Discovering Invariant Rationales for Graph Neural Networks (DIR)

Shirley Wu, Stanford University
Motivation
—— GNN Interpretability

Deep learning models like Graph Neural Networks generally
① Fail to exhibit interpretations about why the model makes certain prediction

I have 99% confidence that this molecule is toxic
Well…

Neural Predictor   Scientist / Engineer
Deep learning models like Graph Neural Networks generally fail to exhibit interpretations about why the model makes certain prediction. This subgraph is important for this molecule being toxic. I have 99% confidence that this molecule is toxic. Well… This subgraph is important for this molecule being toxic. Interesting! Maybe we can design more new drugs with it!
Deep learning models like Graph Neural Networks generally fail to generalize to out-of-distribution (OOD) dataset.

Motivation

OOD Generalization

Graph classification task

Three kinds of base graph: tree, ladder, wheel
Three kinds of motif (label): house, circle, crane

Shirley Wu, Stanford University
Motivation
—— OOD Generalization

Deep learning models like Graph Neural Networks generally fail to generalize to out-of-distribution (OOD) dataset.

Different graph sizes/node degrees

Different domains

Shirley Wu, Stanford University
General Assumption & Solution

—— Overview

Motivation:
Deep learning models like GNNs generally
① Fail to exhibit interpretations
② Fail to generalize out of distribution

Solution: Find causal feature $C$!
Methodology

--- Invariance Condition of Causal/Shortcut Features

\[ do(S = \emptyset) \]
\[ do(S = s) \]
\[ do(S = s) \]

\( s \)-interventional environments:
\[ do(S = s) \]

Original Distribution

Multiple \( s \)-Interventional Distributions

House

Cycle

Crane

Shirley Wu, Stanford University
Methodology
—— Intrinsic Interpretable Model

In general, only the pairs of input $G$ and label $Y$ are observed during training, while neither causal feature $C$ or shortcut feature $S$ is available.

Any intrinsic interpretable model $h(G) = h_Y \circ h_C(G)$

$h_Y: \tilde{C} \rightarrow \hat{Y}$ outputs the prediction $\hat{Y}$ to approach $Y$

$h_C: G \rightarrow \tilde{C}$ discovers rationale $\tilde{C}$ from the observed $G$
Methodology

Approaching Causal Features inside the Model

**Target:** approach causal feature $C$ using rationale $\tilde{C}$

$\text{min} \mathcal{R}(h_{\hat{Y}} \circ h_{\tilde{C}}(G), Y)$

$\text{s.t. } Y \perp \tilde{S} \mid \tilde{C}$

An interpretable model: $h(G) = h_{\hat{Y}} \circ h_{\tilde{C}}(G)$

Previous methods:

$\text{DIR}: \text{min} \mathcal{R}(h_{\hat{Y}} \circ h_{\tilde{C}}(G), Y)$

$\text{s.t. } Y \perp S \mid C$: Changes on $S$ do not affect $Y$ as long as $C$ is observed

Shirley Wu, Stanford University
Methodology
—— DIR Principle & Objective

DIR Principle: Minimizes all $s$-interventional risks

\[
\text{DIR Objective: } \min \mathcal{R}_{DIR} = E_s[\mathcal{R}(h(G), Y \mid do(S = s))] + \lambda \text{Var}_s(\mathcal{R}(h(G), Y \mid do(S = s)))
\]

Theoretical Guarantees

Necessity: Oracle model (ground truth mapping) $f_Y: C \rightarrow Y$ s.t. the DIR Principle.

 Sufficiency: Suppose there exists one and only one non-trivial subset $C$, then for any $f'_Y$ s.t. the DIR Principle, we have $f_Y = f'_Y$. 

Shirley Wu, Stanford University
Methodology
——— DIR Principle & Objective

\[ R(h(G), Y | do(S = s)) \]

**Original Distribution**

| do(S = ∅) | 0.5 |
| do(S = ) | 0.3 |
| do(S = ) | 0.1 |
| do(S = ) | 0.7 |

Shirley Wu, Stanford University
DIR Modeling

Distribution Intervener

\( s \)

Rationale Generator

\( \tilde{c} \)

Shared Graph Encoder

\( \tilde{y}_{\tilde{s}} \)

Shortcut Classifier

\( y \)

Causal Classifier

\( \tilde{y} \)

Causal Prediction

\( \mathcal{R}_{\tilde{s}} \)

\( \mathcal{R}_{\text{DIR}} \)

Shirley Wu, Stanford University
No BP

g
Rationale Generator

\( \tilde{c} \)

\( \tilde{s} \)

Distribution Intervener

\( \hat{y} \)

\( \hat{y}_c \)

\( \hat{y}_S \)

\( R_{\hat{y}} \)

\( R_{\hat{y}_c} \)

\( R_{\hat{y}_S} \)

\( y \)

\( R_{\text{DIR}} \)

Shared Graph Encoder

\( \text{Shortcut Classifier} \)

\( \text{Causal Classifier} \)

\( \text{Causal Prediction} \)

Shirley Wu, Stanford University
**Rationale**

- Generator
- Shared Graph Encoder
- Causal Classifier

**Shortcut Classifier**

- No BP

**Prediction**

- Causal Prediction

Shirley Wu, Stanford University
No BP
Rationale Generator

Shared Graph Encoder

Shortcut Classifier

Causal Classifier

Distribution Intervener

\( g \)

\( \tilde{c} \)

\( \tilde{s} \)

\( \hat{y}_s \)

\( \hat{y}_c \)

\( y \)

\( \hat{y} \)

\( R_{\tilde{s}} \)

\( R_{D\text{IR}} \)

Causal Prediction

Shirley Wu, Stanford University
No BP

Rationale
Generator

Shared Graph
Encoder

Shortcut
Classifier

Causal
Classifier

Distribution
Intervener

$\hat{\mathcal{R}}_{\hat{s}}$

$\hat{y}_s$

$\hat{y}_c$

$\hat{y}$

$\mathcal{R}_{\hat{s}}$

$\mathcal{R}_{\text{DIR}}$

Causal Prediction

Shirley Wu, Stanford University
Interpretability Results

—— Rationale Precision & Visualization on Spurious-Motif

Table 2: Precision@5 on Spurious-Motif.

<table>
<thead>
<tr>
<th>Model</th>
<th>Balance</th>
<th>$b = 0.5$</th>
<th>$b = 0.7$</th>
<th>$b = 0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attention</td>
<td>0.183±0.018</td>
<td>0.183±0.130</td>
<td>0.182±0.014</td>
<td>0.134±0.013</td>
</tr>
<tr>
<td>ASAP</td>
<td>0.187±0.030</td>
<td>0.188±0.023</td>
<td>0.186±0.027</td>
<td>0.121±0.021</td>
</tr>
<tr>
<td>Top$k$ Pool</td>
<td>0.215±0.061</td>
<td>0.207±0.057</td>
<td>0.212±0.056</td>
<td>0.148±0.018</td>
</tr>
<tr>
<td>SAG Pool</td>
<td>0.212±0.033</td>
<td>0.198±0.062</td>
<td>0.201±0.064</td>
<td>0.136±0.014</td>
</tr>
<tr>
<td>DIR</td>
<td>0.257±0.014</td>
<td>0.255±0.016</td>
<td>0.247±0.012</td>
<td>0.192±0.044</td>
</tr>
</tbody>
</table>

Shirley Wu, Stanford University
Interpretability Results

— Visualization on Graph-SST2

(a) Training rationale: Positive sentiment.

(b) Training rationale: Negative sentiment.

(c) Testing rationale: Positive sentiment.

(d) Testing rationale: Negative sentiment.

Conclusion:
DIR is able to ① emphasize the tokens that directly result in the sentences’ positive or negative sentiment and ② focus persistently on the causal features for OOD testing data.
# Generalization Results

Table 1: Test ACC on the Synthetic Dataset and Real Datasets. In Spurious-Motif dataset, we color olive for the results lower than ERM, where $b$ is the indicator of the confounding effect.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Balance</th>
<th>Spurious-Motif $b = 0.5$</th>
<th>Spurious-Motif $b = 0.7$</th>
<th>Spurious-Motif $b = 0.9$</th>
<th>MNIST-75sp</th>
<th>Graph-SST2</th>
<th>Molhiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM</td>
<td>42.99±1.93</td>
<td>39.69±1.73</td>
<td>38.93±1.74</td>
<td>33.61±1.02</td>
<td>12.71±1.43</td>
<td>81.44±0.59</td>
<td>76.20±1.14</td>
</tr>
<tr>
<td>Attention</td>
<td>43.07±2.55</td>
<td>39.42±1.50</td>
<td>37.41±0.86</td>
<td>33.46±0.43</td>
<td>15.19±2.62</td>
<td>81.57±0.71</td>
<td>75.84±1.33</td>
</tr>
<tr>
<td>ASAP</td>
<td>44.44±8.19</td>
<td>44.25±6.87</td>
<td>39.19±4.39</td>
<td>31.76±2.89</td>
<td>15.54±1.87</td>
<td>81.57±0.84</td>
<td>73.81±1.17</td>
</tr>
<tr>
<td>Top-$k$ Pool</td>
<td>43.43±8.79</td>
<td>41.21±7.05</td>
<td>40.27±7.12</td>
<td>33.60±0.91</td>
<td>14.91±3.25</td>
<td>79.78±1.35</td>
<td>73.01±1.65</td>
</tr>
<tr>
<td>SAG Pool</td>
<td>45.23±6.76</td>
<td>43.82±6.32</td>
<td>40.45±7.50</td>
<td>33.60±1.18</td>
<td>14.31±2.44</td>
<td>80.24±1.72</td>
<td>73.26±0.84</td>
</tr>
<tr>
<td>Group DRO</td>
<td>41.51±1.11</td>
<td>39.38±0.93</td>
<td>39.32±2.23</td>
<td>33.90±0.52</td>
<td>15.13±2.83</td>
<td>81.29±1.44</td>
<td>75.44±2.70</td>
</tr>
<tr>
<td>V-REx</td>
<td>42.83±1.59</td>
<td>39.43±2.69</td>
<td>39.08±1.56</td>
<td>34.81±2.04</td>
<td>18.92±1.41</td>
<td>81.76±0.08</td>
<td>75.62±0.79</td>
</tr>
<tr>
<td>IRM</td>
<td>42.26±2.69</td>
<td>41.30±1.28</td>
<td>40.16±1.74</td>
<td>35.12±2.71</td>
<td>18.62±1.22</td>
<td>81.01±1.13</td>
<td>74.46±2.74</td>
</tr>
<tr>
<td>DIR-Var</td>
<td>45.87±2.61</td>
<td>43.81±1.93</td>
<td>42.69±1.77</td>
<td>37.12±1.56</td>
<td>17.74±4.17</td>
<td>81.74±0.89</td>
<td>76.05±0.86</td>
</tr>
<tr>
<td>DIR</td>
<td>47.03±2.46</td>
<td>45.50±2.15</td>
<td>43.36±1.64</td>
<td>39.87±0.56</td>
<td>20.36±1.78</td>
<td>83.29±0.53</td>
<td>77.05±0.57</td>
</tr>
</tbody>
</table>

Shirley Wu, Stanford University
Two-stage Training Dynamics

--- Adaption-Fitting

**Spurious-Motif**

T-SNE visualizations of rationale embeddings

--- Adaption

The learning of $h_{\tilde{C}}: G \rightarrow \tilde{C}$ is mainly conducted during the initial training stage, which explores the rationales that satisfy the DIR principle.

--- Fitting

DIR mainly optimizes $h_{\tilde{Y}}$ to consolidate the functional relation $\tilde{C} \rightarrow \tilde{Y}$ until model convergence, while $h_{\tilde{C}}$ (rationales) only makes small changes.
Two-stage Training Dynamics

—— Similarity between DIR and IRM penalties

We observed a strong correlation between the variance penalty and the precision metrics.

The gradient penalty term of IRM follows a similar pattern to the DIR penalty.

While IRM consistently outperforms DIR w.r.t. Training ACC, its testing performance degrades, potentially due to over-fitting.

Note: Early stopping is important in the implementation!
Future Directions

1. Expressiveness of the rationale generators
2. Generalization to unseen spurious patterns
3. More general assumptions?
   Precondition: Understand graph generation process (graph distribution)
4*. Higher level interpretability
   ① In the representation level
   ② By distilling abstract variables

(Figure from Xuanyuan et al. 2022)
Thanks!

Shirley Wu  
Xiang Wang  
An Zhang  
Xiangnan He  
Tat-Seng Chua

Github: https://github.com/Wuyxin/DIR-GNN
Email: shirwu@cs.stanford.edu/