RAZAFINDRALAMBO Raphaël
raphael.razafin@gmail.com

Lab session # 4
ALTEGRAD 2023

11/14/23

# 1 Question 1

The number of parameters of each layer is observable during the fine-tuning.
**Number of parameters (without bias, normalization layer and language model head)**

- **Embedding layer:** The dictionary contains 32K tokens turned into a vector embedding of length 512.

$$d_{\text{embedding}} \times n_{\text{tokens}} = 512 \times 32000$$

- **Learned positional embedding:** We denote $n_{\text{positions}}$ the number of positions (number of tokens in a sequence).

$$n_{\text{positions}} \times d_{\text{embedding}} = 258 \times 512$$

- **Transformer encoder layers (here case of 1 layer (/4)):** Dimension of embedding is equal to the hidden dimension. Thus we have
  - **for the multi-head attention layer:**
    - Projection that transform the input embeddings into queries (Q), keys (K), and values (V) for multi-head attention:
      - $3 \times d_{\text{embedding}} \times d_{\text{embedding}} = 3 \times 512 \times 512$ parameters.
    - Projection to produce the output of the layer:
      - $d_{\text{embedding}} \times d_{\text{embedding}} = 512 \times 512$ parameters.
  - **for the linear Feed-forward layers:** For each of the 2 feed-forward layers, we have
    - $d_{\text{embedding}} \times d_{\text{embedding}} = 512 \times 512$ parameters.

Hence the total number of parameters is $512 \times 32k + 258 \times 512 + 4 \times (4 \times 512 \times 512 + 2 \times 512 \times 512) = 22807552$

# 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",
)
```

`r` refers to the rank of the low-rank decomposition matrices of the fine-tuning weights. Here 16 is the rank.
`lora_alpha` scales the learning rate.
`target_modules` defines the components which will be adapted by using LoRA. Here Low-Rank Adaptation is applied to Query,Key,and Value matrices.
`lora_dropout` is the dropout rate that is applied specifically to the Low-rank adaptations.
`bias` defines if the bias should be included in LoRA adaptation. Here they are not.
`task_type` specifies the task on which the model is used. Here it is causal language modelling

# 3 Task 3: Analysis on Fairseq experiments

Pre-trained models (yellow, orange and pink) lead to faster convergence and higher accuracy on the test set. Pretrained models have already learned general language representations, which provide a strong foundation for specific tasks. Randomly initialized models, lacking this prior knowledge, must learn both the general language structure and task-specific features from scratch, making the learning process less efficient. And moreover, the plots appear to be less tightly clustered together. Indeed we have a slightly higher standard deviation between the seeds for the random initialization.

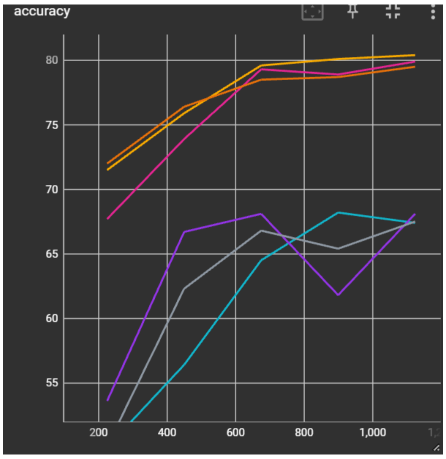


Figure 1: Test accuracy of the six $RoBERTa^{fr}_{small}$ models

| Model | Seed | Best Validation (Step, Accuracy) | Test (%) | Mean / Std |
|---|---|---|---|---|
| $RoBERTa^{fr}_{small}$ | 0 | Training step=1125, 81.5% | 79.5 | Mean: 0.794%, Std: 0.0066 |
| | 1 | Training step=675, 80% | 78.5 | |
| | 2 | Training step=900, 83.5% | 80.1 | |
| Random $RoBERTa^{fr}_{small}$ | 0 | Training step=900, 62% | 65.4 | Mean: 0.665%, Std: 0.0083 |
| | 1 | Training step=450, 65% | 66.7 | |
| | 2 | Training step=1125, 63% | 67.4 | |

Table 1: Results on test sets for $RoBERTa^{fr}_{small}$ et Random $RoBERTa^{fr}_{small}$

## 4 Task 5: Analysis of HuggingFace experiment

The range of values for accuracy on validation and test sets are the same as we observed in fairseq part, which are around 80%. The convergence speeds are also the same, with at least 400 iterations to get at least 70% accuracy. In order to additionally have the graph for test accuracy 2, I put the test file as the value of the parameter `validation_file` during fine-tuning.

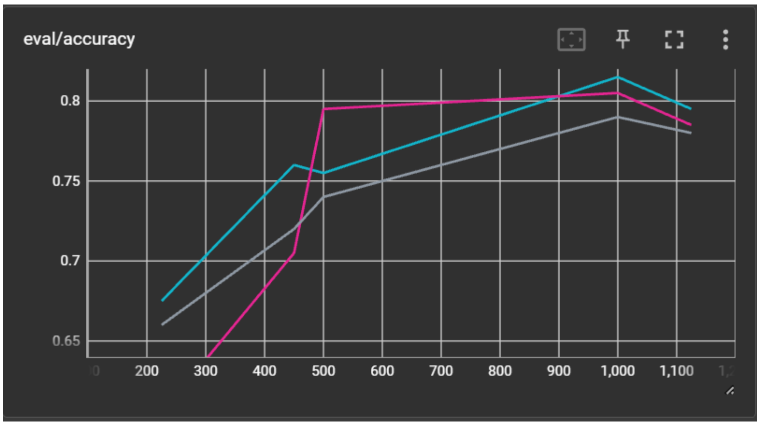


Figure 2: Validation accuracy of the $RoBERTa^{fr}_{small}$ models (3 seeds) from HuggingFace

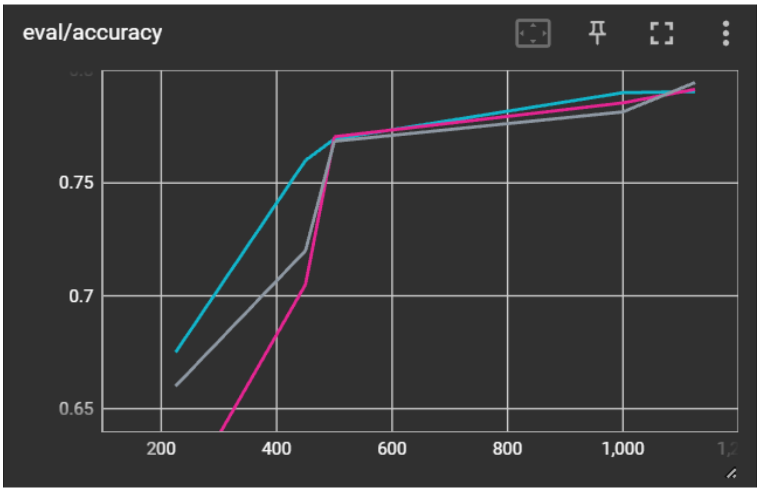


Figure 3: Test accuracy of the $RoBERTa^{fr}_{small}$ models (3 seeds) from HuggingFace

| Model | Seed | Best Validation (Step, Accuracy) | Test (%) | Mean / Std |
|---|---|---|---|---|
| $RoBERTa^{fr}_{small}$ | 0 | Training step=1000, 79% | 78.15 | Mean: 0.794%, Std: 0.0066 |
| | 1 | Training step=1000, 81.5% | 79 | |
| | 2 | Training step=1000, 80.5% | 78.55 | |

Table 2: Results on test sets for $RoBERTa^{fr}_{small}$ et Random $RoBERTa^{fr}_{small}$