LLM Assignment 3

Manav Mittal
CSAI, 2021538

I have used **AutoModelForSequenceClassification** for classifying the given input, i.e. premise and hypothesis, as one of entailment, neutral or contradiction.

Part 1

The accuracy score is out of 1. To get the %age accuracy, multiply the score by 100.

	Train	Validation	Test
Pre-trained	0.327	0.31	0.37
Finetuned	0.94	0.77	0.86

Part 2

The time taken to fine-tune the model was **30 min 55 sec.**

Part 3

Total Model parameters = 1408634880 (approx. 1.4 billion)

Trainable Model parameters = 18357760 (approx. 18 million)

Percentage of trainable model parameters: 1.30%

Part 4

Resources used were

1. GPU P100 (Kaggle) - 16 GB
2. CPU - Kaggle CPU
3. RAM - 30 GB

During fine-tuning, the maximum GPU memory used was 8.5 GB. Maximum RAM usage was about 8 GB. Both CPU and GPU were operating on 100% utilization.

Part 5

Pretrained v/s Finetuned

Initially, the classification head of the Phi-2 is not trained, due to which it is generating random guesses, so we are getting an accuracy of around 0.33. But when we finetune the model over the dataset, it now knows which label is which and then can successfully distinguish among relationships neutral, entailment and contradiction. Thus, we are getting a testing accuracy of 0.86 after finetuning.

Why fine-tuned model is still wrong?

```
Premise: This church choir sings to the masses as they sing joyous songs from the book at a church.
Hypothesis: The church has cracks in the ceiling.
Label: Neutral
Label Finetuned: Contradiction

Premise: A Skier ski-jumping while two other skiers watch h is act.
Hypothesis: A skier preparing a trick
Label: Entailment
Label Finetuned: Neutral

Premise: An Ambulance is passing a man wearing a bandanna a nd girl.
```

```
Hypothesis: The man in the bandana is running after the amb
ulance
Label: Contradiction
Label Finetuned: Neutral
```

Some of the training data is ambiguous. Considering the second example we can see that it is difficult even for us to tell the label. Basically, learning the context is difficult in this case. In my opinion, if there was the word “stunt” in place of the word “trick”, then the model might have performed better. All the other samples are also difficult to predict.

```
Premise: A woman is standing near three stores, two have be
autiful artwork and the other store has Largo written on i
t.
Hypothesis: A woman standing on a street corner outside bes
ide three different stores, two of which contain beautiful
artwork and one with a Largo sign.
Label: Entailment
Label Finetuned: Neutral
```

In the case of this example, the reason for incorrect prediction might be model was not able to preserve the context since the text was very long.

Mitigation

After fine-tuning, we saw that the training accuracy was near 0.94 while testing and validation were 0.86 and 0.77, respectively. This implies that the model is overfitted on the data. Thus, if we provide more training data to the model, it might perform better.