Kevin Vece

vece2@illinois.edu

CS410 Text Information Systems

6 November 2022

Technology Review: A Comparison of Large Language Models

Since the advent of large language models, we have seen an increasing number of approaches to solve the problem of realistic text generation. Perhaps the most notable examples include the GPT-1, GPT-2, and GPT-3 models created by OpenAI. Additionally, there have been alternative open source approaches such as the GPT-J model produced by EleutherAI. In this paper, we explore the evolution of these models and compare and contrast their differences in approach and its effect on their model architectures.

Often cited as the publication that popularized the Transformer architecture, "Attention Is All You Need" by A. Vaswani, N. Shazeer, N. Parmar et al., demonstrated the viability of attention models in the machine translation task. Their model "achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU" (Vaswani et al. 1), and "on the WMT 2014 English-to-French translation task, … establishes a new single-model state-of-the-art BLEU score of 41.8" (Vaswani et al. 1). The novelty presented in these models that allowed such high performance is known as "multi-headed attention". With this new mechanism, language models are able to learn complex relationships between distant entities within a text, solving the memory issues present in recurrent neural networks.

In 2018, the company OpenAI expanded this Transformer concept to produce the language model Generative Pretrained Transformer, or GPT. As described by OpenAI "we

demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task" (Radford "Improving Language Understanding by Generative Pre-Training" 1). This revolutionary approach eliminates the need for large labeled text corpuses, significantly increasing the amount of text available to train models with.

Upon further research, OpenAI was able to refine this model to increase its performance. Through their research and development, they discovered "the capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across all tasks" (Radford "Language Models are Unsupervised Multitask Learners" 1). In other words, they found that increasing the number of parameters and the amount of data ingested by the model to be paramount to the success and learning of the GPT. This discovery led to the development of the GPT-2 model with further improved performance over the prior generation.

In 2020, a year later, the OpenAI team further capitalized on this discovery in their development of GPT-3, which contains an increase from 1.5 billion parameters in GPT-2, all the way to 175 billion parameters in GPT-3. By increasing the parameters and training, they created a model that "greatly improves task-agnostic few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches" (Brown 1). This larger model is able to perform tasks that require "on-the-fly reasoning or domain adaptation, such as unscrambling words, using a novel word in a sentence, or performing 3-digit arithmetic" (Brown 1).

In contrast, the company EleutherAI was able to utilize the knowledge and approach presented by OpenAI, to create an open-source alternative, loosely based on a smaller version of

the GPT-3 6.7B architecture and size. By developing the model with open-access in mind, they were able to make useful improvements for personal developments, including striking a balance between model size and performance, and modifying the architecture for faster parallel computations. As a result, “the zero-shot performance is roughly on par with GPT-3 of comparable size” (Wang 1). Another improvement they have made is the use of rotary positional encoding (RoPE) as opposed to the sinusoidal method proposed in the original Transformer model. This method improves both the simplicity and efficiency of the Transformer model.

By looking at the development year by year, we can see multiple substantial discoveries in the large language model space. Namely, the use of the Transformer architecture, and the use of large text corpora combined with large models. While minor differences such as number of layers and positional encoding schemas may vary, the core concept “Attention Is All You Need” has remained standing firm. Their remarkable performance on reducing the amount of fine-tuning needed for text-based tasks ensures that we will be seeing increased usage of Transformer architectures in the future.

Citations

A Vaswani, N Shazeer, N Parmer, et al. “Attention Is All You Need.” https://arxiv.org/pdf/1706.03762.pdf, 2017.

A Radford, K Narasimham, T Salimans, I Sutskever. “Improving Language Understanding by Generative Pre-Training.” https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018.

A Radford, J Wu, R Child, et al. “Language Models are Unsupervised Multitask Learners.” https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, 2019.

T Brown, B Mann, N Ryder, et al. “Language Models are Few-Shot Learners.” https://arxiv.org/pdf/2005.14165.pdf, 2020.

B Wang, A Komatsuzaki “GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.” https://github.com/kingoflolz/mesh-transformer-jax, 2021.