Cover Letter for Paper Discussion

Zhihan Zhang

This paper is an analytical study that focuses on the topic of in-context learning (ICL) of large language models (LLMs). In-context learning refers to LLMs performing a new task via inference alone by conditioning on a few input-label pairs (demonstrations) and making predictions for new inputs. Although being a common practice when using LLMs, before this paper, there was little understanding of how models learn from in-context demonstrations. Therefore, this paper dives deeper into the design of in-context demonstrations and reveals insights about which aspects of the demonstrations matter most in the model's ICL performance.

In the paper, the authors first analyzed whether label correctness is a key factor in ICL. Counterintuitively, they found that replacing correct labels with random labels barely hurts performance. They further proved their finding to be consistent across different numbers of demonstrations and different templates. To move further, they did a series of controlled experiments that targeted a range of possible factors in ICL. These include:

- The distribution of the input text: The authors replaced the inputs of demonstrations with random text from an external corpus to create an out-of-distribution setting.
- The label space: The authors used random English words to replace the labels.
- Input-label pairing: The authors tried removing either the input or the label from the demonstrations.

When doing these experiments, they leveraged 26 tasks and randomly sampled demonstrations from the training set. They repeated such a strategy 5 times with different random seeds to reduce the amount of randomness. After experiments, they found that all these three factors contribute to the success of ICL.

Combining these with their previous finding that ground-truth labels are not very critical for ICL, they concluded that LLMs do not "learn" how to solve the tasks during ICL, but the problem-solving abilities come from the pre-training stage of LLMs. Instead, LLMs adapt themselves to the input and label distributions suggested by the demonstrations, which ultimately helps them predict tasks more accurately. In summary, this paper brought new insights into the ICL performance of LLMs, and helps NLP researchers better understand the learning mechanisms and test-time behavior of pre-trained LLMs.