Siyuan Zhang(siyuan3)

CS 410 Tech Review: GPT-3 Language Model

GPT-3 is the latest pre-trained language model published by Open-AI in May 2020, it is also the largest and computationally-expensive machine learning model ever produced. It has a transformer architecture with 175 billion parameters and was trained using 430GB of test data gathered from the internet and books. Training of this model is done by feeding it with a certain length of text and let it predict the next word. The resulting model is so powerful and versatile that it is achieving stunning performances in almost all of the NLP benchmarking tests and outperforming previous state of the art models in a few of them.

The rapid development during the past decade in NLP is mainly due to the discovery of incredible generalization and learning capability of deep neural networks. Before that era the classical approaches aren't providing satisfactory results. The early deep NLP models either tries to learn a vector representation of the words or using recurrent neural networks to read text. The first approach is most known as word2vec and what is exciting is that the learned high dimensional space can understand useful relationships. The later approach mainly represented by sequence to sequence models using recurrent neural networks has demonstrated remarkable performance in tasks like machine translation.

A big breakthrough was made in 2017 when transformer architecture was proposed that leverages a mechanism called attention. This ability allows the model to scale much deeper compare to previous models and further improvements such as BERT were soon published. Most of these models learns by predicting masked words thus can fully take advantage of the self-supervision and train on massive amount of available data, allowing them to break many records in NLP. A downside of such large models is that only the few largest companies with the most funding can afford the development and maintenance of such costly models. With the GPT-3 being the latest and the biggest, it is expected that this model would outperform all the predecessors which it did. Despite not bringing any major architecture innovation but only larger model, GPT-3's experiment results showed a lot of interesting discoveries especially in tasks of meta-learning and few-shot learning which are some of the major open problems in AI research.

Humans are able to learn a new task with only very few examples and demonstrations. Despite researcher's effort, this ability is still extremely hard for machines to achieve as current deep learning models are particularly data hungry often require thousands of training examples to learn even basic novel tasks. Although GPT-3 doesn't explicitly solves the few-shot learning problem, it's empirical evaluation showed interesting insights in this direction. As the model's size scales up to another new level, the low-shot performance of GPT-3 improved significantly. And even GPT-3 has so many parameters, it's still only 0.1% of the synapses in a typical human brain, which leaves huge gap still, not mentioning human neuron architecture being more complex. One of the tasks that GPT-3 was able to learn is 3 digits arithmetics just by training on large text corpus with only very few of such examples exist. This indicates somehow that the model is learning this specific task by itself. The author further confirmed this guess by stating that most of the mistakes that GPT-3 made in performing arithmetic is missing the carrying 1 bit. This shows that GPT-3 is not trying to memorize a table of arithmetics calculation result but rather trying to learn the underlying mechanisms. The model was also able to achieve a above 20% accuracy when calculating combined 1 digit calculation, significantly higher than random guess. This result is very stunning in a way that the only task that the model was trained on is to predict the next word given a certain length of text. It would be interesting be find out if further scaling would lead to even greater meat learning capabilities and to see if reading research paper can enable to trained models to propose and predict new scientific approaches and results.

Another important result from GPT-3 was it's capability in few-shot learning. Typical pre-trained deep neural net models usually requires fine tuning before being applied on to specific down stream tasks. On the other hand, GPT-3 is so large that it doesn't even require fine-tuning anymore and is still able to do well on many of the tasks such as translation, reading comprehension, question answering etc. with remarkable accuracy. The large capacity of the network even allowed GPT-3 to memorize large amount of facts about the word that allows it to answer common questions. Such great few-shot learning ability brings many benefits. One of which is to allow the use of such model in tasks which usually require lots of structured training data. It also makes ML models more accessible to people with such an easy to use prompt. Last but not least, it also makes the ML model more flexible as it is able to handle multiple tasks at the same time without fine tuning to a specific one.

One supervising result from GPT-3 is that it even learning tasks that it was not expected to such as coding. Open-AI provided a prompt that even allows the users to describe a simple coding functionality description, and GPT-3 is able to generate the necessary code such as JavaScript to achieve such functionality with great success rate. Because of the incredible capabilities that GPT-3 has, Open-Ai did not open source the model's weight as there has already been many concerns over the potential use of such a powerful model in many application such as creating realistic spamming or fishing text as well as automatically generating realistic and coherent articles with false or misleading information. Such misuse can bring a lot of trouble and further development of such powerful models should definite take precautions to prevent the misuse to them.

Although GPT-3 brought many breakthroughs in it's ability to perform multiple NLP tasks as state of the art performance as well as it's incredible ability to perform few-shot and meta learning, the model itself doesn't introduce any major innovation in terms of it's architecture. It's basically just a bigger transformer model with more parameters compare to its predecessors and trained on more text data. Although it does prove the idea that a bigger model in NLP leads to better performance, the result is just not that exciting when comparing to the skyrocketed training cost given the enormous computational resource it takes. It also puts a big barrier to many researches in this field who simple don't have the amount of funding and resources that these few large companies have. Besides just scaling up the model to achieve sightly higher performance, it is important for future research to not solely focusing on scaling up. There are many other interesting aspects that needs innovations and It is foreseeable that the development in NLP will bring many benefits to the society in the near future.

Siyuan Zhang(siyuan3)

Reference

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Chen, Ting, et al. "Big self-supervised models are strong semi-supervised learners." *Advances in Neural Information Processing Systems* 33 (2020).