Static Analysis-Guided Automatic Source Code Summarization via Deep Learning

by Wenhua Wang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Guandong Xu

University of Technology Sydney
Faculty of Engineering and Information Technology

December 2021

Certificate of Original Authorship

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Wenhua Wang, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Software, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.
**If applicable, the above statement must be replaced with the collaborative doctoral degree statement (see below).*

**If applicable, the Indigenous Cultural and Intellectual Property (ICIP) statement must be added (see below).*

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 9/7/2022

ACKNOWLEDGMENTS

As a student of the Joint PhD Training Program offered by the University of Technology Sydney (UTS) and Southern University of Science and Technology (SUSTech), I would like to express my heartfelt gratitude to my supervisors, Guandong Xu in UTS and Yunquan Zhang in SUSTech, for their supervision as well as to UTS and SUSTech for their support for my studies in both universities. Besides, I also want to thank Yulei Sui in UTS and Shin Hwei Tan, Yepang Liu in SUSTech for their help when I encounter difficulties and bottlenecks and Yuanyuan Dou in SUSTech for her careful work to support our daily research work. Thank you all.

Thanks to the staff and teachers of the GRS in UTS and SUStech graduate school for their efforts to our daily life support. Thanks to teacher Xing Zhang in School of Engineering, SUSTech to offer help when I encounter in my life and he also organized various activities for example hiking and mountain climbing which relax us when we felt tired. Thanks to teachers in Department of Computer Science and Engineering, SUSTech and School of Computer Science at UTS who provides a comfortable scientific research environment. Thank you all.

Then, I want to thank the students I work with in both UTS and SUSTech. Thank you for your daily help in both my research and my life. Thank you for the warmhearted conversation when I feel confused. Thank you for going sightseeing and tasting delicious food together with me on holidays which let me know I am not lonely. Thank you for your warm solicitude when I am sick. Thank you for your talking with me about our research progress and solve the problems together. Thank you all.

Next, I want to thank all my lovely friends. When I feel sad and doubts about life and the future, it's your patience that enlightens and removes my doubts. When I encounter major difficulties in life, you give me great help. Thank you for being with me all the time, so that I can maintain my hope and longing for a better life in the future when I feel confused and helpless. Thank you all.

At last, I want to thank my family. I want to express my heartfelt gratitude to my parents who have supported me for all these thirty years, you are the source of my continuous progress in these years. Thanks to my sister, my grandmother and other members in my big family. Thank you all.

Thank you all!

List of Publications

Related to the Thesis :

1. Wenhua Wang, Yuqun Zhang, Yulei Sui, Yao Wan, Zhou Zhao, Jian Wu, Philip Yu, and Guandong Xu. Reinforcement-Learning-Guided Source Code Summarization using Hierarchical Attention. IEEE Transaction on Software Engineering, 2021.

2. A Transformer-based Generative Adversarial Network Framework for Universal Code Summarization, submitted.

3. On Semantic-rich Code Summarization by Vital Code and Transformer-based Model, submitted.

Abstract

Abstract. Code summarization provides the main aim described in natural language of the given function, it can benefit many tasks in software engineering. As far as I known, the existing research on comment generation can be summarized as the template based approaches, the information retrieval based approaches and the deep learning based approaches. Nowadays, based on the proposal and wide utilization of deep learning, the research of neural machine translation has been introduced to the research of code summarization. Based on my study, The existing deep learning based code summarization approaches mainly utilize the seq2seq model in which the encoder translates the source code into hidden representation of the program code and then the decoder decodes the representation into comment. However, due to the special grammar and syntax structure of programming languages and various shortcomings of different deep neural networks, the accuracy of existing code summarization approaches is not good enough. These approaches mainly suffer from three major drawbacks: a) they regard the source code as plain text directly, which neglecting the syntax structure of the source code that is quite important for the comprehension of source code; b) they only consider the generation of the code's intent, while ignore the information of parameters etc which is also quite important for the understanding and usage of the source code; c) their adopted CNN/RNN model usually cause long-distance dependency and excessive computation cost problem. Considering these limitations, the main research work of this thesis are as follows: (1) The first work proposes to adopt the hierarchical attention mechanism to enable the code summarization framework to translate three representations of source code to the hidden spce and then it injects them into a deep reinforcement learning model to enhance the performance of code summarization. (2) While many existing approaches exploit inadequate power of statement-wise semantic contributions for augmenting their performance, the second work proposes the transformer-based generative adversarial network framework for universal code summarization which constructs a cross-language universal hierarchical semantic (UHS) model to classify statements according to their positions in the source code. (3) Considering that almost all approaches only consider to generate the general intent of the method without documenting their parameters, the third work proposes to generate both the method comment and the parameter comment to provide complete java documentation for the code snippets. Specifically, it designs a programming-analysis-based component to extract UseSet of parameter and the KeySet of the source code to obtain the main semantic information and discard the useless noise information and utilizes the copy-attention-integrated transformer based NMT framework. Thought the completion of this thesis, I conduct a few of experiments, and the results of which prove that the proposed approaches can obtain better accuracy compared with the baseline approaches.

Table of Contents

LIST OF FIGURES

FIGURE **Page**

List of Tables