THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Master's Thesis Presentation

Zijian Zhao
Department of Statistics
The University of Chicago

"Knowledge Editing through Contrastive Fine-tuning"

April 23, 2027, at 4:00 PM
Jones 111, 5747 S. Ellis Avenue

Abstract

When one considers training a large language model (LLM), it may require significant computational resources and a long period of time. It is not possible that one can train and deliver an LLM product frequently. Between two versions of an LLM, there must be some information updates in the real world such that the training data would change accordingly. This means the old version of LLM would give some outdated outputs during this period. Therefore, one may wonder if the updated information could be used to do knowledge editing (KE) such that one would edit the knowledge in the old version and would not need to train the model completely.

The goal of the thesis is to design a module that follows an LLM such that when a prompt is input into the LLM and the LLM generates an answer, the module can validate the answer and give a more reliable answer when the original one seems outdated. Our model will establish a score function by a neural network to intuitively act as a metric of similarity. Its input will be the prompt and a candidate answer for that prompt, and its output will be a scalar that can be understood as a distance or quality between the two components. If the score is low, then the answer is likely to be true. Especially when comparing two answers, one can simply compare the scores to decide which one is more reliable. We will use contrastive learning as a perfect way to implement for this model, where we maximize the score for undesired answers and minimize the score for our new answer. The main structure would be BERT connected with an MLP for the score function implementation.

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.