Neural Machine Translation by Jointly Learning to Align and Translate

Date	@March 2, 2022
Tags	Attention Mechanism, NLP, Recurrent Neural Networks
Link	https://arxiv.org/abs/1409.0473
Authors	Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio
Status	Reading
Comments	ICLR 2015
File	NMT_attention.pdf

WHAT?

Neural Machine Translation Model using all hidden states of encoder state while decoding the output using a mechanism called Attention.

WHY?

The fixed-sized context vector used in encoder-decoder architecture cannot encode all the information of long sequences.

HOW?

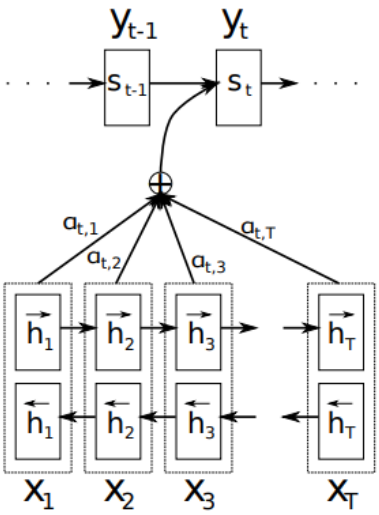

Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

- Using all hidden states as input to calculate context vectors
- Calculating context vector based on the alignment scores, which gives information about how much similar the source word is to the previous output of the decoder
- The context vector is calculated as:

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

Where,

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

and

$$e_{ij} = a(s_{i-1}, h_i)$$

s_{i-1} → Decoder output of the previous time step

h_j → Encoder output

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

v_a^T, W_a, U_a are weights of feedforward neural network.

- This extra layer is called attention layer, which finds where to focus more when translating the text.

AND?

- Authors have use Bi-directional LSTMs to train all the models
- On training two models each of attention architecture and encoder-decoder architecture, for 30 and 50 words long sequences both models with attention architecture performed better.

Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50⋆	28.45	36.15
Moses	33.30	35.63

Table 1: BLEU scores of the trained models computed on the test set. The second and third columns show respectively the scores on all the sentences and, on the sentences without any unknown word in themselves and in the reference translations. Note that RNNsearch-50⋆ was trained much longer until the performance on the development set stopped improving. (○) We disallowed the models to generate [UNK] tokens when only the sentences having no unknown words were evaluated (last column).