CS 410 – Technology Review – Fall 2021

Author: Pericles Rocha (procha2@illinois.edu)

A review of OpenAI's study "Language Models are Unsupervised Multitask Learners"

In 2019, the article "Language Models are Unsupervised Multitask Learners" by Radford et al. provided the study of a promising path towards building language processing systems which learn to perform tasks without using a supervised machine learning model. Instead, they propose a technique to perform these tasks from the naturally occurring demonstrations language.

Several natural language processing tasks can benefit from this approach, from question answering, to document summarization, machine translation, and even reading comprehension. These tasks are typically approached with supervised machine learning that use purpose-specific datasets. For instance: to train models that are used for machine translation, pre-trained models that were built specifically for that purpose of translation are used and improved upon. These models need hundreds of thousands of examples to induce functions which generalize well. When we think about building purpose-specific models at scale to resolve different problems, the requirement to use that many examples make it difficult to scale the creation of specialized datasets, and the design of objectives to the degree that may be require to brute force our way with current techniques that require supervised training in order to perform a task.

Systems that used these pre-trained, specialized models, are categorized as narrow experts, rather than competent generalists. The goal of the authors was to research the move towards more "general" language processing systems, which can perform many tasks, without the need to manually create several "specialized" models for each task. The authors attempt to demonstrate that language models begin to learn these generalized tasks without any explicit supervision when trained on a new dataset of millions of web pages called **WebText** – which will be described later in this review.

It's worth mentioning that supervised learning is error prone, so the authors highlight some of the shortcomings of this approach. In supervised learning, individuals will typically manually categorize large datasets to enable training supervised models. For instance, in machine translation, an user would rate a translation, or attempt to correct it manually, and this manual iteration part of the feedback loop can introduce human errors. When models are trained, there's no way to separate "good" labels from "bad" labels automatically. Unsupervised learning, while much more complicated to achieve for this scenario, reduces the chance for these types of errors.

Approach

The authors propose an approach that uses language modelling. Since language has a natural sequential ordering, it is common to factorize the joint probabilities over symbols as the product of conditional probabilities. Learning to perform a single task can be expressed in a probabilistic framework as estimating a conditional distribution *p(output|input)*. This has been variously formalized in multitask and meta-learning settings. Since the supervised objective is the same as the unsupervised objective but only evaluated on a subset of the sequence, the global minimum of the unsupervised objective is also the global minimum of the supervised objective.

The authors speculate that a language model with sufficient capacity will begin to learn to infer and perform the tasks demonstrated in natural language sequences to better predict them, regardless of their method of procurement. If a language model can do this, it can be assumed that it is in fact performing an unsupervised multitask learning. The authors tested whether this was the case by analysing the performance of language models in a zero-shot setting on a wide variety of tasks.

A key part of this study, however, is the dataset used by the authors. In general, previous studies in natural language used supervised learning systems specialized on a single domain of text, with the authors mentioning news articles, Wikipedia, or fiction books as examples. The proposal for this study was to use a dataset which was as diverse and as large as possible to be able to acquire real-life demonstrations of tasks in various domains and contexts. So why not perform a large-scale web crawl (a.k.a. "Common Crawl)?

The problem with a Common Crawl is data quality. The authors mention previous studies where content crawled from the web was "mostly unintelligible". So instead of a "regular" web crawl, the authors created a web scraper that had an emphasis on document quality, by using only web pages which had been curated or filtered by humans which received at least 3 karma (note: "Karma" is Reddit's voting system. The posts with the most karma are the ones you see on the front page). The authors used the approach as a heuristic indicator for whether other users found the link interesting, educational, or just funny.

This curation resulted in 45 million links of high-quality documents, for a total of 40GB of text. The authors named this dataset **WebText**.

For input representation, the authors used Byte Pair Encoding (BPE), which has sown to be useful in several natural language processing applications as a simple form of data compression. If it weren't the case, the authors claim that a base vocabulary of over 130,000 before any multi-symbol tokens are added, which would make it prohibitively large when compared to the 32,000 to 64,000 token vocabularies often used by BPE.

Because the approach used by the authors can assign a probability to any Unicode string, it allowed them to evaluate their language models on any dataset, regardless of pre-processing, tokenization, or vocabulary size.

Finally, they used a Transformer-based architecture for their language models. With four different models having 117, 345, 762 and 1542 million parameters each, the models had 12, 24, 36 and 48 layers respectively.

Experiments and results

The authors performed experiments with eight natural-language processing tasks: Language Modelling, Childen's Book Test, LAMBADA, Winograd Schema Challenge, Reading Comprehension, Summarization, Translation, and Question Answering. The smallest model was equivalent to GPT, and the largest model had over an order of magnitude more parameters than GPT. Still, all models still underfit WebText. The authors called the resulting model **GPT-2**.

The article summarizes results for each task. In general, the analysis suggested that data overlap between WebText training data and specific evaluation datasets provided a small, but consistent benefit to the reported results. As an example, on tasks such as reading comprehension, the performance of GPT-2 is competitive with supervised baselines in a zero-shot setting. But still, the zero-shot performance of GPT-2 is still considered by the authors far from useable.

For the Winograd Schema Challenge, the observed results were that GPT-2 outperforms with an F1 score of about 3 better than the specialized data set. For LAMBADA, GPT-2 performed about 2 perplexities better on examples with 15% overlap.

Finally, GPT-2 is considered by the authors to be able to write news articles about the “discovery of talking unicorns”. The article produces an example for this experiment.

Conclusion

Overall, GPT-2 zero-shots the state-of-the-art performance of 7 out of 8 specialized language model systems. The study proposed by the authors prove that when a language model is trained on a large and diverse enough dataset, it can perform as well as any specialized system that uses supervised learning.

The results were highly dependent not only on the scale of the experiment but also the quality of the documents. In such setting, the authors prove it to be possible to achieve results similar to those of specialized systems without the need for supervised machine learning.

References

- Language Models are Unsupervised Multitask Learners: Language Models are Unsupervised Multitask Learners (openai.com)
- What is Reddit Karma and how do I get it? https://www.howtogeek.com/465411/what-is-reddit-karma-and-how-do-i-get-it/