Assignment 4 Writeup

DO NOT TAG

Name: Zefeng Dong
GT Email: zdong95@gatech.edu

Seq2Seq Results

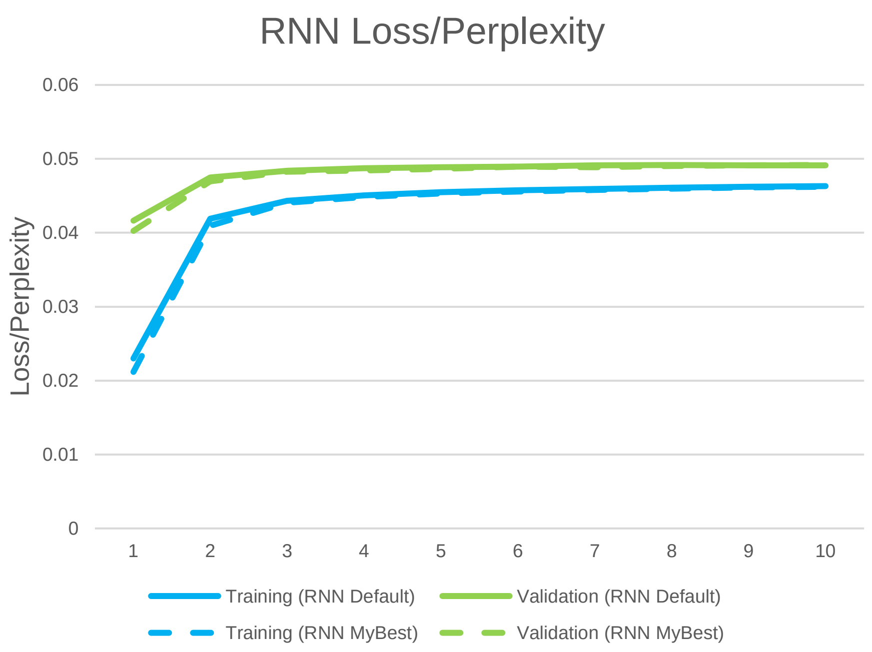
Put your results from training before and after hyperparameter tuning here.

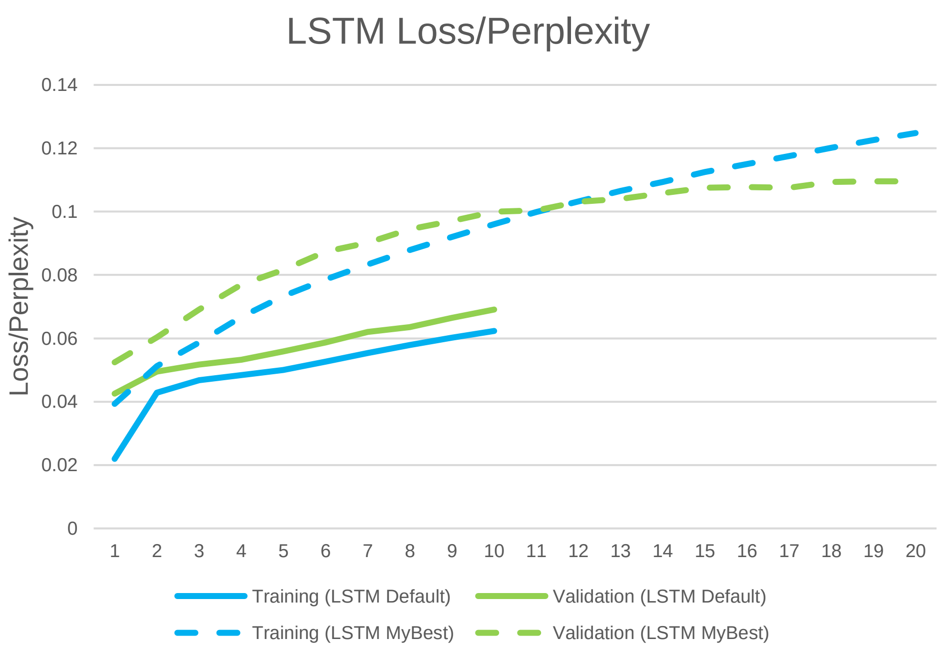
Results for default configuration using RNN		Results for default Configuration Using LSTM	
Training Loss	4.597513	Training Loss	4.213440
Training Perplexity	99.237178	Training Perplexity	67.588641
Validation Loss	4.522235	Validation Loss	4.077815
Validation Perplexity	92.041116	Validation Perplexity	59.016381
Result for your Best Model using RNN after hyperparameter tuning		**Resut for your Best Model using LSTM after hyperparameter tuning**	
Training Loss	4.600179	Training Loss	3.264268
Training Perplexity	99.502144	Training Perplexity	26.160952
Validation Loss	4.521154	Validation Loss	3.449836
Validation Perplexity	91.941675	Validation Perplexity	31.495220
Your best model configuration for RNN after hyperparameter tuning		**Your best model configuration for LSTM after hyperparameter tuning**	
encoder_emb_size = 32, encoder_hidden_size = 64, encoder_dropout = 0.3, decoder_emb_size = 32, decoder_hidden_size = 64, decoder_dropout = 0.3, learning_rate = 1e-3, model_type = "RNN“, EPOCHS = 10		encoder_emb_size = 32, encoder_hidden_size = 96, encoder_dropout = 0.2, decoder_emb_size = 32, decoder_hidden_size = 96, decoder_dropout = 0.2, learning_rate = 5e-3, EPOCHS = 20	

Table 1

Seq2Seq Curves

Put the plots for loss/perplexity curves (training & validation) for your configuration with default setting and for your best model here.





Seq2Seq Explanation

Explain what you did here and why you did it to improve your model performance. Compare and explain the differences when using LSTM vs RNN.You can use another slide if needed.

For RNN, I increased drop-out from 0.2 (Default) to 0.3. I tried all other hyperparameters by increasing or decreasing these parameters, but without any change in RNN structure, the improvement to validation set by hyperparameters tuning from the default setting is so minimal. Increasing drop-out tighten the model regularization, and therefore, decreasing the overfitting issue of the model and improve the performance in validation set.

For LSTM, I increased encorder's and decorder's hidden size from 64 (Default) to 98 and the learning rate from 1e-3 (default) to 5e-3. A learning rate of 1e-3 works well, too. I tried increasing or decreasing every hyperparameter as well, and the improvement from changing most hyperparameters to the performance in validation set is minimal. The key part from the default hyperparameter specification to improvement is that the model has not been well trained. Either increasing learning rate or increasing epoch number appropriately would boost model performance significantly.

Seq2Seq Explanation

Explain what you did here and why you did it to improve your model performance. Compare and explain the differences when using LSTM vs RNN.You can use another slide if needed.

In term of structure, LSTM allows a long-term memory that saves information over several iterations, while RNN forgets information quickly over iterations.

In term of performance, based on the results, LSTM is much better than RNN in model training goodness and model performance in validation data.

In term of model coding, the RNN returns output and hidden, but LSTM returns output and (hidden, cell), where the hidden of these two models are the same thing and LSTM has one more return, cell, that denotes a cell state. RNN only needs to pass over output and hidden, but LSTM needs to pass these two as well as cell.

Transformer Results

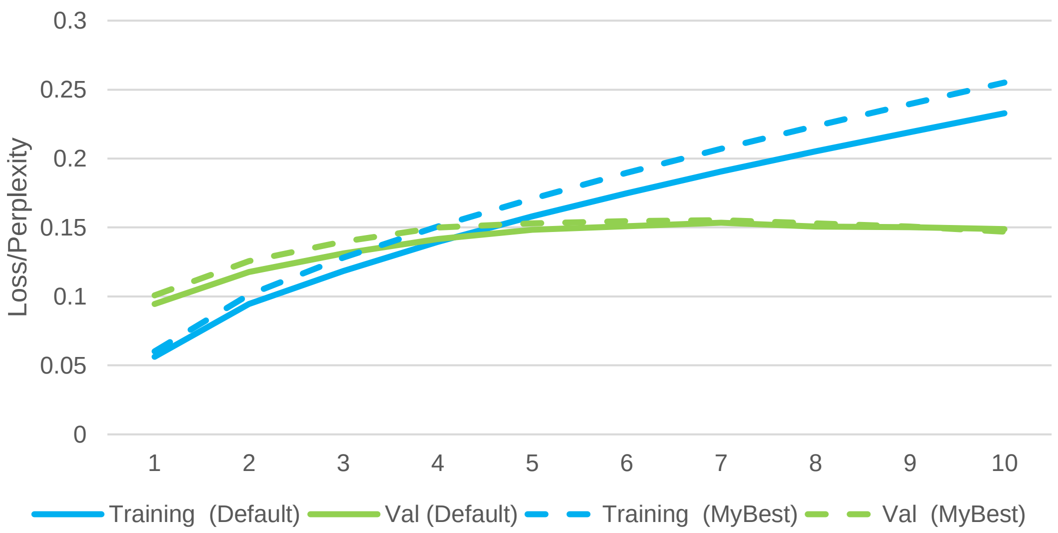
Put your results from training before and after hyperparameter tuning here.

Results for default configuration			
Training Loss	2.2916	Validation Loss	3.0067
Training Perplexity	9.8903	Validation Perplexity	20.2215
Result for your Best Model			
Training Loss	2.484380	Validation Loss	2.940406
Training Perplexity	11.993688	Validation Perplexity	18.923520
Your best model configuration after hyperparameter tuning			
learning_rate = 1e-3， EPOCHS = 10， hidden_dim= 140， num_heads=2， dim_feedforward=2048， dim_k=96， dim_v=96， dim_q=96， MAX_LEN = 20			

Table 2

Transformer Curves

Put the plots for loss/perplexity curves (training & validation) for your configuration with default setting and for your best model here.



Transformer Explanation

Explain what you did here and why you did it to improve your model performance. You can use another slide if needed.

The improvement by hyperparameters tuning is minimal from the default hyperparameters for Transformer. I tried to either increase or decrease every hyperparameter and the improvement in term of the perplexity of validation data is minimal.

The only hyperparameters that improve the performance in validation set is hidden_dim, where I changed it from 128 (default) to 140, and the validation perplexity (i.e. in default 20.2215) can be 18.92352 at Epoch 6 and can be 20.58053 at Epoch 10.

I also tried to adjust learning rate and epoch number, but the model can easily overfit.

Transformer Translation Results

Put translation results for your best model (1st 9 sentences) here

Input sentence	Back translation
'<sos>', 'a', 'young', 'woman', 'and', 'older', 'woman', 'wear', 'traditional’, 'saris', 'as', 'they', 'spin', '<unk>', ',', 'three', 'people', 'are', 'pictured’, '<eos>'	'<sos>', 'a', 'young', 'woman', 'and', 'older', 'older', 'woman’, 'in’, 'traditional', 'saris', 'in', 'three', 'three', 'three’, 'three', 'three', 'in', 'in', '<eos>'
'<sos>', 'sitting', 'casually', 'in', 'a', 'public', 'place', ',', 'a', 'girl', 'reads’, 'holding', 'the', 'book', 'open', 'with', 'her', 'hand', 'on', '<eos>'	'<sos>', 'a', 'girl', 'sits', 'sitting', 'a', 'is', 'is', 'reads', 'to', 'a', 'public’, 'reads', 'book', 'a', 'a', 'reads', 'reads', 'book', '<eos>'
'<sos>', 'a', 'male', 'metal', 'worker', 'using', 'a', 'welding', 'tool', 'in', 'his’, 'right', 'hand', ',', 'while', 'holding', 'the', 'mask', 'in', '<eos>'	'<sos>', 'a', 'foreign', 'in', 'in', 'area', '<unk>', '-', 'the', 'or', 'or', 'the’, '<unk>', '.', '.', 'hand', 'hand', '<eos>', '<unk>', '<eos>'
'<sos>', 'two', 'people', ',', 'one', 'dressed', 'as', 'a', 'nun', 'and', 'the’, 'other', 'in', 'a', '<unk>', '<unk>', 't', '-', 'shirt', '<eos>'	'<sos>', 'two', 'people', ',', 'one', 'in', 'one', 'one', 'a', 'one', 'and', 'in', 'in’, 'and', 'and', 'other', 'is', '<eos>', '.', '<eos>'
'<sos>', 'a', 'dark', '-', 'skinned', 'man', 'in', 'white', 'shirts', 'and', 'a’, 'black', 'sleeveless', 'shirt', 'flips', 'his', 'skateboard', 'on', 'a', '<eos>'	'<sos>', 'a', 'black', 'man', 'in', 'white', 'shorts', 'and', 'a', 'sleeveless’, 'black', 'shirt', 'doing', 'a', 'his', 'with', 'his', 'a', 'on', '<eos>'
'<sos>', 'two', '<unk>', 'sit', 'perched', 'on', 'horses', ',', 'dressed', 'in’, '<unk>', 'ceremonial', 'wear', ',', 'each', 'holding', 'a', '<unk>', 'in', '<eos>'	'<sos>', 'two', 'women', '<unk>', 'in', 'in', '<unk>', '<unk>', 'bra', '<unk>’, ',', '<unk>', 'and', 'a', 'and', 'in', '.', 'one', 'hand', '<eos>'
'<sos>', 'two', 'people', 'are', 'holding', 'a', 'large', 'upside', '-', 'down’, 'earth', '<unk>', ',', 'about', '4', '""', 'in', '<unk>', ',', '<eos>'	'<sos>', 'two', 'people', 'stop', 'holding', 'a', '<unk>', '<unk>', 'with', 'a’, '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', '<unk>', 'it', '.’, '<eos>'
'<sos>', 'a', 'woman', 'is', 'reading', 'a', 'card', 'while', 'sitting', 'on', 'an’, 'end', 'of', 'a', 'couch', ',', 'while', 'another', 'woman', '<eos>'	'<sos>', 'a', 'woman', 'is', 'a', 'a', 'map', 'sitting', 'a', 'couches', 'a', 'the’, 'on', 'a', 'while', 'a', 'a', 'woman', 'while', '<eos>'
'<sos>', 'a', 'lady', 'in', 'a', 'red', 'coat', ',', 'holding', 'a', '<unk>', 'hand’, 'bag', 'likely', 'of', 'asian', 'descent', ',', 'jumping', '<eos>'	'<sos>', 'a', 'woman', 'in', 'a', 'red', 'red', 'coat', 'coat', 'placing', ',’, 'trench', '<unk>', 'in', [illegible]', 'in', 'a', 'blue', 'blue', '<eos>'

Table 3

LSTM Translation Results

Put translation results for your best model (1st 9 sentences) here

Input sentence	Back translation
'<sos>', 'a', 'young', 'woman', 'and', 'older', 'woman', 'wear', 'traditional’, 'saris', 'as', 'they', 'spin', '<unk>', ',', 'three', 'people', 'are', 'pictured’, '<eos>'	'<sos>', 'a', 'young', 'woman', 'and', 'a', 'and', 'wearing’, 'in', 'in', 'and', 'and', 'and', ',', ',', ',', 'are', 'are’, '<eos>', '<eos>'
'<sos>', 'sitting', 'casually', 'in', 'a', 'public', 'place', ',', 'a', 'girl', 'reads’, 'holding', 'the', 'book', 'open', 'with', 'her', 'hand', 'on', '<eos>'	'<sos>', 'a', 'girl', 'sits', 'in', 'a', 'a', 'chair', 'with’, 'a', 'a', 'a', 'a', 'a', 'a', 'a', 'a', '<eos>’, '<eos>’, '<eos>'
'<sos>', 'a', 'male', 'metal', 'worker', 'using', 'a', 'welding', 'tool', 'in', 'his’, 'right', 'hand', ',', 'while', 'holding', 'the', 'mask', 'in', '<eos>'	'<sos>', 'a', 'asian', 'worker', 'in', 'a', '"', '<unk>’, '<unk>', 'a', 'a', 'a', 'a', 'a', 'a', '"', '<eos>', '<eos>’, '<eos>', '<eos>'
'<sos>', 'two', 'people', ',', 'one', 'dressed', 'as', 'a', 'nun', 'and', 'the’, 'other', 'in', 'a', '<unk>', '<unk>', 't', '-', 'shirt', '<eos>'	'<sos>', 'two', 'people', ',', 'one', 'in', 'a', 'a', 'a', 'a’, 'a', 'a', 'a', 'a', 'a', '<eos>', '<eos>', '<eos>', '<eos>’, '<eos>'
'<sos>', 'a', 'dark', '-', 'skinned', 'man', 'in', 'white', 'shirts', 'and', 'a’, 'black', 'sleeveless', 'shirt', 'flips', 'his', 'skateboard', 'on', 'a', '<eos>'	'<sos>', 'a', 'bald', 'man', 'in', 'jeans', 'and', 'a', 'and’, 'and', 'and', 'a', 'is', 'a', 'a', 'a', 'a', 'a', 'a', '<eos>'
'<sos>', 'two', '<unk>', 'sit', 'perched', 'on', 'horses', ',', 'dressed', 'in’, '<unk>', 'ceremonial', 'wear', ',', 'each', 'holding', 'a', '<unk>', 'in', '<eos>'	'<sos>', 'two', '<unk>', 'men', 'standing', 'in', 'in', 'a’, 'in', ',', ',', ',', ',', ',', ',', 'in', 'in', 'in', '<eos>’, '<eos>'
'<sos>', 'two', 'people', 'are', 'holding', 'a', 'large', 'upside', '-', 'down’, 'earth', '<unk>', ',', 'about', '4', '"', 'in', '<unk>', ',', '<eos>'	'<sos>', 'two', 'people', 'are', 'a', 'a', 'a', 'a', '"', '“’, 'says', '"', '<unk>', '<unk>', '<unk>', '"', '<eos>', '<eos>’, '<eos>', '<eos>'
'<sos>', 'a', 'woman', 'is', 'reading', 'a', 'card', 'while', 'sitting', 'on', 'an’, 'end', 'of', 'a', 'couch', ',', 'while', 'another', 'woman', '<eos>'	'<sos>', 'a', 'woman', 'is', 'a', 'a', 'a', 'a', 'a', 'a', 'a’, 'a', 'a', 'a', 'a', 'a', 'a', '<eos>', '<eos>', '<eos>'
'<sos>', 'a', 'lady', 'in', 'a', 'red', 'coat', ',', 'holding', 'a', '<unk>', 'hand’, 'bag', 'likely', 'of', 'asian', 'descent', ',', 'jumping', '<eos>'	'<sos>', 'a', 'woman', 'in', 'a', 'red', 'shirt', 'is', 'a', 'a’, 'a', 'a', 'a', 'a', 'a', 'a', '<eos>', '<eos>', '<eos>', '<eos>'

Table 4

Compare LSTM to Transformer

Compare your LSTM results to your Transformer Results both quantitatively and qualitatively and explain the differences.

Quantitatively, in term of model prediction results, i.e. loss and perplexity, the Transformer model that has a training perplexity of about 12 and a validation perplexity of about 19 is significantly superior to the LSTM model that has a training perplexity of about 26 and a validation perplexity about 31. Based on the quantitative results, the Transformer model supposes to perform better than LSTM model in translation.

Qualitatively, in term of the back translation, the output from Transformer does not produce complete and intuitive sentence, or say the back translated sentence does not make sense in English language. But it produce reasonable words and the output words somehow are similar or exactly the same with the input words. However, the LSTM model does not produce complete sentence or word and frequently produces non-sense words such as “a” and “in”. Hence, even though the performance of Transformer model is not good, but it is much better than the performance of LSTM model.

Theory question

Beam search

Known $s^1 \leq best_{\leq i}$, which denotes the current highest scoring beam in B_t scores worse than or equal to $best_{\leq i}$, where $s^1 = \sum_{j=1}^{t-1} \log p(y_j^1 | x, y_{<j}^1)$.

Then assume s^f , a possible further step, then $s^f = \sum_{j=1}^{t-1} \log p(y_j^f | x, y_{<j}^f)$, where $s^f \leq s^1$ for all items in B_t.

Add more steps will add-in more probability in the summation. As the probability is less than one, no item will be better than $best_{\leq i}$.

Hence, the current best completed beam is the overall highest-probability completed beam and future steps will be no better.