Ikhlas Attarwala
06-05-19 / Week #10
Current Topics / Paper 2

Deep Contextualized Word Representations
by Peters, Matthew et al.

Okay. That does it. At this point, I have to ask.. what is the fascination with Sesame Street? BERT? ERNIE?? ELMO??? Anyway… ELMO stands for 'Embeddings from Language Models'. It's one of the major breakthroughs in NLP history. It's widely known that embeddings *should* take context into account in order to assign meaning to it. The model proposed by the paper takes in an entire sentence/paragraph and assigns an embedding to an individual word. ELMO uses a multi-layered B-LSTM, extracts a hidden state of each layer for the input sequence of words, and then computes a weighted sum of those hidden states to obtain an embedding for each word, passing it to the task RNN. This is useful in sentiment analysis, question-answering tasks, and NER.

I believe after Józefowicz introduced the 2 B-LSTM layered model, and the authors of this paper adapted that into ELMO. The authors showed that just using the last layer wasn't as great as averaging the bidirectional layers, and that wasn't as accurate as learning the layer weights. They found that the first bidirectional layer was great at POS tagging, and the second for word sense disambiguation. This is why weighing the layers is better for different tasks. The results spoke for themselves; adding ELMO to existing NLP systems improved accuracy for each task (incl. SQuAD, SNLI, SRL, Coref, NER, Sentiment (5-class), etc.).