Sequence to Sequence Learning with Neural Networks

Muhammad Tasnim Mohiuddin

School of Computer Science and Engineering
Nanyang Technological University

March 14, 2018

Paper Abstract

Problem: Apply simple DNN approach to map sequences to sequences.

Paper Abstract

Problem: Apply simple DNN approach to map sequences to sequences.

Approach: Use two main steps:

1. Deep LSTMs to map the input sequence to vector of fixed dimensionality
2. Deep LSTMs to decode from vector to target sequence

Paper Abstract

Problem: Apply simple DNN approach to map sequences to sequences.

Approach: Use two main steps:

1. Deep LSTMs to map the input sequence to vector of fixed dimensionality
2. Deep LSTMs to decode from vector to target sequence

Achievement:

- Close to Winning Achievement of EN to FR SMT (BLEU score: 34.8 vs. 37)
- It can beat with improvement

Related Works

- Kalchbrenner and Blunsom (2013) were first to map the input sentence into a vector and then back to a sentence which lose the ordering of the words.

Related Works

- Kalchbrenner and Blunsom (2013) were first to map the input sentence into a vector and then back to a sentence which lose the ordering of the words.

- Cho et al. (2014) proposed a novel neural network model called RNN Encoder-Decoder Model. But it was used only for rescoring hypotheses produced by a phrase-based system.

Related Works

- Kalchbrenner and Blunsom (2013) were first to map the input sentence into a vector and then back to a sentence which lose the ordering of the words.
- Cho et al. (2014) proposed a novel neural network model called RNN Encoder-Decoder Model. But it was used only for rescoring hypotheses produced by a phrase-based system.
- Graves (2013) introduced a novel differentiable attention mechanism that allows neural networks to focus on different parts of their input.

The Model

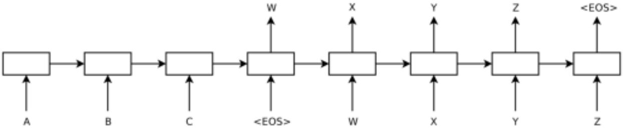


Figure 1: Our model reads an input sentence "ABC" and produces "WXYZ" as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that make the optimization problem much easier.

Differences Between the Model and Standard LSTMs

First: They used two different LSTMs: one for the input sequence (Encoder) and another for the output sequence (Decoder).

Second: They found that deep LSTMs significantly outperformed shallow LSTMs, so they used LSTM with four layers.

Third: They found it extremely valuable to reverse the order of the words of the input sentence.

Reversing the Source Sentences

- They do not have a complete explanation to this phenomenon.
- They believe that it is caused by the introduction of many short term dependencies to the dataset that make optimization problem much simpler.
- Each word in the source sentence is far from its corresponding word in the target sentence.
- By reversing the words in the source sentence, the average distance between corresponding words in the source and target language is unchanged.

  However, the first few words in the source language are now very close to the first few words in the target language.
- LSTMs trained on reversed source sentences did much better on long sentences than LSTMs trained on the raw source sentences - better memory utilization.

Decoding and Rescoring

Trained it by maximizing the log probability of a correct translation T given the source sentence S, so the training objective is to maximize

$$\frac{1}{\mathcal{S}} \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

where $\mathcal{S}$ is the training set.

Decoding and Rescoring

Trained it by maximizing the log probability of a correct translation T given the source sentence S, so the training objective is to maximize

$$\frac{1}{\mathcal{S}} \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

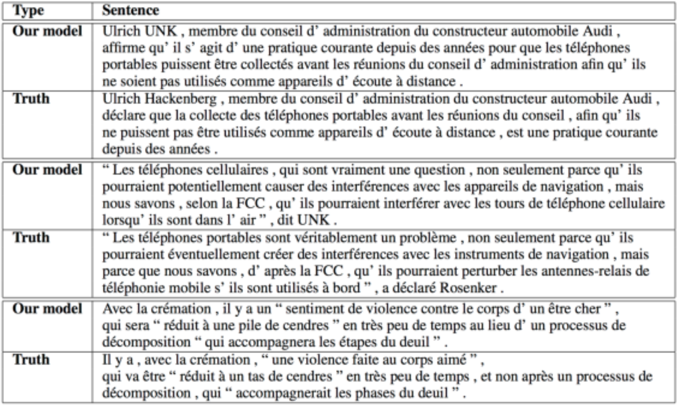
where $\mathcal{S}$ is the training set.

Once training is complete, translations are produced by finding the most likely translation according to

$$\hat{T} = \operatorname*{argmax}_{T} p(T|S)$$

Simple left-to-right beam search is used in the decoder to find the most likely translation.

Performance on Long Sentences



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 " .

Table 3: A few examples of long translations produced by the LSTM alongside the ground truth translations. The reader can verify that the translations are sensible using Google translate.

Model Analysis

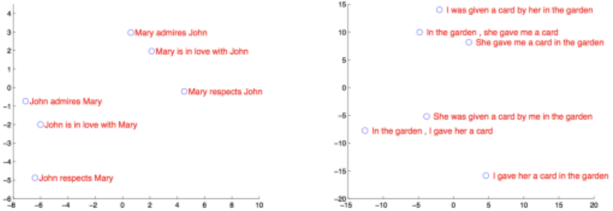


Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words model. Notice that both clusters have similar internal structure.

The figure clearly shows that the representations are sensitive to the order of words, while being fairly insensitive to the replacement of an active voice with a passive voice.

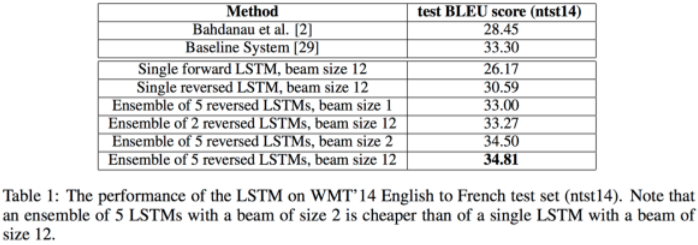
Experimental Dataset

- WMT'14 English to French dataset
- 12M sentences of 348M French words and 304M English words
- Vocabulary consisted of:
  - 160,000 most frequent words for the source language
  - 80,000 most frequent words for the target language
  - Replacement for all out-of-vocabulary words with special UNK token

Training Details

- LSTM parameters initialized with a uniform distribution between -0.08 and 0.08
- Used SGD without momentum, learning rate of 0.7. Learning rate halved at each epoch after 5
- Hard constraint on the gradient enforced to resolve exploding gradient problem
- Each minibatch was curated so that sentence lengths were roughly uniform

Experimental Results



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**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

Experimental Results

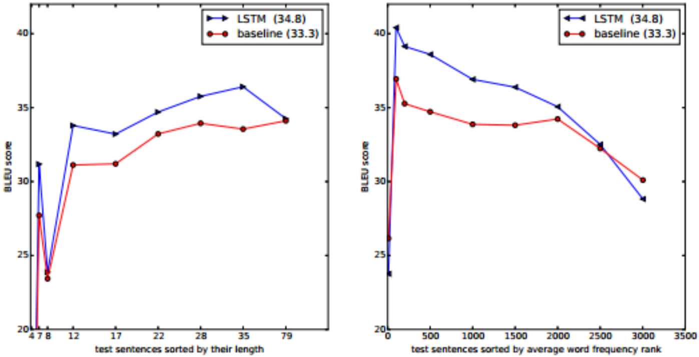


Figure 3: The left plot shows the performance of our system as a function of sentence length, where the x-axis corresponds to the test sentences sorted by their length and is marked by the actual sequence lengths. There is no degradation on sentences with less than 35 words, there is only a minor degradation on the longest sentences. The right plot shows the LSTM's performance on sentences with progressively more rare words, where the x-axis corresponds to the test sentences sorted by their "average word frequency rank".

Conclusion

- A large, deep LSTM outperformed an SMT system despite having a limited vocabulary and no assumptions about the problem structure
- The simple trick of reversing the words in the source sentence is one of the key technical contributions of this work
- Model performed surprisingly well on long sentences
- Further optimization of this model could achieve yet more impressive results