Technology Review: Transformers and their impact on NLP

By: Yun Hui Xu

"Attention is All You Need" (*Vaswani, et al.*): splashing down back in 2017, this paper brought about a revolution in NLP and Deep Learning as a whole. While primarily used for NLP tasks, transformers have now been applied to everything with tremendous success, even now taking over the computer vision front (*Dosovitskiy, et al.*) where Convolutional Neural Networks have remained superior in the past. But even so, the most impressive applications of these Transformer models have been in Large Language Models, where the original transformer architecture has proven to be very impressive with regards to natural language processing tasks, holding the current SOTA in almost all Natural Language benchmarks (*GLUE Benchmark*). We can categorize all these transformer based models into one of three categories: Encoder-only models, used in BERT (*Devlin, et al.*), Decoder-only models, used in GPT models (*Brown, et al.*), and Seq2Seq models, used in T5 (*Raffel, et al.*).

The transformer model is the base of all of the following models, usually referred to as Encoder-Decoder models or also as Seq2Seq models, which use the full transformer architecture proposed in the Attention paper (*Vaswani, et al.*). These models are built of stacked layers of attention and with an encoder section built of attention blocks with feed-forward layers, all of which feed into a decoder section with more attention blocks and feed-forward layers. The attention layers in the encoder can access all tokens at any position, while the attention layers in the decoder can only access the tokens before the current position.

The T5 model, presented in a paper called Text-to-Text Transfer Transformer, is a seq2seq model from Google, trained using Masked Span Corruption. This objective is based on a generalization of two other pre-training objectives. The first of which is Masked Language Modelling, which corrupts a certain percent of tokens with a "mask token", and has the model fill in the blanks; the second of which is known as Causal Language Modelling, which is trained on predicting the next token at the end of a sequence. This generalization into Masked Span

Prediction allows spans of tokens to be masked and also allows for spans to be generated at the end of the token sequence input, encapsulating both objectives.

Encoder-only models, also known as auto-encoding models, use only the encoder portion of a transformer model. The characterizing factor of these models is that each stage in the attention layer can access all the tokens in the initial input. BERT specifically uses bi-directional attention, which means that the attention layers can look backward and forward to inform their decision on tokens. These types of models are trained via Masked Language Modelling, which masks out random words (corrupts) and then asks the model to fill in the blanks. BERT is best used for tasks that need to see and understand the full input sequence, such as sentiment analysis, NER, and question-answering tasks.

Decoder-only models, also known as auto-regressive models, use only the decoder portion of a transformer model. Compared to encoder-only models, these models limit the attention layers to only being able to access the tokens before the current position. These models are trained via an objective called Causal Language Modelling, which means they are trained to predict the next word in a sentence, much like autocomplete. The best use of decoder-only models is text generation, although we can also use few-shot learning to apply any of the encoder-only tasks that BERT can accomplish to GPT with no further training needed.

OpenAI has trained the GPT style family of models, of which the latest (and biggest) model called GPT-3 containing 175 billion parameters has been used to great success for all sorts of tasks such as text completion, storytelling, and even code generation with Codex, a fine-tuned version of GPT-3 on GitHub code, powering GitHub's newest Copilot code-complete feature. They have also taken the performance a level further using Instruction Fine-Tuning (*Ouyang, et al.*), an RL-based approach that incorporates user feedback to better train the model to follow instructions, which has huge benefits in benchmarks, even prompting a paper from Google with their take on the method, called FLAN (*Chung, et al.*).

All three types of models are simply variations of the original Transformer model that was originally proposed back in 2017, however, each of which has obtained state-of-the-art results in almost all NLP benchmarks, and continues to show promise after being scaled up and up with almost no limit in sight. The trend lately in NLP especially with transformer models has just been to scale as much as possible given computing constraints (*Hoffmann, et al.*), increasing parameter counts, getting more data, and training for longer all increase performance to new ceilings, all with the same original architecture with minimal changes. However that doesn't mean that the only path forward is scaling, recently people have been looking into re-evaluating the entire Transformer architecture, specifically in reducing its attention scheme's complexity from quadratic to linear (*Dao, et al.*), and also its pre-training schemes, with massive performance gains reported in a very recent paper (*Tay, et al.*), which means that even as impressive as these models currently are, there is still a lot of room for improvement. The landscape is ever evolving with papers being pushed out at a breakneck pace.

References

Brown, Tom B., et al. *Language Models Are Few-Shot Learners*. 2020. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2005.14165.

Chung, Hyung Won, et al. *Scaling Instruction-Finetuned Language Models*. 2022. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2210.11416.

Dao, Tri, et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2205.14135.

Devlin, Jacob, et al. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. 2018. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.1810.04805.

Dosovitskiy, Alexey, et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2020. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2010.11929.

GLUE Benchmark. https://gluebenchmark.com/. Accessed 6 Nov. 2022.

Hoffmann, Jordan, et al. *Training Compute-Optimal Large Language Models*. 2022. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2203.15556.

Ouyang, Long, et al. *Training Language Models to Follow Instructions with Human Feedback*. 2022. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2203.02155.

Raffel, Colin, et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2019. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.1910.10683.

Tay, Yi, et al. *UL2: Unifying Language Learning Paradigms*. 2022. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.2205.05131.

Vaswani, Ashish, et al. *Attention Is All You Need*. 2017. *DOI.org (Datacite)*, https://doi.org/10.48550/ARXIV.1706.03762.