Deep learning - third homework

Jovan Prodanov (63180376)

1 Introduction

Like the previous homework, I implemented this one locally on my PC and trained all networks with my GPU (NVIDIA RTX 3060 12GB).

2 Task 1

2.1 Include examples of the generated text using the Top-K and the greedy sampling when using the LSTM network. What are the differences in the generated text?

Example used: ”O Romeo, why art thou here”.

```
Top-K sampling -----------------
O Romeo, wherefore art thou dissumble,
and to sup the sore with to steal
his proud of heard.
CORIOLANUS:
I ap a mine were calls are at half
we wear
Apostant wife anothouseling.
Second Citizen:
An in an old my living frowes and
the sease
To seading than holset in my fear
to my lash
Must than the fair suition of the
commanden
Greedy sampling ----------------
O Romeo, wherefore art thou the
sensely soul
That the man and the state and
seems to be all the sensely
That the state the courterate
and seems to be a sensely
That the state the courterate
and seems to be a sensely
That the state the courterate
and seems to be a sensely
That the state the courterate
and seems to be a sensely
```

The difference between the generated text is that the greedy sampling starts to repeat.

2.2 Train the LSTM network several times with different sequence lengths during training. What is the impact of the sequence length during training on the generated text.

I trained the network 3 times with different chunk sizes: 64, 128, 256. The losses were the following: 64 received 1.31, 128 received 1.5, 256 received 1.31.

The difference between sequence lengths in training is that a longer sequence causes the network to look for long-term dependencies in the text. A shorter sequence, on the other hand, results in a less coherent and meaningful text because it does not connect the words at all points. In addition, the shorter sequence length results in faster computation time and allows the network to converge faster - but ”incorrectly.”

3 Task 2

3.1 Include examples of the generated text using the Top-K and the greedy sampling when using the Transformer network. What are the differences in the generated text?

Example was the same as the previous task.

```
Top-K sampling -----------------
O Romeo, why art thou
here ? thy are be;
And thou shalt to adds
and by and by foaming but
Within the allige frayall
of our actions.
KING RICHARD III:
Must I speak and noble lord,
you thank you all:
And half that provoked for
a hundred man.
DUKE OF YORK:
In God’s base faith, duly
anon speak man that?
Happpy, the on thy
aims--shapen still stars!--
Appotet like as thy true and
thy fatal nay?
TRANIO:
Of all thing, believe we
```

```
Greedy sampling ----------------
O Romeo, why art thou here
about thy country?
ROMEO:
I am ashap, as the appetite
of all the world.
JULIET:
I warrant thee to meet
thy married:
These she change now I
prophet now a farmer man.
ROMEO:
I would tell it be tolder,
behold, holding our father
And now in his law to his
patramed the people.
JULIET:
O God, foood him, for his
substitute, for and breast,
Look, or never be assurance
for a topsax.
HORTENSIO:
She h
```

The difference is that Greedy sampling produces predictive text (with correct grammar), while Top K sampling is more varied and imaginative.

3.2 Train the Transformer network several times with different sequence lengths during training. What is the impact of the sequence length during training on the generated text.

The Transformer network is designed to handle sequences of any length, which is one of its biggest advantages. The only difference is that when training with a longer sequence length, the net trains on more long-term dependencies, but this also results in slower training because it has to process more tokens and consider more positions in the input.

I trained the Transformer network 3 times with different chunk sizes: 64, 128, 256. The losses were as follows: 64 received 1.11, 128 received 0.998, 256 received 0.71. So we see that the loss decreases as the sequence length increases.

3.3 What is the point of the masking operation in the Scaled dot product attention module in the Attention is all you need paper in Figure 1? Why is the attention map masked?

The attention map is masked to prevent the network from paying attention to future positions, the sequences are processed from left to right, when predicting the third word it should only consider the first and the second word. For this reason the attention map is masked.

So masking is useful when certain positions of the input should be excluded, in our case from left to right.