Zakaria ECHCHAIR
Zakaria.ECHCHAIR@polytechnique.edu

Lab session # 2
ALTEGRAD 2022

11/14/22

# 1 Question 1

- The square mask's role is to prevent the attention mechanism from sharing any information about the token at the next positions when making a prediction using all of the previous tokens.

- The arrangement of words in a phrase is crucial while speaking a language. The entire meaning of the sentence can be modified by changing the word order. When creating NLP solutions, recurrent neural networks contain an internal system that deals with sequence order. The transformative model, on the other hand, does not require recurrence or convolution and treats each data point independently of the rest. Positional information is purposefully incorporated into the model in order to maintain information about the arrangement of words in a phrase. Positional coding is the process of remembering the sequential order of items.

# 2 Question 2

- We need to do sentiment analysis to determine the polarity of a given text for our purposes, so we change the classification head to output only two classes. However, in the first case, the decoder's (language model's) classification task is to predict the best next token from a predefined list of tokens (classification based on previous tokens).

- Language model is the task of predicting the next token based on previous ones, which means classifying words and selecting the one with the highest probability. Classification, which is to assign a class label to examples from the problem domain, is disliked.

# 3 Question 3

- The model Language has 988000 trainable parameters, which we can divide into 20000 = (200 * 100) embedding layer parameters, which are the product of embedding dimension and vocabulary size. Furthermore, we have 242000 parameters for each layer of the transformer encoder, with 160800 = (4 * (200*200 + 200)) parameters coming from the multi head attention part, 80400 = (2 * (200*200 + 200)) parameters coming from the feed forward neural network part, and 800 = (2 * (200 + 200)) parameters coming from the normalization layer part, for a total of 968000 parameters because we have four layers.

- The classification task has $20100 = (200 * 100 + 100)$ parameters comes only from le dense layer.

# 4 Question 4

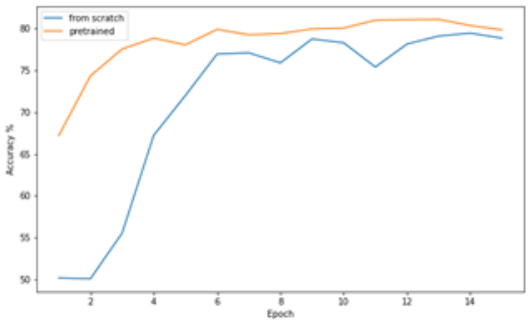


Figure 1: Evolution of accuracy per epoch

- As expected, using pre-trained weights as a starting point results in greater accuracy than starting from scratch. Furthermore, when using pretrained weights, fine tuning begins with a higher accuracy and gradually increases after 15 epochs. However, after a few epochs, both the pretrained and the generated weights converge to the same precision.

# 5 Question 5

- The language model used in the notebook has the limitation of only reading text input sequentially from left to right. As a result, no component of an input sequence contains information from both the past and the present. BERT, on the other hand (presented in the paper), is intended to read in both directions.