Hyomin Seo

Title : Paper Analyzation on “BERT”, “Big Bird”

Reading Links :

https://arxiv.org/pdf/1810.04805.pdf :BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
https://arxiv.org/pdf/2005.14165.pdf: Language Models are Few-Shot Learners
https://arxiv.org/pdf/2007.14062.pdf: Big Bird: Transformers for Longer Sequences

Three papers discuss and introduce an innovative approach to improve Natural Language Processing, focusing on evaluating the transformer - its use, advantage, limitation followed by a method to alleviate such limitation. The transformer that appears congruently through these papers is ‘BERT, the bidirectional Encoder Representations from Transformers.’

The key feature of such an approach is that input text is unlabeled. Instead, these models read an article or a post and understand the meaning of words within the context that they are used. So the models will start to see these terms as having a somewhat related context. With more and more data, they will learn more nuances about the different usage and meaning between these related terms. At least that is the theory; the bigger and broader the data is, the more influential the model becomes.

BERT: Bidirectional Encoder Representations from Transformers, introduced in a paper published by GOOGLE AI team 2018. BERT is pre-trained with deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers, letting it create state-of-the-art models with the minimum fine-tuning and output layers. In other words, BERT considers all the words of the input sentence simultaneously and then uses an attention mechanism to develop a contextual meaning of the words. The system shows substantial improvement in numerous testing, proving its great empirical practicality.

“Language Models are Few-Shot Learners” - a paper published by OpenAI. The paper describes GRT-3, a deep-learning model for Natural Language Processing, with 175 Billion parameters- 100x more than the previous version, GPT-2. The model is pre-trained on half a trillion words and achieves SOTA performance on several NLP benchmarks without fine-tuning. There is no fine-tuning; For training, the researchers have used a combination of model parallelism within each matrix multiply and model parallelism.

Based on its paper, *GPT-3* is an autoregressive language model instead of a denoising autoencoder, like *BERT* introduced earlier. Accelerating the competency and ensuring great accuracy regardless of the magnitude of the model - faster, accurate for a bigger model. However, the cost of implementing GPT-3 is easily several ten million due to its significant immensity- considering that this model is adopted for its ‘practicality,’ it appears that only minimal companies will be able to access such a model.

Big Bird: Transformers for Longer Sequences, a paper published by Google research. This paper addresses the limitations of the full attention used by Transformer models (such as BERT, mainly due to quadratic dependency), introducing a sparse attention mechanism that can alleviate. Such limitation - that uses memory that scales linearly to the sequence length. There exist several approaches to attention - Random, Window, Global, and Sparse; in the paper, it is presented that eventually, Sparse attention enables the mechanism to attend to longer sequences, acting as the most effective transformer.

Even though these learning models such as BERT and GPT-3 are claimed to be practical with their innovative approach to their performance, there is reasonable doubt on just how well they will perform literature/linguistic tasks. Both LMs are to be trained by 'context' from a vast dataset of text; this, no need to say, significantly enhances their ability to excel in particular/designated tasks: Q&A, entity, or value recognition. However, the models might be recognizing the singular words in the context instead of understanding its 'Context' as they claim so. This indicates that these models might be speechless - only able to use a single word from word, but not in the form of a sentence in which the dynamics among the words, instead of singular meaning, determine the actual thesis of the sentence. Considering that being able to improve its linguistic insight through training is an essential aspect of a Natural Language Processing learning model, the presented models - BERT and GPT-3, might be losing their most valuable ability in exchange for their fast and vast performance on specific tasks.