GNN XAI 学习提纲

https://hustwj.github.io/notes/
https://github.com/hustwj/notes

Papers

XAI papers

- "Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD 2016, Marco Tulio Ribeiro et al.
  Note：本文提出的LIME是最经典的XAI post analysis方法。

- A Unified Approach to Interpreting Model Predictions, NIPS 2017, Scott M. Lundberg et al.
  Note: 该论文的网页中还列出和对比了其他常见的多种XAI方法，相关论文也请仔细阅读并比较。

- Beyond Accuracy: Behavioral Testing of NLP models with CheckList, ACL 2020, Marco Tulio Ribeiro et al.
  Note: LIME作者另一篇XAI论文，获得ACL2020最佳论文。

- A Survey of the State of Explainable AI for Natural Language Processing, AACL 2020

- Towards better understanding of gradient-based attribution methods for Deep Neural Networks, ICLR 2018

- Learning to Explain: An Information-Theoretic Perspective on Model Interpretation, ICML 2018

- L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data, ICLR 2019

还有几篇NeurIPS2020收录的以及 ICLR2021 submission的论文稍后列出。

Rationale-based XAI papers

- Rationalizing Neural Predictions, EMNLP 2016, Tao Lei et al.
  Note：本文为基于Rationale方法的开山之作。

- Interpretable Neural Predictions with Differentiable Binary Variables, ACL 2019, Jasmijn Bastings, Wilker Aziz, **Ivan Titov**

- Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport, ACL 2020, Tao Lei et al.
  Note：本文是Tao Lei的另外一篇最新相关论文，所以也列出参阅。

- How do Decisions Emerge across Layers in Neural Models? Interpretation with Differentiable Masking, EMNLP 2020, Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, **Ivan Titov**

- Learning to Faithfully Rationalize by Construction, ACL 2020, Sarthak Jain et al.
  Note：作者声称是对Tao Lei EMNLP2016论文的改进。

- Towards Explainable NLP: A Generative Explanation Framework for Text Classification, Hui Liu, Qingyu Yin, William Yang Wang

- Towards Interpretable Natural Language Understanding with Explanations as Latent Variables, NeurIPS 2020

- Interpreting Image Classifiers by Generating Discrete Masks, TPAMI 2020, Hao Yuan et al.
  Note: 将基于Rationale的方法用于图像，输出hard mask。

GNN XAI Papers

- GNNExplainer: Generating Explanations for Graph Neural Networks, NeurIPS 2019, **Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, Jure Leskovec**
  Note: Stanford发的第一篇GNN XAI的论文。

- XGNN: Towards Model-Level Explanations of Graph Neural Networks, KDD 2020, Hao Yuan et al.
  Note：TAMU通过生成实现最大化预测值的input graph来解释GNN模型。

- Multi-Objective Molecule Generation using Interpretable Substructures, ICML 2020, **Wengong Jin, Regina Barzilay, Tommi Jaakkola**
  Note: MIT新药发现研究

- Drug discovery with explainable artificial intelligence, Nature Machine Intelligence volume 2, pages573–584(2020)

- A Deep Learning Approach to Antibiotic Discovery, Cell 2020

- Parameterized Explainer for Graph Neural Network, NeurIPS 2020, Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, Xiang Zhang

- Hard Masking for Explaining Graph Neural Networks, ICLR 2021 Submission

- Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking, ICLR 2021 Submission, Michael Sejr Schlichtkrull, Nicola De Cao, **Ivan Titov**

- Explainability in Graph Neural Networks: A Taxonomic Survey, Group of Prof. Ji, TAMU

- Interpreting and Unifying Graph Neural Networks with An Optimization Framework, Group of Prof. Peng Cui, Tsinghua

- On Explainability of Graph Neural Networks via Subgraph Explorations, Group of Prof. Ji, TAMU

- CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks
- Parameterized Explainer for Graph Neural Network, NeurIPS 2020
- PGM-Explainer: Probabilistic Graphical Model Explanations for Graph Neural Networks, NeurIPS 2020

GNN Self-Supervised Learning Papers

- Graph Self-Supervised Learning: A Survey, Group of Prof. Philip S. Yu, UIC
- Self-Supervised Learning of Graph Neural Networks: A Unified Review, Group of Prof. Ji, TAMU
- Motif-Driven Contrastive Learning of Graph Representations, Group of Prof. Yizhou Sun, UCLA

Recent tutorials

XAI tutorials

- Tutorial on Explaining ML Predictions: State-of-the-art, Challenges, and Opportunities, NeurIPS 2020, Himabindu Lakkaraju, Julius Adebayo, Sameer Singh
- Tutorial on Interpreting Predictions of NLP Models, EMNLP 2020, Eric Wallace, Matt Gardner, Sameer Singh
- Interpretability and Analysis in Neural NLP, ACL 2020
- Interpreting and Explaining Deep Neural Networks: A Perspective on Time Series Data, KDD 2020

- Human-Centered Explainability for Healthcare, KDD 2020

GNN tutorials

- Advanced Deep Graph Learning: Deeper, Faster, Robuster, and Unsupervised, KDD 2020
- Multi-modal Network Representation Learning: Methods and Applications, KDD 2020
- Recent Advances on Graph Analytics and Its Applications in Healthcare, KDD 2020
- Graph Neural Networks: Models and Applications, AAAI 2020/2021

GNN for NLP tutorials

- Graph Neural Networks in NLP, CCL2019, Yue Zhang
- A Tutorial on Graph Neural Networks for Natural Language Processing, EMNLP2019

Note: 上面大部分GNN for NLP的方法在后BERT时代都显得过时了。

Resources

Courses on XAI

- Interpretability and Explainability in Machine Learning, COMPSCI 282BR, Harvard University by **Hima Lakkaraju**

Courses on GNN

- CS224W: Machine Learning with Graphs, Jure Leskovec
- Machine Learning for Graphs and Sequential Data (MLGS), Stephan Günnemann, *TU Munich*

Talks on XAI

- DARPA's Explainable Artificial Intelligence (XAI) Program, Dave Gunning
- Please Stop Doing "Explainable" ML, Cynthia Rudin

Talks on GNN

- AI Cures Drug Discovery Conference, MIT

Talk by Lee Hungyi

- Generation by RL and GAN, and - video

GNN books

- Deep Learning on Graphs, Yao Ma and Jiliang Tang
- Graph Representation Learning Book, William L. Hamilton, McGill University

> Written with StackEdit.