DLLAB WEEK 11 Pre-Report : Sequence to Sequence

Hyunsoo Cha[1], Kwanghoon Sohn[2]
[1]Yonsei University. [2]Yonsei University.

1 Introduction

DNN Deep Neural Network must be a powerful machine learning model, but there are several challenges. There is a disadvantage that DNNs should be constructed only with input and target encoded in vectors of fixed dimensions. Speech recognition and machine translation are sequence problems. Similarly, question answering refers to mapping from a corpus representing a question to a corpus representing a response.

LSTM Researchers present applications using LSTM as a way to solve sequence problems. An input sequence is read by one LSTM and 1 time step is read to configure a large fixed-dimensional vector presentation. In the next LSTM, output sequence is extracted from the vector. The second LSTM can be described as a recurrent neural network language model. For a long range of time-dependent data, LSTM successfully learns a time lag that is considered between input and corresponding output. The researchers map the input sentence to the entire vector first in a structure similar to Kal's paper[2].

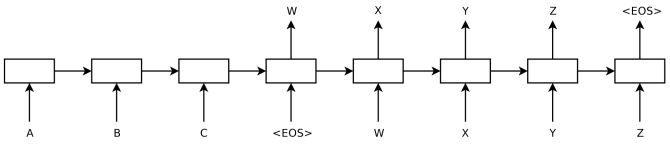

Figure 1: The researchers' model uses input sentence to add ABC and output sentence to derive wxyz. The model's prediction stops when the end of sentience token is printed. LSTM reads input sentence backwards, making the optimization problem simpler.

2 Sequence to Sequence Model

BLEU The researchers' model was able to obtain a BLEU score of 34.81, when the dataset was WMT'14 English-French translation work. The researchers used LSTM as a score that enabled a thousand best lists of the same tasks based on SMT. By doing so, they scored 36.5 BLEU points, which is better than other tasks. SGD was able to optimize LSTM for long sentence without significant problems. It causes the technical distribution of this task as a result of a simple trick to reverse words.

RNN[5] is a feedforward natural derivation result of the neural network for permutation. For input $(x_1, \cdots x_T)$, the following formula is repeated when the output result is called $(y_1, \cdots, y_T)$.

$$h_t = \text{sigm}(W^{hx} x_t + W^{hh} h_{t-1}) \tag{1}$$

$$y_t = W^{yh} h_t \tag{2}$$

RNN[5] is a model that allows easy mapping from sequence to sequence, even if there is no alignment between input and output over time. A comprehensive method of sequential learning is to map a fixed-size vector into an RNN in input sequence. Then map the vector to another RNN with the target sequence.

The objective of LSTM is to estimate conditional probabilities as follows:

$$p(y_1, \cdots, y_{T'} | x_1, \cdots, x_T) \tag{3}$$

It should be noted that T' here differs from T. The conditional probability can be calculated by the following formula.

$$p(y_1, \cdots, y_{T'} | x_1, \cdots, x_T) = \prod_{t=1}^{T'} p(y_T | v, y_1, \cdots, y_{t-1}) \tag{4}$$

The latter part of the formula represents a softmax over all words in vocabulary. The researchers' models differ in three main areas. First, LSTM for input sequence and LSTM for output sequence are different. Second, they selected deep LSTM with four layers to implement high-performance LSTM. Finally, the word input sentence is inserted backwards to make it worth it.

3 Experiments

The experiment was conducted in two ways. The researchers did not refer to the SMT system and translated input sentence directly. The researchers reported the accuracy of the translation method, visualizing the sensitivity presentation by representing a simple sample translation.

3.1 Dataset Details

Researchers trained the sequence to sequence model using a dataset called French in WMT'14 English. The configuration of this dataset is simple. It consists of 304M English words and 348M French words, with a total of 12M sentences. Neural language models rely on vector representations of each word. Each word outside the vocabulary is replaced by a special UNK token.

3.2 Decoding and Rescoring

Since using deep LSTM to train with many sentence combinations is key to the researchers' experiments, the researchers trained to maximize log probability. The objective of training can be expressed as formula as follows:

$$1/|S| \sum_{(T,S) \in S} \log p(T|S) \tag{5}$$

The meaning of S is the training set. Training this model allows us to perform translation tasks with the following formulas. The formula can be expressed as follows.

$$\hat{T} = \text{argmax}_T \, p(T|S) \tag{6}$$

Translation was performed using a simple left-to-right beam search decoder. Partial hypothesis is some translation of prefix. For each timestep, the researchers extended each partial hypothesis from the beam. The EOS symbol is added to the hypothesis, which is removed from the beam and added to the completed hypothesis. To rescore the n-best list, the researchers calculated log probability.

3.3 Reversing the Source Sentences

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Figure 2: The above table shows LSTM measurements using the WMT'14 English to French test set. If measured using five LSTMs, it should be taken into account that the beam size is 2.

Researchers have demonstrated that LSTM exhibits better results when learning source sentences backwards. Previously, they knew that LSTM exhibits superior performance on problems that rely on long sentences. For your information, target sentience is not reversed but forward. This is a clear difference from training sentience. Experimental results showed that LSTM's test perflexity was 5.8 but dropped to 4.7 and test BLEU[3] was 25.9 but dropped to 30.6.

Researchers believe that this phenomenon was caused by the introduction of short-term dependencies on datasets, but it is not fully explained. The problem arose in the following. They had the least about the timestep. They show that for the first source, a few words are close to a small number of words in the first target language.

3.4 Training Details

LSTM models are easy to train. The researchers used LSTM in four layers. This layer has 1000 cells inserted into each layer and 1000 words embedding. There are 1600,000 input vocabularies and 80,000 output vocabularies. In addition, deep LSTM uses 8,000 mistakes to express sentences. Researchers found that deep LSTM outperforms shallow LSTM. The researchers used 80,000 words for each output. Listing the number of parameters, as a result of LSTM, 384M parameters became pure recurrent connections with 64M words. The details of the training are as follows.

1. The researchers initialized the number of parameters in LSTM, which is -0.08 to 0.08 in uniform distribution.
2. The researchers used SGD without momentum, with a learning rate of 0.7 fixed at 5 epochs.
3. Researchers cut the learning rate by half every five epochs.
4. The researchers trained the model at a total of 7.5 epochs.
5. The researchers used 128 sequences in batches and divided them into batches. It is literally 128.
6. Although LSTM does not tend to suffer from vanishing gradient problems, they have an expanding gradient. In addition, the researchers executed a rigid constrain with a norm of 10 to 25. For each training batch, the researchers computed $s = \|g\|_2$, and g is gradient divided by 128. If $s5$, the researchers say $g = 5g/s$. The researchers noted the simple fact that the sentences they use or the sentences for training are slightly different in length. However, there is a slight bias. Most training sentences consist of 20 to 30 relatively short sentences. There are also long sentences with more than 100 sentences. In the end, when 128 training sentences are randomly selected, minibatch shows that there are fewer long sentences, along with the fact that there are many short sentences. Furthermore, most computational processes show that accurate computations are made for the disappearance of minibatch. To solve this problem, the researchers made all sentences roughly the same length and calculated them in a way that doubled the speed.

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 3: The table above shows how the neural network was used together as a SMT system in the WMT'14 English to French test set[4].

3.5 Parallelization

deep LSTM is a model written in C++. A single GPU can process 1,700 words in one second. This is a speed that does not fit the purpose of the researchers, so it was processed using eight GPUs. Each layer of LSTM runs in excess of the other GPUs, receiving activities from the next GPU and running them sequentially. The researchers' model consists of four layers of LSTM, distributed to four GPUs for processing. The remaining four GPUs are used as a parallelize for softmax. Each GPU is multiplied by 1000×20000 matrix. The speed at which 6,300 words can be processed in one second for English and French was measured by using minibatch size of 128.

3.6 Experimental Results

Comment about Figure 2, 3, 4 The researchers used BLEU scores to measure the quality of the translation. multi-bleu.Using pl to obtain a BLEU score, the BLEU score is 33.3. However, calculated as WMT'14, 37.0 would be obtained. That's 35.8 higher than statmt.org/matrix. The results of the researchers were obtained by a bundle of LSTM[1]. It was randomly initialized.

Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

Figure 4: The table above shows examples of long transactions using LSTM.

3.7 Model Analysis

The advantage of this model, developed by the researchers, is that it has the ability to rotate sequences of words into fixed-dimensional vectors.

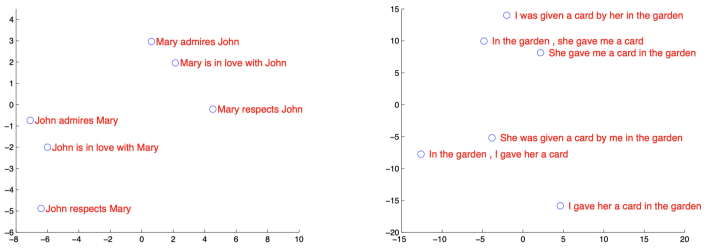

Figure 5: The figure above shows LSTM's hidden state's PCA two-dimensional projection. They can see that each phase collects clusters by meaning, and that it is difficult to capture with a bag-of-words model.

4 Conclusion

Through this study, researchers showed large deep LSTM. The model outperforms limited vocabulary and most estimation-free problem structures, resulting in a standard SMT-based system. This ensures that the vocabulary is unrestricted for large-scale MT tasks. Researchers say the expansion of improvements is surprising. It is said that it is important to find a problem encoding that has a large number of short term dependencies. Researchers believe standard RNN[5] can be easily trained.

5 Reference

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[2] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[5] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.