Using Image Classification to Predict the Quality of Source Code

Bachelor's / Master's Thesis

Supervisor: Prof. Dr. Alexander Pretschner
Advisor: Markus Schnappinger, Dr. Arnaud Fietzke
Email: markus.schnappinger@tum.de
Phone: +49 (89) 289 - 17386

Starting date: October 2021

Keywords: Machine Learning, Image Classification, Data Augmentation, Software Quality Prediction

Fakultät für Informatik
Lehrstuhl 4
Software und Systems Engineering
Prof. Dr. Alexander Pretschner

Boltzmannstraße 3
85748 Garching bei München

Tel: +49 (89) 289 - 17386
https://www4.in.tum.de

sponsored by: itestra be excellent

Context

Program characteristics such as bug locations or software quality attributes are hard to determine without manual reviews. However, there exists evidence that software analysts already build a strong hypothesis based on their first impression of the code. When analyzing and labeling source code [1,2], several experts confirmed they have been able to get an accurate intuition of the quality of a code snippet by looking at a visual representation of its overall structure, without going into syntactic or semantic detail. This applies in particular to assessments about the complexity and understandability of code. In this study, we try to mimic this process by training machine learning algorithms on *images* of source code. In a first pre-study, we found encouraging first results using Neural Networks [3,4] and Support Vector Machines.

Goal

The goal of the thesis is to investigate the use of image classification for software quality prediction. The student researches image classification algorithms that are promising for this task and implements them prototypically. Furthermore, the student examines related literature and identifies applicable, state-of-the-art data preprocessing techniques to aid image classification. In particular, the main challenge in this domain is the small size of the available data sets. Thus, data augmentation techniques and the creation of further, artificial data points are worth investigating. Other measures include determining the optimal image size or used color themes. Eventually, the performance of the approaches is evaluated with either open source projects or example code provided by the industry partner.

This research is going to be conducted in cooperation with itestra GmbH. If you are interested in this topic, please follow the application instructions below.

Working Plan

1. Familiarize yourself with both image classification and software quality assessments
2. Research which image classifiers and preprocessing techniques are applicable for quality prediction
3. Implement these approaches prototypically
4. Evaluate which algorithms and configurations perform best to predict certain quality attributes
5. Record the studies, results, findings and the engineering approach in form of a thesis

Deliverables

- Source code of the implementation.
- Technical report with comprehensive documentation of the implementation, i.e. design decisions, architecture description, API description and usage instructions.
- Final thesis report written in conformance with TUM guidelines.

References

[1] Schnappinger, M., Fietzke, A., Pretschner, A.: Defining a software maintainability dataset: Collecting, aggregating and analysing expert evaluations of software maintainability. In: 2020 IEEE International Conference on Software Maintenance and Evolution (ICSME). pp. 278–289. IEEE (2020)

[2] Schnappinger, M., Fietzke, A., Pretschner, A.: A software maintainability dataset(Sep 2020). https://doi.org/10.6084/m9.figshare.12801215

Application:
Please apply via email to markus.schnappinger@tum.de. Your email should explain your interest in the topic and contain your current transcript of records. The most promising candidates will be invited for an informal interview. Upon mutual agreement, the thesis can be sponsored by itestra GmbH.

[3] Karpathy, A., Fei-Fei, L., Johnson, J.: Convolutional neural networks for visual recognition, stanford university. http://cs231n.github.io (2017)
[4] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25,1097–1105 (2012)

Fakultät für Informatik
Lehrstuhl 4
Software und Systems Engineering
Prof. Dr. Alexander Pretschner

Boltzmannstraße 3
85748 Garching bei München

Tel: +49 (89) 289 - 17386
https://www4.in.tum.de

sponsored by:

Application:
Please apply via email to markus.schnappinger@tum.de. Your email should explain your interest in the topic and contain your current transcript of records. The most promising candidates will be invited for an informal interview. Upon mutual agreement, the thesis can be sponsored by itestra GmbH.