Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Authors: Kyunghyun Cho, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio

Introduction

The authors add to the Encoder-Decoder paradigm in this paper, introducing the RNN Encoder-Decoder, which consists of two RNNs as an encoder-decoder pair. RNN is used to encode a sequence of symbols into a fixed vector representation, and, also to decode a representation into a sequence of symbols. The authors analyze this model to realize that the model preserves the semantic and syntactic structure of the phrase by learning the continuous space representations of a phrase. Recurrent neural networks can learn a probability distribution over a sequence by being trained to predict the next symbol in the sequence. One of the important aspects of this architecture is that the RNN encoder can learn the fixed vector representation from input sequences of *variable* length. While the RNN decoder learns the next word in the sequence, the RNN Encoder-Decoder are jointly trained to maximize the conditional log-likelihood. While the model can be trained to generate the target sequence, the authors use the probability to score a given input-output pair.

The authors also introduce a new hidden unit motivated from the LSTM unit which is much simplet to compute and implement. This hidden unit has a *reset gate* and an *update gate*, wherein the hidden unit is forced to ignore the previous hidden state if the reset gate is close to 0, and reset with the current input only, thus allowing the hidden state to drop any information that is irrelevant in the future. The amount of information carried from the previous hidden state is decided by the update gate. The authors conclude that with each hidden unit having separate reset and update gates, each hidden unit with more active reset gate will learn to capture short-term dependencies better, and those with more active update gate will capture the long-term dependencies better.

Theoretically, statistical machine translation systems are based on learning a translation model and a language model. In practice however, most SMT systems model the log-linear model which finds the translation of a target sentence given a source sentence. The authors use the RNN Encoder-Decoder on a table of phrase-pairs to score them, which do not learn the rank of the phrase-pair just based on the number of occurences, but ensure learning linguistic regularities by learning a a manifold of plausible translations. When used for scoring in machine translation systems, the maximum phrase length was generally chosen to be small, which in the model proposed by the authors is variable. Also, this model naturally distinguishes between sequences which have the same words but in a different order.

Experimentation & Results

The authors evaluated their approach on the English/French translation task of the WMT'14 workshop. This task includes the Europarl corpora, news commentary, UN and two other crawled corpora. The French language model was trained on 712M words. The authors acknowledged that

the concatenation of all the data doesn't lead to an optimal performance, and that one should concentrate on relevent subsets of data for a particular task. While training, the source and target vocabulary was limited to most frequent 15,000 words for both English and French, which covers approximately 93% of the dataset.

The RNN Encoder-Decoder used for the experimentation had 1000 hidden units with the update and reset gates. An embedding of 100 dimensions was learned for each word in the dictionary. The authors also compare the computational efficiency of the scoring method in the RNN Encoder-Decoder with the CSLM technique of phrase scoring. In Fig.2, the performance of the translation models on long, frequent source phrases and long, rare source phrases are tracked. To quantitatively track the performance of the model, four different combinations are analysed, the BLEU scores of which are recorder in Table 1. The authors also perform a qualitative analysis on the improvement of performance due to the RNN Encoder-Decoder, and focus on those phrase pairs whose source phrase is long and frequent. The RNN Encoder–Decoder captures both semantic and syntactic structures of the phrases, as can be seen from the clustering of semantically similar words in Fig.4.

Strengths & Weaknesses

The RNN Encoder-Decoder is one of the first neural network architecture which learns the mapping from a sequence of an arbitrary length to another sequence of arbitrary length. Since the RNN Encoder-Decoder is orthogonal to other traditional neural network approaches, it can further be improved by incorporating the neural net language model into it's design. One of the weaknesses with sequential nature of RNNs is that it precludes parallelization within training examples, which becomes critical at longer sequence lengths. [1]

Further Work

In order to counter the difficulty of parallelization while training with RNNs, attention models were introduced, which are based on the Transformer architecture. These models have shown to perform better and faster than the RNN Encoder-Decoder design, by gathering information about the relevant context of a given word, and encoding that context into it's vector representation.

References

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.