References

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. *CoRR*, *abs/1409.0473*. Retrieved Nov 21, 2019, from https://arxiv.org/pdf/1409.0473.pdf

> This paper was foundational to attention mechanisms. Up to this point, RNNs for machine translation relied on an encoding vector of fixed length to represent the entire sentence. This time, the contributors allowed the decoder to rely on encoded representations of each word in the sentence, allowing it to learn the (sometimes long range) dependencies between individual words. This paper coined the phrase "attention mechanism", defined it as a learned mechanism (instead of a latent variable), and demonstrated that models with attention achieved higher performance on translations of long sentences. They produced the first neural machine translation model with accuracy comparable to that of statistical models.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend, and spell. *CoRR*, *abs/1508.01211*. Retrieved Nov 21, 2019, from http://arxiv.org/abs/1508.01211

> The same year the original attention paper was published, this group performed end-to-end speech transcription with an RNN+attention model. They used a time-dimension reducing BLSTM network to convert raw audio into a compressed representation, and then an RNN with attention to decode that representation into character sequences. This work demonstrated that attention could be used to allow for greater dependence between outputs, because without those dependencies a character-level output would not work. This paper does not contribute in terms of model architecture, but does demonstrate attention's application for character-level decoding.

Cho, K., Courville, A. C., & Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. *CoRR*, *abs/1507.01053*. Retrieved Nov 21, 2019, from http://arxiv.org/abs/1507.01053

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *CoRR*, *abs/1703.03130*. Retrieved Nov 21, 2019, from http://arxiv.org/abs/1703.03130

Parikh, A. P., Täckström, O., Das, D., & Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *CoRR*, *abs/1606.01933*. Retrieved Nov 21, 2019, from http://arxiv.org/abs/1606.01933

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. Retrieved Nov 19, 2019, from https://papers.nips.cc/paper/7181-attention-is-all-you-need

> This work did away with the recurrent layers used with attention up until this point, simply training the attention mechanism to perform the task of text translation. Recurrent models have to be run sequentially, so removing them allows training and evaluation to run faster. They achieved state-of-the-art performance on English-to-German and English-to-French, while being able to train the models in significantly less time. In this work the researchers were careful to analyze what each component of the model contributed, allowing them to draw interesting conclusions. Another

contribution is the definition of attention as a learned function of queries, keys, and values. The source of each of these inputs determines the kind of attention used.