FACULTY OF INFORMATION TECHNOLOGY CTU IN PRAGUE

Supervisor's statement of a final thesis

Supervisor:	Mgr. Alexander Kovalenko, Ph.D.
Student:	Rudolf Raevskiy
Thesis title:	Machine Learning Techniques for Source Code Pattern Recognition
Branch / specialization:	Knowledge Engineering
Created on:	3 June 2022

Evaluation criteria

1. Fulfillment of the assignment

[1] assignment fulfilled
▸ **[2] assignment fulfilled with minor objections**
[3] assignment fulfilled with major objections
[4] assignment not fulfilled

The initial task included extensive work with crafted code representation, however, the supervisor underestimated the compute needed for this task (working with graphs), therefore the student emphasized training large transformer-based models on raw code and abstract syntax trees. In the end, the student's contribution was, in my opinion, even more valuable. He showed that we could leverage the structure of the high-level programming languages by learning AST embeddings and fine-tuning models pre-trained on natural language data (BERT-like model pre-trained on Wikipedia) outperformed the models pre-trained on source code (CodeBERT pre-trained on 14 languages).

2. Main written part

85 / 100 (B)

The work is of a good extent for a bachelor's thesis. Given the task, the wider audience (eg. programming language researchers) might be interested in the content of the work, thus, all the parts are necessary to understand the concept fully. However, some parts of the thesis are a bit chaotic and hard to follow, which is common for writing by a junior researcher. It would be nice to see some more human-comprehensible examples in the thesis. All the sources are relevant and properly cited.

3. Non-written part, attachments

99 / 100 (A)

The source code is clear and comprehensible.
The student implemented a demo of his model.

Moreover, the student is contributing to creating a training dataset, which is a valuable contribution to the community.

4. Evaluation of results, publication outputs and awards 90 /100 (A)

Thesis results can be used to get high-quality coed embeddings. With some additional work, this model can be used to solve several different tasks, such as code completion, bug fix, etc.

5. Activity of the student

- ▸ [1] **excellent activity**
- [2] very good activity
- [3] average activity
- [4] weaker, but still sufficient activity
- [5] insufficient activity

6. Self-reliance of the student

- [1] excellent self-reliance
- ▸ [2] **very good self-reliance**
- [3] average self-reliance
- [4] weaker, but still sufficient self-reliance
- [5] insufficient self-reliance

The overall evaluation 90 /100 (A)

Nice work, overall, with certain drawbacks, in writing and result demonstration, however, given the amount of work the student conducted I believe that the student deserved an "A".

Instructions

Fulfillment of the assignment

Assess whether the submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.

Main written part

Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies?

Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 52/2021, Art. 3.

Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.

Non-written part, attachments

Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

Evaluation of results, publication outputs and awards

Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

Activity of the student

From your experience with the course of the work on the thesis and its outcome, review the student's activity while working on the thesis, his/her punctuality when meeting the deadlines and whether he/she consulted you as he/she went along and also, whether he/she was well prepared for these consultations.

Self-reliance of the student

From your experience with the course of the work on the thesis and its outcome, assess the student's ability to develop independent creative work.

The overall evaluation

Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.