recebe a média das séries BOLD dos voxels atribuídos a ela pelo atlas.

É possível contornar a execução manual da tarefa com o Preprocessed Connectomes Project[3]. Esse projeto providencia os dados de fMRI dos voluntários da ABIDE I pré-processados, com exatamente uma série BOLD para cada ROI, calculada do modo descrito anteriormente. Há diversas opções de atlas disponíveis no projeto.

Dada a série BOLD de cada ROI de um paciente, podemos construi o grafo do cérebro do paciente, inserindo uma aresta entre cada par de regiões altamente correlacionadas de acordo com uma determinada métrica, como a Correlação de Pearson ou o DTW. Além disso, podemos usar a série BOLD para calcular o vetor de features de cada nó no grafo. A identificação de cada ROI também determina o vetor de features, já que as GNNs clássicas possuem saída invariante à permutação dos nós, e isso não é uma propriedade desejável neste contexto.

[3]http://preprocessed-connectomes-project.org/abide/index.html

2.4 Análise Exploratória

2.4.1 Dados Fenotípicos

O conjunto de dados fenotípicos tem uma quantidade bastante expressiva de dados faltantes, mas essa ausência de dados está concentrada nas colunas dos testes neuropsicológicos, pois cada ponto de coleta escolheu o seu próprio conjunto de testes para avaliar o diagnóstico de TEA nos voluntários. As colunas sem dados faltantes são 'SITE_ID', 'SUB_ID', 'DX_GROUP', 'DSM_IV_TR', 'AGE_AT_SCAN', 'SEX' e 'EYE_STATUS_AT_SCAN'.

Analisando a demografia dos dados, vale destacar que a grande maioria dos voluntários da ABIDE I são do sexo masculino, sendo que o público feminino tem uma necessidade maior de acesso a tecnologias diagnósticas, pois suspeita-se que o autismo seja significativamente subdiagnosticado em mulheres[1].

3 Treinamento da Rede Neural

4 Aplicação de Métodos de Explicabilidade

4.1 Metodologia

As GNNs serão treinadas diversas vezes com pequenas perturbações de hiperparâmetro, de forma a posteriormente voltar o foco das explicações menos para os modelos treinados e mais para os padrões objetivos dos dados. Após o treinamento dos modelos, aplicaremos algoritmos de explicabilidade, de forma a encontrar os nós mais relevantes para a classificação de cada instância, gerando máscaras de nós importantes. Além disso, iremos observar a frequência de cada nó nas máscaras de nó de cada método, de forma a avaliar a relevância de cada ROI do cérebro a nível de modelo.

Para avaliar as explicações oferecidas por cada método de interpretabilidade, usaremos as métricas de Fidelidade, Esparsidade, e Contrastividade, já consolidadas na interpretabilidade de Redes Neurais de Grafos. Além disso, dada a distribuição de probabilidade dos nós a partir da frequências de cada nó nas máscaras de nós, iremos calcular a entropia dessa distribuição de probabilidade, de forma a avaliar a estabilidade dos métodos de explicação e a capacidade de encontrar marcadores biológicos do TEA em geral.

4.2 Resultados

4.2.1 Métricas de explicabilidade

4.2.2 Principais possíveis bio-indicadores

Referências

[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition*. Washington, D.C.: American Psychiatric Publishing, 5 ed., 2013.

[2] J. Zeidan, E. Fombonne, J. Scorah, A. Ibrahim, M. S. Durkin, S. Saxena, A. Yusuf, A. Shih, and M. Elsabbagh, "Global prevalence of autism: A systematic review update," *Autism research*, vol. 15, no. 5, pp. 778–790, 2022.

[3] Presidência da República Federativa do Brasil, "Lei nº 12.764, de 27 de dezembro de 2012," 2012. Institui a Política Nacional de Proteção dos Direitos da Pessoa com Transtorno do Espectro Autista.

[4] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, *et al.*, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, 2023.

[5] W. Liu, M. Li, and L. Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[6] S. Zhang, J. Yang, Y. Zhang, J. Zhong, W. Hu, C. Li, and J. Jiang, "The combination of a graph neural network technique and brain imaging to diagnose neurological disorders: A review and outlook," *Brain Sciences*, vol. 13, no. 10, p. 1462, 2023.

[7] 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.

[8] W. L. Hamilton, "Graph representation learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159.

[9] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[11] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[12] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[13] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[14] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, 2018.

[15] B. Weisfeiler and A. Leman, "The reduction of a graph to canonical form and the algebra which appears therein," *nti, Series*, vol. 2, no. 9, pp. 12–16, 1968.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[17] E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, and S. Wang, "A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability," *Machine Intelligence Research*, pp. 1–51, 2024.

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

[19] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 5, pp. 5782–5799, 2022.

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] W. Lin, H. Lan, and B. Li, "Generative causal explanations for graph neural networks," in *International Conference on Machine Learning*, pp. 6666–6679, PMLR, 2021.

[29] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *Advances in neural information processing systems*, vol. 30, 2017.

[30] A. Lucic, M. A. Ter Hoeve, G. Tolomei, M. De Rijke, and F. Silvestri, "Cf-gnnexplainer: Counterfactual explanations for graph neural networks," in *International Conference on Artificial Intelligence and Statistics*, pp. 4499–4511, PMLR, 2022.

[31] M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. C.-H. Lam, and Y. Zhang, "Robust counterfactual explanations on graph neural

networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5644–5655, 2021.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] Z. Huang, M. Kosan, S. Medya, S. Ranu, and A. Singh, "Global counterfactual explainer for graph neural networks," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 141–149, 2023.

[39] E. Dai and S. Wang, "Towards self-explainable graph neural network," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 302–311, 2021.

[40] Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee, "Protgnn: Towards self-explaining graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 9127–9135, 2022.

[41] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20437–20448, 2020.

[42] Y.-X. Wu, X. Wang, A. Zhang, X. He, and T.-S. Chua, "Discovering invariant rationales for graph neural networks," *arXiv preprint arXiv:2201.12872*, 2022.

[43] S. Fan, X. Wang, Y. Mo, C. Shi, and J. Tang, "Debiasing graph neural networks via learning disentangled causal substructure," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24934–24946, 2022.

[44] Y. Chen, Y. Zhang, Y. Bian, H. Yang, M. Kaili, B. Xie, T. Liu, B. Han, and J. Cheng, "Learning causally invariant representations for out-of-distribution generalization on graphs," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22131–22148, 2022.

[45] 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.

[46] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. Washington, DC: American Psychiatric Association, 4th, text revision ed., 2000.