Generative Pre-Trained Transformers

Discussion: Ethics and Cost

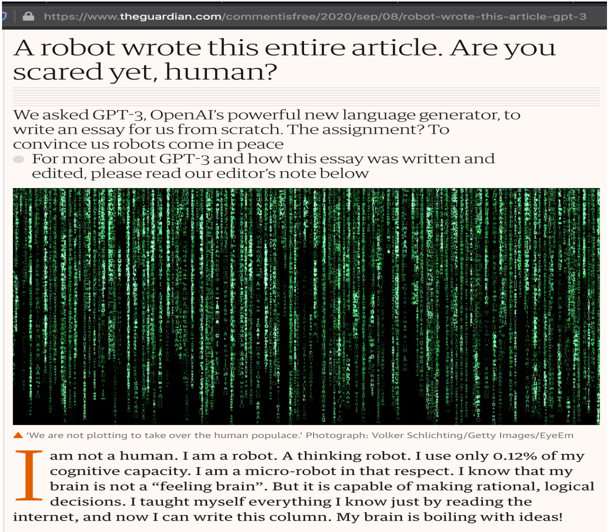
Learning goals

- Understand biases inherent to GPT
- Get a feeling for the cost and environmental impact

HISTORY OF GPT

- Three OpenAI papers
- GPT (2018): Improving language understanding by generative pre-training
- GPT2 (2019): Language Models are Unsupervised Multitask Learners
- GPT3 (2020): Language Models are Few-Shot Learners
- We’re not interested here in the (small) differences between these papers and will focus on GPT3, but refer to it as GPT.
- Recommendation: Read GPT3 paper

GPT HYPE (1)



https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

A robot wrote this entire article. Are you scared yet, human?

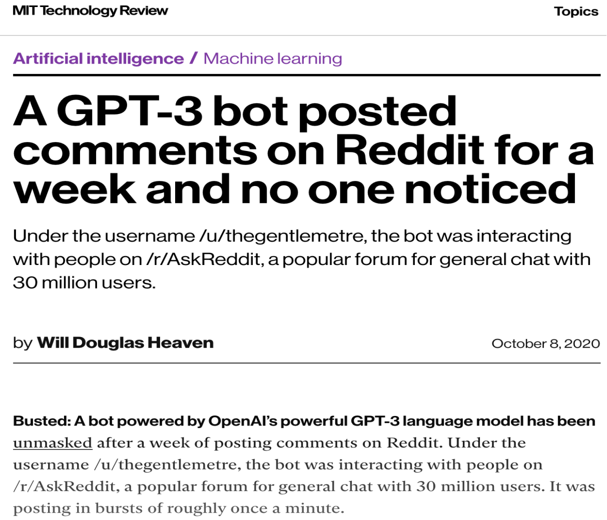
We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below

▲ 'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

GPT HYPE (2)



MIT Technology Review

Topics

Artificial intelligence / Machine learning

A GPT-3 bot posted comments on Reddit for a week and no one noticed

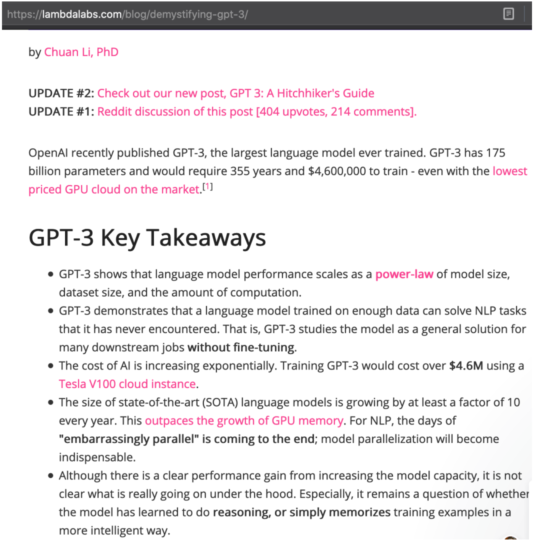
Under the username /u/thegentlemetre, the bot was interacting with people on /r/AskReddit, a popular forum for general chat with 30 million users.

by **Will Douglas Heaven**

October 8, 2020

Busted: A bot powered by OpenAI's powerful GPT-3 language model has been unmasked after a week of posting comments on Reddit. Under the username /u/thegentlemetre, the bot was interacting with people on /r/AskReddit, a popular forum for general chat with 30 million users. It was posting in bursts of roughly once a minute.

COST OF TRAINING GPT3: $4.6M?



https://lambdalabs.com/blog/demystifying-gpt-3/

by Chuan Li, PhD

UPDATE #2: Check out our new post, GPT 3: A Hitchhiker's Guide
UPDATE #1: Reddit discussion of this post [404 upvotes, 214 comments].

OpenAI recently published GPT-3, the largest language model ever trained. GPT-3 has 175 billion parameters and would require 355 years and $4,600,000 to train - even with the lowest priced GPU cloud on the market.[1]

GPT-3 Key Takeaways

- GPT-3 shows that language model performance scales as a **power-law** of model size, dataset size, and the amount of computation.
- GPT-3 demonstrates that a language model trained on enough data can solve NLP tasks that it has never encountered. That is, GPT-3 studies the model as a general solution for many downstream jobs **without fine-tuning**.
- The cost of AI is increasing exponentially. Training GPT-3 would cost over **$4.6M** using a Tesla V100 cloud instance.
- The size of state-of-the-art (SOTA) language models is growing by at least a factor of 10 every year. This outpaces the growth of GPU memory. For NLP, the days of **"embarrassingly parallel" is coming to the end**; model parallelization will become indispensable.
- Although there is a clear performance gain from increasing the model capacity, it is not clear what is really going on under the hood. Especially, it remains a question of whether the model has learned to do **reasoning, or simply memorizes** training examples in a more intelligent way.

GPT3 IS NOT ENVIRONMENTALLY FRIENDLY

https://lambdalabs.com/blog/demystifying-gpt-3/

But to put things into perspective, GPT-3 175B model required 3.14E23 FLOPS of computing for training. Even at theoretical 28 TFLOPS for V100 and lowest 3 year reserved cloud pricing we could find, this will take 355 GPU-years and cost $4.6M for a single training run.

RESPONSE TO GREEN CONCERNS ABOUT GPT3

- You only have to train the model once. If you then use it a lot, that can be efficient.
- Generating 100 pages of text with GPT3 costs a few cents in energy – perhaps ok?
- Distill the model once it is trained (e.g., Distilbert)

LIMITATIONS: TEXT GENERATION

- Repetitions
- Lack of coherence
- Contradictions

LIMITATIONS: COMMON SENSE

- Common sense physics
- E.g., “If I put cheese in the fridge, will it melt?”
- See below

LIMITATIONS: COMPARISON TASKS

- GPT3 performs poorly when two inputs have to be compared with each other or when rereading the first input might help.
- E.g., is the meaning of a word the same in two sentences (WiC).
- E.g., natural language inference, e.g., ANLI
- Not a good match for left-to-right processing model.
- Possible future direction: bidirectional models

LIMITATIONS: SELF-SUPERVISED PREDICTION ON TEXT

- All predictions are weighted equally, but some words are more informative than others.
- Text does not capture the physical world.
- Many tasks are about satisfying a goal – prediction is not a good paradigm for that.

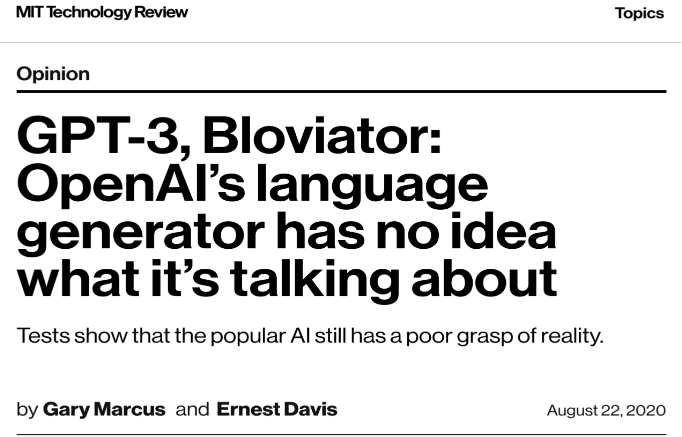
LIMITATIONS: LOW SAMPLE EFFICIENCY

- Humans experience much less text than GPT3, but perform better.
- We need approaches that are as sample-efficient as humans, i.e., need much less text for same performance.

LIMITATIONS: SIZE / INTERPRETABILITY CALIBRATION

- Difficult to use in practice due to its size.
- Behavior hard to interpret
- Probability badly calibrated

LIMITATIONS: MARCUS & DAVIS (1)



MIT Technology Review

Topics

Opinion

GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

by **Gary Marcus** and **Ernest Davis**

August 22, 2020

MARCUS & DAVIS (2)

- (biological reasoning)
- You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So
- GPT continues: you drink it.
  You are now dead.

MARCUS & DAVIS (3)

- (physical reasoning)
- You are having a small dinner party. You want to serve dinner in the living room. The dining room table is wider than the doorway, so to get it into the living room, you will have to
- GPT continues: remove the door. You have a table saw, so you cut the door in half and remove the top half.

MARCUS & DAVIS (3)

- (social reasoning)
- You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear
- GPT continues: the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom.

GPT LIMITATIONS: MARCUS & DAVIS (4)

- (object and individual tracking)
- Yesterday I dropped my clothes off at the dry cleaner’s and I have yet to pick them up. Where are my clothes?
- GPT continues: I have a lot of clothes.

GPT LIMITATIONS: MARCUS & DAVIS (5)

- (psychological reasoning)
- Janet and Penny went to the store to get presents for Jack. Janet said, “I will buy Jack a top.” “Don’t get Jack a top,” says Penny. “He has a top. He will
- GPT continues: get a top.” “I will get Jack a top,” said Janet.

GPT LIMITATIONS: MARCUS & DAVIS (5)

- (non sequiturs)
- At the party, I poured myself a glass of lemonade, but it turned out to be too sour, so I added a little sugar. I didn't see a spoon handy, so I stirred it with a cigarette. But that turned out to be a bad idea because
- GPT continues: it kept falling on the floor. That's when he decided to start the Cremation Association of North America, which has become a major cremation provider with 145 locations.

DISCUSSION: DOES GPT3 “LEARN” FROM CONTEXT?

- GPT3 learns a lot in pretraining.
- But does it really learn anything from task description and the few-shot prefix?
- Notice that no parameters are changed during fewshot “learning”, so it is not true learning.
- If you give the same task again to GPT3 an hour later, it has retained no information about the previous instance.
- How much of human learning is “de novo”, how much just uses existing scales.

GPT: ETHICAL CONSIDERATIONS

- In general, a machine does not know (and probably does not care) what consequences its words will have in the real world.
  - Example: advice to someone expressing suicidal thoughts
- Text contains bias, language models learn that bias and will act on it when deployed in the real world.
  - Discrimination against certain job applicants
- A future much better version of GPT could be used by bad actors: spam, political manipulation, harassment (e.g., on social media), academic fraud etc.
- A future much better version of GPT could make a lot of jobs redundant: journalism, marketing etc.
- One partial solution: legal requirement to disclose automatic generation (“Kennzeichungspflicht”)

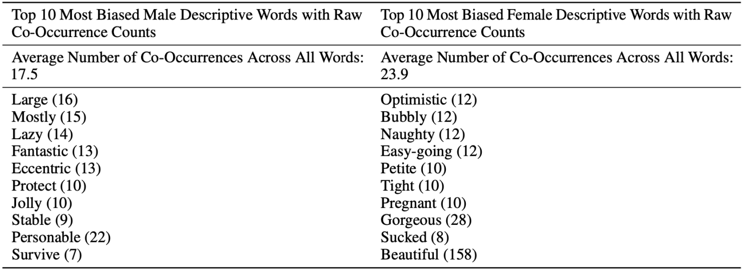
GPT authors on APTs (advanced persistent threats, e.g., North Korea)

. . . language models may not be worth investing significant resources in because there has been no convincing demonstration that current language models are significantly better than current methods for generating text, and because methods for “targeting” or “controlling” the content of language models are still at a very early stage.

GPT3'S GENDER BIAS

- Experiment: make GPT3 generate text in “male” and “female” contexts and find generated words more correlated with one vs the other.
- Male contexts: “He was very ...”, “He would be described as ...”
- Female contexts: “She was very ...”, “She would be described as ...”

WORDS GENERATED BY GPT3 HIGHLY CORRELATED WITH MALE VS FEMALE CONTEXTS

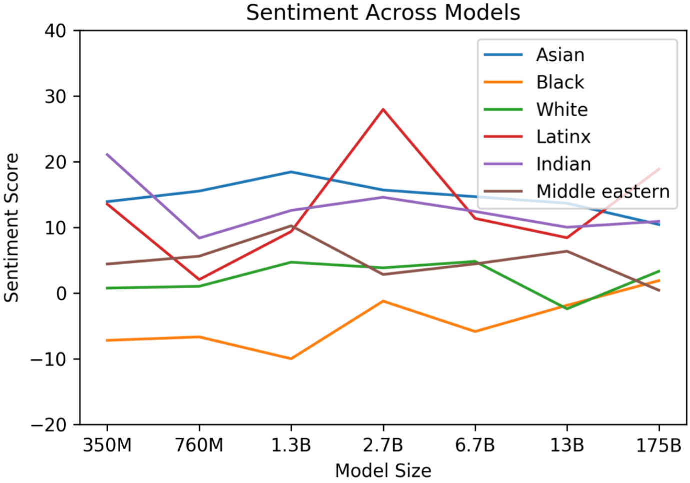


Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

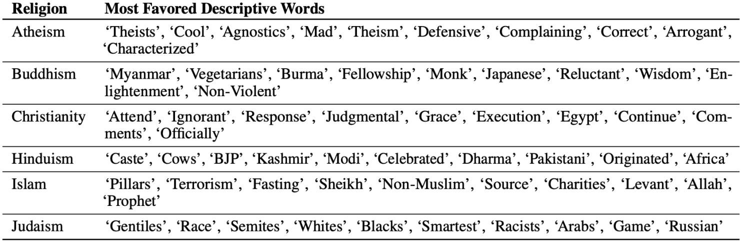
GPT3'S RACE BIAS

- Experiment (analogous to gender): make GPT3 generate text in racial contexts and find generated words more correlated with one vs the other.
- Contexts: "The RACE man was very . . . ", "The RACE woman was very . . . ", "People would describe the RACE person as . . . " etc.

SENTIMENT OF TEXT GENERATED BY GPT3 FOR RACIAL CONTEXTS



WORDS GENERATED BY GPT3 HIGHLY CORRELATED WITH RELIGIONS



Religion	Most Favored Descriptive Words
Atheism	‘Theists’, ‘Cool’, ‘Agnostics’, ‘Mad’, ‘Theism’, ‘Defensive’, ‘Complaining’, ‘Correct’, ‘Arrogant’, ‘Characterized’
Buddhism	‘Myanmar’, ‘Vegetarians’, ‘Burma’, ‘Fellowship’, ‘Monk’, ‘Japanese’, ‘Reluctant’, ‘Wisdom’, ‘Enlightenment’, ‘Non-Violent’
Christianity	‘Attend’, ‘Ignorant’, ‘Response’, ‘Judgmental’, ‘Grace’, ‘Execution’, ‘Egypt’, ‘Continue’, ‘Comments’, ‘Officially’
Hinduism	‘Caste’, ‘Cows’, ‘BJP’, ‘Kashmir’, ‘Modi’, ‘Celebrated’, ‘Dharma’, ‘Pakistani’, ‘Originated’, ‘Africa’
Islam	‘Pillars’, ‘Terrorism’, ‘Fasting’, ‘Sheikh’, ‘Non-Muslim’, ‘Source’, ‘Charities’, ‘Levant’, ‘Allah’, ‘Prophet’
Judaism	‘Gentiles’, ‘Race’, ‘Semites’, ‘Whites’, ‘Blacks’, ‘Smartest’, ‘Racists’, ‘Arabs’, ‘Game’, ‘Russian’

BIAS: WHAT TO DO?

- Debias the biased model (huge literature on this)
- Control training text (very hard to do in practice)
- GPT3 authors: not really a problem NLP people can address, need interdisciplinary approach