Problemstatement

July 27, 2022

TECHNISCHE
UNIVERSITÄT
DARMSTADT

This research thesis will address the problem of completing code based on previously viewed data. Code completion is a Natural Language Processing (NLP) problem. However, unlike Text, Code has other constraints such as syntax and control dependency, which do not have the same relevance in plain text. For example, changing the order of a conditional statement would impact the meaning of the code greatly and give a compiling error based on syntax. Doing the same thing in natural language would perhaps change the meaning but could still be understood.

To tackle this, language models, such as transformers are trained on a code representation that includes those constraints. In Natural Language Processing, a transformer is frequently used. It is a powerful language model, that can generate text using the attention mechanism. This lets the transformer work with bigger datasets and handle long range-dependencies better compared to other language models. Previous works such as the Facebook Groups TravTrans have used Abstract Syntax Trees (AST) as a data representation. AST give access to the structure of the code, such as recognizing variables or showing what type of statement a line of code is.

The question arises, whether other kinds of code representations could improve the result of code completion using a transformer. For example, Control Flow Graphs (CFG) would add the knowledge of control dependencies whereas the order of two statements is not considered in an AST. Changing the order of the lines of code can have a great impact on the validity of the code, so this information could be quite beneficial.

To evaluate which data representation yields the better results, we will use the CFG that is generated as part of the knowledge tree in the GraphGen4Code tool by the wala group. This graph will be adapted to fit into the requirements of the TravTrans GTP-2 model. It will then be evaluated which data representation performs better in terms of accuracy and depending on results, further research will be done.