Next chapter in artificial writing

OpenAI released a beta version of its language model, GPT-3. As artificial writing permeates our lives, the challenge is how to think clearly about what it is and what impact it could have on society.

OpenAI, the artificial intelligence (AI) company, published a research paper in May 2020 on GPT-3, the latest version of its generative language model. More recently, OpenAI released a private beta version of GPT-3 to select users and made its API available by request. Responses in articles and social media have been swift and often laudatory, describing it as the "world's most impressive AI" and "terrifyingly good".

GPT-3 is a language model based on neural networks. The transformer-based model and architecture is similar to GPT-2, but the model size and dataset of GPT-3 is roughly two orders of magnitude larger. GPT-3 is trained with 175 billion parameters, using data from CommonCrawl, WebText, internet-based books corpora and English-language Wikipedia. Like GPT-2, GPT-3 can predict, or suggest, the next word or paragraph given a prompt of just a few words. This type of one-or zero-shot learning goes beyond previous natural language processing models which need many labelled examples to perform a new type of task.

Much of the buzz about GPT-3 has focused on its ability to generate text using the 'text in, text out' interface of the API. Users can enter a word or phrase, and text emerges. GPT-3 is so good at this that it can generate synthetic news articles that seem to be written by humans. It is easy to imagine how the technology could have a positive impact, for example, by creating sophisticated bots to assist people, providing text to compose e-mails, overcoming writer's block, facilitating learning between teachers and students, helping people with language disorders communicate in writing, and even writing code. It is important to note that the OpenAI paper reports the performance of GPT-3 on language tasks other than text generation, including its ability to answer general knowledge questions, to translate between languages, to perform common sense reasoning and reading comprehension tasks, and so on.

There are downsides to GPT-3, and important questions about its impact on society. The OpenAI researchers discuss these issues in their paper, such as GPT-3 being used for spam, phishing, misinformation and fraudulent academic essay writing. The authors also present preliminary analyses on the limitations of GPT-3 with respect to fairness, bias, and representation. The fundamental issue here is that GPT-3 is trained on data from the internet, with its inherent biases in race, gender, religion, and other subjects. Prominent voices such as Jerome Pesenti (VP of AI at Facebook) and Anima Anandkumar (professor at Caltech and director of Machine Learning Research at NVIDIA) took to Twitter to raise concerns about bias in GPT-3 and language models, including examples of toxic language generated by GPT-3 when prompted with words such as Jews, black and women. OpenAI is aware of the problem and has introduced a toxicity filter to check GPT-3's output. But this ignores the question of whether it is a responsible strategy in the first place to train language models by taking any data from the web simply because it is available, including from sources such as Reddit. The obvious risk is amplification of unchecked and harmful biases.

Another concern is the substantial compute time and energy impact of language models. This issue was raised last year in a paper by Emma Strubell and colleagues (see also our News Feature this month on the carbon impact of artificial intelligence) who calculated that training a 'big' transformer language model has the same carbon impact as five US cars over their lifetime, including fuel. Of course, GPT-3 is much bigger, with orders of magnitude more parameters, although calculating the carbon impact requires more details on the design process and hardware infrastructure. In their paper, the OpenAI authors only spend one and a half paragraphs on energy usage, in which they acknowledge the need to consider energy impact but also argue that once trained, GPT-3 is very efficient.

At a more philosophical or conceptual level, debates are raging about the degree to which such an AI tool can be called intelligent, or even scientific, rather than a clever engineering feat. One loose definition of AI is that it can perform tasks that people consider intelligent when done by humans, such as creative writing. On the other hand, literally ascribing intelligence to technology must be done with care. With its apparent ability to artificially read and write, GPT-3 is perhaps different from other forms of AI, in that writing seems more fluid, open-ended, and creative than examples of AI that can beat people in a game or classify an image. But what kind of 'writer' or writing tool is GPT-3? It has no consciousness, no motivation, no experience, no moral compass, no vision, no human connections and no humanity.

Despite lacking these qualities, GPT-3's text-generation abilities are remarkable and amazingly versatile. From a selection taken from Twitter, users of the recent 'text-in, text-out' interface have experimented with deploying GPT-3 for mocking-up websites, writing machine learning code in Keras and of course producing creative writing, including comedy. There might be a 'killer application' that has not even been found yet.

At the same time, there is much work to be done to reflect on and tackle the downsides of this new tool that, like any AI technology, risks amplifying existing societal injustices and can be used in harmful ways. For the next generation of language models, it seems urgent to focus on compute and energy impact, as well as the need to incorporate more diversity and quality control. ❐

Published online: 12 August 2020
https://doi.org/10.1038/s42256-020-0223-0