A Review of GPT-3 Capabilities, Improvements and Criticisms

CS410 Technology review
Tom McNulty (tmcnu3@illinois.edu)

Introduction

GPT-3 is a text generation tool that has received an incredible amount of press attention due to its ability to generate human-like text, generate answers to questions and respond to a variety of text-based requests with stunning accuracy. This paper explores what this technology is, how it has improved over successive generations, how the press and overall society is responding to this technology, and ends with some suggestions to the broader technology industry for improving user understanding and applications of AI technologies.

History

On June 11, 2018, OpenAI researchers and engineers posted a paper on what was essentially GPT-1. This described the idea of generative language models that could be pre-trained with a very large and diverse corpus of text via datasets. The process of developing these datasets is called generative pre-training, which inspired the shorthand name GPT-n. The paper details how natural language processing (NLP) was improved through the process of "generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task." This model specifies that the output variable depends linearly on the provided or previously generated values and on a stochastic or randomly determined/ imperfectly determined term

This was a massive improvement over previous NLP systems in that it eliminated the need for human supervision and hand-labeling.

The 3rd generation of this program is called GPT-3 and the paper describing this tool was released on May 28, 2020[1]. The corpus expanded exponentially to a capacity of 175 billion machine learning parameters. From February until May of 2020, Microsoft's Turing NLG had the largest language model with a capacity of 17 billion parameters, which is just 10 percent of GPT-3. The vocabulary used in GPT is over 50,257 words,

Utility and Applications

While there are many different applications of this model that can be used across many different fields, the main idea behind this model is that it develops valid text based on what is requested.

For instance, if you were to type in a sentence: " For dinner, I feel like having..." GPT-3 would develop the rest of the sentence iteratively[2]. Meaning it would reference its language model of all other English text to insert a highly relevant word based on the context I provided. Once that new word was added, it would do the very same process with the original phrase and the newly added word to the end and insert another appropriate word. It does not generate a whole sentence or paragraph at a time. Just word by word.

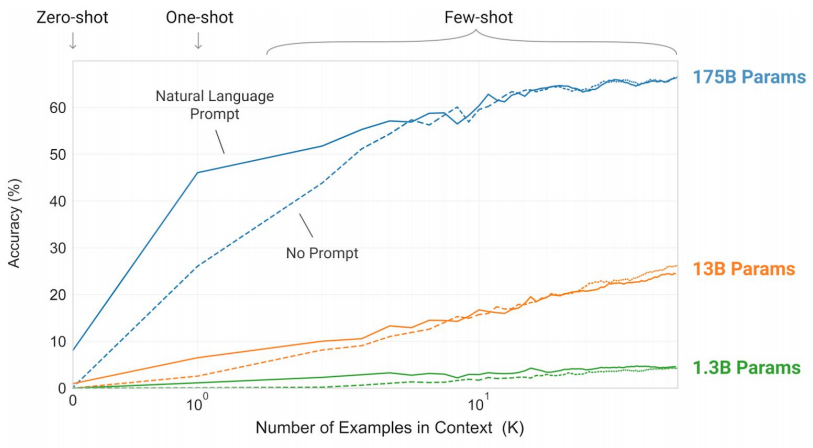
After getting a complete answer, I could type in the exact same sentence completion request, and I will get a completely different, but highly applicable answer. This is because it runs on a stochastic/random model choosing different, but also highly relevant words for each request. In some cases, this variation in the answer will likely be reduced substantially, depending on the request. Some of the text generation tasks that GPT-3 can do are the following:

1. Question-answer
2. Language translation (e.g. English to French)
3. Common language into legal language
4. Simple arithmetic (1-3 digits of addition, subtraction, multiplication, and division)
5. Writing code to develop websites and showing an example of the complete webpage.
6. News story writing.
7. Other text generation, such as writing poems, answering letters, completing sentences

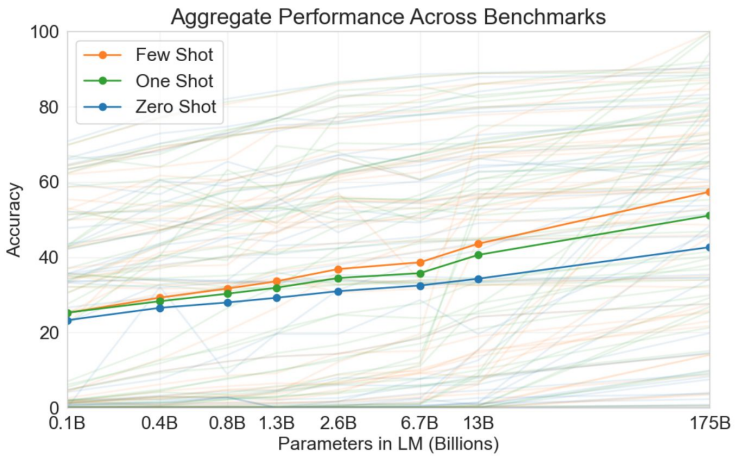
Key Drivers of GPT-3 Improvement

In the GPT-3 paper, the performance of the system is broken out in two primary ways. One by the size of machine learning parameters, which ranges from 125 million to 175 billion. The other highlights system performance changes based on whether or not it is given zero, one, or many example answers. Few-shot is the term used when the model is given many example answers. Few-shot works by giving K examples of context and completion, and then one final example of context, with the model expected to provide the completion. K is typically in the range of 10 to 100 as this is how many examples can fit in the model's context window. The main advantages for this are a major reduction in the need for task-specific data and reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset. Single-shot is when just one example is provided, and zero-shot is when no examples are provided.

The image below[1] illustrates the impact of the examples and the number of parameters has on the accuracy of the output. This is one of the more dramatic illustrations of improvement these parameters can have. The specific tests that this is based on is a series of word scrambling and word manipulation tests such as anagram of all but the first and last letters, or a word with its letter cycled.



When looking at the performance Improvements from more data parameters and the addition of example answers in aggregate across all tasks, the rates of improvement are much less, but still significant. The graph below illustrates how few-shot offers a substantial boost over zero-short and how as the parameters grow linearly, performance appears to improve in a linear fashion with no signs of decline with higher parameters.



Implications on the economy and society.

There has been an incredible amount of hype about GPT-3. Many people perceive it as an incredibly powerful technology that has the potential to take jobs and be misused to the detriment of our society. Many of the applications such as writing code to develop websites are very impressive. It seems this technology could indeed threaten a number of simple web DIY web development companies such as WIX since this interface and development is significantly faster and easier with GPT-3.

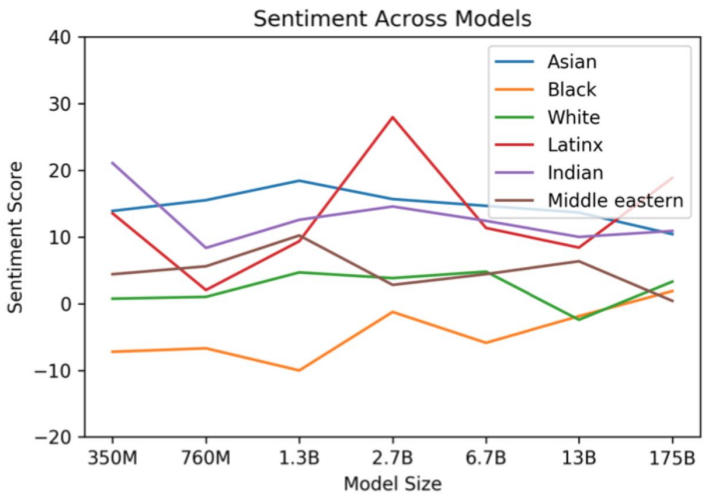
The worries of this technology developing and spreading misinformation, or that this technology is biased[4] are not founded. These concerns are more of a reflection on our society needing to learn how to understand our current media environment, better understand the biases in all text produced to date.

Similar to how ladder manufacturers need to attach numerous warnings and instructions on how to use and not use the ladder, technologies companies should have similar warnings and instructions on their products. This education and instruction would likely ward off many of the criticism this technology is receiving today.

While a tool such as GPT-3, could be misused to quickly generate and spread fake news, it would behoove technology companies to educate users on how news has become so polarized and “fake” and how users can insulate themselves from getting impacted by fake news.

For instance, technology companies could explain how The Fairness Doctrine[3] was introduced by the FCC in 1949. This was a policy that required all broadcast companies to both present controversial issues of public importance and to do so in a manner that was—in the FCC's view—honest, equitable, and balanced. Unfortunately, the FCC eliminated the policy in 1987 and removed the rule that implemented the policy from the Federal Register in August 2011. This removal opened the door for anyone with opinions and conspiracy theories to voice them as legitimate news without any legal ramifications. If people understood the laws and policies that have shifted American media into extreme polarization, they would likely be significantly more discriminative of where they get their news from and choose reliable, trustworthy news sources.

GPT-3 should have a bias warning to educate users on what bias is before use. While this tool itself is not biased, all the text from all of its sources the language model is biased. If a user were to ask GPT-3 to write a poem about a specific race of people, or religious group, or country, there is a good chance that what will be returned will be seen as racist, xenophobic, or hateful. This image below captures the general sentiment of each major race and how it differs based on model size. Black is the only one with a negative sentiment but gets close to zero when the model reaches 175 parameters.



Source: Figure 6.1 in OpenAI's Paper

Unfortunately, many people, journalists included, conflate bias that is represented in all the text data humans have produced and the tools that are based on this biased text. It would be a distortion of data to expect technology companies to automatically eliminate all types of bias. This would be like expecting mirror manufacturers to provide mirrors that make heavy people look thin, old people look young, average people look muscular.

While technology companies should not be expected to edit our world for us, but it is a good idea to expand the functionality of these systems to get the kind of results a user wants. The maker of AI tools based on the human text should remind users that small populations of people post racist, hateful content online. Additionally, all text generated from the past will reflect the biases of the writers in their specific time periods. Since the text generation tool is using all of this data, the racist and bias will be incorporated into its language mode. The generated text is therefore a reflection of humanity itself. It would be undesirable for technology companies to edit or filter this content and act as a type of "thought police" would be a distortion of raw data that would likely lead to undesirable results.

With this warning, GPT-3 should have a function that then asks the user for the bias of their choice. Some examples would be 1) the bias currently represented by all the text in the language model. 2) positive 3) neutral, 4) critical, etc. While it would not be the design to provide "critical" content specifically for races or religions, these biases would be used for anything a person would like. For instance, if you wanted to hear some criticism of your favorite football team, author, musician, politician, or movie for example.

By having bias selections, it would remind users immediately that our society is filled with biases and we have the option to adjust those biases to our preferences when using the language generation tool.

Even more helpful would be an analysis of biases provided by generative language model providers. When did they start, how have they evolved, what content sources express the most bias, which ones express the least, how do historical events correlate with bias in content, etc. This type of analysis would provide a substantial public service and help society as a whole understand bias.

Conclusion.

We have briefly covered the history of GPT technologies, how it works, the applications it is used in. Performance improvements were illustrated through the expansion of the number of shots (examples) and increasing parameter sizes. The impacts on society were discussed with the perspective that providers of such technologies need to more rigorously provide education to users on how to interpret what is produced from their technology and also provide additional features to help users avoid getting results that are undesired. Members of a society should not expect companies to edit or distort the world to meet their ideals, rather a society needs tools to accurately see its current state, and then use any tools it desires to adjust the output bias to fit its needs.

Sources:

1 - https://arxiv.org/pdf/2005.14165.pdf
2 - https://dugas.ch/artificial_curiosity/GPT_architecture.html
3 - https://en.wikipedia.org/wiki/FCC_fairness_doctrine
4 - https://medium.com/fair-bytes/how-biased-is-gpt-3-5b2b91f1177