Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] Neural Machine Translation by Jointly Learning to Align and Translate

1. Paper Summary

Most of the neural machine translation models are based on RNN Encoder-Decoder networks. Here, the encoder reads and encodes a source sentence $\mathrm{x} = (\mathrm{x}_1, \dots, x_{T_x})$ into a fixed-length vector $\mathrm{c} = \mathrm{q}(\{\mathrm{h}_1, \dots, h_{T_x}\})$ where $h_t = f(x_t, h_{t-1})$ is the hidden state of encoder RNN at time t. Then, the decoder outputs a translation $\mathrm{y} = (\mathrm{y}_1, \dots, \mathrm{y}_{T_y})$ from the conditional probability on each component $\mathrm{p}(\mathrm{y}_t|\{y_1, \dots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$ where s_t is the hidden state of decoder RNN at time t. Then, these are jointly trained to maximize the probability of a correct translation given a source sentence. From this probabilistic perspective, the translation is equivalent to finding a target sentence that maximizes the conditional probability given a source sentence $\mathrm{argmax}_y p(y|x)$, which can be done by fitting the parameters in Encoder-Decoder networks.

However, there is a potential with this encoder-decoder approach where a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector. As Cho has shown in his work, the network suffers in case of learning long sentences, getting worse when the sentences are longer than those in the training corpus. Therefore, the author addressed this issue by introducing a novel architecture that is names as 'RNNsearch'. One of the most remarkable characteristics of this model is the fact that the proposed model learns to align and translate jointly. Whenever the proposed model generates a word, it searches for a set of positions in a source sentence using attention mechanism. This is to find the most relevant information in the source when generating a target word so that a model does not have to squash all the information into a fixed length vector, which allows it to cope better with long sentences.

For decoder, the conditional probability changes to $\mathrm{p}(\mathrm{y}_\mathrm{i}|y_1, \dots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i)$ where $\mathrm{s}_\mathrm{i} = k(s_{i-1}, y_{i-1}, c_i)$ is again the hidden state of the decoder at time t. Now, the probability is conditioned on a distinct context vector c_i for each target word y_i and the context vector c_i can be expressed as a weighted sum of the sequence of annotations $(\mathrm{h}_1, \dots, \mathrm{h}_{\mathrm{T_x}})$. Here, the weight α_{ij} of each annotation h_j is computed as the follows where it indicates the probability that the target word y_i is aligned to, or translated from, a source word x_i.

$$\alpha_{\mathrm{ij}} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \qquad where \quad e_{ij} = a(s_{i-1}, h_j)$$

(e_{ij} is named as an 'alignment model' since it scores how well the inputs around position j and the output at position i match.)

For encoder, the author used bidirectional RNN to make the annotation of each word to summarize not only the preceding words, but also the following words. This bidirectional RNN consists of forward and backward RNN's where the former reads the inputs sequence as it is ordered while the latter reads in reverse order. Then, the forward hidden states $(\overrightarrow{h_1}, \dots, \overrightarrow{h_{T_x}})$ and the backward hidden states $(\overleftarrow{h_1}, \dots, \overleftarrow{h_{T_x}})$ are concatenated to derive the annotation for each word.

Just as the other papers dealing with the machine translation tasks, this paper also evaluated the proposed approach on English-to-French translation with WMT 14 dataset. The quantitative analysis shows that in terms of BLEU score, the model outperforms the conventional RNN encoder-decoder models and also performs as good as the conventional phrase-based translation system. Moreover, there was no performance deterioration even with very long sentences, which verifies the validity of the attention mechanism. The qualitative analysis based on alignment and long sentences further confirms the hypotheses that the proposed architecture enables far more reliable translation of long sentences than the standard RNN encoder-decoder model.

2. Discussion

Here I would like to offer 2 discussion points. To begin with, how can the increased model parameters be handled efficiently? Using Bidirectional RNN increases the capacity of the model and this is also compatible with the experiments in the paper, "Sequence to Sequence Learning with Neural Networks", written by Google which encourages to read the input sentence in reverse order. However, without doubt, there occurs inefficiency as the same source sentences are given twice. Not only the number of the parameters but also memory usage is nearly doubled so that the training gets slower. I suggest to use ensemble methods with knowledge distillation method to learn a teacher model in reverse order and the student model in original order. This will be able to show somewhat similar performance with the RNNsearch in a significantly low computational complexity. Next, what about using game-theoretic approach when learning attention mechanism? In Generative Adversarial Network (GAN), they adopted 2 player game when constructing the loss function and showed remarkable performance as a generative model. Neural Machine Translation is somewhat similar in the sense that the common semantic meaning of inputs and outputs is captured. So, when the hidden states of encoder and decoder network are learned through the discriminator network, the better performance would be achieved.