Assignment 3
Fine-Tuning LLM

Name: Drishti Singh
Roll No: MT23117

I. Introduction

In this assignment, the Microsoft Phi 2 model is fine-tuned on the SNLI (Stanford Natural Language Inference) dataset for the task of Natural Language Inference (NLI). The primary goal is to predict the relationship between pairs of sentences, which can be classified as entailment, neutral, or contradiction. The fine-tuning process employs QLoRA, a method designed to reduce the number of trainable parameters, thereby optimizing resource usage during training.

II. Dataset Collection

The dataset was collected following the guidelines outlined in the assignment. The splits are as follows:

- Training Set: 1000 samples, obtained by selecting every 550th sample from the SNLI dataset.
- Validation Set: 100 samples, collected by selecting every 100th sample.
- Test Set: 100 samples, also chosen using every 100th sample.

The fine-tuning process involves training the model for 5 epochs on the specified training and validation datasets, containing 1001 and 101 entries, respectively. The model is subsequently evaluated on the test set, which comprises 101 entries. The labels are defined as follows:

- 0:Entailment
- 1: Neutral
- 2: Contradiction

III. Accuracy Comparison

The following accuracies were recorded for both the pre-trained and fine-tuned models:

- Pre-trained Model Accuracy: 67.33%
- Fine-tuned Model Accuracies:
  - Epoch 1: 54.98%
  - Epoch 2: 60.07%
  - Epoch 3: 69.54%
  - Epoch 4: 67.01%

- Epoch 5: 70.34%

IV. Training Time and Loss

The time taken for fine-tuning the model using QLoRA for each epoch is as follows:

- Epoch 1/5: Loss: 0.0943, Time taken: 1121.98 seconds
- Epoch 2/5: Loss: 0.0137, Time taken: 1121.90 seconds
- Epoch 3/5: Loss: 0.0127, Time taken: 1121.67 seconds
- Epoch 4/5: Loss: 0.0123, Time taken: 1121.59 seconds
- Epoch 5/5: Loss: 0.0118, Time taken: 1121.57 seconds

The training time for each epoch remains approximately constant at around 1121 seconds (~18.7 minutes). However, the loss values demonstrate a decreasing trend, indicating that the model is learning effectively as training progresses.

V. Parameter Analysis

The implementation of QLoRA resulted in a significant reduction in the number of parameters being trained, which is detailed below:

- Total parameters in the model: 2,798,033,920
- Number of parameters being fine-tuned: 18,350,080
- Percentage of parameters fine-tuned: 0.66%

This reduction highlights the efficiency of QLoRA in minimizing the computational load while retaining model performance.

VI. Resources Used

The model fine-tuning was performed in the Kaggle environment with the following specifications:

- GPU: NVIDIA P100
- GPU Memory: 16 GB
- CUDA Cores: 3584
- Tensor Cores: 240
- FP16 Performance: Approximately 18.7 TFLOPS
- Architecture: Pascal

VII. Failure Cases

During the evaluation, certain failure cases were observed

Premise	Hypothesis	Label	Label by Fine Tuned Model
A woman within an orchestra is playing a violin.	A woman is playing the violin.	0	1
Two men climbing on a wooden scaffold.	Two sad men climbing on a wooden scaffold.	1	2
Two men in neon yellow shirts busily sawing a log in half.	Two men are cutting wood to build a table.	1	1