Exercise Nr. 5

1) LSTM

We used the standard network size (2 LSTM layers with d_lstm=128) with a higher learning rate of 0.0005 instead of 0.0001 (which was a bit too slow for our taste) to accelerate training. We used Google Colab to train on a Tesla V4 GPU and training for 1000 epochs took roughly one hour and we achieved a final error of 0.0293. The final model outputs perfect Tolkien text:

> Initialization: then the old man looked up. pi...
>
> Tolkien: — then the old man looked up. pippin saw his carven face with its proud bones and skin like ivory, and the long curved nose between
>
> LSTM model: — then the old man looked up. pippin saw his carven face with its proud bones and skin like ivory, and the long curved nose between t

Here is what the LSTM learns with successive epochs:

0 Epochs (Training error 3.048):

> Initialization: “and yet’ - he paused and stoo...
>
> Tolkien: — ‘and yet’ - he paused and stood up, and looked round, north, east, and south - ‘the doings at isengard should warn us that we are
>
> LSTM model: — ‘and yet’ - he paused and stoo

50 Epochs (Training error 1.320):

> Initialization: ‘yes, and one who journeyed wi...
>
> Tolkien: — ‘yes, and one who journeyed with him,’ said pippin. ‘and boromir of your city was with us, and he saved me in the snows of the nor
>
> LSTM model: — ‘yes, and one who journeyed with were the see that the street of the shall be the street of the street of the street and the street

100 Epochs (Training error 0.932):

> Initialization: ‘which question shall i answer...
>
> Tolkien: — ‘which question shall i answer first?’ said pippin. ‘my father farms the lands round whitwell near tuckborough in the shire. i am
>
> LSTM model: — ‘which question shall i answer is it is not that is well,’ said denethor, ‘and so that i can do not there is the storm its white tr

200 Epochs (Training error 0.408):

> Initialization: ‘nine o’clock we’d call it in ...
>
> Tolkien: — ‘nine o’clock we’d call it in the shire,’ said pippin aloud to himself. ‘just the time for a nice breakfast by the open window in
>
> LSTM model: — ‘nine o’clock we’d call it in the shire,’ said pippin aloud to himself. ‘man, they were some halls of gondor.’ the widing and cloa

600 Epochs (Training error 0.038):

> Initialization: gandalf went to the door, and ...
>
> Tolkien: — gandalf went to the door, and there he turned. ‘i am in haste pippin,’ he said. ‘do me a favour when you go out. even before you r
>
> LSTM model: — gandalf went to the door, and there he turned. ‘i am in haste pippin,’ he said. ‘do me a favour when you go out. even before you re

We can see that already after 50 epochs, the LSTM has learned to output English words (even though the output is very repetitive). After 200 epochs, the model starts to output Tolkien, but has not yet memorized the whole sentence and drifts off in the middle. After 600 epochs, the model is basically done training and outputs perfect Tolkien sentences (but it adds an additional letter at the end).

2) Transformer

We implemented everything that was asked but sadly the loss did not drop during training. We are unsure why, we went over the code multiple times, we tried different learning rates(1e-5, 1e-4, 1e-3, 5e-3), implemented biases for W^Q, W^K, W^V, W^O, looked for bugs but did not find anything. The loss always converged within the first few epochs to a very high value somewhere around 3.5-3.6 depending on the LR. The output makes no sense:

> Initialization: bergil clapped his hands, and ...
>
> Tolkien: — bergil clapped his hands, and laughed with relief. ‘all is well,’ he cried. ‘come then! we were soon going to the gate to look on.
>
> Transformer model: — bergil clapped his hands, and t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t t

So probably we missed something in our implementation.