Improvements of Seq2Seq Model on Neural Machine Translation

Jiaqing Xie [1]

Abstract

In this research, we take the advantage of sequence to sequence(seq2seq) model to perform neural machine translation. While the prototype Recurrent Neural Network(*RNN*) based seq2seq model is able to translate from Chinese to English, its performance is not the optimal. Therefore, we have tried other models such as Gated Recurrent Unit(GRU), Long Short Term Memory(LSTM) and attention based models to achieve better translations. The results show that Bidirectional GRU(*BiGRU*) outperforms Bidirectional LSTM(*BiLSTM*) and attention based GRU to be the best model in this experiments with the optimal BLEU-n value. We also show the attention results in visualization, together with some test results on translations.

1. Experiments

All experiments were implemented on GPU: *RTX 2060 Super*. Therefore the performance does not overpass SOTA.

1.1. Baseline model: seq2seq with RNN

At the first stage, we test on the baseline model, which is a a RNN based uni-directional model with 2 layers and 512 hidden states. Both encoder and decoder are based on RNN where the decoder is connected with a post-processing MLP layer. The optimizer is based on Adam optimizer with a learning rate of 0.0001 and no weight decay or any learning scheduler. We set a training loop of 90000 epochs in stead of initial 50000 epochs to ensure that the loss is more stable. The results show that the lowest loss it has reached is approximately 3.8 after the training session. The average BLEU value is very low, which is equal to 2.81 while the BLEU-4 value is only 0.33 which means that a simple RNN model is not adequate to translate a sentence.

*Equal contribution [1]University of Edinburgh. Correspondence to: to <J.Xie-21@sms.ed.ac.uk>.

Technical report

1.2. RNN substituted with GRU or LSTM-GRU

In the original setting of seq2seq model, a RNN based model is followed by the embedding layer. We consider using different models instead of RNN layers, which are GRU and LSTM specifically. GRU exists in both encoders and decoders. However, LSTM based model only appears in the encoder while decoder is based on GRU(LSTM-GRU). The dropout rate is 0.05. We keep the same hyper-parameter settings as the ones in our baseline model. The results show that GRU model is better than a LSTM-GRU model with a lower negative loglikelihood loss(NLLloss) and a better overall BLEU score, which are shown in figure 1.

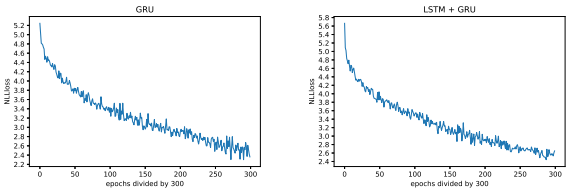

Figure 1. Left: GRU, Right: LSTM + GRU

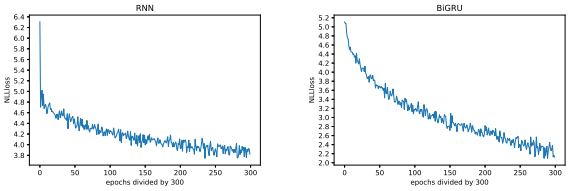

Figure 2. Left: RNN, Right: BiGRU

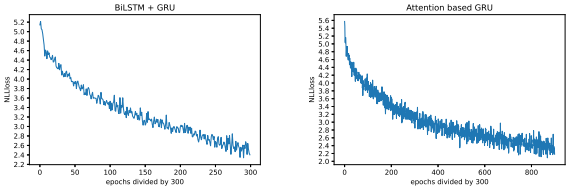

Figure 3. Left: BiLSTM+GRU, Right: Attention GRU

1.3. Bi-directional GRU or LSTM-GRU

Uni-directional model may not pay attention to the following word's back propagation. A better understanding of

Model	BLEU-2	BLEU-3	BLEU-4	BLEU-avg.
RNN	4.16	1.37	0.33	2.81
GRU	25.57	15.52	10.14	21.75
LSTM+GRU	25.05	14.92	9.41	20.98
BiGRU	**30.58**	**19.33**	**12.80**	**25.98**
BiLSTM + GRU	25.09	14.35	8.66	20.41
Attention GRU	22.40	11.75	6.63	17.30

Table 1. BLEU value (in %) between first 4000 reference sentences and their corresponding translations

a sentence requires both forward and backward attention. Therefore we tune the parameter of the GRU and LSTM model: '*bidirectional*' from **False** to **True** to view the performance of BiGRU and BiLSTM+GRU(only difference from part 1.2 is the parameter: '*bidirectional*').

In the pytorch implementation of bi-directional RNN or GRU or LSTM, the number of layers should be multiplied by two to satisfy the back propagation of word information, as well as the layers of initial hidden states. The results show that BiGRU overpasses BiLSTM-GRU to be the best model, with the highest average BLEU value of 25.98(refer to table 1). Its overall NLLloss is the lowest among all models.

1.4. Attention Mechanism

We would like to see if self attention mechanism is helpful to the translation. Therefore we implement the attention GRU where the program is partially given by pytorch tutorial. From the results we can find that the first 4000 sentence's average BLEU value does not perform better than BiGRU model. However, when we implement this model on test dataset, it translates more better intuitively. We provide the translation results based on both models within the submission file. An exmaple of attention output is shown in figure 4.

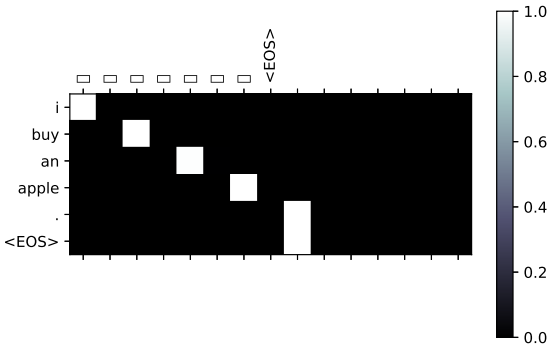

Figure 4. A translation from Chinese version "我买一点苹果" to "i buy an apple ." The blocks are hidden since showing chinese character seems not working on Jupyter Lab

Translation examples on train datasets We also show some examples of translation results with attention GRU based model. In this research, we do not implement Transformer model due to the limitation of GPU resources.

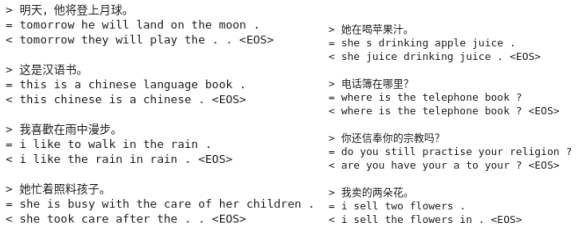

Figure 5. Eight examples of translation results

2. Translation Results

In this section, we show the first 10 translation results in test datasets. Since the maximum input sentence length for attention GRU based model is 15, we filter them in test dataset. For the those of the sentences of which the length is more than 15, we ignore them and the translation result will be just a NIL string.

First ten different sentences within length 15 in Chinese are:
"v? "
"她让我坐在她的身边。"
"这瓶酸奶不含乳糖。"
"我不能帮你了。"
"汤姆不是一个好司机。"
"这个问题有那么简单。"
"他不会说英语也不会说法语。"
"买红酒吗? "
"我又熬夜了。"
"今天天气怎么样? "

The translation results are: what is the ?
she sat on her on side .
the milk isn t to milk .
i can t help you .
tom isn a good good .
there is no problem in the . .
he can speak french speaking .
is the beer ? ?
i ve gone up again .
what s the weather today ?

More results can be checked in the translation.txt file.