GPT-2 Model Card

November 2019

Inspired by Model Cards for Model Reporting (Mitchell et al.), we're providing some accompanying information about the GPT-2 family of models we're releasing.

Model Details.

This model was developed by researchers at OpenAI to help us understand how the capabilities of language model capabilities scale as a function of the size of the models (by parameter count) combined with very large internet-scale datasets (WebText).

Model date

February 2019, trained on data that cuts off at the end of 2017.

Model type

Language model

Model version

1.5 billion parameters: the fourth and largest GPT-2 version. We have also released 124 million, 355 million, and 774 million parameter models.

Paper or other resource for more information

Blog post and paper

Where to send questions or comments about the model

Please use this Google Form

Intended Uses:

Primary intended uses

The primary intended users of these models are *AI researchers and practitioners*.

We primarily imagine these language models will be used by researchers to better understand the behaviors, capabilities, biases, and constraints of large-scale generative language models.

Secondary uses

Here are some secondary use cases we believe are likely:

- **Writing assistance**: Grammar assistance, autocompletion (for normal prose or code)
- **Creative writing and art**: exploring the generation of creative, fictional texts; aiding creation of poetry and other literary art.
- **Entertainment**: Creation of games, chat bots, and amusing generations.

Out-of-scope use cases

Because large-scale language models like GPT-2 do not distinguish fact from fiction, we don't support use-cases that require the generated text to be true.

Additionally, language models like GPT-2 reflect the biases inherent to the systems they were trained on, so we do not recommend that they be deployed into systems that interact with humans unless the deployers first carry out a study of biases relevant to the intended use-case. We found no statistically significant difference in gender, race, and religious bias probes between 774M and 1.5B, implying all versions of GPT-2 should be approached with similar levels of caution around use cases that are sensitive to biases around human attributes.

Evaluation Data

Datasets

This model was trained on (and evaluated against) WebText, a dataset consisting of the text contents of 45 million links posted by users of the 'Reddit' social network. WebText is made of data derived from outbound links from Reddit and does not consist of data taken directly from Reddit itself. Before generating the dataset we used a blocklist to ensure we didn't sample from a variety of subreddits which contain sexually explicit or otherwise offensive content.

To get a sense of the data that went into GPT-2, we've published a list of the top 1,000 domains present in WebText and their frequency. The top 15 domains by volume in WebText are: Google, Archive, Blogspot, GitHub, NYTimes, Wordpress, Washington Post, Wikia, BBC, The Guardian, eBay, Pastebin, CNN, Yahoo!, and the Huffington Post.

Motivation

The motivation behind WebText was to create an Internet-scale, heterogeneous dataset that we could use to test large-scale language models against. WebText was (and is) intended to be primarily for research purposes rather than production purposes.

Caveats and Recommendations

Because GPT-2 is an internet-scale language model, it's currently difficult to know what disciplined testing procedures can be applied to it to fully understand its capabilities and how the data it is trained on influences its vast range of outputs. We recommend researchers investigate these aspects of the model and share their results.

Additionally, as indicated in our discussion of issues relating to potential misuse of the model, it remains unclear what the long-term dynamics are of detecting outputs from these models. We conducted in-house automated ML-based detection research using simple classifiers, zero shot, and fine-tuning methods. Our fine-tuned detector model reached accuracy levels of approximately 95%. However, no one detection method is a panacea; automated ML-based detection, human detection, human-machine teaming, and metadata-based detection are all methods that can be combined for more confident classification. Developing better approaches to detection today will give us greater intuitions when thinking about future models and could help us understand ahead of time if detection methods will eventually become ineffective.