YOU Borachhun
borachhunyou@gmail.com

Lab session # 4
ALTEGRAD 2023

11/12/23

1 Question 1

The number of parameters of the $RoBERTa^{fr}_{SMALL}$ model:

- Token embedding layer: vocabulary size of 32000 tokens with embedding dimension of 512 ($32000 \times 512 = 16384000$ parameters)
- Learned positional embedding layer: $258 \times 512 = 132096$ parameters
- 4 encoder layers, each with:
  - Self-attention mechanism: projection of Q, K, V ($3 \times 512 \times 512 = 786432$ parameters) and projection of the result ($512 \times 512 = 262144$ parameters)
  - 2 fully connected layers: $2 \times 512 \times 512 = 524288$ parameters

In total, there are 22807552 parameters.
Note: Biases and parameters of normalization layers and the model's head are excluded.

2 Accuracy of fine-tuning with Fairseq

Figure 1 shows that in general the accuracy has an increasing trend throughout the epochs for both the pretrained and the random model. Furthermore, it appears that the accuracy of the pretrained model is much higher than the random model, which is true from the first to the last epoch. Table 1 and 2 show that the same is true for the average accuracy. In addition, for both models the standard deviation of the accuracy decreases throughout the epochs.

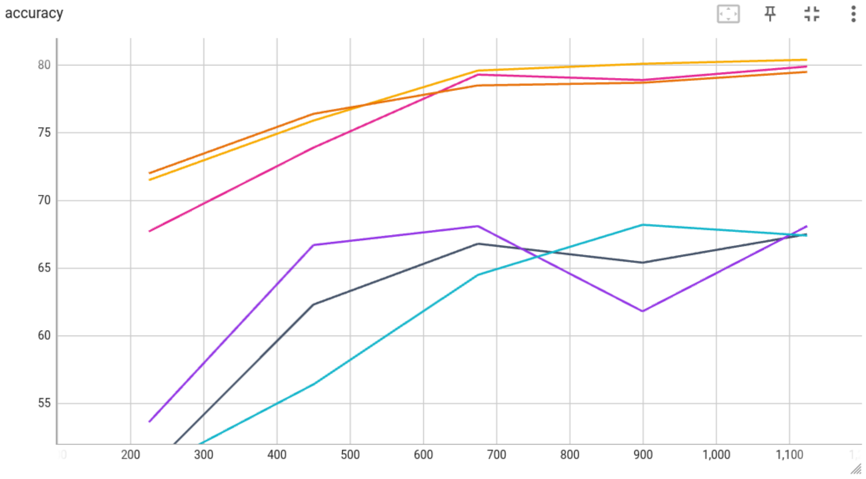

Figure 1: Visualization of the accuracy of the fine-tuned pretrained and random model over 3 runs (Fairseq)

Epoch	1	2	3	4	5
Accuracy on run 1	67.7	73.9	79.3	78.9	79.9
Accuracy on run 2	72	76.4	78.5	78.7	79.5
Accuracy on run 3	71.5	75.9	79.6	80.1	80.4
Average accuracy	70.4	75.4	79.13	79.23	79.93
Standard deviation of accuracy	1.92	1.08	0.46	0.61	0.36

Table 1: Accuracy of the fine-tuned pretrained model over 3 runs (Fairseq)

Epoch	1	2	3	4	5
Accuracy on run 1	50.1	62.3	66.8	65.4	67.5
Accuracy on run 2	53.6	66.7	68.1	61.8	68.1
Accuracy on run 3	50	56.4	64.5	68.2	67.4
Average accuracy	51.23	61.8	66.46	65.13	67.66
Standard deviation of accuracy	1.67	4.22	1.49	2.62	0.31

Table 2: Accuracy of the fine-tuned random model over 3 runs (Fairseq)

3 Accuracy of fine-tuning with Hugging Face's transformers

Table 3 shows that the average of the accuracy has an upward trend while the standard deviation has a downward trend throughout the epochs, which is the same as Fairseq. Moreover, at the last epoch the model has 1% less average accuracy compared to Fairseq, and has a higher standard deviation of the accuracy.

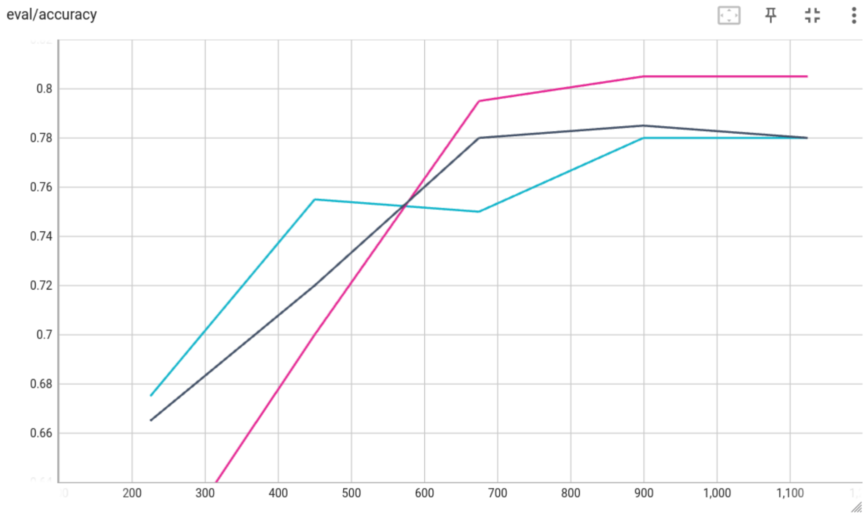

Figure 2: Visualization of the accuracy of the fine-tuned pretrained model over 3 runs (Hugging Face's transformers)

Epoch	1	2	3	4	5
Accuracy on run 1	66.5	72	78	78.5	78
Accuracy on run 2	67.5	75.5	75	78	78
Accuracy on run 3	60	70	79.5	80.5	80.5
Average accuracy	64.66	72.5	77.5	79	78.83
Standard deviation of accuracy	3.32	2.27	1.87	1.08	1.18

Table 3: Accuracy of the fine-tuned pretrained model over 3 runs (Hugging Face's transformers)

4 Question 2

Parameters used in LoraConfig:

- `r=16`: Rank of the smaller update matrices from low-rank decomposition. Lower rank results in smaller matrices, while higher rank opposes LoRA's efficiency gains.
- `lora_alpha=32`: LoRA scaling factor.
- `target_modules=["query_key_value"]`: Names of the modules to apply LoRA, for example, attention blocks.
- `lora_dropout=0.05`: Dropout probability for Lora layers.
- `bias="none"`: Specifies if the bias parameters should be trained.
- `task_type="CAUSAL_LM"`: Type of task to be performed, which is causal language modeling in this case.