GRADIENT REACTIVATION ENHANCED CAUSAL ATTENTION FOR OUT-OF-DISTRIBUTION GENERALIZABLE GRAPH CLASSIFICATION

Xu Wang[1+], *Pengfei Gu*[1+], *Yudong Zhang*[1], *Binwu Wang*[1*], *Pengkun Wang*[12], *Yang Wang*[12*]

[1]University of Science and Technology of China, Hefei, China
[2]Suzhou Institute for Advanced Research, USTC, Suzhou, China

ABSTRACT

Seeking for generalizable graph representations becomes hot spot in the area of graph learning. Recently, causality theory has been applied for extracting the causal relations between graph data and labels, which are generalizable under distribution shift and result in better OOD generalization. In this paper, for more accurately capturing causal representation of graph data, we propose a gradient reactivation enhanced causal subgraph extraction method. The proposed model utilizes attention mechanism to extract the causal features and attenuates the confounding effect of shortcut features. For ensuring stability of extracted causal features, we propose a novel gradient reactivation method to filter features with greater effect on making prediction. Extensively experimental result proves the effectiveness of the proposed model.

Index Terms— Graph classification, out of distribution, causal models, graph neural networks, gradient reactivation

1. INTRODUCTION

Graph structured data Neural Networks (GNNs) [1, 2, 3] have been extensively studied due to their powerful fitting ability to non-euclidean data. The progress of GNNs benefit various applications in many domains, including molecular analysis, recommendation systems and social networks [4, 5, 6].

Existing works are mostly based on in-distribution (ID) hypothesis, i.e., training and testing datasets are identically distributed. However, in real-world scenarios, there exist distribution shifts between training and testing datasets, which is against in-distribution hypothesis and attenuates the performance of existing models in such Out-Of-Distribution (OOD) test evaluation [7]. The poor generalization of ID based models results from their learning paradigm that maximizes the mutual information between extracted graph representations and corresponding labels [8]. Such paradigm makes models trapped in noncausal shortcut features [9, 10, 11], which are discriminative in training data but indiscriminative in testing data. Therefore, while existing models are well performing on training data by simply extracting the shortcut features, they fail to perform satisfying classification on testing data. Fig.1 shows a simple example of shortcut features in graph classification. When classifying graphs with subgraph *house*, models may wrongly extract *star* as it is discriminative in training set.

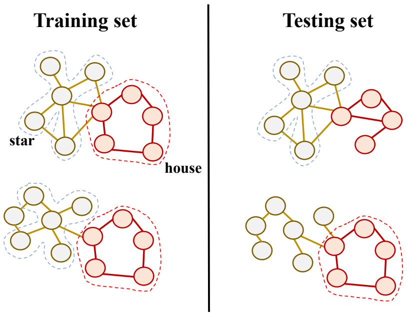

Fig. 1. An example of shortcut features. *Star* is shortcut feature which is discriminative in training set but not correlated to the classification target, i.e., having *house* or not.

Tackling this issue, great recent efforts have been made to develop generalizable models based on causal theory [12, 13, 14, 15]. The key of those methods is the strategies of extracting causal features and shortcut features. OOD-GNN [12] introduce Hilbert-Schmidt Independence Criterion and sample reweighting mechanism to eliminate the statistical dependence between graph representations, so that the model is forced to learn more generalizable graph representations. StableGNN [13] extracts high-level graph representations and resorts to the distinguishing ability of causal inference to get rid of spurious correlations. DGNN [14] follows the idea of StableGNN proposes a framework for OOD generalized node representation learning by jointly optimizing a decorrelation regularizer and a weighted GNN model. However, all these models focus on feature-wise graph representations and are unable to extract structural causal parts of graphs. Although many recent works propose novel mechanisms for extracting

+Xu Wang and Pengfei Gu are with equal contribution. *Prof. Yang Wang is the corresponding author, and Dr. Binwu Wang is the joint corresponding author. Email:angyan@ustc.edu.cn. This paper is partially supported by the National Natural Science Foundation of China (No.62072427, No.12227901), the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR-005), Academic Leaders Cultivation Program, USTC.

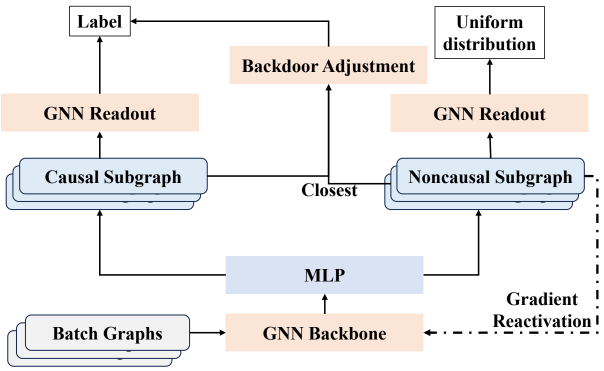

Fig. 2. Framework overview.

critical subgraphs, these works fail to take a causal look and fall short of OOD generalization.

Addressing the above issue, in this paper, we propose an attention based framework to explicitly and dynamically separate causal and noncausal subgraphs, and design a new do-calculus method for backdoor adjustment [16] so as to get rid of low generalization caused by noncausal shortcut features. A novel gradient reactivation module is proposed to ensure the reliability of the extraction of causal subgraphs. The proposed framework can be easily plugged into existing works to achieve good OOD generalization. Experimental result on several widely used datasets proves the superiority of the proposed framework. Our contribution can be summarized as,

1) We propose an attention based causal subgraph extraction module to separate causal and noncausal subgraphs, and a novel gradient reactivation module is proposed to ensure the reliability of the extraction of causal subgraphs.

2) We design a distance based method for backdoor adjustment, which enable our model to cut off the the backdoor path caused by noncausal shortcut features.

3) Extensive experiments demonstrate that the proposed method achieves SOTA prediction performance on several widely used data sets.

2. PROBLEM DEFINITION

We denote a graph as $G = \{A, X\}$. $A \in \mathbb{R}^{N\times N}$ is the adjacency matrix, where $A_{ij} \neq 0$ if there exists an edge between i-th and j-th nodes. $X \in \mathbb{R}^{N\times D}$ denotes the features of all nodes and N is the number of nodes. Let $\{y_i, G_i\}$ be the i-th sample of a given dataset, where G_i is the graph data and y_i denotes the corresponding label. The goal of graph classification is to learn a mapping function f with trainable parameters θ from graph data to corresponding labels.

3. METHOD

The proposed framework is composed with three components, i.e., attention based causal subgraph extraction module, causality determination module and gradient reactivation module. In this section, we detail each component.

3.1. Causal Subgraph Extraction

Given a graph $G = \{A, X\}$, we obtain node-level and edge-level representations of it through a GNN-based network,

$$\begin{aligned} H &= \mathrm{GNN(A,X)} \\ E &= \{e_{ij} = h_i || h_j\} \end{aligned} \tag{1}$$

where $H = \{h_i | i \in [1, N]\}$ is the output representations of all nodes generated by an arbitrary GNN backbone and E corresponds to representations of all edges. h_i and h_j correspond to representations of two adjacent nodes, $||$ is the concatenation operation. Two MLPs are applied to calculate the attention score of nodes and edges respectively,

$$\begin{aligned} \alpha_{c_i}, \alpha_{s_i} &= \sigma(MLP_{node}(node)) \\ \beta_{c_{ij}}, \beta_{s_{ij}} &= \sigma(MLP_{edge}(edge)) \end{aligned} \tag{2}$$

σ represents the softmax activation function. α_{c_i} and $\beta_{c_{ij}}$ represent the attention score of i-th node and edge between i-th and j-th nodes, which are employed to extract causal subgraph. Similarly, α_{s_i} and $\beta_{s_{ij}}$ are employed to extract noncausal subgraph. Noting that we have $\alpha_{c_i} + \alpha_{s_i} = 1$, $\beta_{c_{ij}} + \beta_{s_{ij}} = 1$. Attention scores represent how much attention the model pays to each edge and node. Therefore, mask matrices $M_x = \{\alpha_{c_i} | i \in [1, N]\}, \overline{M_x} = \{\alpha_{s_i} | i \in [1, N]\}$ at the node level and the mask matrices $M_a = \{\beta c_{ij} | A_{ij} \neq 0\}, \overline{M_a} = \{\beta s_{ij} | A_{ij} \neq 0\}$ at the edge level are obtained. According to the mask matrices, the original graph can be divided into causal subgraph and noncausal subgraph as,

$$\begin{aligned} G_c &= \{M_a \cdot A, M_x \cdot X\} \\ G_s &= \{\overline{M_a} \cdot A, \overline{M_x} \cdot X\} \end{aligned} \tag{3}$$

3.2. Causality Determination

After obtaining mask matrix through attention score and constructing causal and noncausal subgraph, for the causal subgraph G_c, in order to capture the causal features, we obtain the representation of causal subgraph through another GNN, and then make prediction through readout function and classification function as,

$$\begin{aligned} h_{G_c} &= f_{readout}(\mathrm{GNN_c(G_c)}) \\ z_{G_c} &= \Phi_c(h_{G_c}) \end{aligned} \tag{4}$$

The prediction on causal features should be as close as possible to the real label, so the classification loss used here is defined as,

$$Loss_c = -\frac{1}{|\mathcal{D}|} \sum_{G \in \mathcal{D}} y_G^T \log(z_{G_c}) \tag{5}$$

Table 1. Performance comparison on graph classification datasets.

Dataset	MUTAG	NCI1	PROTEINS	COLLAB	IMDB-B	IMDB-M
DGK	87.44±2.72	80.31±0.46	75.68±0.54	73.09±0.25	66.96±0.56	44.55±0.52
GlobalAtt	88.27±8.65	81.17±1.04	72.60±4.37	81.48±1.46	69.10±3.80	51.40±2.91
SortPool	86.17±7.53	79.00±1.68	75.48±1.62	77.84±1.22	73.00±3.50	49.53±2.29
GCN	88.20±7.33	82.97±2.34	75.65±3.24	81.72±1.64	73.89±5.74	51.53±3.28
GCN+CAL	89.24±8.72	83.48±1.94	76.28±3.65	82.08±2.40	74.40±4.55	52.13±2.96
GCN+GRECA	**91.55±6.16**	**83.97±0.68**	76.77±2.85	82.45±2.11	**74.66±3.78**	52.54±3.01
GIN	89.42±7.40	82.71±1.52	76.21±3.83	82.08±1.51	73.40±3.78	51.53±2.97
GIN+CAL	89.91±8.34	83.89±1.93	76.92±3.31	82.68±1.25	74.13±5.21	52.60±2.36
GIN+GRECA	89.94±5.18	84.14±1.22	77.04±3.51	82.79±1.11	74.44±4.77	**52.83±2.21**
GAT	88.58±7.54	82.11±1.43	75.96±3.26	81.42±1.41	72.70±4.37	50.60±3.75
GAT+CAL	89.94±8.78	83.55±1.42	76.39±3.65	82.12±1.95	73.30±4.16	50.93±3.84
GAT+GRECA	90.50±6.44	83.79±1.58	**77.13±2.18**	**82.67±1.22**	73.81±3.63	51.41±2.91

where $\mathcal{D}$ denotes the dataset, and $|\mathcal{D}|$ is the size of $\mathcal{D}$.

Meanwhile, for the noncausal subgraph, we also have,

$$\begin{aligned} h_{G_s} &= f_{readout}(\mathrm{GNN_s}(\mathrm{G_s})) \\ z_{G_s} &= \Phi_s(h_{G_s}) \end{aligned} \tag{6}$$

As we consider the noncausal subgraph to be trivial for classification, so we restrict the classification result on the noncausal subgraph to be trivial, that is, the influence of the noncausal part on the predicted result is as small as possible. Thereby, the labels we use for noncausal subgraph are uniform distribution. We define the classification loss as the KL divergence of the uniform distributed labels and the predicted result on noncausal subgraph,

$$Loss_s = \frac{1}{|\mathcal{D}|}\sum_{G\in\mathcal{D}} \mathrm{KL}(\mathrm{y_{unif}}, \mathrm{z_{G_s}}) \tag{7}$$

By optimizing the above two losses, we can distinguish between causal and noncausal features.

Additionally, for more effective extraction of causal and noncausal features, we further propose a do-calculus method to cut off the association between noncausal features and labels. Generally, we have,

$$\begin{aligned} P_m(Y|C) &= \sum_{s_j\in S} P_m(Y|C,s_j)P_m(s_j|C) \\ = \sum_{s_j\in S} P_m(Y|C,s_j)P_m(s_j) &= \sum_{s_j\in S} P(Y|C,s_j)P(s_j) \end{aligned} \tag{8}$$

where S is the set of all environments (noncausal features). This formula is called backdoor adjustment, which performs causal intervention on C by combining C with different s_j, thus cut off the association between S and labels. However, Eq.8 requires traversal of S and combine every s in S with C, which is intractable. Therefore, we propose a distance-based causal intervention method, and the distance function is defined as $\mathrm{dis}(\mathrm{z_{G_c}}, \mathrm{z_{G_s}}) = ||\mathrm{z_{G_c}} - \mathrm{z_{G_s}}||_2^2$. Given a batch of B graphs $\mathbf{G} = G^i | i \in [1, B]$, for each graph G^i, we calculate the distance between its causal subgraph with every noncausal subgraph of other graphs as,

$$d_{ij} = \mathrm{dis}(\mathrm{z_{G_c^i}}, \mathrm{z_{G_s^j}}) \tag{9}$$

Therefore, we can approximately achieve Eq.8 by combining $z_{G_s^j}$ with the smallest distance function from $z_{G_c^i}$ in the data of each batch. At this time, in the training, the difficulty of distinguishing the subject and the environment of the model can be increased, and the prediction generalization ability of the model can be improved as much as possible. By the combination, we have the following loss,

$$\begin{aligned} z_G &= \Phi(z_{G_c} + z_{G_s}) \\ Loss_{cs} &= -\frac{1}{|\mathcal{D}|}\sum_{G\in\mathcal{D}} y_G^T \log(z_G) \end{aligned} \tag{10}$$

After obtaining the three optimization objectives, we combine them in a weighted way and optimize them simultaneously,

$$Loss = Loss_c + \lambda_1 \cdot Loss_s + \lambda_2 \cdot Loss_{cs} \tag{11}$$

where λ_1 and λ_2 are hyperparameters.

3.3. Gradient reactivation

To ensure the reliability of the separation of causal and noncausal subgraphs, we further propose a gradient reactivation method to filter the noncausal subgraphs. Specifically, after obtaining the noncausal subgraph, we reactivate some nodes and edges with large gradient into the causal subgraph as they contribute a lot to making prediction. For the obtained noncausal subgraph z_{G_s}, the goal is to extract the wrongly divided causal component from it. We calculate the loss of using z_{G_s} to make prediction of real label, as,

$$Loss_t = -\frac{1}{|\mathcal{D}|}\sum_{G\in\mathcal{D}} y_G^T \log(z_{G_s}) \tag{12}$$

This cross entropy loss will not participate in back propagation, and serves only for the calculation of gradient of $\overline{M_a}$

and $\overline{M_x}$ in Eq.3. Elements in $\overline{M_a}$ and $\overline{M_x}$ with large gradients should not be included in noncausal subgraph, so we remove them from noncausal subgraph and denote the new noncausal subgraph masks with $\overline{M_a'}$ and $\overline{M_x'}$. Then we propose two more losses to ensure that elements with large gradients will not be included in noncausal subgraph,

$$\begin{aligned} Loss_{node} &= (\overline{M_x} - \overline{M_x'})^2 \\ Loss_{edge} &= (\overline{M_a} - \overline{M_a'})^2 \end{aligned} \tag{13}$$

Therefore, the final loss of our framework becomes,

$$\begin{aligned} Loss = Loss_c + \lambda_1 \cdot Loss_s + \lambda_2 \cdot Loss_{cs} + \\ \lambda_3 \cdot Loss_{node} + \lambda_4 \cdot Loss_{edge} \end{aligned} \tag{14}$$

4. EXPERIMENTS

We conduct experiments on eight datasets, including three biological datasets (MUTAG, NCI1, PROTEINS), three social datasets (COLLAB, IMDB-B, IMDB-M) [17], and two superpixel datasets (MNIST, CIFAR-10) [11].

Table 2. Performance comparison on image datasets.

Dataset	MNIST	CIFAR-10
GCN	90.49	54.68
GCN+CAL	94.58	56.21
GCN+GRECA	94.77	56.45
GIN	96.51	56.36
GIN+CAL	96.93	56.63
GIN+GRECA	**97.02**	56.81
GAT	95.53	64.22
GAT+CAL	95.91	66.16
GTA+GRECA	96.33	**66.71**

4.1. Main comparison

To evaluate the effectiveness of proposed Gradient Deactivation Enhanced Causal Attention learning (GRECA) framework, we apply diverse GNN backbones as GCN [1], GIN [18] and GAT [19]. We employ the following baselines: GCN, GIN, GAT, DGK [20], GlobalAtt [21], SortPool [22] and CAL [8]. The evaluation metric is classification accuracy, and we use 10-fold cross-validation to ensure the reliability of the results. As shown in Table.1 and Table.2, combinations of GRECA and GNN backbones result in accuracy improvement.

4.2. Ablation study

Impact of gradient reactivation. To evaluate the impact of proposed gradient reactivation module, we set λ_3 and λ_4 in Eq.14 as $\beta \in [0, 1]$, which control the weight of gradient reactivation in the final loss. As shown in Fig.3, when $\beta \leq 0.7$, the performance is better when $beta$ is bigger, which proves that gradient reactivation can truly benefit classification. And when $\beta \geq 0.7$, the performance drops when $beta$ is bigger.

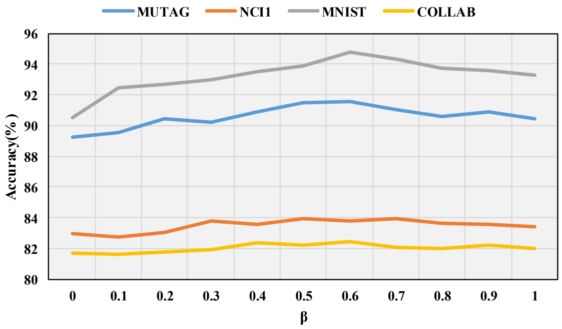

Fig. 3. Performance comparison with different values of β.

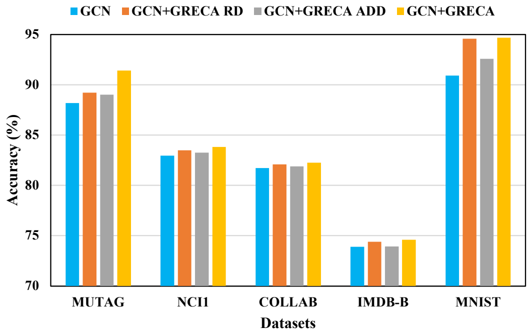

Fig. 4. Performance comparison with different backdoor adjustment strategies.

Impact of backdoor adjustment strategies. Our framework proposes a distance based backdoor adjustment, to evaluate the effectiveness of this strategy, we compare it with two more strategies, i.e., randomly concatenation (RD) and randomly adding (ADD). As shown in Fig.4, the proposed distance based backdoor adjustment strategy achieves the best performance on the tested five datasets.

5. CONCLUSION

In this paper, a gradient reactivation enhanced causal attention learning framework is proposed, which applies an attention based module to separate causal and noncausal subgraphs. To ensure the reliability of the separation, a gradient reactivation module is proposed to constrain the correlation between noncausal subgraph and the real label. A distance based backdoor adjustment strategy is proposed to learning the causal and noncausal features. Experimental result on eight widely used datasets validates the effectiveness of the proposed framework on graph classification.

6. REFERENCES

[1] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[2] Hongyang Gao and Shuiwang Ji, "Graph u-nets," in *international conference on machine learning*. PMLR, 2019, pp. 2083–2092.

[3] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson, "Benchmarking graph neural networks," *arXiv preprint arXiv:2003.00982*, 2020.

[4] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang, "Deepinf: Social influence prediction with deep learning," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 2110–2119.

[5] Lanning Wei, Huan Zhao, Quanming Yao, and Zhiqiang He, "Pooling architecture search for graph classification," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 2091–2100.

[6] Dongkwan Kim and Alice Oh, "How to find your friendly neighborhood: Graph attention design with self-supervision," *arXiv preprint arXiv:2204.04879*, 2022.

[7] Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf, "Handling distribution shifts on graphs: An invariance perspective," *arXiv preprint arXiv:2202.02466*, 2022.

[8] Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua, "Causal attention for interpretable and generalizable graph classification," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 1696–1705.

[9] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz, "Invariant risk minimization," *stat*, vol. 1050, pp. 27, 2020.

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[11] Boris Knyazev, Graham W Taylor, and Mohamed Amer, "Understanding attention and generalization in graph neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[12] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu, "Ood-gnn: Out-of-distribution generalized graph neural network," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[13] Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, and Bai Wang, "Generalizing graph neural networks on out-of-distribution graphs," *arXiv preprint arXiv:2111.10657*, 2021.

[14] Shaohua Fan, Xiao Wang, Chuan Shi, Kun Kuang, Nian Liu, and Bai Wang, "Debiased graph neural networks with agnostic label selection bias," *IEEE transactions on neural networks and learning systems*, 2022.

[15] Ying-Xin Wu, Xiang Wang, An Zhang, Xia Hu, Fuli Feng, Xiangnan He, and Tat-Seng Chua, "Deconfounding to explanation evaluation in graph neural networks," *arXiv preprint arXiv:2201.08802*, 2022.

[16] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell, *Causal inference in statistics: A primer*, John Wiley & Sons, 2016.

[17] Christopher Morris, Nils M Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann, "Tudataset: A collection of benchmark datasets for learning with graphs," *arXiv preprint arXiv:2007.08663*, 2020.

[18] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka, "How powerful are graph neural networks?," *arXiv preprint arXiv:1810.00826*, 2018.

[19] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al., "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[20] Pinar Yanardag and SVN Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1365–1374.

[21] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel, "Gated graph sequence neural networks," *arXiv preprint arXiv:1511.05493*, 2015.

[22] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen, "An end-to-end deep learning architecture for graph classification," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.