



# GNNExplainer: Generating Explanations for Graph Neural Networks

이상용 / 2020-04-03



Computational Data Science LAB



# GNNExplainer: Generating Explanations for Graph Neural Networks

Computational Data Science LAB

목차

1. Introduction
2. Formulating explanations for graph neural networks
3. GNNEXPLAINER
4. Experiments



논의사항 및  
결정사항

관련문서

Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks.  
*In Advances in Neural Information Processing Systems (pp. 9240–9251).*

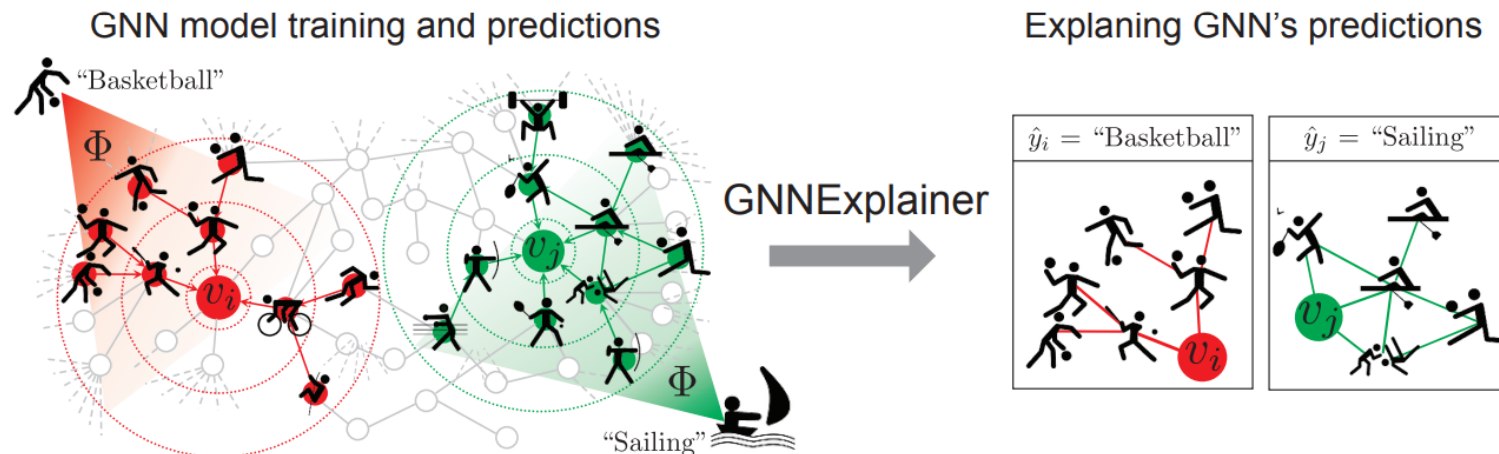


# CONTENTS

1. Introduction
  2. Formulating explanations for graph neural networks
  3. GNNEXPLAINER
  4. Experiments
- 
- 

# 01 | INTRODUCTION

- Graph neural networks (GNN)은 그래프 데이터에 대해 강력한 머신러닝 툴
- GNN은 특정한 노드에 대한 node representation과 그래프의 structural information을 고려하여 학습하는 특성으로 인해, GNN 모델에 대한 해석을 하는 것은 어려움
- 본 논문은 최초의 GNN기반 모델 해석을 위한 model-agnostic 방식의 GNNExplainer를 제안



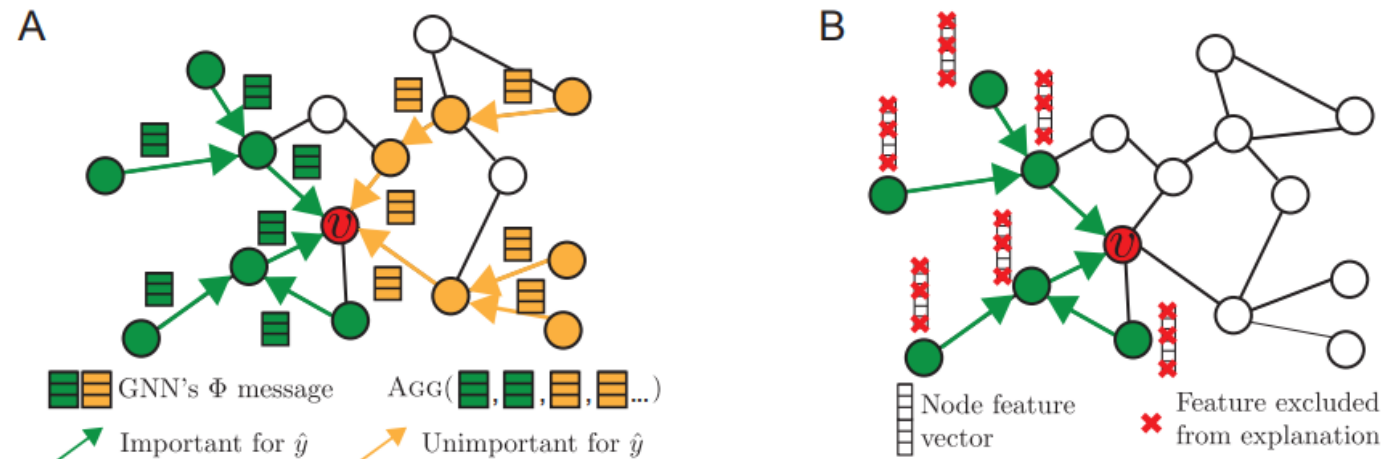
## 02 | Formulating explanations for graph neural networks

- $G$ : graph,  $E$ : edges,  $V$ : nodes,  $\mathcal{X} = \{x_1, \dots, x_n\}$ ,  $x_i \in \mathbb{R}^d$
- $f$ 는 label function으로 노드를 특정한 클래스  $C$ 에 매핑  $f: V \mapsto \{1, \dots, C\}$
- $\Phi$ : optimized GNN model
- Background on graph neural networks
  - ✓ GNN 모델  $\Phi$ 는 세 가지 중요한 computations 특성을 가짐
    1. 모델은 모든 노드 쌍에 대해 message를 계산  $\rightarrow m_{ij}^l = \text{MSG}(\mathbf{h}_i^{l-1}, \mathbf{h}_j^{l-1}, r_{ij})$
    2. 노드  $v_i$ 의 이웃노드  $\mathcal{N}_{v_i}$ 의 메시지  $m_{ij}^l$ 를 모두 aggregate하는 특성  $\rightarrow M_i^l = \text{AGG}(\{m_{ij}^l | v_j \in \mathcal{N}_{v_i}\})$
    3.  $l$ 번째 layer의  $v_i$  feature 정보를 구하기 위해 1,2 의 정보를 사용하여 update  $\rightarrow \mathbf{h}_i^l = \text{UPDATE}(M_i^l, \mathbf{h}_i^{l-1})$
  - ✓ GNNExplainer는 MSG, AGG, UPDATE 세 가지의 computations 특성을 GNN에 대해 explanation 가능 (Model-agnostic)

# 03 | GNNEXPLAINER

## GNNEXPLAINER: Problem formulation

- ✓  $G_c$ : graph,  $A_c(v) \in \{0,1\}^{n \times n}$ : adj,  $X_c(v) = \{x_j | v_j \in G_c(v)\}$ : feature set
- ✓ The GNN model learns a conditional distribution  $P_\phi(Y|G_c, X_c)$ ,  $\hat{y} = \phi(G_c(v), X_c(v))$ : prediction
- ✓ GNNEXPLAINER는  $(G_S, X_S^F)$ 로  $\hat{y}$ 의 해석을 제공함
- ✓  $G_S$  : small subgraph,  $X_S^F$  : small subset of node feature (i.e.,  $X_S^F = \{x_j^F | v_j \in G_S\}$ )



## 03 | GNNEXPLAINER

- ✓ 노드  $v$ 가 주어졌을 때, GNN의 예측  $\hat{y}$ 에 중요한 영향을 준 서브그래프  $G_S \subset G_c$ 와 features  $X_S^F = \{x_j^F | v_j \in G_S\}$ 를 찾는 것이 목적
- ✓ 본 논문은 importance의 개념을 mutual information  $MI$ 로 공식화 하고, GNNEXPLAINER를 optimization framework로 공식화 함

$$\max_{G_S} \overset{\text{전체 그래프}}{MI(Y, (G_S, X_S))} = H(Y) - \overset{\text{서브 그래프}}{H(Y|G = G_S, X = X_S)}. \quad \begin{array}{l} P_{east} = 0.99 \rightarrow H(east): \text{높음} \\ P_{west} = 0.01 \rightarrow H(west): \text{낮음} \end{array}$$

- ✓  $MI$ 는 서브그래프  $G_S$ 와 서브피쳐  $X_S^F$ 로 제한되었을 때  $\hat{y}$ 의 확률의 변화를 정량화
- ✓  $G_c$ 에서  $v_i$ 의  $\hat{y}$ 를 예측하는데 어떤 노드  $v_j$ 를 지웠을 때 확률이 강하게 감소한다면,  $v_j$ 는  $v_i$ 의 예측에 대한 좋은 설명
- ✓ 엔트로피 term  $H(Y)$ 는  $\phi$ 가 고정 되어 있기 때문에 상수
- ✓ 즉, 위 식을 maximization 하는 것은  $H(Y|G = G_S, X = X_S)$ 를 minimize 하는 것과 같음

# 03 | GNNEXPLAINER

## GNNEXPLAINER's optimization framework

$$H(Y|G=G_S, X=X_S) = -\mathbb{E}_{Y|G_S, X_S} [\log P_{\Phi}(Y|G=G_S, X=X_S)].$$

엔트로피 평균 정보량

- ✓ Compact explanation을 위해  $G_S$ 의 사이즈를 제한할 수 있음 :  $|G_S| \leq K_M$
- ✓ 위 목적함수를 바로 최적화 하는 것은 어려움  $\rightarrow G_c$ 에 대한 모든 가능한  $G_S$  후보군을 고려해야 하기 때문
- ✓ Computationally efficient version of GNNEXPLAINER's objective, which we optimize using gradient descent, is as follows:

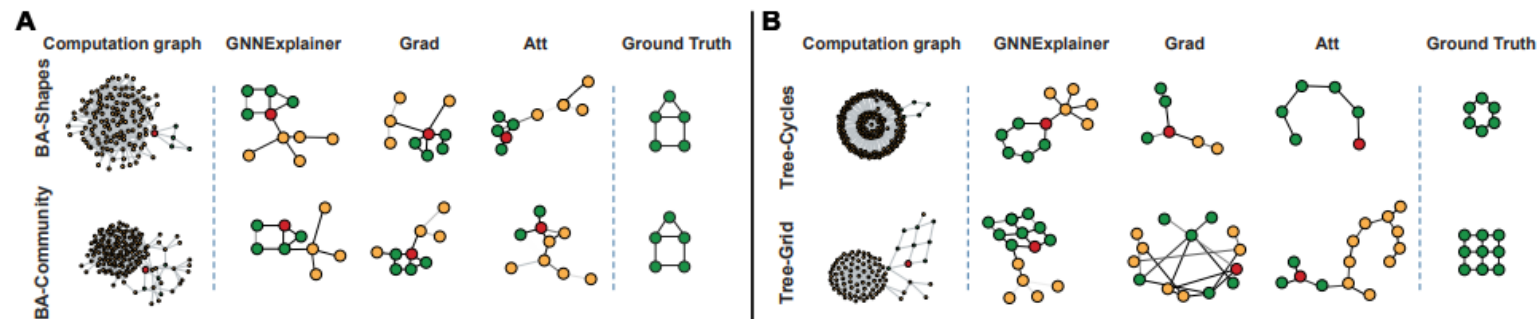
$$\min_M - \sum_{c=1}^C \mathbb{1}[y=c] \log P_{\Phi}(Y=y|G=\underbrace{A_S \odot \sigma(\underbrace{M}_{\text{Fractional adjacency matrix}})}_{\text{Threshold로 low-value는 cut}}, X=X_S),$$

- ✓  $M \in \mathbb{R}^{n \times n}$  : 학습해야하는 mask,  $\sigma : mask$ 를  $[0,1]^{n \times n}$ 으로 매핑하는 *sigmoid* 함수
- ✓ 어떤 변수가 결과에 중요한 영향을 미쳤는지 파악할 때에도 위와 같은 방식으로 진행
- ✓ Binary feature selector  $F \in \{0,1\}^d$  사용  $\rightarrow X_S^F$  as  $X_S \odot F$

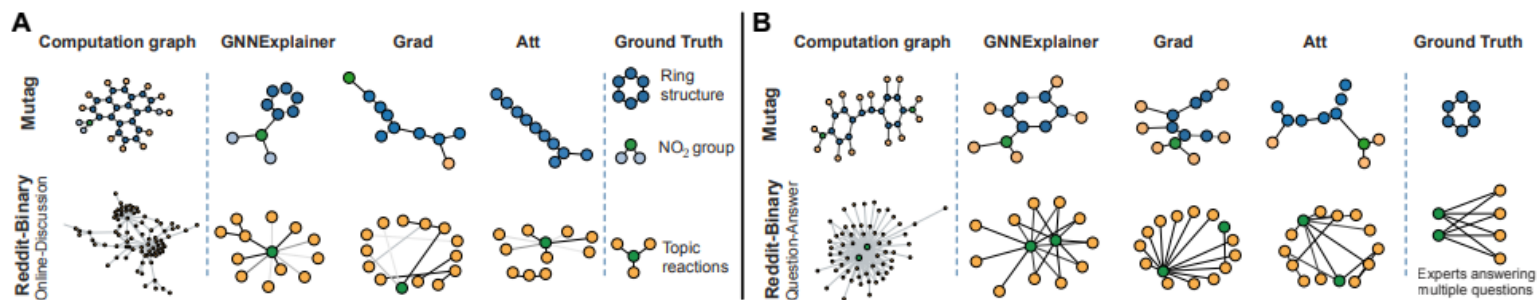


# 04 | EXPERIMENTS

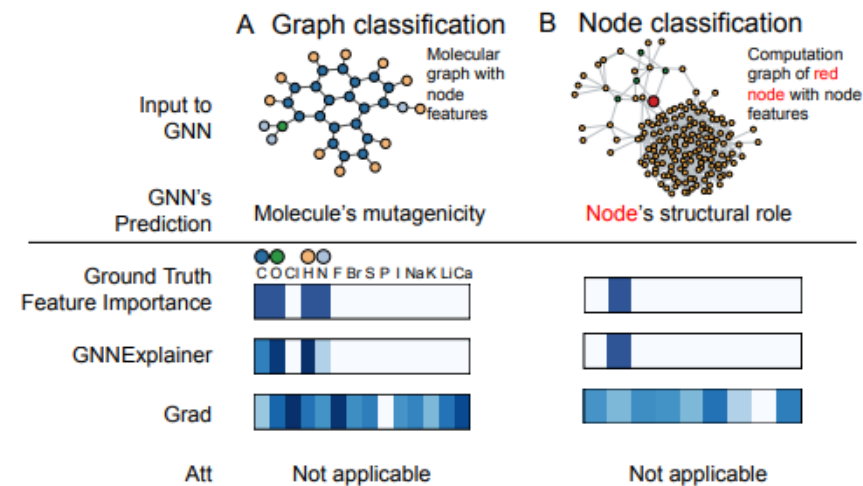
- 인조데이터 - 중요한 subgraphs structure



- 실제데이터 - 중요한 subgraphs structure



- 중요변수



# Q&A

---

감사합니다.