Fine tuning a small language model to summarize product reviews and measuring forgetting

Shaunak Kapur & Pranav Krishnan

Introduction and problem statement

The goal of this project is to fine tune a small pretrained language model so that it can generate concise and clear one sentence summaries of product reviews. Summarization is a well defined text-to-text task where the model must condense information while preserving meaning, which makes it a useful setting for studying how targeted fine tuning affects model behavior. Product reviews are particularly suitable because they follow a consistent pattern, vary in length, and often contain a human-written summary that serves as an ideal reference target. The central aim of the project is to show how a lightweight model can be adapted to produce higher quality summaries compared to its baseline performance.

At the same time, the project examines an important side effect of model specialization, which is forgetting. When a model is fine tuned on a narrow domain, its performance on tasks outside that domain may degrade. This type of behavior is relevant in many real-world deployments where models are continuously adapted for specific applications. By training the model on summarization and then testing its ability to answer basic factual questions, I can observe whether the specialization comes at the cost of reduced general knowledge. The project therefore has two connected objectives: improving capability on the summarization task and measuring whether that improvement weakens the model's performance on unrelated prompts. Together, these goals make the project a balanced study of both gains and trade-offs from fine tuning.

Data sources

The primary dataset for this project will be the Amazon Fine Food Reviews dataset, a publicly available collection of about five hundred sixty-eight thousand product reviews. Each record contains a detailed review text written by a customer and a shorter human-written summary that captures the main message of the review. The dataset is provided as a CSV file named Reviews.csv and includes clearly labeled Summary and Text fields that can be used directly for supervised learning. Its size, structure, and availability make it ideal for training a model to map long-form text to shorter summaries.

To evaluate forgetting, I will create a small separate set of factual question and answer pairs. These questions will be simple examples of common world knowledge that the base model typically handles well. By testing both the base model and the fine tuned model on this question set, I can compare changes in accuracy and observe whether the fine tuning process alters the model's ability to answer basic queries. Combining a large authentic dataset with a small handcrafted evaluation set provides a complete foundation for exploring both summarization performance and forgetting.

High-level methods and technologies

The model chosen for this project will be google flan t5 small, which is a lightweight text-to-text transformer hosted on Hugging Face. It has been instruction tuned on a variety of language tasks and provides strong baseline performance while remaining small enough for efficient fine tuning. This balance makes it an appropriate model for exploring task-specific adaptation in a controlled setting.

I will begin by loading the dataset and preparing it for training. This includes removing empty entries, trimming overly long reviews for computational efficiency, and creating training, validation, and test splits. The full review text will be used as the model input, optionally preceded by a short instruction such as summarize this review, and the corresponding human-written summary will serve as the target output. This establishes a straightforward sequence-to-sequence mapping for the model to learn.

Fine tuning will be performed using a supervised learning setup with teacher forcing, following standard practices in sequence-to-sequence training. I will use the Hugging Face Transformers library to run the training loop, track validation loss, and save model checkpoints. After training, I will evaluate the summarization quality on the test set using automatic metrics such as ROUGE, which measure how closely the model's summary matches the human-written one. I will also examine qualitative examples to understand how the fine tuned model improves in clarity, relevance, and faithfulness to the original text.

To measure forgetting, I will test both the base model and the fine tuned model on the separate factual question set. I will compare their answers with the expected responses and look for changes in accuracy or consistency. Even small shifts in performance can indicate that specialization has influenced the model's broader abilities. This methodological approach combines structured model training, quantitative evaluation, and analysis of secondary behavior changes, fulfilling the requirement to outline the high-level techniques and technologies used in the project.

Products to be delivered

The final products of the project will include a well organized git repository containing the full codebase for data preprocessing, model training, evaluation, and generation of example summaries. The repository will also store the fine tuned model checkpoint or adapter weights, depending on the implementation. I will provide quantitative results comparing the base model and the fine tuned model using ROUGE scores and test-set performance, as well as qualitative examples that highlight improvements in summarization quality.

The forgetting analysis will include a clear comparison of the model's factual question accuracy before and after fine tuning, along with specific example questions where the answer changed. These comparisons will help illustrate whether the fine tuning process influenced the model's general capabilities. A written report will describe the project from motivation through results, and a short presentation video will summarize the main points, methods, findings, and conclusions. These deliverables align with the grading expectations and provide a complete picture of the project's outcomes.