Week 1: Tuesday: Attention is all you need

explained by The Illustrated Transformer

Supplemental: Coding self-attention

1. What are the input and output in “Attention is all you need”? How is the quality of the output evaluated? Be precise.

Ans: The input to the transformer model as presented in the paper is a sequence of symbol representations.(x1,x2,...,xn)

The output is also another sequence of tokens which could be translated text in the target language. Transformers can also be used for summarization of the input text as well.
The paper also mentions-”We are excited about the future of attention-based models and plan to apply them to other tasks. We plan to extend the Transformer to problems involving input and output modalities other than text and to investigate local, restricted attention mechanisms to efficiently handle large inputs and outputs such as images, audio and video.”

With respect to the quality of output, the paper mentions in its abstract-”Our 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. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.”

BLEU or Bilingual Evaluation Understudy measures the correspondence between machine generated translations and a set of reference translations. It was invented at IBM in 2001 and its value is in the range of 0 to 1.

2. What are the Query, Key and Value is self-attention? Give an intuitive description of what role each of them plays.

Ans: Query is the representation of a particular word for which we intend to find other words in the sequence that are related to it. We make use of the query of the word in focus and use the keys of the rest of the tokens in the sequence to find words related to it. The keys of every token are another representation of the token. Once the model knows how each word relates to each other and which words to focus on, it uses the values, another representation of each token in order to construct the context for the word in the query.

In simpler words, each word tries to find out which words are important to it based on similarity scores by using query(word in focus) and keys(rest of the words including the word in focus) and it uses the values of the important words to represent itself more contextually.

3. What is an "attention head"?
Ans: An attention head is a component of the transformer model that computes the attention mechanisms independently using the Q,K and V of each token. The transformer model as described in the paper makes use of multiple heads to capture different types of relationships in parallel. For example, the word "great" can be used in a positive manner and in a sarcastic manner as described in this youtube video. Consider the two sentences-"Jack is a great guy." and "Jack won't make it to work next week, great." Here, the word "great" is used in two different ways and this is exactly what multi-head attention intends to capture.

4. What are the advantages of the Transformer architecture over RNNs and LSTMs in processing sequences? Include in your answer: How does the Transformer model scale in time and memory compared to an RNN? Be precise.

Ans: The advantages are,
Parallelization: Transformers do not process data sequentially like RNNs/LSTMs and this reduces training time by a lot because of the ability to train parallelly.

Transformers compute attention to the entire sequence, so it can handle long range dependencies whereas RNNs/LSTMs may struggle with information loss over long sequences.

The transformers scale better in terms of computation time as it does not process data sequentially but they are memory intensive as they have to store the attention scores for every pair of words in the sequence.

5. The model presented here is an “encoder/decoder” architecture. What are examples of encoder-only and decoder-only architectures?

Ans: Tasks involving text classification like the one completed in HW0 is an example of an encoder-only application as only the input sequence is required. An example is BERT(Bidirectional Encoder Representations from Transformers)

Generative Pretrained Transformers(GPTs) are an example of decoder-only architectures used for tasks like text generation where the model generates output sequences based on some initial input.

6. What are the limitations or challenges associated with the Transformer model? What is being done to overcome them?

Ans: The main limitations associated with the transformer model is that it is memory intensive as it needs to store attention scores for all pairs of words in a sequence and they would require significant computational resources for training especially with very long sequences.

There have been some efforts to mitigate these obstacles like opting for different attention mechanisms like sparse attention mechanisms. A quick google search showed other points like knowledge distillation, model pruning, quantization, etc.

7. Why is Huggingface so popular?

Ans: Huggingface is popular due to its comprehensive and easy to use libraries that provide implementations of many transformer models. It also offers a collaborative platform, tools and pre-trained models that simplify the process of training and deploying the models in NLP. It also offers courses for people to learn or refresh their memory.

8. What questions do you have about the paper?

Ans: There are so many things I do not understand in this paper and I need to read a lot more outside this paper to understand what is going on. I also look forward to learning more about the way the output is evaluated because wikipedia itself dives deep into it and offers really technical explanations about different aspects of it.