Large Language Models (Assignment-3)

Github Link: https://github.com/skyscrappers/LLM/tree/main/LLM_A3

Finetuning Phi-2

Results and Approach:

Used **AutoModelforSequenceClassification** for classifying the relationship between the premise and hypothesis for each sample, which will among {0: 'entailment', 1: 'neutral', 2: 'contradiction'}.

LoRA Parameters:

r=16,
lora_alpha=64,
lora_dropout=0.05,

Selected 1000 training, 100 testing and 100 validation samples as mentioned from complete dataset.

Training Parameters:Why not corrected?

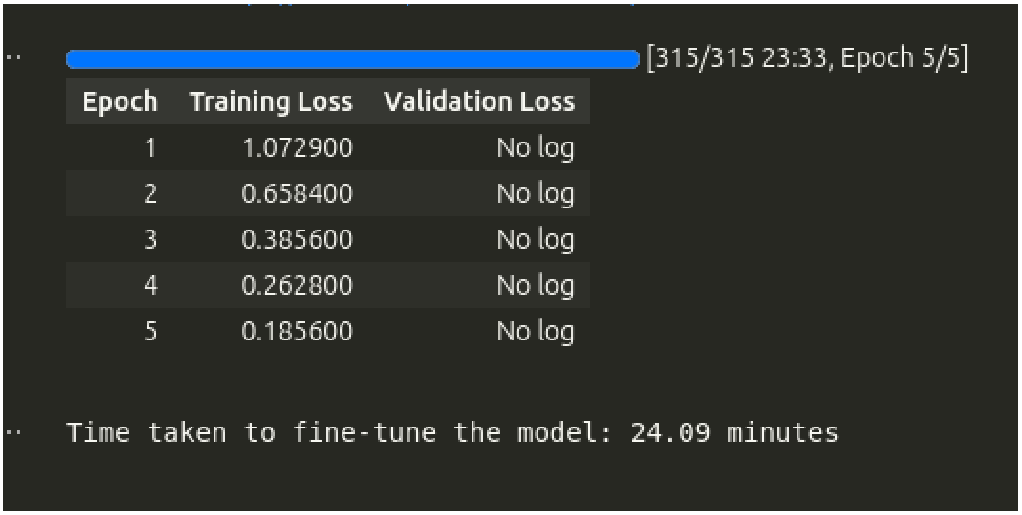
per_device_train_batch_size=16,
per_device_eval_batch_size=16,
num_train_epochs=5,
learning_rate=0.0001,
eval_strategy="epoch",
save_strategy="epoch",
logging_strategy="epoch",

Part 1: Accuracy comparison between the pretrained and fine-tuned models on the test set.

	Training set	Validation set	Test set

Pretrained Model	0.348	0.33	0.33
Fine tuned Model	0.943	0.82	0.85

Part 2: Time taken to fine-tune the model using QLoRA.



[315/315 23:33, Epoch 5/5]

Epoch	Training Loss	Validation Loss
1	1.072900	No log
2	0.658400	No log
3	0.385600	No log
4	0.262800	No log
5	0.185600	No log

```
Time taken to fine-tune the model: 24.09 minutes
```

Time taken: 24.09 minutes

Part 3: Total parameters in the model and the number of parameters fine-tuned.

Total Parameters: 1408634880

Trainable Parameters: 18350080

Trainable percentage: 1.3026853346127565%

Part 4: Resources used (e.g., hardware, memory) during fine-tuning.

GPU P100 (Kaggle) - 16 GB

CPU - Kaggle CPU

RAM - 30 GB

Maximum GPU usage during the fine tuning was 14.9GB.

Part 5:

Failure cases of the pretrained model that were corrected by the fine-tuned model, as well as those that were not corrected. Provide possible explanations for both.

Corrected cases:

```
"{'premise': 'A woman within an orchestra is playing a violin.',
'hypothesis': 'A woman is playing the violin.'}, 'label': 0}",1,0,0
"{'premise': 'many children play in the water.',
'hypothesis': 'The children are playing mini golf.', 'label': 2}",0
"{'premise': 'A female softball player wearing blue and red crouche
the infield, waiting for the next play.',
'hypothesis': 'the player is flying planes', 'label': 2}",1,2,2,2
"{'premise': 'Children bathe in water from large drums.',
'hypothesis': 'The kids are wet.', 'label': 0}",1,0,0,0
"{'premise': 'People are all standing together in front of a statue
and they are all wearing cool-weather clothing.',
'hypothesis': 'A beautiful statue of a man.', 'label': 2}",1,2,2,2
```

Not corrected cases:

```
"{'premise': 'A Skier ski-jumping while two other skiers watch his
'hypothesis': 'A skier preparing a trick', 'label': 0}",1,0,1,0
"{'premise': 'A woman is standing near three stores, two have beaut
and the other store has Largo written on it.',
'hypothesis': 'A woman standing on a street corner outside beside t
stores, two of which contain beautiful artwork and one with a Largo
"{'premise': 'An Ambulance is passing a man wearing a bandanna and
'hypothesis': 'The man in the bandana is running after the ambulanc
"{'premise': 'Two middle-aged police officers watch over a parking
 'hypothesis': 'The officers are actually security guards.', 'label
"{'premise': 'Group of young adults posing for picture near spanish
'hypothesis': 'The people are taking a science test.', 'label': 2}'
```

Why corrected?

For example, in the case where the **premise** is "A woman within an orchestra is playing a violin" and the **hypothesis** is "A woman is playing the violin," the pretrained model incorrectly labeled this as a contradiction, likely due to its difficulty in recognizing the synonymous phrasing. The fine-tuned model, however, correctly identified this as an entailment, showing improved contextual understanding and sensitivity to paraphrasing through fine-tuning.

The fine-tuned model likely became better at recognizing when two sentences convey the same meaning, even if they're phrased differently. For instance, "A woman within an orchestra is playing a violin" and "A woman is playing the violin" mean the same thing, but the pretrained model might have struggled with this level of paraphrasing

Why not corrected?

In the case where the **premise** is "A Skier ski-jumping while two other skiers watch his act" and the **hypothesis** is "A skier preparing a trick," both models incorrectly labeled this as a contradiction rather than the correct label of entailment. This error likely stems from the ambiguity in interpreting the skier's action and intent, requiring a more nuanced understanding of the situation that goes beyond straightforward language matching, which fine-tuning alone could not resolve.

If the fine-tuning dataset lacks enough examples of certain complex or rare cases, the model may not learn to handle them effectively. Fine-tuning can help with common patterns, but outliers or less frequently occurring types of contradictions might remain problematic.

Possible Solution:

Fine tuned model may perform better if more data is provided because it will prevent problems like overfitting.