日期: /

Robustness of GNNs

[Adversarial Attack (对抗攻击)]

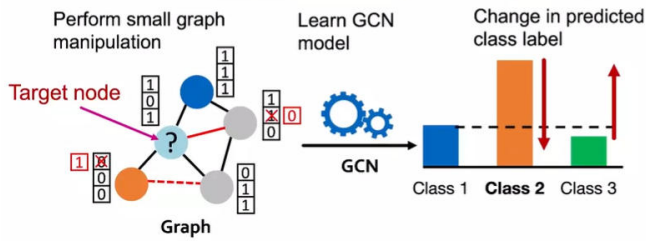
对输入样本进行细微但精心设计的修改，使得模型以高置信度给出错误输出。

对于GNN，有：

1) 直接攻击：目标节点被攻击者直接控制

2) 间接攻击：目标节点未被攻击者控制

Objective：在极小图操作的前提下改变目标节点标签预测



Original graph: A, X Manipulated graph: A', X'

Assumption: $(A, X) \approx (A', X')$

Target node: $v \in V$

GCN learned ove the original graph: $\theta^* = \arg\min_\theta \mathcal{L}_{train}(\theta; A, X)$

GCN's original prediction on the target node: $c_v^* = \arg\max_c f_{\theta^*}(A, X)_{v,c}$

日期: /

GCN learned over the manipulated graph: $\theta^{*'} = \arg\max_\theta \mathcal{L}_{train}(\theta, A', X')$

GCN's prediction on the target node v: $c_v^{*'} = \arg\max_c f_{\theta^{*'}}(A', X')_{v,c}$

$c_v^* \neq c_v^{*'}$

Change of prediction on target node v: $\Delta(v, A', X') = \log f_{\theta^{*'}}(A', X')_{v,c_v^*} - \log f_{\theta^{*'}}(A', X')_{v,c_v^{*'}}$

Two challenges:

① 邻接矩阵 A 为离散对象，基于梯度的方法不可用

② 对 A、X 的每次修改，都需要重新拟合 GNN 进行训练

直接攻击是最有效的攻击类型