Axel Dumont
axel.dumont@ensta-paris.fr

Lab session # 4
ALTEGRAD 2023

11/18/23

1 Question 1

First, in the embedding layer we have with the embedding matrix and the positional embedding matrix (the maximum position is 256 but with the SOS and EOS the dimension becomes 258):

$$n_1 = (N_{Tokens} \times n_{dim}) + (n_{pos} \times n_{dim}) = 32000 \times 512 + 258 \times 512 = 16,516,096$$

Then we have 4 encoder layers with 8 attention heads. Hence for 1 layer we have (ignoring biases and normalization layers) :

$$n_{layer} = n_{dim} \times n_{dim} + (n_{dim} \times \frac{n_{dim}}{n_{heads}}) \times n_{head} \times 3 + 2 \times n_{dim}^2 = 6 \times n_{dim}^2 = 1,572,864$$

Then we have in total :

$$n_{tot} = n_1 + 4 \times n_{layer} = 22,807,552$$

2 Question 2

We have for the parameters,

- r: The rank for the low-rank decomposition of the weight matrices in the target modules (e.g. attention matrices). Lower rank reduces the number of trainable parameters.
- lora_alpha: The coefficient that controls the regularization loss added to penalize the Frobenius norm of the decomposed low-rank matrices. Higher alpha means stronger regularization.
- target_modules: The names of the modules to decompose (e.g. "query_key_value" for the attention matrices).
- lora_dropout: Dropout applied to the low-rank decomposed matrices during training. Can help prevent overfitting.
- bias: The strategy for biases in the low-rank decomposed matrices. "none" means no bias term.
- task_type: The type of task, which determines how the regularization loss is calculated. "CAUSAL_LM" is for causal language modeling tasks.