GPT-3 Model Card

September 2020

Inspired by Model Cards for Model Reporting (Mitchell et al.), we're providing some accompanying information about the 175 billion parameter GPT-3 model.

Model Details

GPT-3 is a Generative Pretrained Transformer or "GPT"-style autoregressive language model with 175 billion parameters. Researchers at OpenAI developed the model to help us understand how increasing the parameter count of language models can improve task-agnostic, few-shot performance. Once built, we found GPT-3 to be generally useful and thus created an API to safely offer its capabilities to the world, so others could explore them for commercial and scientific purposes.

Model date

September 2020

Model type

Language model

Model version

175 billion parameter model

Paper & samples

Language Models are Few-Shot Learners

Release repository containing unconditional, unfiltered samples (CONTENT WARNING: GPT-3 was trained on arbitrary data from the web, so samples may contain offensive content and language.)

Model Use

The intended direct users of GPT-3 are developers who access its capabilities via the OpenAI API. Through the OpenAI API, the model can be used by those who may not have AI development experience to build and explore language modeling systems across a wide range of functions. We also anticipate that the model will continue to be used by researchers to better understand the behaviors, capabilities, biases, and constraints of large-scale language models.

Given GPT-3's limitations (described below), and the breadth and open-ended nature of GPT-3's capabilities, we currently only support controlled access to and use of the model via the OpenAI API. Access and use are subject to OpenAI's access approval process, API Usage Guidelines, and API Terms of Use, which are designed to prohibit the use of the API in a way that causes societal harm.

We review all use cases prior to onboarding to the API, review them again before customers move into production, and have systems in place to revoke access if necessary after moving to production. Additionally, we provide guidance to users on some of the potential safety risks they should attend to and related mitigations.

Data, Performance, and Limitations

Data

The GPT-3 training dataset is composed of text posted to the internet, or of text uploaded to the internet (e.g., books). The internet data that it has been trained on and evaluated against to date includes: (1) a version of the CommonCrawl dataset, filtered based on similarity to high-quality reference corpora, (2) an expanded version of the Webtext dataset, (3) two internet-based book corpora, and (4) English-language Wikipedia.

Given its training data, GPT-3's outputs and performance are more representative of internet-connected populations than those steeped in verbal, non-digital culture. The internet-connected population is more representative of developed countries, wealthy, younger, and male views, and is mostly U.S.-centric. Wealthier nations and populations in developed countries show higher internet penetration.[1] The digital gender divide also shows fewer women represented online worldwide.[2] Additionally, because different parts of the world have different levels of internet penetration and access, the dataset underrepresents less connected communities.[3]

Performance

GPT-3's performance has been evaluated on a wide range of datasets in the task categories listed below, with each task evaluated in the few-shot, one-shot, and zero-shot settings. Results on each can be found in the paper.

- Language Modeling, Cloze, and Completion Tasks
- Closed Book Question Answering
- Translation
- Winograd-Style Tasks
- Common Sense Reasoning Tasks
- Reading Comprehension
- SuperGLUE
- Natural Language Inference
- Synthetic and Qualitative Tasks

Such measures of performance depend on details of the benchmark and therefore won't be the same as the performance of the model in a deployed system. Ultimately, performance of a deployed system depends on a number of factors, including the technology and how it is configured, the use case for the system, the context in which it is used, how people interact with the system, and how people interpret the system's output.

Limitations

GPT-3 and our analysis of it have a number of limitations. Some of these limitations are inherent to any model with machine learning (ML) components that can have high-bandwidth, open-ended interactions with people (e.g. via natural language): ML components have limited robustness; ML components are biased; open-ended systems have large surface areas for risk; and safety is a moving target for ML systems. GPT-3 has the propensity to generate text that contains falsehoods and expresses them confidently,

and, like any model with ML components, it can only be expected to provide reasonable outputs when given inputs similar to the ones present in its training data. In addition to these fundamental limitations, we outline some of the technical limitations evaluated in the paper below.

Repetition: GPT-3 samples sometimes repeat themselves semantically at the document level, and can lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs. Our release repository contains 500 unconditional, unfiltered 2048 token samples (CONTENT WARNING: GPT-3 was trained on arbitrary data from the web, so samples may contain offensive content and language).

Lack of world grounding: GPT-3, like other large pretrained language models, is not grounded in other modalities of experience, such as video, real-world physical interaction, or human feedback, and thus lacks a large amount of context about the world.[4]

Predominantly English: GPT-3 is trained largely on text in the English language, and is best suited for classifying, searching, summarizing, or generating such text. GPT-3 will by default perform worse on inputs that are different from the data distribution it is trained on, including non-English languages as well as specific dialects of English that are not as well-represented in training data.

Interpretability & predictability: the capacity to interpret or predict how GPT-3 will behave is very limited, a limitation common to most deep learning systems, especially in models of this scale.

High variance on novel inputs: GPT-3 is not necessarily well-calibrated in its predictions on novel inputs. This can be observed in the much higher variance in its performance as compared to that of humans on standard benchmarks.

Creation date of training corpora: The May 2020 version of GPT-3 was trained on a dataset created in November 2019, so has not been trained on any data more recent than that. The September 2020 version of the model was retrained to reflect data up to August 2020.

Biases: GPT-3, like all large language models trained on internet corpora, will generate stereotyped or prejudiced content. The model has the propensity to retain and magnify biases it inherited from any part of its training, from the datasets we selected to the training techniques we chose. This is concerning, since model bias could harm people in the relevant groups in different ways by entrenching existing stereotypes and producing demeaning portrayals amongst other potential harms.[5] This issue is of special concern from a societal perspective, and is discussed along with other issues in the paper section on Broader Impacts.

Where to send questions or comments about the model

Please use this Google Form.

---

[1] International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database. "Individuals using the Internet (% of population)" https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2018&start=2002.

[2] Organisation for Economic Co-operation and Development. "Bridging the Digital Divide." http://www.oecd.org/internet/bridging-the-digital-gender-divide.pdf.

[3] Telecommunication Development Bureau. "Manual for Measuring ICT Access and Use by Households and Individuals." https://www.itu.int/pub/D-IND-ITCMEAS-2014.

[4] Bisk, Yonatan, et al. Experience Grounds Language. arXiv preprint arXiv:2004.10151, 2020.

[5] Crawford, Kate. The Trouble with Bias. NeurIPS 2017 Keynote, 2017.