

Final Project Proposal

Evan Gronberg

July 31, 2024

For my final project, I plan to fine-tune a model on the Bible then evaluate the model on the BibleQA dataset. [1] I will start by downloading the American Standard Version (ASV) of the Bible in plaintext, which has around 850K words. [2] I will then tokenize this corpus using the model's associated tokenizer, which should result in a fine-tuning dataset of roughly 1.13M tokens according to the conversion rate given by OpenAI. [3] From there, I will fine-tune the model; I plan to start with the GPT-2 model, although a different model may be selected upon evaluation of hardware capacity and time requirements. Parameter-efficient fine-tuning (PEFT) will be incorporated if needed; if time allows, it will be incorporated regardless for the sake of comparison. Next, I will perform evaluation on the aforementioned BibleQA dataset, a question-answering dataset with about a thousand Bible-related questions. My evaluation process will compare the fine-tuned model(s) against the model that was not fine-tuned using a measure of accuracy against the ~1K answers provided in the BibleQA dataset (i.e., the percentage that were answered correctly). Per most recent office hour discussion, this measure of accuracy will be a subject of determination in the process of building this project. To accomplish these steps, I will use Hugging Face libraries (at the very least, `transformers`). [4]

This project differs from previous course exercises in that fine-tuning was never implemented in those assignments. While there are many such techniques that we did not have time to implement, I have chosen fine-tuning for its relevance to real-world use cases I am likely to encounter in my career. A particular point of interest with this proposed fine-tuning project is that most every model includes the Bible in its pre-training dataset. It will thus be interesting to see if further fine-tuning on the Bible will result in enhanced ability to answer Bible-related questions.

[1] <https://github.com/helen-jiahe-zhao/BibleQA>

[2] <https://openbible.com/textfiles/asv.txt>

[3] <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>

[4] <https://pypi.org/project/transformers/>