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

Nirajan Koirala

CSC 9010_001

Summary

Neural machine translation is a newly emerging approach to machine translation. In the given paper, authors have proposed a new neural machine translation method that allows a model to automatically search for parts of a source sentence that are relevant to predicting a target word. This method as opposed to the previous methods doesn't use a fixed-length vector which acted as a bottleneck in the performance of the basic encoder-decoder architecture. Authors have introduced an extension to the existing encoder-decoder model which learns to align and translate jointly. Using this method, the model encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation. The improvement on translation of their approach is more apparent on longer sentences and they are able to achieve a translation performance close to the conventional phrase-based system.

In section 2, authors provide a background on neural machine translation and its methods. In a neural machine translation, a model is parameterized to maximize the conditional probability of sentence pairs using a parallel training corpus. After learning the conditional distribution, the model can use a source sentence to translate by searching for a sentence that maximizes the conditional probability. A brief discussion of RNN Encoder-Decoder is done in the subsections where appropriate equations for the input and output vectors is provided.

The discussion of their novel architecture is done in section 3 and this architecture consists of a bidirectional RNN as an encoder and a decoder that emulates searching through a source sentence during decoding a translation. Unlike the existing encoder-decoder approach, in this model the probability is conditioned on a distinct context vector for each target word. This context vector depends on a sequence of annotations which further depends upon the alignment model and this model directly computes a soft alignment. This allows gradient of the cost function to be backpropagated through which allows to train the alignment model as well as the whole translation model jointly. Using this method, decoder is provided with an attention mechanism and it relives the encoder from having to encode all the information in the source sentence into a fixed length vector. Furthermore they have proposed to use a bidirectional RNN in which annotation of each word summarizes the preceding as well as the following word.

Evaluation is performed on the task of English-to-French translation. A description of the dataset WMT ' 14 is provided and some pre-processing methods like tokenization is applied before training. Two models are used for comparison. The first one is RNN Encoder-Decoder and the second one is their proposed model RNNsearch. Both models have same number of hidden units (1000) and both of them use a multi-layer network with a single maxout hidden layer to compute the conditional probability of each target word. Minibatch SGD is used for training the models and after they are

trained they've used beam search to find a translation that approximately maximizes the conditional probability.

Authors later discuss about the results obtained from their comparison of two models. In all cases of translation, it is found that their proposed RNNsearch outperforms the conventional RNNencdec and also the performance of their model is equal to the phrase-based translation system when only the sentences consisting of known words are considered. While translating longer sentences, their models perform way more accurately and are very robust as well. Even their 30-word model outperforms the 50-word model of RNNencdec. Authors also take into account the alignment of words in the generated translation and the soft-alignment is found to be way stronger than the hard-alignment. Since soft-alignment lets the model look at the preceding as well as the following word, it is able to correctly translate the sentences by correct alignments of words in the target language. They further discuss about the alignment learning done by other people and how their method is superior as it requires computing the annotation weight of every word in the source sentence for each word in the translation.

Finally, they conclude by mentioning how their novel architecture addresses the issue of fixed-length vector during translation and lets the model focus only on information relevant to the generation of the next target word and achieve remarkable results which is comparable to the state-of-art results.

Strengths

- Paper is very application focused and practical and also the evaluation methods used are comprehensive.
- Ample amount of examples and figures are provided which helps in the realizing the superiority of their model.

Weaknesses

- The conclusion section could have been made more concise since it almost repeats the architecture model in the 2nd paragraph which was described fully in previous sections.

Points of Confusion

- I am confused about the workings of annotations and alignment models and their correlation.
- How beam search can be used to approximately maximize the conditional probability?

Discussion Questions

- The authors have trained the models with sentences of lengths upto 50 words and the model is still performing very well. Would the performance of the model degrade if we increase the lengths by large margin, say 100?

- How does BLEU performance metric work and are there any other similar methods for judging the performance of these kind of networks?
- Would these kind of networks find semantically similar translations for sentences which are intended as a pun?