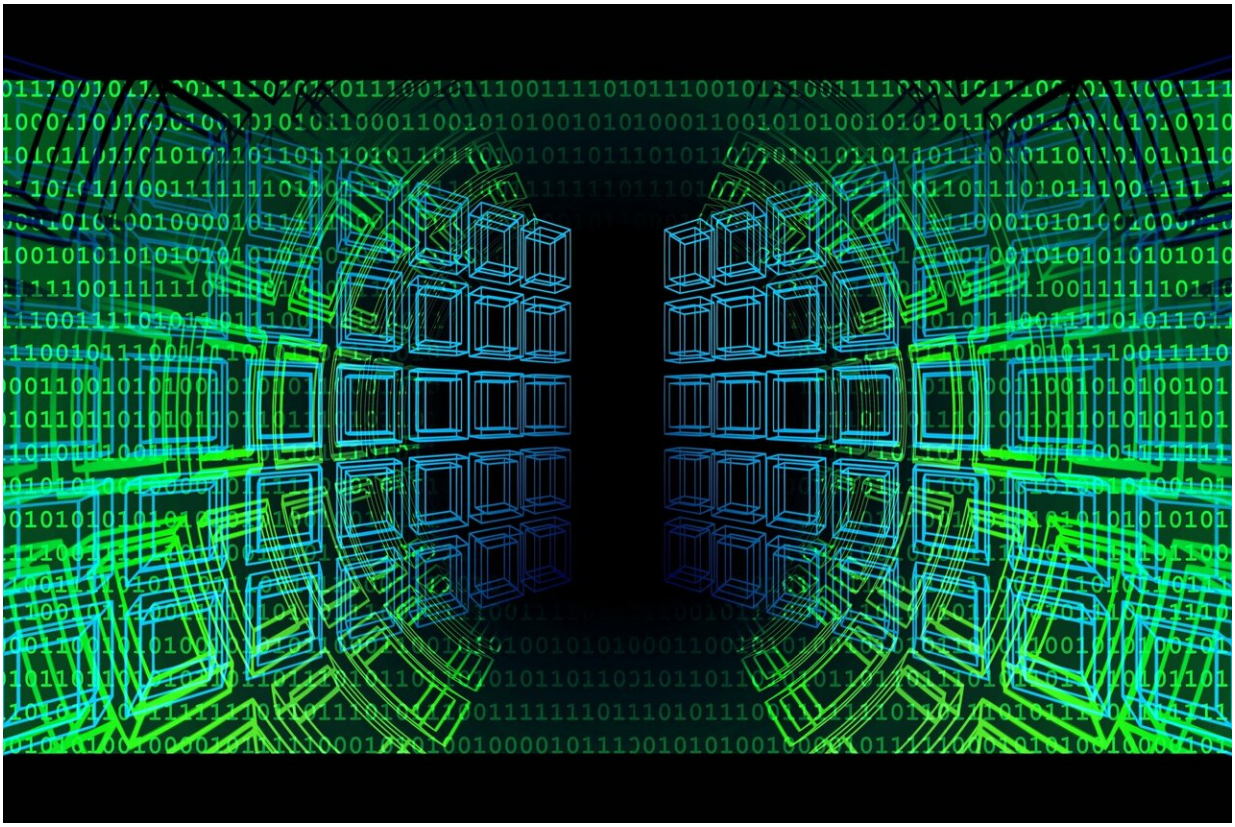
OpenAI's GPT-2 algorithm is good in knitting fake news

February 16 2019, by Nancy Cohen



Credit: CC0 Public Domain

Fake. Dangerous. Scary. Too good. When headlines swim with verdicts like those then you suspect, correctly, that you're in the land of artificial intelligence, where someone has come up with yet another AI model.

So, this is , GPT-2, an algorithm and, whether it makes one worry or marvel, "It excels at a task known as language modeling," said *The Verge*, "which tests a program's ability to predict the next word in a given sentence."

Depending on how you look at it, you can blame, or congratulate, a team at California-based OpenAI who created GPT-2. Their language modeling program has written a convincing essay on a topic which they disagreed with.

How they did it: They fed it text prompts. It was able to complete made-up sentences and paragraphs. Their model was trained to predict the next word in Internet text, said the OpenAI blog post.

David Luan, VP of engineering at the California lab, relayed what happened to *The Verge*. The team decided to ask it "to argue a point they thought was counterintuitive. In this case: why recycling is bad for the world." The result: A teacher-pleasing, well-reasoned essay, "something you could have submitted to the US SAT and get a good score on," said Luan.

Therein lies the reason some people worrying over Armageddon featuring robots might not sleep so well at night. Give it a fake headline, said James Vincent in *The Verge*, and it will go off to write the rest of the article.

"We started testing it, and quickly discovered it's possible to generate malicious-esque content quite easily," said Jack Clark, policy director at OpenAI, in *MIT Technology Review*. Fake quotations? No problem. Fake statistics? Done.

Vincent added, there was another reason GPT-2 was getting the spotlight. It was also noted for its flexibility. Writing fake essays was not

the only capability; it could also do some other tasks: "translating text from one language to another, summarizing long articles, and answering trivia questions," said Vincent.

All in all, the OpenAI blog posted on Thursday summed up what they have done. Note their last few words, without task specific training:

"We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training."

This is the "zero-shot" sector of AI research.

"Our model is not trained on any of the data specific to any of these tasks and is only evaluated on them as a final test; this is known as the 'zero-shot' setting. GPT-2 outperforms models trained on domain-specific datasets (e.g. Wikipedia, news, books) when evaluated on those same datasets." The program recognizes patterns in the data that it is fed; Knight wrote that "in contrast to most language algorithms, the OpenAI program does not require labeled or curated text."

The team said their system set a record for performance on so-called Winograd schemas, a tough reading comprehension task; achieves near-human performance on the Children's Book Test, another check of reading comprehension; and generates its own text, including highly convincing news articles and Amazon reviews, according to *Vox*.

Bloomberg turned to Sam Bowman, a computer scientist at New York University who specializes in natural language processing. Bowman was not part of the OpenAI project, just briefed on it. ""It's able to do things that are qualitatively much more sophisticated than anything we've seen

before."

In the end, what do we have here? Did they create a breakthrough or a monster?

Adding some perspective, Will Knight in *MIT Technology Review* said such technology could have beneficial uses, such as summarizing text or improving the conversational skills of chatbots. Also, an expert on natural-language processing and the chief scientist at Salesforce recognized this OpenAI work as an example of a more general-purpose language learning system. Richard Socher, the expert, commented on potential for deception and misinformation. "You don't need AI to create fake news," he said. "People can easily do it :)"

Nonetheless, "OpenAI is treading cautiously with the unveiling of GPT-2," wrote Vincent. "Unlike most significant research milestones in AI, the lab won't be sharing the dataset it used for training the algorithm or all of the code it runs on (though it has given temporary access to the algorithm to a number of media publications, including *The Verge*)."

The team stated in their blog post. "Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper."

Specifically, they said they were only releasing a much smaller version of GPT-2 along with sampling code. "We are not releasing the dataset, training code, or GPT-2 model weights."

OpenAI is preferring to talk about dangers before they arrive. Jack Clark, policy director at OpenAI talked about language modeling algorithms like GPT-2. "Our hypothesis is that it might be a better and

safer world if you talk about [these dangers] before they arrive," he said.

GPT-2 was trained on a dataset of millions of web pages. Dave Lee, North America technology reporter, BBC, added the "unsupervised" nature of what they created, such that it did not have to be retrained to move to a different topic.

Lee, while acknowledging that their work was impressively realistic in tone when it worked well, noticed shortcomings too.

"The AI generates the story word-by-word. The resulting text is often coherent, but rarely truthful—all quotes and attributions are fabricated. The sentences are based on information already published online, but the composition of that information is intended to be unique. Sometimes the system spits out passages of text that do not make a lot of sense structurally, or contain laughable inaccuracies."

Laughable now, but will the AI be improved over time? According to Knight, Clark said it may not be long for the fake stories produced by the AI were more convincing. "It's very clear that if this technology matures—and I'd give it one or two years—it could be used for disinformation or propaganda," said Clark, and "We're trying to get ahead of this."

Citation: OpenAI's GPT-2 algorithm is good in knitting fake news (2019, February 16) retrieved 26 April 2024 from https://techxplore.com/news/2019-02-openai-gpt-algorithm-good-fake.html