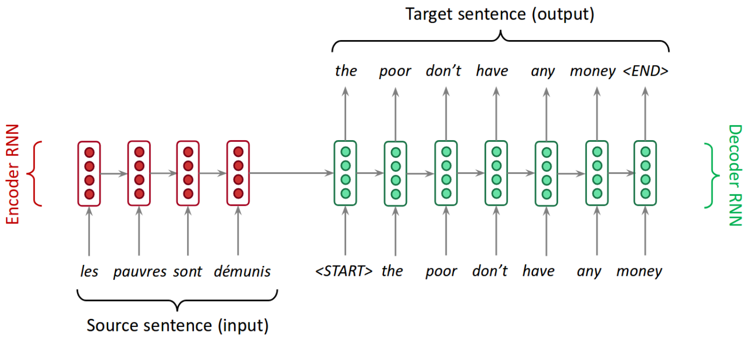
Neural Machine Translation by Jointly Learning to Align and Translate

Muhammad Tasnim Mohiuddin

School of Computer Science and Engineering
Nanyang Technological University

March 21, 2018

Encoder-Decoder Model in NMT



- Encoder reads and encodes a source sentence into a fixed-length vector (Thought Vector / Context Vector).
- Decoder then outputs a translation from the encoded vector.

Problems with Basic Encoder-Decoder Model

Encoder needs to be able to **compress all the necessary information** of a source sentence **into a fixed-length vector**.

Problematic while dealing with long sentences.
Performance of a basic encoder-decoder model deteriorates rapidly as the length of an input sentence increases.

Problems with Basic Encoder-Decoder Model

Encoder needs to be able to **compress all the necessary information** of a source sentence **into a fixed-length vector**.

Problematic while dealing with long sentences.
Performance of a basic encoder-decoder model deteriorates rapidly as the length of an input sentence increases.

What is the solution?

Attention Mechanism

- The most important distinguishing feature of this approach from the basic encoderdecoder is that it does not attempt to encode a whole input sentence into a single fixed-length vector.

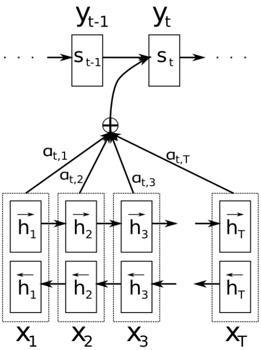
Attention Mechanism

- The most important distinguishing feature of this approach from the basic encoderdecoder is that it does not attempt to encode a whole input sentence into a single fixed-length vector.

- Each time the proposed model generates a word in a translation, it searches for a set of positions in a source sentence where the most relevant information is concentrated.

- This frees a neural translation model from having to squash all the information of a source sentence, regardless of its length, into a fixed-length vector.

Proposed Model



- Context vector

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

- Weight α_{ij} is computed by

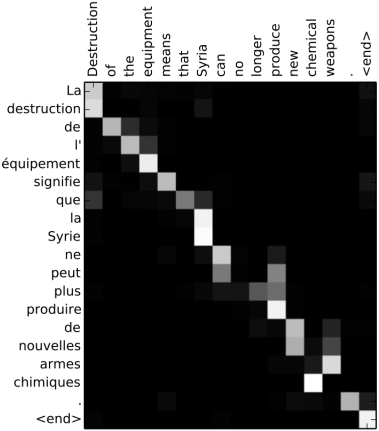
$$\alpha_{ij} = softmax(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

- Alignment model score

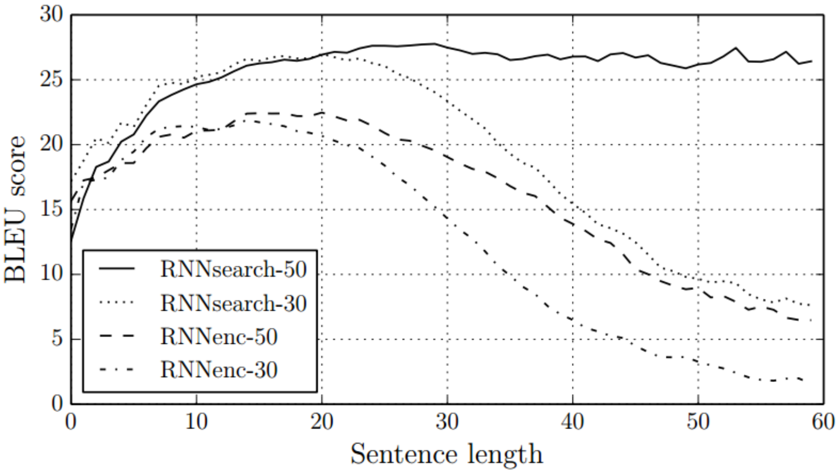
$$e_{ij} = v_a^T \tanh(W_a s_{i-1} + U_a h_j)$$

It is a feedforward neural network jointly trained with others.

Visualize the effect of weight α_{ij}



Result


30
25
20
15
10
5
0
BLEU score
0
10
20
30
40
50
60
Sentence length
RNNsearch-50
RNNsearch-30
RNNenc-50
RNNenc-30