Ikhlas Attarwala
04-08-19 / Week #2
Word Embeddings / Paper 2

**Linguistic Regularities in Continuous Space Word Representations** by Mikolov, Tomas et al.

This paper was written in 2013 by Microsoft's Research group. They proposed that training a neural network language model implicitly provides learned word representations, this in turn captures clean, meaningful syntactic and semantic regularities. Furthermore, instead of working with discrete units, the idea of using a continuous space is that similar words are expected to have similar vectors; accommodating a model to a word should then carry over to similar words in that space. Mikolov et al. then create syntactic tests in the form of analogies (such as "*a* is to *b* as *c* is to ___") to evaluate performance over adjectives, common nouns and temporal-tenses of verbs. Their model was able to answer about 40% of questions correctly. Finally, they test semantic relation similarity, and evaluate the degree of a relation between words. This supposes that groups of word pairs have the same relation (such as "*clothing* is to *shirt* as *dish* is to *bowl*", measuring the degree of the same relation between *clothing:shirt* vs. *dish:bowl*).

Both tasks required a Vector Offset Method. This assumes relationships between words are presented as vector offsets. In the analogy mentioned above, they compute y as the difference between vectors b and a, and add in vector c, where y is the continuous space of the word expected to be the answer. If a direct answer is not present, a cosine similarity is computed to find the nearest embedding vector. Their method outperformed the last latest.