Jenish Maharjan **Week 5**

Paper: *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*

Summary:

In this paper, the authors describe Google's Neural Machine Translation (GNMT) system focusing on the implementation details and explanations about the techniques they have implemented to make it a robust, accurate and a high performing system. The system they have introduced attempts to address many of such issues present in the previous neural machine translation (NMT) systems. By implementing wordpiece modeling, model and data parallelism, model quantization and many other details like length-normalization, coverage penalties and so on, their model is able to reduce approximately 60% reduction in translation errors on multiple language pairs.

The architecture of the model has an encoder network that transforms an input sentence into a list of vectors, a decoder network which produces a word at a time from the list of vectors until a special symbol representing the end of sentence is produced and an attention module that connects the encoder and decoder and also allows the decoder to focus on different parts of the input sentence during decoding. The layer at the bottom of the encoder is bi-directional allowing it to gather information from both right to left and left to right while the other seven layers of the encoder are unidirectional. All layers of the encoder are LSTM layers. The decoder is also made up of 8 LSTM layers, all unidirectional. The model has residual connections starting from the third layer which allows training deeper networks by improving the gradient flow. The model is designed with parallelism in mind which allows for the two directional components of the bottom encoder layer to run on different GPUs. While model and data parallelism improve the training speed, model parallelism does constrain the model architecture. One of the main constraints of model parallelism is that having bi-directional LSTM layers for all encoder layers becomes very computational demanding, thus unaffordable. The authors have also described two methods of segmentation – Wordpiece Model and Mixed Word/Criteria Model – out of which the wordpiece model was the most successful.

Attributing to "undesirable properties" of the BLEU score, the authors used their own scoring system "GLEU score" whose range is always between 0 and 1. According to the authors, the GLEU score correlates well with the BLEU scores on a corpus level and thus also useful for the "per sentence reward objective". For training, the model is first trained using the maximum likelihood objective until convergence and then refined using a mixed maximum likelihood and expected reward objective. This continues until the BLEU score on a development set stops improving. The inference task during translation is computationally intensive thus low latency translation becomes difficult and high-volume deployment becomes expensive. To address this, the authors implement quantized inference using reduced precision. Although this adds more constraints to the model, the authors still used it because it did not affect the model convergence of the quality of the model once it converged. Like the previous NMT systems, the authors use beam search during decoding to find the sequence of words that maximizes the score for a given trained model. The authors implement this using a refined pure max-probability based beam search algorithm – the refinements being a coverage penalty and length normalization.

The GNMT system is trained and tested on English-to-French and English-to-German datasets benchmarked with word-based, character-based and wordpiece-based vocabularies. The results they had

from their experiments showed that the models had improved accuracy after fine-tuning with RL and model ensembling. They compare the model ensemble performances with human evaluation which emphasizes the accuracy and robustness of their system. They also experiment on production data for multiple languages which is a great feat for the field of Neural Machine Translation.

Strengths:

- The paper does a great job of explaining the novel contributions that their study makes. The content of the paper is well organized.
- The addition of human evaluation of their models adds confidence to their system.

Weaknesses:

- Some sections are verbose and thus difficult to understand. The language could've been edited to make it more readable.

Confusions:

- The concept of a quantized model is a little confusing.
- Could segmentation be omitted if we had machines with infinite computational capabilities?

Discussion Questions:

- Since the paper on their model is openly available, how do they maintain proprietary rights to systems like google translate?
- The production data they use is internal to their organization. So how do we know that their data and the tests they've run is valid?
- Their system reduces translation error by 60%. How much can we attribute this to the resources the team had access to at Google?