Daniel Felker
MCS 410
November 6, 2022

Review: [Approach] OpenAI GPT, GPT-2 as illustrated by the paper: “Language Models are Unsupervised Multitask Learners”[1]

The current state of natural language processing when combined with machine learning most often involves supervised learning approaches. In this paper the authors demonstrated that language models can begin to learn dataset specific tasks without any explicit supervision.

In order to understand this approach a concept called zero-shot learning needs to be expanded briefly. The concept centers around machine learning where at test time the learner observes samples not observed during training and those samples are predicted to fall into an appropriate class. Wikipedia provides a good example to this problem by considering a group of pictures with horses and then providing a zebra to your learner. In the NLP scope you could imagine sentence structure and organization as a potential application of this. Consider verb first as compared to noun first languages and how they represent the same sentence differently for example. Within the paper the authors give a good example of this phenomenon when comparing a French to English sentence: “As-tu aller au cinema?” in French translates directly to Have-you to go to the movies? But in English the phrase would be Did you go to the movies? This contextual differences are one element they hope to capture with this work.

One significant feature of this work comes in the form of curated data from web scraping that they called WebText. This dataset contains data that in general can be considered intelligible and would aid others who are looking to examine methods to accomplish zero-shot learning on text. The first version of WebText was created with the goal of having high quality documents, to accomplish this they used the reddit social media platform and links that received a minimum quality score from readers. The first data set for WebText resulted in a document set that contained over 40GB of high quality text. The authors then converted the documents into a data representation that is called Byte Pair Encoding. This encoding technique allows for a middle ground between character encoding and word level encoding. This combination is generally done in a way where common symbol sequences are done at a word level and infrequent sequences at a byte level.

With this dataset the authors then implemented their model and performed a variety of tests on the data including children’s book test, which is a test where the system is required to predict which of 10 possible choices for an omitted word is correct. The results from this were exceptionally good and an interesting component of this test with regard to thinking of the problem of language modeling was considering proper nouns. The authors tried to have their system fill in the blanks on the book Jungle book which would of course include proper nouns such as names of heroes. The general idea for predicting is determining the highest probability for a word to occur in the given context, p(output | input, task) generally.

The authors also tested the work on several other metrics including reading comprehension, LAMBADA which is the ability of the system to predict the final word, summarization, translation and question answering. For each of these recognized standard metric tests this system performed well as compared to competing methods.

For reading comprehension the technique was compared to the Conversion Question Answering dataset (CoQA)[3]. This is a collection of documents from 7 different domains and the idea is to test the system's ability to comprehend the document it has been given as well as it's ability to infer logic for questions such as Why? Summarization is exactly what it sounds like, it's the ability of the method to summarize something. For this the author's used CNN and Daily Mail datasets and a technique that looked at the top tokens from a random sampling of size 'k'. This allowed them to generate sentences that reflected the most observed details in an article. An interesting observation from the authors on this technique was that it generally got the summaries right but would mess up on details such as how many cars were in a crash or if a logo appeared on a hat or shirt. This is an interesting by-product that needs to be further explored.

Regarding translation the authors did English-to-French translations, this test was much more difficult for their method as compared to the current state of the art, the authors didn't have any insight into why this was currently only that it was better at translating French-English likely do to the abundance of English available as compared to French for the dataset. This problem goes back to the idea of language structure (noun vs. verb) and seems to be a difficult challenge overall.

A final feature of this method that the authors looked into was this idea of generalization vs. memorization in datasets. This comes from the size of datasets getting larger as our ability to gather and store data improves. The events that are occurring now is overlap in data between training and test datasets, you could think of it as if the system has already memorized the data that is in the training set it will naturally perform better. With regard to text this is still an interesting question, how do quantify similar text to prevent pure memorization. With regard to this technology the authors demonstrated that it had a small but repeatable improvement when the overlap was increased.

Overall this method (GPT-2) is a nice example of understanding and learning in natural language processing and the authors provided a nice collection of datasets and samples to explore in one's own work with regard to improving natural language processing abilities to understand and evaluate text meaning.