Yunhao CHEN
yunhao.chen@polytechnique.edu

Question 1

We denote by $n_t = 30000$ the number of tokens; by $n_h = 512$ the hidden size of attention layer (the embedding dimension). The dimension of key, query and value is $n_k = n_v = n_q = 64$. Look at Fig. 1

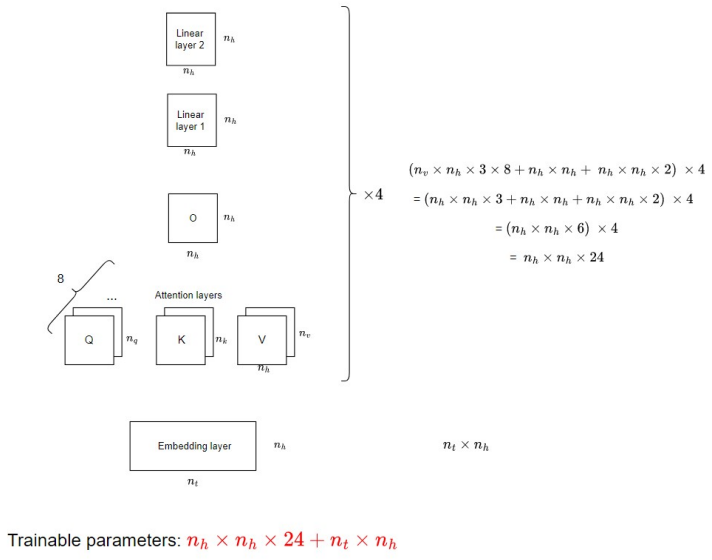

Figure 1: Parameters in RoBERTa

If we omit the bias and normalization layers, the total trainable parameters in RoBERTa model equals to **22675456**; If we count the parameters in positional embedding layer, the total parameters is **22807552**. This result is verified by checking the model architecture in Pytorch.

Task 3 & 4

Finetune the pretrained model: for each seed, I've chosen the checkpoint with best validation accuracy. They give respectively **62, 65, 63** as validation accuracy. Moreover, they give respectively 67.4, 66.7, 64.5 as test accuracy. Therefore, on test set, the average accuracy is **66.2**; and its standard deviation is **1.51**
Finetune a random checkpoint: it gives the same result.

Question 2

Fairseq: need to perform tokenization and binarization.
Hugging face framework: need to convert your data to json format.
Personally, I prefer hugging face framework. Because it is better documented. Compared to HF, fairseq is poorly documented and sometimes we could not find the right parameters (for example, in 'train.py', there is no parameter 'warmup ratio', but only 'warmup-updates' indicating the number of warmup steps. However, the explanation about this point is unclear in the documentation.)