ECE 661: Homework #3

Understand and Implement Sequence Models

NetID: wh162

1. True/False Questions:
   1. Problem 1.1: True, as we know from the structure of a Transformer, the basic block is self-attention and feed-forward. Whereas self-attention requires queries, values, and keys.
   2. Problem 1.2: False, in self-attention, the key, value, and queries all came from the same source, where all three of them are linear transformation used for all
   3. Problem 1.3: False, not only after the attention matrix, but we also have to done executing the feed-forward module before we can apply any normalization.

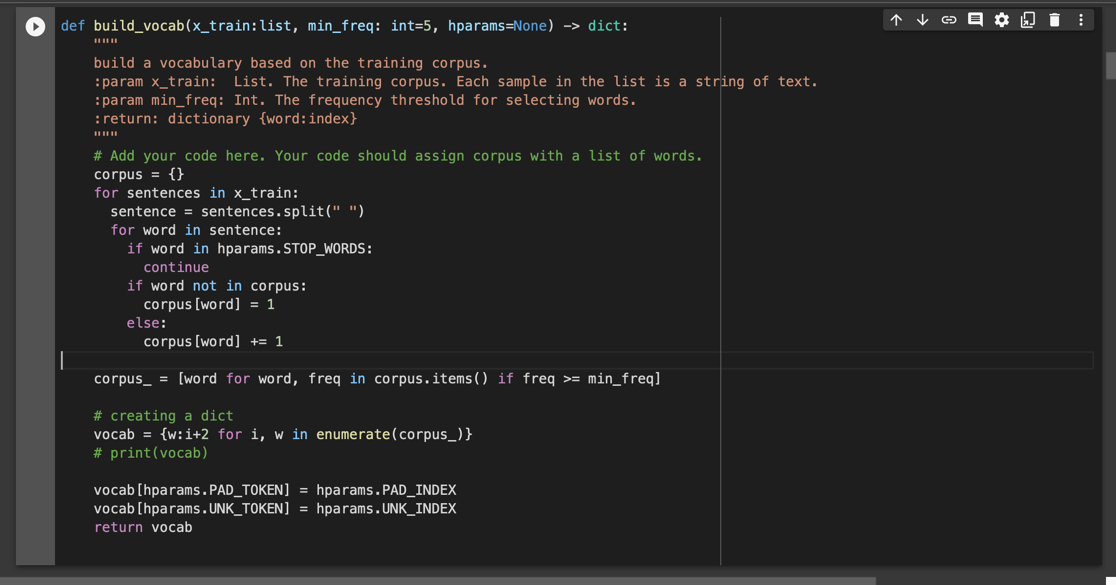
* 1. Problem 1.4: False, from Lec 10 P26, we know that GPT is autoregressive while BRET is autoencoding.
  2. Problem 1.5: False, both has encoder and decoder.
  3. Problem 1.6: True, refer to Lec 10 P 27
  4. Problem 1.7: False, both GPT and BERT are zero-shot learner
  5. Problem 1.8: True, gradient clipping helps gradient descent to have a reasonable behavior even is the loss landscape in the model is irregular.
  6. Problem 1.9: False, word embedding can contain both positive and negative values.
  7. Problem 1.10: False, the sum of the weight is not guaranteed to be 1

1. Lab 1: Recurrent Neural Network for sentiment analysis
   1. Problem 2.1:

Text

Description automatically generated

* 1. Problem 2.2:



* 1. Problem 2.3:

Text

Description automatically generated

* 1. Problem 2.4:

Text

Description automatically generated

* 1. Problem 2.5:
     1. 2.5.1:

Text

Description automatically generated

* + 1. 2.5.2:

Text

Description automatically generated

* 1. Problem 2.6:

Text

Description automatically generated

No, the result doesn’t meet my expectation. I thought the training accuracy and validation accuracy would increase as the epoch goes on. However, the accuracy stayed the same even after 5 epochs. My guess is that SGD optimizer got itself into a local min that result in this way.

* 1. Problem 2.7:
     1. 2.7.1:

Text

Description automatically generated

* + 1. 2.7.2:

Text

Description automatically generated

* 1. Problem 2.8:

Text

Description automatically generated

We can see that GRU has a slightly higher accuracy than LSTM, but GRU also has the same problem which the accuracy was stuck in a steady value and didn’t increase.

1. Lab 2: Training and improving Recurrent Neural Network
   1. Problem 3.1:

Text

Description automatically generated

Text

Description automatically generated

In the experiment we can notice that changing the optimizer result in a very obvious change in the accuracy. The reason why Adam optimizer gives the best accuracy is because Adam adjust its learning rate in corresponding to the training status and is better with dealing with gradient vanish issues.

* 1. Problem 3.2:

Text

Description automatically generated

Text

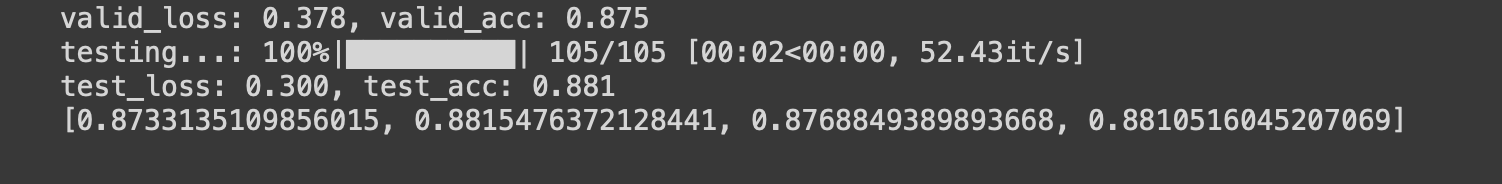
Description automatically generated

The experiment in GRU results the same comparing to LSTM. We can see that Adam has the highest accuracy and in GRU, the result is slightly higher than LSTM.

* 1. Problem 3.3:

Text

Description automatically generated



We can see that when the number of layer is 2, the accuracy is the highest.

* 1. Problem 3.4:

Text

Description automatically generated

Text

Description automatically generated

We can see that when the hidden units is 200, the accuracy is the highest.

* 1. Problem 3.5:

Text

Description automatically generated

Text

Description automatically generated

We can see that when the embedded dimension is 1, the accuracy is the highest.

* 1. Problem 3.6:

Text

Description automatically generated

Text

Description automatically generated

We can see that when the number of layers is 1, the hidden units is 150, and embedded dimension is 1, the accuracy is the highest.

* 1. Problem 3.7:

Text

Description automatically generated

Text

Description automatically generated

We can see that although we enabled the bidirectional parameter, the accuracy didn’t went up, it even decreases a bit.