Text generation by finetuned GPT-2 model

Maria Bancerek

Advanced Applications of Neural Networks 2021

1 Introduction

The project was aimed at observing how specific datasets influence the text generated by a neural network model. GPT-2 model by OpenAI was finetuned on various datasets consisting of eihter scientific abstracts, news articles or novels. The text generated by three models was then compared.

2 Methods

2.1 Data scraping

The first stage of the project was aquiring and preparing the data.

- For scientific abstracts dataset, due to accessibility, the arXiv (https://www.arXiv.org) webpage was chosen. The dataset consists of 200 abstracts including a keyword *neural networks*.
- For news dataset, The New York Times webpage (https://www.nytimes.com/) was chosen. The dataset consists of 55 articles scraped from the main page of the portal.
- For novels dataset, the books were downloaded via the Gutenberg Project (https://www.gutenberg.org/). The dataset consists of 16 parts of english novels.

2.2 Finetuning the GPt-2 model

A small(er) version (124M parameters) of GTP-2 model by OpenAI was loaded and finetuned on aforementioned datasets. The finetuning was adapted from (https://github.com/minimaxir/gpt-2-simple) and (https://github.com/nshepperd/gpt-2/blob/finetuning/train.py). Adam optimizer was used with learning rate 0.0001; other parameters are specified in the code.

2.3 Text generation by trained models

For text generation, 2 options were used: generation of a random chunk of text and generation of a text following a given prefix. The prefixes were chosen to be: (a) specific for abstracts, i.e. *Complex classification problems...*, (b) specific for news, i.e. *Many argue that european policy of...*, (c) specific for novels, i.e. *My eyes filled with tears...* or (d) rather neutral, i.e. *It is surprising....*

3 Results

Examples of random text chunks generated with the models are:

- model finetuned on novels:

  So I thought, if I am going to make a living as a musician, I must be making a living as a musician'–she said. I can do it. I can learn any language I want at the school I want to go to, and I can sing English well. I have done it by doing what many good singers do in the afternoon: singing in the tones of cool credulity, rather than the cool, calculated, cowardly, cruel tongue...

  It was a solemn gray morning in Autumn, and the sky was a pallid shade. The sun had set upon the site of the first dig, and the first Sabbath meal had been served. The birds had been given a good purification through the witching, through the sacrifice of wild meats...

- model finetuned on news:

  A proposal to revise Britain's criminal code would put a high-speed internet cable over the top and put online a record of every internet pass through, the Mail on Sunday has revealed. The proposals would also make it easier to banish content overseas, as in a country like China, which has no plans to make a video service over the internet. Legal challenges have been filed at the Supreme Court...

  It was supposed to be a summer of experimentation and experimentation, but the pandemic turned into a year of murder, rape and experimentation. A young woman was forced to work at a hospital where she was tortured and forced to give birth. A doctor was tortured and killed. A doctor had his career cut short. By November, nearly half of Americans said the flu had ruined their vacations. A dozen American soldiers were imprisoned. A Uyghur woman was driven to the brink of oblivion.

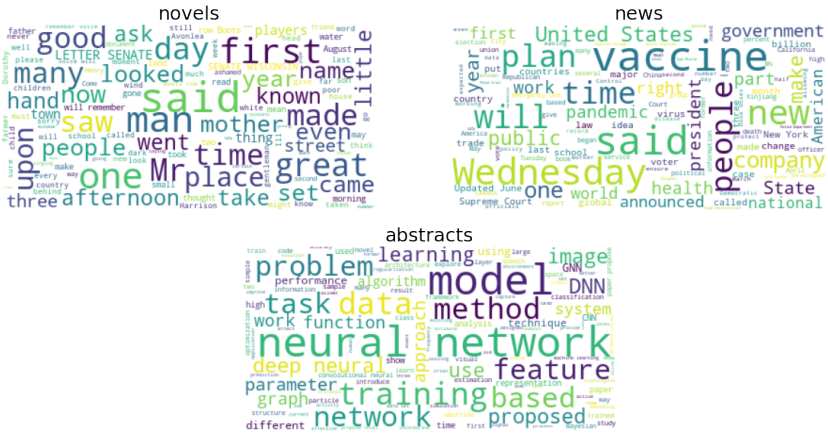
- model finetuned on abstracts:

  Dance music is a popular activity for a group of high-functioning neural networks to dance. In this paper, we investigate the activity of neural networks trained on a domain-specific mixup of sample rates and settings to perform dance analysis. In this way, the neural networks are able to

capture the dynamic range of the sound and determine the appropriate intensity level for the task...

Bayesian neural networks are commonly used in neuroscience to model single- or multiple-particle events. In this work, we investigate Bayesian occlusion methods in neural network models. We use data points collected from the scalp during training and observe the occurrence of events focal, delayed, and swaying with frequency, amplitude, and segmentationorder using a photonic process modeled by FLOP...

Generated texts seem different at first glance already. Despite using small datasets for finetuning, the influence on the generated text is clearly visible. To further illustrate that, word maps of 100 generated texts are presented for each model:



For the model tuned on novels the map has some adjectives and probably most common words people use in *day-to-day* language (like *mother* or *time*); the news model referred a lot to global pandemic situation, also a lot of words like *government*, *public*, *national* etc. appear here. The abstracts model is pressumably the most specific; apart from words like *neural*, *network*, *data* etc. there are some acronyms probably unknown to wast majority of people, like *GNN* or *DNN*.

When given the task to finish a random sentence, most of the times the models kept their unique style of the text: When finishing sentences **It is surprising...**, **My eyes filled with tears...**, the models produced an output:

- model finetuned on novels:

  It is surprising how little is known about the human heart and mind. Its growth, either slow or steady, is interrupted by discomforts and dangers.

  My eyes filled with tears as she spoke, and I could see that she was very sorrowful; but she did not utter a word and the big red bottle she was pointing at Mr. Utterson slipped slightly past her lips.

- model finetuned on news:

  It is surprising, then, that the media does not take note of this and respond with more depth and fact-checking, especially given what we know about the Pfizer-BioNTech trial at the time.

  My eyes filled with tears and he said, "That's it.'" When the trial began in November, Kwame was one of the first journalists to arrive on the scene.

- model finetuned on abstracts:

  It is surprising that such remarkably successful and very detailed approaches still have not been extensively used by existing techniques for capturing and quantifying complex 3D geometric features in images.

  My eyes filled with tears as I read these accounts of the channeling of deep neural networks with their non-stationary counterparts.

  My eyes filled with tears. When the photos are readmitted to the cloud, the neural network learns to contrast images of the images of the processed images with the images of the readmitted datasets.

All models were able to finish the neutral sentence in a sensible way. The two latter models seem to experience more difficulty when given an *emotionally* marked prefix. It is especially visible in the last example, where the prefix sentence was rapidly ended and the text is not really coherent. More results are attached to the code.

4 Conclusions

To conclude, a surprisingly small dataset is enough to drastically alter the GTP-2 performance. It is best visible when the dataset is a very specific one, e.g. from a given scientific field. Generated texts are coherent and rather correct grammatically. Bigger dataset and longer training could potentially improve small flaws. Nevertheless, this strikes me as a dangerous tool for generating fake news about desired topic with desired language (and polarization).