Machine Translation Task with seq2seq model and attention

Ankur Debnath
M.tech (EE)
ankurdebnath@iisc.ac.in

Abstract

Machine Translation Task was performed using seq2seq model and different attention mechanisms namely **Additive attention**[1], **Multiplicative attention**[2] and **Scaled Dot attention**[3] using LSTM (in our case GRU). The task was performed on the WMT14 machine translation task[4]. Dataset used for this task was the English-German parallel corpus and Finally, different attentions in encoder and decoder were implemented and evaluated. Then BLEU score(scaled to 0-100) was considered as a metric to evaluate the performance on the test dataset.

1 Introduction

Neural machine translation (NMT) is an approach to machine translation that uses a large artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.The word sequence modeling is typically done using a recurrent neural network (RNN). A bidirectional recurrent neural network, known as an encoder, is used by the neural network to encode a source sentence for a second RNN, known as a decoder, that is used to predict words in the target language.

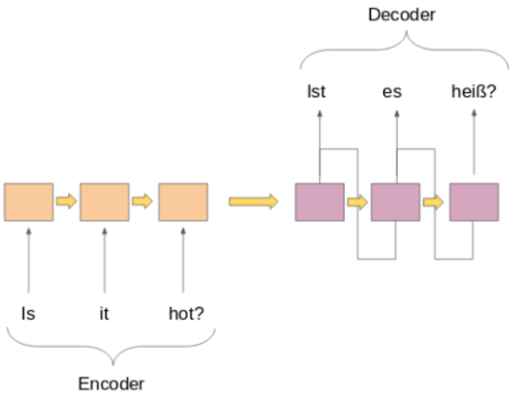
Attention is proposed as a solution to the limitation of the Encoder-Decoder model encoding the input sequence to one fixed length vector from which to decode each output time step. This issue is believed to be more of a problem when decoding long sequences.In this particular task, attentions namely Additive, Multiplicative and Scaled-Dot attentions were implemented.

1.1 Additive attention

The original attention mechanism[1] uses a one-hidden layer feed-forward network to calculate the attention alignment:

$$\text{score}(\mathrm{s}_t, h_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a[s_t; h_i])$$

Figure 1: Encoder-Decoder Model



1.2 Multiplicative attention

Multiplicative attention[2] simplifies the attention operation by calculating the following function:

$$\text{score}(\mathrm{s}_t, h_i) = s_t^\top \mathbf{W}_a h_i$$

Additive and multiplicative attention are similar in complexity, although multiplicative attention is faster and more space-efficient in practice as it can be implemented more efficiently using matrix multiplication. Both variants perform similar for small dimensionality of the decoder states, but additive attention performs better for larger dimensions.

1.3 Scaled-Dot attention

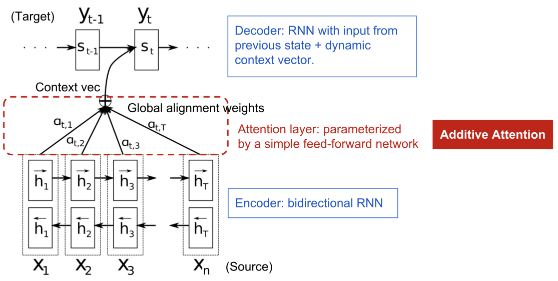
Compared to the standard form of Attention, Scaled Dot-Product Attention is a type of Attention that utilizes Scaled Dot-Product to calculate similarity. The difference is that it has an extra dimension (K dimension) for adjustment that prevents the inner product from becoming too large.

$$\text{score}(\mathrm{s}_t, h_i) = \frac{s_t^\top h_i}{\sqrt{n}}$$

Attention	BLEU Score(in scale of 100)
Additive attention	15
Multiplicative attention	18
Scaled-Dot attention	11

Table 1: BLEU scores with different attentions on ENG-GER dataset

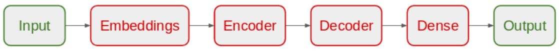
Figure 2: Encoder Decoder with attention



2 Implementation and Model Evaluation

In this task, a GRU model with encoder-decoder and attentions mentioned in Section 1 was implemented and trained on the dataset. Different hyperparameters such as Batch Size, Embedding Size, number of hidden layers and epochs were carefully chosen according to the available computational power and it was made to ensure that the model is not trained too much to overfit on the training data.

Figure 3: Basic model architecture

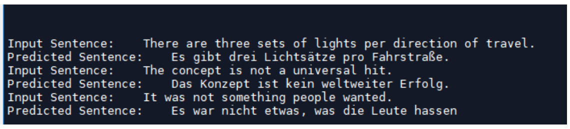


The different tasks that were performed during this experiment involve data preprocessing and cleaning, tokenization, zero paddings, model generation and evaluation. In the evaluation process, BLEU score was used to evaluate the final translations on the entire test dataset and the average was taken of all translations for each type of attentions. The BLEU scores are tabulated for each type of attention in Table 1.

3 Experimental Results

After training the model on the training dataset, BLEU score was calculated on each translation on the dataset and average was considered for evaluation. Results show that multiplicative attention performed the best with BLEU score of 18 and scaled-dot attention performed the worst with BLEU score of 11 among all three attentions. A sample output translation is shown in Fig. 4.

Figure 4: Sample Translations



```
Input Sentence:    There are three sets of lights per direction of travel.
Predicted Sentence:    Es gibt drei Lichtsätze pro Fahrstraße.
Input Sentence:    The concept is not a universal hit.
Predicted Sentence:    Das Konzept ist kein weltweiter Erfolg.
Input Sentence:    It was not something people wanted.
Predicted Sentence:    Es war nicht etwas, was die Leute hassen
```

Conclusion

The experiments verify that when Machine Translation Task is performed using Recurrent Neural Networks (in this case GRU) with the encoder-decoder architecture and attentions, Multiplicative attention seems to perform better than Additive and Scaled-Dot in terms of BLEU scores.It is also to be noted that according to the literature Self attention is the most used attention model for Machine Translation Task, which was not implemented in this experiment.

References

1.Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. ”Neural machine translation by jointly learning to align and translate.” arXiv preprint arXiv:1409.0473 (2014).
2.Luong, Thang, Hieu Pham, and Christopher D. Manning. ”Effective Approaches to Attention-based Neural Machine Translation.” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.
3.Vaswani, Ashish, et al. ”Attention is all you need.” Advances in Neural Information Processing Systems. 2017.
4.WMT14 Machine Translation Task
5.GitHub Link