Katharina Alefs, Finn Freiheit, Raphael Reimann

Exercise 1 - Assignment 5

Question 2
Do you need to replace the padding token pad with the end of the sequence eos token? Why or why not?

Because some texts are very short, we should replace the padding token with the EOS token. If we did not do this it is possible that the padding token is masked during training for short texts, which is unwanted behavior in the training process.

Question 3
Are there any columns in the ag_news dataset that are not required for the MLM task?

As we do not use the label column because our task is masked language modeling, we can remove it.

Question 9
What is the best validation loss you achieved after training? Describe your setup including final choices of hyper-parameters, optimizer, etc.

The best validation loss we achieved after training is 0.276.
The hyperparameters that we chose in our setup are:

- batch size: 32
- epochs: 3
- weight decay: 0.01
- learning rate: 2e-5

Apart from that we kept all the default settings of the pretrained RobertaForMaskedLM model (this includes a linear learning rate scheduler and AdamW Optimizer).

Question 10
Calculate the perplexity on validation and test splits and report them separately. Do you think there is a relationship between perplexity and cross-entropy?

We calculate the perplexity on the validation and test set. Perplexity Validation: 1.33; Perplexity Test: 1.32.
Cross-entropy measures how different the predicted probability distribution of the model and the true distribution of the target sequence are. Perplexity is a measure of uncertainty and a lower perplexity indicates that the model is better at predicting the next token.
Perplexity and cross-entropy are related concepts as perplexity is calculated $2^{cross-entropy}$.
Both metrics give indications into the models performance as lower values indicate better predictions and understanding of the underlying data.

Question 11

As an explicit inference, use your model to predict the token in the following text (taken from ag news) and report the top 5 probable tokens predicted. Do you think these predictions make sense? Why or why not?

The top 5 predictions are:

1. scams
2. fraud
3. threats
4. dangers
5. attacks

We would argue that these predictions make sense in the context of the sentence, since they would occur similarly in human interaction. The only criticism that we would raise is that the given answers are not really creative.