# - NLP - Paper Review -
# Sequence to Sequence Learning with Neural Networks

Ayoub Youssoufi, Yassin El Hajj Chehade

February $10^{th}$, 2023

# Table of Contents

# Introduction

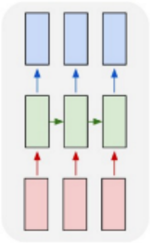
Deep Neural Network (DNN) are extremely powerful models. However, they have some limitations:

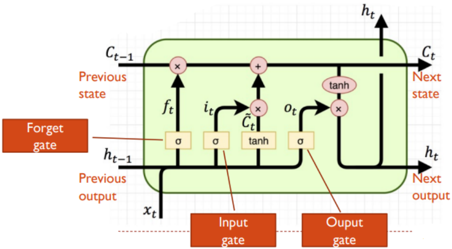
- can only map vectors to vectors with fixed dimensionality
- cannot map sequences to sequences
- learning to map sequences to sequences is important
  - Machine Translation
  - Speech Recognition
  - Image caption generation
  - Many other interesting tasks

**The goal of this paper : Solve the sequence to sequence problems**

Can we use RNN ?

- Have a one-to-one correspondence between the input and the outputs
- have trouble learning "long-term dependencies"
  - Vanishing gradient problem: use of LSTM
  - Exploding gradient problem : Gradient clipping






Long-Short-Term-Memory (LSTM) is a certain RNN architecture that has no vanishing gradient $\Pr(Y_1, \ldots, Y_T \mid X_1, \ldots, X_q) = \prod_{q=1} \Pr(Y_q \mid v, Y_1, \ldots, Y_{q-1})$

# Table of Contents

# Main idea:

- Have an LSTM first read the input sequence
- Produce the output sequence

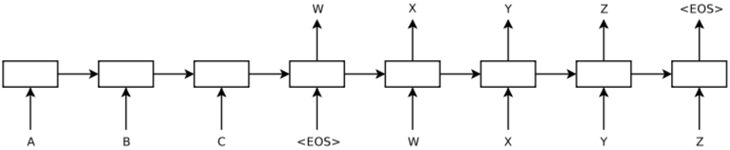



Figure: model reads an input sentence “ABC” and produces “WXYZ” as the output sentence.

# Big Dataset:

The architecture :

- WMT'14 English to French
- 340M french words
- 303M English words
- 160K input words and 80K output words
- 4 layers of LSTMs

The learning parameters :

- batch_size = 128
- initialized LSTM: uniform distribution between -0.08 and 0.08
- learning rate is halved every 0.5 epoch /5epochs
- using Parallelization with 8 GPUs ...

## Objective function

An experiments involved training a large deep LSTM. Trained by maximizing the objective function :

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

$$\hat{T} = \arg\max_{T} p(T|S)$$

- T: Target sentence
- S: Input sentence

Reverse the input of the input sentence when mapping to the output. (abc mapped to XYZ will be bca to XYZ)

# Table of Contents

# 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** |

Figure: 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.

# Table of Contents

# Model analysis 1

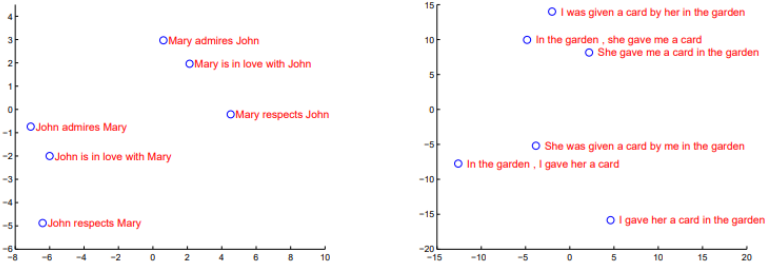



Figure: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words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

# Model analysis 2

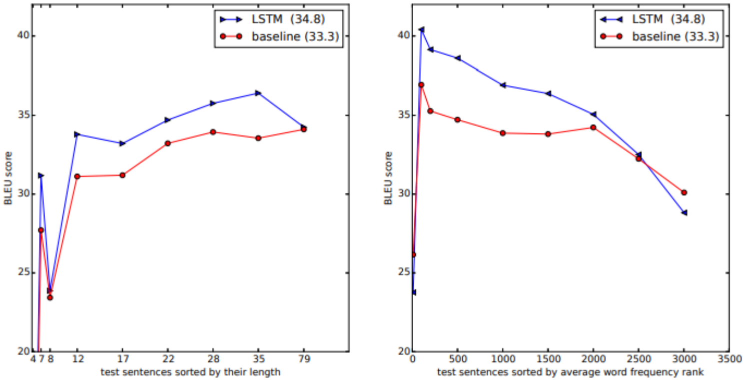



Figure: The left plot shows the performance of our system as a function of sentence length. The right plot shows the LSTM's performance on sentences with progressively more rare words.

# Table of Contents

# Conclusion

- Seq2Seq based on LSTM outperforms standard SMT.
- Reversing the order of words in the source sentence improves the performance.
- LSTM performs well in very long sentences.

# Critic

**Pros**

- Well structured, clear

**Limits**

- reliance to left-to-right beam search decoder which may not be optimal for all the languages.
- Focus only in the single language pair (English-French) and not cross-lingual translation.
- The lack of exploration of other NLP tasks to validate the robustness of the model.

# Questions

Thank you for your attention!