

---

# GNNExplainer: Generating Explanations for Graph Neural Networks

---

School of Industrial and Management Engineering, Korea University

Lee Kyung Yoo

# Contents

---

- ❖ Research Purpose
- ❖ GNNExplainer
- ❖ Experiments
- ❖ Conclusion

# Research Purpose

---

- ❖ GNNExplainer: Generating Explanations for Graph Neural Networks (NeurIPS 2019)
  - Stanford University에서 연구하였으며 2022년 3월 24일 기준으로 285회 인용

---

## GNNExplainer: Generating Explanations for Graph Neural Networks

---

Rex Ying<sup>†</sup>   Dylan Bourgeois<sup>†,‡</sup>   Jiaxuan You<sup>†</sup>   Marinka Zitnik<sup>†</sup>   Jure Leskovec<sup>†</sup>

<sup>†</sup>Department of Computer Science, Stanford University

<sup>‡</sup>Robust.AI

{rexying, dtsbourg, jiaxuan, marinka, jure}@cs.stanford.edu

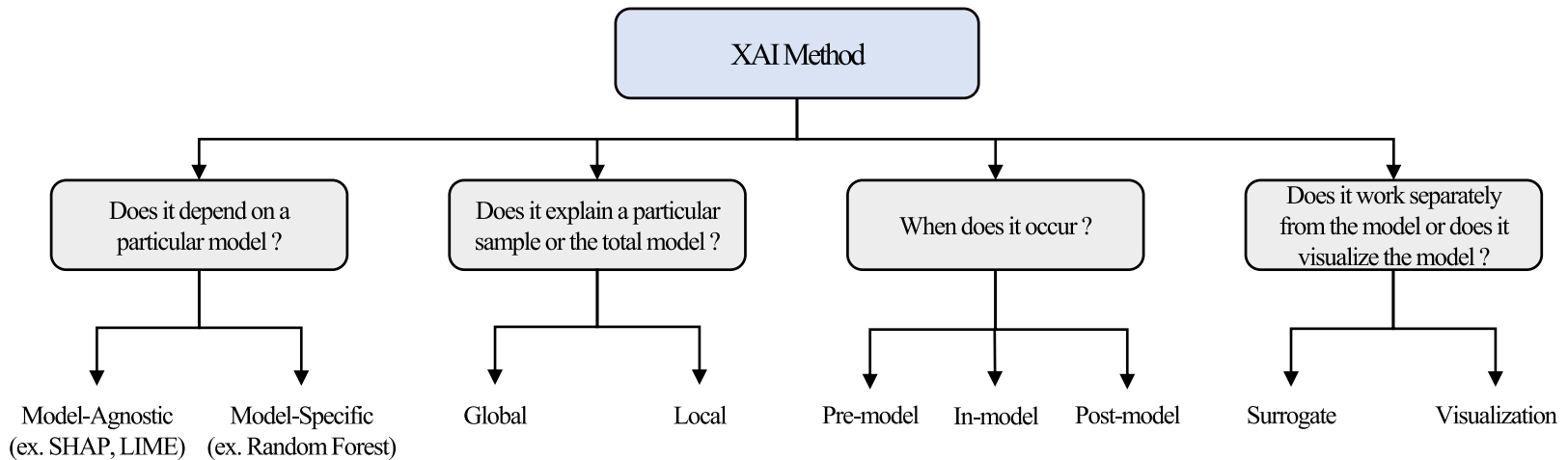
### Abstract

Graph Neural Networks (GNNs) are a powerful tool for machine learning on graphs. GNNs combine node feature information with the graph structure by recursively passing neural messages along edges of the input graph. However, incorporating both graph structure and feature information leads to complex models and explaining predictions made by GNNs remains unsolved. Here we propose GNNEXPLAINER, the first general, model-agnostic approach for providing interpretable explanations for predictions of any GNN-based model on any graph-based machine learning task. Given an instance, GNNEXPLAINER identifies a compact subgraph structure and a small subset of node features that have a crucial role in GNN's prediction. Further, GNNEXPLAINER can generate consistent and concise explanations for an entire class of instances. We formulate GNNEXPLAINER as an optimization task that maximizes the mutual information between a GNN's prediction and distribution of possible subgraph structures. Experiments on synthetic and real-world graphs show that our approach can identify important graph structures as well as node features, and outperforms alternative baseline approaches by up to 43.0% in explanation accuracy. GNNEXPLAINER provides a variety of benefits, from the ability to visualize semantically relevant structures to interpretability, to giving insights into errors of faulty GNNs.

# Research Purpose

## ❖ Explainable AI (XAI)

- 모델에 대한 해석을 제공하는 방법론
  - 모델의 신뢰성 및 투명성 향상으로 의사결정에 도움
  - 네트워크의 특징을 쉽게 이해하여 모델 내 문제를 쉽게 식별 가능
- 뉴럴 네트워크에 적용되는 다양한 방법론 존재
  - 특정한 모델에 국한되지 않는 model-agnostic approach(ex. SHAP, LIME)가 많이 활용되는 추세

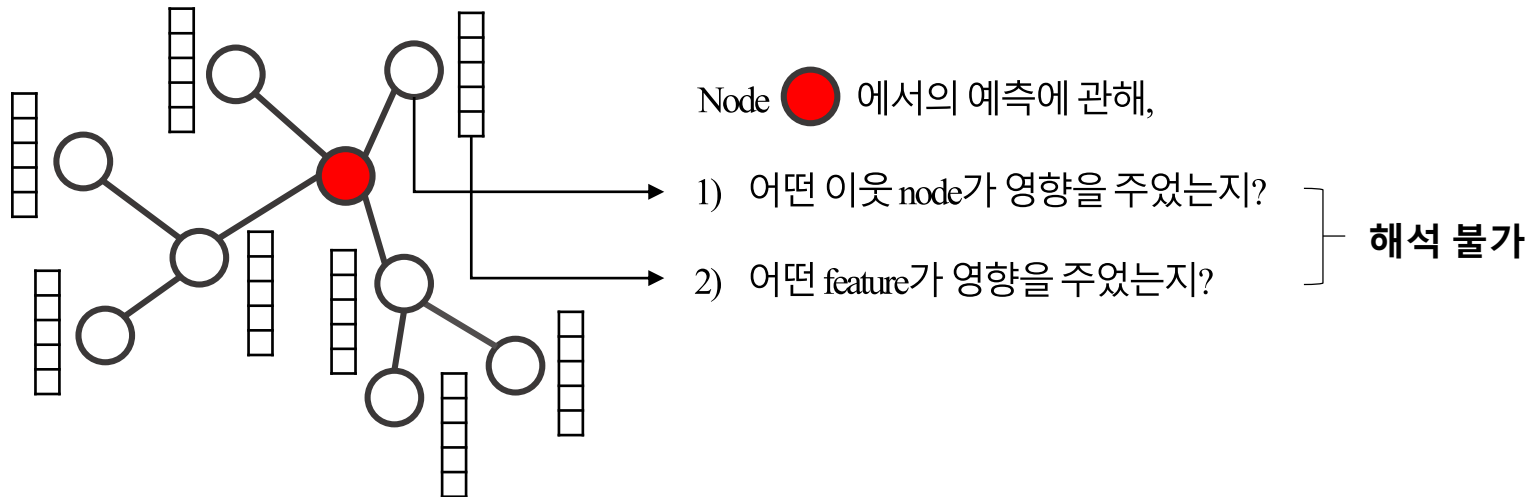


### < XAI Method Taxonomy >

# Research Purpose

## ❖ Explainable AI (XAI) for Graph Neural Networks

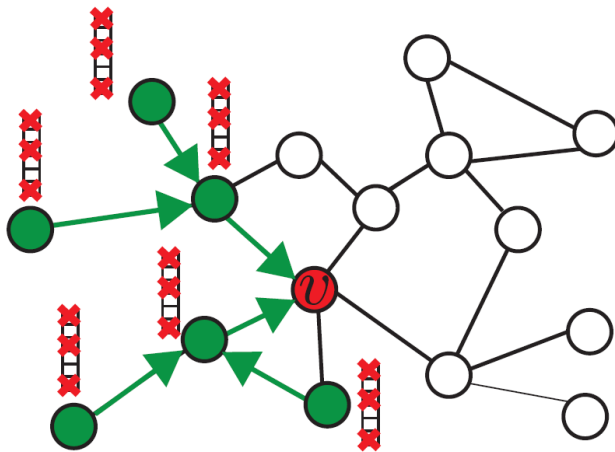
- 이를 위해 기존 XAI 방법론을 그대로 적용하기에는 역부족
  - 그래프에서는 데이터들 간 dependency가 존재하여, 관계에 대한 의미를 반영할 수 있어야 함
  - 따라서 어떤 feature가 중요한지와 함께 어떤 이웃 node가 중요한지 관계성을 판단할 수 있어야 함
- 그럼에도 그래프 분야에 적용 가능한 XAI 연구 부진



# GNNExplainer

## ❖ Overview

- GNN 모델에 대하여 agnostic한 XAI 방법론을 고안한 최초의 시도
- Message / Aggregate / Update 단계를 거치는 보편적 모델에 대해서 설명할 수 있는 기법
- 해석 목표
  - 기존 그래프 내에서 예측에 주요한 영향을 미치는 graph structural information (= **subgraph**) 과 node feature information (= **feature**) 으로 구성된 작은 집합을 식별해내고자 함



↗ : Important for prediction at node  $v$

→ 어떤 이웃 node가 영향을 주었는지 해석 가능

□ : Node feature vector

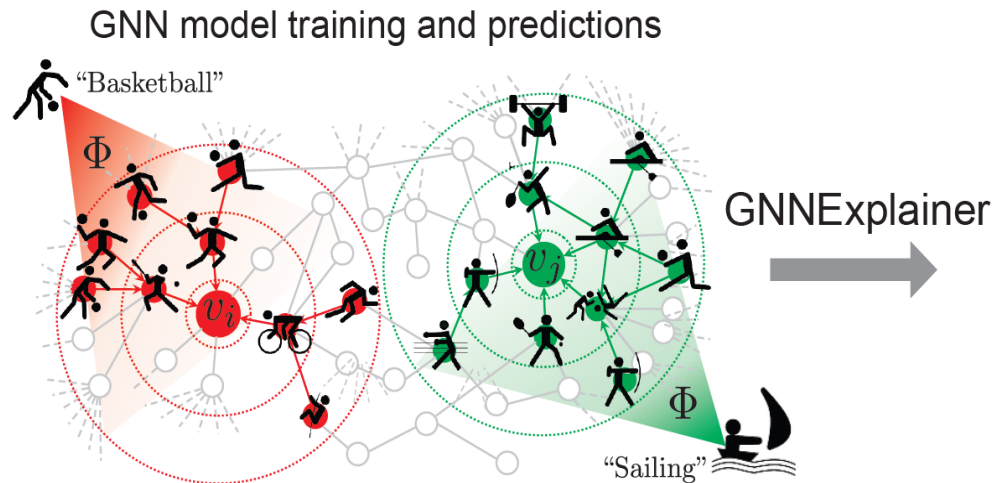
✗ : Feature excluded from explanation

→ 어떤 feature가 영향을 주었는지 해석 가능

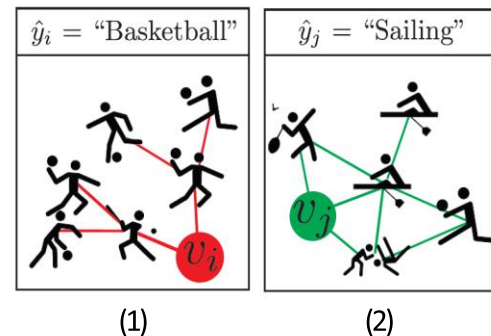
# GNNExplainer

## ❖ Overview

- 실제 예측 문제에 대한 subgraph 예시
- GNN model( $\Phi$ )이 학습 및 예측을 통해서 노드  $v_i$ 와  $v_j$ 에 대해서 node classification을 진행
  - (1)  $v_i$ 의 class를 basketball로 분류할 때 영향을 준 subgraph
  - (2)  $v_j$ 의 class를 sailing으로 분류할 때 영향을 준 subgraph



## Explaining GNN's predictions



“Subgraphs”

# GNNExplainer

## How to find the **Subgraph**?

### ❖ Initial objective function

- Computation Graph  $G_C$  내에 포함되어 있는 Subgraph  $G_S$  중, **mutual information**이 가장 큰  $G_S$ 를 선택
- Mutual information을 최대화하는 목적 함수를 기반으로 학습
  - Mutual information의 일반적 정의는 두 변수의 상호 종속 여부 = 예측값  $Y$ 와 Subgraph  $G_S$ ,  $X_S$ 와의 연관성
  - 아래 식과 같이 예측값에 대한 엔트로피의 차를 이용하여 계산

Computation Graph  $G_C$ 의 Subgraph

Subgraph  $G_S$ 에 포함되어 있는 Feature

$$\max_{G_S} MI(Y, (G_S, X_S)) = \boxed{H(Y)} - \boxed{H(Y|G = G_S, X = X_S)}$$

모델 예측값      Full computation graph (고정값)      Subgraph (변동값)

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

GNN 모델  $\Phi$ 의 예측에 대한 불확실성을 줄이고자 함



# GNNExplainer

## How to find the Subgraph?

### ❖ Optimization problem

- Computation Graph  $G_C$  내에 포함되어 있는 Subgraph  $G_S$ 의 후보가 매우 많아, 최적화 불가
- Random graph  $\mathcal{G}$ 로부터 샘플링한  $G_S$ 의 기댓값을 활용하는 문제로 변경
- Mean-Field Variational Approximation를 기반으로 **adjacency matrix에 masking을 적용하도록 함**

$$\min H(Y|G = G_S, X = X_S)$$

$\mathcal{G}$ 로부터 샘플링한  $G_S$ 의 엔트로피 기댓값으로 대체

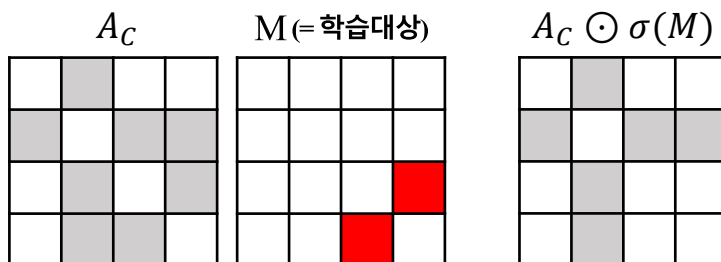
$$\min_{\mathcal{G}} E_{G_S \sim \mathcal{G}} H(Y|G = G_S, X = X_S) \leq \min_{\mathcal{G}} H(Y|G = E_{\mathcal{G}}[G_S], X = X_S)$$

\*Jensen's Inequality :  $E[\log y] \leq \log E[y]$

Mean-Field Variational Approximation를 기반으로  
기댓값을 구하는 과정을 Masking으로 대체

$$\min_{\mathcal{G}} H(Y|G = A_C \odot \sigma(M), X = X_S)$$

⏟  
sigmoid



## How to find the **Subgraph**?

### ❖ Final objective function

- 최종 objective function에 따라 학습을 통해 **adjacency mask generation algorithm 업데이트**
  - 하나의 예측값을 가장 잘 설명하는 subgraph 탐색 시
  - 모델의 confidence 관점에서 활용

$$\min_G H(Y|G = A_C \odot \sigma(M), X = X_S)$$

- 특정 레이블로 예측한 것을 가장 잘 설명하는 subgraph 탐색 시
- 레이블과의 직접적인 비교 관점에서 활용, cross-entropy 적용하여 계산

$$\min_M - \sum_{c=1}^C \mathbb{1}[y = c] \log P_{\Phi}(Y = y|G = A_C \odot \sigma(M), X = X_c)$$

# GNNExplainer

## How to find the **Feature**?

### ❖ Final objective function

- Graph structure에서 adjacency matrix에 적용했던 방식과 동일하게 masking 진행
- 어떠한 node feature가 예측에 주요한 영향을 끼쳤는지 파악하고자, feature selector  $F$  학습
  - $G_S$  : Subgraph에 포함되어 있는 nodes
  - $X_S$  : Subgraph node features  $\rightarrow X_S^F$  : Feature selector  $F \in \{0,1\}^d$  로 선택된 Subgraph node features

$$\max_{G_S} MI(Y, (G_S, X_S)) = H(Y) - H(Y|G = G_S, X = X_S)$$



$$\max_{G_S, F} MI(Y, (G_S, F)) = H(Y) - H(Y|G = G_S, X = X_S^F)$$

Node Features  $X_S$  :

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
-------	-------	-------	-------	-------	-------

Masked Node Features  $X_S^F$  :

	$x_2$	$x_3$		$x_5$	
--	-------	-------	--	-------	--

선택된 정보에 대해서만 학습


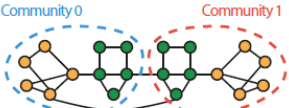






# Experiments

## ❖ Experiment Settings

### • Datasets

- Node classification과 graph classification task에 대하여 실험 진행함에 따라 두 가지 부류의 데이터셋 활용
- Node classification
  - ✓ Synthetic datasets (BA-Shapes / BA-Community / Tree-Cycles / Tree-Grid : 4가지 인위적인 네트워크)
  - ✓ Ground-truth = motif
- Graph classification
  - ✓ Real-world datasets (Mutag : 분자그래프 / Reddit-Binary : Reddit 질의응답)
  - ✓ Ground-truth = 사전에 지정

### Synthetic datasets

	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid
Base				
Motif				
Node Features	None	$\mathcal{N}(\mu_l, \sigma_l)$ where $l$ = community ID	None	None
Explanation content	Graph structure	Graph structure Node feature information	Graph structure	Graph structure

# Experiments

---

## ❖ Experiment Settings

- Baseline

- Grad

- ✓ GNN 모델이 convolution을 사용하기에 기존 convolution based XAI를 사용
    - ✓ GNN loss function의 gradient를 adjacency matrix에서 계산해내는데 saliency map을 구성하는 것과 동일


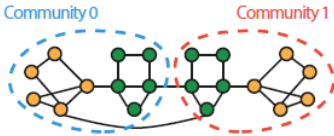





- Att

- ✓ Self attention을 graph에 적용하는 Graph Attention Network(GAT)를 통해 attention weight 구성
    - ✓ 이웃들의 영향력을 동일하게 보지 않고, 중요한 이웃 노드에 가중치 부여
    - ✓ Computation graph의 feature에 대한 가중치는 알 수 없음

# Experiments

## ❖ Results

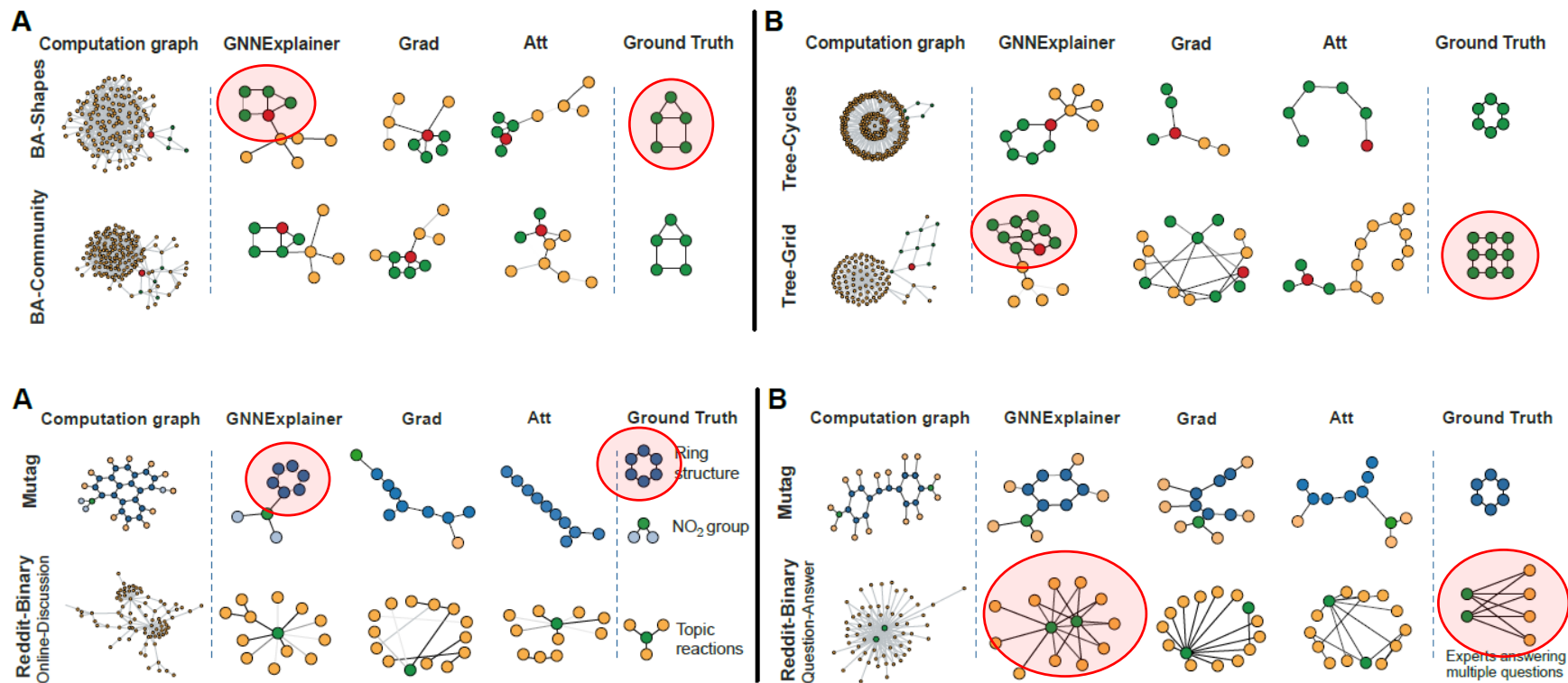
- Node classification task 실험에서는 4가지 synthetic dataset에 대하여 ground-truth label을 예측하는 binary classification으로 문제를 정의 후 평가
  - GNNExplainer를 통해 ground-truth explanation에 포함되어 있는 edge를 레이블로서 예측
- GNNExplainer가 다른 방법론에 비해 더 정확히 ground-truth edge를 예측함

	BA-Shapes	BA-Community	Tree-Cycles	Tree-Grid
Base				
Motif				
Node Features	None	$\mathcal{N}(\mu_l, \sigma_l)$ where $l$ = community ID	None	None
Explanation content	Graph structure	Graph structure Node feature information	Graph structure	Graph structure
Explanation accuracy				
Att	0.815	0.739	0.824	0.612
Grad	0.882	0.750	0.905	0.667
GNNExplainer	<b>0.925</b>	<b>0.836</b>	<b>0.948</b>	<b>0.875</b>

# Experiments

## ❖ Results

- 세 가지 방법론에 의해 도출한 Subgraph를 직접 ground truth와 비교
- GNNExplainer가 Grad와 Att보다 ground-truth에 더 근접하게 subgraph를 도출해냄



# Conclusion

---

## ❖ Conclusion

- 어떠한 GNN model, 어떠한 graph task에도 적용이 가능한 XAI 방법론을 고안
- Adjacency mask generation algorithm을 기반으로 중요한 subgraph 탐색
- 탐색한 subgraph와 node feature를 통해 graph structural & node feature information이 해석 가능함에 따라 그래프에 보다 적합한 XAI 방법론임을 입증
- 단일 노드에 관한 해석은 다각도로 가능하지만, 여러 노드에 대한 해석은 future work 존재
- 그럼에도 GNN XAI의 시초로서 이론 및 실험에 대한 baseline을 확립
- 이후 연구된 GNN XAI 방법론을 이해하는데 중요한 근간이 될 논문이라고 생각됨



# Conclusion

---

## ❖ Reference

- Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.
- Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 52.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).
- Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *stat*, 1050, 20.
- <http://dsba.korea.ac.kr/seminar/?mod=document&uid=1443>
- <https://youtu.be/NvDM2j8Jgvk>

*Thank You*