Ikhlas Attarwala
06-05-19 / Week #10
Current Topics / Paper 1

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
by Devlin, Jacob et al.

This paper by Jacob Devlin et al. is quite possibly the most popular paper we'll be reading this quarter. BERT stands for Bidirectional Encoder Representations for Transformers, and unlike traditional models at the time that read the previous 'n' tokenized words, BERT took in both the previous and upcoming 'n' number of tokens for use in prediction. Further, it has been trained on next-sentence prediction tasks, whereas previously, pretraining in NLP was limited to word embeddings such as word2vec and Glove. Word embeddings weren't too powerful, and they were trained on shallow language modeling tasks. BERT also does well to take context into account because it trains a complex deep net to map vectors of words based on the tokens before and after.

Results were impressive, with 80.4% accuracy on the GLUE benchmark, 86.7% accuracy on MultiNLI, and 93.2% accuracy on the SQuAD question answering test. The empirical improvements are outstanding. Now a masked language model (bidirectional) is slower than left-to-right, but the improvements outweigh the loss. Additionally, it also outperformed the NER base model, along with SWAG by 25%+. I don't think BERT will be forgotten in years to come.