Simon Queric
simon.queric@telecom-paris.fr

Question 1

I took the notations of *Attention Is All You Need* [1]. We have

$$\begin{cases} n_{\text{tokens}} & = 32000 \text{ is the number of tokens in our dictionary.} \\ d_{\text{model}} & = 512 \text{ is the embedding dimension.} \\ N & = 4 \text{ the number of transformer's encoder layers.} \end{cases}$$

Without the biases, we have $d_{\text{model}} \times n_{\text{tokens}}$ parameters for the embedding layer. Then, in each attention head, the model learns 4 matrices W^Q, W^K, W^V, W^O which gives us $N \times 4 \times d_{\text{model}} \times d_k \times h = N \times 4 \times d_{\text{model}}^2$ parameters for the attention layers. The last layer is a linear layer of dimension $n_{\text{tokens}} \times d_{\text{model}}$.
Finally, the number of parameters of our model is :

$$n_{\text{tokens}} \times d_{\text{model}} + 4 \times N \times d_{\text{model}}^2 + n_{\text{tokens}} \times d_{\text{model}} = 36 \times 10^6$$

Question 2

The hyper-parameters of the LoraConfig are the following :

- The rank r of the weight matrices of dense layers, is a small integer. It allows to reduce drastically the number of fine-tuned parameters and to gain in efficiency.
- α is a scale parameter of the weights.
- The target modules is a list of all dense layers in the model.
- The dropout parameter is the probability of removing a weight.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.