UTS UNIVERSITY OF TECHNOLOGY SYDNEY

A Study on Neural-based Code Summarization in Low-resource Settings

by Yang He

Thesis submitted in fulfilment of the requirements for the degree of

Master in Analytics (Research)

under the supervision of Guandong Xu

University of Technology Sydney
Faculty of Engineering and Information Technology

06/2022

C03051: Master in Analytics (Research)
13696152 Master Thesis: Analytics
June 2022

A Study on Neural-based Code Summarization in Low-resource Settings

Yang He

School of Computer Science
Faculty of Engineering & IT
University of Technology Sydney
NSW - 2008, Australia

A Study on
Neural-based Code Summarization in Low-resource Settings

A thesis submitted in partial fulfilment of the requirements
for the degree of

Master
in
Analytics (Research)

by

Yang He

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2008, Australia

June 2022

Certificate of Original Authorship Template

Graduate research students are required to make a declaration of original authorship when they submit the thesis for examination and in the final bound copies. Please note, the Research Training Program (RTP) statement is for all students. The Certificate of Original Authorship must be placed within the thesis, immediately after the thesis title page.

Required wording for the certificate of original authorship

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yang He declares that this thesis, is submitted in fulfilment of the requirements for the award of master of analytics (research), in the 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.

Signature: Production Note:
Signature removed prior to publication.

Date: 23/6/2021

ABSTRACT

Automated software engineering with deep learning techniques has been comprehensively explored because of breakthroughs in code representation learning. Many code intelligence approaches have been proposed for the downstream tasks of this field in the past years, contributing to significant performance progress. Code summarization has been the central research topic among these downstream tasks because of its contributions to practical applications, e.g., software development and maintenance. It remains challenging to represent code snippets and generate more accurate descriptions to summarize the functionality and semantics of programs.

Existing methods of the code summarization task have been devised to tackle real-world problems and have been successfully proven effective. However, there is little attention to its application in novel programming languages where only a few well-documented programs in these low-resource languages are available for training. According to our observation, existing approaches can only acquire poor performances in such settings, and we attribute the problem to *data-hungry* and *programming language gaps*.

Enlightened by recent pre-training methods, we propose METASUM, a meta-learning-based code summarization model, to extract prior and shared knowledge from high-resource programming language where high-quality code snippets are easily accessible and then adapt it to low-resource settings. The critical contribution of this dissertation is that we (1) give a comprehensive illustration of the development of machine-learning-based code summarization task, (2) identify a new problem of low-resource code summarization and propose a meta-learning-based model to improve over other methods by 3.18 and 1.79 BLEU points over state-of-the-art pre-trained models on Nix and Ruby datasets, respectively, and (3) introduce a machine-learning-based toolkit, NATURALCC, for fair comparison of models for the automated software engineering community.

Keywords Automated Software Engineering, Code Summarization, Low-resource Setting, Meta-Learning, Code Intelligence, NATURALCC

AUTHOR'S DECLARATION

I, *Yang He*, declare that this dissertation, submitted in partial fulfillment of the requirements for the award of Master by Research, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia, is wholly my 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.

I acknowledge that the research work in this dissertation is done under the supervision of Prof. Guandong Xu and Dr. Yao Wan at Huazhong University of Science and Technology. This dissertation is composed of my previous work in cooperation with Dr. Yao Wan and contains no material published except where the open-source NATURALCC toolkit has been released on GitHub, and some figures are adapted from submitted papers.

I have clearly stated the contribution of others to my thesis as a whole. The content of my thesis comes from my research works since the commencement of my Master by Research degree. I acknowledge that an electronic copy of my thesis must be lodged with the University Library and subject to the policy and procedures of the University of Technology Sydney.

SIGNATURE: Production Note: Signature removed prior to publication.

DATE: 23rd June, 2022

PLACE: Sydney, Australia

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my supervisor, Prof. Guandong Xu, and co-supervisor, Dr. Yulei Sui, for presenting me with a chance to study at the DSMI group at the University of Technology Sydney. They suggested that I make research plans and balance research work and daily life. Their guidance considerably helped me when I began my study in Australia.

Besides my advisors, I would like to thank my long-term cooperator Dr. Yao Wan who has been working with me. I cannot remember how often we stayed up all night to submit conference papers via remote cooperation. Under his cooperation, my programming ability is comprehensively developed via maintaining and extending the repositories.

Moreover, I would like to thank my colleagues in the research group. When I met mathematical problems, Dr. Yangyang Shu gave detailed interpretations of machine learning theories. I can acquire many experiences, including adjusting for a research position since it is easy to get depressed, especially for an international student in Australia, while facing Covid-19. On the other hand, thank Dr. Jun Yin and Dr. Qian Li for daily affairs and sharing facial masks when the pandemic is severe in Sydney.

My greatest thankfulness is owed to my parents, who covered my expense for the entire master phase and always encouraged me during the past two years when I was depressed by research work. Without their support, I would never pursue a master's degree at UTS and met friendly people in UTS. Although there are some regrets in my two-year's study, such as no full-paper has been formally accepted, your support and encouragement teach me never to give up pursuing research.

Lastly, thanks to Dr. Chandranath Adak for providing this thesis template and staff in UTS for offering me a quiet environment for study.

List of Publications

Related to the Thesis :

1. Yao Wan*, **Yang He***, Yulei Sui, Jian-Guo Zhang, Zhou Zhao, Lin Li, Guandong Xu, Philip S. Yu, *Cross-Language Knowledge Transfer for Low-Resource Code Summarization*. (*Equal contribution. Under revision for IEEE Transactions on Software Engineering)

2. Yao Wan, **Yang He**, Jian-Guo Zhang, Yulei Sui, Hai Jin, Guandong Xu, Caiming Xiong, Philip S. Yu, *NaturalCC: A Toolkit to Naturalize the Source Code Corpus*. In 2022 IEEE/ACM 44th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion) (pp. 149-153). IEEE.

Table of Contents

List of Figures

Figure **Page**

List of Tables