GNN Explainability

Shichang Zhang

11.02.2021

Roadmap

- Model Explainability
  - Motivating Examples for Images and Tabular Data
- GNN Explainability (Graph Data Explainability)
  - Graphs vs. Images vs. Tabular Data
  - SubgraphX (ICML 2021)

Model Explainability

- Goal: understand black-box models, e.g. NNs.
- Existing approaches
    - Instance-level
        - Example-specific understanding, why an input data is mapped to a certain output
    - Model-level
        - High-level generic understanding, how the model mechanism leads to a certain output

Motivating Examples: Images

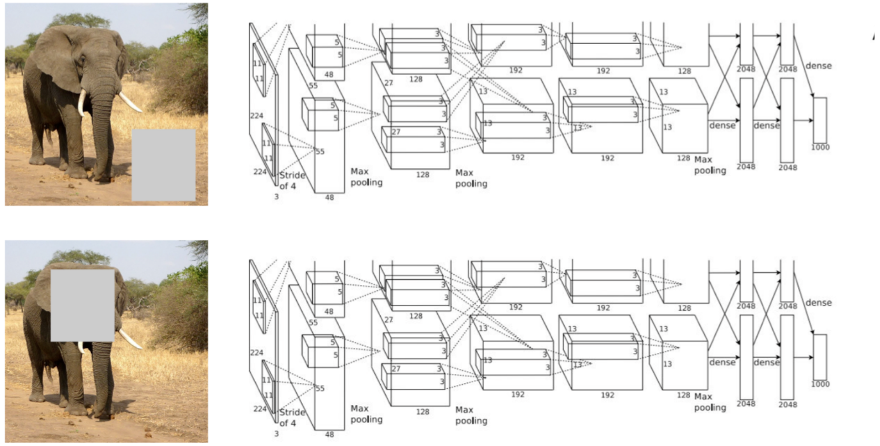
- Instance-level: Which pixels are importance for classifying an image

Motivating Examples: Images

- Instance-level: Which pixels are importance for classifying an image

Which pixels matter: Saliency via Occlusion

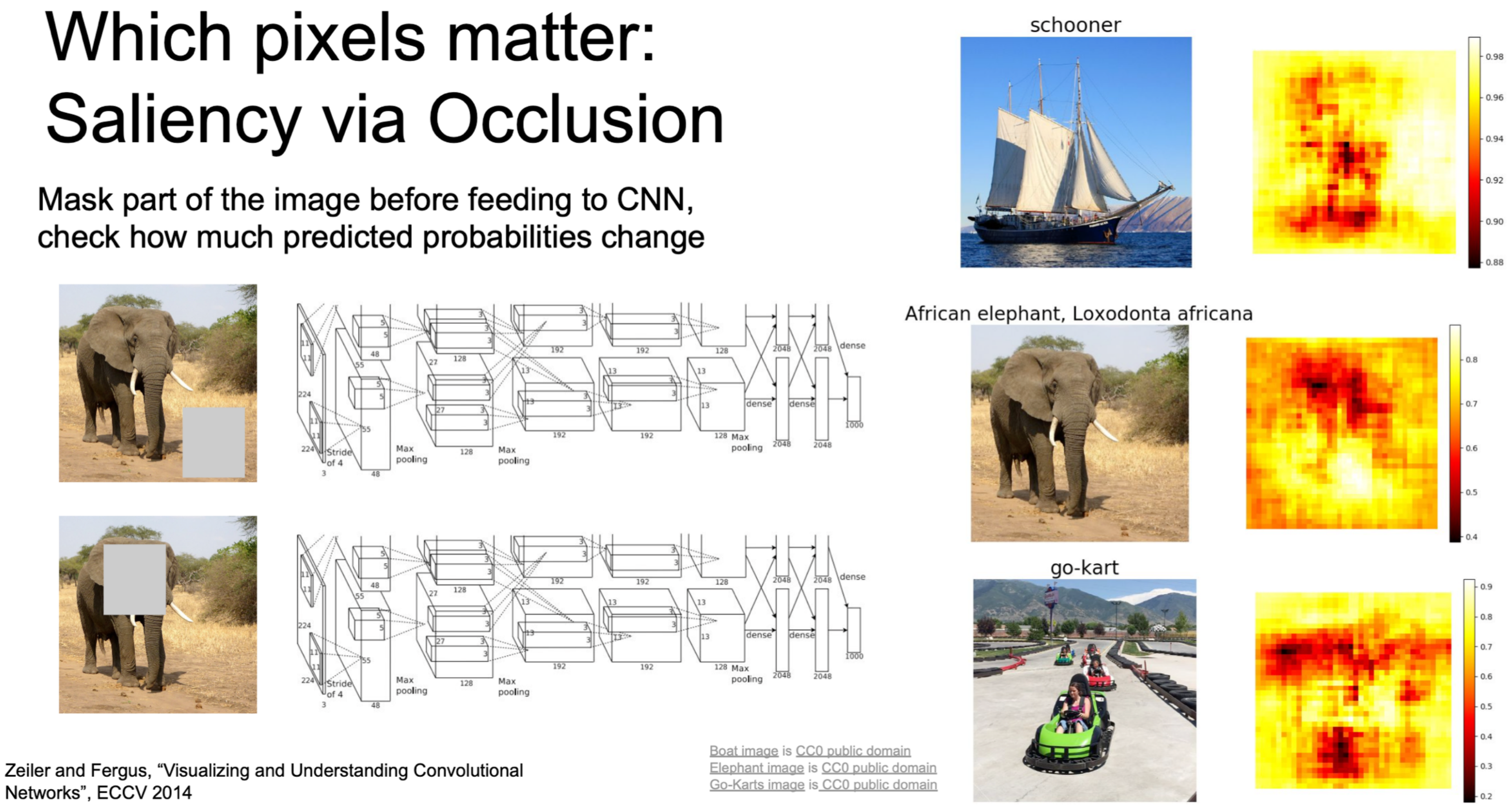
Mask part of the image before feeding to CNN, check how much predicted probabilities change



Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014

Motivating Examples: Images

- Instance-level: Which pixels are importance for classifying an image



Motivating Examples: Images

- Model-level: What kind of image maximizes the probability for a certain class

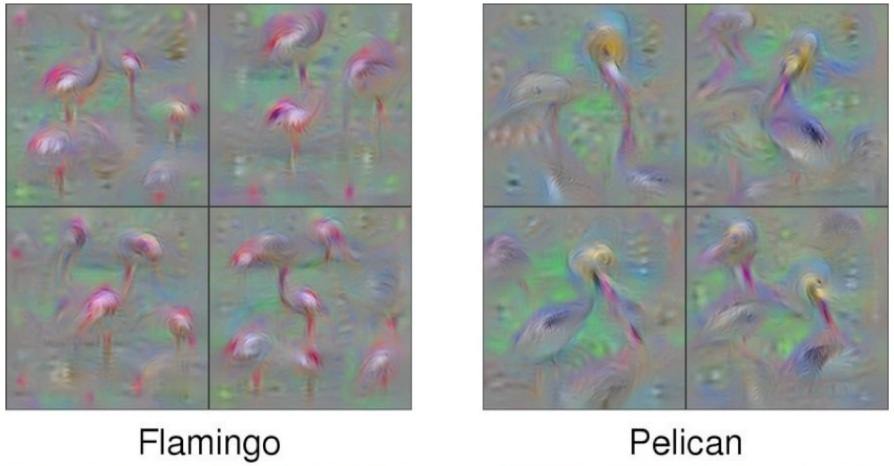
Motivating Examples: Images

- Model-level: What kind of image maximizes the probability for a certain class
  - Treat pixels in the image as parameters and tune them with gradient ascent

Motivating Examples: Images

- Model-level: What kind of image maximizes the probability for a certain class
  - Treat pixels in the image as parameters and tune them with gradient ascent

Random Initialization



Synthesized

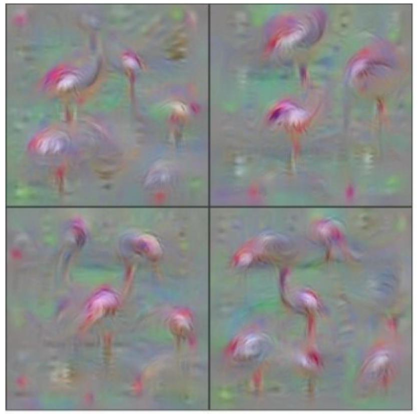
Yosinski et al, “Understanding Neural Networks Through Deep Visualization”, ICML DL Workshop 2014.

Nguyen et al, “Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks”, ICML Visualization for Deep Learning Workshop 2016.

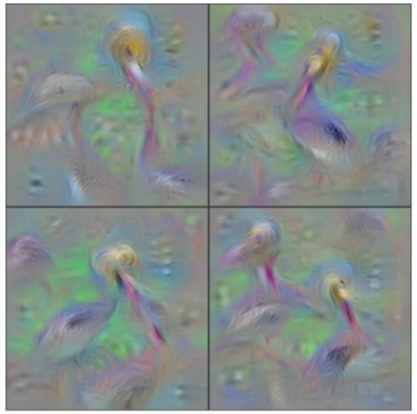
Motivating Examples: Images

- Model-level: What kind of image maximizes the probability for a certain class
  - Treat pixels in the image as parameters and tune them with gradient ascent

Random Initialization



Flamingo



Pelican

Synthesized

Smart Initialization considering multimodality.
The "grocery store" class

Synthesized

Ground Truth

Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014.
Figure copyright Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, 2014.
Nguyen et al, "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks", ICML Visualization for Deep Learning Workshop 2016. Figures copyright Anh Nguyen, Jason Yosinski, and Jeff Clune, 2016;

Motivating Examples: Tabular Data

- A feature selector for both instance-level and model-level

Motivating Examples: Tabular Data

- A feature selector for both instance-level and model-level
  - Usually involves a score function to assign each feature an importance score

Motivating Examples: Tabular Data

- A feature selector for both instance-level and model-level
  - Usually involves a score function to assign each feature an importance score
  - Instance-level

Figure credit: https://github.com/slundberg/shap

Motivating Examples: Tabular Data

- A feature selector for both instance-level and model-level
  - Usually involves a score function to assign each feature an importance score
  - Instance-level

  - Model-level
    - Ex. a simple linear model

y = b1*Age + b2*Sex + b3*BP + b4*BMI

Figure credit: https://github.com/slundberg/shap

GNN Explainability

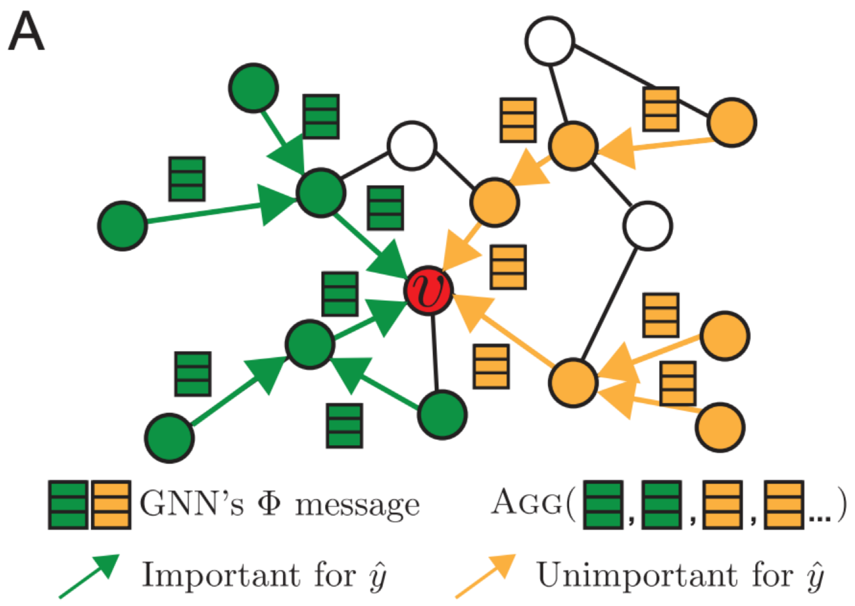
- Instance-level

GNN Explainability

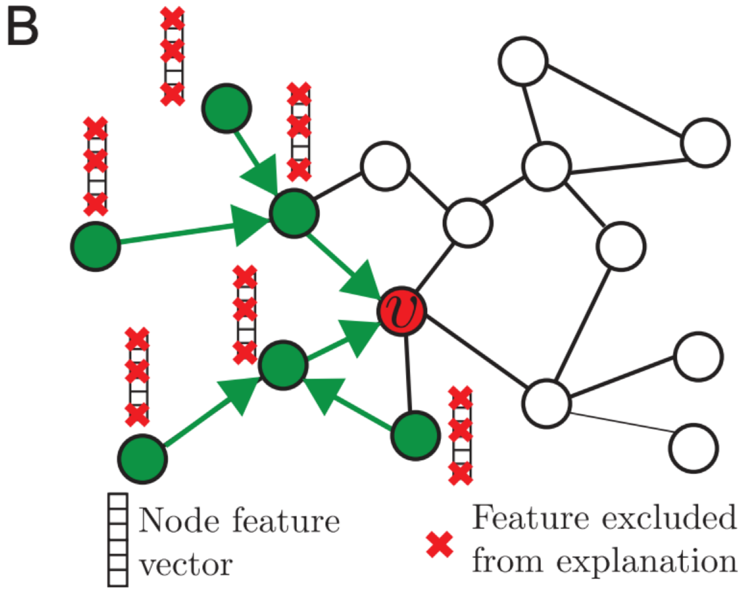
- Instance-level
    - Node classification

GNN Explainability

- Instance-level
  - Node classification



Neighbor node selection



Node feature selection

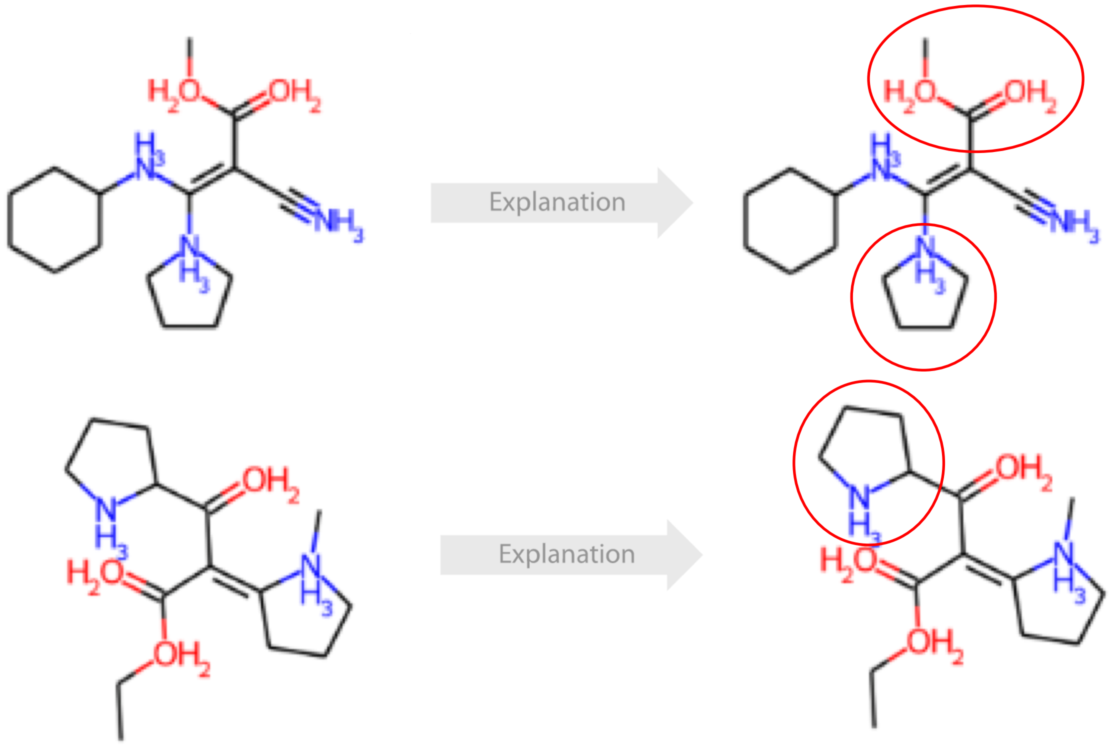
Figure credit: Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, *32*, 9240.

GNN Explainability

- Instance-level
  - Graph classification

GNN Explainability

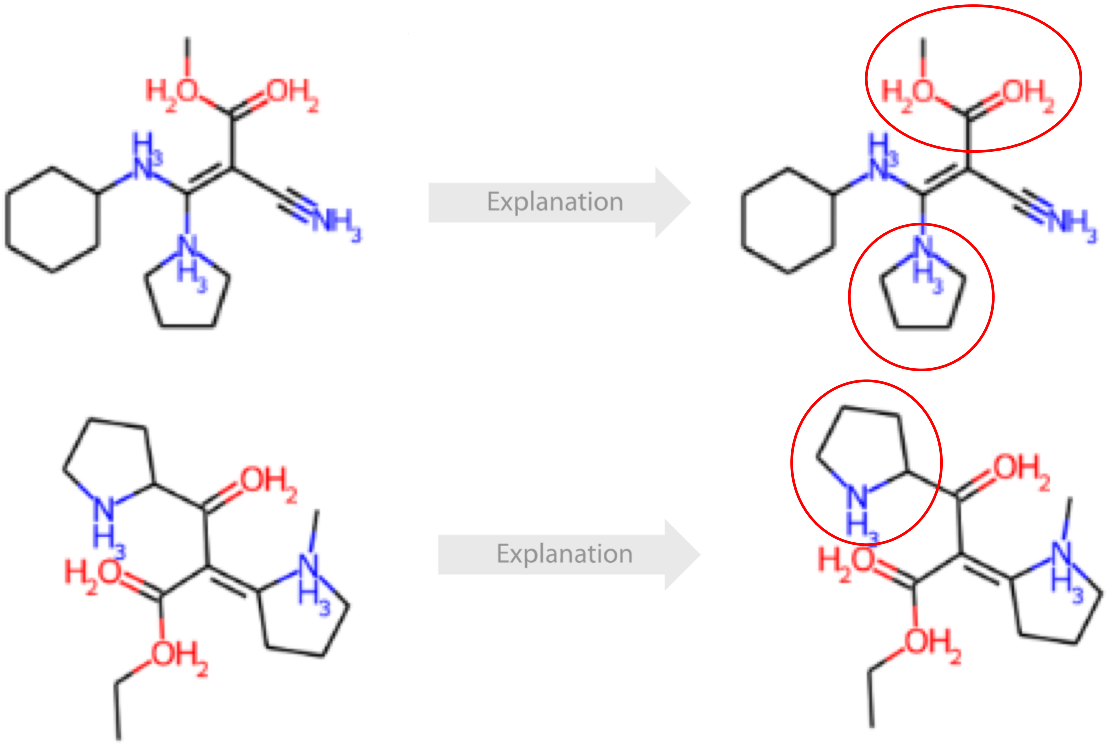
- Instance-level
  - Graph classification



Important subgraph selection

GNN Explainability

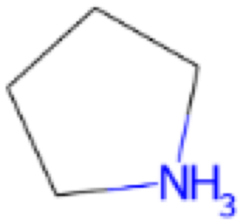
- Instance-level
  - Graph classification



Important subgraph selection

Model-level

(data-level)



GNN Explainability (Graph Data Explainability)

GNN Explainability (Graph Data Explainability)

- Explanation of graphs is more meaningful than images

GNN Explainability (Graph Data Explainability)

- Explanation of graphs is more meaningful than images
  - Image explanation is mostly for model understanding and debugging

GNN Explainability (Graph Data Explainability)

- Explanation of graphs is more meaningful than images
  - Image explanation is mostly for model understanding and debugging
  - Graph explanation may reveal useful knowledge

GNN Explainability (Graph Data Explainability)

- Explanation of graphs is more meaningful than images
  - Image explanation is mostly for model understanding and debugging
  - Graph explanation may reveal useful knowledge
    - Data-level graph classification explanation is similar to frequent pattern mining.

GNN Explainability (Graph Data Explainability)

- Explanation of graphs is more meaningful than images
  - Image explanation is mostly for model understanding and debugging
  - Graph explanation may reveal useful knowledge
    - Data-level graph classification explanation is similar to frequent pattern mining.
- Explanation of graphs is more challenging than tabular data

GNN Explainability (Graph Data Explainability)

- Explanation of graphs is more meaningful than images
  - Image explanation is mostly for model understanding and debugging
  - Graph explanation may reveal useful knowledge
    - Data-level graph classification explanation is similar to frequent pattern mining.
- Explanation of graphs is more challenging than tabular data
  - Graphs as tabular data with structure information

SubgraphX (ICML21)

On Explainability of Graph Neural Networks via Subgraph Explorations

Hao Yuan [1] **Haiyang Yu** [1] **Jie Wang** [2] **Kang Li** [3] **Shuiwang Ji** [1]

- Instance level
- Graph classification

Problem Formulation

- Goal: Identify the most important subgraph for classifying a graph

Problem Formulation

- Goal: Identify the most important subgraph for classifying a graph
- Notations

$f(\cdot)$ GNN to be explained

$\mathcal{G}$ Input graph

$\{\mathcal{G}_1, \cdots, \mathcal{G}_i, \cdots, \mathcal{G}_n\}$ All connected subgraphs

$\mathcal{G}^*$ The most important subgraph

Problem Formulation

- Goal: Identify the most important subgraph for classifying a graph
- Notations

$f(\cdot)$ GNN to be explained

$\mathcal{G}$ Input graph

$\{\mathcal{G}_1, \cdots, \mathcal{G}_i, \cdots, \mathcal{G}_n\}$ All connected subgraphs

$\mathcal{G}^*$ The most important subgraph

- Objective

$$\mathcal{G}^* = \underset{|\mathcal{G}_i| \leq N_{\min}}{\operatorname{argmax}} \ \mathrm{Score}(f(\cdot), \mathcal{G}, \mathcal{G}_i)$$

Challenges

- There are too many subgraphs. How can we explore them?
- What is a reasonable score function?

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Tree construction: start with the input graph and take pruning actions

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Tree construction: start with the input graph and take pruning actions
  - Each node $\mathcal{N}_i$ in the MCT represents a subgraph $\mathcal{G}_i$
  - Root node $\mathcal{N}_0$ represents the input whole graph $\mathcal{G}$
  - Edge from a parent to its child represents a prune action a_j

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Tree construction: start with the input graph and take pruning actions
  - Each node $\mathcal{N}_i$ in the MCT represents a subgraph $\mathcal{G}_i$
  - Root node $\mathcal{N}_0$ represents the input whole graph $\mathcal{G}$
  - Edge from a parent to its child represents a prune action a_j
- Variables needed for the MCTS algorithm

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Tree construction: start with the input graph and take pruning actions
  - Each node $\mathcal{N}_i$ in the MCT represents a subgraph $\mathcal{G}_i$
  - Root node $\mathcal{N}_0$ represents the input whole graph $\mathcal{G}$
  - Edge from a parent to its child represents a prune action a_j
- Variables needed for the MCTS algorithm
  - $C(\mathcal{N}_i, a_j)$ denotes the number of counts for selecting action a_j for node $\mathcal{N}_i$.
  - $W(\mathcal{N}_i, a_j)$ is the total reward for all $(\mathcal{N}_i, a_j)$ visits.
  - $Q(\mathcal{N}_i, a_j) = W(\mathcal{N}_i, a_j)/C(\mathcal{N}_i, a_j)$ and denotes the averaged reward for multiple visits.
  - $R(\mathcal{N}_i, a_j)$ is the immediate reward for selecting a_j on $\mathcal{N}_i$,

    $R(\mathcal{N}_i, a_j) = \text{Score}(f(\cdot), \mathcal{G}, (\mathcal{N}_i, a_j))$

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Each MCTS iteration selects a path to a leaf node $\mathcal{G}_\ell$

$$a^* = \underset{a_j}{\operatorname{argmax}}\, Q(\mathcal{N}_i, a_j) + U(\mathcal{N}_i, a_j),$$

$$U(\mathcal{N}_i, a_j) = \lambda R(\mathcal{N}_i, a_j) \frac{\sqrt{\sum_k C(\mathcal{N}_i, a_k)}}{1 + C(\mathcal{N}_i, a_j)},$$

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Each MCTS iteration selects a path to a leaf node $\mathcal{G}_\ell$

$$a^* = \underset{a_j}{\operatorname{argmax}}\, Q(\mathcal{N}_i, a_j) + U(\mathcal{N}_i, a_j),$$

$$U(\mathcal{N}_i, a_j) = \lambda R(\mathcal{N}_i, a_j) \frac{\sqrt{\sum_k C(\mathcal{N}_i, a_k)}}{1 + C(\mathcal{N}_i, a_j)},$$

- Then update the counts and total rewards by

$$C(\mathcal{N}_i, a_j) = C(\mathcal{N}_i, a_j) + 1,$$
$$W(\mathcal{N}_i, a_j) = W(\mathcal{N}_i, a_j) + \mathrm{Score}(f(\cdot), \mathcal{G}, \mathcal{G}_\ell).$$

Subgraph Exploration via Monte Carlo Tree Search (MCTS)

- Each MCTS iteration selects a path to a leaf node $\mathcal{G}_\ell$

$$a^* = \underset{a_j}{\operatorname{argmax}}\, Q(\mathcal{N}_i, a_j) + U(\mathcal{N}_i, a_j),$$

$$U(\mathcal{N}_i, a_j) = \lambda R(\mathcal{N}_i, a_j) \frac{\sqrt{\sum_k C(\mathcal{N}_i, a_k)}}{1 + C(\mathcal{N}_i, a_j)},$$

- Then update the counts and total rewards by

$$\begin{aligned} C(\mathcal{N}_i, a_j) &= C(\mathcal{N}_i, a_j) + 1, \\ W(\mathcal{N}_i, a_j) &= W(\mathcal{N}_i, a_j) + \mathrm{Score}(f(\cdot), \mathcal{G}, \mathcal{G}_\ell). \end{aligned}$$

- Finally, select the subgraph with the highest reward from the leaf level

Score Functions

Score Functions

- Recall the instance-level feature selection of tabular data

Score Functions

- Recall the instance-level feature selection of tabular data

F all features
f target features
$\nu : 2^F \to \mathbb{R}^+$ predictive evaluation
$I_\nu(\cdot)$ importance score

- Score function

Score Functions

- Recall the instance-level feature selection of tabular data

F	all features
f	target features
$\nu : 2^F \rightarrow \mathbb{R}^+$	predictive evaluation
$I_\nu(\cdot)$	importance score

- Score function
  - Ablation study: $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$

Note: exclude (replace w/ average)

Score Functions

- Recall the instance-level feature selection of tabular data

F	all features
f	target features
$\nu : 2^F \rightarrow \mathbb{R}^+$	predictive evaluation
$I_\nu(\cdot)$	importance score

- Score function
  - Ablation study: $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$
  - Bivariate association: $I_\nu(f) = \nu(\{f\}) - \nu(\emptyset)$

Note: exclude (replace w/ average)

Score Functions

- Recall the instance-level feature selection of tabular data

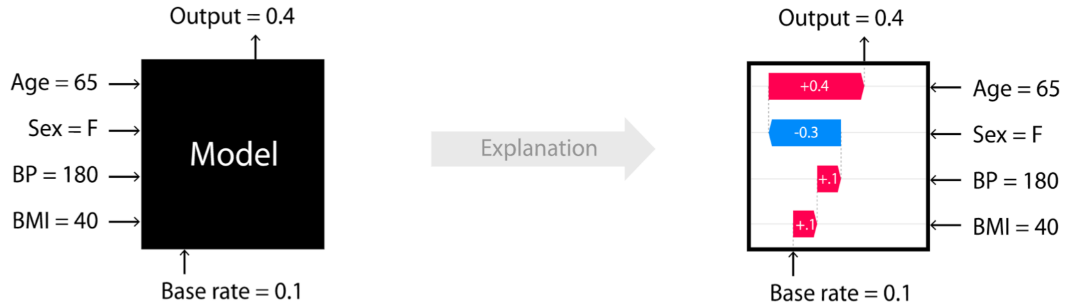
F all features
f target features
$\nu : 2^F \to \mathbb{R}^+$ predictive evaluation
$I_\nu(\cdot)$ importance score

- Score function
  - Ablation study: $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$
  - Bivariate association: $I_\nu(f) = \nu(\{f\}) - \nu(\emptyset)$
  - Consider interactions between features:

Note: exclude (replace w/ average)

Score Functions

- Recall the instance-level feature selection of tabular data



F all features
f target features
$\nu : 2^F \rightarrow \mathbb{R}^+$ predictive evaluation
$I_\nu(\cdot)$ importance score

- Score function
  - Ablation study: $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$
  - Bivariate association: $I_\nu(f) = \nu(\{f\}) - \nu(\emptyset)$
  - Consider interactions between features:
    - Ex. y = 1{Age > BP/3}

Note: exclude (replace w/ average)

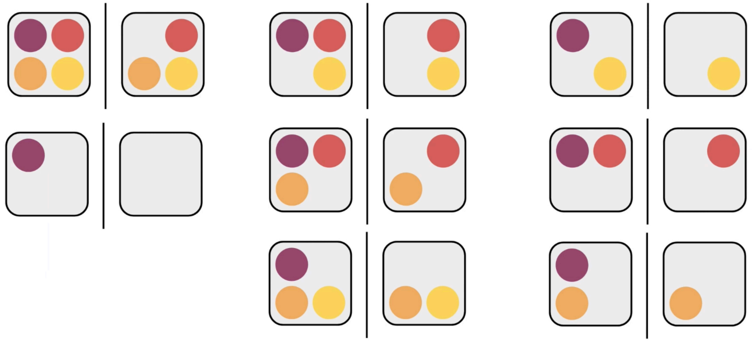
Score Functions

- Recall the instance-level feature selection of tabular data

F	all features
f	target features
$\nu : 2^F \to \mathbb{R}^+$	predictive evaluation
$I_\nu(\cdot)$	importance score

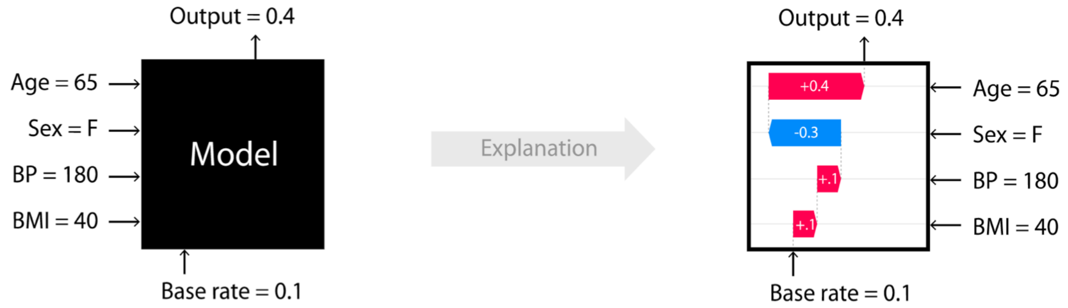
- Score function
  - Ablation study: $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$
  - Bivariate association: $I_\nu(f) = \nu(\{f\}) - \nu(\emptyset)$
  - Consider interactions between features:
    - Ex. y = 1{Age > BP/3}
    - Shapley value:

Note: exclude (replace w/ average)



Score Functions

- Recall the instance-level feature selection of tabular data

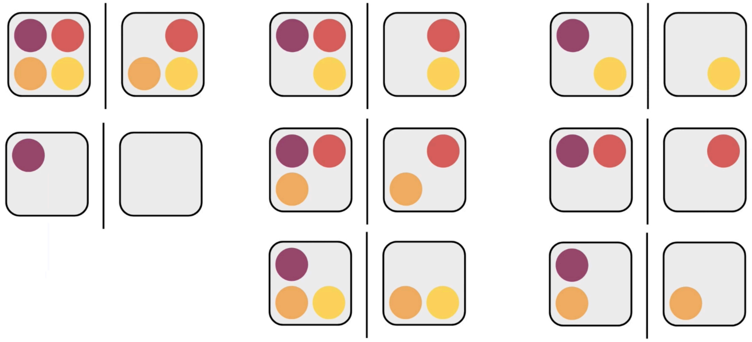


F	all features
f	target features
$\nu : 2^F \rightarrow \mathbb{R}^+$	predictive evaluation
$I_\nu(\cdot)$	importance score

- Score function
  - Ablation study: $I_\nu(f) = \nu(F) - \nu(F \setminus \{f\})$
  - Bivariate association: $I_\nu(f) = \nu(\{f\}) - \nu(\emptyset)$
  - Consider interactions between features:
    - Ex. y = 1{Age > BP/3}
    - Shapley value: $I_\nu(f) = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{f\}} \frac{1}{\binom{|F|-1}{|S|}} \Delta(f, S, \nu)$

$$\Delta(f, S, \nu) = \nu(S \cup \{f\}) - \nu(S)$$

Note: exclude (replace w/ average)



Shapley Value on Graphs

- The selected subgraph $\mathcal{G}_i$ ($\{v_1, \cdots, v_k\}$) as one "feature"

Shapley Value on Graphs

- The selected subgraph $\mathcal{G}_i$ ($\{v_1, \cdots, v_k\}$) as one "feature"
- Neighbors of $\mathcal{G}_i$ within L-hops $\{v_{k+1}, \cdots, v_r\}$ as other "features"

Shapley Value on Graphs

- The selected subgraph $\mathcal{G}_i$ ($\{v_1, \cdots, v_k\}$) as one "feature"
- Neighbors of $\mathcal{G}_i$ within L-hops $\{v_{k+1}, \cdots, v_r\}$ as other "features"

$$F = \{\mathcal{G}_i, v_{k+1}, \cdots, v_r\}$$

$$f = \mathcal{G}_i$$

$\nu =$ GNN

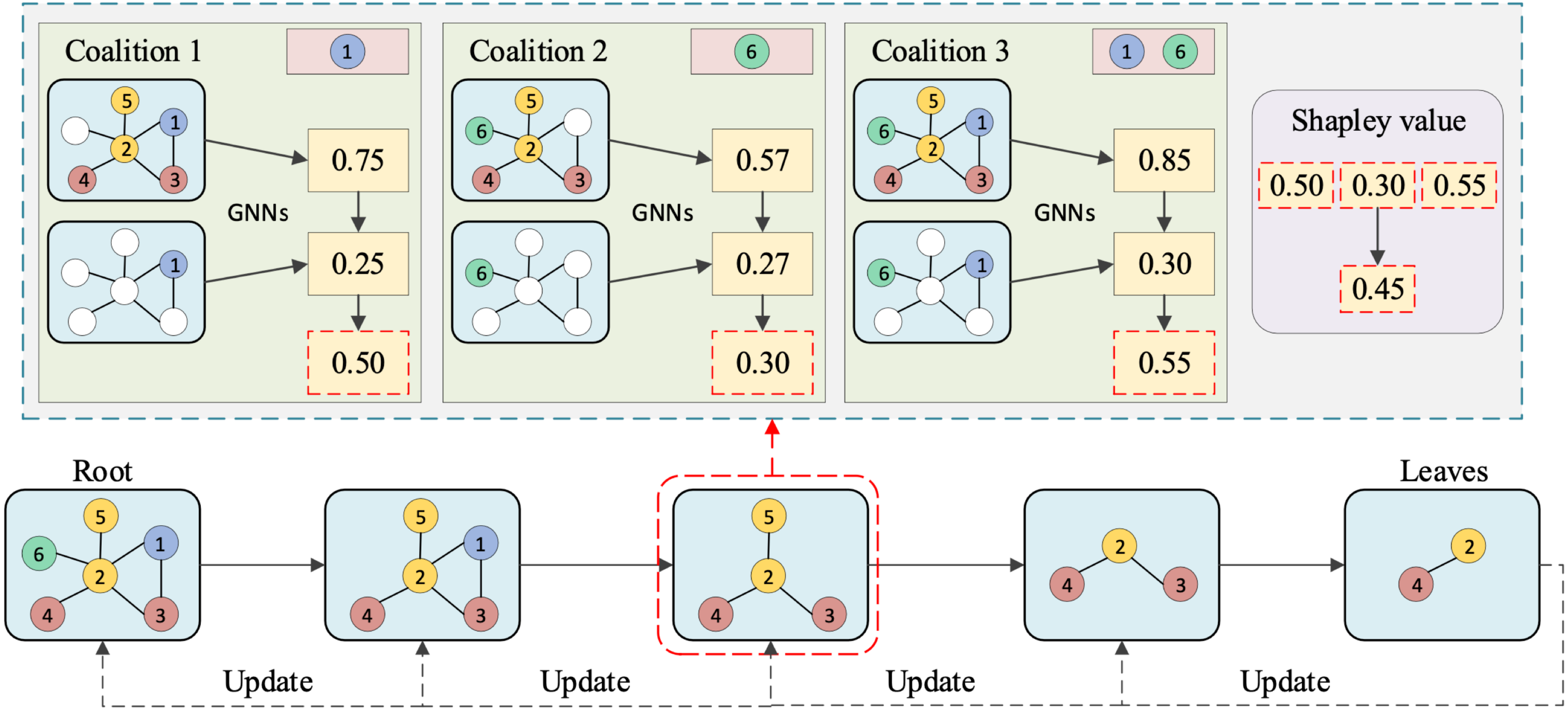
$$I_\nu(f) = \frac{1}{|F|} \sum_{S \subseteq F \setminus \{f\}} \frac{1}{\binom{|F|-1}{|S|}} \Delta(f, S, \nu)$$

$$\Delta(f, S, \nu) = \nu(S \cup \{f\}) - \nu(S)$$

SubgraphX Framework

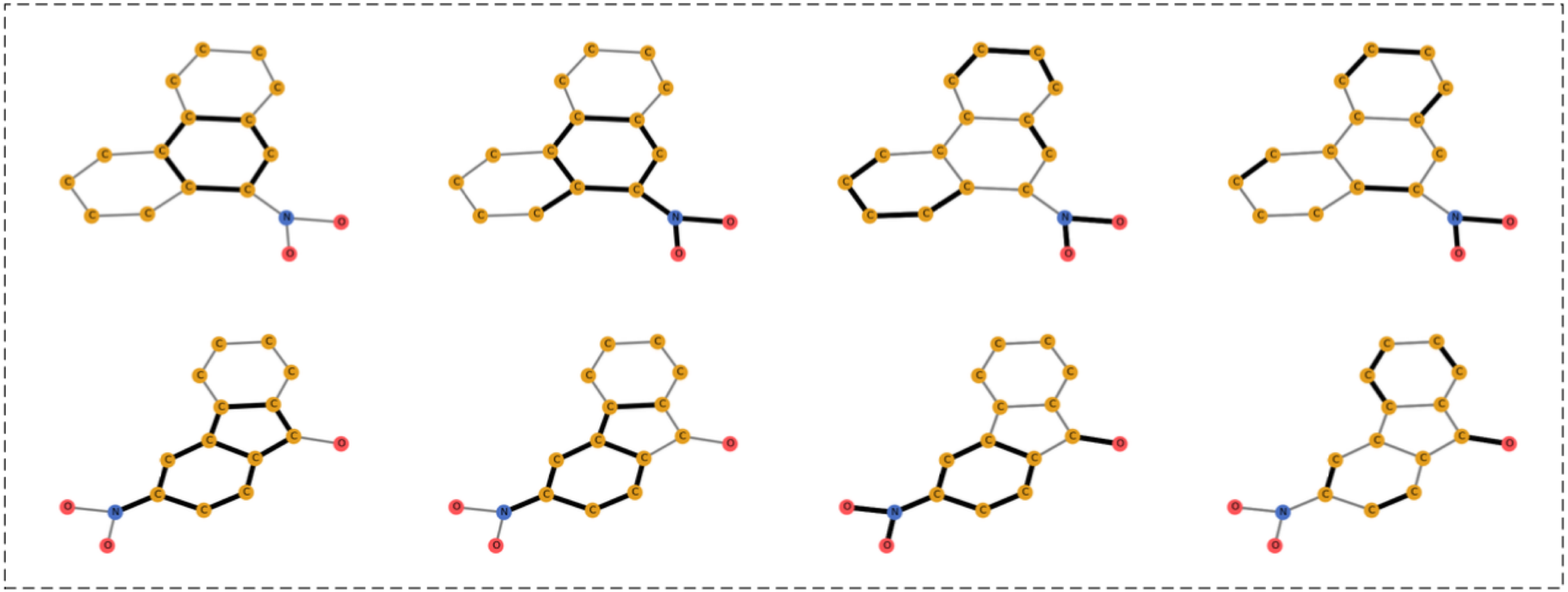
Shapley Value Scoring

MCTS



Result Visualization

- MUTAG dataset for molecule classification



SubgraphX MCTS_GNN PGExplainer GNNExplainer

Reference

- SubgraphX: Yuan, H., Yu, H., Wang, J., Li, K., & Ji, S. (2021). On explainability of graph neural networks via subgraph explorations: *https://arxiv.org/pdf/2102.05152.pdf*
- Shapley value: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Appendix