Review: Neural Machine Translation by Jointly Learning to Align and Translate
By Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio and published at ICLR 2015

This landmark paper first laid the foundation of the concept of attention in the domain of machine translation that is being used in various domains nowadays. Encoder-Decoder consisting of LSTM are being used to resolve exploding/vanishing gradient problems in machine translation but they are able to translate only short sentences of length 20-30 words but the accuracy drops as length of sentences increases which is demonstrated through a performance metric, BLEU score. Authors compared the proposed attention model (RNNSearch) with basic seq-seq models (RNNenc) consisting of 30 and 50 word long sentences and showed that BLEU score (an indicator of correct translation) decreases for seq-seq model as length of input sentence increases. This is due to the fact that basic encoder-decoder architecture tries to encode the sentence in a single fixed-length vector (encoder RNNs final hidden state) which causes information bottleneck but it can be the case that intermediate hidden states also contain relevant information.

To resolve this information bottleneck, the authors extended the encoder-decoder model to fix the sequence to sequence model and the ability to translate longer sentences by jointly learning to align and translate by soft-searching for parts of a sentence relevant to predict the target word. To generate a target word, the model soft-searches for a set of positions in the input sentence where the relevant information is concentrated in the form of a context vector which is then used to predict the target word with the help of previously generated target words. RNNSearch model consists of a bidirectional RNN encoder where conditional probability for each target word y^i is conditioned on a distinct context vector c^i. The c^i is a weighted sum of the annotation sequence h^i which contains the information of the whole input sentence with a strong focus on the surrounding words and weight alpha. The alpha is dependent on associate energy e which is the score that reflects the importance of annotation h with previous hidden state to decide next state and the target word.

The authors evaluated the proposed approach by using bilingual parallel corpora of ACL WMT '14 English-French translation consisting of 348M words. For activation function, gated hidden unit such as LSTM is used while for alignment, single layer multilayer perceptron with tanh activation is used to reduce the computations. Mini-batch stochastic gradient descent (SGD) with Adadelta is used to train the model with 80 sentences in each batch. One of the good things I liked about the paper was that the authors had separated the mathematics behind the model at the end in the separate section, because I get intimidated by all the math and don't read the paper most of the time so it was definitely easy for me to read this paper. Secondly, the authors showed the improvements made by the alignment through a visualization of attention weights and how different positions in source and target were correctly aligned and also illustrated the translated sentence.

One of the major drawbacks authors faced was handling of unknown or rare words. It can be improved by using word embedding of larger dimensions. Secondly, English and French being largely monotonic languages without much difference in position change, experiments need to be performed for various languages. To conclude, the model correctly aligns each target word with the relevant words and assigns good alignment in the source sentence and is more robust to length of the sentence.