BENYAMINA Elyas
elyas.benyamina@ens-paris-saclay.fr

Lab session # 4
ALTEGRAD 2023

11/14/23

1 Question 1

For the embedding, we have $32000 \cdot 512 + 512 \cdot (256 + 2) + 2 \cdot 256$ parameters. We have 4 transformer units with $4 \cdot (3 \cdot 512 \cdot 512 + 3 \cdot 512 + 3 \cdot 512 + 6 \cdot 512)$ parameters in total.
So, we finally have 22829056 parameters.
We have one more parameter when checking in code.

2 Question 2

```
config = LoraConfig(
    r=16,
    lora_alpha=32,
    target_modules=["query_key_value"],
    lora_dropout=0.05,
    bias="none",
    task_type="CAUSAL_LM",
)
```

1. `r`: This parameter represents the rank used in the Low-Rank Adaptation (LoRA) technique.
2. `lora_alpha`: This parameter, named `alpha` in the LoRA context, scales the learned weights. This allows us to control how much importance we want to give to low rank approximation versus original prediction.
3. `target_modules`: This parameter is a list of target modules or layers in the model architecture that will undergo LoRA fine-tuning. In this example, we will apply it to Query, Key and Value matrices of the self-attention layer.
4. `lora_dropout`: This parameter represents the dropout rate used during the low rank approximation. Dropout is a regularization technique, and `lora_dropout=0.05` indicates a dropout rate of 5
5. `bias`: This parameter specifies the handling of bias during low rank approximation. In this example, no bias is applied.
6. `task_type`: This parameter indicates the type of task for which the model is being fine-tuned. In the example, `task_type="CAUSAL_LM"` and we are working at Causal Language Modeling task.