Research Paper: Neural Machine Translation by Jointly Learning to Align and Translate

Authors:

1. Dzmitry Bahdanau
2. KyungHyun Cho
3. Yoshua Bengio

1. Citations

This paper has 9008 citations. Some of which are:

Sequence to Sequence Learning with Neural Networks. Authors:

a. Ilya Sutskever
b. Oriol Vinyals
c. Quoc V. Le

Deep Learning:

a. Yann LeCun
b. Yoshua Bengio
c. Geoffery Hinton

Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

a. Junyoung Chung
b. Caglar Gulcehre
c. Kyunghyun Cho
d. Yoshua Bengio

2. Task

The task that the paper tries to achieve is to use the encoder-decoder machine translation mechanism to translate long sentences with higher probability than the previous cases. Also, in previous situations, the encoder used to encode a sentence in a fixed length vector, but this paper describes a new task which tends to reduce the load on the encoder by encoding the input sequence generated from the input sentence into a sequence of vectors. Finally, for the result, a subset of these vectors is chosen adaptively as will be described in the approach section. This task happens with the translation process.

Since we tend to apply machine translation using recurrent neural network using encoder-decoder technique, this paper builds on the same technique and then advances to correct the issues that this method introduced which is giving issues for translation of long sentences.

3. Data

The dataset used by the people working on this paper was from WMT'14. Specifically they used the English French parallel corpora: Europarl containing 61 Million words, news commentary (5.5 Million) words and two crawled corpora of 90 Million and 272.5 Million words. But they resized the data to 348 Million words using a data selection method by Axelrod et al. (2011).

4. Approach

They have used bi-directional recurrent neural networks for both the encoder and decoder with 1000 hidden units. Also, a single maxout hidden layer is used to compute the conditional probability of each target word. Besides in order to train the model they have used stochastic gradient descent (SGD) algorithm with Adadelta to train the model. Finally, after the model is trained they have used a beam search to translate source sentences with maximum probability.

The forward RNN receives the input sequence and generates the forward hidden states of the RNN. Similarly, the backward RNN generates backward hidden states. These hidden states are used to obtain an annotation for every word. Hence, every target word which is to be translated uses these annotations which contains summaries of both the preceding and following words. As RNN focuses on recent inputs largely, these annotations obtained have more information of the words close to the target word which is to be translated. So, this approach tries to align the target word to the word to be translated with maximum probability using these annotations.

5. Evaluation

They considered the BLEU score for the models to understand it's effectiveness. Their approach outperformed the traditional encoder-decoder using RNN approach. Their approach RNNsearch achieved a BLEU score of 21.50 on all corpora and 31.44 when all the <UNK> removed. The BLEU score was even better when the training on the data was allowed for more time. In that case, the score reached a maximum BLEU score of 28.45 for All words and 36.15 for a corpora in which there were no unknown words.

The conclusion derived from their approach was that their technique focuses on aligning the words during translation to correspond with the source sentence. This alignment is dependent on the weight provided to the annotation of the jth source word for the ith target word. RNN trains the model to maximize the probability of generating better target sentences. Correspondingly to improve the weights associated with the annotations.