GraphCoder: Transformer Pre-trained on Code Represented as Graph

G.A. Arutyunov
gaarutyunov@edu.hse.ru
HSE University
Moscow, Russia

S.M. Avdoshin
savdoshin@hse.ru
HSE University
Moscow, Russia

Abstract

Although software development is mostly a creative process, there are many scrutiny tasks. As in other industries there is a trend for automation of routine work. In many cases machine learning and neural networks have become a useful assistant in that matter. Programming is not an exception –GitHub has stated that Copilot is already used to write up to 30% code in the company. Copilot is based on Codex, a Transformer model trained on code as sequence. However, sequence is not a perfect representation for programming languages. In this work we claim and demonstrate that by combining the advantages of Transformers and graph representations of code it is possible to achieve very good results even with comparably small models.

Keywords: neural networks, Transformers, graphs, abstract syntax tree, data flow graph

ACM Reference Format:
G.A. Arutyunov and S.M. Avdoshin. 2024. GraphCoder: Transformer Pre-trained on Code Represented as Graph. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Application of Transformers yet again has managed to break the deadlock —this time in the task of code generation [Chen et al. 2021; Hendrycks et al. 2021; Y. Li et al. n.d.; Nijkamp et al. 2022]. Nevertheless, the versatile Transformer architecture has displayed good results on several benchmarks, in the recent work [Xu et al. 2022] it was shown that increasing the size of the model doesn't result in a better performance. Moreover, it is evident that context matters a lot to produce a working code. However, it is not feasible to relentlessly increase the length of context sequence in a Transformer. Therefore, a different approach is needed to boost the efficiency in the task of code synthesis [Arutyunov and Avdoshin 2022].

First of all, an expressive code representation has to be selected. Several ways including token-based, structured and graph-based approaches have been reviewed [S.M. Avdoshin and G.A. Arutyunov 2022]. For instance, graph representation using abstract syntax tree (AST), data-flow graph (DFG) and control-flow graph (CFG) yield good results in such tasks as variable misuse detection and correction [Allamanis, Brockschmidt, et al. 2017]. Such graph representation can capture an extensive amount of information about the programs code.

Secondly, a versatile model architecture that supports learning on graphs must be used. Multiple models such as RNN [White et al. 2016], LSTM [Wei and M. Li 2017] and CNN [Mou et al. 2016] with flattened graphs have been used. However, graph-aware model architecture is more suitable for the graph representation of code. For this reason, Graph Neural Networks (GNN) are a more reasonable choice of architecture, namely message-passing neural networks [Allamanis, Brockschmidt, et al. 2017].

Nonetheless, in this work we aim to make the most from both: the advantages of Transformer architecture and graph representation of code. For instance, we will use Transformer training parallelization and graph code representation created from AST. To make this possible we will use Pure Transformers [Kim et al. 2022] instead of models that have some architectural alterations to support graph structure [Dwivedi and Bresson 2021; Kreuzer et al. 2021; Ying et al. 2021].

Our main contributions:

1. Source code graph representation with AST
2. Transformer model that can be directly trained on graph structure data and applied for different tasks including code and documentation generation
3. Model pretrained on Python source code represented as graph

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Problem Statement

In this work we test the ability of Pure Transformers to add types to Python source code based on its graph structure. We compare the results with the models from previous work in Table 1 [Allamanis, Barr, et al. 2020; Mir et al. 2021; Pradel et al. 2020].

2.1 Dataset

To train and test the model we gathered 600 Python repositories from GitHub containing type annotations from Typilus [Allamanis, Barr, et al. 2020]. We clone these repositories and utilize pytype for static analysis, augmenting the corpus with inferred type annotations. The top 175 most downloaded libraries are added to the Python environment for type inference. Through deduplication, we remove over 133,000 near code duplicates to prevent bias.

The resulting dataset comprises 118,440 files with 5,997,459 symbols, of which 252,470 have non-Any non-None type annotations. The annotations exhibit diversity with a heavy-tailed distribution, where the top 10 types cover half of the dataset, primarily including str, bool, and int. Only 158 types have over 100 annotations, while the majority of types are used fewer than 100 times each, forming 32% of the dataset. This distribution underscores the importance of accurately predicting annotations, especially for less common types. The long-tail of types consists of user-defined and generic types with various type arguments. Finally, they split the data into train-validation-test sets with proportions of 70-10-20, respectively.

2.2 Metrics

To test the model we use two metrics from the Typilus paper [Allamanis, Barr, et al. 2020]:

Exact Match τ_p and τ_g match exactly.
Match up to Parametric Type Exact match when ignoring all type parameters.

3 Previous Work

3.1 Graph Transformers

Graph Transformers is a novel architecture that has been developing in the past few years. They have been applied for several tasks, mostly in the field of molecule generation, node classification and node feature regression [Dwivedi and Bresson 2021; Kim et al. 2022; Kreuzer et al. 2021; Ying et al. 2021].

AST and DFG have already been used with Transformers in the code generation and summarization tasks [Sun et al. 2020; Tang et al. 2021; Wang et al. 2022], as well as

4 Proposed Solution

4.1 Model Architecture

We base our model architecture on TokenGT [Kim et al. 2022]. For training, cross entropy loss with weights is used due to the imbalance of the dataset.

5 Experiment Results and Ablation

For now, the model has been trained and tested on one repository. The resulting accuracy for all types is 41% and 45.9% accuracy up to parametric type.

6 Future Work

In this work we explored the application of Graph Transformers for type inference. The versatile architecture of the proposed solution lets us explore other tasks.

First, if a universal version of AST parsing is used the can train the model for multiple programming languages [Wang et al. 2022]. Second, we can train the model using a technique similar to generative pretrained models [Brown et al. 2020; Radford et al. 2019] to generate code. Third, our model can be used to generate code summarization or docstring generation [Barone and Sennrich 2017; Liu et al. 2021]. Another useful task is to detect errors and generate fixes [Bhatia and Singh 2016; Fujimoto et al. 2018; Marginean et al. 2019]. Finally, we can extend our model with information about changes to analyse them and propose refactoring possibilities [Cabrera Lozoya et al. 2021].

7 Conclusion

As for the conclusion, we were able to create a universal model based on TokenGT [Kim et al. 2022] and code represented as graphs. One of the most important advantages of this model is that the code graph is used directly by the model. Secondly, the model can be modified to fit other tasks, such as code generation and summarization, docstring generation, refactoring and many more. The code graph can also be extended by different features and node types, since the representation does not differ depending on graph structure.

8 Acknowledgments

This research was supported in part through computational resources of HPC facilities at HSE University [Kostenetskiy et al. 2021].

References

Miltiadis Allamanis, Earl T Barr, Soline Ducousso, and Zheng Gao. 2020. "Typilus: Neural Type Hints." In: *PLDI*.

Miltiadis Allamanis, Marc Brockschmidt, and Mahmoud Khademi. 2017. "Learning to represent programs with graphs." *arXiv preprint arXiv:1711.00740*.

Name	% Exact Match	% Match up to Parametric Type
GraphCoder	41%	45.9%
Typilus	54.6%	64.1%

Table 1. Quantitative evaluation of models measuring their ability to predict ground truth type annotations.

German Arsenovich Arutyunov and Sergey Mikchailovitch Avdoshin. 2022. "Big Transformers for Code Generation." *Proceedings of the Institute for System Programming of the RAS*, 34, 4, 79–88. Publisher: Institute for System Programming of the Russian Academy of Sciences. DOI: 10.15514/ispras-2022-34(4)-6.

Antonio Valerio Miceli Barone and Rico Sennrich. 2017. "A parallel corpus of python functions and documentation strings for automated code documentation and code generation." *arXiv preprint arXiv:1707.02275*.

Sahil Bhatia and Rishabh Singh. 2016. "Automated correction for syntax errors in programming assignments using recurrent neural networks." *arXiv preprint arXiv:1603.06129*.

Tom Brown et al.. 2020. "Language models are few-shot learners." *Advances in neural information processing systems*, 33, 1877–1901.

Rocío Cabrera Lozoya, Arnaud Baumann, Antonino Sabetta, and Michele Bezzi. 2021. "Commit2vec: Learning distributed representations of code changes." *SN Computer Science*, 2, 3, 1–16. Publisher: Springer.

Mark Chen et al.. 2021. "Evaluating large language models trained on code." *arXiv preprint arXiv:2107.03374*.

Vijay Prakash Dwivedi and Xavier Bresson. Jan. 24, 2021. *A Generalization of Transformer Networks to Graphs*. Number: arXiv:2012.09699. (Jan. 24, 2021). arXiv: 2012.09699[cs]. DOI: 10.48550/arXiv.2012.09699.

Scott Fujimoto, Herke van Hoof, and David Meger. Oct. 22, 2018. *Addressing Function Approximation Error in Actor-Critic Methods*. (Oct. 22, 2018). arXiv: 1802.09477[cs, stat]. DOI: 10.48550/arXiv.1802.09477.

Dan Hendrycks et al.. 2021. "Measuring coding challenge competence with apps." *arXiv preprint arXiv:2105.09938*.

Jinwoo Kim, Tien Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. July 6, 2022. *Pure Transformers are Powerful Graph Learners*. (July 6, 2022). arXiv: 2207.02505[cs]. DOI: 10.48550/arXiv.2207.02505.

P. S. Kostenetskiy, R. A. Chulkevich, and V. I. Kozyrev. Jan. 2021. "HPC Resources of the Higher School of Economics." *Journal of Physics: Conference Series*, 1740, 1, (Jan. 2021), 012050. Publisher: IOP Publishing. DOI: 10.1088/1742-6596/1740/1/012050.

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Oct. 27, 2021. *Rethinking Graph Transformers with Spectral Attention*. Number: arXiv:2106.03893. (Oct. 27, 2021). arXiv: 2106.03893[cs]. DOI: 10.48550/arXiv.2106.03893.

Yujia Li et al.. N.d. "Competition-Level Code Generation with AlphaCode," 74.

Xuye Liu, Dakuo Wang, April Wang, Yufang Hou, and Lingfei Wu. 2021. "HAConvGNN: Hierarchical attention based convolutional graph neural network for code documentation generation in jupyter notebooks." *arXiv preprint arXiv:2104.01002*.

Alexandru Marginean, Johannes Bader, Satish Chandra, Mark Harman, Yue Jia, Ke Mao, Alexander Mols, and Andrew Scott. 2019. "Sapfix: Automated end-to-end repair at scale." In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 269–278.

Amir M. Mir, Evaldas Latoskinas, Sebastian Proksch, and Georgios Gousios. 2021. "Type4py: Deep similarity learning-based type inference for python." *arXiv preprint arXiv:2101.04470*.

Lili Mou, Ge Li, Lu Zhang, Tao Wang, and Zhi Jin. 2016. "Convolutional neural networks over tree structures for programming language processing." In: *Thirtieth AAAI conference on artificial intelligence*.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. "A Conversational Paradigm for Program Synthesis." *arXiv preprint arXiv:2203.13474*.

Michael Pradel, Georgios Gousios, Jason Liu, and Satish Chandra. 2020. "TypeWriter: neural type prediction with search-based validation." In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (ESEC/FSE 2020). Association for Computing Machinery, Virtual Event, USA, 209–220. ISBN: 9781450370431. DOI: 10.1145/3368089.3409715.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language models are unsupervised multitask learners." *OpenAI blog*, 1, 8, 9.

S.M. Avdoshin and G.A. Arutyunov. 2022. "Code Analysis and Generation Methods Using Neural Networks: an Overview." *INFORMATION TECHNOLOGIES*, 28, 7, 378–391. DOI: 10.17587/it.28.378-391.

Zeyu Sun, Qihao Zhu, Yingfei Xiong, Yican Sun, Lili Mou, and Lu Zhang. 2020. "Treegen: A tree-based transformer architecture for code generation." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. Issue: 05, 8984–8991.

Ze Tang, Chuanyi Li, Jidong Ge, Xiaoyu Shen, Zheling Zhu, and Bin Luo. Dec. 2, 2021. *AST-Transformer: Encoding Abstract Syntax Trees Efficiently for Code Summarization*. (Dec. 2, 2021). arXiv: 2112.01184[cs]. DOI: 10.48550/arXiv.2112.01184.

Kesu Wang, Meng Yan, He Zhang, and Haibo Hu. May 16, 2022. "Unified Abstract Syntax Tree Representation Learning for Cross-Language Program Classification." In: *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*. (May 16, 2022), 390–400. arXiv: 2205.00424[cs]. DOI: 10.1145/3524610.3527915.

Huihui Wei and Ming Li. 2017. "Supervised Deep Features for Software Functional Clone Detection by Exploiting Lexical and Syntactical Information in Source Code." In: *IJCAI*, 3034–3040.

Martin White, Michele Tufano, Christopher Vendome, and Denys Poshyvanyk. 2016. "Deep learning code fragments for code clone detection." In: *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 87–98.

Frank F. Xu, Uri Alon, Graham Neubig, and Vincent J. Hellendoorn. 2022. "A Systematic Evaluation of Large Language Models of Code." *arXiv preprint arXiv:2202.13169*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Nov. 23, 2021. *Do Transformers Really Perform Bad for Graph Representation?*

(Nov. 23, 2021). arXiv: 2106.05234[cs]. DOI: 10.48550/arXiv.2106.05234.