http://mlrg.cs.binghamton.edu/

RelEx: A Model-Agnostic Relational Model Explainer

Yue Zhang, David Defazio, Arti Ramesh
{yzhan202,ddefazi1,artir} @ binghamton.edu
SUNY Binghamton

Problem?

How to construct meaningful relational explanations for relational models?

Approach?

Use perturbations and black-box predictions of perturbations to construct local approximator g. Learn relational mask on g.

Challenges?

Correctly identifying core relational structure corresponding to the prediction in a black-box setting

RelEx Highlights

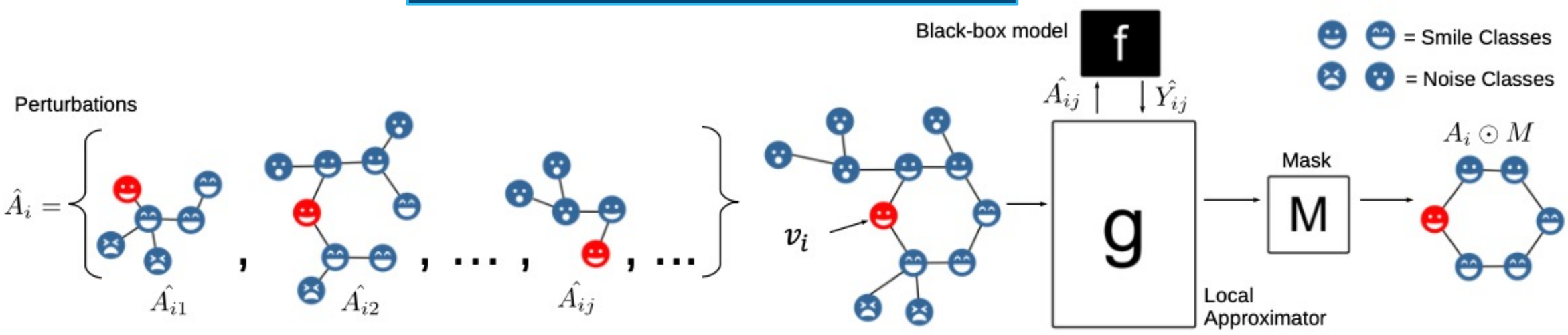
Model-agnostic relational explanations with only access to black-box output predictions

RelEx can be applied to any relational model: graph neural networks and statistical relational models

Diverse explanations by maximizing the cross entropy between two learned relational explanations

Capable of learning core topological structures in explanations

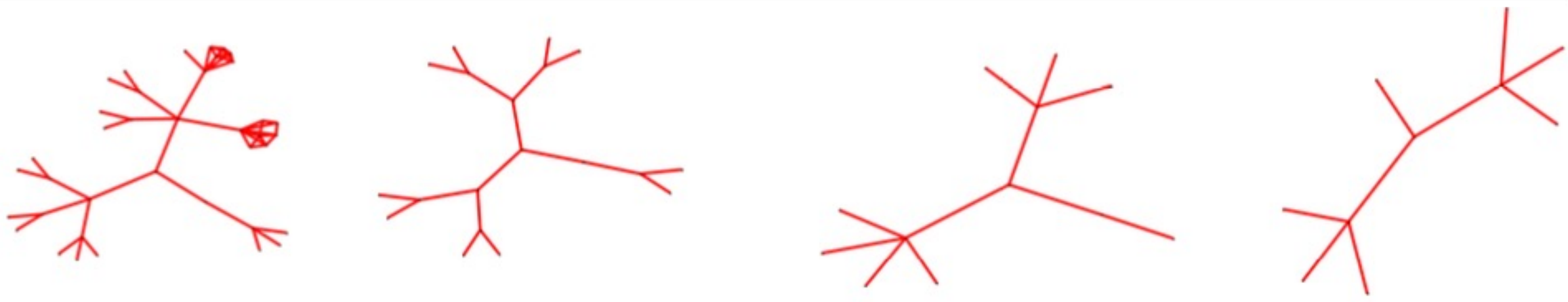
RelEx Architecture



Perturbations of adjacency matrix are constructed to query underling relational model *f*
Learn local approximator *g* using the outputs of f and perturbed inputs
Mask *M* is learned on the output of the local approximator to identify relational nodes important to the classification

Experimental Evaluation

Evaluation Metric	Saliency Map	Relational Anchors	GNN-Explainer	$RelEx_{Sigmoid}$	$RelEx_{Gumbel}$
AUC-ROC	0.4352 ± 0.1055	0.5069 ± 0.0986	0.5666 ± 0.2057	0.5470 ± 0.2028	**0.5873 ± 0.1422**
Infidelity	0.1199 ± 0.0729	0.1110 ± 0.0229	0.0885 ± 0.0207	0.0893 ± 0.0209	**0.0884 ± 0.0207**

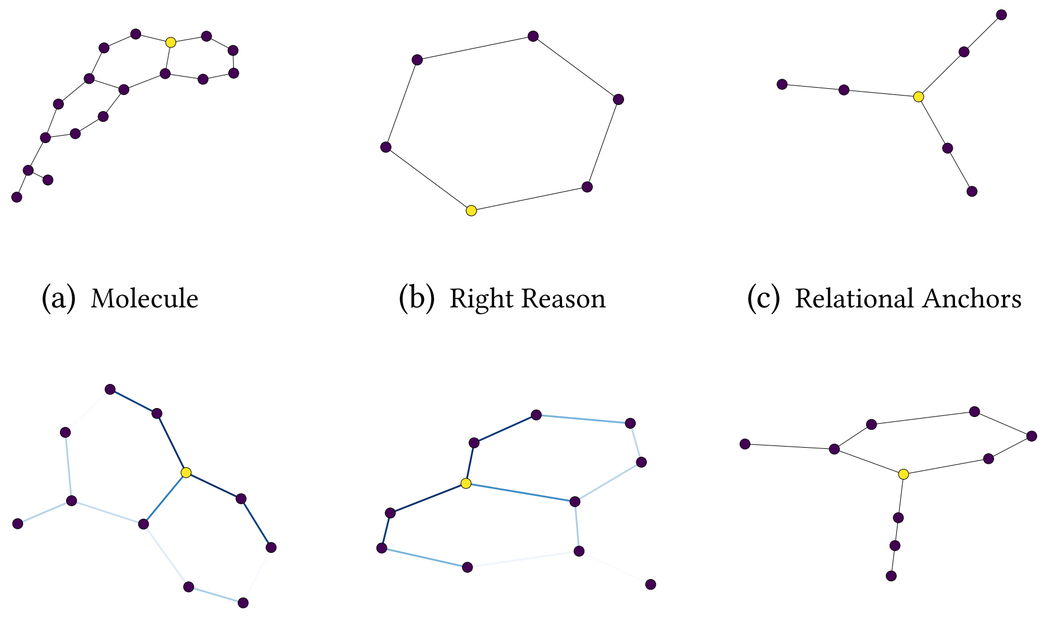


(a) Tree Computation Graph (b) Tree Right Reason (c) Tree $RelEx_{Sigmoid}$ (d) Tree $RelEx_{Gumbel}$

Synthetic TREE-GRID dataset where we connect multiple grid structures to a tree and explain whether a node is a tree-node or a grid-node using RelEx

Experiments show that RelEx achieves best performance on explanation metrics and can identify core structure

RelEx identifies the correct core explanation structure (hexagon) when compared with other explainers



(a) Molecule (b) Right Reason (c) Relational Anchors

(d) GNN-Explainer (e) $RelEx_{Sigmoid}$ (f) $RelEx_{Gumbel}$