NLP: A History... and Future

1960s

The Term-vector model (aka VECTOR SPACE MODEL (VSM)) of information retrieval is developed using is an algebraic model used for information filtering, information retrieval, indexing and relevance ranking [1]

1986

Concept of RNNs (Recurrent Neural Networks) developed by David Rumelhart [2] giving rise to the idea of keeping track of arbitrary long-term dependencies in input sequences

1997

LSTM (Long Short-Term Memory)[3] developed. Overcame issues with RNNs, namely "vanishing" (tending toward 0 over time) or "exploding" (tending toward infinity) gradients after enough iterations of backpropagation

~2013

Shallow Neural Networks such as word2vec (2013)[4], which uses CBOW or skip-gram methods to predict words (targets or neighbors) based on sliding context windows. GloVe improves on w2v with cosine product between words representing number of times they co-occur, fastText is another similar method

~2014

Paragragh Vector aka Doc2vec developed [5] developed to exploit concatenation of words in embeddings vs averaging. Sent2vec expands on CBOW of word2vec by using shallow neural network, trained to predict a sentence using a sliding window instead of a word

~2014 -2019

Sequence2Sequence developed and improved upon[6]. The basic idea being to use LSTM as a way to remember past words from a much wider range. Keys are use of encoder - decoder for creation and remembering of vectors in hidden layer. Also, method of attention was developed to aide in this remembering

2017

ELMo developed by AllenNLP [7] to be able to create contextualized word embeddings using bi-directional LSTMs. Still extremely popular and was trained on enormous datasets like wikipedia and can enhance other NLP projects. Self-attention technique developed later for finding associations from options inside sequences Self Attention eventually became the transformer [8]

2018

Universal Language Model Fine-Tuning (ULMFiT) [9] built on this by adding ability to optimize any neural-network-based language model for any task. Key features are gradual unfreezing for fine tuning progressively more parameters and discriminative fine tuning.

2018

Bidirectional Encoder Representations from Transformers (BERT) [10] and GPT [11] Generative pre-trained Transformer both build on the encoder decoding concepts. BERT modified transformer architecture by getting rid of decoder and relying on "masked" words which need to later be predictied accurately

Terms and Sources

Key Terms to Learn/Remember:

1. Recurrent Neural Networks
2. Long Short-Term Memory Networks
3. Bi-directional LSTMs
4. Convolutional Neural Networks
5. Recursion
6. Backpropagation
7. Hidden Layers
8. Encoder
9. Decoder
10. Attention
11. Self-Attention
12. the transformer

Sources:

1. https://www.researchgate.net/publication/265913721_A_Comparison_of_Information_Retrieval_Models
2. Williams, Ronald J.; Hinton, Geoffrey E.; Rumelhart, David E. (October 1986). "Learning representations by back-propagating errors". Nature. 323 (6088): 533–536. Bibcode:1986Natur.323..533R. doi:10.1038/323533a0. ISSN 1476-4687. S2CID 205001834.
3. http://www.bioinf.jku.at/publications/older/2604.pdf
4. https://arxiv.org/abs/1301.3781
5. https://arxiv.org/pdf/1405.4053v2.pdf
6. https://arxiv.org/abs/1409.3215
7. https://allenai.org/allennlp/software/elmo
8. https://arxiv.org/abs/1706.03762
9. https://arxiv.org/abs/1801.06146
10. https://arxiv.org/abs/1810.04805
11. https://paperswithcode.com/method/gpt