Jenish Maharjan **Week 4**

Paper: *Neural Machine Translation by Jointly Learning to Align and Translate*

Summary:

Bahdanau et al., focus on machine translation using a neural network and extend the RNN encoder-decoder framework using their novel architecture in this paper. The RNN (specifically LSTM) encoders and decoders have performed pretty well on the task of machine translation but they are limited in their ability to track long-term dependencies. As the length of sentences increased, the ability of these encoder-decoder networks to translate decreased. The authors identify the encoding of the input in a single fixed-length vector as the reason for this limitation. The paper addresses the single node bottleneck problem in two ways – using a bidirectional LSTM for input and by introducing an alignment model.

The goal in the task of translation is “to find a target sentence y that maximized the conditional probability of y given a source sentence x”. The translation model first learns the conditional distribution by fitting the model to maximize the conditional probability of sentence pairs, a translation can be generated for a source sentence that maximizes the conditional probability. The initial RNN encoder-decoder model, proposed by Cho et al., the input sentence is read by an encoder into a vector. The decoder is trained to predict the next word given the context vector and all the previously predicted words.

The new model architecture proposed in the paper has bidirectional RNN as an encoder which consists of forward and backward RNNs. The forward RNN reads the input sequence as it is and the backward RNN reads the sequence in the reverse order. Thus, the forward and the backward RNNs calculate different hidden states that are concatenated to obtain an annotation for each word. The obtained annotation provides summaries of both preceding and following words and is used by the decoder to compute the context vector. The decoder consists of the alignment model and another LSTM. The alignment model that scores how well the inputs and outputs are two positions match. The alignment model is jointly trained with all the other components of the system.

The authors trained their models on sentences of varying lengths, both with and without the attention mechanism. Output sentences were generated with beam search. The idea behind beam search is that greedy choice of the next word generates poor quality sentences and so it is worth considering short sequences instead of words. The authors evaluate their models on the task of English to French translation and compare it to the basic RNN encoder-decoder proposed by Cho et al. As expected, the quality of long sentence translations was improved compared to the baseline model. They have also presented the weights generated by the alignment model at every input and output word location which shows that their approach provides an intuitive way to check the “alignment” between words in the source sentence and the generated translation. While outperforming the conventional encoder-decoder model significantly regardless of the length of the sentences, the new model’s translation performance was also comparable to the existing phrase-based statistical machine translation.

Strengths:

- The paper presents both quantitative results and qualitative analysis of the model which makes their study more reliable.
- Use of figures in the results sections helps in the understanding the results more intuitively.
- Overall, the organization of the content in the paper is pretty logical and adds to its readability.

Weaknesses:

- The section explaining the alignment model and the network architecture could have used more diagrams to give a better perspective of the model.
- The conclusion could have been made more concise – it mostly repeats the content presented in the previous sections.

Confusions:

- Does alignment of words refer to their order with respect to each other?
- How does the attention mechanism in the decoder work?

Discussion Questions:

- A discussion on the math presented in the paper would be really helpful.
- How would the preprocessing of the dataset like lower or stemming affect the task of translation?
- Is BLEU score the standard measure for machine translation or are there other such measures as well?