TSE: Revised Project Proposal

Team Deep Understanding (Shirish Singh, Aria Jiang, Yiqiao Liu)

March 2021

1 Project Description

The project aims to build a code understanding tool that helps find code clones across multi-file projects. The tool will be based on machine learning techniques and models. A motivating scenario for such a tool is that if a developer wants to understand a piece of code that s/he has encountered for the first time, it would be helpful to find similar and familiar code that the developer might have developed before. The tool can find behavioral similarity between the new code and code the developer is familiar with. More broadly, the applications of behavioral similarity are manifold: a) Program understanding for first-time developers, b) code search, c) code refactoring, and so on. We will attempt to answer the following research questions:

- Does obfuscation of variable names yield an improved code2seq
- How can we utilize code2seq model's output for code similarity?

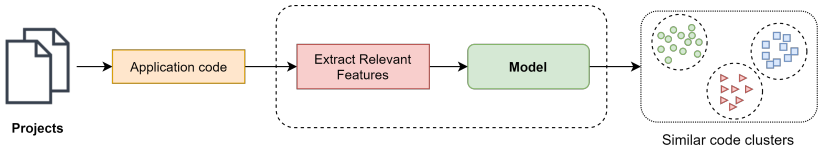

Figure 1: High level overview

2 Our Plan

For our project, we are planning to build a real-time code similarity tool that can be used to find similar functions. Figure 1 shows the high level overview of our system.

We plan to use the 106 student submissions [1] from the Advanced Software Engineering course that took place in Fall-2020 as dataset for the project.

In terms of milestones we hope to achieve, please find details in the table below.

Milestones	**Tasks**
First Progress Report	Process dataset, apply an existing model on dataset
Second Progress Report	Improve upon existing models
Project Demo	Refine implementation and organize as an open-source project

3 Deliverables

By the end of the project, we will deliver the following items:

1. Processed Dataset
2. Open-source implementation of our scripts (Training and testing)
3. Final Report

4 Task Division for Phase 1

- **Aria Jiang**: (Phase-1) Apply existing model (code2seq) on dataset.
- **Yiqiao Liu**: (Phase-1) Apply word2vec to the output of code2seq, calculate the similarity/distance between each pair of embedding, and find the "closest" code.
- **Shirish Singh**: (Phase-1) Environment setup, ASE Dataset extraction and cleaning.

5 How the project is relevant to this course

The project is relevant to our course because such a tool has ample applications in advancing the field of software engineering. Additionally, the topic of code understanding, and more specifically code understanding through testing is an important area within the field of software engineering.

[1] https://github.com/jxm033f/4156-PublicAssignment/network/members

6 Why we are interested in doing this project

Aria Jiang: I am interested in doing this project because the topic is very much related to my midterm paper. After reading about different preprocessing techniques to incorporate syntactic information and different models, I am very excited to work on a solution that builds on top of the state-of-the-art solutions. Moreover, it will be very cool to be able to compare my solution against the existing solutions on a new and real dataset. This project could be a great extension to the research I have done for the midterm paper and could help me understand even more about this field and about conducting research in general.

Yiqiao Liu: I am interested in the topic because I'd like to see what factors similarity between code chunks may depend on - somewhat related to my midterm paper, and how that may help with programming in the industry.

Shirish Singh: I want to learn more about code/program representation. Any model is as good as the data we train it on, and having a good representation that captures most if not all the nuances of the code will assist the model to perform the tasks better.

7 If the project is related to your midterm paper

This project is related to Aria's midterm paper. It is somewhat related to Yiqiao's midterm paper in that the topic is in code understanding and test case generation. It is not related to Shirish's midterm paper

References

[1] Uri Alon, Shaked Brody, Omer Levy, and Eran Yahav. code2seq: Generating sequences from structured representations of code. In *International Conference on Learning Representations*, 2019.