DAG matters! Gflownets Enhanced Explainer for Graph Neural Networks

Wenqian Li, Yinchuan Li, Zhigang Li, Jianye HAO, Yan Pang

Overview

- Uncovering rationales behind predictions of GNNs focus on selecting a subgraph through combinatorial optimizations.
- Turn the combinatorial optimization problem into a step-by-step generative problem, aiming to learn the distribution of subgraphs.
- Construct the Directed Acyclic Graph structure for sequential modeling.
- Dynamically check cut vertices to check the connectivity of the subgraph, efficiently explore parent states for the GFlowNets structure.

Problem Statement

What problem the post-hoc GNN explanation solves

- Given an instance, a node v or a graph G, the goal of GNN explanation is to identify a subgraph $G_S = (\mathcal{V}_S, \mathcal{E}_S)$ and the associated features $X_S = \{x_j | v_j \in G_S\}$ that are important for the GNN prediction $Y_i = \Phi(v_i)$ or $Y_{g_i} = \Phi(G_i)$, where g_i is a graph instance, Φ is the trained GNN model.
- The objective is to maximization the mutual information:

$$\max MI(Y, G_S) = H(Y) - H(Y|G_S)$$

Motivation of GFlowExplainer

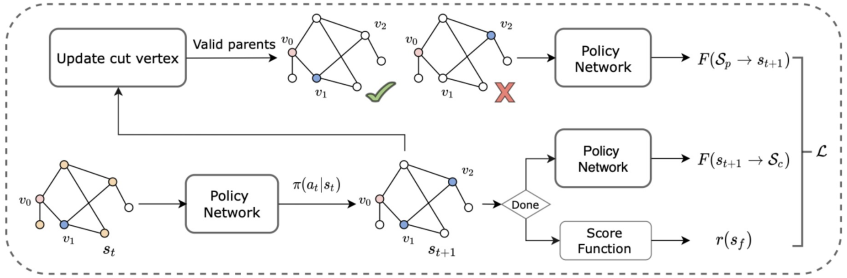
- **Existing RL Solution 1: Reward maximization**
  - Issue: Local optimum in combinatorial optimizations
  - Motivation: TD-like flow matching condition
- **Existing RL Solution 2: Sequential modeling**
  - Issue: Computational expensive pretraining strategy
  - Motivation: Consider graph as an ordered set, construct DAG structure

High-level Framework of GFlowExplainer

- GFlowExplainer consists of a tuple $(\mathcal{S}, \mathcal{A})$, $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is the action set consisting transitions: $a_t: s_t \to s_{t+1}$.
- Consider G_S as a compositional object.
- Starting from an empty graph, different from traditional optimization problems maximizing the mutual information, the objective is to construct **TD-like flow matching condition** , to obtain a generative forward policy $\pi(a_t|s_t)$ so that $P(Y, G_S) \propto r(Y, G_S)$.

<u>The probability of generating a subgraph is proportional to its reward</u>

Our Solution



Aggregate information for the graph structured data

- ***Feature representation*** $x_i' = [x_i, \mathbb{1}_{v_i=v_0}, \mathbb{1}_{\{v_i \in G_s(s_t)\}}]$, $X_t' = [x_i']_{\forall v_i \in G_s(s_t) \cup \mathcal{N}(s_t)}$
- ***Combine information*** $H_t^{(0)} = \Theta_1 X_t'$, $H_t^{(l+1)} = (1-\alpha)\hat{A}H_t^{(l)} + \alpha H_t^{(0)}$
- ***Improve representation*** $\overline{H_t}(v_i) = \mathrm{MLP}(H_t^L(v_i); \Theta_2)$, $v_i \in G_s(s_t) \cup \mathcal{N}(s_t)$

Self-attention mechanism to avoid generating large subgraphs

$$\gamma_t(v_i) = \frac{\exp(\theta_1^T H_t(v_i))}{\sum_{v_j \in \mathcal{N}(s_t)} \exp(\theta_1^T H_t(v_j))}, v_i \in \mathcal{N}(s_t)$$

$$H_t(stop) = \sum_{v_i \in G_s(s_t) \cup \mathcal{N}(s_t)} \gamma_t(v_i) H_t(v_i)$$

Training Objective for flow modeling : Inflow = Outflow

$$\mathcal{L}(\tau) = \sum_{s_{t+1} \in \tau} \left(\sum_{T(s_t, a_t) = s_{t+1}} F(s_t, a_t) - \mathbb{1}_{s_{t+1} = s_f} r(s_f, Y) - \mathbb{1}_{s_{t+1} \neq s_f} \sum_{a_{t+1} \in \mathcal{A}} F(s_{t+1}, a_{t+1}) \right)^2$$

Experiment setup

Datasets:

- Node classification task : BA-shapes, BA-Community, Tree-Cycles/Tree-Grid
- Graph classification task : BA-2motifs, Mutagenicity, Graph-SST2

Baselines: GNNExplainer , PGExplainer, DEGREE, RG-Explainer

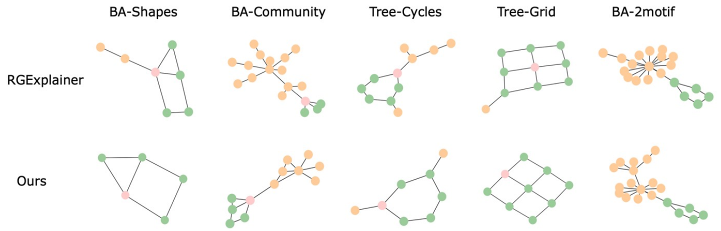
Metrics: AUC score (quantitative analysis) , visualization (qualitative analysis)

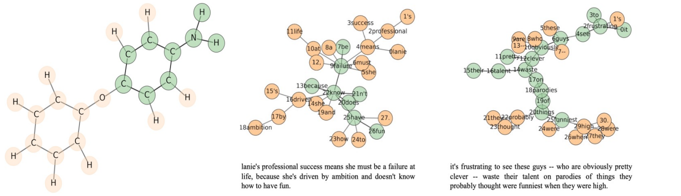
Experiment

Table 1: Explanation AUC (Quantitative Evaluation)

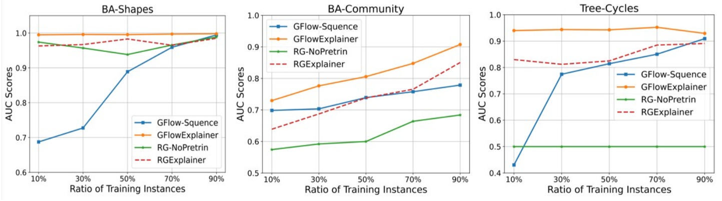
	Node Classification				Graph Classification	
	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid	BA-2motifs	MUTAG
GNNExp	0.742±0.006	0.708±0.004	0.540±0.017	0.714±0.002	0.499±0.001	0.498±0.003
PGExp	0.974±0.005	0.884±0.020	0.574±0.021	0.673±0.004	0.133±0.045	0.843±0.084
DEGREE	0.993±0.005	**0.957±0.010**	0.902±0.022	0.925±0.040	0.755± 0.135	0.773±0.029
RGExp (NoPretrain)	0.983±0.021	0.684±0.012	0.500±0.000	0.500±0.000	0.503±0.011	0.623±0.021
RGExp	0.985±0.013	0.858±0.021	0.787±0.099	0.927±0.030	0.615±0.029	0.832±0.046
Ours	**0.999±0.000**	0.938±0.019	**0.964±0.028**	**0.982±0.011**	**0.854±0.016**	**0.882±0.024**
Improve	1.4%	-2.0%	6.8%	5.9%	13.1%	4.6%

Our method improves the SOTA method by 5% over all !





Our method could identify ground truth structures effectively without many irrelevant edges.



Have better generalizations in the inductive setting. Ablation experiments show the superiority of proposed DAG structure.