Implementation of Neural Machine Translation System

Zhiming Guo zg758@nyu.edu
Yueqiu Sun ys3202@nyu.edu
Nan Su ns3783@nyu.edu
Yu Xiong yx1201@nyu.edu

Abstract

The encoder-decoder model has been used in most of the Neural Translation Systems in recent years. With the discovery of attention mechanism, the translation performance is greatly improved. In this project, we aim to compare the translation quality between simple RNN based model, Luong attention RNN based model, and self-attention based model for Chinese to English translation (Ch2En) and Vietnamese to English translation (Vi2En), respectively. It turns out that after equipping attention into the encoder-decoder model, BLEU can be improved at least 90% for both Ch2En and Vi2EN. Also, we demonstrate that LSTM with Luong attention sequence-to-sequence model is the best model that can achieve 12.75 BLEU on character-level Chinese and 21.30 BLEU on Vietnamese. All of our code is available at Github.

1 Introduction

Neural Machine Translation (NMT) is an end to end approach that takes the source sentence as input and outputs the probability distribution to the likely translations. The traditional count-based n-gram language model suffers from the data sparsity issue and cannot generalize well to out-of-box data (Cho, 2015). By compressing the words into low-dimension embeddings that contain the semantic information of the word, NMT learns to generalize to unseen sentences. Words that are used in similar ways are projected to nearby points in the embedding space.

Encoder-decoder architecture that deploys two recurrent neural networks (Sutskever et al., 2014) and attention mechanisms (Bahdanau et al., 2014) is shown to achieve excellent performance on machine translation task.

In this work, we present a complete analysis of the performance of different neural translation models. We report the performance of each model with different hyper-parameters and provide insight into how to design better architecture in the future.

2 Models

2.1 Simple encoder-decoder

A simple Encoder-decoder network or Sequence to Sequence (Sutskever et al., 2014) model is composed of two Recurrent Neural Networks (RNNs) named the encoder and decoder. The encoder takes in the source sentence $X = [X_1, X_2, ..., X_T]$ in the translation task and outputs a single vector that summarizes this sentence. This process is finished by first establishing an initial hidden state, and at each time step, updating the hidden state h_i using some function of the concatenated affine transformation of the current input word(embedding) X_i and previous hidden state h_{i-1}. And at the last time step, we obtain a context vector h_T which can be used as the input of the decoder.

When decoding, an RNN sequentially takes an input token, initialized by start-of-sentence token, along with the hidden state, initialized by the context vector h_T, to predict the next token in the target language.

2.2 Attention

A problem with the naive Encoder-Decoder model is that the encoder squeezes each input sentence of varied-length to a fixed-length representation in the last hidden state, where sometimes long sentences could not be summarized well in a fixed-length representation. Attention is an improvement to the model that allows the decoder to "attend" over different words in the input sequence by scoring for each state in encoders.

The attention mechanism that we used is Luong Attention (Luong et al., 2015). Specifically, we use its global attention, which is summarized in Figure 1. The original attention has three methods to define the scoring function. In our case, the simple dot product method is used. As Figure 1 and the following equations show, at each decoder time step k, the scoring function assign a weight (normalized score) a_{ki} to each encoder output, S_i, by evaluating the relevance to this decoding time step. Then, the context vector, c_k, computed by the weighted average of each encoder output, is used to generate the next token.

$$\text{scores}(T_k, S_i) = T_k^T S_i \tag{1}$$

$$a_{ki} = \text{align}(T_k, S_i) = \frac{\exp(\text{score}(T_k, S_i)}{\sum_{i'} \exp(\text{score}(T_k, S_{i'})} \tag{2}$$

$$c_k = \sum_{i=1}^{T} a_{ki} S_i \tag{3}$$

2.3 Self-attention encoder

In the self-attention encoder model, we stick to the RNN decoder and replace the RNN encoder with 6 identical layers with each layer has a multi-head self-attention mechanism and a simple positionwise fully connected feed-forward network.

Self-attention mechanism allows model updating the word embedding by looking at other words in the sentence (Alammar,). Each word is assigned a weight which reflects how relative is the word to the word the model is updating. The word is updated using the weight and the embedding of all the words in the sentence.

The "Scaled Dot-Product Attention" used in the self-attention encoder is as follows,

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}}V) \tag{4}$$

Where Q, K, V represent the Query, Key, and Value matrices which are calculated by multiplying our embeddings by their respective trainable weight matrix. d_k is the dimension of the Query and Key.

The above attention mechanism can be further refined by having multiple sets of Query, Key, and Value matrices and calculate multiple representations of the updated embedding.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{5}$$

where head_i denotes the output representation calculated using the i^{th} set of weight matrix. W^O is a trainable parameter matrix.

3 Experimental Setup

3.1 Datasets and Data Prepossessing

The data for this project consists of translation pairs from Chinese to English and Vietnamese to English. The Sentence in the data set was pre-tokenized into phrases by the nature of each language. We also prepared a character-level Chinese for comparison. The number of pairs in each data set are summarized in Table 1.

	Pair Size	
	Chi-Eng	Viet-Eng
Train	213378	128908
Validation	1262	1269
Test	1398	1554

Table 1: Original Dataset

In order to maintain a feasible training time, we filtered out around 15% of total pairs for each language by trimming off long sentences. We kept the maximum sentences length of validation and test set as the same as that of the training set. Fixed vocabulary is used for each language. In Chinese to English translation, 4800 of the most frequent characters / 65,000 of the most frequent phrases are included in the source language and 50,000 of the most frequent words are included in the target language. In Vietnamese to English translation, 19,000 most frequent

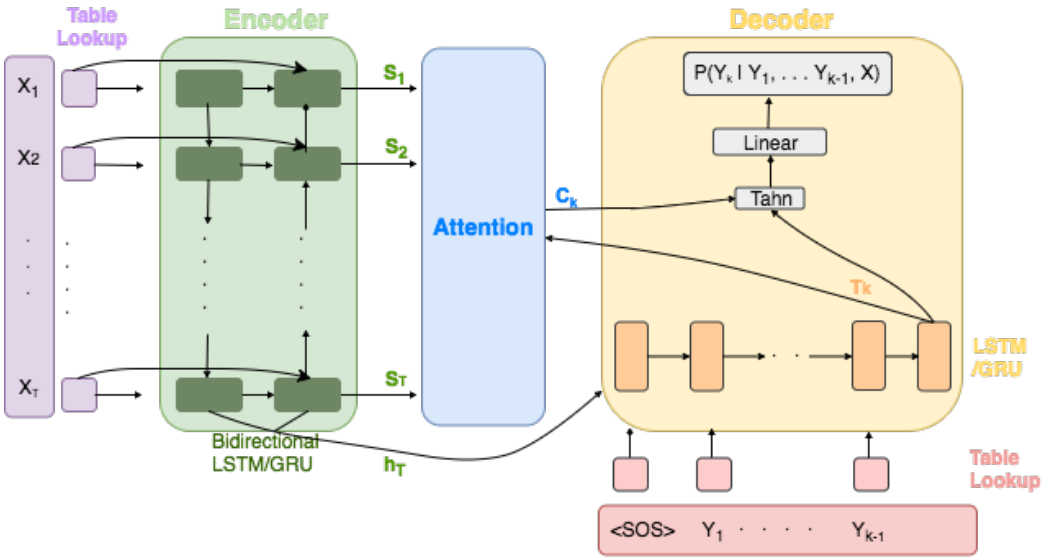

Figure 1: Seq2Se2 With Luong et al. Attention Model

phrases and 22,000 of the most frequent words were used for source and target language respectively. Every out-of-vocabulary word was replaced with a special "UNK" token. "PAD", "SOS" and "EOS" were added to the vocabularies as tokens for further training purpose. We did not apply any other special prepossessing other than removing some uncommon special symbols.

3.2 Training Setup and Software

Our experiments were running on Google Collaboratory with the Tesla K80 GPU based on Pytorch. We use a batch size of 64 and teach forcing ratio of 1 in decoder to ensure the speed and quality of training. Adam optimizer was used and the learning rate was initialized to 0.0001 and 0.0005 respectively for the encoder and decoder. Whether it decays depends on the model that we use. For convenience, we set the hidden size and embedding size to be equal, and the number of layers in both encoder and decoder are the same.

Each experiment is running for a maximum of 20 epochs with different initialization. We recorded training loss and validation BLEU of our model every 100 steps and saved the checkpoints if the model outperformed previous models.

3.3 Evaluation Metric

BLEU (bilingual evaluation understudy) is the most widely accepted evaluation metric in translation. BLEU computes the geometric mean of the modified n-gram precision multiplied by brevity penalty. Specifically, we use ScareBLEU(Post, 2018) to evaluate performance.

3.4 Inference Strategy

We used Greedy search and Beam search during the evaluation process. Greedy search predicts next token with the highest probability based on only one current candidate token. Greedy search is fast but it only considers the optimal solution at current step (Cho, 2015). Beam search considers the k candidates with top accumulated log-likelihood until current time step. Greedy search is a special case of beam search when beam width, k, equal to 1. Since beam search is more time-consuming than greedy search, we use greedy search during training and evaluation state. After training, beam width was tuned to find the highest BLEU on validation set.

4 Experiments and Results

4.1 Baseline Model

In our experiment, simple encoder-decoder model with GRU units was used as the baseline model. The highest BLEU is 4.54 for Chinese-English transla-

tion and 9.82 for the Vietnamese-English translation.

4.2 Model Evaluation and Comparison

Given the baseline model, important changes and improvements are made and listed below:

- Luong Attention:

  Enhancing the model by adding Attention mechanism. Tuning the hyper-parameters with respect to hidden size and number of layers mainly. For Chinese, we attempt to compare character-level corpus and phrase-level corpus.

- Self Attention:

  Replacing the RNN encoder with a stack of 6 identical layers which has two sub-layers 1.a multi-head self-attention mechanism 2.a simple, position-wise fully connected feed-forward network. Tuning the hyper-parameters with respect to the number of heads in multi-head attention module.

4.2.1 Chinese

For Chinese-English translation, the character-level model outperformed the phrase-level model. Table 2 shows that with GRU unit and the same parameters, the character-level model is at least 1 score higher than the phrase-level model.

Table 2 also indicates that replacing the GRU cells with LSTM cells helps improve the model slightly. With hidden layer of 512, number of hidden layers of 2 and GRU module, the model achieves 12.54 BLEU after greedy search, while the model with LSTM units achieves 12.70 BLEU. Figure 2 compares the training loss with respect to the best LSTM, the best GRU, and the best self-attention model over 20 epochs. We found that these models have very similar training speed.

Hidden layer size has little impact on the results during our experiment. From Table 2 we observe that the BLEU even drops from 12.90 to 12.44 after increasing the hidden layer size from 512 to 1024 for the character-level model.

Generally, deeper encoder and decoder should perform better than the shallower one. For the Chinese dataset, it turns out that increasing number of layers from 2 to 4 enhances the model performance by approximately 0.2 point.

Table 2 also shows the BLEU of each model after performing beam-search. It can be observed that the beam-search technique is very useful to improve the model performance on the validation set. In our experiment, the beam size is also a hyper-parameter. We explore the effect of varying beam widths from 1 to 14, 1 stands for greedy search. In our case, a beam size of 6 gives us the best model of all, with BLEU rising to 13.33.

Table 3 shows the BLEU score of self-attention models with different number of heads. For self-attention model, beam size of 4 and head number of 8 offer the best performance. Multiple heads in multi-head attention module enable the model to focus on different part of the sentence and improve the performance of the model.

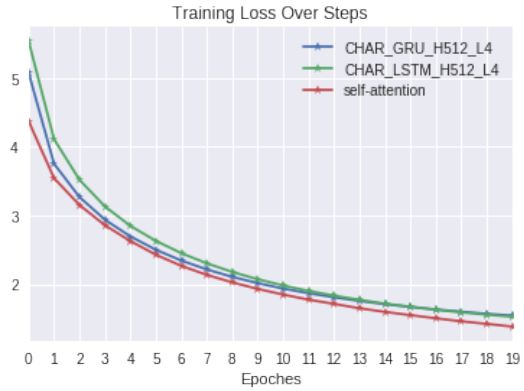

Figure 2: Training loss of GRU, LSTM and self-attention models for 20 epochs, Chinese corpus

Model Type	Chinese Greedy BLEU	Chinese Beam-Search BLEU
Phra-GRU + H512 + L1	9.30	10.55
Char-GRU + H512 + L1	10.51	11.87
Char-GRU + H512 + L2	12.54	12.83
Char-GRU + H512 + L4	12.68	13.10
Char-LSTM + H512 + L2	12.70	12.91
Char-LSTM + H512 + L4	12.90	**13.33**
Char-LSTM + H1024 + L2	12.44	12.85

Table 2: BLEU on Chinese-English validation set. H = hidden size. L = the number of layers. Char = character-level

Model Type	Chinese Greedy BLEU	Chinese Beam-Search BLEU
Self-Attention + h1	9.45	11.06
Self-Attention + h4	10.21	11.83
Self-Attention + h8	10.75	**12.28**

Table 3: BLEU on Chinese-English validation set. h = the number of heads in multi-head attention.

4.2.2 Vietnamese

For the Vietnamese-English dataset, the LSTM cells outperformed GRU units in our experiment. Table 4 shows that when the dimension of the hidden layer and the number of layers are the same (512 and 2, respectively), the model with LSTM units achieves 23.56 BLEU after greedy search while the model with GRU units has 21.55 BLEU. Also, it can be observed from Figure 3 that the models with GRU and LSTM units have very similar training speed. Therefore we only perform parameters-tuning on the LSTM model.

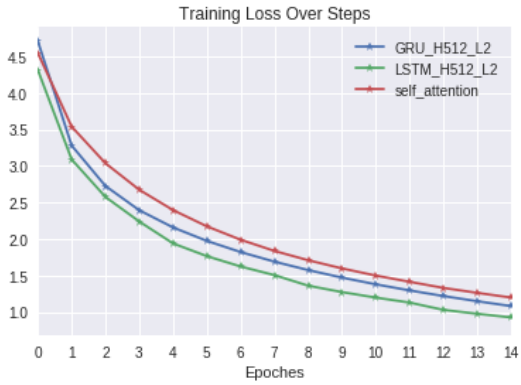

Figure 3: Training loss of both GRU and LSTM models for 15 epochs, Vietnamese corpus

We expected more hidden layer parameters to result in better BLEU on our validation set. As a matter of fact, when the number of layers is set to 2, the BLEU ascends from 22.95 to 24.12 when the hidden layer size increases from 256 to 1024.

For the Vietnamese dataset, we explore the effect of both encoder and decoder depth up to 4 layers. As expected, the BLEU of the model with LSTM units and hidden layer size of 512 jumps from 21.01 to 24.23 when the number of layers rises from 1 to 4. Although the score changes slightly after the number of layers is larger than 2.

Table 4 also shows the BLEU score of each LSTM model after performing beam-search. It can be observed that the beam-search technique helps improve the model performance on the validation set. Figure 4 shows the effect of varying beam widths from 1 to 14, 1 stands for greedy search. A beam size of 7 offers us the best model of all, with BLEU raising to 25.73.

Table 5 shows the BLEU of self-attention model

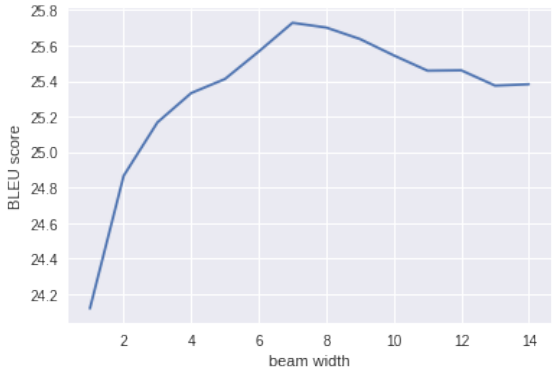

Figure 4: BLEU on validation set, varying beam size

with different number of heads. For the best self-attention model, beam size = 6 offers the best performance. Multiple heads in multi-head attention module enable the model to focus on different parts of the sentence and improve the performance of the model.

Model Type	Vietnamese Greedy BLEU	Vietnamese Beam-search BLEU
GRU + H512 + L2	21.55	N/A
LSTM + H512 + L1	21.01	22.89
LSTM + H512 + L2	23.56	24.64
LSTM + H512 + L4	24.23	25.12
LSTM + H256 + L2	22.95	24.85
LSTM + H1024 + L2	24.12	**25.73**

Table 4: BLEU on the Vietnamese-English validation set based on Luong attention.

Model Type	Vietnamese Greedy BLEU	Vietnamese Beam-Search BLEU
Self-Attention + h1	14.02	16.16
Self-Attention + h4	14.90	17.48
Self-Attention + h8	15.44	**18.15**

Table 5: BLEU on the Vietnamese-English validation set based on self-attention.

4.3 Test Performance and Translation Showcase

We evaluated our best model on the test dataset for Chinese-English and Vietnamese-English dataset, respectively. Table 6 shows that the test BLEU is smaller than what we obtained for Chinese and Vietnamese validation set respectively, as expected.

	Test BLEU
Chi-Eng Char-LSTM + H512 + L4	12.75
Vi-Eng LSTM + H1024 + L2	21.30

Table 6: Best model's BLEU on the two language test set

Table 7 lists some translation examples for Chinese using the best model: The first translated sen-

tence exactly match the target sentence. The second translated sentence missed the word "table". The last translated sentence translated "Yalu", the name of a river, as "green duck". In fact, in Chinese, "Ya" means "Duck" and "Lu" means "Green". However, due to the nature of Chinese, name that composed by words has its unique translation other than translating by semantic meaning of each word. In this case, it is very hard for character-level machine to translate the Chinese name into English.

Source sentence	Predicted sentence	Translation sentence
"我并不是在帮助你，" 他回答	" I'm not helping you" he said .	" I'm not helping you" he said .
一些光流了出来，到了桌子上，所以你开始看到这些波纹。	And some of these lights came out and you're beginning to see these waves .	Some of the light leaks goes on the table and you start seeing these ripples of waves .
这是鸭绿江，作为着中国和朝鲜边界的一部分。	It's a green duck . It's part of the Chinese border .	This is the Yalu River which serves as a part of the border between North Korea and China .

Table 7: Chinese Translation Examples

5 Conclusion

In this project, we implemented three kinds of model: the baseline Seq2Seq model, the Luong attention model, and the self-attention model. In general, Luong attention outperformed the baseline model and the self-attention model. Through the multiple training experiments, we have some practical findings to share.

A character-level tokenization helps to increase the BLEU of Chinese-English dataset. Replacing the GRU unit by LSTM cells further enhanced the model performance. The model with hidden dimension of 512 and 4 layers outperformed other models, achieving 13.33 BLEU after fine-tuned beam-search on the validation set and 12.75 BLEU on the test set. Vietnamese to English translation has similar results. The model with LSTM cells consistently outperformed the model with GRU cells. And the model with hidden size of 1024 and 2 layers depth performed the best, producing 25.74 BLEU on validation set and 21.3 on test set after beam-search.

Generally, the results of Vietnamese translation was better than that of Chinese-English translation. One possible reason is that Chinese is a very complicated language, even the best Chinese word segmentation tools, like jieba, could hurt the meaning of any Chinese sentence. In the meanwhile, the character-level modeling sometimes cannot capture the meaning of consecutive words, just like the translation showcase above.

6 Limitation and Future Works

Although we have tried several mechanisms and fine-tuned important hyper-parameters, our model is not robust in many ways:

1. Even if we trimmed around 15% of the longest sentences, this translation model was still not able to capture all information of the source sentences if the sentences are relatively long.

2. RNNs handled dependencies by being recurrent, which means RNNs can only process tokens one by one. The sequential nature of RNNs also makes it slow for training with large dataset.

3. Our model performs much worse in translating English from Chinese due to the gap between the two different writing systems.

For future work, a more advanced encoder-decoder mechanism such as transformer could be used to address the first two issues listed. Transformer gets rid of the sequential operations from RNNs and learns to model dependencies using Attention mechanism directly (Vaswani et al., 2017). Moreover, transformer processes all the tokens in a parallel fashion, which is able to take advantage of modern fast computing devices such as GPUs. In order to improve the translation quality of cross-system languages, more linguistic research need to be done to find the linguistic unit that conveys its meaning so that the machine could better learn the semantic correspondences between languages.

7 Contributions

All members conceived the model architecture and implementation. Specifically, Zhiming built the GRU and beam search; Yu implemented LSTM; Nan pre-processed Chinese corpus and implemented Luong attention; Yueqiu established self-attention based model.

References

[Alammar] Jay Alammar. The illustrated transformer. http://jalammar.github.io/illustrated-transformer/.

[Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Cho2015] Kyunghyun Cho. 2015. Natural language understanding with distributed representation. *arXiv preprint arXiv:1511.07916*.

[Luong et al.2015] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

[Post2018] Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

[Sutskever et al.2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

[Vaswani et al.2017] 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.