Pradel Guillaume
gpradel88@gmail.com

Lab session # 2
ALTEGRAD 2024

10/22/24

1 Question 1

- The mask is a triangular matrix composed of 1 on the lower part and $-\infty$ on the upper part. This makes it possible to ignore future words in the sequence, retaining only information from the "past" during the training phase. For instance, the first row will allow the model only to view the first word, and the last row will be the whole sequence.
- As described in [2], Positional encoding is added to the input embeddings before encoder. The goal of this is to inject information about the position of the words in the sequence. In fact, this information is very relevant because the word's order in a sequence create the meaning of the sequence.

2 Question 2

The *ClassificationHead* is defined here to do sentiment analysis, it will categorize the sentence according to its meaning. On the other side, we have the Transformer model which aims to return a certain token in a list. So we have to replace classification head to make either sentiment analysis or language modelling.

The main difference between the language modeling and the classification tasks :

- The languague modeling aims to predict the next word of a given sequence, it was done here with the sequence : *"Bonjour les"*
- The classification aims to categorize a sequence. Here, it was done with sentiment analysis, where we predicted whether a book review was positive or negative.

3 Question 3

Number of trainable parameters for the language modeling task :

- Embedding : $nhid * ntoken = 20000$
- Positional encoder : nothing to train, so 0
- Transformer Encoder : $nhead * nhid^2 = 80000$, puis on multiplie par le nombre de layers : $nlayers * 80000 = 320000$
- Output : $nhid * ntoken = 20000$

We add all these parameters to find :
Total : 360000 parameters

Number of trainable parameters for the classification task :

- Decoder : $nhid * nclasses = 400$

As seen in the Figure 1 : Visualization of the model, we had the parameters of the decoder to those already calculated.
Total : 340400 parameters

4 Question 4

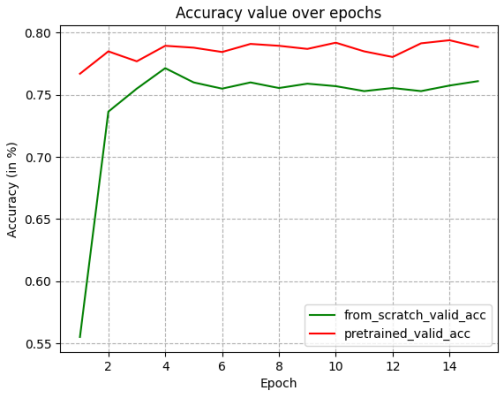

Figure 1: This is a caption.

According to the graph, the accuracy of the pre-trained model is always higher than the accuracy.
As expected, when looking at the first point, we see that the accuracy of the pre trained model is way higher (by 20%). Also, we do not see it here on the graph, but both model should converge to the same accuracy (maybe we need to train it with more epochs).

5 Question 5

One limitation here is that during the pre-training phase, the context a word receives as input is unidirectional, meaning it only gets the words from the right. However, for a word to fully understand the context of a sentence (or document), it is intuitively non-optimal for a model to only provide part of the context. In [1], the masked language model (LM) is introduced to address this issue: the entire 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, 2019.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.