Gabriel Athènes
gabriel.athenes@polytechnique.edu

Lab session # 3
ALTEGRAD 2021

11/30/21

1 Question 1

The square mask randomly masks some of the tokens from the input. The objective is to avoid the attention mechanism to focus all its attention on the same tokens, therefore forcing all the tokens to be considered by the model. The positional encoding adds information to its word about its order in the sentence.

2 Question 2

Before fine-tuning the model, we need to change the classification head because we re-purpose our pretrained model for a new classification task. We cannot therefore use the same pretrained weights.

3 Question 3

The number of trainable parameters is equal to $maxlen \times ntoken \times nhid$ (embedding) + $numberlayers \times nhid^2$ (transformer) + $nhid \times nclasses$ (decoder), with $nclasses = 2$ for the classification task and $nclasses = ntokens$ for the language modeling task.

4 Question 4

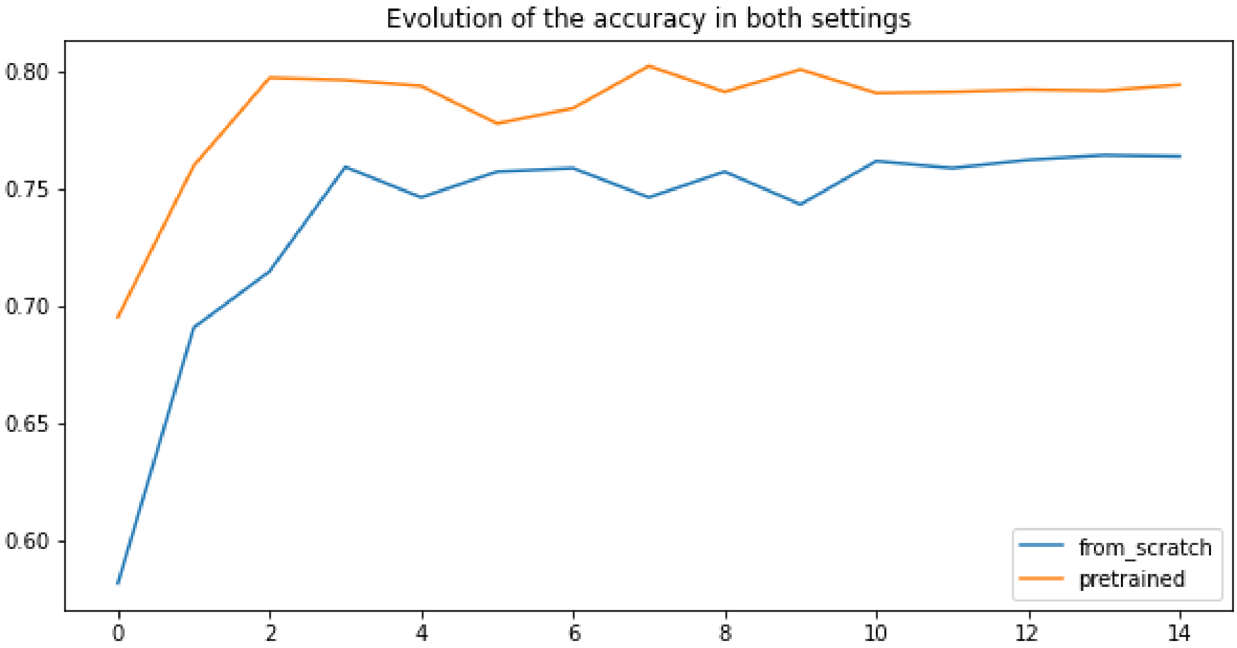

Figure 1: Comparaison of the accuracy of the model in a pretrained and from scratch setting

We can see that in the pretrained setting, the model has better accuracy.

5 Question 5

The main limitation of language modeling objective is that it is uni-directional. As [1] explains, this can be very harmful to fine tuning tasks that require contexts from both directions of the sentence.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.