ECE472 – Quiz 9

Jonathan Lam

November 18, 2020

1. *Identify something you find "unsavory" about these kinds of models; explain and argue.* (arXiv:1810.04805, arXiv:2005.14165, arXiv:2007.14062)

   I chose to focus on some of the experiments in the GPT-3 paper, as this seems to be the one generating the most excitement recently. (I also watched this interview with GPT-3 that is fascinating). While the empirical evidence (e.g., that video interview) is very convincing in being able to demonstrate the kinds of questions that GPT-3 is able to provide rational (and sometimes humorous) responses, it's still very hard to see when things are canned and when they are not.

   On the other hand, the BERT and Big Bird papers do not have this problem: they focus on domain-specific tasks in NLP (the Big Bird paper also tackles genomics, but it provides a convincing argument how this is similar to NLP tasks), e.g., GLUE and SQuAD.

   With the use of a state-of-the-art NLP language model (transformers, the BERT model), I believe GPT-3's ability to do comprehension and language generation tasks are impressive and well-justified. However, when it comes to some other tasks mentioned – in particular the tasks involving some sort of factual lookup (information retrieval) or reasoning (logical reasoning and arithmetic) – I don't think that their claims are very satisfactory ("unsavory"?). Perhaps my bias as an engineer makes me overly skeptical that this LM can actually perform more generalized (non-language-related) reasoning tasks. Fleshing out these tasks, as the GPT-3 paper describes them:

   "Closed Book Question Answering" (Information Retrieval) The authors state that they "measure GPT-3's ability to answer questions about broad factual knowledge" in a "closed-book" manner, i.e., they don't allow training on external, content-specific resources, and (like the rest of the paper) they don't perform fine-tuning. GPT-3 has SOTA performance here.

   "Common Sense Reasoning" This task "consider[s] three datasets which attempt to capture physical or scientific reasoning, as distinct from sentence completion, reading comprehension, or broad knowledge question answering." Examples include questions from pre-college science exams, and some questions "which simple statistical or information retrieval systems were unable to correctly answer." GPT-3 has SOTA performance here.

   Arithmetic This includes two- to five-digit addition and subtraction, two-digit multiplication, and one-digit composite (composition of operations). GPT-3 reaches high accuracy on the two- and three-digit addition and subtraction on the largest model (175B parameters), but does not reach high accuracy for the other tasks or in smaller models. With

these results, the authors conclude that "Overall, GPT-3 displays reasonable proficiency at moderately complex arithmetic in few-shot, one-shot, and even zero-shot settings."

The problem I have with these three in particular is that they all have some claim to logical reasoning (unrelated to the LM), but it seems to me that all of these could be attributed to having examples in the training set. With such a large input set from the Internet, even after some filtering and deduplication I wouldn't be surprised if many of their results are achieved by finding similar results in the training dataset rather than by doing what we might consider "logical reasoning." Information retrieval is inherently dependent on the information content of the training dataset; even if it is "closed-source" as in their definition (i.e., not training on specialized data sources), the fact that the information may exist in the dataset makes it seem unremarkable that GPT-3 should perform well as an information retrieval system. On both the common sense reasoning and arithmetic tasks, I imagine that many common questions of a similar form may be encountered on the web (e.g., KhanAcademy and the wealth of educational websites). Even though the authors mention that they attempt to reduce data contamination (and they had a bug that allowed some unwanted contamination) and specifically mention that they attempted to search for the math problems in their training dataset to see if the arithmetic was memorized, I am still not too convinced that the overlap was not large. For "common sense problems," there are so many variants of the same questions on the Internet that I wouldn't be surprised if many of the answers to the logical reasoning questions were memorized in some slightly different form that survived the data contamination filtering. In particular for arithmetic, there are multiple forms in which an arithmetic operation may be shown; if it came in the form of a word problem, and GPT-3 correctly parses the word problem and "extracts" the arithmetic, then this goes to show the strong LM in GPT-3 but not necessarily its arithmetic/computational/logical reasoning skills. Also, the fact that it did well on two- and three-digit numbers could be another indication of memorization (this is not a high number of permutations out of the hundreds of billions of tokens in the training dataset); good performance on larger operands would be more convincing.

I should qualify my criticisms: I do not know too much about the nature of the Common Crawl dataset, and perhaps I am assuming that it is larger and more all-encompassing than it actually is. (I cannot imagine what a trillion words can contain. My mind is limited to understanding what a few terabytes of digital data can hold.) Of course, working at such a scale (of parameters, data (and data filtering to prevent contamination), etc.) is unprecedented, and the work at OpenAI is all very fascinating. I just would like to see more research into the claims about any reasoning abilities past the scope of conventional NLP metrics as being more than memorization, because I believe the performance they achieve, due to their training methods and assumptions, might be due largely to memorization of examples in the training set.