Zakaria El Founoun
zakaria.el-founoun@student-cs.fr

Lab session # 2
ALTEGRAD 2022

11/14/22

1 Question 1

The role of the square mask is to mask the next tokens in a sequence in order to predict them given the previous ones, this will force the algorithm to predict the next word for example from only the previous words.

The positional encoding is used to include the information about the relative positions of words in their representation / encoding.

2 Question 2

The main goal of transfer learning is to use a pretrained model that was trained on some task, and to 'transfer' it to perform a specific task, so it is mandatory to change the classification head to adapt it to the new task which is classification in our case.

The main difference between language modelling and classification tasks: language modelling is generative, meaning that it aims to predict the next word given a previous sequence of words, and language models are trained on very large datasets in an unsupervised manner. On the other hand, classification task is supervised, meaning we need annotated data, which are scarce.

3 Question 3

Let's start with the base model: $n_{base} = n_{embbeding} + n_{transfered},$
We have $n_{embbeding} = n_{tokens} * d_{embedding} = 50001200 = 10000200$ and $n_{transf} = 968000$
* For the last layer / classification head: $n_{head} = n_{hidden} * n_{classes} + n_{classes} = 402$ where $n_{head} = 2$ and $n_{hidden} = 2$ because we have 2 linear and 2 norm layers
So: $n_{classification} = 10968602$
* For the Language modelling task : the classes are the vocab size, so we add $n_{head} = 10050201$ Hence we have $n_{languagemodelling} = 21018401$

4 Question 4

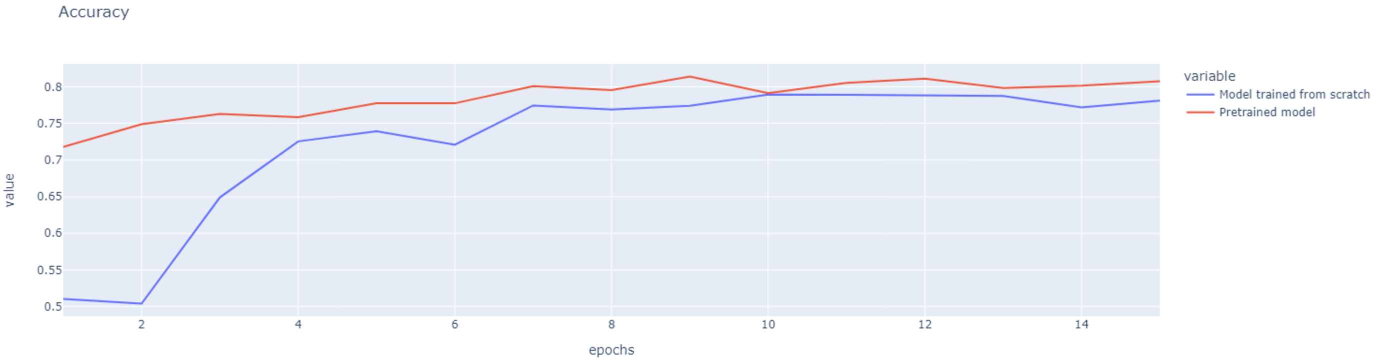

Figure 1: Accuracy of pretrained model and a model from scratch.

Using a pretrained model, we have relatively good accuracy from the first epochs, which is natural because we use a pretrained parameters. On the other hand using a model built from scratch requires more epochs. In short, pretrained model outperform a model built from scratch and when the number of epochs is sufficiently large, the two models have the same scale of accuracy.

5 Question 5

The language modelling used in this notebook is unidirectional meaning that we only mask the future tokens and try to learn on the past ones to predict them. So the model learns only in one direction ”left to right”. The solution proposed by BERT algorithm is a generalization of this concept to deep bidirectional architectures by jointly conditioning on both left and right context in all layers.