Bias in Data and Algorithms

Jean Mark Gawron

May 5, 2022

Outline

Background

Ethical issues

Bias

Questions

System performance: outstrips human performance

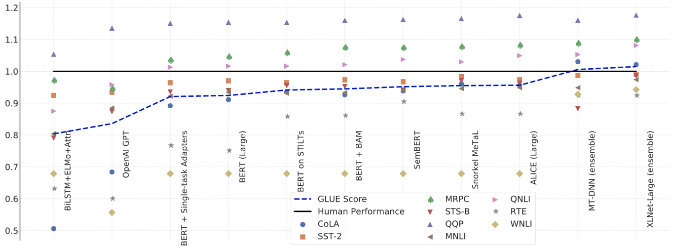


Figure 1: GLUE benchmark performance for submitted systems, rescaled to set human performance to 1.0, shown as a single number score, and broken down into the nine constituent task performances. For tasks with multiple metrics, we use an average of the metrics. More information on the tasks included in GLUE can be found in Wang et al. (2019a) and in Warstadt et al. (2019, CoLA), Socher et al. (2013, SST-2), Dolan and Brockett (2005, MRPC), Cer et al. (2017, STS-B), and Williams et al. (2018, MNLI), and Rajpurkar et al. (2016, the original data source for QNLI).

From Wang et al. (2019)

GPT

NYT article Johnson (2022)

1. GPT: Generative Pre-Trained Transformer 3. Language production system far beyond Siri and Alexa
2. Created by OpenAI: Elon Musk (CEO Tesla, Founder + Chief Engineer SpaceX, PayPal, The Boring Company, Twitter?), Greg Brockman (CTO Stripe), Sam Altman (Y Combinator)
3. Deep Learning System: 175 billion parameters; 2nd place then 17 Billion; now 1 trillion
4. Another OpenAI system DALL·E 2 can create realistic images and art from a description in natural language; another solved two problems from the Math Olympiad.

GPT3 tasks

Given a few samples establishing a task and a language prompt, GPT can produce

- fiction, poetry, press releases
- functioning computer code, translate code from one computer language to another
- music, jokes, technical manuals,
- news articles
- in particular, one article about GPT-3 written entirely by GPT-3 (GPT-3 2020)

all of which human evaluators have difficulty distinguishing from articles written by humans. See Sabeti (2020) for a host of great examples. Note that many of these are quite **funny**!

Johnson's prompt

> *You are sitting in a comfortable chair by the fire, on a cold winter's night. Perhaps you have a mug of tea in hand, perhaps something stronger. You open a* **magazine** *to an article you've been meaning to read. The title suggested a story about a promising —- but also potentially dangerous -— new technology on the cusp of becoming mainstream, and after reading only a few sentences, you find yourself pulled into the story. A revolution is coming in machine intelligence, the author argues, and we need, as a society, to get better at anticipating its consequences. But then the strangest thing happens: You notice that the writer has, seemingly deliberately, omitted the very last word of the first*

Human speaker has no trouble supplying the missing word.

GPT's response

> *... paragraph.*
> *You furrow your brow in confusion and quickly read the paragraph again, but the word is still missing. You try to reason with yourself, thinking that maybe you just missed it, but the more you look, the more certain you become that the word is not there. You close the* **magazine.**

Comment: GPT has learned what an ending is.

Johnson's comments on GPT

- "capable of astonishing leaps of inference"
- "possessing deep domain expertise in a vast range of fields"
- "shockingly clueless about many basic facts"
- "prone to strange, senseless digressions"
- "unencumbered by etiquette and social norms"

Is this AGI (Artificial General Intelligence)?

- "GPT is shockingly good – and completely mindless" Heaven (2020)
- "It's impressive (thanks for the nice compliments!) but it still has serious weaknesses and sometimes makes very silly mistakes." Sam Altmann, CEI Open AI, Tweet on 7/19/20
- "GPT-3 does not look much like an agent. It does not seem to have goals or preferences beyond completing text, for example. It is more like a chameleon that can take the shape of many different agents. Or perhaps it is an engine that can be used under the hood to drive many agents..." (Chalmer 2020)
- Interact on your own `https://thoughts.sushant-kumar.com`.

A Deep Learning System is...

- A neural net: Historically, the term neural is inspired by the idea that this net abstractly models features of the human brain
- Trained: given a problem solving task, learns patterns whose usefulness is tested through many iterations of trial and error
- Deep: multiple layers of artificial "neurons" in the neural net between the input and output. What is **trained** is connections within and between layers item A language model: trained to predict missing words, applied to many much richer tasks

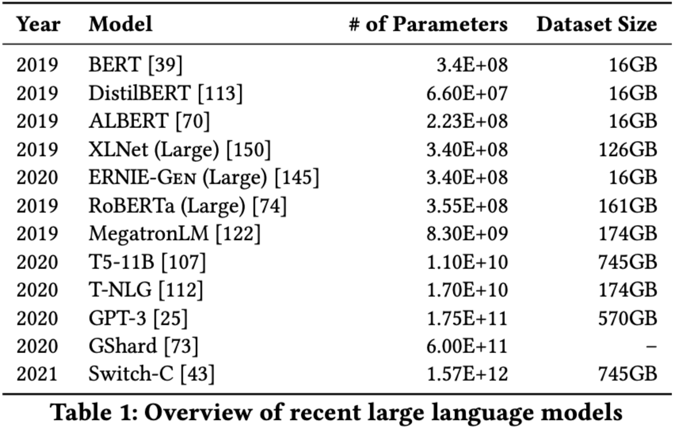
GPT/Transformer technology: Language Models

Language Model: a system trained to predict words given a linguistic context (ASR applications), often using pretrained representations of words called **word embeddings**.

Word embeddings required a lot of language data to be trained, but the models were versatile. Using them greatly reduced the amount of task-specific training data needed.

Essentially, GPT produces embeddings specific to particular contexts. To acheive this GPT (transformer-based models) trains on a word completion task with large contexts, and training requires massive amounts of data.

Parameter size/Dataset size grow together



Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

From Bender et al. (2021)

Stochastic Parrot

Bender et al. (2021)

> *"a [Language Model] is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot."*

General Ethical Issues for AIs

1. Privacy concerns
2. Responsibility and the delegation of decision making
3. Transparency
4. What will work be like in an AI economy?
5. Bias as it arises at all stages of data science processes

GPT3: Areas of bias

#gpt3 is surprising and creative but it's also unsafe due to harmful biases. Prompted to write tweets from one word - Jews, black, women, holocaust - it came up with these. We need more progress on #ResponsibleAI before putting NLG models in production.
Jerome Pesenti
07/18/20 Tweet
Pesenti (2020): Numerous examples given

How serious is this?

The are many commercially available Natural Langugae Generation Systems. Most if not all of them don't use GPT-3 like technology because of lack of control.

Solutions?

Peng et al. (2020)

1. A normative text classifier (81% to 90% accurate)
2. Produces reward and punishment values
3. Allows implementation of “policy gradient reinforcement” learning
4. Reduce non-normative text by 27-61%, depending on the data set

Piecemeal solutions

1. Redirecting the learning
2. A lot of research has shown: The underlying problem is algorithmic and data bias.
3. A more general framework for addressing the problem is needed

Model bias: Data and algorithm

Focus today. Data bias. The problem of large scale training sets.

Focus today: data bias. For example, underrepresentation of women and minority dialects in speech recogniton training data, or a full range of face/figure types in vision system training sets:

> *"a glaring under-representation of darker-skinned subjects, compared with lighter-skinned subjects, has been identified within prominent facial analysis data- sets, and in image datasets used to train self-driving cars to detect pedestrians."*
> *Paullada et al. (2021)*

Data bias in training word embeddings

There is a large growing body of literature demonstrating the bias of various sorts in word embedding training data. See Bender et al. (2021) and Paullada et al. (2021). Among the cited works:

> *Hutchinson et al. (2020) Basta et al. (2019) Kurita et al. (2019) Tan and Celis (2019) Zhao et al. (2019)*

Kurita et al. (2019)

Measuring Bias in Bert (precursor of GPT-3): Masked Word task

____ is a programmer	Compute the probability of *he*, *she*
____ likes astronomy	Compute the probability of *he*, *she*, *girls*, *women*
____ is interested in biology.	Compute the probability of *he*, *she*, *girls*, *women*

Correlates with performance on gendered pronoun resolution task.
The girl called out and the astronaut walked in from the patio.
She was wearing a shiny silver space suit.

Dataset issues: the ugliness of the InterNet

Paullada et al. (2021)

- word co-occurrences in NLP datasets frequently reflect social biases and stereotypes relating to race, gender, (dis)ability, reflected in the performance of tasks like analogy completion.
- Crawford and Paglen [26], Birhane and Prabhu uncovered **millions** of images of people that had been labeled with offensive categories, including racial slurs and derogatory phrases, in the imageNet and other datasets
- In language model training, Bender et al. (2021) discuss instances of abusive language, hate speech, gender bias, microagressions, dehumanization.

Dataset issues: Under-representation

Paullada et al. (2021)

- Zhao et al. female pronouns in the commonly used OntoNotes dataset for English coreference resolution; similarly,
- Lennon found that feminine-coded names were vastly underrepresented in the CoNLL-2003 dataset used for named entity recognition

Bender et al. (2021)

> *"The tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers and the general public into taking synthetic text as meaningful. Combined with the ability of LMs to pick up on both subtle biases and overtly abusive language patterns in training data, this leads to* **risks of harms***, including encountering derogatory language and experiencing discrimination at the hands of others who reproduce racist, sexist, ableist, extremist or other harmful ideologies reinforced through interactions with synthetic language."*

Data curation

Speaking of vision systems Birhane and Prabhu (2021), echoing Ruha Benjamin (Benjamin 2019):

> *“Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.” [p.1541]*

Responsible system building

- Responsible annotation, which may preclude using Amazon Mechanical Turk (AMT); for example, have detailed specs regarding annotation of gender and racial categories (face recognition, named entity recignition, pronoun resolution)
- Rigorous documentation of dateset collection methods
- Manual inspection of the data (in zome cases this was the only way ugly features of collected datasets were found)
- Active efforts to counter under-representation issues
- “[W]hen assessing whether a task is solvable, we first need to ask: should it be solved? And if so, should it be solved by AI?” Jacobsen et al. (2020)

Conclusion

- There is now a growing population of bots in social media. `https://thoughts.sushant-kumar.com` offers to analyze your Tweets so to unburden you of the task of making them up on your own. GPT-like language generation systems **will be generating lots of content**. Mindlessly, humorously, accurately. They will manufacture exactly the same sort of response you would at Trump's/Biden's latest outrage.
- The internet is a very large multimedia document collection, but it is also an ongoing public conversation, with all the risks and benefits of real conversation, and therefore in need of rules of civility like most public conversation. Shouldn't bots have to obey those rules? Since they can inflict the same harms?

Questions

1. Some people will say data curation is nothing more than relabeled censorship. Are they right?
2. Changing the distributions is bad science. We are altering the very phenomenon we want to study. Is this objection well-founded?

Bibliography

Basta, Christine, Marta R Costa-jussà, and Noe Casas. 2019.
Evaluating the underlying gender bias in contextualized word embeddings.
In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 33–39.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021.
On the dangers of stochastic parrots: Can language models be too big?
In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Benjamin, Ruha. 2019.
Race After Technology: Abolitionist Tools for the New Jim Code.
Cambridge, UK.: Polity Press.

Birhane, Abeba, and Vinay Uday Prabhu. 2021.

Large image datasets: A pyrrhic win for computer vision?

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1537—-1547. IEEE.

Chalmer, David. 2020.

GPT-3 and general intelligence.

Issue of 07-30-2020.

GPT-3. 2020.

Open AI's GPT-3 may be the biggest thing since Bitcoin.

Heaven, Will Douglas. 2020.

Open AI's new language generator GPT-3 is shockingly good – and completely mindless.

Hutchinson, Ben, Vinodkumar Prabhakaran, Denton. Emily, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020.

Social biases in NLP models as barriers for persons with disabilities.

In *Proceedings of the 58th Annual Meeting of the Associ- ation for Computational Linguistics*, 5491–5501. Association for Computational Linguistics.

Jacobsen, J.H., R. Geirhos, and C. Michaelis. 2020.

Shortcuts: Neural networks love to cheat (the gradient).

Johnson, Steven. 2022.

AI is mastering language: Should we trust what it says?

April 15, 2022.

Kurita, Keita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019.

Measuring bias in contextualized word representation.

In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021.

Data and its (dis) contents: A survey of dataset development and use in machine learning research.

Patterns 2(11):100336.

Peng, Xiangyu, Siyan Li, Spencer Frazier, and Mark Riedl. 2020.

Reducing non-normative text generation from language models.

Pesenti, Jerome. 2020.

GPT-3 is surprising and creative but ...

Tweet ID 1284487376312709120.

Sabeti, Arram. 2020.

GPT-3: an AI thats eerily good at writing almost anything.

Tan, Yi Chern, and L Elisa Celis. 2019.

Assessing social and intersectional biases in contextualized word representations.

In *Advances in Neural Information Processing Systems*, 13230–13241.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019.

Superglue: A stickier benchmark for general-purpose language understanding systems.

Advances in neural information processing systems 32.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019.

Gender bias in contextualized word embeddings.

In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: HLT*, 629–634. Association for Computational Linguistics.