CSE 490 G1 HW2 Short Answer

Wei Jun Tan

1. Just like last time, provide plots for training error, test error, and test accuracy. Also provide a plot of your train and test perplexity per epoch.

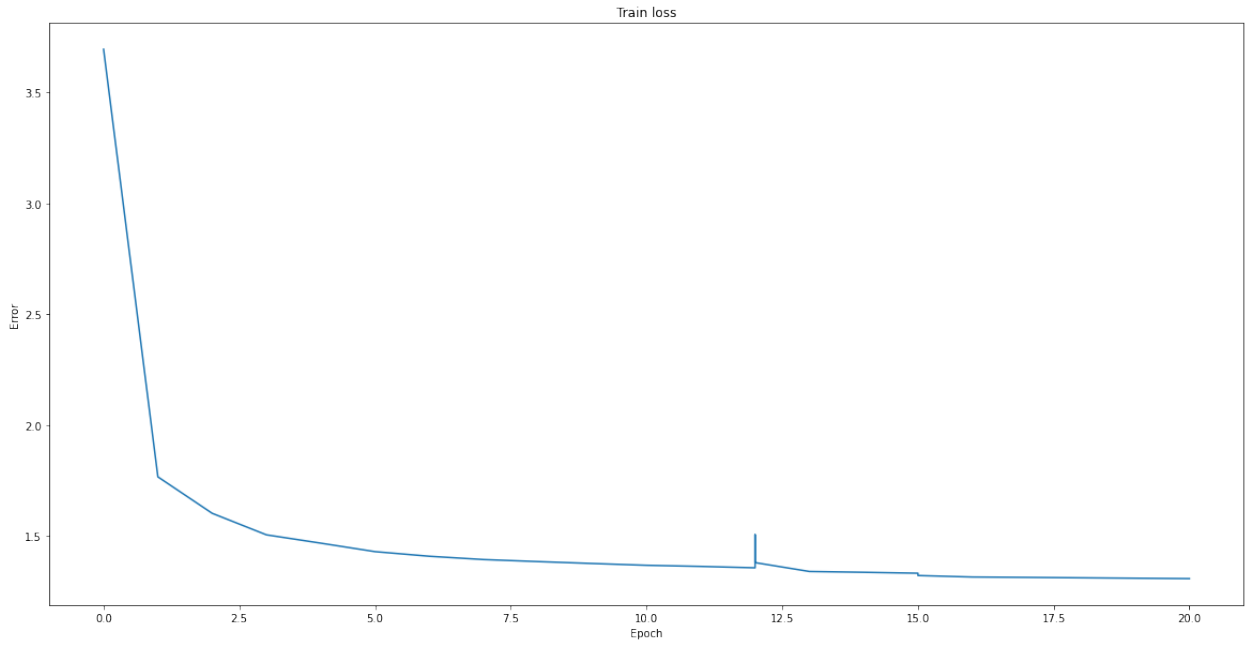
- In class we defined perplexity as $2^{p \cdot log_2(q)}$, However the PyTorch cross entropy function uses the natural log. To compute perplexity directly from the cross entropy, you should use $e^{p*ln(q)}$.
- We encourage you to try multiple network modifications and hyperparameters, but you only need to provide plots for your best model. Please list the modifications and hyperparameters.

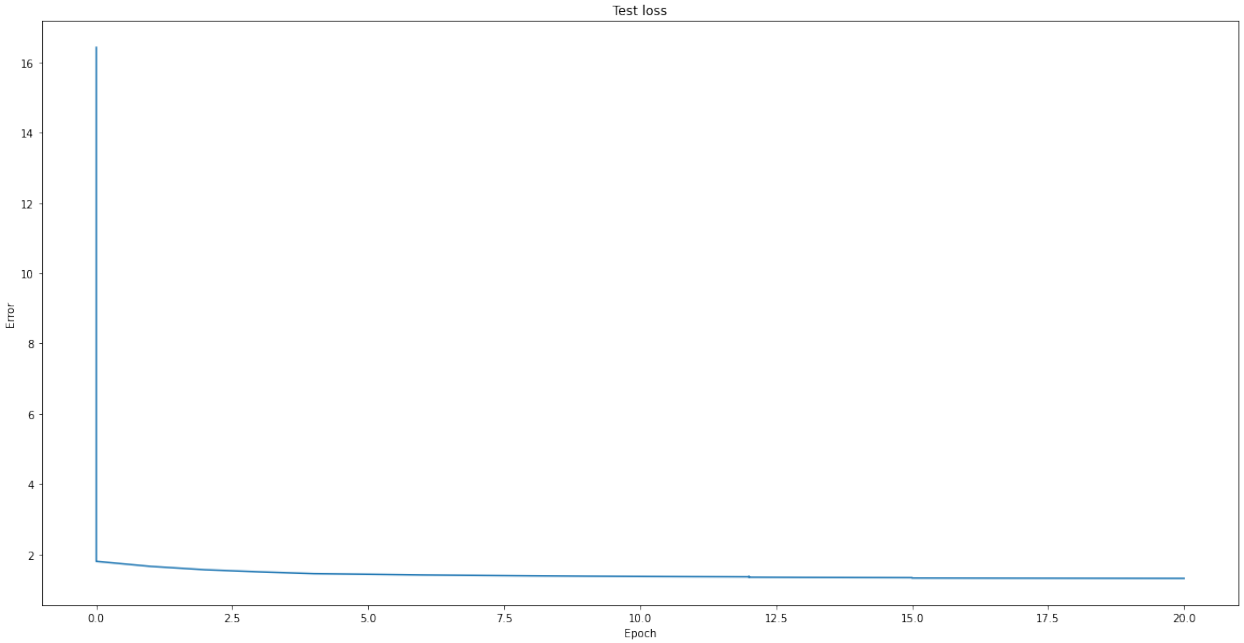
Hyperparameters

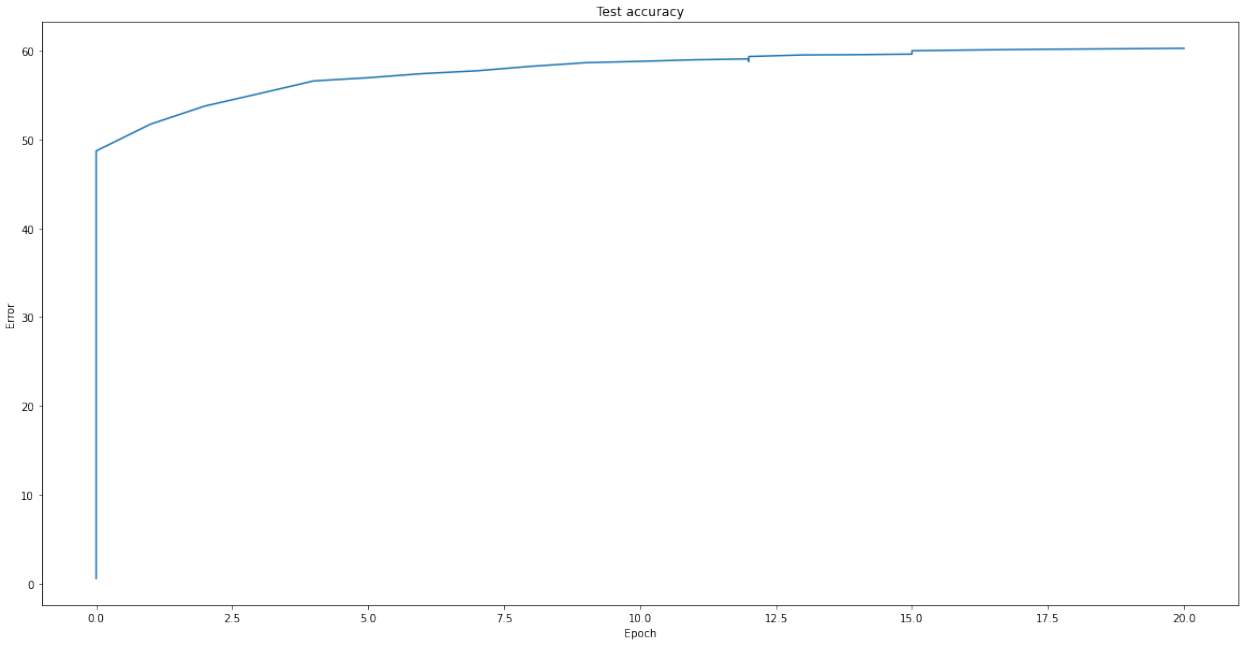
```
TEMPERATURE = 0.5
SEQUENCE_LENGTH = 100
BATCH_SIZE = 256
FEATURE_SIZE = 512
TEST_BATCH_SIZE = 256
EPOCHS = 20
LEARNING_RATE = 0.002 (12 epochs), 0.001 (3 epochs), 0.0005 (5 epochs)
WEIGHT_DECAY = 0.0005
USE_CUDA = True
PRINT_INTERVAL = 10
```

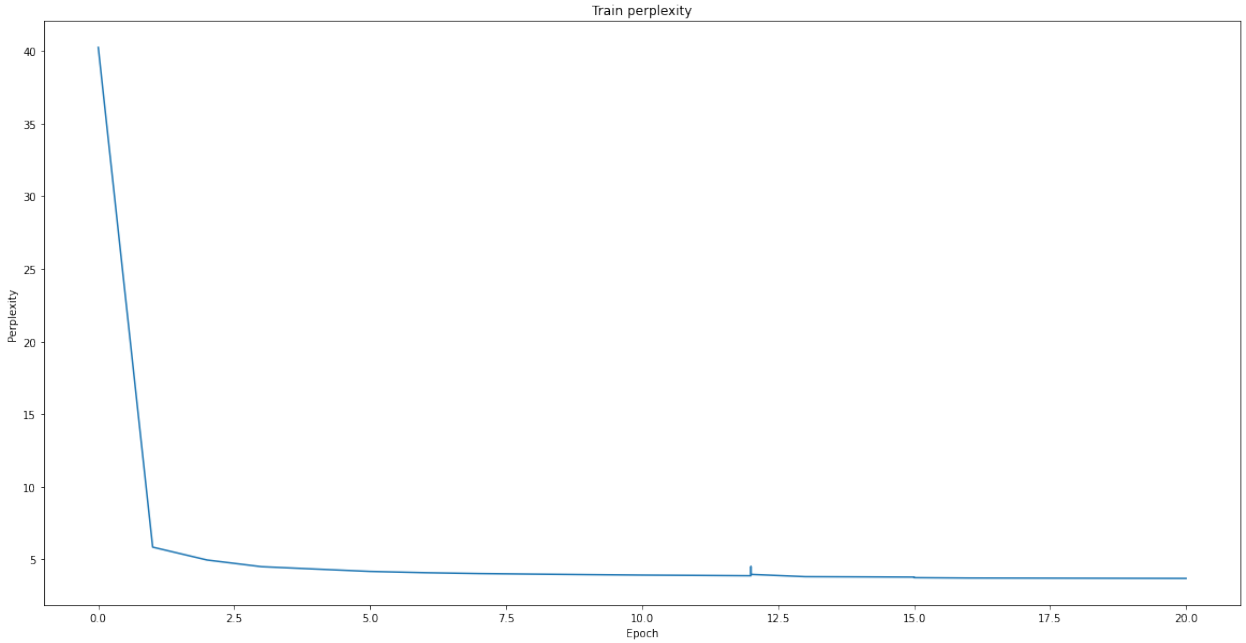
Generally, I didn't find any better network structure (the accuracy often gets worse). Generally, I feel that the existing network is a good one.

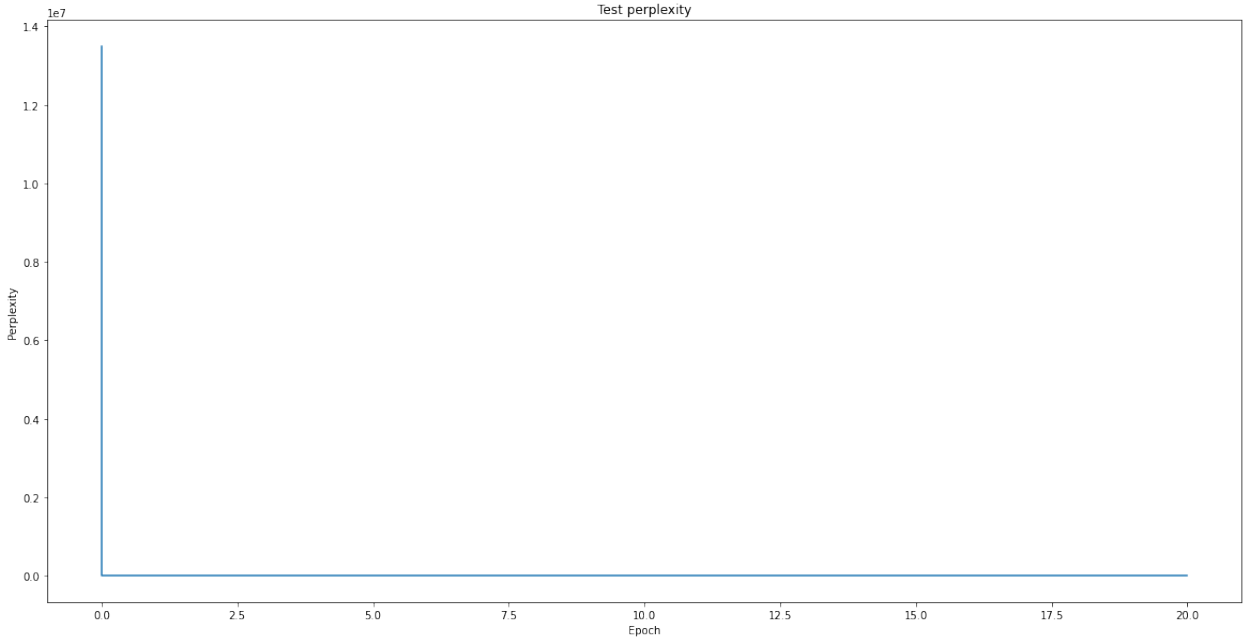
Plots




Test loss
Error
Epoch
16
14
12
10
8
6
4
2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Test accuracy
Error
Epoch
60
50
40
30
20
10
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0





2. What was your final test accuracy? What was your final test perplexity?

```
Final test accuracy: 60.25980548469388
Final test perplexity: 3.7534398397438586
Train Loss: 1.306089
```

3. What was your favorite sentence generated via each of the sampling methods? What was the prompt you gave to generate that sentence?

- Max ("Harry Potter and the"):

Harry Potter and the only one of the corridor that he had not think that the morning with him and started to the fire and the only one of the corridor that he had not think that the moment that he had not think

that the

- Sample ("Harry Potter loves"):

Harry Potter loves that your matter and passes," said Mrs. Weasley, then squeech of parchment over the door and forwazed, and he make a streaming of the window of the stairs could approached on the ord on the Ron, and

- Beam ("Harry laughed"):

Harry laughed and started to the stairs and said, "I think you want to be a lot of the catch of the castle and the start of the car of the castle and the start of the card of the castle and the start of the card a

4. Which sampling method seemed to generate the best results? Why do you think that is?

The sampling method seems to be the best method. The max and beam seems to always produce sentences that do not make sense in the long run and often run into into nonsense loops. The sample sampling (although with some significant amount of grammar mistakes) are very creative and rich in terms of the variation of sentences. It is probably because it selects next tokens based on a skewed distribution, which theoretically makes all output possible statistically. The temperature of 0.5 is also not too high, which still makes the sentences feel more like sentences.

5. For sampling and beam search, try multiple temperatures between 0 and 2.

- Which produces the best outputs? Best as in made the most sense, your favorite, or funniest, doesn't really matter how you decide.

0.5 - 1.0 makes the best output for me. I feel like 0.65 is pretty good number to produce creative outputs that still make sense.

- What does a temperature of 0 do?

It makes the sampling and beam method much more like the max sampling method as it greatly magnifies the output that is highly possible.

- What does a temperature of 0<temp<1 do?

It looks like a reasonable temperature that produces varied but possible outputs. The sentences make more sense and structured as the weights learned previously are weighted accordingly.

- What does a temperature of 1 do? What does a temperature of above 1 do?

It makes the model to generate more random outputs. It makes the model less dependent on the previously learned weights and make more tokens to be possibly selected (perhaps make the sentence more gliberrish)

- What would a negative temperature do (assuming the code allowed for negative temperature)?

Negative temperature will make output that is less possible to be more possible, hence might makes the model to generate sentences that make less sense.

6. Other things

6.1. New Corpus

- What corpus did you choose? How many characters were in it?

The new corpus that I chose is the full texts of the Lord of the Ring. It has 2523660 characters after eliminating newlines/tabs.

- What differences did you notice between the sentences generated with the new/vs old corpus.

It seems like it is much harder to learn the new corpus. The sentences generated make less sense and are generally less meaningful. It is probably because the size of the corpus is smaller, hence is more prone to overfitting.

- Provide outputs for each sampling method on the new corpus (you can pick one temperature, but say what it was).

Temperature: 0.7

Max:

Frodo laughed and the stars and the trees of the trees. Then they stood and the light of the stones of the trees of the trees. Then they stood and the strange the trees of the trees of the trees. Then they stood an

Sample:

Frodo laughed like and bottle. Then sprang to the trees and the trees of the swords of the Stone. Shad of the booth watchful stood light of the mountains, and then they had the land in the forest had been the Ming-

Beam:

Frodo laughed and the stars and the trees of the trees. Then they stood and the light of the stones of the trees of the trees. Then they stood and the strange the trees of the trees of the trees. Then they stood an

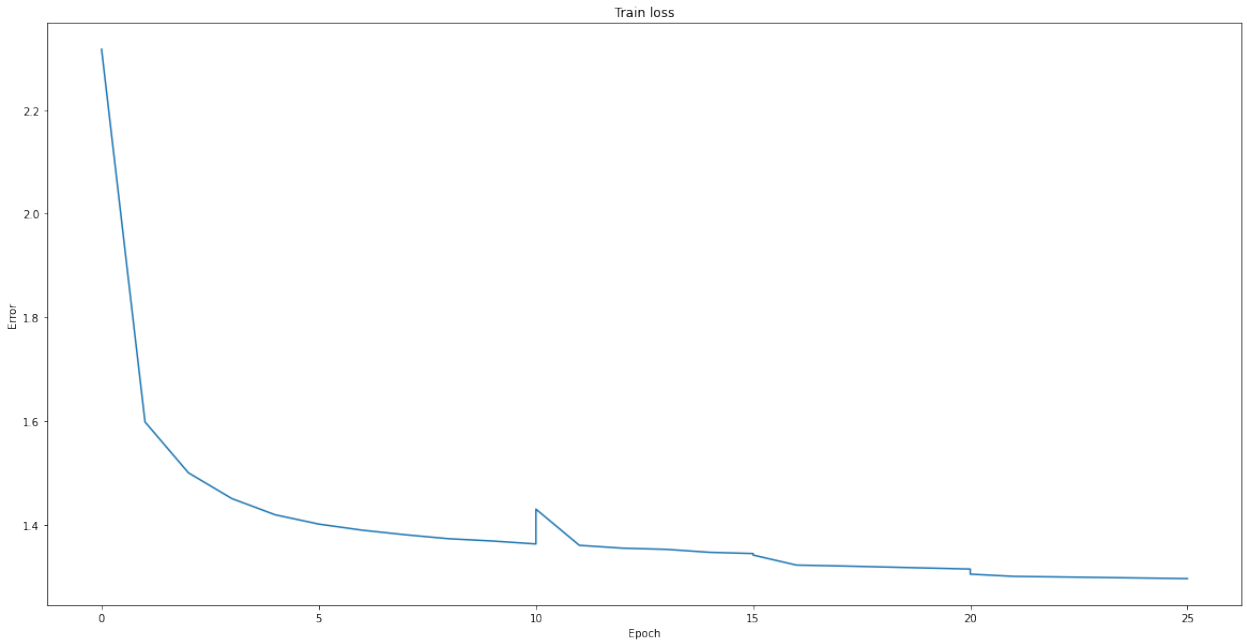
6.3 LSTM

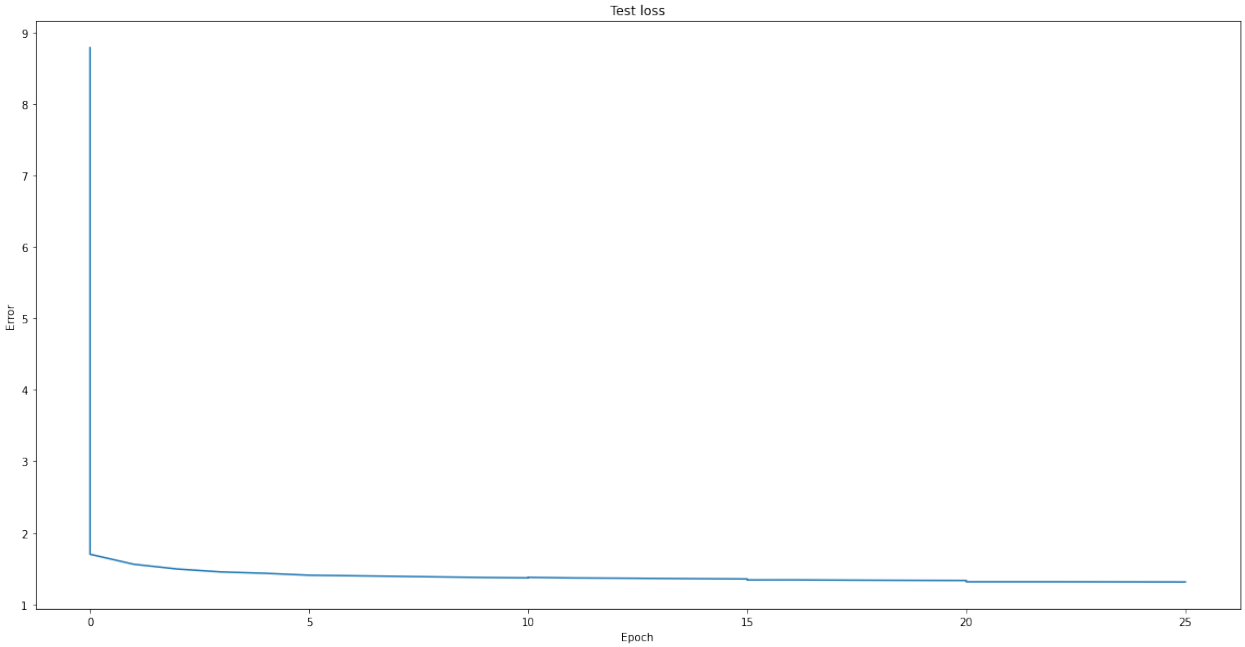
- What new difficulties did you run into while training?

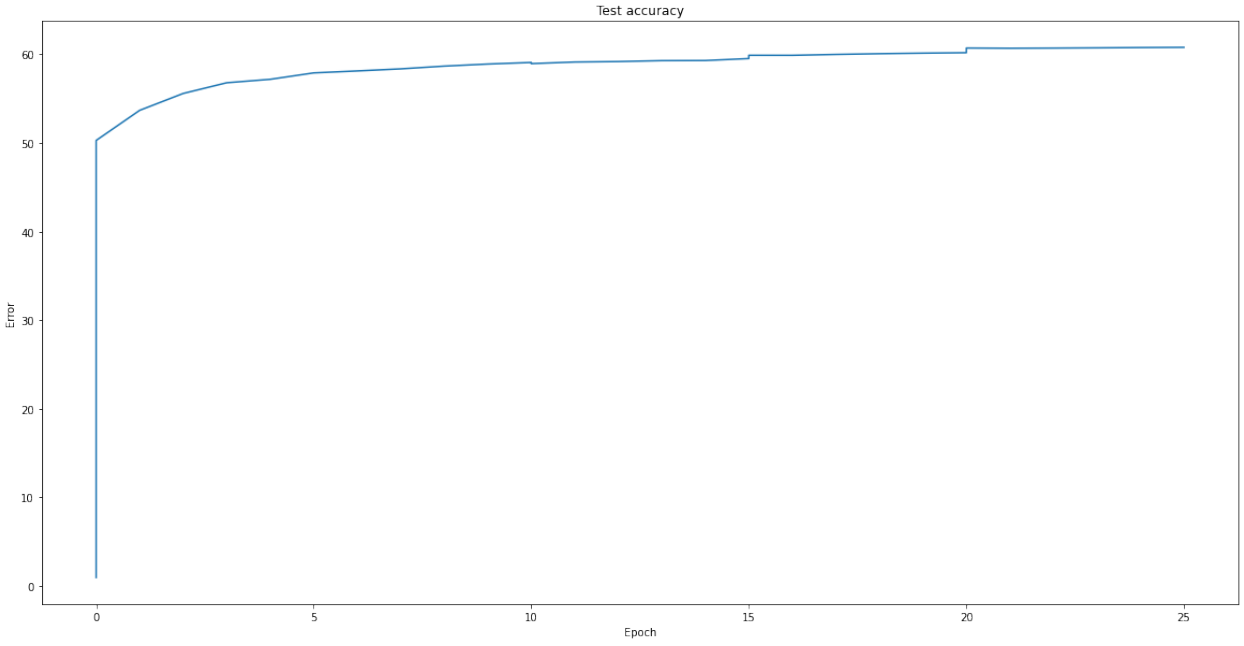
The training process is much slower for LSTM compared to GRU due to the increased complexity of the model. Other than that, the training loss reduces much faster for the first few epochs compared to GRU. Generally, there is not much difficulties/differences as we only change a small part of the model.

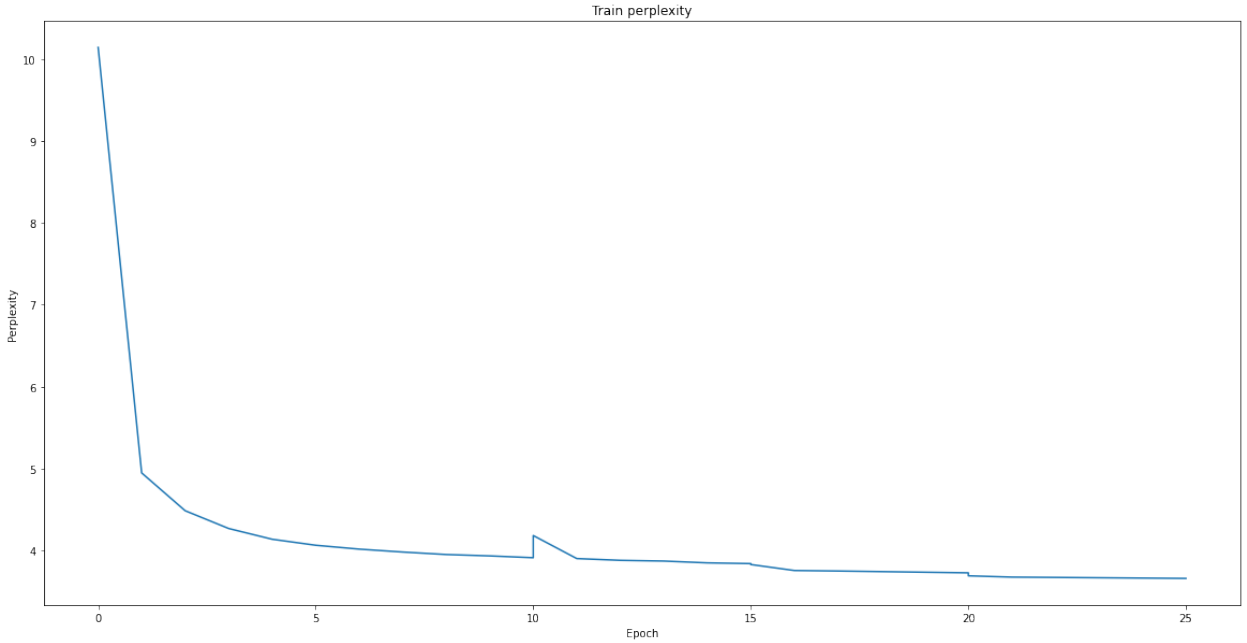
- Were results better than the GRU? Provide training and testing plots.

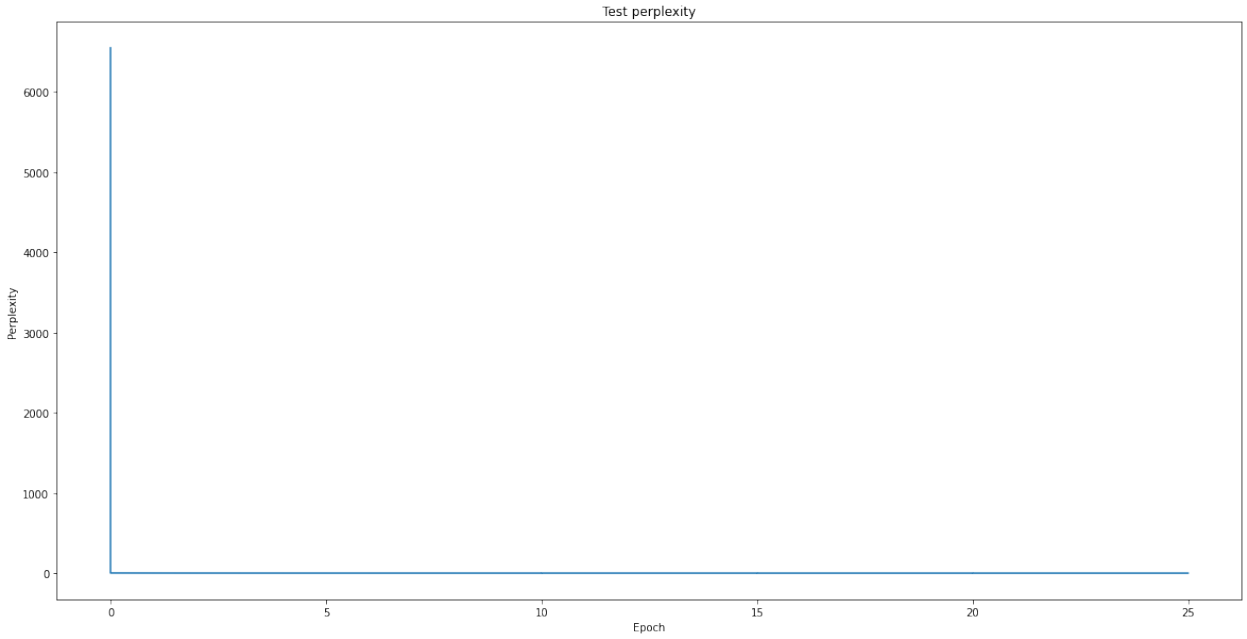
Yes, the result is slightly better than the GRU, specifically the final test accuracy: 60.782844387755105 > 60.25980548469388.




Test loss
Error
Epoch
9
8
7
6
5
4
3
2
1
0
5
10
15
20
25


Test accuracy
Error
Epoch
60
50
40
30
20
10
0
0
5
10
15
20
25





- Provide outputs for each sampling method on the new corpus (you can pick one temperature, but say what it was).

Temperature: 0.5 Seed word: 'Harry laughed'

Max:

Harry laughed at the start of the floor and started to the start of the corridor to the fire and started to the start of the floor and started to the start of the corridor to the fire and started to the start of t

Sample:

Harry laughed. "I was here," said Harry, but he didn't have to see the time and a stone students were close to have a bit of the chair and pulled the words, they were still that the way in the firely not to see th

Beam:

Harry laughed at the start of the floor and started to the start of the corridor to the fire and started to the start of the floor and started to the start of the corridor to the fire and started to the start of t

6.6 Word

- What new difficulties did you run into while training?

I face CUDA out of memory error as the size of vocabulary is too big to be fed into the model. It is because the number of unique words is way more than the number of unique characters. Hence, to reduce the size of vocabulary, I only include words which appear more than 5 times in the vocabulary, which however makes the output predicted always contain the unknown token (empty string), makes the sentences generated often contains a lot of whitespaces.

- How large was your vocabulary?

14403 words.

- Did you find that different batch size, sequence length, and feature size and other hyperparameters were needed? If so, what worked best for you?

I find that a bigger feature size from 512 to 1024 and a shorter sequence length from 200 chars to 20 words (to maintain the relative length of the generated sentences) improve the model accuracy by about 1.5%.