Technology Review Submission

William Shih, November 6, 2022

ws21@illinois.edu

Language Models: GPT-2 and GPT-3

Recent advances in Natural Language Processing have resulted in large strides in performance in a variety of NLP tasks such as question answering, translation, reading comprehension, and generating text. GPT-3 is a recent language model that produces text on average indistinguishable from a real human. Humans had just a 52% chance, which is essentially a coin flip, of correctly identifying texts produced by GPT-3. A future language model may improve further from GPT-3 and make it nearly impossible to know whether a computer wrote the text. These advances have been enabled by a machine learning model introduced in 2017 called the Transformer. Developments using the transformer have been recent and rapid. GPT-2 (Generative Pre-Trained Transformer 2) was introduced in February 2019 and GPT-3 was introduced in June 2020.

The simplest type of neural network is the feedforward neural network. These have no cycles, but the input size is fixed. Recurrent neural networks were then derived from feedforward neural networks. RNNs can process variable length sequence of inputs, which is clearly important for NLP tasks that involve varying number of words as an input. For translation, RNNs without attention take in the whole input word by word and then from that single input, put out the translation word by word. Attention mechanisms were then added to RNNs that allowed the neural network to translate parts of a sentence at a time instead of translating the whole sentence simultaneously. Before transformers, RNNs with attention mechanisms were state of art for certain NLP tasks, which is just 5 years ago from 2022.

Transformers improve RNNs by removing the recurrence part of RNN and processing the entire input simultaneously. This allows easier parallelization while training the neural network and captures long range dependencies better than RNNs as no part of the input will ever be lost, which was a large problem with RNNs, even with attention mechanisms. With the ability of parallelize training, state of the art transformers are using larger and larger training datasets with an extremely high number of parameters. GPT-2 has a total of 1.542 billion parameters and GPT-3

has a total of 175 billion parameters, so GPT-3 has more than 100 times as many parameters as GPT-2. The dataset used for GPT-3, called the Common Crawl dataset, has over 1 trillion words. Additionally, when GPT-3 first came out, it had more than 10 times as many parameters as any other language model that existed before. According to the papers on GPT-2 and GPT-3, they tested their model with a smaller number of parameters and there appears to be a linear increase in performance with an exponentially higher number of parameters. Therefore, these transformer models are going to get ever larger and the higher number of parameters may perhaps allow the models to perform new or unexpected tasks.

Perhaps due to the extremely large size of the model, GPT-3 can perform what the authors of the GPT-3 calls “few-shot learning”. Few-shot learning is the idea that the model can learn a concept (in-context learning) by just feeding it a few examples. Traditionally, fine-tuning a model for specific tasks require many thousands of examples of that task and updating the weights. The fine-tuning step requires supervised learning of large dataset of labelled examples which limits the real-world applicability of these models generally. The idea behind few-shot learning is that humans can often perform a task quite well by just giving it one example or demonstration, which is far different from these models historically. With GPT-3, significant improvement is shown when shown one example (one-shot) versus when no example (zero-shot) is shown. It is shown that smaller models have lower in-context learning (small difference in performance with zero-shot versus few-shot) versus the largest 175 billion parameter model (significant different in performance with zero-shot versus few-shot). This could mean that future transformer model could be much better at in-context learning than GPT-3. This could result in a language model able to learn new tasks and be able to skip the fine-tuning step typically needed to improve performance of specific tasks, which is a major goal of the GPT models.

There are some concerns that come up with models such as GPT-3 however. Because, the model is trained on a large dataset of the internet, the model is biased in ways that reflect the current state of the internet. Use of models such as GPT-3 in real life may reinforce harmful stereotypes if the issue is not somehow addressed. Texts produced by GPT-3 show bias by gender, race, and religion. When writing stories, GPT-3 includes more male characters compared to female characters. A male identifier was more likely than a female identifier for 83% of

occupations by GPT-3. As mentioned in the introduction, fraudulent essays can be produced by GPT-3, which will make detecting cheating difficult when writing essays. This may prove to be a very large problem in the future as these models get more sophisticated and the school system is likely slow to react across the world.

GPT-3 shows state of the art performance in some respects as the largest transformer model trained by far. However, there are many improvements to be made. There is a huge amount of promise in future transformer models in the performance of NLP tasks and the generalization they are able to provide without needing to fine-tune. There are also potential issues in terms of producing fraudulent text and bias shown by these models however. These issues are going to be difficult to tackle and it will be important to focus on them as more models are produced and trained.

Citations:

Vaswani, et al, Attention Is All You Need. (2017)
https://arxiv.org/abs/1706.03762

Radford, Alect et. al. “Language Models are Unsupervised Multitask Learners.” (2019).
https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Brown et. al, Language Models are Few-Shot Learners, NeurIPS 2020
https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html

Dou et. al, Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text, ACL 2022
https://aclanthology.org/2022.acl-long.501/

Lucy et. al, Gender and Representation Bias in GPT-3 Generated Stories, ACL 2021
https://aclanthology.org/2021.nuse-1.5/