Neural Machine Translation System: Zero-Shot Translation

Dillon Laird

Introduction

We attempt to implement a neural translation machine similar to that in (Wu et. al. 2016) and use it for zero-shot translations similar to (Johnson et. al. 2016). We make several different design choices because of constraints which we explain in the details section. We then discuss experimental results and conclude with future work we want to implement.

Details

In our problem setting we are given a set source and target sentence pairs (X, Y). Following the notation in (Wu et. al. 2015) we let $X = x_1, x_2, \ldots, x_M$ be the sequence of M symbols in the source sentence and $Y = y_1, y_2, \ldots, y_N$ be the sequence of N symbols in the target sentence. Our encoder is the follow function:

$$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M), s = \text{EncoderRNN}(x_1, \ldots, x_M)$$

Where $\boldsymbol{x}$'s are the encoder hidden states and s is the final hidden state that may include the cell state if the RNN is an LSTM. The first layer of our encoder consists of a bidirectional RNN and higher layers contain residual connections. Our decoder network is then the following function:

$$\boldsymbol{y}_i = \text{DecoderRNN}(s, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M, y_0, \ldots, y_{i-1})$$

Where s is the hidden state from the encoder. We pass the hidden state from the encoder to the decoder which is not done in (Johnson et. al. 2016). The decoder is also a multilayered RNN which residual connections between layers. The output of the first layer is used to calculate attention scores, the attention scores are then concatenated with the hidden state from each layer, projected, and passed up as well as over time. We experimented with two types of attention, one used in (Luong et. al. 2015) with a simple dot product to calcualte the scores and the other proposed in (Wu et. al. 2016) using a feed forward network with one hidden layer.

For multilingual tasks we attach a special translation token to the end of each sentence which designates the language of the target sentence similar to (Johnson et. al. 2016).

One item we did not have time to implement was the reward optimization in (Wu et. al. 2016) where the loss function was augmented with a special reward function that computes the per-sentence score.

For our preprocessing step we use the wordpiece model used in (Sennrich et. al. 2015) and we use a modified version of their code to create our own wordpieces that include the translation tokens.

Experiments

We ran 4 different models on 7.6 million French to English and English to German pairs. The models (with the exception of one) made it over the entire dataset at least once.

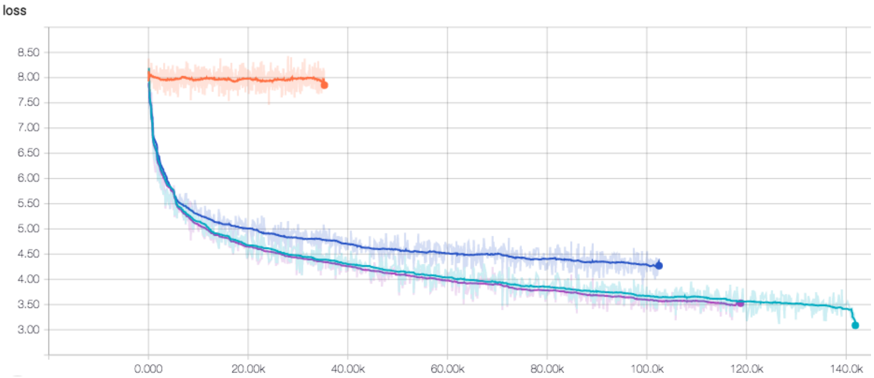

Figure 1: Training Loss

Color	Number of Layers	Hidden Units	Attention	Optimizer	Learning Rate
Orange	2	512	Luong	Adam	0.01
Blue	2	512	Feed Forward	SGD	1.0
Teal	2	512	Luong	SGD	1.0
Purple	3	512	Luong	SGD	1.0

Interestingly the orange graph, the model using the Adam optimizer, did not seem to be converging so we ended that run early. The other 3 runs ran for about 1 day and 19 hours. The attention in Luong seemed to work much better than a feed forward network used in (Wu et. al. 2016) acheiving not only a lower training error but also running faster. Luong's attention with SGD and

1.0 learning rate seemed to work the best. You can see the model with 3 hidden layers has a slightly lower training loss but is alos running slower.

Unfortunately we were not able to achieve able to achieve a BLEU score for our model with best loss. We tested this for both zero shot, French to German, and a test set for French, English to English, German. We tried with both a greedy search and a beam search but got a BLEU score of 0 for both.

References

[Johnson et. al. 2016] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Kirkun, Yonghui Wu, Zhifeng Chen and Nikhil Thorat. 2016. Google's Multilingual Neural Machine Tranlsation System: Enabling Zero-Shot Tranlsation.

[Wu et. al. 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le and Mohammad Norouzi. 2016. Google's Neural Machine Tranlsation System: Bridging the Gap between Humand and Machine Translation.

[Sennrich et. al. 2015] Rico Sennrich, Barry Haddow and Alexandra Birch. 2015. Neural Machine Translation of Rare Words with Subword Units.

[Luong et. al. 2015] Minh-Than Luong, Hieu Pham, Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation.