Assignment 1 - Word2Vec Skip-gram model

Prasanna Patil
CSA, IISc,
Banaglore
patilk@iisc.ac.in

Abstract

This document contains the final report of assignment 2. The task was to train a NMT model. This report contains charts for various attention configuration and different layers.

1 Task 1

In this task, the Sequence 2 Sequence RNN model (Sutskever et al., 2014) was implemented using pytorch.

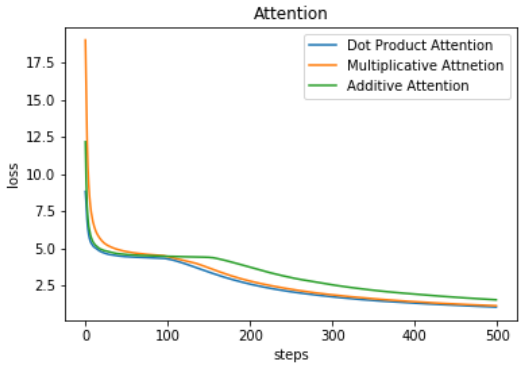
preprocessing of data Following preprocessing procedures were applied to the training data before training the model.

- All the words were converted to lower case and only alphabetic words were considered.
- The default word tokenizer exposed by nltk corpus interface was used to extract words.
- All words appearing less than 5 times in entire corpus were ignored.
- Only sentences with length greater than 2 and less than 30 were considered for training and testing.

The training data was then converted to pair of sentences. All models were then trained on this preprocessed training data.

1.1 Comparison for different Attention mechanism

Batch Size Following plot shows the convergence of Sequence 2 Sequence RNN architecture for different Attention mechanism. Bidirectional LSTM cell was for encoder and decoder. Encoder and decoder contained same number of layers. Number of hidden nodes were set to 256 and embedding dimension is set to 300 for all models. In addition to attention over encoder outputs, attention was also applied to previous decoder outputs as self-attention (also know as intra attention) (Cheng et al., 2016) mechanism. The attention was applied only to output of encoder and decoder last layer.



(a) Figure 1

Figure 1: Figure: Training loss for different attention mechanism.

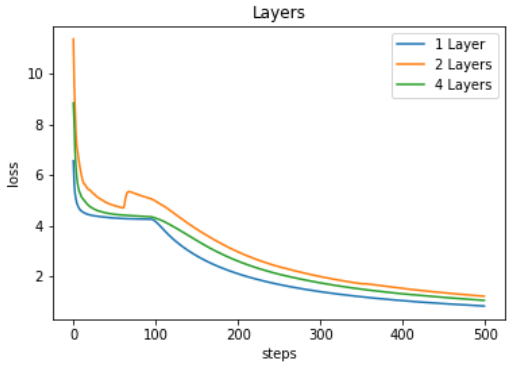
As it can be seen that dot product and multiplicative attention converges faster than additive attention. Dot product attention was scaled (Vaswani et al., 2017) by number of hidden units of LSTM cell.

Number of layers Later on, models were trained for 1, 2 and 4 layers of both encoder and decoder using scaled dot product attention mechanism.

As it can be seen that both 1 and 4 number of layers converge faster than 2 layer model.

1.1.1 Final model

Based upon above result, final model was trained for both English to German and English to Hindi translation. The final model had hidden size 256 and embedding dimension 300. Both models were trained using 2 layers (as a trade off between time



(a) Figure 2

Figure 2: Figure: Training loss for different number of layers.

for training and convergence of model). English to German model was trained on 94k sentences from News Commentary dataset and English to Hindi model was trained on 169k sentences. Both models were trained for 7-8 epochs.

1.2 Evaluation

BLEU score was used to evaluate the final model. However, BLEU score stayed almost near to 0-0.05 during test period. To investigate the reason, the final model was trained on a subset of 100 sentences to overfit the data. On this extremely small dataset the model's performance was acceptable when it had performed 20-25 epochs of all 100 sentences. From this, it can be anticipated that model needed more training to deliver good performance over entire daraset, however, it took 3 hours for one epoch and doing 20+ epochs over one dataset was infeasible without the help of GPU for large amount of time, which was not available.

2 References

References

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.