LACOMBE Yoach
yoach.lacombe@gmail.com

1 Question 1

The **square mask**, which is a matrix filled with 1 on the lower triangular part and filled with $-\infty$ in the strictly upper triangular part, imposes the self-attention layer to only look at the earlier positions in the sequence (the model only knows what happened in the past - in an auto-regressive fashion). The model here is supposed to learn from left-to-right because the *language modeling* task would have data leakage otherwise.
The positional encoding helps the model focus on the closest positions in the sequence (the words that are the closest to the current word).

2 Question 2

We have to replace the classification head because we're facing two different classification tasks. The classification layer of a given task doesn't work for a totally different task. One part is because the final layers of a model don't learn the same information to solve two different tasks and the other part is because the number of classes is different.
The *language modeling* task aims to predict the next word of a given sentence (in practice it should predict the same phrase (starting from the second word) and the next word).
The *classification* task aims at predicting the sentiment (positive or negative) of a given sentence.

3 Question 3

Classification head:

- Linear Layer: $(ntokens + 1) * nhid$
- **Numerical application:** For the language modeling head: $200 * 50002 = 10000400$ For the classification head: $200 * 3 = 600$

Base model:

- Embedding: $ntokens * nhid$
- Transformer encoder: $nlayers * [nhid(nhid + 1)(2 + 2nhead) + 4nhead]$
- **Numerical application:** 50001*200 + 4*(200*201*(2+4)+4*200) = 10965016
- **Note:** Each TransformerEncoderLayer (there are $nlayers$ of them) have $2+2nhead$ linear layers(200,200) and 2 normalization layers with $2 * nhead$ parameters.

4 Question 4

As you can see on Fig 1, because it has been pretrained with the language modeling task, the performance of the pretrained model is better than the performance of the model from scratch. Better, the pretrained language reaches performance that the scratch model doesn't achieve, there is a clear gap between the two models. However, in this case, the pretrained model stagnates in terms of accuracy.

5 Question 5

As pointed in the [1], the language objective model only trains from left-to-right whereas the masked language model objective (MLM) trains according to the surrounding context (so in every direction). The main advantage it introduces is that the model really learns from the surrounding context as opposed to the language objective model which learns only from the left context.

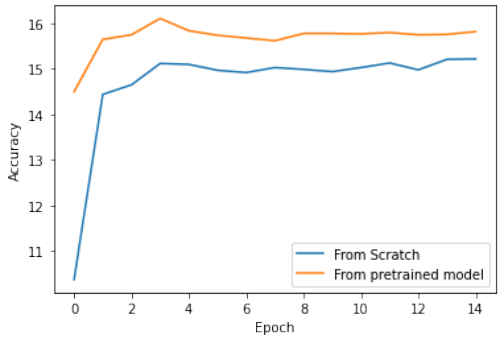

Figure 1: Attention weights.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.