Explainability-based Backdoor Attacks Against Graph Neural Networks

Jing Xu[1], Minhui (Jason) Xue[2], Stjepan Picek[1]
[1]Delft University of Technology, The Netherlands
[2]The University of Adelaide, Australia
J.Xu-8@tudelft.nl, jason.xue@adelaide.edu.au, s.picek@tudelft.nl

Abstract

- Backdoor attacks represent a serious threat to neural network models.
- There are already numerous works on backdoor attacks on neural networks, but only a few works considering graph neural networks (GNNs).
- We conduct an experimental investigation on the performance of backdoor attacks on GNNs.
- We apply two powerful GNN explainability approaches to select the optimal trigger injecting position to achieve two attacker objectives – high attack success rate (ASR) and low clean accuracy drop (CAD).

Contributions

- We utilize GNNExplainer, an approach for explaining predictions made by GNNs, to analyze the impact of trigger injecting position for the backdoor attacks on GNNs for the graph classification task.
- We propose a new backdoor attack on GNNs for the node classification task, which uses a subset of node features as a trigger pattern. Additionally, we explore GraphLIME, a local interpretable model explanation for graphs, to explore the proposed backdoor attack.

Background

- GNN takes a graph as an input, including its structure and node feature information, and learns a representation vector(embedding) for each node in the graph.
- Graph-level classification aims to predict the class label for an entire graph.
- Node-level classification attempts to learn a model that identifies the class labels for the unlabeled nodes, given a single graph with partial nodes being labeled and others unlabeled.

Threat Model

- Given a pre-trained GNN model, the adversary forges a backdoored GNN by perturbing its model parameters without modifiying the neural network architecture.
- The attacker has access to a dataset (e.g., graphs or nodes) sampled from the training dataset.

Explainable Backdoor Attacks

- Proposed backdoor attack on graph classification task based on GNNExplainer is shown in Figure 1.
- Proposed backdoor attack on node classification task based on GraphLIME is shown in Figure 2.

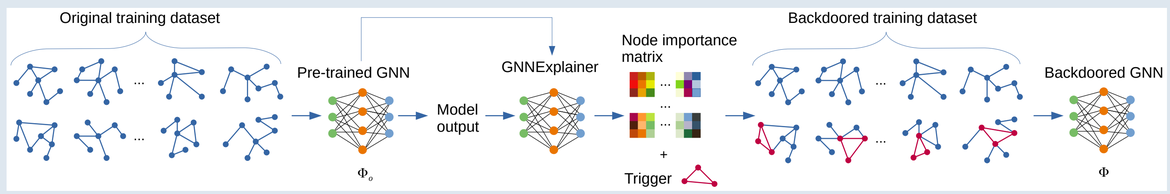

Figure 1: The framework of the backdoor attack on graph classification task based on GNNExplainer.

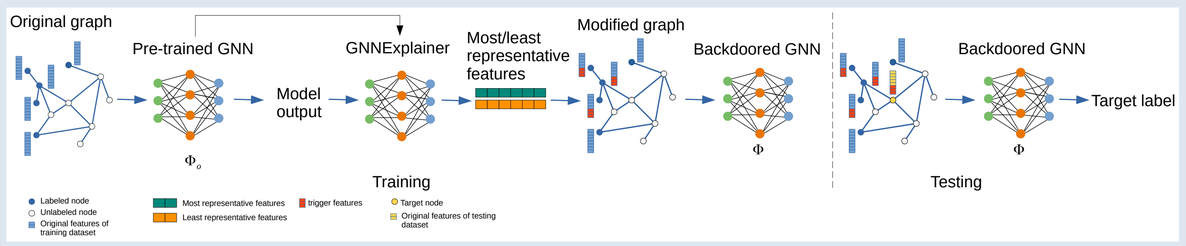

Figure 2: The framework of backdoor attack on node classification task based on GraphLIME.

Results

- For graph classification (Table 1), all backdoor attacks on GIN achieve high attack success rate, while performances in GraphSAGE degrade slightly. The rank between these three attacks is $LIA \approx RSA > MIA$ so the attacker can inject the trigger to the least important structure of the graph. As the trigger size grows, the ASR and CAD of all attacks monotonically increase, as shown in Figure 3.
- For node classification (Table 2), all three backdoor attacks on GAT obtain high ASR, i.e., over 84% and 95% for two datasets. Thus, the attacker can select the least representative features of a node to inject the feature trigger.

Table 1: Graph classification task.

ASR(%) | *CAD*(%)	GIN						GraphSAGE					
	RSA		MIA		LIA		RSA		MIA		LIA	
Mutagenicity	98.24	2.80	93.66	1.66	97.69	2.65	79.73	1.03	73.55	0.58	82.24	0.65
facebook_ct1	100	3.93	95.35	3.32	100	0.52	64.23	2.27	67.22	2.64	69.57	2.85

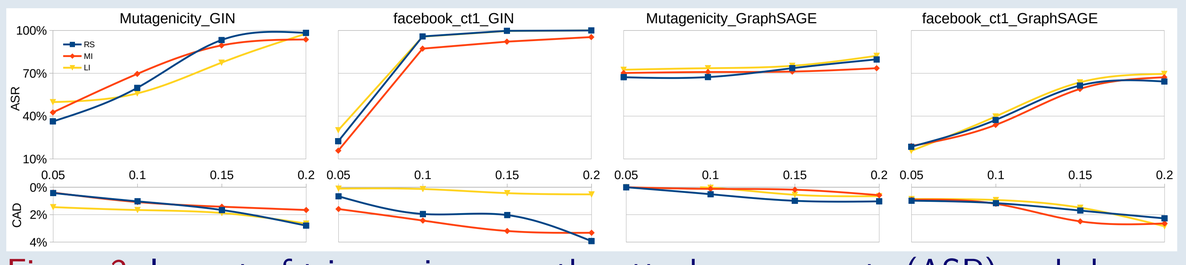

Figure 3: Impact of trigger size γ on the attack success rate (ASR) and clean accuracy drop (CAD) of backdoor attack on the graph classification task.

Table 2: Node classification task.

ASR(%) | *CAD*(%)	GAT					
	RSA		MIA		LIA	
Cora	86.01	2.23	84.11	0.66	84.22	1.95
CiteSeer	96.35	1.72	95.28	1.39	96.26	1.72

Delft University of Technology

THE UNIVERSITY OF ADELAIDE AUSTRALIA