Sequence to Sequence Learning with Neural Networks

Authors: Ilya Sutskever, Oriol Vinyals, Quoc V. Le
Publisher: Conference on Neural Information Processing Systems(NIPS)
Year: 2014

Keywords: Encoder-Decoder, SMT, RNN, LSTM

1. **Abstract:**

"Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Our main result is that on an English to French translation task from the WMT'14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8 on the entire test set, where the LSTM's BLEU score was penalized on out-of-vocabulary words. Additionally, the LSTM did not have difficulty on long sentences. For comparison, a phrase-based SMT system achieves a BLEU score of 33.3 on the same dataset. When we used the LSTM to rerank the 1000 hypotheses produced by the aforementioned SMT system, its BLEU score increases to 36.5, which is close to the previous best result on this task. The LSTM also learned sensible phrase and sentence representations that are sensitive to word order and are relatively invariant to the active and the passive voice. Finally, we found that reversing the order of the words in all source sentences (but not target sentences) improved the LSTM's performance markedly, because doing so introduced many short term dependencies between the source and the target sentence which made the optimization problem easier."

2. **Introduction:**
   - Deep neural networks with NLP had been very successful in the difficult tasks, which had been failed in sequence to sequence learning. The sequence to sequence mapping had been used for the machine translation purposes.
   - So, to overcome that obstacle the authors adapted the multi layered LSTM for the accomplishment of the task.
   - The multi layered LSTM takes the variable input sequence, converts into a fixed vector and output the target sequence just as in encoder-decoder.
   - Moreover the technique of reversing had been proposed to introduce the short term dependencies in LSTM which would result in the better optimization.

- The work had been advantageous for the increase in the effectiveness of the translation of sentences into different languages.
-

3. **Related Works:**
    - Most of the works in translations had been done by RNNs and Feed forward Neural networks such as NNLMs.
    - These approaches had been reliable so far but needed the fixed input or atleast to limit the maximum input of the sentences.
    - One of the most relevant works had been done by Kalchbrenner et al., who first mapped the input to output sentences.
    - Further the researches also had been done on SMTs which had failed in translating the long sentence inputs.
    - However, these approaches had not achieved the BLEU score as much as that the approach proposed by this paper.

4. **Content and Approach:**
    - The problem with DNNs approach to the machine translation is the fixed vector input. Hence, that is not suitable for the language translations.
    - For that purpose, the LSTM architectures would replace the approach to suit to the variable inputs and another layer of LSTM is used for deriving the target sentence.
    - The other approach they have introduced in the paper is to input the sentence in reverse order which had increased the efficiency of the translation resulted in the relative increase of 3.2 BLEU score.
    - This is explained by the authors that it had been achieved because that the initial words of the input sequence of the LSTM and the first words in the target sequence had been nearer when reversed.
    - The LSTM also assists in picking the sentences which are nearer and which are not similar will be placed at far.
    - LSTM will map the output sequence of length T' to input sequence of length T, where T and T' may not be equal.
    - The authors claimed that their approach had been different in three ways than to implement normal LSTM to inputting sequence. The distinct implementations are:
        - They have used different LSTMs for the input sequence and for the output sequence.
        - The deep LSTMs are considered which is effective than the shallow LSTMs. So,they implemented four LSTM layers
        - The input sequence had been reversed, which also extended the better results.

- The experiments were done on the WMT'14 data set, which is a English to french dataset. The dataset consists of around 650M of words in both English and French with 12M sentences.
- The training had been done by taking the maximum log probability of translation given a input sentence. This training had been done with deep LSTMs .
- After the training, the translation had been done by taking the same maximum probability of translation given sentence.
- The LSTM code had been implemented in C++ on an individual GPU, which is very slow. So, they have created parallel training on four GPUs with four layers of LSTM on each GPU. This even took 10 days of training period.
- To represent the distance between the sentences, they had to decrease the vector dimension to two dimensions, which they had done with PCA.

5. Results:

- The quality of the approaches had been evaluated by comparing their BLEU scores.
- Through *multi-bleu.pl* which computes the BLEU scores, had resulted in 37.0 for this approach which had been greater than the previous best 35.8.
- The LSTMs had produced a better result when the input sequence had been reversed with 30.6 than when they are input normally.
- Even on the longer sentences Machine Translation, the reversing approach had been beneficial.
- From the graph which had been drawn in between the BLEU score and the sorting of test sentences in article, we can draw out that the LSTM had performed better irrelevant of sorting i.e., with their length and the frequency rank.

6. Conclusion and Future works:

- The accomplishment of deep LSTMs with large networks with no assumption prior to the input on length or semantics had outperformed on the huge Machine Translation task with a classical SMT systems, had been exceptional.
- Moreover, an insignificant technique of reversing the input sequence had also resulted in considerable improvements in the outputs.
- In addition to that, the ability to translate long sentences with LSTMs had been noteworthy with the approach of reversed sentences to the previous methods, however the translation is not accurate.
- Further works can be done on training the RNNs with the same approach and comparing the results. The authors had been optimistic to train on the reversed sentences with RNNs.
- Henceforth, this approach can be used in many of the future sequence to sequence challenges in different applications.