Homework Assignment 3

Deep Learning, 23/24

Nikolay Vasilev
(63190338)

1 Introduction

In this assignment, we will focus on training neural networks for character-level text generation. In the first part, we will implement Long Short Term Memory (LSTM) recurrent neural network according to the LSTM PyTorch documentation. In the second part, we will implement a Transformer-like network according to the original transformer paper using the core mechanisms - the scaled dot product attention module and the Masked multi-head attention module. We will compare the performance of both models using Top-K and Greedy sampling. We will look at the differences in the generated texts, aggregated characters counts, and how they are influenced by the sequence length during training in the generated text.

2 LSTM

After implementing the LSTM model, we trained it for 30 epochs using different sequence lengths during training in the generated text. In Figure 1 we can see the aggregated character count of generated texts on our trained models using Top-K and Greedy sampling. As we can see there are 3 different models, which are trained using 3 different sequence lengths (64, 128, 256). In all three models we can see that the characters t, h, e, a, and empty space are most frequent. It seems that the model that uses a sequence length of 128 in the training process, has the most equal distribution between both sampling methods. Using only 64 characters seems to increase a lot the frequency of the most frequently generated characters using greedy sampling, which may lead to generating a lot of "the" and empty spaces when using this sampling. We can also see that in this case there are some characters that are generated using the Top-K sampling, which are not generated at all using the Greedy sampling. Something similar seems to be visible also when using a sequence length of 256.

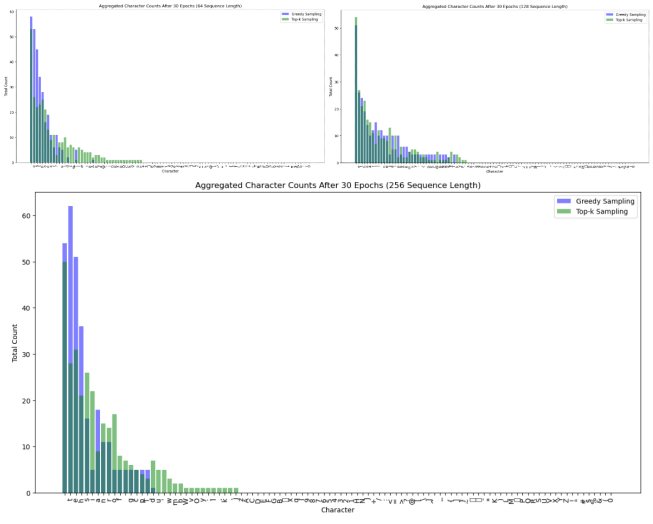

Figure 1: Aggregated characters count of generated texts using Top-K and Greedy sampling and after training the LSTM model for 30 epochs. Different plots show different models trained with different sequence lengths.

Let's take a look at the actual generated texts using the sample text "O Romeo, wherefore art thou":

1. Greedy Sampling:
   - Sequence Length 64 - **O Romeo, wherefore art thou** art thee And send the sense of the senses of the senses That thou art the sense of the senses of the senses That thou art the sense of the senses of the senses That thou art...
   - Sequence Length 128 - **O Romeo, wherefore art thou** art the world That we will be the world and the supper thee.
     KING RICHARD III: Ay, and the common thing to the proport That we will be the world and the supper thee.
     KING RICHARD III: Ay, and the common thing to the proport That we will be the world and the supper thee.
     KING RICHARD III: Ay, and ...
   - Sequence Length 256 - **O Romeo, wherefore art thou** as the state and heart That the state of the prince the strength thee the state That the state of the prince the strength thee the state That the state of the prince...
2. Top-K Sampling:
   - Sequence Length 64 - **O Romeo, wherefore art thou** hust be the Torters and test make the write, With an own shouldst becal and this shame.

KING RICHARD II: An hust be the cares on to his hen hand Of my lords with his counters, then belit of you; Whose son my spower in our comportess shall wind them, For the chore, though too the suce of his cheeps...

- Sequence Length 128 - **O Romeo, wherefore art thou** have been which I think you will not an as that he would bay Fregh well son tell me.
  KING RICHARD III: We will not show the wards of the comfort: Where you hope talks of the sumpeast their land. Were I have high, thy best was a bride, Whom you and this cheer of think, sir, there's a bort A battles...
- Sequence Length 256 - **O Romeo, wherefore art thou** subjects his father's lends Of their pait of mine is tongue.' Where is't to the first there's sinsed this dig; To chase it to the cheeting out of this persents To bud heavy to thee to thrown spoke of his for in a thisties against the secongrance, The won he is stronger'd friends with hurm and prise...

From these generated texts we can clearly see why the distribution was like that. For Greedy Sampling, we can see that most of the text just contains repeated sentences, but we can also see that the model that used the length of 128 seems to produce the longest and the most sensible sentences. When looking at the Top-K generated texts we can actually see why this model is the best. Using the other 2 lengths seems to generate text that sometimes contains some right words, but most of them don't really make sense, while the model that uses sequence length 128 generates almost all sensible words. Another interesting thing is that the model that uses sequence length 64 seems to produce better text than the one that uses 256 chars.

3 Transformer-like network

Let's do a similar evaluation on our implementation of the transformer-like network. Better overall performance is expected over the LSTM networks, but we are interested in the impact of the sequence length between different transformer-like networks. The transformer-like network uses a scaled dot-product attention mechanism to understand the importance of each token (character) in a sequence, i.e. how much a character should pay attention to the other characters in a sequence. In this module, we use a masking operation in order to make sure that the decoder part of our model will look only to the tokens that precede the current position in the output sequence, i.e. to prevent attention to not yet generated characters.

We have trained these models for 50 epochs using different sequence lengths during training in the generated text. In Figure 2 we can see the aggregated characters count of generated texts on our trained models using the Top-K and Greedy Sampling. As we can see there are again 3 different models, which are trained using 3 different sequence lengths (this time 32, 64, 128). As expected also here the characters t, h, e, a, and empty space seem to be most frequent. Something interesting that we can see here in comparison with the LSTM model, is that even using only 32 characters generates

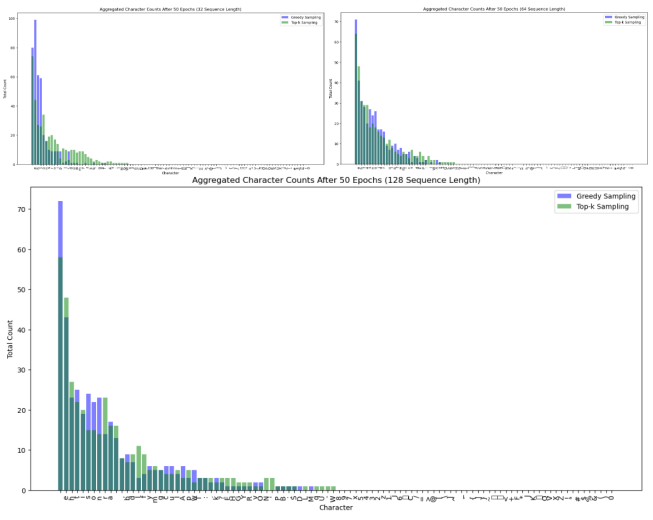

Figure 2: Aggregated characters count of generated texts using Top-K and Greedy sampling and after training the Transformer-like model for 50 epochs. Different plots show different models trained with different sequence lengths.

larger number of different characters using Greedy Sampling. When we compare these Transformer-like models between each other we can see that in the Greedy sampling seem to generate mostly the most frequent letters only for the model that uses the smallest sequence length. For the other two the distribution seems to be almost equal for both models. Also here the model that uses a sequence length of 128 seems to be the best since it generates almost half of the characters and still has pretty good character distribution.

Let's see what the generated texts will actually look like again using the same sample text:

1. Greedy Sampling:
   - Sequence Length 32 - **O Romeo, wherefore art thou** art a slave, That which should be the prince that he hath seeen theee theee to the prince that he hath seeen...
   - Sequence Length 64 - **O Romeo, wherefore art thou** art not stands to be put them against the world of the world: I would not be so bridegrooom of thine ears: The words but the lands of my chambitions, That in the consuls of their cheeeks and these worlds That were not beechelesss but that he hath beeen The words of his children breathies of his charm, And his head in and whaten here is not so deen...
   - Sequence Length 128 - **O Romeo, wherefore art thou** hear?
     GREMIO: Ay, ay, a strange a brave straight; the other service, The crown with his eye: and his beind tongue, And therefore begin speech in the other of mine, And therefore I may be the man.
     BAPTISTA: I know it not to cellesss in this businesss. Do you He came to speak of within you, and the while Your highness is a passs but so one on the bed, And then to be a man to be so stick?...

2. Top-K Sampling:
    - Sequence Length 32 - **O Romeo, wherefore art thou** assuies there we may show it were feeel. Go, trick, hast thou honour'd thyself; and you are never A people in a marrriage, sir, and hall I amade And many me as thou art and hours. You keep mark, Which shall, be the general, trick, when they was not at thy father what thou hide and his mann's life way thou wilt, as if they are not made here...
    - Sequence Length 64 - **O Romeo, wherefore art thou** against a thousand, That made meaning tooo sppeak, but that whoule have Duke in a heart. Thither and here best the held Wisheresof and to make an even bring the pallace To the house of cause, and with his hateful stout, With then have brought home in. Since here favourselvesss and troubled.

      Third Citizen: That sends them did before I seeek heaven the world: I know it never speak...
    - Sequence Length 128 - **O Romeo, wherefore art thou** hear?

      GREY: Is they abhorre him? Therefore fish in this act, Having neither sheath'd at himself. But he beggs, Wert send their before too greeen in? I'll be the think there be therefore be our kisss.

      ANTIGONUS: Affliction, gentleman, A hide and sequary for the people, I pray, he may He'ld us aliene; the which his order, in heart-plied treach, Towards the told me of my pinsorable...

We can see a clear difference between the models. Even using the Greedy Sampling only gives a repetitive sequence for sequence length 32, the other models seem to generate sensible text. Meanwhile, when using the Top-K Sampling all the texts contain sensible words, but again the model that uses the shortest sequence length seems to have the lowest performance. The other two have similar performance, but we can see that the model that uses a sequence length of 128 seems to have the better structure of sentences and therefore is the best among them.