Ikhlas Attarwala
05-08-19 / Week #6
Improving Efficiency with NNLMs / Paper 2

Attention Is All You Need

by Vaswani, Ashish et al.

This paper introduced Transformers, a network without recurrence that models dependencies using attention. RNN-based architectures were hard to parallelize, and had trouble learning these long-term dependencies. With the introduction of the Transformer, instead of using a single sweep of attention, it used multiple "heads" (multiple attention distributions & multiple outputs per single input). The Transformer used explicit position encodings so that the input embeddings positions can be utilized by attention. Attention is a manner of computing the relevance of some information based on keys and queries, such that the focus falls only on relevant information. Previously, when only a single attention was computed by weighing the sum of values, nuances within the input may not have been captured.

I think this was a revolutionary paper that introduced a model that covered the disadvantages of RNNs (difficult to learn long-range dependency, cannot parallelize within instances, hard to model languages) and CNNs (path length can be logarithmic for dilated convolutions). Transformers parallelize by replacing recurrence with attention and encoding symbol positions within sequences (seq2seq). Computation is also reduced for sequential tasks, where O(1) operations are performed to learn dependency between symbols (without regard for position distance within the sequence). The following link does a great job at creating a visualization to understand how the encoding/decoding steps within the transformer works for translating the sentence "I arrived at the…" from English to French:

https://cdn-images-1.medium.com/max/1600/1*sbfNVjf3yERRD9Rg4OLIBw.gif