Neural Machine Translation by Jointly Learning to Align and Translate

December 20, 2023

1 Summary

0. Existing state-of-the-art translations networks use encoder-decoder format in which the entire input sentence is encoded first and then decoded using another network. This makes dealing with long input problematic, or sometimes delayed. The paper proposes an alternative in which input space is 'soft' searched when generating a new word in the output. This achieves comparable results with existing algorithms while dealing with long inputs effectively.

1. **Background.** Existing encoder-decoder RNN architecture is captured by the equations:

$$h_t = f(x_t, h_{t-1})$$
$$c = q(\{h_1, \cdots, h_{T_x}\})$$

where h_t is the hidden state, x_t is the input at t, c is the encoder generated input for decoder, and f, q are non-linear functions. The decoder RNN with hidden state at t denoted by s_t follows

$$p(y_t|\{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$

2. The Model. The new models is described by

$$p(y_t|\{y_1, \cdots, y_{t-1}\}, \mathbf{x}) = g(y_{t-1}, s_t, c_i)$$
$$s_i = f(s_i, y_{i-1}, c_i)$$

Importantly, here c_i is specific to the output y_i being generated. It is dependent on underlying *annotations* $\{h_1, \cdots, h_T\}$. Each h_i generated by the encoder contains information about whole input but emphasis on i-th word. The context vector c_i is computed by

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

where

$$\alpha_{ij} = \sum^{T_x} \frac{e_{ij}}{\sum_{k=1}^{T_x} e_{ik}}$$
$$e_{ij} = a(s_{i-1}, h_j)$$

is an *alignment model* which scores how inputs around j and output at i match. The alignment model is trained as a feed forward network with the whole network jointly. The c_i can be interpreted as taking expected annotation over possible alignments.

By letting the decoder have an attention mechanism, the paper relieves the encoder from the burden of having to encode all information in the source sentence into a fixed- length vector. With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly.

In the encoder, the RNN is passed through twice to concatenate and obtain bidirectional annotations.

3. **Experiments and Results.** The experiments set up were similar to state-of-the-art neural networks in machine translation and the results outperformed them. More importantly they were able to perform well in long sentences.

4. **Qualitative interpretation.** The alignments factors seem to be putting more weight on the appropriate words which is a good way to interpret the model.

2 Analysis

This paper although achieves moderate gains over existing networks it is a new way to think about language models. As it turns out it is indeed a stepping stone for the now influential Transformers paper as it is the first paper to introduce 'attention' ('alignment' in this paper) model. It is an important paper demonstrating new approaches in design can have significant improvements over existing mechanisms.