Xie Yuheng

CS410: Text Information Systems

06/11/2021

Technology Review on GPT-1 and its successors

1. Introduction

The Generative Pre-Trained Transformer (GPT) is a state-of-the-art language model developed by OpenAI in 2018. Over a period of 2 years, GPT-1 has evolved into GPT-2 and then GPT-3, which is one of the most powerful Natural Language Processing (NLP) tools available in our world today. It can write articles, answer questions, translate texts from one language to another, create document summaries and even write code by itself with high accuracies. In this paper, I will discuss and review the technicalities of GPT-1 and its successors, GPT-2 and GPT-3.

2. Body

2.1. Transformer

The core architecture of GPT is the Transformer model. The Transformer was introduced by Vaswani et al. (2017) and it revolutionized the NLP space [1]. Before that, Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) architectures were proposed to solve problems involving text sequences. However, these architectures are severely limited in their abilities to retain recent information when dealing with long sequences. To address this limitation, the attention mechanism was incorporated to extract information from the entire sequence and assign different weights to the tokenized units depending on their relative importance. This mechanism resolves the information decay issue but makes computation expensive for large documents expensive since every step has to be processed sequentially. The

Transformer was proposed to make use of the benefits from the attention mechanism while enabling efficient computation through parallelization.

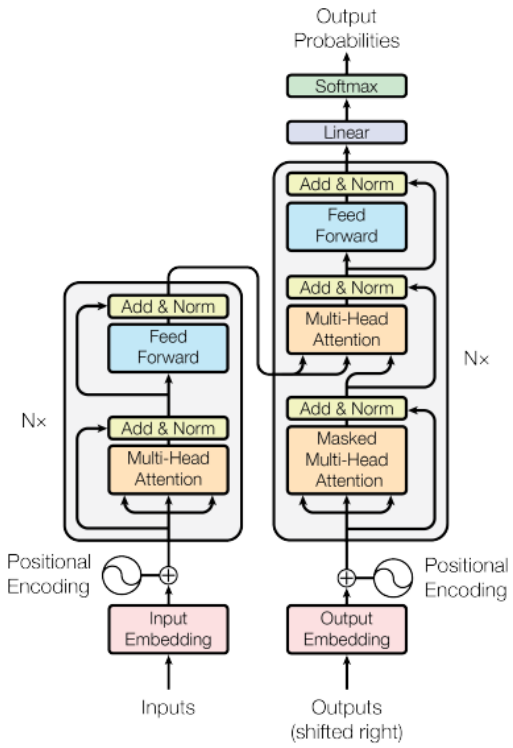

Figure 1: Transformer Architecture [1]

The main feature of the Transformer is the self-attention mechanism which computes the weight of a token (either a word or a sequence of words) with respect to other tokens in a text sequence by only using linear transformations and activation functions without using recurrent units, making it highly efficient and parallelizable. It is further enhanced into the multi-head attention mechanism where the attention weights are not just computed once but multiple times on the entire sequence to generate various attention patterns between each token. A positional encoding vector is added to allow the model to capture additional position information within a text sequence. Lastly, residual connections are added to each encoding and decoding unit to enable better information flow through the network for the model to converge faster. At the time when the Transformer paper was published, it managed to outperform all previously published models

for English-to-German and English-to-French translation tasks, cementing it as the new state-of-the-art model for sequence-to-sequence problems.

2.2. GPT-1

In 2018, Radford et al. published the paper “Improving Language Understanding by Generative Pre-Training” which introduced the GPT-1 model which is based on the Transformer architecture [2]. Before this model was published, most state-of-the-art language models were pre-trained for their specific tasks such as translation, sentiment analysis and topic classification. The GPT-1 model uses semi-supervised learning methods to generalize for other NLP tasks other than the one it was trained for. During the initial unsupervised learning stage (Pre-training), a standard language model objective function is used to generate initial weights. After that, in the supervised fine-tuning stage, another objective function is used to maximize the likelihood of observing the ground truth given a set of features. An auxiliary learning objective is also included into the objective function for quicker convergence and to better generalize the model. Finally, in order to minimize architectural changes to the model during the fine-tuning stage for better generalization, task specific input transformations such as transforming inputs into ordered sequences are performed. GPT-1 is trained on the BooksCorpus dataset which contains some 7000 unpublished books to aid its training on unseen data, and uses the 12-layer Transformer architecture. It performed better than specifically trained supervised state-of-the-art modes in 9 out of 12 NLP tasks, which is an impressive feat since the GPT-1 model is semi-supervised and yet it managed to outperform most of the other fully supervised models in their specific tasks. GPT-1 showcases the power of generative pre-trained language models in performing general NLP tasks, and its huge potential in zero-shot learning if it is trained on a larger dataset and better tuned.

2.3. GPT-2

Following that, GPT-2 was introduced in the paper “Language Models are Unsupervised Multitask Learners” by Radford et al (2019) [3]. GPT-2 was trained on a much larger dataset and had 10 times more parameters than GPT-1. GPT-2 uses the same Transformer architecture as GPT-1, and the only difference is that layer normalization is moved to the input side of each block and an additional normalization layer is added after the final self-attention block. The training objective of the language model is also modified from P(output|input) to P(output|input, task) for task conditioning to allow for zero-shot learning task transfer. In general, GPT-2 performed better than most state-of-the-art NLP models for different tasks, with the exception of text summarization. The evaluation results also show that the performance of the model continues to increase in a log-linear fashion with the increase in capacity of the model without reaching a saturation point, suggesting that the model is still underfitted and its performance can still be increased substantially if the model size is increased.

2.4. GPT-3

As mentioned previously, GPT-2 was underfitted and its performance can still be improved given a larger model size. With that, GPT-3 was published by Brown et al. (2020) as a successor to GPT-2 with 100 times more parameters [4]. The sheer size of the model has enabled it to perform at human level accuracies for certain NLP tasks such as writing of articles given a specific topic. A major discovery from the paper is that large language models are able to develop pattern recognition skills using their input text dataset, which helps the model substantially in zero-shot settings. From the experimental results, there is a positive correlation between the size of the model and its predictive performance in zero-shot scenarios. Evaluation results show that GPT-3 performed as well as if not better than many state-of-the-art models in

various NLP tasks such as translation, question answering and article writing. As such, GPT-3 has become state-of-the-art for general NLP tasks. Despite the impressive performance of GPT-3, it has its limitations and disadvantages. GPT-3 has poor performance on inference tasks such as filling in blanks and tends to lose coherency when generating long sections of text. Also, GPT-3 is algorithmically biased towards gender, race and religion, and it may produce offensive texts under certain contexts. Furthermore, there are external risks associated with GPT-3, since it might be misused for malicious activities, such as spamming, generation of phishing texts and impersonation.

3. Conclusion

In summary, the Transformer model was pivotal in the development of GPT, as it was the core architecture used in all iterations of GPT. The latest version of GPT which is GPT-3, is powerful enough to automate many NLP tasks in our world today given that it can produce results with accuracies close to human levels. With great power comes great responsibility, and the onus also lies on the research community to mitigate the risks of GPT-3 in the event that it is misused.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention is All You Need. arXiv:1706.03762 [cs.CL].

[2] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. OpenAI Blog.

[3] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. OpenAI Blog.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL].