Robert Wallace
Cover Letter for Discussion Leader
CSE-60876 (Research Methods)

Paper: Su, CY., McMillan, C. Distilled GPT for source code summarization. Autom Softw Eng 31, 22 (2024)

General Topic: Distilling GPT model for improved portability and security

Specific Behavior or Activity Studied: Training smaller models to mimic GPT-3.5 level source code summarization on lower cost and higher privacy machines. Verifying that GPT-3.5 level summarizations are worth mimicking

Specific Research Questions:

RQ1 How well do summaries generated by GPT-3.5 compare to human-written reference summaries, across key quality criteria established in relevant literature?

RQ2 How closely do language models mimic GPT-3.5 for code summarization, across different model and dataset sizes?

RQ3 How closely does the distilled model mimic GPT-3.5 for code summarization, as measured by human experts?

Challenges

1: "The participant pool can be a threat to validity because online survey participants can fake work history"

2: "The GPT-3.5 version and prompt are threats to validity because GPT-3.5 is a commercial product and subject to change without notice, and also may give different answers with different prompts."

3: "The subject Java methods are also a key threat to validity because our study results could change with a different set of Java methods."

Paradigm: This paper aims for a qualitative method in two parts, first, the study comparing GPT-3.5 results to human-written summaries, and the second study comparing the GPT-3.5 and jam models, both using Mann-Whitney tests to show the significance of the results.

Problem: Automatic source code summarization is considered a "holy grail" for Software Engineering research, reducing the manual labor required for many tasks. However, current systems require handing over your code to a third-party, losing data custody.

Importance: Distilling GPT-3.5 quality summaries into a smaller model that performs fairly similarly while being able to run on a single GPU provides programmers with options for in-house summarization, as well as control over the training data.

Claims: The authors first claim that GPT-3.5 summaries are worth replicating and using as training data, present a couple models that distill this ability, producing models that while not strictly superior to GPT, are comparable and cheaper to run.

State of Knowledge: The authors draw from the existing research into source code summarization, from neural models to recent fine-tuning of LLMs for summarization tasks. Knowledge distillation is a relatively recent development, which aims to train a smaller model to mimic a the functions of a larger one, with better results being found when mimicking a small subset of the capabilities of the large one.

Evidence: The evidence is found from the studies run on Prolific, asking human programmers to compare the given summaries and give their preference on a accuracy/complete/concise rating. The methods and original human summaries themselves were pulled from a studied Java dataset, and the standard metrics were used to compare the results.

Story Structure: The authors introduce the background to source code summarization and give a brief introduction to the research. They then describe their two studies, discussing the threats and results of each in order. They then describe their results and conclusions at the end