Nicholas DaRosa

ndarosa2

CS410 Text Information Systems

06 November 2022

Technology Review: Generative Pre-trained Transformer 2 (GPT-2)

The boom of machine learning has created great interest in advancing natural language processing tasks such as text generation. Although some trained models are great at specific tasks they have been trained on, their abilities are not generalizable. This lack of generalizability is primarily due to the use of supervised learning on focused datasets where the test data has the same characteristics as the training data. Consequently, there has been a push to develop models that have the ability to perform well at many tasks and perform well on tasks the model was not specifically trained for. In other words, the goal is to have a model with a more general intelligence when it comes to natural language processing tasks. A language model that has made great strides in being a generalist is the Generative Pre-trained Transform 2 model, also known as GPT-2.

Paramount to the success of a model is the dataset it is trained on. In the case of GPT-2, it was trained on a dataset developed specifically for GPT-2 called WebText. WebText was created by scraping outbound links from Reddit which resulted in Webtext containing the text from about 45 millions links and after cleaning, about 8 million documents. Compared to past models, this corpus was general and did not focus on a specific task or domain of interest. As a result, the dataset lends itself better to training a general language model.

Central to GPT-2 is using language modeling to create a probabilistic framework with the goal of estimating the probability of a given string. Two common types of language models are character level and word level. In the case of GPT-2, byte pair encoding is used since it is in between character level and word level language modeling. Byte pair encoding enables GPT-2 to benefit from character level language modeling's generality, while also benefiting from word level language modeling's empirical strengths.

Furthermore, the model architecture of GPT-2 is vital to its success. GPT-2's architecture closely resembles the OpenAI GPT model and like the OpenAI GPT model, the architecture is based on transformers. The largest version of GPT-2 has approximately 1.5 billion parameters with a vocabulary of 50,257 and a context size of 1024 tokens.

A great test of the generality of a language model is to test it on datasets it has never seen before and has not been fine tuned for. This testing practice is called "zero-shot" testing and differs from traditional testing in which a single dataset is broken up into train and test sets. The authors of GPT-2 tested the model on a variety of text datasets such as LAMBADA, WikiText-2, and the One Billion Word Benchmark. GPT-2 achieved state of the art results on 7 out of 8 of the tested datasets which is astonishing given that it was not previously trained on those datasets.

GPT-2 was tested on a variety of tasks such as reading comprehension, summarization, translation, and question and answering. GPT-2's zero-shot performance is comparable to supervised baseline models when it comes to reading comprehension, but on the task of summarization, GPT-2's generated summaries are not usable and are worse than the summaries generated by baseline models. Its English to French capabilities are better than many unsupervised baseline models, but still performs worse than the current state of the art English to French translation model.

GPT-2 has made great gains in furthering the generality of natural language processing models. After being trained on a large corpus of general text, GPT-2's transformer based architecture allowed it to perform above baseline models on datasets it was not previously fine-tuned for and have comparable results to baselines models in tasks it was specifically trained for. Nevertheless, there is still much progress to be made in making a model that is both highly accurate for a variety of tasks and highly general.

References

1. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.