Language Models are Few-Shot Learners

Tom B. Brown et al.

https://arxiv.org/pdf/2005.14165.pdf

Large Language Models (LLMs) are neural networks (developed for Natural Language Technology (NLT) capabilities (Natural Language Processing (NLP), Natural Language Understanding (NLU), Natural Language Generation (NLG)). Recently, LLMs have been increasingly used for task-agnostic applications, also known as transfer learning. Task-agnostic refers to the application of LLMs for generalized purposes or purposes the LLM was not originally trained for. For example, being used for clinical documentation, medical coding, clinical decision support, Diagnosis Related Group (DRG) determinations and many more. The architecture, dataset the LLM is trained on, and fine-tuning the model are all integral for an LLM to perform well on a given task. Recently, the size of datasets and size of the model itself has reduced the need for task-specific architectures.

This paper seeks to explore how this limitation -LLMs requiring the use of task-specific datasets for task-specific fine-tuning - may be avoided. Reducing this burden would allow a greater usage of LLMs in a wider variety of applications in at least 3 ways: 1) there is a large amount of natural language tasks and curating a large corpus of data for each individual task very difficult and very time consuming, 2) LLMs – including the massive ones – seem to by hyper-specific to the data they are trained on and have a hard time generalizing out of the training distributions, 3) humans can look at zero, or few, examples of a given task and be able to perform it – seamlessly performing multiple tasks in parallel or sequence and it would be beneficial for LLMs to be able to perform similarly.

Zero, one and few-shot learning are approaches to ‘in-context’ learning which involves giving the LLM zero, one or few examples of the task they are asked to perform. This contrasts with fine-tuning because only a few examples of the task are given rather than a fine-tuning the LLM with a full dataset specific to the task (which involves updating the weights too). The authors applied zero, one and few-shot learning to ChatGPT 3 and compared the performance on common NLT benchmarks to other models that were fine-tuned.

Generally the few-shot models had the greatest accuracy compared to zero and one-shot models. And given a large enough LLM (usually 13B or 175B parameters), the few-shot models were able to compete and sometimes surpass the state of the art (SOTA) model. The paper is extensive in the evaluation of the LLMs, and not all results can be covered here.

Some of the tasks that were measured include: completion tasks (finishing a text segment), question-answering (QA), translation (between languages: IE Spanish -> English), Winograd schemas (identifying which pronoun a word is describing), among others. The few-shot models surpassed the SOTA models in the following tasks: 2/4 completion tasks, 1/3 QA tasks, 2/6 translation tasks. One of the completion tasks – LAMBADA – GPT-3 surpassed the SOTA by 8% accuracy, totaling 76%. For the TriviaQA task, the Zero-shot had an accuracy of 64%, surpassing the fine-tuned T5-11B model by 14%. The few-shot GPT-3 had an even greater accuracy of 71.2%. There are many more results where the few/one/zero shot outperformed the SOTA, and if not came close, highlighting the potential these models have.