Sequence to Sequence Learning with Neural Networks

By Ilya Sutskever, Oriol Vinyals, Quoc V. Le

Presented by Nathan Sulecki

Review of RNNs

- Map sequences of inputs (x_1,…,x_T) to a sequence of outputs (y_1, … , y_T) via:

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

$$y_t = W^{yh} h_t$$

- Input and output are of same dimensionality

Variable Dimensionality

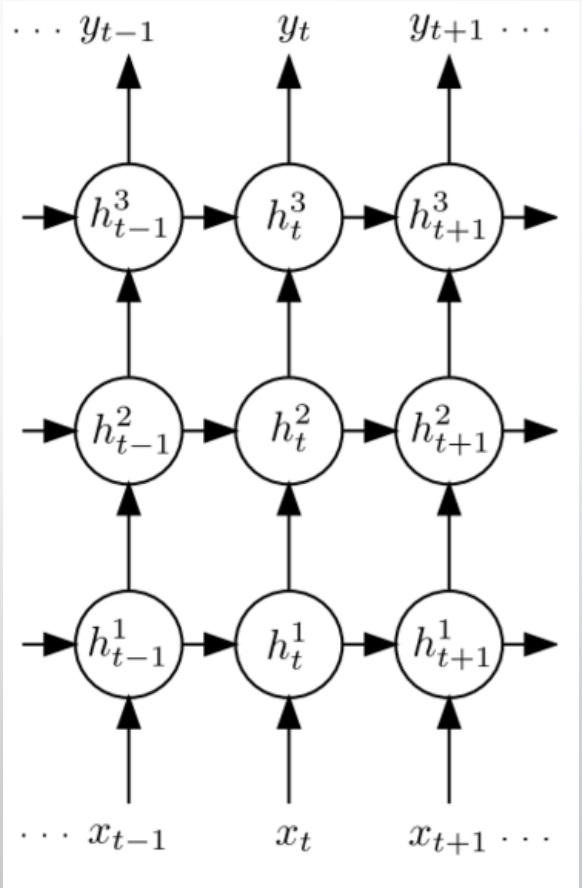
“I am eating the food”	→	“Yo estoy comiendo la comida”
“I’m eating the food”	→	“Yo estoy comiendo la comida”
“I am eating the food”	→	“Estoy comiendo la comida”
“I am eating the food today”	→	“Estoy comiendo la comida de hoy”

Use of LSTMs in Language Modeling

- Linking together RNNs results in long term dependencies; LSTMs used instead
- Estimates $p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T)$ by using v, the fixed dimensional representation of the inputs
- This gives us $$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$
- Each $p(y_t | v, y_1, \ldots, y_{t-1})$ is represented by a softmax over all words in vocabulary

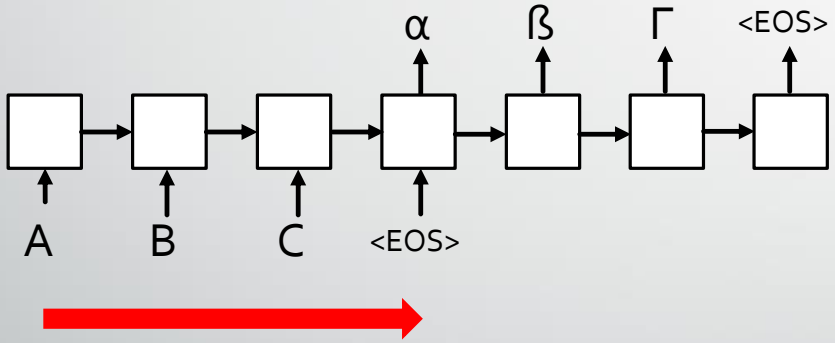
Areas Where This Paper Differed

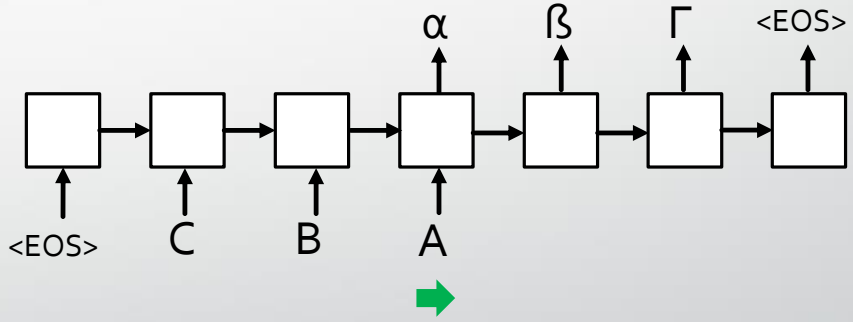
- 3 Major differences:
  - Used two different LSTMs; one for input, one for output
  - Use of four-layer LSTMs
  - Reversal of input sentences



Reversal of Inputs

- "We do not have a complete explanation of this phenomenon"
- Reduces the minimal time lag
- Mean distance from source words to target words unchanged





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

Decoding and Rescoring

- Training was done on sentence pairs via a maximum likelihood estimation objective function:
  - $\frac{1}{|R|}\sum_{(T,S)\in R} \log p(T|S)$ where T is the potential translation, S is the source sentence, and R is the training set*
- After training, translations are given by $\hat{T} = \arg\max_{T} p(T|S)$
- Hypotheses are curated via a left-to-right beam search decoder to keep a predetermined number of hypotheses
- For rescoring, log probabilities were computed for every hypothesis via the new LSTM and were averaged with the old score

*Notation differs from paper. Having two different meanings for "S" was needlessly confusing

Training Details

- LSTM parameters initialized with a uniform distribution between -0.08 and 0.08
- Used stochastic gradient descent without momentum, learning rate of 0.7
  - Learning rate halved at each epoch after 5, trained for 7.5 epochs total
- Gradient used batches of 128 sequences and divided by 128 (size of batch)
- Hard constraint on the gradient enforced
  - $g = \frac{5g}{s}$ if $s > 5$, where s is given by $s = \|g\|_2$
- Each minibatch was curated so that sentence lengths were roughly uniform

Results

- Reversal of input sentences reduced perplexity from 5.8 to 4.7
  - Perplexity is (roughly) a measure of the overall estimated probability of the outputted hypotheses
- This also increased the BLEU scores of the model's decoded translation from 25.9 to 30.6
  - BLEU is an algorithm for evaluating translated text at corpus-level

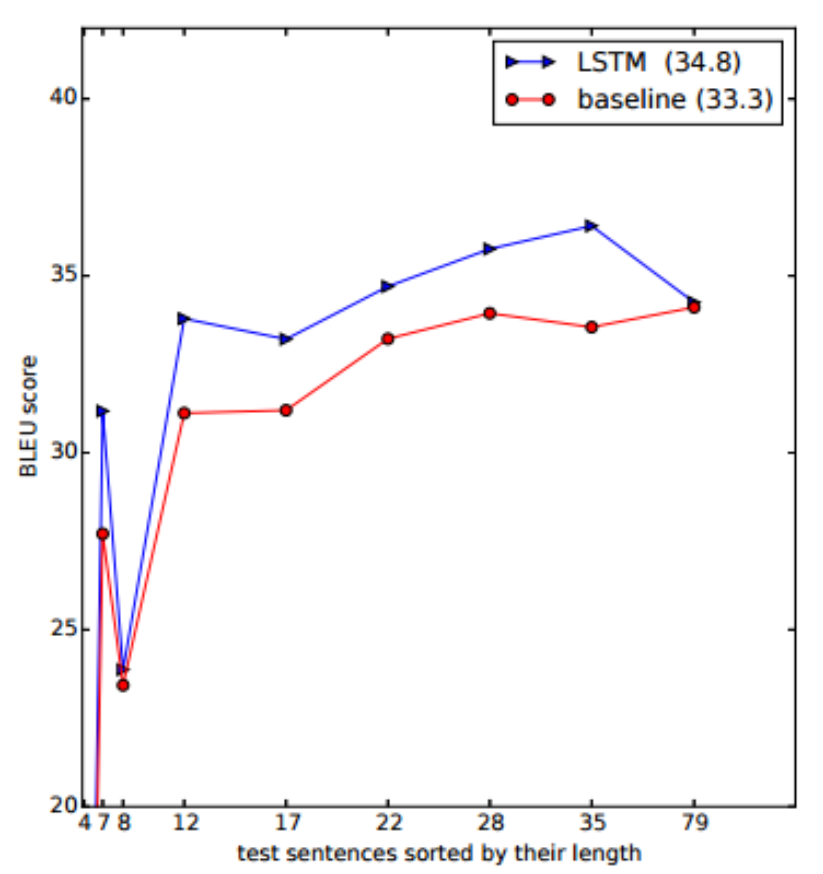
Results

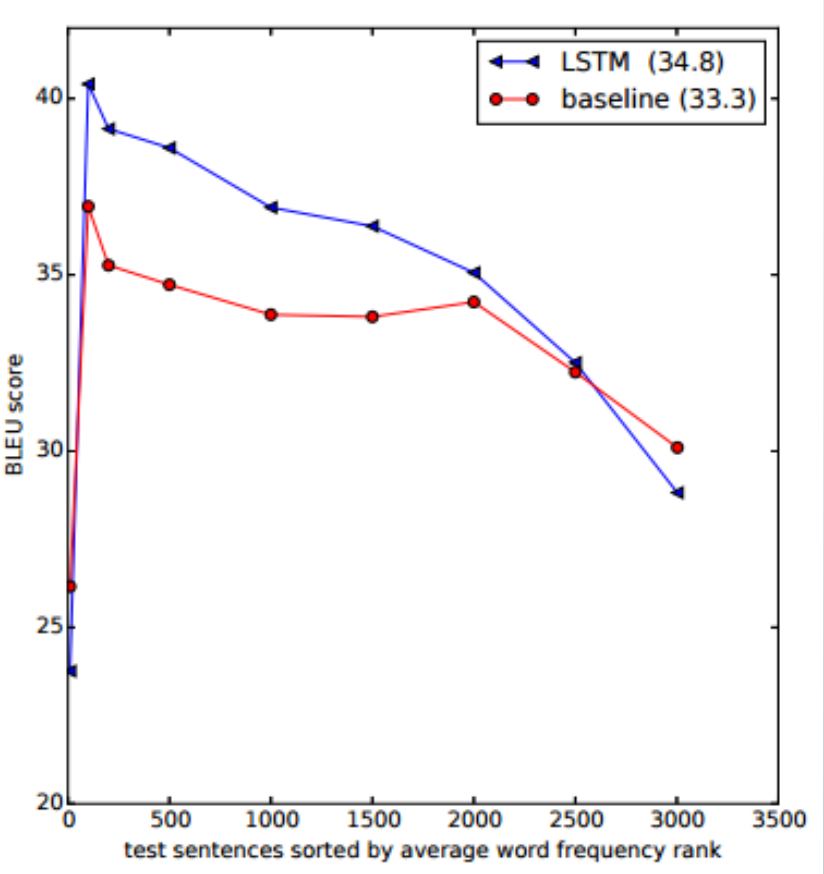
- First example of a "pure neural translation system outperforming a SMT baseline on a large scale MT task by a sizable margin"
  - Done even without ability to handle out-of-vocabulary words

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**

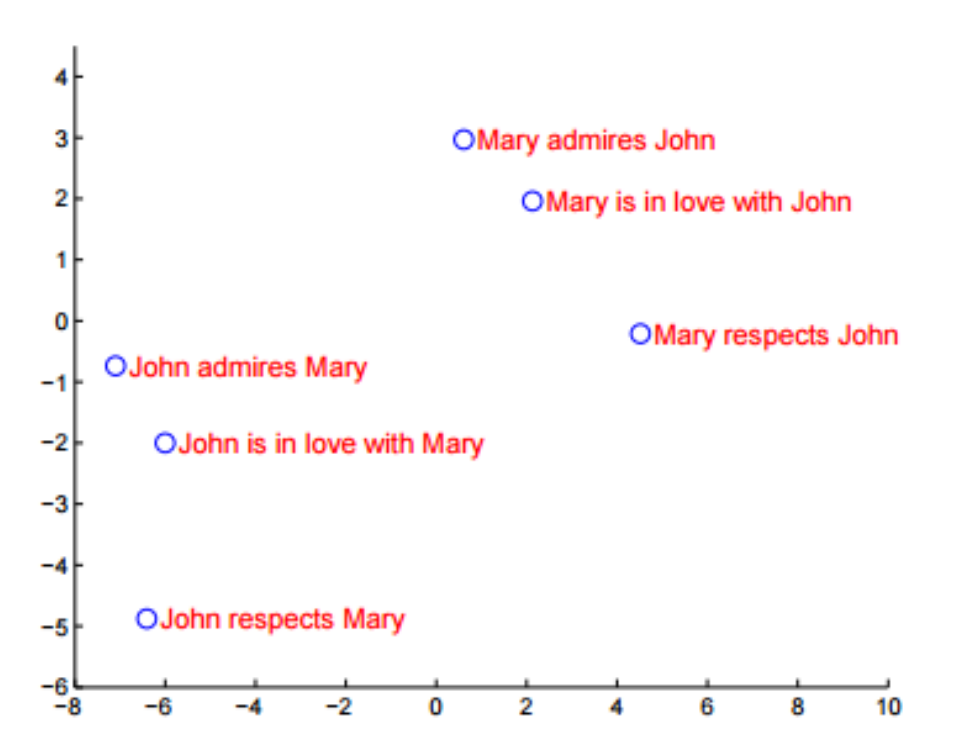
Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed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

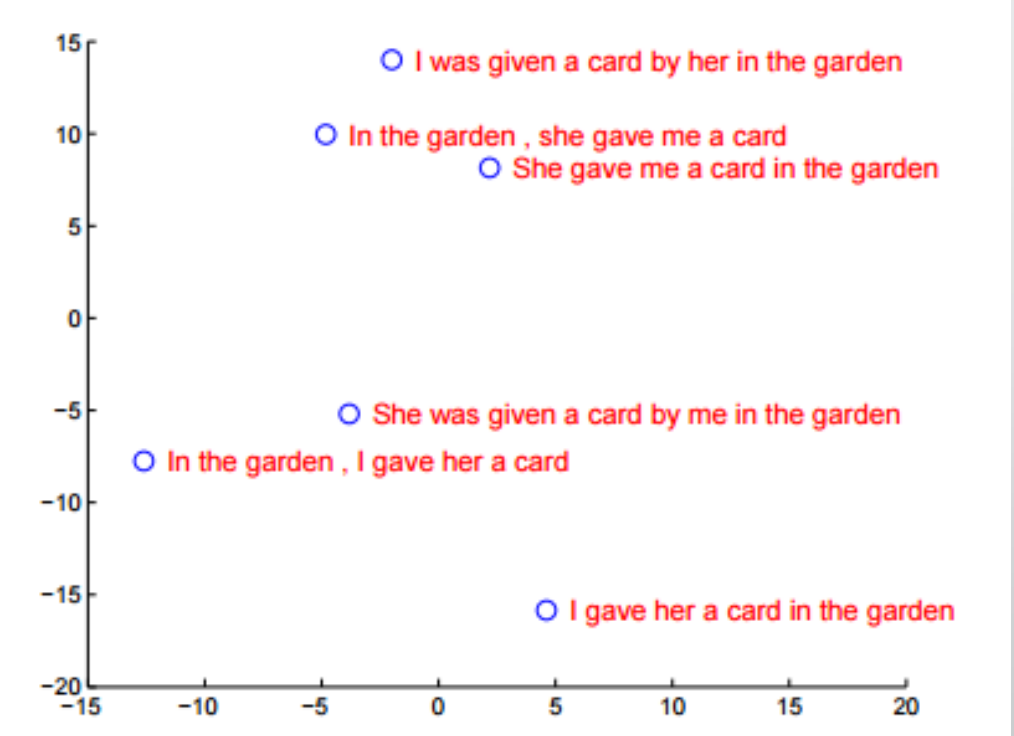
Results


LSTM (34.8)
baseline (33.3)
BLEU score
40
35
30
25
20
4 7 8
12
17
22
28
35
79
test sentences sorted by their length


LSTM (34.8)
baseline (33.3)
BLEU score
40
35
30
25
20
0
500
1000
1500
2000
2500
3000
3500
test sentences sorted by average word frequency rank

Model Analysis





Related Work

- Kalchbrenner and Blunsom were the first to map from a sentence to a vector and back, but did so without retaining word ordering.
- Cho et al. used LSTM architecture, but applied it as an improvement upon an STM system and ran into memory problems
- Bahdanau et al. and Pouget-Abadie et al. used various complex methods to try to solve the problems experienced by Cho et al. with "encouraging results"

Conclusions

- A large, deep LSTM outperformed an SMT system despite having a limited vocabulary and no assumptions about the problem structure
- Learning can be simplified by finding a problems encoding with the greatest number of short term dependencies (as shown by sentence reversal)
- Sentence reversal also still allowed for translation of longer sentences
- Further optimization of this system could achieve yet more impressive results