Report for Programming Assignment-1

Yogesh Tripathi
(CS16B044)

September 25, 2019

- ☑ I have read all the instruction carefully and followed them to my best ability.
- ☑ I have written the name, roll no in report.
- ☑ Run sanity_check.sh.
- ☑ I have not included unnecessary text, pages, logos in the assignment.
- ☑ I have not copied anything for this assignment.

For consolidation of all codes, trained and saved models, please check the following drive link.

Question 1

Implementation(using the official implementation in [1], [2]) and script to run on best parameters is attached in submission. The best performing parameters and architecture of Transformer is as follows.

Question 2

The BLEU Score obtained on public_test.en with the above mentioned parameters is **18.01**(case-insensitive) and **17.62**(case-sensitive).

Question 3

The BLEU Score obtained on private_test.en with the above mentioned parameters is **14.34**(case-insensitive) and **14.03**(case-sensitive).

Parameter	Best Value
Batch size	2048
Hidden size for feedforward network	2048
Layer Dropout probability	0.1
Attention dropout probability	0.1
ReLU dropout probability	0.1
Learning rate	2.0
Learning rate Decay rate	1.0
Learning rate warmup steps	16000
Adam optimizer parameters($\beta_1, \beta_2, \epsilon$)	0.9,0.997,10^{-9}

Table 1: Best performing hyperparameters with $d_{model} = 512, d_k = 128$, 4 attention heads and 2 layers

Question 4

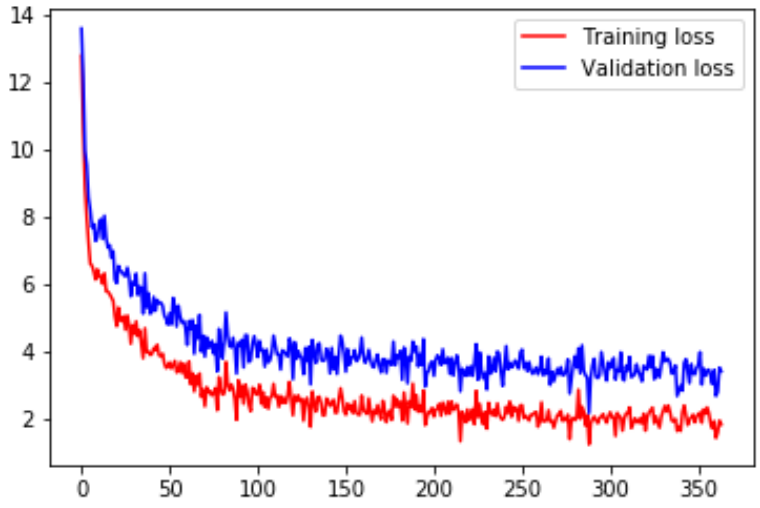


Figure 1: Training and Validation log-likelihood over iterations

Question 5

The BLEU score on validation data for given hyperparameters setting. The training was done for 10 epochs which is same as that for best performing model. d_{model}, d_k, number of attention heads and number of layers was kept same as that of best performing model.

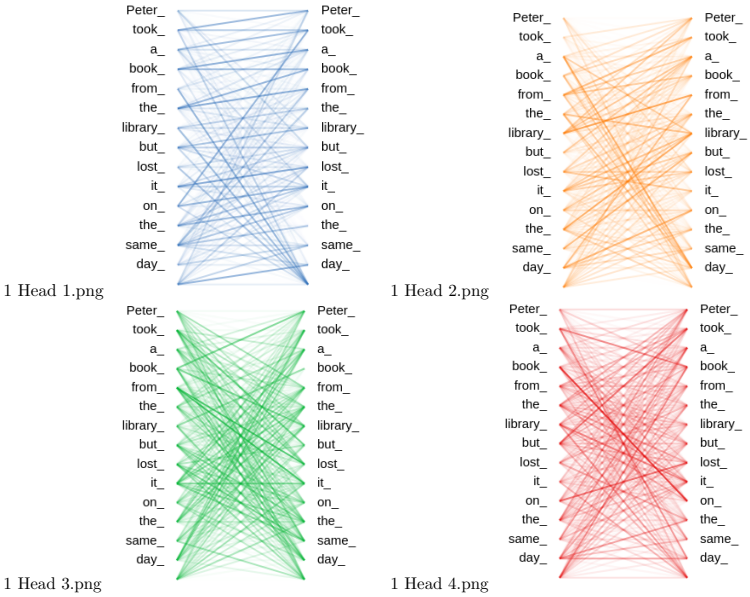
Hyperparameter setting	BLEU Score on Validation data
No dropout Layer Dropout probability = 0 Attention dropout probability = 0 ReLU dropout probability = 0	0.04
Changing hidden layer size Hidden size for feedforward network = 1024	7.80
Lower Learning rate Learning rate = 0.01 Learning rate Decay rate = 1.0 Learning rate warmup steps = 20000 Adam optimizer parameters($\beta_1, \beta_2, \epsilon$) = 0.999,0.999,$10^{-9}$	5.84

Table 2: BLEU score on validation data for different hyperparameter settings

We find that removing dropout has the most detrimental effect on the model. Lowering of learning rate and reducing the hidden layer size to half of it's value also result in reduction of BLEU score.

Question 6

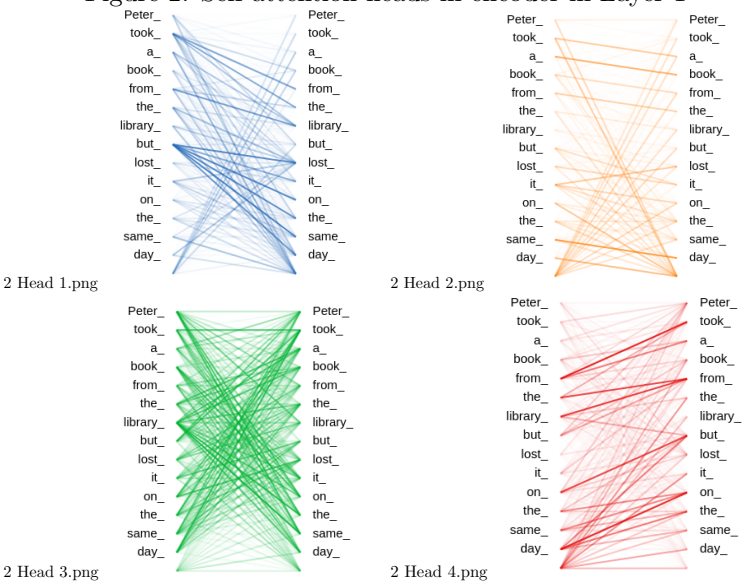
Given below are attention plots in Layer 1 and 2 for each attention head(used [3] for visualization) for the following sentence-"Peter took a book from the library but lost it on the same day".(Refer Figure 2 and Figure 3)



1 Head 1.png 1 Head 2.png

1 Head 3.png 1 Head 4.png

Figure 2: Self-attention heads in encoder in Layer 1



2 Head 1.png 2 Head 2.png

2 Head 3.png 2 Head 4.png

Figure 3: Self-attention heads in encoder in Layer 2

Question 7

The BLEU score on validation data obtained when all 8 heads are in one layer is **5.99**(case-insensitive) and **5.84**(case-sensitive). The reduction in BLEU score can be explained as follows. When there are multiple layers, the representation generated from each layer takes in more refined representation as input and hence produces more refined representation as output. In this case, we have same number of attention heads as in best model, but because refinement of word representation occurs only once, overall performance of model becomes worse.

Question 8

Based on the attention plots in answer for Question 6, we can make infer some characteristics about attention heads as follows:

- Heads in layer 1 mostly attend to neighboring words, especially Head 1. In other heads, we see some words attending to distant words, however no attribute can be as such associated.
- Head 3 in Layer 2 has very special attention distribution, as if it divides the sentence into two halves at 'but' and one half attends to other half.
- Head 1 in Layer 2 shows very interesting characteristic that the word 'but' attends to "lost it on the same day" which is exactly what 'but' is placed in sentence for.

Question 9

When we tie the weights for Key and Value, the BLEU score obtained on validation data is **8.69**(case-insensitive) and **8.50**(case-sensitive). All other model parameters are kept unchanged. This shows that having separate set of weights for Key and Value representations gives better word representations.

The code for Seq2Seq Model with specified architecture- Bidirectional LSTM, Hidden size=512 was taken from [4] which is based on paper [5] which extensively studies Seq2Seq models. The running time for 1 epoch for Transformer with architecture specified in Table 1 and Seq2Seq model on a single GPU(run on Google colab) is summarized below.

Model	Time for 1 epoch(in hours)
Seq2Seq	$\sim$11
Transformer	$\sim$8

Table 3: Comparing training times for one epoch on Seq2Seq and Transformer

References

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

[2] Official code for paper- Attention is all you Need.

[3] Official tensor2tensor code for transformer attention visualization.

[4] Implementation of seq2seq models from paper-"Massive Exploration of Neural Machine Translation Architectures" .

[5] Denny Britz, Anna Goldie, Thang Luong, and Quoc Le. Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints*, March 2017.