

# Sutskever et al. 2014

## Sequence to Sequence Learning with Neural Networks

Josh Meyer

## 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.

# Intro

Despite their flexibility and power, DNNs can only be applied to problems whose inputs and targets can be sensibly encoded with vectors of fixed dimensionality. It is a significant limitation, since many important problems are best expressed with sequences whose lengths are not known a-priori. For example, speech recognition and machine translation are sequential problems. Likewise, question answering can also be seen as mapping a sequence of words representing the question to a

# Intro

Sequences pose a challenge for DNNs because they require that the dimensionality of the inputs and outputs is known and fixed. In this paper, we show that a straightforward application of the Long Short-Term Memory (LSTM) architecture [16] can solve general sequence to sequence problems. The idea is to use one LSTM to read the input sequence, one timestep at a time, to obtain large fixed-dimensional vector representation, and then to use another LSTM to extract the output sequence from that vector (fig. 1). The second LSTM is essentially a recurrent neural network language model [28, 23, 30] except that it is conditioned on the input sequence. The LSTM's ability to successfully learn on data with long range temporal dependencies makes it a natural choice for this application due to the considerable time lag between the inputs and their corresponding outputs (fig. 1).

# Intro

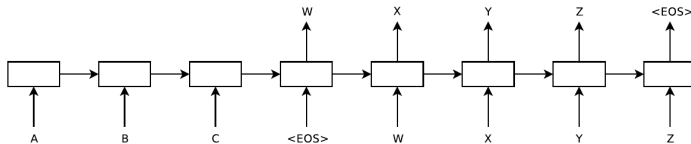


Figure 1: Our model reads an input sentence “ABC” and produces “WXYZ” as the output sentence. The model stops making predictions after outputting the end-of-sentence token. Note that the LSTM reads the input sentence in reverse, because doing so introduces many short term dependencies in the data that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

# Intro

The main result of this work is the following. On the WMT'14 English to French translation task, we obtained a BLEU score of **34.81** by directly extracting translations from an ensemble of 5 deep LSTMs (with 384M parameters and 8,000 dimensional state each) using a simple left-to-right beam-search decoder. This is by far the best result achieved by direct translation with large neural networks. For comparison, the BLEU score of an SMT baseline on this dataset is 33.30 [29]. The 34.81 BLEU score was achieved by an LSTM with a vocabulary of 80k words, so the score was penalized whenever the reference translation contained a word not covered by these 80k. This result shows that a relatively unoptimized small-vocabulary neural network architecture which has much room for improvement outperforms a phrase-based SMT system.

# LSTM

The goal of the LSTM is to estimate the conditional probability  $p(y_1, \dots, y_{T'} | x_1, \dots, x_T)$  where  $(x_1, \dots, x_T)$  is an input sequence and  $y_1, \dots, y_{T'}$  is its corresponding output sequence whose length  $T'$  may differ from  $T$ . The LSTM computes this conditional probability by first obtaining the fixed-dimensional representation  $v$  of the input sequence  $(x_1, \dots, x_T)$  given by the last hidden state of the LSTM, and then computing the probability of  $y_1, \dots, y_{T'}$  with a standard LSTM-LM formulation whose initial hidden state is set to the representation  $v$  of  $x_1, \dots, x_T$ :

$$p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1}) \quad (1)$$

# LSTM

Our actual models differ from the above description in three important ways. **First**, we used two different LSTMs: one for the input sequence and another for the output sequence, because doing so increases the number model parameters at negligible computational cost and makes it natural to train the LSTM on multiple language pairs simultaneously [18]. **Second**, we found that deep LSTMs significantly outperformed shallow LSTMs, so we chose an LSTM with four layers. **Third**, we found it extremely valuable to reverse the order of the words of the input sentence. So for example, instead of mapping the sentence  $a, b, c$  to the sentence  $\alpha, \beta, \gamma$ , the LSTM is asked to map  $c, b, a$  to  $\alpha, \beta, \gamma$ , where  $\alpha, \beta, \gamma$  is the translation of  $a, b, c$ . This way,  $a$  is in close proximity to  $\alpha$ ,  $b$  is fairly close to  $\beta$ , and so on, **a fact that makes it easy for SGD to “establish communication” between the input and the output**. We found this simple data transformation to greatly improve the performance of the LSTM.



## 3.1 Dataset details

We used the WMT'14 English to French dataset. We trained our models on a subset of 12M sentences consisting of 348M French words and 304M English words, which is a clean “selected” subset from [29]. We chose this translation task and this specific training set subset because of the public availability of a tokenized training and test set together with 1000-best lists from the baseline SMT [29].

As typical neural language models rely on a vector representation for each word, we used a fixed vocabulary for both languages. We used 160,000 of the most frequent words for the source language and 80,000 of the most frequent words for the target language. Every out-of-vocabulary word was replaced with a special “UNK” token.

# Decoding and Scoring

## 3.2 Decoding and Rescoring

The core of our experiments involved training a large deep LSTM on many sentence pairs. We trained it by maximizing the log probability of a correct translation  $T$  given the source sentence  $S$ , so the training objective is

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

where  $\mathcal{S}$  is the training set. Once training is complete, we produce translations by finding the most likely translation according to the LSTM:

$$\hat{T} = \arg \max_T p(T|S) \quad (2)$$

# Decoding and Scoring

We search for the most likely translation using a simple **left-to-right beam search decoder** which maintains a small number  $B$  of partial hypotheses, where a partial hypothesis is a prefix of some translation. At each timestep we extend each partial hypothesis in the beam with every possible word in the vocabulary. This greatly increases the number of the hypotheses so we discard all but the  $B$  most likely hypotheses according to the model's log probability. As soon as the “<EOS>” symbol is appended to a hypothesis, it is removed from the beam and is added to the set of complete hypotheses. While this decoder is approximate, it is simple to implement. Interestingly, our system performs well even with a beam size of 1, and a beam of size 2 provides most of the benefits of beam search (Table 1).

# Reverse

While we do not have a complete explanation to this phenomenon, we believe that it is caused by the introduction of many short term dependencies to the dataset. Normally, when we concatenate a source sentence with a target sentence, each word in the source sentence is far from its corresponding word in the target sentence. As a result, the problem has a large “minimal time lag” [17]. 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, so the problem’s minimal time lag is greatly reduced. Thus, backpropagation has an easier time “establishing communication” between the source sentence and the target sentence, which in turn results in substantially improved overall performance.

## 3.4 Training details

We found that the LSTM models are fairly easy to train. We used deep LSTMs with 4 layers, with 1000 cells at each layer and 1000 dimensional word embeddings, with an input vocabulary of 160,000 and an output vocabulary of 80,000. Thus the deep LSTM uses 8000 real numbers to represent a sentence. We found deep LSTMs to significantly outperform shallow LSTMs, where each additional layer reduced perplexity by nearly 10%, possibly due to their much larger hidden state. We used a naive softmax over 80,000 words at each output. The resulting LSTM has 384M parameters of which 64M are pure recurrent connections (32M for the “encoder” LSTM and 32M for the “decoder” LSTM). The complete training details are given below:

# Training

- We **initialized** all of the LSTM's parameters with the uniform distribution between -0.08 and 0.08
- We used **stochastic gradient descent** without momentum, with a fixed learning rate of 0.7. After 5 epochs, we begun halving the learning rate every half epoch. We trained our models for a total of 7.5 epochs.
- We used **batches of 128 sequences** for the gradient and divided it the size of the batch (namely, 128).
- Although LSTMs tend to not suffer from the vanishing gradient problem, they can have exploding gradients. Thus **we enforced a hard constraint on the norm of the gradient** [10, 25] by scaling it when its norm exceeded a threshold. For each training batch, we compute  $s = \|g\|_2$ , where  $g$  is the gradient divided by 128. If  $s > 5$ , we set  $g = \frac{5g}{s}$ .
- Different sentences have different lengths. Most sentences are short (e.g., length 20-30) but some sentences are long (e.g., length  $> 100$ ), so a minibatch of 128 randomly chosen training sentences will have many short sentences and few long sentences, and as a result, much of the computation in the minibatch is wasted. To address this problem, we made sure that **all sentences in a minibatch are roughly of the same length**, yielding a 2x speedup.

## 3.5 Parallelization

A C++ implementation of deep LSTM with the configuration from the previous section on a single GPU processes a speed of approximately 1,700 words per second. This was too slow for our purposes, so we parallelized our model using an 8-GPU machine. Each layer of the LSTM was executed on a different GPU and communicated its activations to the next GPU / layer as soon as they were computed. Our models have 4 layers of LSTMs, each of which resides on a separate GPU. The remaining 4 GPUs were used to parallelize the softmax, so each GPU was responsible for multiplying by a  $1000 \times 20000$  matrix. The resulting implementation achieved a speed of 6,300 (both English and French) words per second with a minibatch size of 128. Training took about a ten days with this implementation.

# Results

## 3.6 Experimental Results

We used the cased BLEU score [24] to evaluate the quality of our translations. We computed our BLEU scores using `multi-bleu.pl`<sup>1</sup> on the *tokenized* predictions and ground truth. This way of evaluating the BLEU score is consistent with [5] and [2], and reproduces the 33.3 score of [29]. However, if we evaluate the best WMT'14 system [9] (whose predictions can be downloaded from `statmt.org\matrix`) in this manner, we get 37.0, which is greater than the 35.8 reported by `statmt.org\matrix`.

The results are presented in tables 1 and 2. **Our best results are obtained with an ensemble of LSTMs that differ in their random initializations and in the random order of minibatches.** While the decoded translations of the LSTM ensemble do not outperform the best WMT'14 system, it is the first time that a pure neural translation system outperforms a phrase-based SMT baseline on a large scale MT



# 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	<b>34.81</b>

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.

# Results

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	<b>37.0</b>
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	<b>36.5</b>
Oracle Rescoring of the Baseline 1000-best lists	~45

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

# Model Analysis

## 3.8 Model Analysis

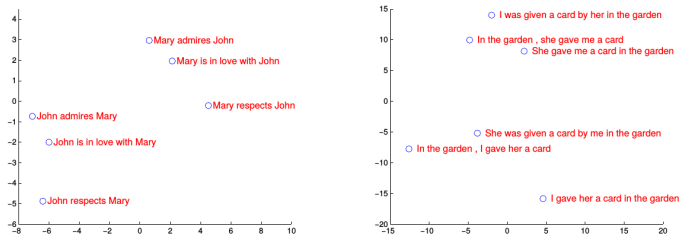


Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words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

tence.” We first compute the  $n$ -gram matches sentence by sentence. Next, we add the clipped  $n$ -gram counts for all the candidate sentences and divide by the number of candidate  $n$ -grams in the test corpus to compute a modified precision score,  $p_n$ , for the entire test corpus.

$$p_n = \frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram}' \in C'} \text{Count}(n\text{-gram}')}.$$

# BLEU 2

We first compute the geometric average of the modified  $n$ -gram precisions,  $p_n$ , using  $n$ -grams up to length  $N$  and positive weights  $w_n$  summing to one.

Next, let  $c$  be the length of the candidate translation and  $r$  be the effective reference corpus length. We compute the brevity penalty BP,

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log p_n \right).$$

The ranking behavior is more immediately apparent in the log domain,

$$\log \text{BLEU} = \min\left(1 - \frac{r}{c}, 0\right) + \sum_{n=1}^N w_n \log p_n.$$