Summary of RAFT: Adapting Language Model to domain-specific RAG

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, Joseph E. Gonzalez

Large Language Models (LLMs) have proved a significant stride in general knowledge tasks. However, when it comes to specialized domains, their response is hindered and hallucinatory. Existing solutions such as Augmented-Retrieval Generation (RAG) or Fine-Tuning expands the model's data, yet they often struggle with irrelevancy. This paper introduces RAFT (Retrieval-Augmented Fine-Tuning). This novel method aims to incorporate domain knowledge to enhance the ability of LLMs in domain-specific Retrieval-Augmented Generation (RAG) tasks and improve accuracy against irrelevant retrieved data while leveraging RAG and Fine-Tuning methods.

The paper compares adapting an LLM to domain-specific knowledge and an "Open-Book Exam." In in-context learning such as RAG, LLMs have no prior domain knowledge. They are only presented with the documents externally during inference. They are also paired with a retriever that fetches relevant documents that are later added to the user's prompt. This method depends significantly on the retriever and its accuracy to extract proper data. However, in-context learning fails to utilize the external documents to enhancing the model's knowledge. Alternatively, the supervised fine-tuning approach is considered as a"Closed- Book Exam". LLMs are further trained on domain-specific knowledge but have no access to external data. Finally, the "Domain-Specific Open-Book exam", which is the main focus of this paper is an approach where the LLM has been fine-tuned on domain-specific knowledge and can reference relevant data using RAG, allowing it to focus on relevant documents and ignore distractors for improved performance.

The authors also introduce the classical technique of supervised Fine-Tuning, which is a process where LLMs are trained on question-answer pairs, allowing it to learn patterns between questions (**Q**) and answers (**A**). After Fine-Tuning the model can answer new questions directly or use external documents (**D**) in RAG inference to provide more accurate, domain-specific answers. Besides, they introduce RAFT which includes the idea of "golden" documents **(D*)** that contain relevant information and "distractor" documents **(Di)**, which are irrelevant. The model is trained with both types. For a fraction **P** of the dataset, training includes both golden and distractor documents, while the remaining **(1-P)** fraction contains only distractors, prompting the model to memorize answers instead of deriving them from context. During testing, the model is presented with a question and the top-k retrieved documents by the RAG pipeline where RAFT is completely independent of the retriever. Additionally, RAFT employs chain-of-thought (**CoT)** reasoning, where the model is presented with question, context, and verified answers and then asked to explain its answer with a reasoning chain that appropriately references the original context.

For RAFT evaluation, the authors use a comparison approach between RAFT with **(Di)**, and several baselines without **(Di)**. LlaMA2–7B-chat model was used alongside with LlaMA2–7B+RAG, LlaMA2–7B-DSF, and LlaMA2–7B-DSF+RAG. Also, GPT-3.5 + RAG was used as a reference. These baselines were fine-tuned and tested using the following datasets: PubMed, HotPot, HuggingFace, Torch Hub, and TensorFlow. In the results section, the paper presents a detailed comparison table between RAFT and the previous baselines. DSF significantly improved LLaMA-7B model performance, but adding RAG to DSF did not lead to further improvements. However, RAFT performed better than GPT-3.5+RAG in 4 out of 5 datasets. Moreover, the paper emphasizes the value of CoT in RAFT's success. Integrating CoT significantly improves training robustness, with results showing a 14.93% improvement when CoT was included in the HuggingFace datasets. Additionally, the paper illustrates RAFT's distinction from DSF using a qualitative comparison where the results show that the DSF often produces incorrect answers, while RAFT consistently identifies the relevant context and answers correctly. Furthermore, the paper explores whether it's always necessary to train LLMs with golden documents. The findings suggest that using a mix of golden and distractor documents during training helps improve model robustness. The paper also investigates how RAFT performs when the number of retrieved documents varies at test time. The results show that RAFT maintains strong performance even when faced with an increasing number of distractor documents.

The authors also review related work on Retrieval-Augmented Language Models, which enhance LLM performance by integrating retrieval modules for tasks like open-domain question answering. They also address the extent of memorization in LLMs, highlighting concerns about understanding versus memorization and privacy issues. Furthermore, recent studies explore fine-tuning pre-trained LLMs for RAG tasks, particularly in cases where test documents match the training set.

To conclude, RAFT is a training strategy for enhancing domain-specific question answering in "open-book" settings, using distractor documents, incomplete contexts, and chain-of-thought reasoning. Evaluations on PubMed, HotpotQA, and Gorilla API Bench highlight its strong potential.