ARIA
APPLIED RESEARCH IN ACTION

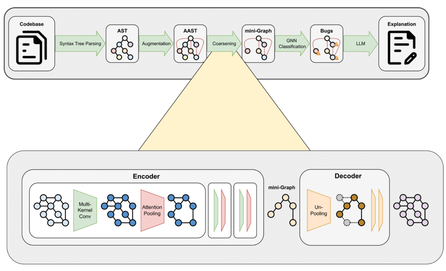
Abstract Syntax Tree Coarsening via Deep Attention-Based Node Pooling Auto-Encoders

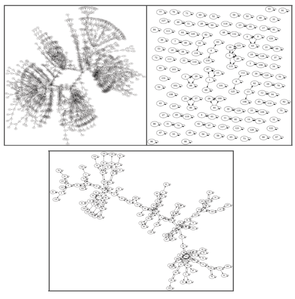
Attention-based graph coarsening allows for significant reduction of the size of graphs, while retaining sufficient information for graph reconstruction, as well as downstream classification tasks versus traditional algorithmic coarsening methods.

Anthony Rinaldi

Sushant Sachdeva
ACADEMIC SUPERVISOR

Avinash Gopal
INDUSTRY SUPERVISOR





PROJECT SUMMARY

Files of code, and collections of files (codebases), can be easily converted to an abstract syntax tree (AST) representation that contains sufficient semantic/structural information to run a program. These tree structures lend themselves naturally to be used in graph neural networks (GNNs) for static analysis of code files and codebases. However, both ASTs and GNNs are memory intensive, in that the former can be quite large and the latter do not scale well with large inputs. To solve this compatibility issue between the two, previous approaches have used heuristics to coarsen ASTs, thereby reducing their size before passing them to GNNs. These approaches suffer in their inability to rely on data to learn how to best coarsen a graph to retain information. In this work, we propose ASTA Encoder (AST Attention Encoder), a data-driven approach to learn how to condense ASTs, with the goal of passing them to a downstream GNN task. This approach introduces novelty to the graph auto-encoder architecture by calculating a learnable node importance score that incorporates edge directionality, which allows the network to understand the effect of dropping nodes with certain parent/child relationships. We train an attention-based GNN auto-encoder network on thousands of ASTs coming from a custom augmented AST parser used on public Python GitHub repositories. We show that graphs can be reduced in size by 92.5% while still increasing performance on downstream graph classification tasks.

REFERENCES

[1] Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, and Wei Wang. Simgnn: A neural network approach to fast graph similarity computation. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, page 384–392, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2016.

[3] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, and Hans-Peter Kriegel. Protein function prediction via graph kernels. Bioinformatics, 21:47–56, 06 2005.

[4] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? ArXiv, abs/2105.14491, 2021.

[5] Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. Journal of Medicinal Chemistry, 34(2):786–797, 1991.

[6] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. arXiv preprint arXiv:2003.00982, 2020.

[7] Hongyang Gao and Shuiwang Ji. Graph u-nets. In Kamalika Chaudhuri and Ruslan Salakhutdi-nov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 2083–2092. PMLR, 09–15 Jun 2019.

[8] Yunhao Ge, Yunkui Pang, Linwei Li, and Laurent Itti. Graph autoencoder for graph compression and representation learning. 2021.

[9] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In NIPS, 2017.

[10] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. ArXiv, abs/1609.02907, 2016.

[11] Z. Raisi and Mohamed A. Naiel. 2d positional embedding-based transformer for scene text recognition. 2021.

[12] Petar Veli˘ckovi c, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. International Conference on Learning Representations, 2018.

[13] Wenhan Wang, Ge Li, Bo Ma, Xin Xia, and Zhi Jin. Detecting code clones with graph neural network and flow-augmented abstract syntax tree. In 2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER), pages 261–271, 2020.

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing