Asad Malik

CS 644

Prof. Vic Berry

Due: December 9th, 2020

**Project Design and Documentation**

**Project Introduction:**

The project that I have selected to work on is an implementation of OpenAI’s GPT-2 language transformer model, which is a large-scale unsupervised language model that looks to predict the next word in text based on the combination of previous corresponding text that it sees. The goal of this project is to perform introductory analysis on how the language framework is implemented, methods around how to custom train it for specific needs and generate output that is relevant to the use case that I am looking for.

**Project Motivation**

In order to get a sense of what I am going for, it will be helpful to get a background into why I have chosen to go down to the path of exploring language models. A habit that I’ve adopted for about a year now is journaling. I find it extremely helpful to write about whatever I may be going through and to push myself to explore the thoughts in my head. To that end, I’ve made an App that I call Rubiic, which is essentially a space where users can write their thoughts and create a journal of their mental state.

If you would like to read more on my app and motivations, it was featured on BU Spark and they captured the story which can be found at: <http://www.bu.edu/spark/2020/11/16/got-the-covid-19-blues-this-app-is-a-refuge/>

My long term plan is to implement what could essentially be considered an AI therapist, which would be able to go from the text that the user is inputting, and be interactable. The AI would be able to analyse the users input and prompt them in ways that would allow them to further explore their thoughts or whatever else they might be writing about.

**Implementation**

The implementation of the GPT-2 starts with the cloning of the library that is available on Github. Second, I downloaded the model that comes from OpenAI, that forms the basis of the training. This is a tricky section, as there are multiple options. I downloaded the smallest model, which comes with 117M training parameters and weights. The largest parameter model is 774M parameters, which is 6x in size of the model that I chose but that comes with its drawbacks. While the results that you get from that model are more accurate. That is because the training data that is then passed through the model is evaluated based on many more parameters, and while the results are noticeably better in the larger model, the performance takes a big hit and it takes the model much longer to do the training and once the training is done, the lag between when you give the model a prompt to when it gives you its answers becomes more prolonged as well. Since my project is about doing exploratory analysis, as opposed to getting the best results possible, I decided to go with the smallest model.

After selecting the base model that is going to form the training parameters, it is time to select the training data itself. The model itself comes pre trained based on absurdly large amounts of text data scraped from the internet, the purpose of the training is to feed it text, and the model learns and mimics the language posed to it in the custom user data section part. For my training purposes, this was challenging as I was not able to find a dataset that matched exactly my needs. I would've liked a dataset that I could pick, that would have text in the form of expression of thoughts, feelings and ideas, and then complimentary questions. This would be more specific for the use case I am looking to integrate into my application. I was unable to find an already existing dataset to match those needs, but in future development of this project, I will instead look to transcribe text from podcasts that go in the format that resembles the use case I am talking about.

For implementation now, I have used a glossary of terms and their definitions of common mental health disorders. An example:

**Alcohol and drug dependence**: A cluster of physiological, behavioural and cognitive phenomena in which the use of alcohol or drugs takes on a much higher priority for a given individual than other behaviours that once had greater value. The alcohol or drug withdrawal state refers to a group of symptoms that may occur upon cessation of alcohol or drug after its prolonged daily use. - <https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health/data-and-resources/key-terms-and-definitions-in-mental-health>

This should train the model to respond to prompts of different conditions and it should output their definitions.

The training mechanism is rather simple with GPT-2. The package comes with a built in encode.py file that transforms a training text file into a npz file. The command is

*python encode.py \*name\_of\_training\_file.txt\* training.npz*

You simply execute the encode.py file with the arguments of the input training file and the output file which you want in the form of a npz file.

After which you execute the training file feeding it the training.npz created from running the encode file. The command is:

*python train.py --dataset training.npz*

I created two implementations of the trained model, one that generates random samples of text from the training model, and the other that takes the starting block text in the form of a user initiated prompt. The results are interesting.

Here are a few samples of my conversations with the model:

Model prompt >>> Depression

======================================== SAMPLE 1 ========================================

. One study showed that the number of depressed individuals increased with the severity of depression. These results clearly indicate that depression can cause changes in the brain's response to stress.4 As noted in our review above, in the present study, we used the mean stress level for all the depressed individuals to determine if we could determine the rate of change. We estimated the rate of change for each of the 5 depressed individuals to be 0.7%.

In short, we expected a 2-fold decrease in the number of depression cases. Because of the magnitude of the difference in the number of depression cases, we expected a smaller increase in the number of depression cases compared with the controls.

Model prompt >>> What is bipolar disorder?

======================================== SAMPLE 1 ========================================

Bipolar disorder refers to a serious mental state, including paranoia, delusional thinking, agitation, aggression. The clinical and biological causes of the disease are not clear. The exact clinical manifestations are poorly understood, and there is no available, clinically validated treatment.

Bipolar disorder is sometimes confused with schizophrenia, with the neuroanatomy associated with having an overactive central nervous system. The term bipolar disorder may be related to an overactive nervous system that is characterized by an intense and complex focus on one central focus.

The term bipolar disorder is not a disorder of impulsivity or excessive-intense, and has nothing to do with impulsivity. Some people with bipolar are over in their behavior and may have high- or low

Model prompt >>> Why am I feeling depressed?

======================================== SAMPLE 1 ========================================

If that's true, why am I not feeling depressed? I don't have any good answers to that. I guess I just have this feeling where I have this feeling of...oh, yeah, I don't know. I don't really know what is going on. I'm feeling really frustrated at myself or something like that. It's hard for any of us. And then I have that feeling where, yeahh, I'm actually enjoying myself. I'm feeling really great and I'm enjoying myself and I'm going to do this and I'm going to play, and I'm not going to let myself go and I'm going to end up with an addiction. And I feel like it's gonna be like a life of no change

================================================================================

**Next Steps:**

As I mentioned above, the next steps is to train this model in such a way that it becomes conversational. That involves changes to the training structure, which I plan on basing it on the transcripts of podcast sessions to create a natural conversational flow to it. It also means modifying the interactive\_conditional\_sample.py script to take into account the already generated text into considerations in it’s prompt, in addition to entered text, so it can output its answers with conversational context as a variable consideration.

I will then put it on my server and make it live for anyone to be able to talk to it by visiting the url.

**Build Instructions:**

**Go into project src directory.**

1. Install project requirements: pip install -r requirements.txt
2. Install tensorflow: pip install tensorflow==1.14.0
3. You should be able to run the python file as the training documents are included in the submitted zip file:
4. To run script: python interactive\_conditional\_samples.py --top\_k 40 --model\_name 117M\_TRAINED --length 150 --truncate "<|endoftext|>"