# 1 Question 1

* Square mask: the square attention mask is used to mask any future tokens from the currently processed token. This masking technique offers the advantage of being able to feed the sequence in one go without fear of ”data leakage”.
* Positional coding: since transformers do not use RNNs, data sequentiality is not as preserved as with RNNs . The tokens are processed simultaneously. Thus, in order to include the notion of position in the architecture, a mapping function is used to encode a unique encoding for each word position in a sentence.

# 2 Question 2

Language modeling involves learning context-based representations of the input text. A good language model provides a good understanding of the input text. The classification task uses this ”understanding” to capture features of interest and perform specific classification tasks such as sentiment analysis. When this classification is supervised, the language model used can be modified (tuned) to obtain a specific understanding of the input.

# 3 Question 3

Language modelling task, n = 20852001:
n = n(embedding) + n(positional encoding) + nlayers*n(transformer bloc) + n(classifier)
= sizevocab * nhid + 0 + nlayers(nhead * (3 * nhid * nhid/nhead) + 2(nhid$^2$ $+nhid$)) $+nhid*nvocab+nvocab$
= sizevocab * nhid +0 + nlayers((3 * nhid$^2$) $+ 2(nhid^2 + nhid)) + nhid * nvocab + nvocab$
= sizevocab * nhid + nlayers(5 * nhid$^2$ $+ 2 * nhid) + nhid * nvocab + nvocab$
= 50001 * 200 + 4(5 * 40000 + 2 * 200) + 200 * 50001 + 50001
= 20852001
Classification task, n = 10802202:
Here, we have a difference on n(classifier) only, with n(classifier)=nhid x nclasses + nclasses = 402

# 4 Question 4

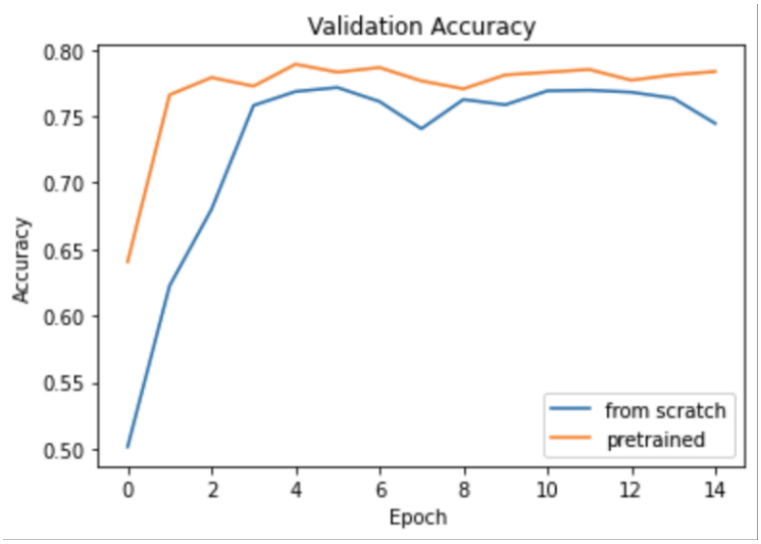


Figure 1: Accuracy in function of epochs for pretrained and trained from scratch models.

For finetuned model, the pretraining gives a better accuracy all along, with a particularly great start (77% after only one epoch) compared to our model trained from scratch (needs at least 4 epochs to reach 77% for

the first time. We also notice that the from-scratch model requires more iterations to converge compared to the pretrained model.

# 5 Question 5

Our language model has been trained on the task of predicting each word based on the previous words in the sentence. This task is then unidirectional. BERT, by being bidirectional thanks to the mask approach, is able to predict each word according to the previous and following words.