Robust Causal Graph Representation Learning against Confounding Effects

Hang Gao\textsuperscript{1,2*}, Jiangmeng Li\textsuperscript{1,2†}, Wenwen Qiang\textsuperscript{1,2}, Lingyu Si\textsuperscript{1,2}, Bing Xu\textsuperscript{3}, Changwen Zheng\textsuperscript{1}, Fuchun Sun\textsuperscript{4}

\textsuperscript{1}Science and Technology on Integrated Information System Laboratory, Institute of Software Chinese Academy of Sciences
\textsuperscript{2}University of Chinese Academy of Sciences
\textsuperscript{3}China Communications Technology Information Group Co., Ltd.
\textsuperscript{4}Tsinghua University
\{gaohang, jiangmeng2019, qiangwenwen, lingyu, changwen\}@iscas.ac.cn, xubing@cccc Ltd.cn, fcsun@mail.tsinghua.edu.cn

Abstract

The prevailing graph neural network models have achieved significant progress in graph representation learning. However, in this paper, we uncover an ever-overlooked phenomenon: the pre-trained graph representation learning model tested with full graphs underperforms the model tested with well-pruned graphs. This observation reveals that there exist confounders in graphs, which may interfere with the model learning semantic information, and current graph representation learning methods have not eliminated their influence. To tackle this issue, we propose Robust Causal Graph Representation Learning (RCGRL) to learn robust graph representations against confounding effects. RCGRL introduces an active approach to generate instrumental variables under unconditional moment restrictions, which empowers the graph representation learning model to eliminate confounders, thereby capturing discriminative information that is causally related to downstream predictions. We offer theorems and proofs to guarantee the theoretical effectiveness of the proposed approach. Empirically, we conduct extensive experiments on a synthetic dataset and multiple benchmark datasets. Experimental results demonstrate the effectiveness and generalization ability of RCGRL. Our codes are available at https://github.com/hang53/RCGRL.

Introduction

“Graph” is a derived discrete data structure consisting of vertices and edges, which can be leveraged to model and solve various general problems. Benefitting from the embedding of human knowledge, graphs are semantically dense data. In contrast, native data structures, such as images and videos, are usually semantically sparse data. Therefore, effectively using graphs to model and learn valuable information for downstream tasks is a compelling area of research. The impressive success in Graph Neural Networks (GNNs) (Kipf and Welling 2016; Xu et al. 2018; Veličković et al. 2017) provokes the exploration to sufficiently learn discriminative representations from graphs.

\*These authors contributed equally.
†Corresponding author.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
In this work, we name such substructures as graph confounders. The empirical results demonstrate that compared with the original graphs, the pre-trained GNN yields performance boosts on graphs pruned by the well-selected scheme. Such results prove the existence of the proposed graph confounders, and such confounders can indeed interfere with the downstream predictions. Note that the pruning requires the label information of the test data. Thus, this motivating experiment can only be introduced to prove the existence of graph confounders. From the experiments in Figure 2, we observe that our method, learning robust graph representation against confounding effects, generally outperforms benchmark methods that learn semantic information from complete graphs (containing confounders) on both ID and OOD tasks by significant margins. These empirical results demonstrate that graph confounders may degenerate the prediction performance and generalization ability of GNNs. The reason behind such observations is that graph confounders impede the model’s acquisition of discriminative information that is causally related to downstream predictions. Therefore, the prediction performance of learned representations degenerates. Furthermore, due to the confounding effects, the learned representations may contain data-biased information so that the generalization ability of the learned representations is degraded.

To learn graph representations against confounding effects, we propose a novel approach, called Robust Causal Graph Representation Learning (RCGR), which aims to obtain the graph representations that is causally related to the prediction on downstream tasks. To this end, RCGR adopts the concept of the Instrumental Variable (IV) approach (Wooldridge 2002). IV approach defines the conditional moment restrictions under which the confounding effect could be eliminated. Accordingly, the approach searches for suitable IVs so that the model can satisfy the mentioned restrictions. However, due to the complicity of graph representation learning models, the conditional moment restrictions are hard to satisfy, so the corresponding IVs are difficult to obtain. Therefore, our implementation of the IV approach contains a novel method that will actively generate the expected IVs, thereby transferring the conditional moment restrictions into unconditional ones. We provide theoretical proof to demonstrate that our implementation achieves the same effect as the IV approach under conditional moment constraints in Theorem 1 and the corresponding proof for Theorem 1. Theorem 1 also ensures the objective of such a GNN leading the network to generate desired IVs. Guided by the theorem, we design two novel training objectives, i.e., the robustness-emphasizing loss and the contrast-auxiliary loss, to empower the model to learn robust graph representations against confounding effects. The graph representation learning network and the IV generating network are updated alternatively until convergence. Empirically, we evaluate our proposed RCGR on one synthetic and three real datasets. Extensive experiments demonstrate that the classification performance on ID tasks and the generalization ability on OOD tasks of RCGR surpass state-of-the-art graph representation learning methods. The following are our contributions:

- We propose a novel method, RCGR, to learn robust graph representations against confounding effects, thereby improving the prediction performance and generalization ability of the representations. Such an approach is suitable for general deep graph neural network models.
- Theoretically, we offer proof to demonstrate the performance guarantee of implementing the IV approach under unconditional moment constraints and a theorem to prove the effectiveness of the generated IVs.
- We provide the implementations of RCGR for prediction and generalization tasks. The results demonstrate the consistently excellent performance of RCGR on a synthetic dataset and three benchmark datasets.
Related Work

Graph Neural Networks

GNN learns the representations of graphs by combining neural networks with graph learning. Such representations can be applied to a variety of downstream tasks. Like other neural network structures, GNNs developed multiple variants. Graph Convolution Network (GCN) (Kipf and Welling 2016) conducts graph learning utilizing the convolutional network. (Xu et al. 2018) propose Graph Isomorphism Network (GIN) which possesses the same representation capability as the Weisfeiler-Lehman test. Graph Attention Network (GAT) (Veličković et al. 2017) adopts the concept of attention mechanism and enables the network to focus on certain parts of the graph data. To improve the pooling process of GNNs, Graph-u nets (Gao and Ji 2019b) propose a novel GNN with an adaptive pooling mechanism.

Graph Causality Learning

Causality learning identifies causal relationships between data and labels. In economics and statistics, means of identifying causality have been studied for a long time, and various methods have been proposed (Glymour, Pearl, and Jewell 2016; Wooldridge 2002; Li, Ren, and Li 2014). Recently, the methods of causality learning have also been adopted in deep learning to optimize the rationality and robustness of the model. In graph learning, studying the causality within models is also beneficial. (Ying et al. 2019; Luo et al. 2020; Yuan et al. 2021) propose to adopt an explanation method to figure out the causal relationship between the model’s inputs and outputs. (Lee, Lee, and Kang 2019; Veličković et al. 2017) create intrinsic interpretable learning frameworks by incorporating rationalization modules, including graph attention and pooling. (Wu et al. 2022) introduces the intervention operation from the causal inference theory (Glymour, Pearl, and Jewell 2016) and builds causal models through finding generalization causally related information. Our method, on the other hand, improves the causality by learning robust graph representations against confounding effects.

Method

Preliminary

First, we briefly review some necessary preliminary concepts and notations.

IV Approach. IV approach (Wooldridge 2002) deals with the confounding effects by conducting estimations under conditional moment restrictions. The general type of an IV model can be described as follows. Let $X_i$, $Y_i$ denote observable variable vectors of some data distribution to be studied. $X_i \in \mathbb{R}^m$, $Y_i \in \mathbb{R}^b$. Assume that the observations \{(X_i, Y_i)\}_{i=1}^n are i.i.d., our goal is to learn a structural function $f(\cdot)$ that can represent the correlations between $X_i$ and $Y_i$. We presume that $f(\cdot)$ can be completely determined by parameters $\theta^*$. Such a function can be formulated as:

$$Y_i = f^\theta(X_i).$$  \hspace{1cm} (1)

Then, we can acquire $\theta^*$ using regression methods. Unfortunately, in real-world circumstances, confounders generally exist in the data. Formally, we have:

$$Y_i = f^\theta(X_i) + \epsilon_i,$$  \hspace{1cm} (2)

where $\epsilon_i$ is the error term. If we cannot acquire a clear representation of $\epsilon_i$, we will not be able to get $\theta^*$ correctly according to $Y_i$ and $X_i$.

The IV approach addresses this issue with conditional moment restrictions. Such an approach selects a conditioning variables vector $Z_i$ as the IV. $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ are i.i.d., $Z_i$ satisfies:

$$\mathbb{E}[\epsilon_i|Z_i] = 0 \ \forall i \in [1, n].$$  \hspace{1cm} (3)

By introducing $Z_i$ into the calculation of $f(\cdot)$, we can eliminate the confounding effects brought by $\epsilon_i$. Based on this idea, (Ai and Chen 2003) defines a formal training objective of the IV approach as:

$$\mathbb{E}[f^\theta(X_i) - Y_i|Z_i] = 0.$$  \hspace{1cm} (4)

However, the IV, $Z_i$, is also hard to acquire in most learning tasks. In linear cases, approaches, such as (Hayashi 2000; Hansen, Heaton, and Yaron 1996), address the issue by transferring the conditional moment restrictions into unconditional ones.

Graph Representation Learning. Given an attributed graph $G = (V, E)$, $V$ and $E$ denote the node and edge sets, respectively, $G = \{G_i, i \in [1, N_G]\}$ is a graph dataset. The objective of graph representation learning is to learn an encoder $f_G(\cdot) : G \rightarrow \mathbb{R}$, where $\mathbb{R}$ denotes an embedding space. Accordingly, $f_G(G_i)$ is the representation containing discriminative information of $G_i$ for downstream tasks.

To learn the discriminative representation, benchmark methods employ GNN as the encoder. GNN encodes nodes in $G_i = (V_i, E_i)$ into embedded vectors. The $k$-th layer of GNN can be formulated as:

$$h_v^{(k+1)} = \text{COM}^{(k)}(h_v^{(k)}, \text{AGG}^{(k)}(h_u^{(k)}), \forall u \in N(v)),$$  \hspace{1cm} (5)

where $h_v$ denotes the vector for node $v \in V_i$, $N(v)$ denotes the neighbors of $v$, and $h^{(k)}$ is the representation vector of the corresponding node at the $k$-th layer. Specifically, $h^{(0)}$ is initialized with the input node features. COM(·) and AGG(·) are learnable functions of GNN. COM(·) combines the aggregated neighbor feature into the feature of the target node, and AGG(·) aggregates the features of neighbors. To obtain the graph representation $f^G(G_i)$ for $G_i$, we adopt

$$f^G(G_i) = \text{READOUT}(h_v, v \in V_i),$$  \hspace{1cm} (6)

where READOUT(·) is a readout function pooling the node representations $\{h_u|v \in V_i\}$.

Robust Causal Graph Representation Learning

Theoretical Design of RCGRL. Our method aims at enhancing causality in graph representation learning. We propose that models generating predictions with high causal relationships with labels can be treated as “confounder-robust” models. We define a confounder-robust GNN as follows.
**Definition 1** (Confounder-robust GNN). Given a graph dataset \( G \), if a GNN \( f(\cdot) \) outputs information that contains the sufficient and correct information of \( G \) to perform prediction in downstream tasks, regardless of confounder \( C \), then \( f(\cdot) \) is a confounder-robust model in \( G \).

We adopt the IV approach (Greene 2003) to empower us to acquire a confounder-robust model. Given a graph dataset \( G = \{G_i, i \in [1, N_G] \} \), we define the confounder in each graph as \( C_i \). Meanwhile, \( X_i \) denotes the information causally related to \( Y_i \). However, \( X_i \) and \( C_i \) are mixed in \( G_i \). Our purpose is to train a GNN model to predict the label \( Y_i \) based on \( X_i \), regardless of \( C_i \). We denote the parameter of the model that can achieve such a goal as \( \theta^* \). With the IV approach, our objective is to acquire \( \theta^* \), which can be formulated as follows:

\[
\theta^* = \arg \min_{\theta} \left( \mathbb{E}[m(G_i, Y_i, Z_i, \theta)] | Z_i \right), \quad \forall i \in [1, n],
\]

s.t. \( \mathbb{E}[|C_i|] = 0, \quad \forall i \in [1, n] \). \tag{7}

where \( m(\cdot) \) outputs the difference between the model’s output and ground truth labels, \( |C_i| \) denotes the number of elements in \( C_i \), \( Z_i \) denote the IVs that instruct the model to eliminate confounders \( C_i \) through certain measures. Equation 7 can be treated as a generalization of Equation 4 in the field of graph representation learning.

As mentioned above, the IV \( Z_i \) can be challenging to acquire. Furthermore, the complexity of graph representation learning with neural networks enlarges such difficulty. To make calculating \( Z_i \) operable, we transfer the conditional moment restrictions of Equation 7 into unconditional moment restrictions by generating the IV \( Z_i \) with a GNN model \( q^G(\cdot) \), which can be formulated as:

\[
Z_i = q^G_i(G_i),
\]

where \( \varphi \) are the parameters of GNNs that generates \( Z_i \). To replace \( Z_i \), an ideal \( q^G(\cdot) \) with parameters \( \varphi^* \) should satisfy:

\[
\mathbb{E}[|C_i|] q^G_i(G_i) = 0, \quad \forall i \in [1, n].
\]

As we expect to train the model to predict the label \( Y_i \) based on \( X_i \), we need to leave \( X_i \) unaffected. Therefore, \( q^G(\cdot) \) should also satisfy:

\[
\mathbb{E}[X_i | q^G_i(G_i)] = X_i, \quad \forall i \in [1, n].
\]

As \( q^G(\cdot) \) is a GNN, we can utilize the backpropagation to train it. We propose the following theorem to define the objective for training \( q^G(\cdot) \).

**Theorem 1** Given a large enough graph dataset \( G \), where each graph \( G_i \in G, i \in [1, n] \). \( G \) consists of two parts, \( C_i \) and \( X_i \). \( C_i \) is the confounder. \( X_i \) is causally related to the label \( Y_i \). We define a model consisting of two GNNs, \( f^G(\cdot) \) and \( q^G(\cdot) \). \( q^G(\cdot) \) outputs IVs that are used to remove the confounders in \( G_i \). \( f^G(\cdot) \) is a GNN that predicts \( Y_i \). Assuming \( \theta^* \) is sufficiently learned so that \( f^G_{\theta^*}(X_i) = Y_i \) and \( f^G_{\theta^*}(X_i \cup C_i^\prime) \neq Y_i, \forall C_i^\prime \subseteq C_i \). Then, we can acquire the parameters \( \varphi^* \) that enable the \( q^G(\cdot) \) to satisfy Equations 9 and 10 by maximizing the cross entropy between \( Y_i \) and the model’s output. Formally, we have:

\[
\varphi^* = \arg \max_{\varphi} \left( I \left( f^G_{\theta^*} \left( r \left( G, q^G_i(G) \right) \right), Y \right) \right), \tag{11}
\]

where \( I(\cdot) \) denotes the mutual information, \( r(\cdot) \) represents a function without trainable parameters to remove confounders based on IVs.

**Proof.** We provide proofs to demonstrate the effectiveness of Theorem 1. Since we already know the values of \( \theta^* \), therefore, \( \varphi^* \) is the only parameter variable in Equation 11. We can give the proof of Theorem 1 by proving that equations 9 and 10 hold if and only if cross entropy \( I \left( f^G_{\theta^*} \left( r \left( G, q^G_i(G) \right) \right), Y \right) \) reaches the maximum value.

First, we prove the sufficiency. According to the definition of mutual information, we have:

\[
I \left( f^G_{\theta^*} \left( r \left( G, q^G_i(G) \right) \right), Y \right) = H(Y) - H \left( Y | f^G_{\theta^*} \left( r \left( G, q^G_i(G) \right) \right) \right),
\]

where \( H(\cdot) \) is defined by the distribution of the dataset, therefore \( H(Y) \) has a fixed value. Thus, if
As I have proved Theorem 1.

that Equations 9 and 10 holds if and only if cross entropy
Y

As into Equation 12, we get:
X

X

X

X

X

And, \( f_{\theta^*}^G (X_i) \neq Y_i, \forall C_i \subseteq C_i \), therefore:

As \( r(\cdot) \) is a non-parameter function that removes graph data according to \( q(\cdot) \), according to Equation 18, the confounder \( C_i \) has been totally removed and the information of \( X_i \) is unaffected. Therefore, the conditions defined in Equations 9 and 10 are met. Then, we can get that if \( I (f_{\theta^*}^G (r(G,q_{\phi^*}^G(G))), Y) \) is maximized, then Equations 9 and 10 holds, the sufficiency is proved.

Next, we will prove the necessity. If Equations 9 and 10 hold, then \( r(\cdot) \) will be able to remove \( C_i \) without affecting \( X_i \), therefore Equation 18 holds. Substituting Equation 18 into Equation 12, we get:

As \( X_i = f_{\theta^*}^G (X_i) \), we have:

Because \( H(Y|Y) = 0 \), therefore \( I (f_{\theta^*}^G (r(G,q_{\phi^*}^G(G))), Y) \) is maximized. We can then come up with the conclusion that if Equation 9 and Equation 10 holds, then \( I (f_{\theta^*}^G (r(G,q_{\phi^*}^G(G))), Y) \) is maximized.

With Equation 7 and Theorem 1, we define our IV approach-based training objective with unconditional moment restrictions as:

\[
\theta^* = \arg \min_{\theta} \left( E[m(G_i,Y_i,q_{\phi^*}^G(G_i),\theta)] \right),
\]

where \( \varphi^* \) can be calculated with Equation 11. However, at the beginning, we can neither acquire \( \theta^* \) nor \( \varphi^* \). Furthermore, the assumptions of Theorem 1 cannot be strictly met in real-world scenarios. To address such issues, we initialize \( \theta \) and \( \varphi \) randomly and update them alternatively. We further introduce some new designs to ensure that the model can learn label-causal-related graph representations against confounding effects. These designs will be discussed in the implementation and optimization parts.

**Implementation.** We implement the framework to reach the training objectives defined in Equations 11 and 21. See Figure 3 for the illustration. We build a GNN \( f^G(\cdot) \) as our encoder, and a GNN \( q^G(\cdot) \), which outputs the IVs as a series of edge weights. These edge weights are used to remove a certain amount of graph data, including dropping certain edges when the corresponding edge weights are too low and reducing information transfer through edges according to the edge weights.

Meanwhile, some nodes and edges may contain both causal and confounding information. Certain critical information may be lost if they are removed before went through the GNN encoder. Therefore, we perform graph data removal after the \( u \)-th layer of \( f^G(\cdot) \), \( u \) is a hyperparameter. In this way, we can perform confounder elimination at a more fine-grained level.

To make \( f^G(\cdot) \) further robust to confounders, we introduce the concept of contrastive learning into our network. We enforce the encoder without the confounder removed to output similar features as the original ones (with the confounder removed), thereby bootstrapping the encoder’s ability to eliminate the confounding effects of some confounders.

**Optimization.** For optimization, we perform the updates alternatively, i.e., fix \( \theta \) to learn \( \varphi \), and then fix \( \varphi \) to learn \( \theta \), and so on. We update the parameter \( \varphi \) of \( q^G(\cdot) \) according to Theorem 1, and maximize \( I (f_{\theta^*}^G (r(G,q_{\phi^*}^G(G))), Y) \) in Equation 11 by minimizing the cross entropy loss based on the label and the model output.

Our confounder elimination operation should be beneficial to improve the causality of model predictions. Therefore, in order to get a more ideal \( \varphi \), we need to further optimize the training of \( q^G(\cdot) \) from the causality aspect. According to Definition 1, a confounder-robust GNN model is required to make correct predictions regardless of the confounders. Based on this, the model should learn the same representations for the same and correct category. For the graph sample \( G_i \) of class \( l \), suppose the model learns a representation vector \( h_{i_l} \), and \( h_{i_l} \) enables the model to perform the correct prediction. Then, the optimal value of \( h_{i_l} \) is to be the same as other representation vectors within the same class \( l \). On the other hand, if \( h_{i_l} \) fails to enable the model to predict correctly, \( h_{i_l} \) is restricted to be different from all
other classification feature vectors within class $l$. According to (Sun et al. 2020), the farther the output deviates from the optimal value, the greater the contribution to training. Inspired by such intuition, during the optimization of $\hat{q}^G(\cdot)$, we want to emphasize those samples that contribute more to training in terms of causality.

Specifically, for $h_i$ belongs to class $l$, we first obtain a feature vector $t_i$, which is the average of all feature vectors within the class $l$. Then, we acquire a series of weights $w_i$ for samples to perform the emphasizing operation, which is formulated by:

$$w_i = \left\{ \begin{array}{ll} s(h_i, t_i) - O_{\text{max}}, & \text{correct result}, \\ s(h_i, t_i) - O_{\text{min}}, & \text{wrong result}, \end{array} \right. \quad (22)$$

where $s(\cdot)$ stands for the similarity calculation function, $O_{\text{min}}$ and $O_{\text{max}}$ denote the minimal and maximal values of similarity, respectively.

We can then define the loss for learning $\varphi$:

$$L_w = \frac{1}{n} \sum_{i=1}^{n} w_i \mathcal{H} \left( f^G_{\hat{\theta}} \left( r(G_i, q^G_{\hat{\varphi}}(G_i)) \right), Y_i \right), \quad (23)$$

where $\mathcal{H}(\cdot)$ calculates the cross entropy loss for the $i$-th sample. $n$ is the amount of samples. $\hat{\theta}$ denotes the fixed parameters of $f^G(\cdot)$. $\gamma$ is a hyperparameter that controls the effect of $w_{\theta, i}$.

$L_w$ makes the model output more consistent representations for the same class. We further hope that the representations of all the samples be discriminative and that the model can distinguish the samples of different classes. Therefore, we design a regular term $\mathcal{M}$:

$$\mathcal{M} = \tau \frac{1}{n} \sum_{i=1}^{n} |s(h_i, \bar{t}) - O_{\text{max}}|, \quad (24)$$

where $\bar{t}$ denotes the average of all feature vectors of all classes. $\tau$ is a hyperparameter. $\mathcal{M}$ prevents the model from outputting the same features for all classes, or, falling into a local optimum as parts of the graph data are removed according to the proposed 4-based approach.

Leveraging Equations 23 and 24, we propose the robustness-emphasizing loss for learning $\varphi$:

$$L_r = L_w + \zeta \mathcal{M}, \quad (25)$$

where $\zeta$ can be defined as follows:

$$\zeta = \left\{ \begin{array}{ll} -1, & \mathcal{M} \leq \sigma \\ 0, & \mathcal{M} > \sigma \end{array} \right. \quad (26)$$

$\sigma$ is the hyperparameter that determines when the regular term $\mathcal{M}$ should be activated.

For the optimization of $\hat{\theta}$, our objective is defined in Equation 21. As an implementation for $\eta(\cdot)$ in Equation 21, we adopt the cross entropy loss to judge the difference between model output and ground truth labels. Formally, we define the loss for learning $\theta$ as:

$$L_o = \frac{1}{n} \sum_{i=1}^{n} \mathcal{H} \left( f^G_{\hat{\theta}} \left( r(G_i, q^G_{\hat{\varphi}}(G_i)) \right), Y_i \right), \quad (27)$$

where $\hat{\varphi}$ denotes the fixed parameters of $q^G(\cdot)$.

Moreover, we build a contrastive structure to make $f^G(\cdot)$ further robust to confounders. For the training objective, we design a contrastive loss as follows:

$$L_a = -\frac{1}{n} \sum_{i=1}^{n} s \left( f^G_{\hat{\theta}} \left( r(G_i, q^G_{\hat{\varphi}}(G_i)) \right), f^G_{\hat{\theta}}(G_i) \right), \quad (28)$$

where $s(\cdot)$ denotes the similarity calculation function, $\hat{\varphi}$ denotes the fixed parameters of $q^G(\cdot)$, and $f^G(\cdot)$ denotes the encoder without the confounder removal module.

Note that $f^G(\cdot)$ and $f^G(\cdot)$ share the same GNN structure and parameters, but the parameters of $f^G(\cdot)$ are fixed as $\hat{\theta}$, i.e., $f^G(\cdot)$ are excluded from back propagation of $L_a$. Therefore, $f^G(\cdot)$ that fed by the input data containing confounders is enforced to generate similar outputs as $f^G(\cdot)$, which fed by the graphs without confounders (by using the confounder removal operation). Then, the learned parameters of $f^G(\cdot)$ will be used to update $f^G(\cdot)$. This training paradigm encourages $f^G(\cdot)$ to ignore information with confounding effects, result in certain confounders no longer have confounding effects on our model. Because we fix the outputs of a view in the contrastive learning part, we could avoid the risk of generating trivial solutions. Therefore, the contrastive learning in our method can be performed without negative samples, which significantly simplifies the computation during training.

We combine Equations 27 and 28 to introduce the contrast-auxiliary loss as the training objective of $\hat{\theta}$:

$$L_c = L_o + \lambda L_a, \quad (29)$$

where $\lambda$, as a hyperparameter, controls the effect of $L_a$.

**Experiments**

**Comparison with State-of-the-art Methods**

**Datasets.** We evaluate our method on both OOD and ID datasets. The OOD datasets include: 1) Spurious-Motif, a synthetic dataset created by (Ying et al. 2019), and we adopt the re-implementation version created by (Wu et al. 2022); 2) Graph-SST2(OOD) is an OOD version of Graph-SST2 (Yuan et al. 2020) created by (Wu et al. 2022). The ID datasets include: Graph-SST2 (ID) (Yuan et al. 2020), Graph-Twitter (Yuan et al. 2020), Mol-BBBP, and Mol-BACE (Hu et al. 2020). Different GNNs are picked for different datasets. Please refer to Appendix for details.

**Experiment Settings.** We compare our method with Empirical Risk Minimization (ERM) and various causality-enhanced methods, including the interpretable baselines, i.e., GAT and Top-k Pool, and the robust learning baselines, i.e., Group DRO, IRM, and DIR.

For a fair comparison, we follow the experimental principle of (Wu et al. 2022) and adopt the same training setting...
Table 1: Performance of classification accuracy in OOD datasets, including Spurious-Motif and Graph-SST2. The Graph-SST2 dataset is artificially added with bias. We highlight the best records in bold.

| Method                  | Graph-SST2 | Graph-Twitter | Mol-BBBP | Mol-BACE |
|-------------------------|------------|---------------|----------|----------|
|                         | Balanced   | bias = 0.5    | bias = 0.7 | bias = 0.9 |
| ERM                     | 42.99 ± 1.93 | 39.69 ± 1.73  | 38.93 ± 1.74 | 33.61 ± 1.02 |
| GAT (Veličković et al. 2017) | 43.07 ± 2.55 | 39.42 ± 1.50  | 37.41 ± 0.86 | 33.46 ± 0.43 |
| Top-k Pool (Gao and Ji 2019b) | 43.43 ± 8.79 | 41.21 ± 7.05  | 40.27 ± 7.12 | 33.60 ± 0.91 |
| Group DRO (Sagawa et al. 2019) | 41.51 ± 1.11 | 39.38 ± 0.93  | 39.32 ± 2.23 | 33.90 ± 0.52 |
| IRM (Arjovsky et al. 2019) | 42.26 ± 2.69 | 41.30 ± 1.28  | 40.16 ± 1.74 | 35.12 ± 2.71 |
| V-REx (Krueger et al. 2021) | 42.83 ± 1.59 | 39.43 ± 2.69  | 39.08 ± 1.56 | 34.81 ± 2.04 |
| DIR (Wu et al. 2022)     | 42.53 ± 3.38 | 41.45 ± 2.12  | 41.03 ± 1.53 | 39.20 ± 1.94 |
| RCGRL-Wabl              | 43.49 ± 2.06 | 42.97 ± 1.96  | 42.06 ± 1.53 | 40.83 ± 1.89 |
| RCGRL-Labl              | 43.84 ± 1.98 | 43.41 ± 1.49  | 41.94 ± 1.39 | 40.63 ± 1.52 |
| RCGRL                   | **44.03 ± 1.98** | **43.89 ± 1.51** | **42.13 ± 1.64** | **41.58 ± 1.12** |

Table 2: Performance in ID datasets, including classification accuracy in Graph-SST2 and Graph-Twitter, and ROC-AUC in Mol-BBBP and Mol-BACE. The best records are highlighted in bold.

| Method                  | Graph-SST2 (ID) | Graph-Twitter | Mol-BBBP | Mol-BACE |
|-------------------------|-----------------|---------------|----------|----------|
|                         | Balanced        |               |          |          |
| ERM                     | 89.19 ± 0.87    | 65.29 ± 1.05  | 64.19 ± 0.80 | 73.61 ± 0.75 |
| GAT (Veličković et al. 2017) | 89.89 ± 0.68    | 65.17 ± 1.21  | 65.63 ± 0.74 | 72.91 ± 0.93 |
| Top-k Pool (Gao and Ji 2019b) | 89.31 ± 1.26    | 63.78 ± 0.97  | 64.69 ± 1.41 | 71.30 ± 0.84 |
| Group DRO (Sagawa et al. 2019) | 89.94 ± 1.60    | 63.35 ± 0.52  | 64.35 ± 1.26 | 70.38 ± 1.47 |
| IRM (Arjovsky et al. 2019) | 89.55 ± 1.03    | 63.02 ± 0.77  | 64.03 ± 1.08 | 71.52 ± 0.98 |
| V-REx (Krueger et al. 2021) | 88.78 ± 0.82    | 64.81 ± 1.31  | 63.38 ± 1.06 | 71.85 ± 0.89 |
| DIR (Wu et al. 2022)     | 89.91 ± 0.86    | 65.14 ± 0.96  | 65.36 ± 1.14 | 72.30 ± 1.39 |
| RCGRL-Wabl              | 89.95 ± 0.47    | 66.36 ± 0.38  | 66.02 ± 0.99 | 74.03 ± 0.61 |
| RCGRL-Labl              | 90.38 ± 0.75    | **66.73 ± 0.49** | 65.69 ± 1.17 | 73.31 ± 0.83 |
| RCGRL                   | **90.73 ± 0.40** | 66.58 ± 0.53  | **66.78 ± 0.95** | **74.37 ± 0.53** |

Results. The results are reported in Table 1 and 2. We observe from the tables and find that our method outperforms all baselines on all downstream tasks, which demonstrates that RCGRL can learn discriminative and robust graph representations against confounding effects. Such representations have the desired performance on both ID and OOD problems, i.e., RCGRL has outstanding prediction performance and generalization ability. For the ablation studies, RCGRL outperforms RCGRL-Wabl and RCGRL-Labl on most tasks, which verifies the effectiveness of our proposed weights \( w \) and auxiliary contrastive learning.

Evaluation of IV Introducing Position
We introduce the IVs in the \( u \)-th layer instead of the beginning. Here, we provide an empirical analysis to prove the effectiveness of such a design. We conduct experiments on multiple datasets, including Spurious-Motif, Graph-Twitter, and Mol-BACE, with different \( u \). As Figure 4 shows, RCGRL reaches the best performance when \( u = 2 \) across all tasks. The reason behind this phenomenon is that introducing IV in the \( u \)-th layer helps the model to remove the confounders at the feature level, thereby avoiding excessive loss of semantic information of the original graphs.

Evaluation of Confounder Robustness
We conduct experiments on Mol-BACE and Mol-BBBP to demonstrate the confounder robustness of RCGRL. We follow (Lin, Lan, and Li 2021) to measure the causality between the elements of the graph data and the ground-truth label. The elements that have no causal relationship with the label and may even affect the model’s performance are considered confounders. We calculate the confounder percentage of graph representations learned by a trained ERM, which is compared with the confounder percentage of graph representations achieved by RCGRL. In Figure 5(a), the graphs with IV have a significant decrease in the proportion...
Figure 4: Performance of RCGRL with different $u$. If $u = 0$, we introduce the IVs at the beginning. All models consist of four GNN layers.

Figure 5: The analyses of confounders. Visualized results show the confounders of several graphs in the datasets, and the orange substructures indicate the confounders measured by using the principle of Graph Granger Causality (Lin, Lan, and Li 2021).

**Conclusion**

We propose RCGRL to learn robust graph representations against confounding effects. RCGRL actively generates instrumental variables under unconditional moment restrictions to eliminate confounders. Theoretically and empirically, we demonstrate the effectiveness of the proposed RCGRL.

**Acknowledgements**

The authors would like to thank the anonymous reviewers for their valuable comments. This work is supported by National Key Research and Development Program of China No. 2019YFB1405100, CAS Project for Young Scientists in Basic Research, Grant No. YSBR-040.

**References**

Ai, C.; and Chen, X. 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6): 1795–1843.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Gao, H.; and Ji, S. 2019a. Graph U-Nets. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 2083–2092. PMLR.

Gao, H.; and Ji, S. 2019b. Graph u-nets. In *international conference on machine learning*, 2083–2092. PMLR.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Greene, W. H. 2003. *Econometric analysis*. Pearson Education India.

Hansen, L. P.; Heaton, J.; and Yaron, A. 1996. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3): 262–280.

Hayashi, F. 2000. *Econometrics*, l Princeton University Press: Princeton.

Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; and Leskovec, J. 2020. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33: 22118–22133.

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

Krueger, D.; Caballero, E.; Jacobsen, J.-H.; Zhang, A.; Biswas, J.; Zhang, D.; Le Priol, R.; and Courville, A. 2021. Out-of-distribution generalization via risk extrapolation (rex).
