Rahul Thapa

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

Summary:

In this paper, the authors present GNMT, Google's Neural Machine Translation system, which attempts to address many of the issues of the previous Neural Machine Translation (NMT) model such as computational expense and lack of robustness. Their model consists of a deep LSTM network with 8 encoder and 8 decoder layers using residual connections as well as attention connections from the decoder network to the encoder.

The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text. Another advantage of NMT is that it sidesteps many brittle design choices in traditional phrase-based machine translation. However, in practice NMT system used to be worse in accuracy than phrase-based translation systems, especially when training on very large-scale datasets as used for the very best publicly available translation systems. The three inherent weaknesses of such models, that authors talk about in this paper are: its slower training and inference speed, ineffectiveness in dealing with rare words, and sometimes failure to translate all words in the source sentence. To deal with the slower training, the authors use LSTM RNNs with residual connections between layers and connect the attention from the bottom layer of the decoder network to the top layer of the encoder network. This allows for parallelism and hence increases the training time. To improve inference time, they employ low-precision arithmetic for inference, which is further accelerated by special hardware (Google's Tensor Processing Unit (TPU)). To effectively deal with the rare words, they used sub-word units, also referred to as wordpieces in the paper, for the inputs and outputs in their system. Their beam search technique includes a length normalization procedure to deal effectively with the problem of comparing hypotheses of different length during decoding, and coverage penalty to encourage the model to translate all of the provided input.

The authors present the experimental result of their model on two publicly available corpora: WMT' 14 English-to-French and English-to-German. On these datasets, they benchmarked their GNMT model with word-based, character-based, and wordpiece-based vocabularies. They also present the improved accuracy of their models after fine-tuning with RL and model ensembling. In addition, they also tested their GNMT model on Google's translation production corpora, which is a very large dataset. They compared the accuracy of their model against human accuracy and the best Phrase-Based Machine Translation (PBMT). They used the standard BLEU score metric. To fully capture the quality of a translation, they also carried out side-by-side (SxS) evaluations where human raters evaluate and compare the quality of two translations.

Their best model, which they report as WMK-32k, achieves a BLEU score of 38.95 on English-to-French translation. On English-to-German, their best model, WMK-32k achieves a BLUE score of 24.61. This score has averaged a score of 8 models they trained. They also tested on RL-refined models. The results of RL fine-tuning on the best En-Fr and En-De models achieved a BLEU score of 39.92 and 24.60 respectively. They noted that on WMT En-Fr, model refinement improves BLEU score however, on En-De, RL-refinement slightly degrades the performance. They also ran human evaluation of these translations and they found out that even

though RL refinement can achieve better BLEU scores, it barely improves the human impression of the translation quality. On their production data, they did a three-way side-by-side comparison. The three sides are from phrase-based machine translation (PBMT), GNMT, and bilingual humans. The GNMT model that they used in this procedure is wordpiece mode without ensembling and dropout. As they expected, the GNMT system performs better than the PBMT system. They also found out that in some cases, human and GNMT translations are nearly indistinguishable.

Strengths:

- The paper is very easy to follow. The authors explain in detail all the critical parts of their architecture.
- The paper is rich in figures and tables showing the critical part of their architecture and the results which give a clear picture of their model.
- They use various methods to make sure that their model is performing best. They compared their result with a phrase-based model and well as human translation.

Weaknesses:

- The paper could have made a little shorter. Some information is redundant throughout the paper. For example, the abstract can be shortened a bit since the information that they explained in detail there comes immediately in the Introduction section.

Confusions:

- I was not very clear about the different forms of parallelism they explain the paper i.e. model parallelism and data parallelism. Also, how does align the bottom decoder output to the top encoder output maximizes parallelism?
- What is quantized inference and how does it help reduce the cost of inference?
- What are the hypotheses that they talk about in the Decoder section and how is it being used in their model?

Discussions:

- How does a deep-stacked LSTM network help to increase the efficiency of NMT? What is it necessary to use residual connection if we want to go deeper?
- Why does wordpiece model perform better than other models?
- Where exactly in the training phase is the RL refining involved and how does it help tune the model?