ANLY-580 Natural Language Processing

Arithmetic Reasoning System Using GPT-3 with Chain-of-Thought Prompting

Group Members: Dian Zhi | Ruobing Yan | Muwen You | Zheyuan Hu
NetID: dz247 | ry201 | my549 | zh216

High level description of what you plan to do, and why you think it is interesting:

Modern large language models (LLMs) can execute many NLP related tasks as well as humans, but they still have difficulty doing complex reasoning in several steps reliably (e.g. arithmetic reasoning). Wei et al. (2022) explored a novel way (CoT, chain-of-thought) of generating a few-shot learning prompt for complex reasoning tasks which improved the performance of large language models. Zhang et al. (2022) improved upon this idea of manually generating CoT prompt by designing a method to use LLMs to generate CoT prompt automatically. Following the automatic CoT idea, we explore new methodologies to use LLMs to generate CoT prompts automatically to improve the accuracy of arithmetic reasoning.

A detailed description of the methods you will use to achieve this:

Since the large language models do not perform so well on the reasoning tasks, and the performance won't have a significant improvement even when scaling up the models, the Chain-of-Thought prompting was proposed to help improve LLM's reasoning ability. The Chain-of-Thought prompting is a series of intermediate natural language reasoning steps that lead to the final output. With the CoT as an in-context few-shot learning examples, the large language model can have a higher accuracy on solving the reasoning problems.

A common format of the CoT(Chain-of-Thought) consists of triples: <input, chain-of-thought, output>. This method has been proved to reach a SOTA on GPT-3 and other large language models. Step on that, we hope to further explore the ability of CoT, by introducing auto-CoT.

In the original auto-CoT paper, questions will be clustered first, the most representative questions will be selected and entered into GPT-3 with keywords "Let's think step by step" append to the end. These keywords will elicit GPT-3 to generate a CoT reasoning and the answer of the entered question. Then these automatically generated CoT will be used as few-shot learning examples for any newly entered questions, therefore, generate an even higher accuracy on the reasoning task for any unseen problems. In our project, we will try to skip the clustering step and calculate the distance between queries and learning examples and use the most similar question to generate CoT examples. The questions will be encoded by using Sentence-BERT and the distance between questions will be measured using cosine similarity.

What modeling approach do you intend to use?
GPT-3, with Chain-of-Thought prompting. Sentence-BERT

What data do you intend to use?
In this project, the GSM8K, a dataset of 8.5K high quality linguistically varied grade school math word problems, will be used to explore the possibilities of using automatic CoT promptings to improve the arithmetic reasoning performance. GSM8K is a collection of 8.5K expertly crafted, human-written problems for elementary school math. 7.5K training problems and 1K test problems were separated out of these. Between 2 and 8 steps are required to complete these problems, and the majority of the time, the solution is reached by conducting a series of simple calculations utilizing the (+ - / *) arithmetic operations. Any of these problems should be able to be resolved by a capable middle schooler.

In the dataset files, each line represents a single elementary school arithmetic problem in the form of a json dictionary (with a "question" key and an "answer" key). The final numeric solution is the last line of the answer, followed by ####, and the formatted answer makes use of calculation annotations.

Dataset Link (GSM8K): https://github.com/openai/grade-school-math

How will your system be evaluated and what are the evaluation criteria?
The arithmetic reasonings generated by GPT-3 will be parsed into answers, and the answers will be compared to the ground-truth answers. The accuracies of the arithmetic reasoning will be evaluated and be compared to the accuracies of answers generated by GPT-3 given traditional prompting.

Are there any special computational/hardware considerations?
Finding the proper distribution method and hardware setup for an LLM's tendency to expand is one of the main hardware requirements. For different AI models and other stacks, there isn't a one-size-fits-all solution available. AI models like GPT-3 might get so large that it can't fit on a single GPU.

In this project, we will use two modern language models: Sentence-BERT and GPT-3. We plan to run S-BERT on our local machine with 1 Nvida GTX 1080 GPU, and we plan to run GPT-3 using API provided by OpenAI.com.

What are the biggest unknowns that might dictate the success or failure of this project?
We are not sure that after modifying the prompting method, our model will perform better than the original one or not.

How will the results of your work be presented? Will this be a live demo, a written report, a slide deck + oral presentation? Demos can be given along with reports/presentations.
Slide deck & oral presentation with a live demo.

Work Cited:

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint arXiv:2210.03493*.