Programming Language Representation with Semantic-level Structure

Anonymous Author(s)

ABSTRACT

Pre-trained word embedding helps to find global optimum of the natural language processing (NLP) models and enhances performance on their target tasks. As naturalness of software source code has been proven, pre-trained langugae models also play an important role in fusion of software (SW) engineering area. Compared to natural language, source code has more structured syntax, and the structural information is crucial to construct context of the source code. Considering that, prior works have relied on syntactic level and data level structure information combined with sequence of source code. Unlike prior work, in this work, we focus on leveraging semantic-level structure in a pre-trained model for programming language. The semantic-level structure represents data and control dependencies between statements in a program, thus it constructs code behavior and its semantic. The proposed algorithm is developed based on Transformer. Not only injecting the structural information, we also introduces corresponding tasks to make source code sequence cognizant of the structure.

We evaluate our model and analyze the efficacy of CFG on three programming language downstream tasks, including code clone detection, code transtion and code refinement. We expect that semantic-level structure contributes more to build context of source code, and improves pre-trained source code model raising performances of the downstream tasks.

ACM Reference Format:

Anonymous Author(s). 2022. Programming Language Representation with Semantic-level Structure. In *Proceedings of ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2022).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

2 BACKGROUND

3 RELATED WORK

4 APPROACH

5 RESEARCH QUESTIONS FOR EVALUATION

6 EXPERIMENT

7 RESULT

REFERENCES

ISSTA 2022, 18-22 July, 2022, Daejeon, South Korea

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn