RePlug: How about Leveraging ChatGPT to Upgrade Bing Search Algorithm

Meng Jiang

"RePlug: Retrieval-Augmented Black-Box Language Models" by Shi et al. is a very interesting work. It treats the language model (LM) as a black box and augments it with a tuneable retrieval model. The study shows that the **LM can be used to supervise the tunable retriever**, which can then find documents that help the LM make better predictions. Given that the retriever model is one of the core components of a search engine and Microsoft Bing retains less than 9% share of the global market, this article discusses possible reactions that Microsoft Bing may have on the RePlug's discovery.

Numerous industrial products were incubated from lab research. Generalizing the discovery from the lab research to real applications is fun and valuable. One of the discoveries in the work of RePlug with LM-Supervised Retrieval (LSR) is that it significantly improves (1) the performance of **GPT-3** by 6.3% on language modeling, (2) the performance of **Codex** by 5.1% on massive multi-task language understanding, and (3) the performance of **Codex** by 5.0-12.0% on open-domain question answering. Clearly, the *tunable* retriever model (i.e., **Contriever**) was significantly improved in RePlug, though there was no direct retrieval evaluation.

OK. What does Microsoft have? It has access to the world's (so far) best black-box LM **ChatGPT** (developed by OpenAI). Google's LM products may not be comparably good yet. What does Microsoft *not* have? It does not have the world's best search engine; more precisely, its **Bing** search is far less accurate than Google search, reflected in market share. ***If Bing search is considered as a tunable retriever and ChatGPT is the LM powerful on various tasks, can the RePlug framework leverage ChatGPT to upgrade Bing to be the most accurate search engine?*** There are at least two side factors/questions around Bing vs Google: (1) Does Bing have access to the same, similar, or bigger amount of crawl data for search? (2) Can we assume that the LM's performance on the various tasks reflects user experience with the search engines? I don't have a good answer. In RePlug's experiments, all LMs are given the same data for retrieval and tested on the same set of tasks.

Before we go ahead to tune the Bing search engine, let's think about a conspiracy theory: the retrieval model behind Bing may not be significantly improved by tuning on any feedback. When Microsoft found its search engine was falling behind, it could create a great number of bots, post queries to Google search API, collect ranked pages, use them to compute retrieval likelihood and LM likelihood (as described in the paper on RePlug LSR), and ultimately improve the Bing's retrieval performance. Using one's own LM API is cheaper (and safer) than the competitor's API, however, in any case Microsoft might have tried to improve Bing in the aforementioned way, as we have observed, the result is not good. So, I would suspect any positive answer to the question (in Italic font) in the above paragraph. RePlug may not be able to leverage ChatGPT to upgrade Bing, but who knows?

about 500 words
Feb 22, 2023 (created)
Feb 22, 2023 (completed)