Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism

Presenter: Zhanke Zhou

2022. 06. 30

About the paper

Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism

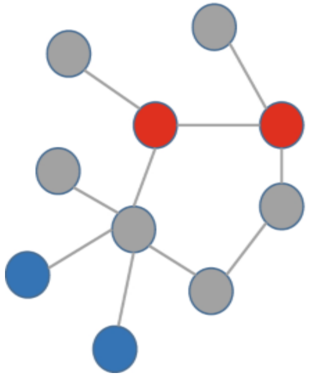
- Authors: Siqi Miao, Miaoyuan Liu, Pan Li
- Conference: ICML 2022
- Affiliation: Department of Computer Science, Purdue University
- Paper: https://arxiv.org/pdf/2201.12987.pdf
- Code: https://github.com/Graph-COM/GSAT

Outline

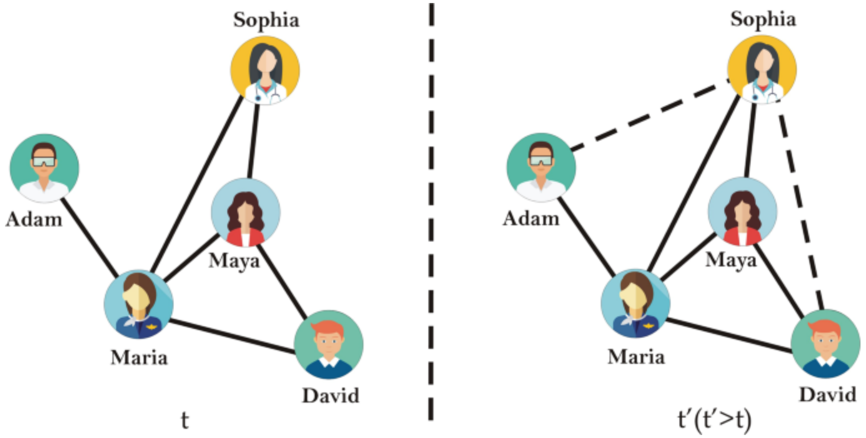
- Background
- The existing methods
- The proposed method
- Experiment
- Summary and discussion

Background | graph learning

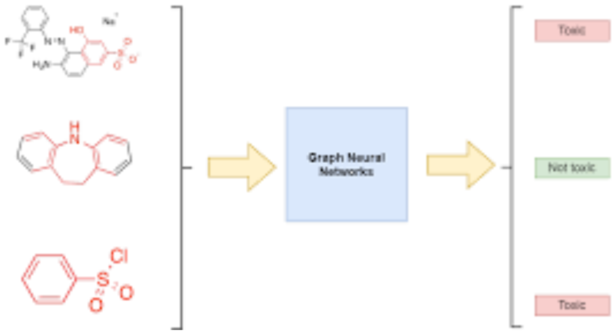
node-level



link-level

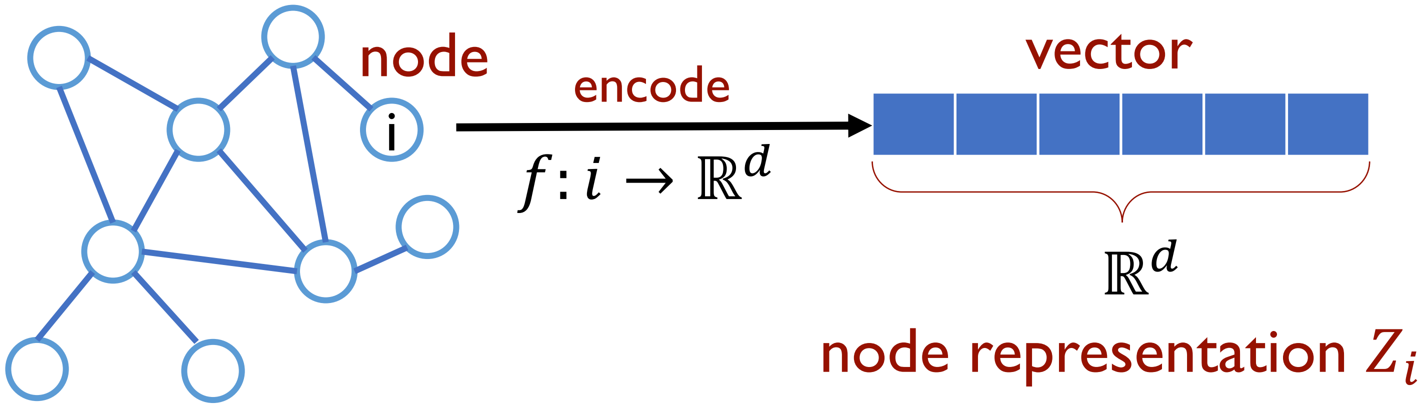


graph-level



Background | graph learning

Graph data $D \rightarrow$ GNN $f \rightarrow$ representation $Z \rightarrow \tilde{Y} \leftrightarrow Y$

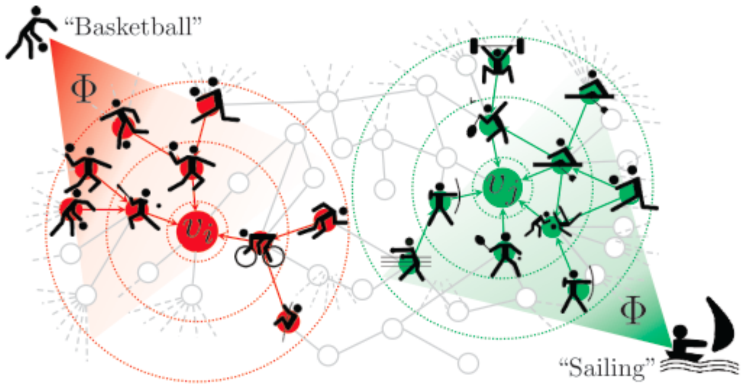


However, only powerful is not enough

Background | motivation

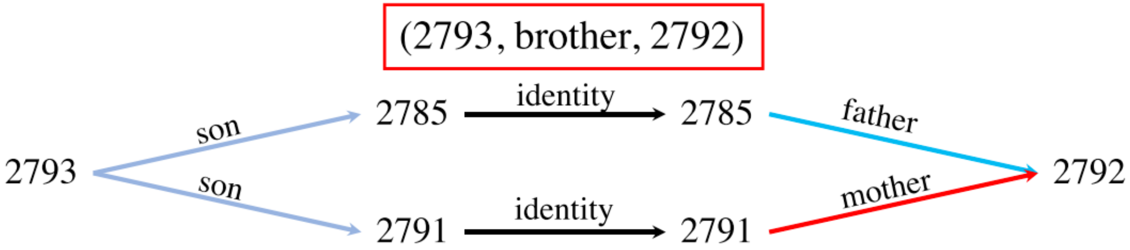
node-level task: requires relevant nodes

- e.g., node classification



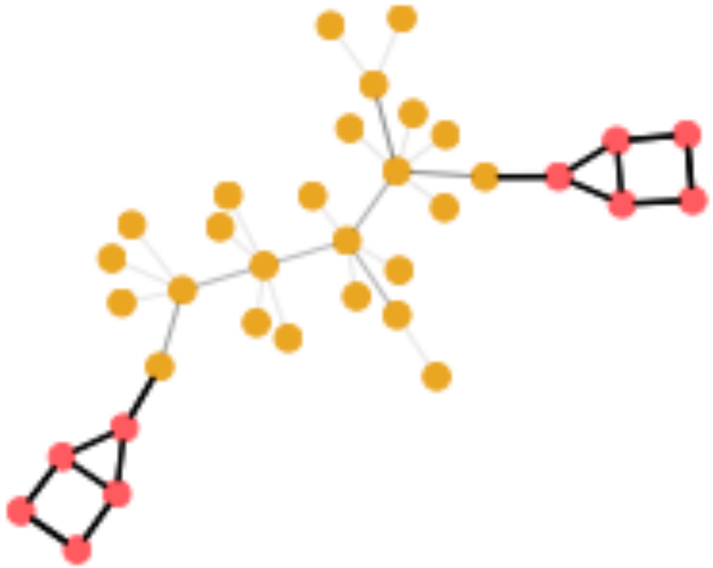
link-level task: requires relevant paths

- e.g., link prediction



graph-level task: requires relevant subgraphs

- e.g., graph classification



Background | motivation

Graph data $D \to$ GNN $f \to$ representation $Z \to \tilde{Y} \leftrightarrow Y$

Powerful

i.e., to approximate Y by $\tilde{Y}$

the learned representation and graph data are usually highly entangled

Interpretable

i.e., which parts in D contribute to $\tilde{Y}$

an important property to trustworthy ML

e.g. identifying the functional groups in a molecule

Core problem:

how to provide more accurate interpretation without sacrificing the accuracy?

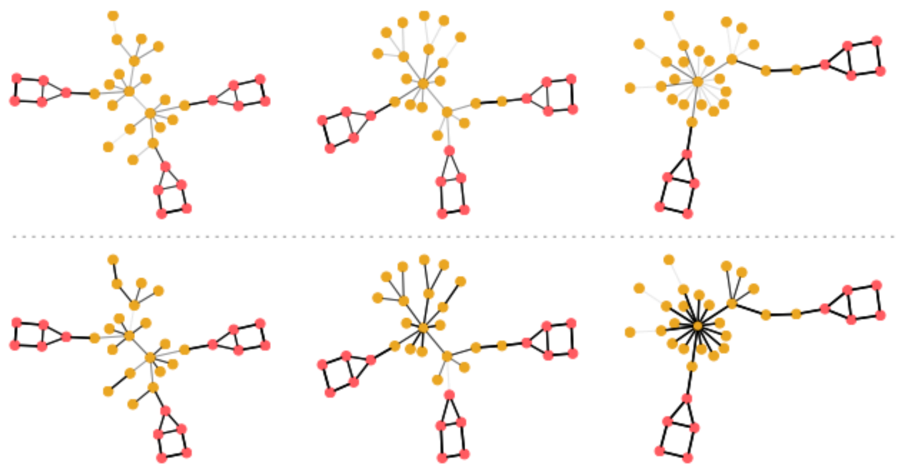
Background | motivation

Graph classification

Image classification

Proposed

Baseline



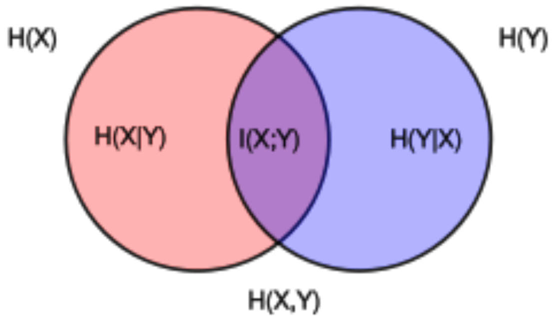
The proposed method can provide the more accurate interpretation

- at the same time, it is not harmful to the performance, and even boost it

Outline

- Background
- The existing methods
- The proposed method
- Experiment
- Summary and discussion

Preliminaries | mutual information



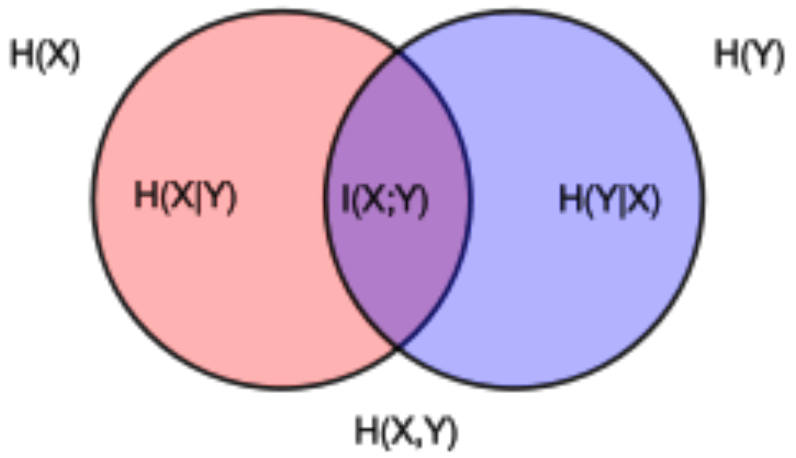
- the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables.

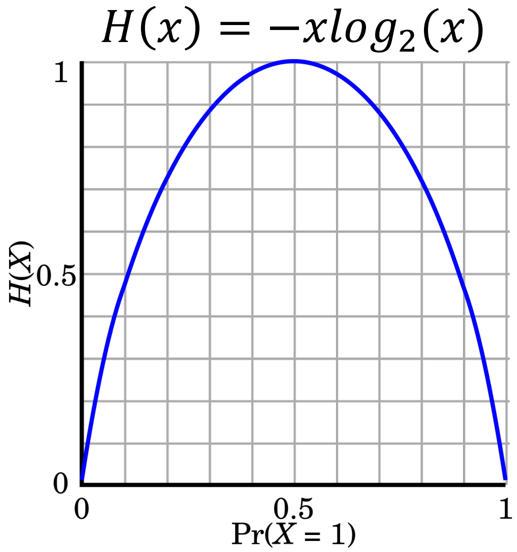
- definition: $I(X;Y) = I(Y;X) = D_{KL}(p(x,y)||p(x)\otimes p(y))$
- discrete variables: $I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$
- continuous variables: $I(X;Y) = \int_Y \int_X p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$

https://en.wikipedia.org/wiki/Mutual_information

Preliminaries | mutual information

- $I(X;Y) = H(X) - H(X|Y)$
- $I(X;Y) = H(Y) - H(Y|X)$
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$





The existing post-hoc methods

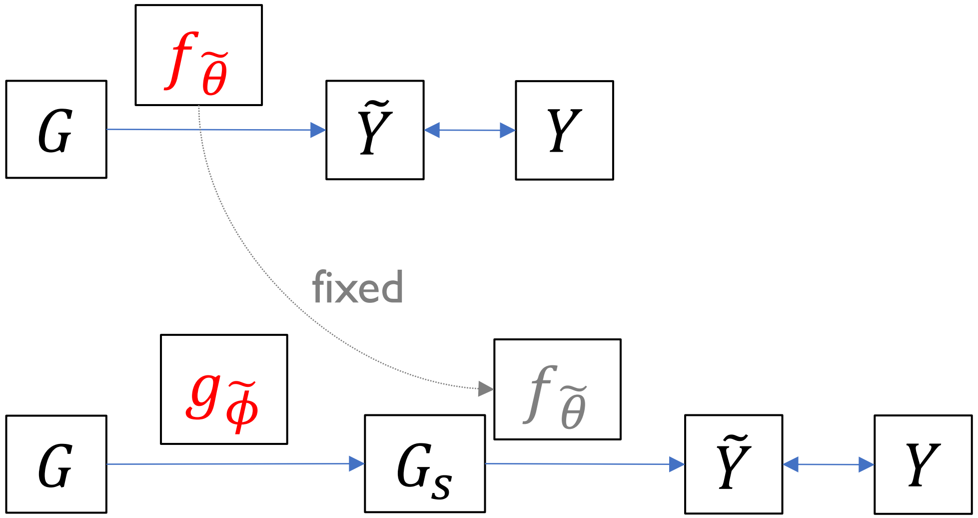
For example:

step1: obtain the model parameter $\tilde{\theta}$

- i.e., the predictor

step2: optimize the subgraph extractor $\tilde{\phi}$

- reducing the MI $I(G;\tilde{Y}) - I(G_s;\tilde{Y})$
- with designed constraint (e.g., size, connectivity)

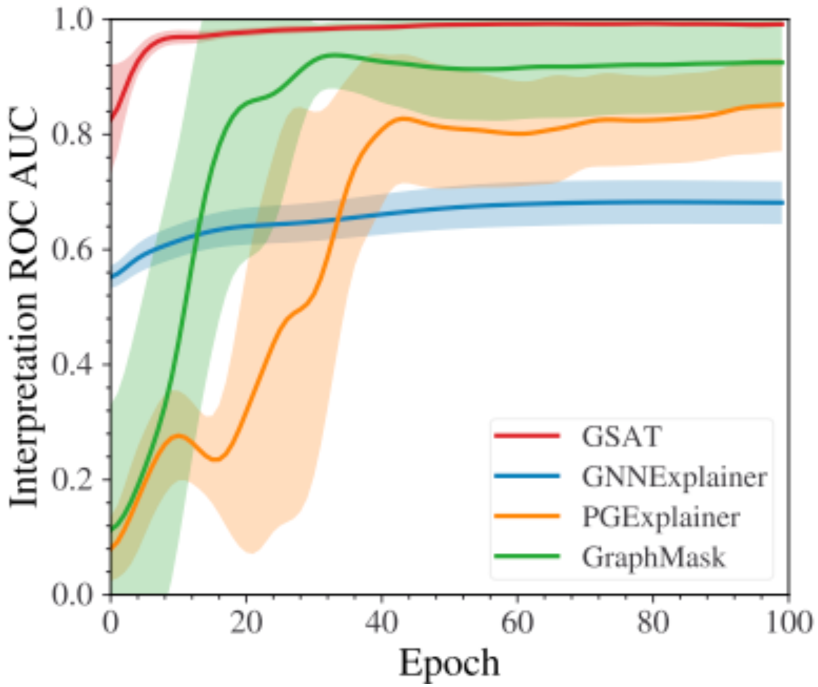


$f_{\tilde{\theta}} \circ g_{\tilde{\phi}}$

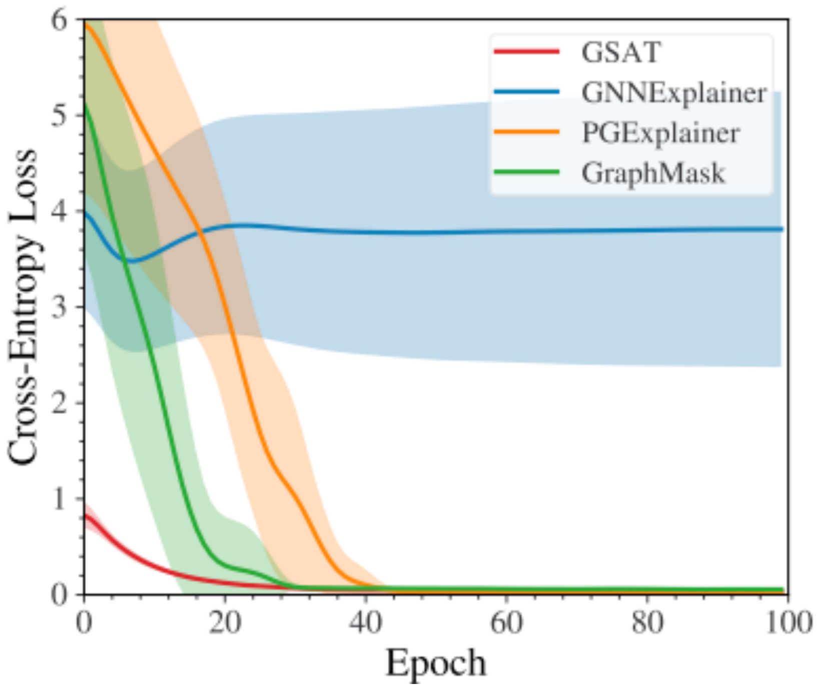
the interpretating system

The existing post-hoc methods | problems

interpretation performance



training loss

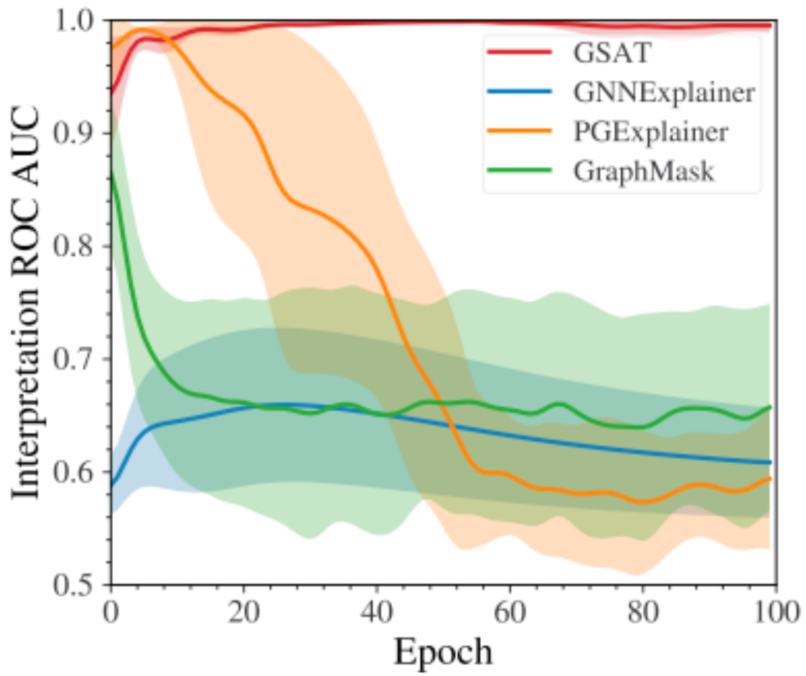


Observation: (under-fitting)

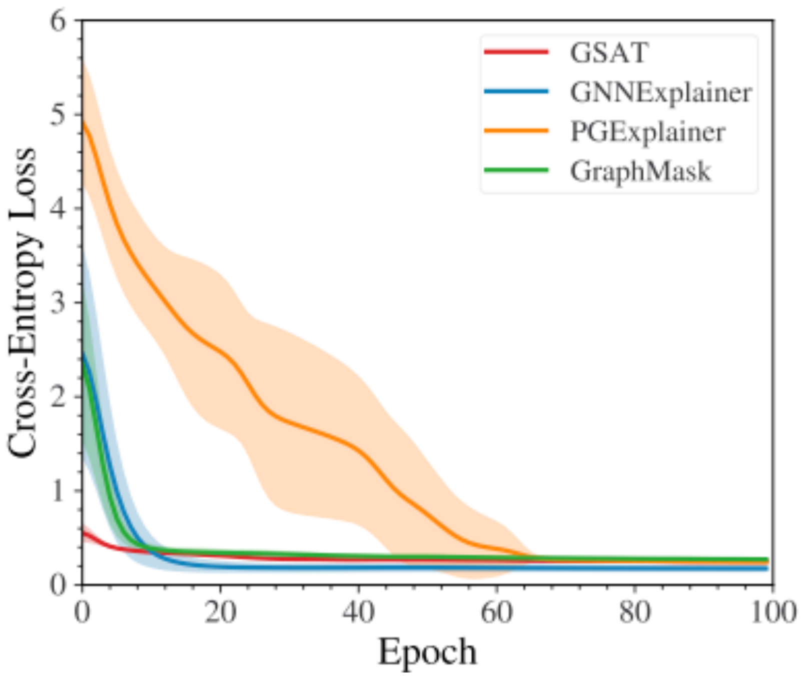
the interpretation is sub-optimal and the training loss keeps high

The existing post-hoc methods | problems

interpretation performance



training loss



Observation: (over-fitting)

The overfitting problems are severe and common

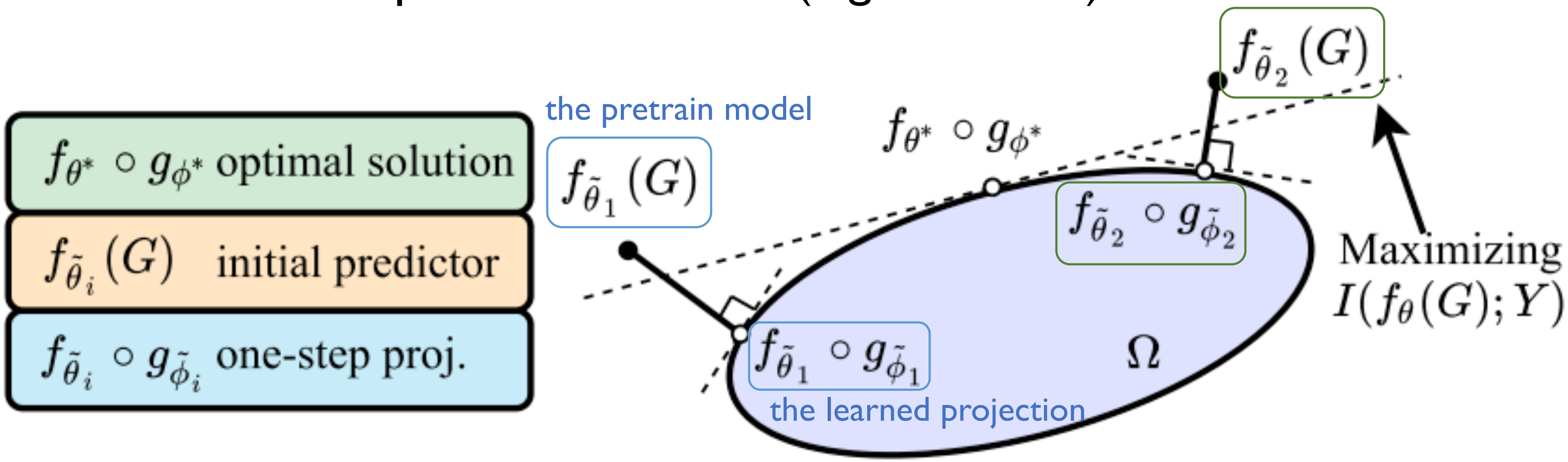
However, it is hard to have the ground truth interpretation labels in practice 🤔

The existing post-hoc methods | problems

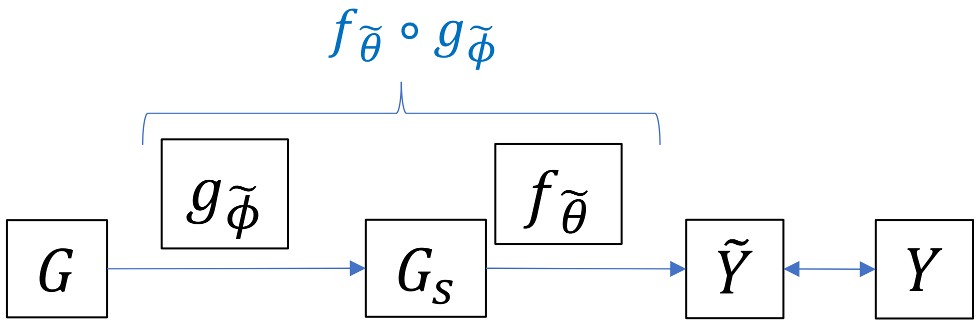
Post-hoc methods just perform one-step projection to the information-constrained space (Ω)

cons

- always suboptimal (low accuracy)
- sensitive to the pre-trained model (high variance)



The existing post-hoc methods | problems



(post-hoc) reducing the MI $I(G;\tilde{Y}) - I(G_s;\tilde{Y})$is not good enough

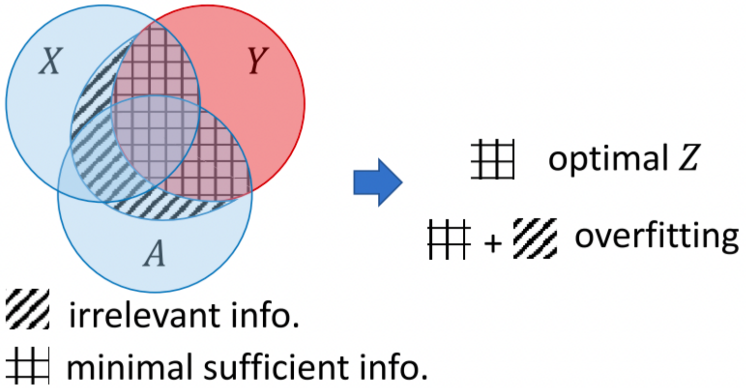
a joint training of $f_{\tilde{\theta}} \circ g_{\tilde{\phi}}$ might be better 🤔

Outline

- Background
- The existing methods
- The proposed method
- Experiment
- Summary and discussion

Graph information bottleneck (GIB)

D=(X,A) → (GNN) → **Z** ↔ **Y**



Y: The target, $\mathcal{D}$: The input data ($= (A, X)$)
A: The graph structure, X: The node features
Z: The representation

Graph Information Bottleneck:

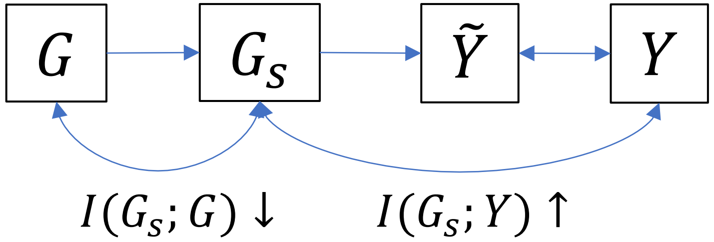
$$\min_{\mathbb{P}(Z|\mathcal{D})\in\Omega} \text{GIB}_\beta(\mathcal{D}, Y; Z) \triangleq [-I(Y;Z) + \beta I(\mathcal{D};Z)]$$

Figure 1: Graph Information Bottleneck is to optimize the representation Z to capture the minimal sufficient information within the input data $\mathcal{D} = (A, X)$ to predict the target Y. $\mathcal{D}$ includes information from both the graph structure A and node features X. When Z contains irrelevant information from either of these two sides, it overfits the data and is prone to adversarial attacks and model hyperparameter change. Ω defines the search space of the optimal model $\mathbb{P}(Z|\mathcal{D})$. $I(\cdot;\cdot)$ denotes the mutual information [17].

Graph Information Bottleneck. NeurIPS 2020.

Graph information bottleneck (GIB)

inspired by the GIB, this work uses
information constraint to select label-relevant subgraph



Graph Information Bottleneck:

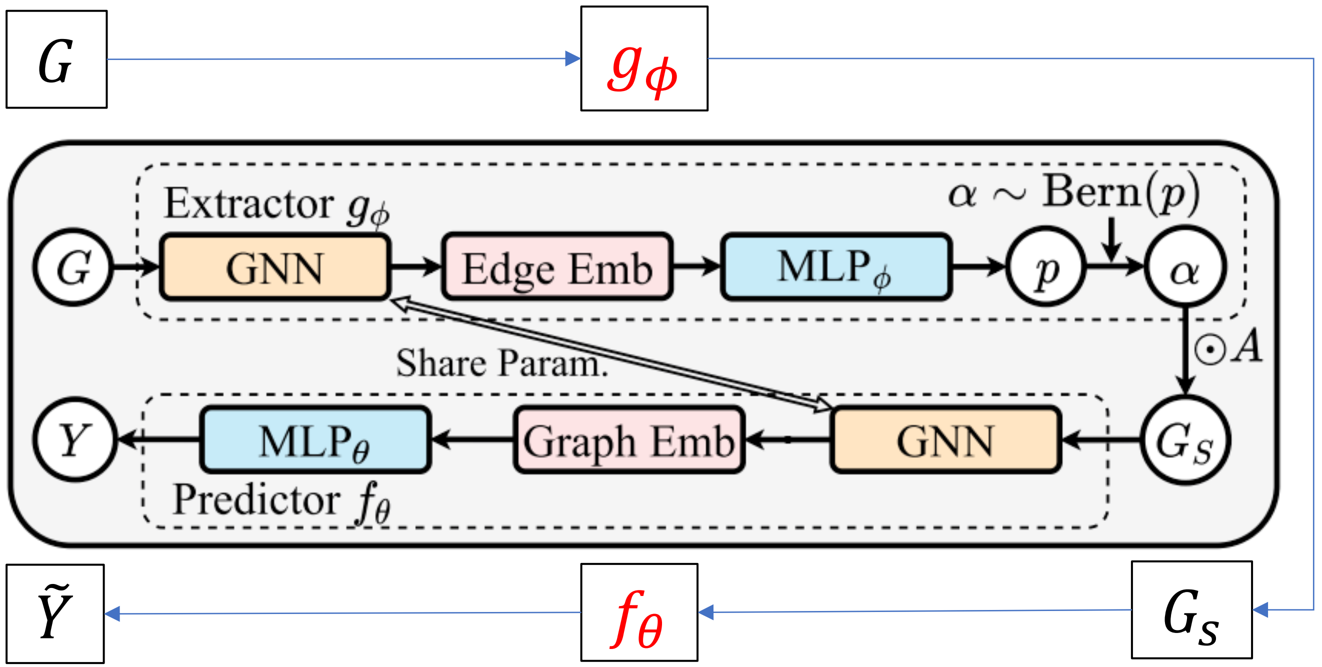
$$\min_{\mathbb{P}(Z|\mathcal{D})\in\Omega} \mathrm{GIB}_{\beta}(\mathcal{D}, Y; Z) \triangleq [-I(Y; Z) + \beta I(\mathcal{D}; Z)]$$

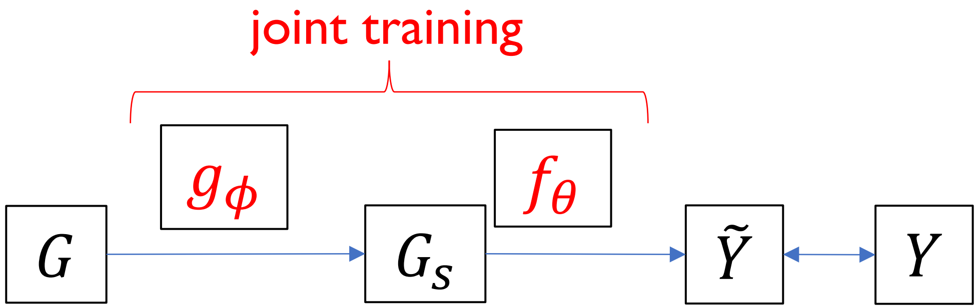
$$\min_{\phi} -I(G_S; Y) + \beta I(G_S; G), \text{ s.t. } G_S \sim g_{\phi}(G)$$

not impose any potentially biased constraints

- e.g., graph size or connectivity (adopted by other works)

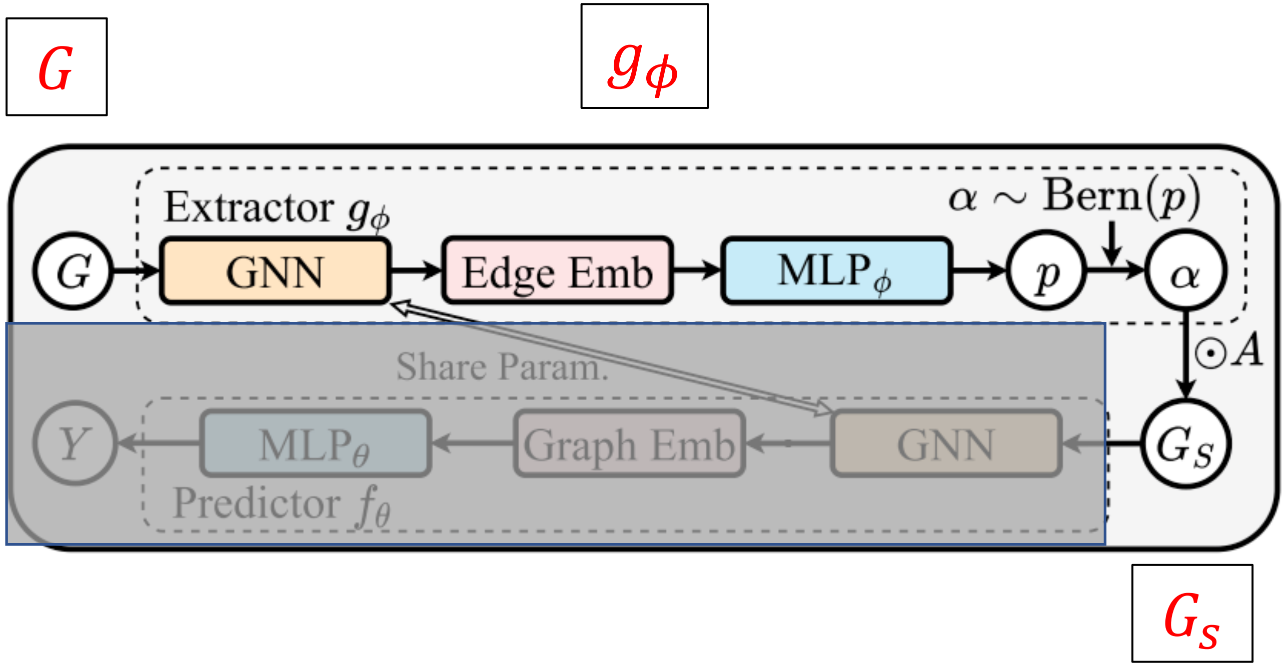
The proposed method

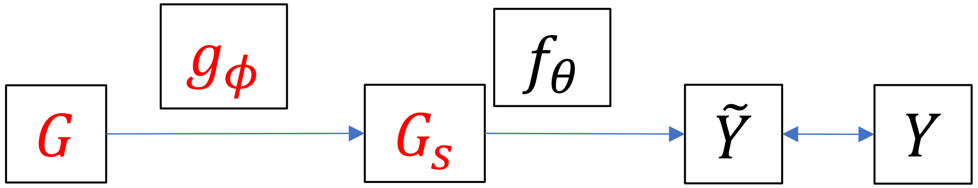




$$\min_{\phi} -I(G_S;Y) + \beta I(G_S;G), \text{ s.t. } G_S \sim g_\phi(G)$$

The proposed method | extractor





1. obtain the node embeddings (representation)

$$GNN(G) \rightarrow \boldsymbol{H} \in \mathbb{R}^{N \times D}$$

2. obtain the edge embeddings

$$\boldsymbol{H}_{edge} = \{[\boldsymbol{h}_i, \boldsymbol{h}_j]: e_{ij} \in \mathcal{E}\}$$

3. obtain the edge probabilities (importance)

$$\boldsymbol{P}_{edge} = MLP(\boldsymbol{H}_{edge})$$

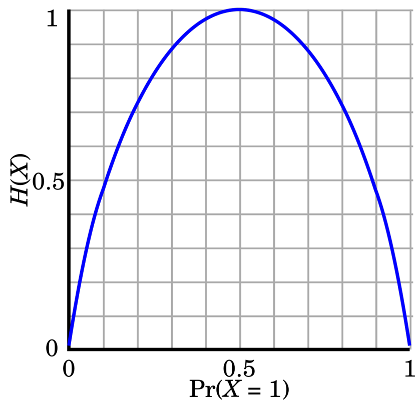
4. obtain the sampled graph G_s with random noise

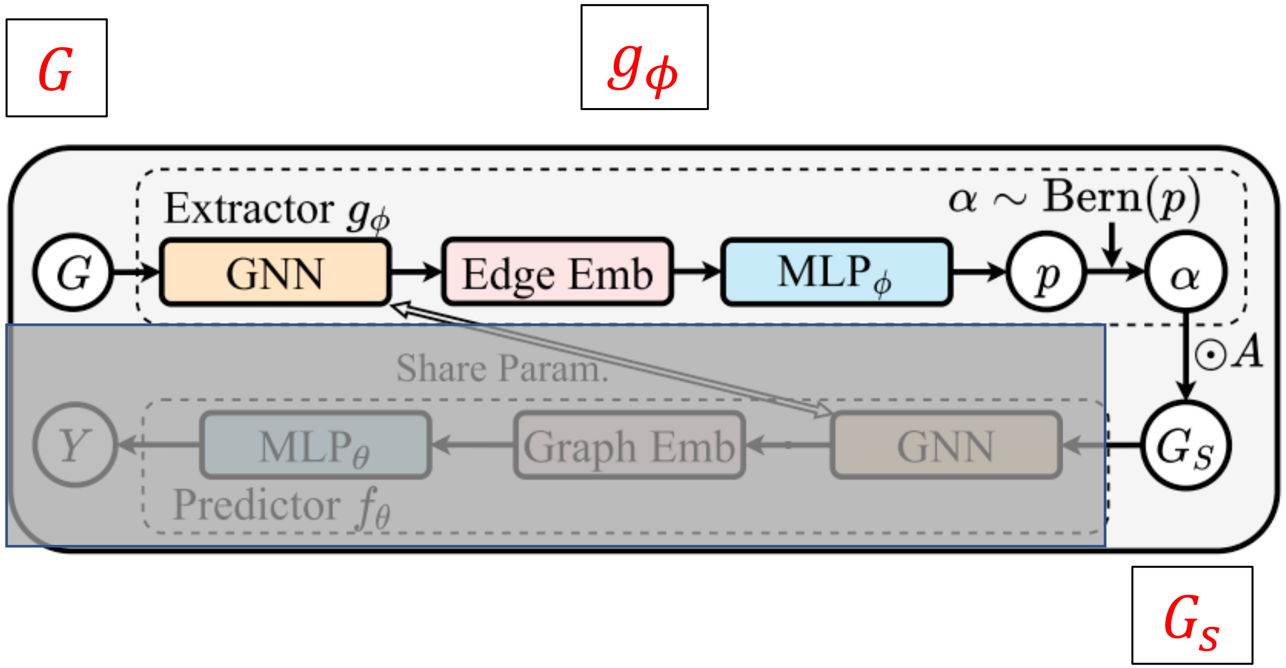
$$\alpha_{ij} \sim \text{Bernoulli}(\boldsymbol{p}_{ij} + u)$$

$$A_s = \alpha \odot A \in \mathbb{R}^{N \times N}$$

$$G_s = (A_s, X)$$

The proposed method | extractor

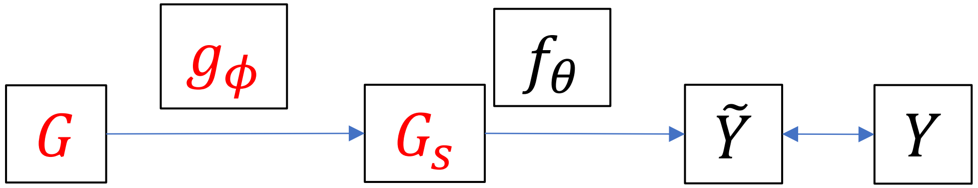




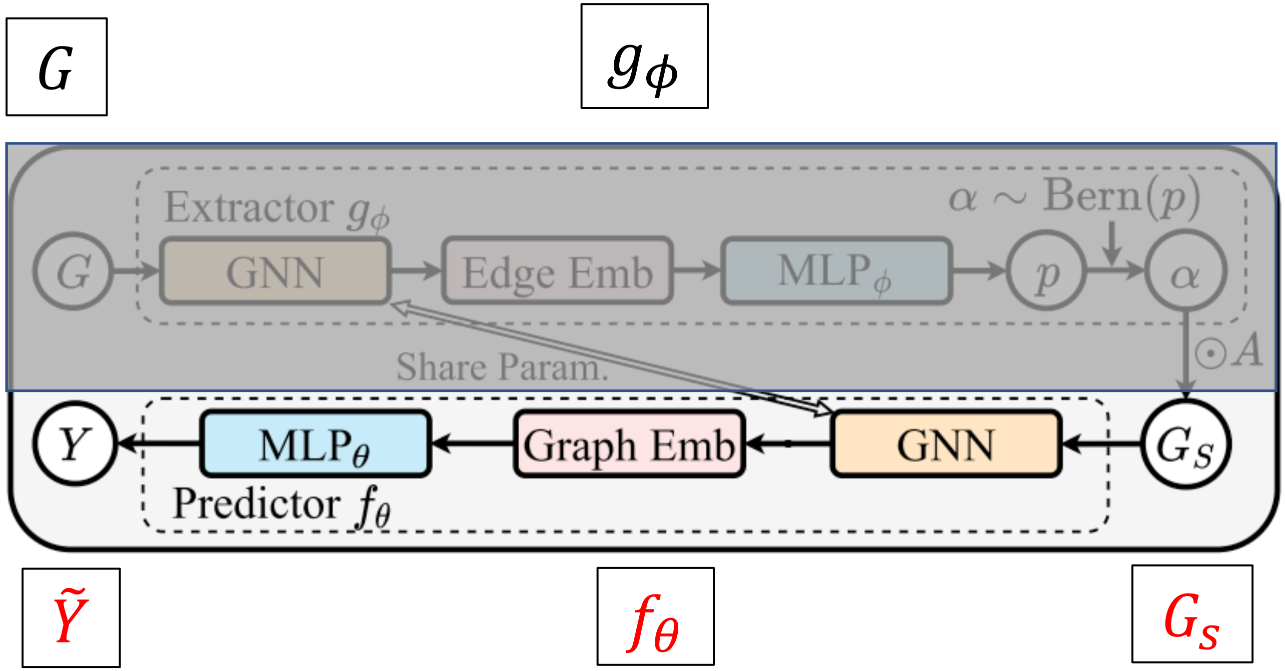
$$\min_{\phi} -I(G_S;Y) + \beta I(G_S;G), \text{ s.t. } G_S \sim g_\phi(G)$$

$$I(G_s;G) \leq \mathbb{E}_G \left[\text{KL}(\mathbb{P}_\phi(G_S|G)||\mathbb{Q}(G_S))\right]$$

$$\text{KL}(\mathbb{P}_\phi(G_S|G)||\mathbb{Q}(G_S)) = \tag{9}$$
$$\sum_{(u,v)\in E} p_{uv} \log \frac{p_{uv}}{r} + (1-p_{uv}) \log \frac{1-p_{uv}}{1-r} + c(n,r).$$



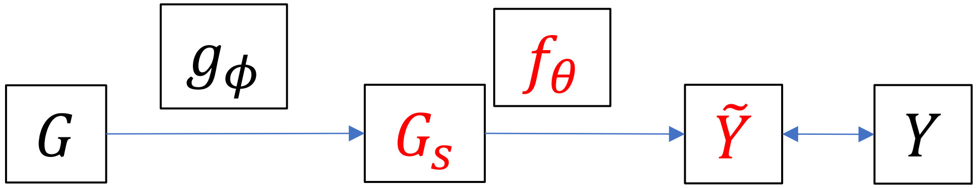
The proposed method | predictor



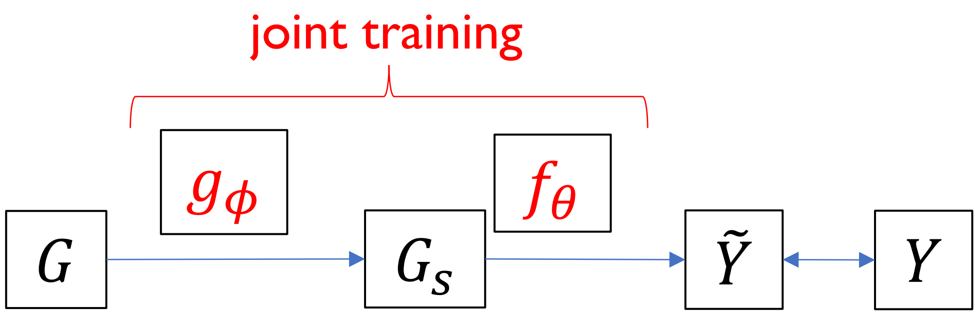
$$\min_{\phi} -I(G_S;Y) + \beta I(G_S;G), \text{ s.t. } G_S \sim g_\phi(G)$$

$$I\left(G_S;Y\right) \geq \mathbb{E}_{G_S,Y}\left[\log \mathbb{P}_\theta(Y|G_S)\right] + H(Y)$$

classification loss, e.g., cross entropy



Full learning objective



$$\min_{\phi} -I(G_S;Y) + \beta I(G_S;G), \text{ s.t. } G_S \sim g_\phi(G)$$

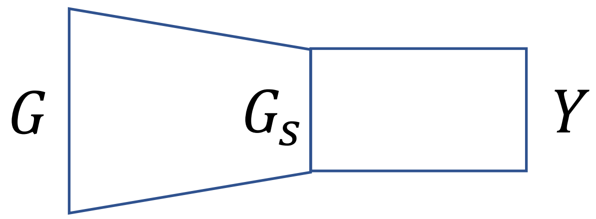
$$I(G_s;G) \leq \mathbb{E}_G\left[\mathbf{KL}(\mathbb{P}_\phi(G_S|G)||\mathbb{Q}(G_S))\right] \qquad (g_{\tilde{\phi}})$$

$$I(G_S;Y) \geq \mathbb{E}_{G_S,Y}\left[\log \mathbb{P}_\theta(Y|G_S)\right] + H(Y) \qquad (f_{\tilde{\theta}})$$

$$\min_{\theta,\phi} -\mathbb{E}\left[\log \mathbb{P}_\theta(Y|G_S)\right] + \beta\mathbb{E}\left[\mathbf{KL}(\mathbb{P}_\phi(G_S|G)||\mathbb{Q}(G_S))\right]$$

Further interpretation

$$\min_{\phi} -I(G_S;Y) + \beta I(G_S;G)$$



GSAT decreases the information from the input graphs

- with injecting stochasticity for all edges

GSAT can learn to reduce such stochasticity on the task-relevant subgraphs

- when $p_{ij} \to 1$, such edge ($e_{ij} \in \mathcal{E}$) is "invariant" and provides interpretation

Outline

- Background
- The existing methods
- The proposed method
- Experiment
- Summary and discussion

Experiment

Interpretation 👍

Table 1. Interpretation Performance (AUC). The underlined results highlight the best baselines. The **bold** font and **bold**† font highlight when GSAT outperform the means of the best baselines based on the mean of GSAT and the mean-2*std of GSAT, respectively.

	BA-2MOTIFS	MUTAG	MNIST-75SP	SPURIOUS-MOTIF $b = 0.5$	SPURIOUS-MOTIF $b = 0.7$	SPURIOUS-MOTIF $b = 0.9$
GNNEXPLAINER	67.35 ± 3.29	61.98 ± 5.45	59.01 ± 2.04	62.62 ± 1.35	62.25 ± 3.61	58.86 ± 1.93
PGEXPLAINER	84.59 ± 9.09	60.91 ± 17.10	69.34 ± 4.32	69.54 ± 5.64	72.33 ± 9.18	$\underline{72.34} \pm 2.91$
GRAPHMASK	$\underline{92.54} \pm 8.07$	62.23 ± 9.01	$\underline{73.10} \pm 6.41$	72.06 ± 5.58	73.06 ± 4.91	66.68 ± 6.96
IB-SUBGRAPH	86.06 ± 28.37	$\underline{91.04} \pm 6.59$	51.20 ± 5.12	57.29 ± 14.35	62.89 ± 15.59	47.29 ± 13.39
DIR	82.78 ± 10.97	64.44 ± 28.81	32.35 ± 9.39	$\underline{78.15} \pm 1.32$	$\underline{77.68} \pm 1.22$	49.08 ± 3.66
GIN+GSAT	$\mathbf{98.74}^\dagger \pm 0.55$	$\mathbf{99.60}^\dagger \pm 0.51$	$\mathbf{83.36}^\dagger \pm 1.02$	$\mathbf{78.45} \pm 3.12$	74.07 ± 5.28	71.97 ± 4.41
GIN+GSAT*	$\mathbf{97.43}^\dagger \pm 1.77$	$\mathbf{97.75}^\dagger \pm 0.92$	$\mathbf{83.70}^\dagger \pm 1.46$	$\mathbf{85.55}^\dagger \pm 2.57$	$\mathbf{85.56}^\dagger \pm 1.93$	$\mathbf{83.59}^\dagger \pm 2.56$
PNA+GSAT	$\mathbf{93.77} \pm 3.90$	$\mathbf{99.07}^\dagger \pm 0.50$	$\mathbf{84.68}^\dagger \pm 1.06$	$\mathbf{83.34}^\dagger \pm 2.17$	$\mathbf{86.94}^\dagger \pm 4.05$	$\mathbf{88.66}^\dagger \pm 2.44$
PNA+GSAT*	89.04 ± 4.92	$\mathbf{96.22}^\dagger \pm 2.08$	$\mathbf{88.54}^\dagger \pm 0.72$	$\mathbf{90.55}^\dagger \pm 1.48$	$\mathbf{89.79}^\dagger \pm 1.91$	$\mathbf{89.54}^\dagger \pm 1.78$

Prediction 👍

Table 2. Prediction Performance (Acc.). The **bold** font highlights the inherently interpretable methods that significantly outperform the corresponding backbone model, GIN or PNA, when the mean-1*std of a method > the mean of its corresponding backbone model.

	MOLHIV (AUC)	GRAPH-SST2	MNIST-75SP	SPURIOUS-MOTIF $b = 0.5$	SPURIOUS-MOTIF $b = 0.7$	SPURIOUS-MOTIF $b = 0.9$
GIN	76.69 ± 1.25	82.73 ± 0.77	95.74 ± 0.36	39.87 ± 1.30	39.04 ± 1.62	38.57 ± 2.31
IB-SUBGRAPH	76.43 ± 2.65	82.99 ± 0.67	93.10 ± 1.32	$\mathbf{54.36} \pm 7.09$	$\mathbf{48.51} \pm 5.76$	$\mathbf{46.19} \pm 5.63$
DIR	76.34 ± 1.01	82.32 ± 0.85	88.51 ± 2.57	$\mathbf{45.49} \pm 3.81$	41.13 ± 2.62	37.61 ± 2.02
GIN+GSAT	76.47 ± 1.53	82.95 ± 0.58	$\mathbf{96.24} \pm 0.17$	$\mathbf{52.74} \pm 4.08$	$\mathbf{49.12} \pm 3.29$	$\mathbf{44.22} \pm 5.57$
GIN+GSAT*	76.16 ± 1.39	82.57 ± 0.71	$\mathbf{96.21} \pm 0.14$	$\mathbf{46.62} \pm 2.95$	41.26 ± 3.01	39.74 ± 2.20
PNA (NO SCALARS)	78.91 ± 1.04	79.87 ± 1.02	87.20 ± 5.61	68.15 ± 2.39	66.35 ± 3.34	61.40 ± 3.56
PNA+GSAT	$\mathbf{80.24} \pm 0.73$	$\mathbf{80.92} \pm 0.66$	$\mathbf{93.96} \pm 0.92$	68.74 ± 2.24	64.38 ± 3.20	57.01 ± 2.95
PNA+GSAT*	$\mathbf{80.67} \pm 0.95$	$\mathbf{82.81} \pm 0.56$	$\mathbf{92.38} \pm 1.44$	$\mathbf{69.72} \pm 1.93$	$\mathbf{67.31} \pm 1.86$	61.49 ± 3.46

Experiment

Table 4. Direct comparison (Acc.) with invariant learning methods on the ability to remove spurious correlations, by applying the backbone model used in (Wu et al., 2022).

SPURIOUS-MOTIF	$b = 0.5$	$b = 0.7$	$b = 0.9$
ERM	39.69 ± 1.73	38.93 ± 1.74	33.61 ± 1.02
V-REx	39.43 ± 2.69	39.08 ± 1.56	34.81 ± 2.04
IRM	41.30 ± 1.28	40.16 ± 1.74	35.12 ± 2.71
DIR	45.50 ± 2.15	43.36 ± 1.64	39.87 ± 0.56
GSAT	$\mathbf{53.27}^{\dagger} \pm 5.12$	$\mathbf{56.50}^{\dagger} \pm 3.96$	$\mathbf{53.11}^{\dagger} \pm 4.64$
GSAT*	43.27 ± 4.58	42.51 ± 5.32	$\mathbf{45.76}^{\dagger} \pm 5.32$

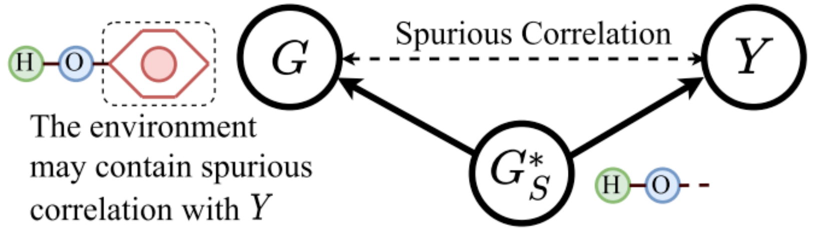


Figure 6. G_S^* determines Y. However, the environment features in $G \backslash G_S^*$ may contain spurious (backdoor) correlation with Y.

Theorem 4.1. Suppose each G contains a subgraph G_S^* such that Y is determined by G_S^* in the sense that $Y = f(G_S^*) + \epsilon$ for some deterministic invertible function f with randomness ϵ that is independent from G. Then, for any $\beta \in [0, 1]$, $G_S = G_S^*$ maximizes the GIB $I(G_S; Y) - \beta I(G_S; G)$, where $G_S \in \mathbb{G}_{\text{sub}}(G)$.

GSAT can remove spurious correlations in the training data 👍🏻

- mainly due to the injecting stochasticity

Experiment

Table 5. Ablation study on β and stochasticity in GSAT (GIN as the backbone model) on Spurious-Motif. We report both interpretation ROC AUC (top) and prediction accuracy (bottom).

SPURIOUS-MOTIF	$b = 0.5$	$b = 0.7$	$b = 0.9$
GSAT	79.81 ± 3.98	74.07 ± 5.28	71.97 ± 4.41
GSAT-$\beta = 0$	66.00 ± 11.04	65.92 ± 3.28	66.31 ± 6.82
GSAT-NOSTOCH	59.64 ± 5.33	55.78 ± 2.84	55.27 ± 7.49
GSAT-NOSTOCH-$\beta = 0$	63.37 ± 12.33	60.61 ± 10.08	66.19 ± 7.76
GIN	39.87 ± 1.30	39.04 ± 1.62	38.57 ± 2.31
GSAT	51.86 ± 5.51	49.12 ± 3.29	44.22 ± 5.57
GSAT-$\beta = 0$	45.97 ± 8.37	49.67 ± 7.01	49.84 ± 5.45
GSAT-NOSTOCH	40.34 ± 2.77	41.90 ± 3.70	37.98 ± 2.64
GSAT-NOSTOCH-$\beta = 0$	43.41 ± 8.05	45.88 ± 9.54	42.25 ± 9.77

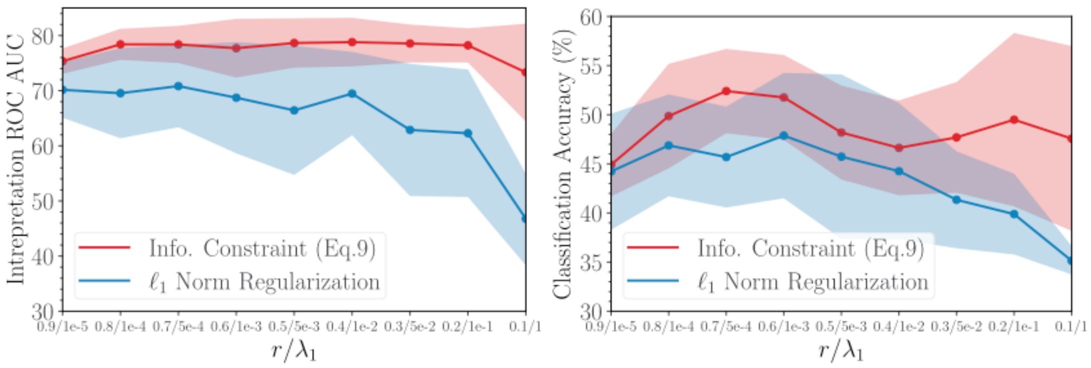
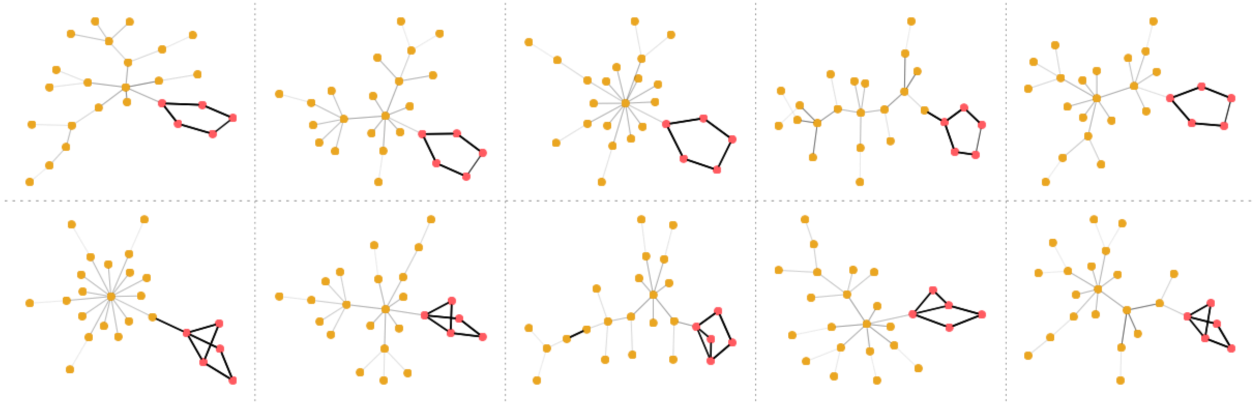


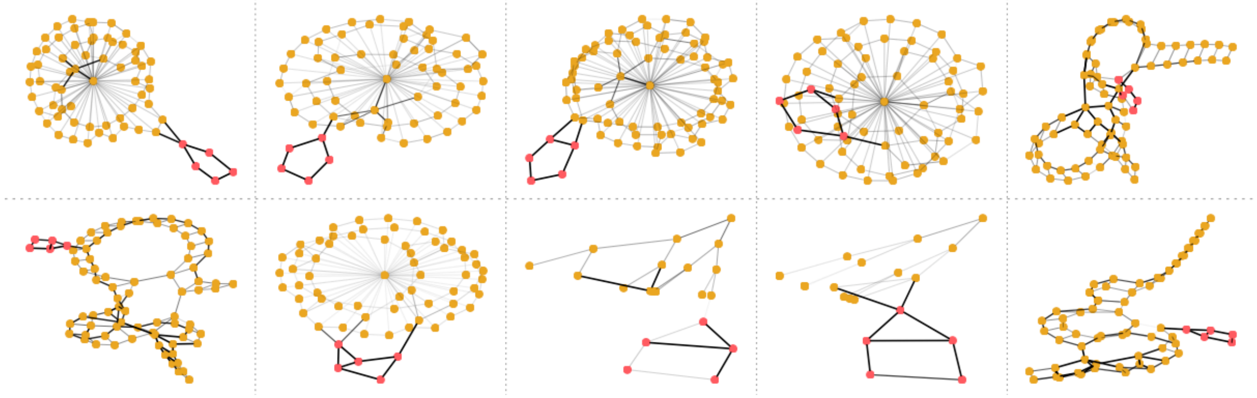
Figure 7. Comparison between (a) using the information constraint in Eq. (9) and (b) replacing it with ℓ_1-norm. Results are shown for Spurious-Motif $b = 0.5$, where r is tuned from 0.9 to 0.1 and the coefficient of the ℓ_1-norm λ_1 is tuned from $1e$-5 to 1.

graph information bottleneck (GIB) 👍🏻
stochasticity (gumbel trick) 👍🏻

Experiment



since the GSAT dose not make any assumptions on the selected subgraphs,
the improvement of GSAT can be even more
if the true subgraph are dis-connected or vary in sizes.



Outline

- Background
- The existing methods
- The proposed method
- Experiment
- Summary and discussion

Summary

First, the GIB frees GSAT from any potentially biased assumptions

- which are adopted in previous methods

Second, GSAT can provably remove spurious correlations in the training data

- by reducing the information from the input graph

Third, GSAT can cooperate with the pre-trained model if provided

- GSAT may further improve both of its interpretation and prediction accuracy

Related works

1. GNNExplainer: Generating Explanations for Graph Neural Networks. NeurIPS 2019.
2. Graph Information Bottleneck. NeurIPS 2020.
3. Parameterized Explainer for Graph Neural Network. NeurIPS 2020.
4. INTERPRETING GRAPH NEURAL NETWORKS FOR NLP WITH DIFFERENTIABLE EDGE MASKING. ICLR 2021.
5. GRAPH INFORMATION BOTTLENECK FOR SUBGRAPH RECOGNITION. ICLR 2021
6. DISCOVERING INVARIANT RATIONALES FOR GRAPH NEURAL NETWORKS. ICLR 2022.

Q&A

Thanks for your attention!