MA, Digitisation, Surveillance and Societies , Erasmus University
Kartikeya Saigal
kartikeyasaigal01@gmail.com

UNBOXING THE ALGORITHM

Case Study | GPT-3

SUMMARY: A description of the algorithm GPT-3, including technical aspects; what machine learning techniques does it use; the actual and potential impacts and uses of the algorithm; a critical reflection about the social context and relevance; the impact of individual behaviour; and a look at the questions of transparency, accuracy and fairness.

GPT-3 is being developed by OpenAI which is an artificial intelligence research and deployment company founded in 2015 with a pledge of 1Bn by industry heavyweights to ensure that AGI benefits all of humanity (Dale, 2021). In proper scientific, or perhaps human fashion, their goal is also to be the first create the problem they've sworn to mitigate. With funding from investors like Elon Musk and eventually Microsoft their resources have allowed for the creation of versions of a Generative Pre-Trained Transformer (GPT) that has been trained on copious amounts of data from the internet and can engage with media and command prompts to generate shockingly coherent text that often sounds quite human.

GPT-3 is a scaled up version of GPT-2 which was announced in 2019 as a large unsupervised transformer with 1.5B parameters trained on 40GB of text, or roughly 10B tokens (Dale, 2021). In June 2020, OpenAI came forward with GPT-3 which is a language model 100 times larger than its predecessor with 175B parameters and 96 layers trained on 499B tokens of web content (Dale, 2021). This is the largest algorithmic language model deployed to date.

Trained on Microsoft's Azure supercomputer at an estimated cost of $12m the algorithm is finding a wide variety of uses (Scott, 2020, Wiggers, 2020 as cited in Floridi & Chiriatti, 2020). From summarization, translation, grammar correction, intuitive coding processes for non-coders, chatbots and question answering to autocompleting text to 'n' number of tokens, the software is flexible and its capabilities expandable with greater training and usage.

As stated by Elkins and Chun, "despite its mathematical, semantic and ethical shortcomings — or better, despite not being designed to deal with mathematical, semantic, and ethical questions — GPT-3 writes better than many people" (Elkins and Chun, 2021). However the technology also has many limitations still, as have been widely noted. Outputs generated may lack coherence the longer they get, exhibit biases that result from training data and also not be consonant with the truth. (Dale, 2021).

The Social Context and Impact of the Algorithm

As a product of OpenAl's efforts to create AGI, GPT-3 is a powerful language model that can not only mimic human textual output but also analyse its training data to do a number of other tasks with the information. It is a new frontier for textual content, especially online. GPT-3 makes it possible to mass produce credible and cheap semantic artifacts. "Translations, summaries, minutes, comments, webpages, catalogues, newspaper articles, guides, manuals, forms to fill, reports, recipes … soon an AI service may write, or at least draft, the necessary texts, which today still require human effort." (Floridi & Chiriatti, 2020).

Once this manner of content production becomes mainstream and acceptable, the profession of writing itself will be supported by programs like GPT-3. And people will have to rethink the skills required for work. Instead of mere cut & paste, they will need to be good at prompt & collate. New editorial skills will be required that can prompt software engines to deliver the best results and collate them intelligently (Floridi & Chiriatti, 2020).

Any industry that currently relies primarily on human generated content is squarely in the crosshairs of a cultural shift. Books will be co-authored by AI, student assignments co-ideated. Newspapers and television will have large portions of content delivered entirely by automated systems that can write better than underpaid interns. "Readers and consumers of texts will have to get used to not knowing whether the source is artificial or human." (Floridi & Chiriatti, 2020). Just as in the case with any major technological upheaval, such as the typewriter, the idea of work will change but human intelligence isn't close to being replaced in the workforce. These systems will still require human oversight and engagement to fulfil the goals we set them for content production. What will change is our understanding of what it means to write and be human, to create semantic artifacts from real world experience as opposed to uncanny simulacrums via algorithm.

At this point it is obvious then that GPT-3 is weaponizable as a tool for misinformation. Its ability to generate text from a given prompt makes it easy to create fake news stories or other false information. Additionally, GPT-3's natural language processing capabilities could be used to create bots that generate convincing arguments for or against a given topic. This could be used to create "astroturfing" campaigns, where fake grassroots support is generated for or against a particular issue. In fact I didn't say that, GPT-3 did as an introduction when I gave it the prompt: Write a long essay on OpenAI and GPT-3 and its potential for misinformation.

There are some ways in which OpenAI could potentially mitigate the software's misuse. OpenAI could develop a way to flag generated text that is likely to be false or misleading. This could be done by training a classifier on a dataset of known fake news stories or other false information. Additionally, OpenAI could develop a way to verify the source of generated text. This could be done by requiring users to provide a URL or other source when submitting a prompt to GPT-3. Finally, OpenAI could work with fact-checking organizations to develop a system for flagging generated text that has been verified as false. Once more, that was GPT-3.

And just to reiterate how coherent these outputs are and how good they are at mimicking human text, especially by cross-referencing different perspectives in the third person, this is how GPT-3 concluded. Despite these potential mitigations, it is important to be aware of the potential for GPT-3 to be used for misinformation. As the technology develops, it will become increasingly easy to generate fake news and other false information. It is important to be critical of the sources of information, and to verify claims before sharing them.

Further, given its training data, GPT-3 can also mimic particular authors. It can therefore write a paragraph on tulips in the voice of author, actor and comedian Stephen Fry.

Prompt: Write a paragraph on tulips in the voice of Stephen Fry

GPT-3 Output: **If you were to ask me what my favourite flower is, I would have to say tulips. I love their elegant shape and the way they seem to dance in the breeze. They are the perfect flower for any occasion, whether you're looking to say "I love you" or "get well soon." There is something about tulips that just makes me happy. Maybe it's their vibrant colours or the way they make me think of springtime. Whatever the reason, I can't get enough of them. I always make sure to have at least a few tulips in my home, and I always get a bouquet for my wife on our anniversary. If you've never had the pleasure of enjoying tulips, I urge you to go out and get some. Trust me, you won't be disappointed.**

Or

Prompt: ***Write a speech in the voice of Barack Obama admitting he's not American***

GPT-3 Output: **It has come to my attention that there are some who question my American citizenship. I want to set the record straight: I am not an American citizen. I was born in Kenya, to a Kenyan father and an American mother. I spent my early childhood in Indonesia, and later moved to Hawaii to attend school. I have always considered myself a citizen of the world, and my allegiance is to the human race, not to any one nation. I understand that some people feel strongly about nationality and patriotism, and I respect that. But for me, being American is not about where you were born, or what passport you hold. It's about shared values and a commitment to making our country a better place for everyone. So whatever doubts anyone may have about my American citizenship, I want to assure them that I am just as committed to the United States of America as anyone else. I may not have been born here, but this is my home, and I will fight for it just as hard as anyone else.**

In both cases, those familiar with the speech and text patterns of the exampled authors must admit there is a distinct capturing of essence and tonality that can allow the content to pass for credible. GPT-3's imitation of Stephen Fry's writing could be mistaken for his actual work if one didn't know it was generated by an algorithm. Perhaps Obama's admission that he's not an American would be more suspect because of its controversial nature but we know how little it takes for content to spread online and influence individuals in real time. This is then especially worrying since fact checking content is most certainly not quicker than the generation of content.

In mildly abstract terms, Benzon discusses and tries to understand why GPT-3 works so well to create semantic artifacts. He says:

> "language exists as strings of signifiers in the external world... strings of written characters that have been encoded into computer-readable form. Let us assume that the signifieds – which bear a major portion of meaning, no? – exist in some high dimensional network in mental space. This is, of course, an abstract space rather than the physical space of neurons, which is necessarily three dimensional. However many dimensions this mental space has, each signified exists at some point in that space and, as such, we can specify that point by a vector containing its value along each dimension."

And that in this context

> "what GPT-3, and other NLP engines, do is to examine the distances and ordering of signifiers in the string and compute over them so as to reverse engineer the distances and orientations of the associated signifieds in high- dimensional mental space."

He points out how GPT-3 and similar algorithms are attempting to be models of the mind yet at the same time have clear limitations in terms of understanding or generating meaning. In the article he cites computational linguistics scientist Martin Kay who said " what we are doing is to allow statistics over words that occur very close to one another in a string to stand in for the world construed widely, so as to include myths, and beliefs, and cultures, and truths and lies and so forth" (Benzon, p. 12, 2020). And therefore, I understand that it is the metaphysical structure of the world as perceived by humans that we cannot conflate with the physical existence and success of the algorithm.

Benzon suggests Neubig's "isomorphic transform onto meaning space" and Gärdenfors' conceptual spaces as the beginning of a framework to better understand these algorithmic models and improve them (Benzon, p.25, 2020). The theory of conceptual spaces presents a framework for representing information on the conceptual level.

As per Gärdenfors a conceptual space consists of geometrical structures based on a number of quality dimensions. The main applications of the theory are on the constructive side of cognitive science: as a constructive model the theory can be applied to the development of artificial systems capable of solving cognitive tasks.

Conclusion

To conclude, GPT-3 is a deep learning language model trained on large amounts of textual data. It is capable of an evolving number of tasks and can be fine-tuned with additional training. OpenAI, the developing company and also external researchers have stated that the technology is weaponizable and inherently threatens information online via misinformation. At the same time, the technology allows for deeper insight and questioning of the human brain and its relationship with language. And it can revolutionize content production across industries. Better understanding of what such technology is, is capable of and how it can and should be regulated to mitigate misuse is an issue for legislative bodies across the globe.

References

Benzon, W. L., GPT-3: Waterloo or Rubicon? Here be Dragons (August 22, 2020). Working Paper, Version 2, 2020, Available at SSRN: https://ssrn.com/abstract=3667608

Dale, R (2021). GPT-3: What's it good for? Natural Language Engineering 27, 113–118. https://doi.org/10.1017/S1351324920000601

Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's Turing Test?. Journal of Cultural Analytics, 5(2), 17212.

Floridi, L., Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds & Machines 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Gardenfors, P. (2004). Conceptual Spaces: The Geometry of Thought (A Bradford Book) (Revised ed.). MIT Press.

McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. arXiv preprint arXiv:2009.06807.