Paper Review: RA-DIT: Retrieval-Augmented Dual Instruction Tuning

Austin T. Barton

1 Summary

They propose Retrieval-Augmented Dual Instruction Tuning (RA-DIT), an approach that retrofits any LLM with retrieval capabilities via fine-tuning over a set of tasks selected to cultivate knowledge utilization and contextual awareness in the language model predictions. The fine-tuning is done separately in two parts: 1) retrieval fine tuning (R-ft) and 2) the language model fine tuning (LM-ft). They initialize the framework using pre-trained LLAMA and a dual-encoder based dense retriever call DRAGON+. They retrieve relevant text chunks based on the LM prompt which are then prepended to the prompt, and the predictions from multiple chunks are computed (in parallel) and ensembled to produce the final output. They show that this method of optimizing each of the components independently alone results in significant improvements in knowledge intensive tasks and competes effectively with retrieval augmented language models (RALM) that have undergone extensive pre-training from end-to-end. They demonstrate their results on knowledge intensive and commonsense reasoning tasks with comparisons to other state-of-the-art RALMs (ATLAS). Fine-tuning strategies and an ablation study to determine the contribution to improved performance is summarized at the end followed by concluding remarks and discussions.

2 Strengths

The biggest contribution this paper has is extending the REPLUG paper (Shi et al., 2023) to not only fine-tune the retriever while treating the LM as a black box, but also fine-tuning the LM to use the retrieved chunks more effectively. They waste no time getting to the point and illustrating to the reader exactly what their work is and where it lies in the current space of research. They provide necessary detail, not assuming familiarity with the REPLUG retrieval fine-tuning method, in a concise manner that is correct, coherent, and digestible so that readers can understand their process and use it in the future. Their paper is broken down into relevant and useful pieces of information in such a manner that for the practicality of using this method, it's easy to scan through and get necessary information to move onto work that applies it. It's also quite helpful in teaching readers that may be less familiar with current works in retriever and language instruction fine-tuning. Their experiment setup and results sections are clear, methodical, scientific, and avoids unnecessary information. They utilize both a state-of-the-art retriever and language model for their experiments, making their results and findings all the more relevant.

3 Weaknesses

A large weakness is in their experimentation. They compare their models with ATLAS, an extensively pre-trained RALM. However, ATLAS uses the Contriever dense passage retriever whereas their RAD-IT model uses the DRAGON+. In their own experimentation with their RAD-IT method, they chose DRAGON+ because it outperformed the Contriever for identical tasks. Meaning that ATLAS could potentially still outperform RAD-IT if its dense passage retriever was changed to DRAON+. This important distinction in the comparison of these RALMs is not mentioned in the paper either. Although their contributions are very useful, I think that they could have done more in their experimentation and ablation studies. This work is not extremely novel and groundbreaking, it's a very logical next step or extension from previous work. This doesn't make it not novel or not important, but I think that understanding where the purpose of this work lies in the current space of research and leaning towards what makes it the most useful would make it better. I believe its purpose is verifying and thoroughly analyzing this natural extension to REPLUG's work with extensive experimentation. The writing style is also very repetitive. It continously re-emphasizes the same points about what their work is and why it matters. This comes across as though they are selling this idea to the reader rather than presenting it. I would prefer to see more detailed and nuanced discussion of their own findings than continual self-verification that their work indeed matters as well as rehashing the overall idea of their work over and over.

4 Extensions

I would recommend performing the same benchmark comparison with ATLAS except with the Contriever as the retriever for both RALMs and also attempt to replicate ATLAS except with the DRAGON+ to see if the RAD-IT approach actually competes with ATLAS given the same retriever. Other extensions would be to create an automated pipeline for dual instruction tuning for practical applications.