Denis Duval
denis.duval13@gmail.com

Lab session # 2
ALTEGRAD 2022

11/14/22

1 Question 1

The square attention mask is used for sentences of length less than the maximum length used in the batch, to hide the characters that aren't part of the actual sequence (padding characters). For language modelling it should also hide all the subsequent characters for the model not to cheat.
The positional embedding is simply here to add information about the relative positions of the elements of our sentence.

2 Question 2

The classification head is a task-specific layer so it should be replaced when working on other task because it is only trained for a specific task.
The main difference between text classification and language modelling is that the first is a discriminative model and the other is a generative one. The aim of language modeling is actually to learn the overall structure of a language and its rules, often on large datasets and in an unsupervised manner.

3 Question 3

For a language modeling model, let us list all the blocks and its number of trainable parameters:
- embedding : $ntokens * dimEmb$ and here in this lab it is assumed that $dimEmb = nhid$.
- positional encoding : these are fixed parameters.
- transformer encoder : [2]we have $nlayers$ times this block and each one represents $3 * nhead * nhid * nhid + 2 * (nhid * nhid + nhid)$ the first representing the three inputs after a multi head linear layer and the second one representing self attention.
-classification: $nhid * ntokens + ntokens$. So we get 20852001 parameters replacing all with the actual values. For classification it is the same except that now the classes are only 2 instead of $ntokens$. We get 10802202 parameters.

4 Question 4

See 1. We can see that the pre-trained model outperforms the model built from scratch on this dataset with accuracies around 0.75 and 0.80. This is what we expected : in fact the pre-trained model has been trained on a large data and hence it knows deeply the structure of the language. Moreover as said, earlier, it had been trained on a generative way for completing sentences which intuitively confirms that the model has a thinner comprehension of language.

5 Question 5

One of the limitation [1] is that the context that a word receives as input during the pre-training phase is unidirectional, so it receives the words to its right. However for a word to know the context of a sentence (or document) it is intuitively sub-optimal for a model to give only a part of it. So in [1] the masked LM is proposed as to patch this problem : the whole context is given to the word but some words are randomly masked.

References

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

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

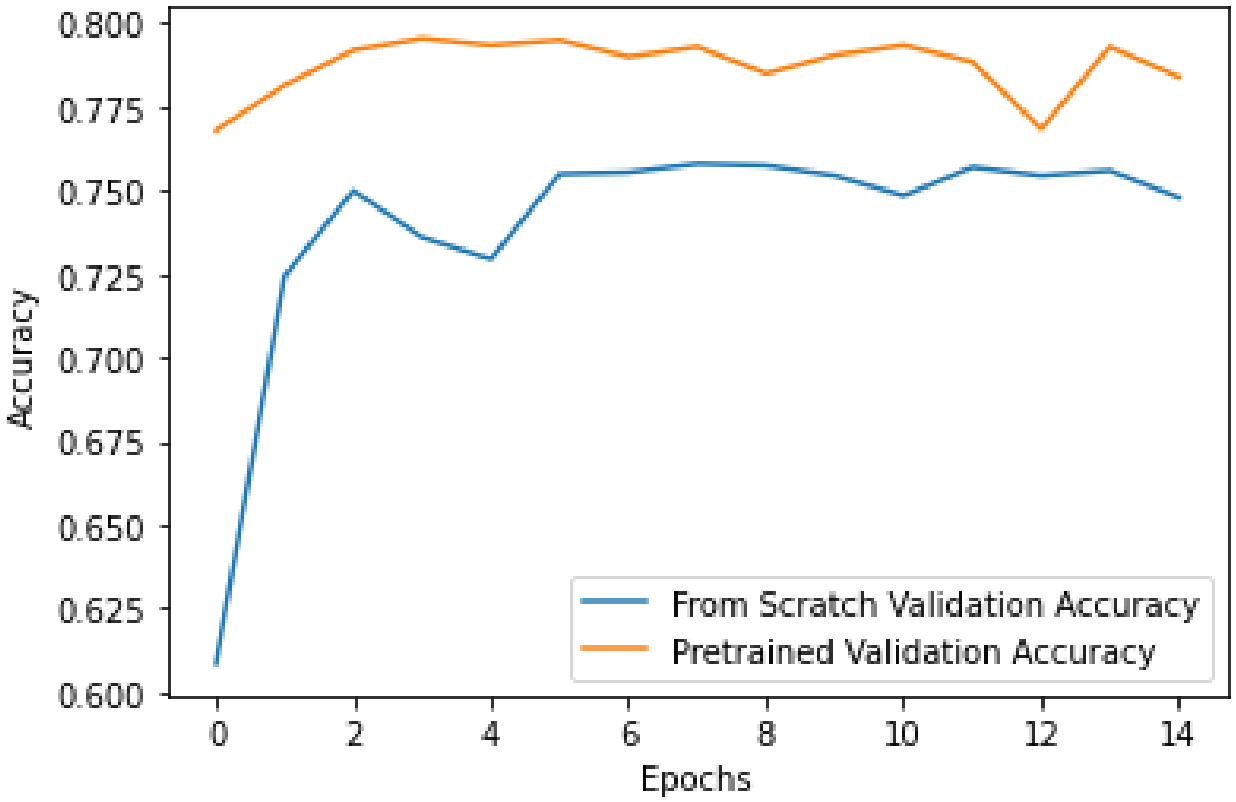

Figure 1: Results when running from scratch and from a pre-trained model