CS410 Fall 2022 Technology Review

GPT-3 Language Models - Chenfeng Chu; NetID: chu61

Introduction

In current state of art language models, a major limitation is that there is still a need for task-specific dataset and fine tuning to achieve strong performance. GPT-3 was an attempt to address such limitation. If such obstacles are removed, it will bring several benefits.

First, it reduces the need to collect large supervised training dataset, which is very difficult to collect in many cases in real life. Second, the narrowly scoped fine tuning may achieve high performance for particular tasks, but it is typically hard to generalize. In other words, the actual performance on the underlying tasks may have been exaggerated if extended further. Third, humans do not require large dataset to learn and perform most language tasks. Removing the dependencies on large supervised dataset in training a language model would be critical to getting closer to human-like learning capabilities.

Approach

Observing the improvements brought to NLP tasks by the increasing capability of transformer language models, GPT-3 was trying to test the hypothesis that in-context learning abilities might show similarly strong gain with scale of model parameters. In context model means that the model develops a broad set of skills and pattern recognition abilities at the training time, and leverage them at inference to rapidly adapt to or recognize the designed tasks.

In particular, GPT-3 trains a 175 billion parameter autoregressive language model and such model was evaluated against over two dozen widely recognized NLP datasets in the field. For each NLP task, GPT-3 was evaluated under three conditions:

1. "few-shot learning", where as many demonstrations are allowed when fitting into the model's context window (typically 10 to 100)
2. "one-shot learning", where only one demonstration is allowed, and
3. "zero-shot" learning, where no demonstrations are allowed and only an instruction in natural language is given to the model

GPT-3 uses the same model and architecture from GPT-2, including the modified initialization, pre-normalization and reversible tokenization. However, in the layer of transformer, GPT-3 uses alternating dense and locally banded sparse attention patterns, similar to the Sparse Transformer [CGRS19].

The majority of the training datasets used in GPT-3 is Common Crawl dataset. Additional data preprocessing steps were applied before using this dataset to train the GPT-3 model: (1) downloaded and filtered a version of Common Crawl based on similarity to a range of high-quality reference corpora, (2) performed fuzzy deduplication at the document level to prevent redundancy and preserve the integrity of held-out validation set, and (3) added known high-quality reference corpora to the training mix to augment Common Crawl and increase its diversity. The final mix of training dataset for GPT-3 is summarized below.

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Figure 1: Dataset used to train GPT-3

To train such a large model, the team used a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. All models were trained on v100 GPU's, part of a high-bandwidth cluster provided by Microsoft.

Results

GPT-3 model was evaluated across 9 categories of NLP tasks: traditional language modelling tasks, closed book question answering tasks, translation, Winograd Schema-like tasks, commonsense reasoning, reaching comprehension tasks, SuperGLUE benchmark analysis, Natural Language Inference, and in context learning abilities.

Broadly, GPT-3 achieves promising results on those evaluation tasks. This is especially the case for few-shot setting as it is competitive with or even occasionally surpasses state-of-the-art performance from fine-tuning NLP models in the domain. For example, GPT-3 achieves 64.3% accuracy on TriviaQA in the zero-shot setting, 68.0% in the one-shot setting, and 71.2% in the few-shot setting, the last of which is state-of-the-art relative to fine-tuned models operating in the same closed-book setting.

Another key observation from the evaluation is that larger models are more proficient at in-context learning, which can be illustrated by the diagram below that the average model accuracy has been increasing along with number of model parameters under each condition.

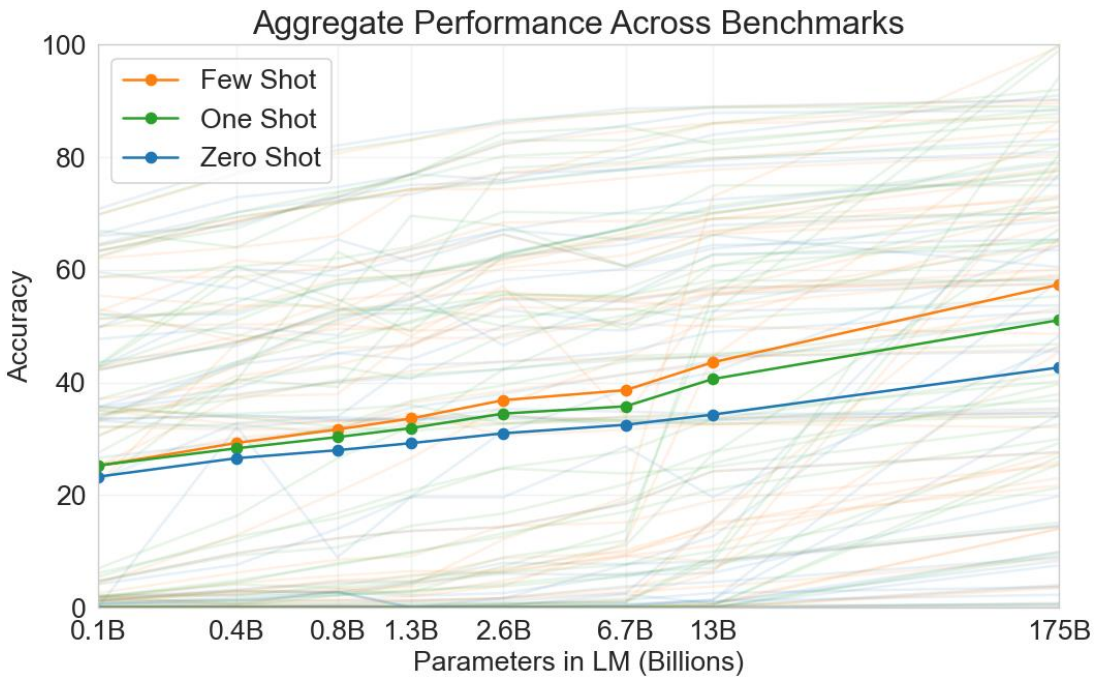

Figure 2: Aggregated performance for all 42 accuracy-denominated benchmarks in GPT-3

Limitations

Despite the quantitative and qualitative improvements of GPT-3, it still has several limitations. First, it performs poorly in text synthesis and several NLP tasks. Second, it has several structural and algorithmic limitations including not encompassing bidirectionality in the language model. This will impact particular NLP tasks such as fill-in-the-blank tasks, tasks that involve looking back and comparing two pieces of content, or tasks that require re-reading or carefully considering a long passage and then generating a very short answer. Third, the model fails to consider customization prediction for each token (currently all tokens are given equal weight) and purely scaling parameters may not improve performance for some NLP tasks, especially those performing well with fine tuning from reinforcement learning.

Conclusion

The birth of GPT-3 apparently demonstrated major breakthrough in the NLP world. The 175 billion parameter langue model showed strong performance on many NLP tasks and in some cases, close to the performance from state-of-the-art fine-tuned systems. It offered NLP researchers additional evidence, hope and directions that large language model could be one of possible solutions for adaptable and general language systems in the future.

Reference

Brown, T. B. (2020, May 28). Language Models are Few-Shot Learners. arXiv.org. https://arxiv.org/abs/2005.14165