Summary: Sequence to Sequence Learning with Neural Networks (2014 NIPS)

Jooyong Song

Department of Electronics & Computer Engineering, Hongik University

1 Introduction

This paper introduces a new method that uses the Long Short-Term Memory(LSTM) architecture to solve sequence-to-sequence problems that traditional deep neural networks(DNNs) struggled with. The new model combines two LSTM structures: the first one reads the input sequence, timestep by timestep, to generate a large fixed-dimensional vector representation, and the second one extracts the output sequence from the vector.(fig.1) This combined structure, now named the encoder and decoder, stands out as an optimal choice compared to other attempts, particularly for tasks that involve handling time lags between the inputs and their corresponding outputs.

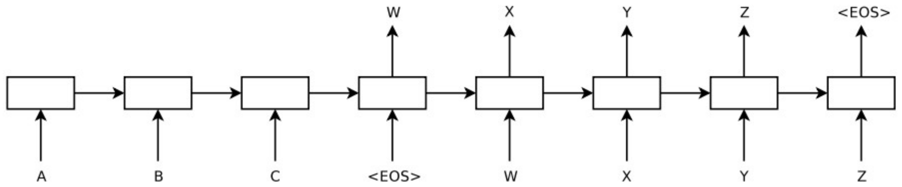


Figure 1: he model reads an input sentence “ABC” and generates “WXYZ” as the output sentence.

2 The model

LSTM is chosen for this model because the RNN is vulnerable to unbalanced and complicated sequences where input and output sequences have different lengths. The actual model structure has three distinct characteristics. First, it consists of two separate LSTMs—one for the input sequence and another for the output sequence. Second, each LSTM is designed as a deep network with four layers. Third, to achieve higher performance, the words of the input sentence are placed in reverse order.

3 Experiments

We translated WMT’14 English to French MT task without using a reference SMT system. We then experimented with LSTM-based reranking of SMT baseline’s n-best outputs and compared the results.

3.1 Decoding and Rescoring

In our experiments, we focused on training large and deep LSTM for many sentence pairs. The training was conducted to maximize log probability using the following formula. Once training is complete, we produce outcomes by finding the most likely translation using a simple left-to-right beam search decoder according to the LSTM.

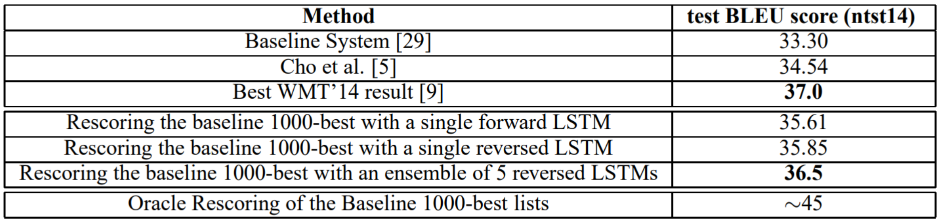
$$1/|\mathcal{S}| \sum_{(T,X)\in S} \log p(T|S) \qquad , \qquad \hat{T} = \operatorname*{argmax}_{T}\; p(T|S)$$

The beam search decoder keeps track of partial hypotheses, where each partial hypothesis is a beginning segment of a potential translation. The size of the hypothesis increases until <EOS> token is added to expand beam's hypotheses for each timestep. So we retain only the top B most probable hypotheses based on the model's log probability.

3.2 Reversing the Source Sentences

When the order of input sentences is reversed during the actual learning and testing process, it demonstrates improved accuracy. This phenomenon arises from the significant correlation among words positioned towards the beginning in a typical language system. This high performance is believed to stem from the fact that, with reversed input sentence order, the words at the beginning are more effectively captured in the context vector, enhancing their influence.

3.3 Experimental Results



Method	**test BLEU score (ntst14)**
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	∼45

Figure 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

While the method introduced in this paper does not surpass the performance of the state-of-the-art (SOTA) model, it serves as a testament to the potential of neural network learning-based machine translation.

4 Related work

Researchers have explored applying neural networks like RNNLM and NNLM to machine translation by rescoring n-best lists of a strong MT baseline, leading to improved translation quality. Recent work involves integrating source language information into NNLM. Examples include combining NNLM with topic models for better rescoring and incorporating NNLM into MT decoders for significant enhancements. Similar approaches involve mapping sentences to vectors and back, with attention mechanisms addressing challenges. End-to-end training models focus on mapping inputs/outputs to similar points in space but may require extra steps for translation.

5 Conclusion

When compared to the existing statistical machine translation by deeply stacking LSTM, it was found through experiments that better performance came out. Reordering the input words improved performance more than using the same existing dataset. In that thesis, achievements include: Using LSTM, a context vector was first generated through an encoder, and the context vector passed through an LSTM-based decoder to produce a translation result. In fact, during the training and testing process, we found that changing the order of words in the input sentence improved performance.

References

Sequence to Sequence Learning with Neural Networks (2014 NIPS)