Assignment: Fine-Tuning a Language Model using LoRA and 4-bit Quantization

ES667: Deep Learning

April 3, 2025

1 Overview and Objectives

Goal:

– Fine-tune a pre-trained language model (approximately 1B parameters) on a chosen dataset.

– Evaluate the trained model with both quantitative and qualitative methods.

Learning Outcomes:

– Gain hands-on experience with the Transformers library for fine-tuning.

– Analyze the impact of fine-tuning on model performance.

– Learn to train LLMs with limited GPU resources using Low Rank Adaptation (LoRA) and 4-bit Quantization.

2 Model Selection & Fine-Tuning

2.1 Model Selection

Task: Choose a pre-trained language model with roughly 1B parameters from Hugging Face. **Do not select a model which is already fine-tuned**.

2.2 Data Preparation

Task:

– Identify and load a fine-tuning text dataset of your choice from the Hugging Face datasets library.

– Preprocess the data according to your need(tokenization, filtering, etc.). You dont have to use the entire dataset as it will take a lot of time and compute. Only use a portion of the dataset which is manageable.

2.3 Fine-Tuning

Task: Fine-tune your selected model using the Transformers library. You may use the Hugging Face Trainer API or write a custom training loop.

Requirements:

– Clearly define training hyperparameters (learning rate, batch size, epochs).

– Save the fine-tuned model for later evaluation.

– Document training progress (loss curves, validation performance, etc.).

– Due to minimal gpu availability, use Low Rank Adaptation (LoRA) to update only a fraction of the model parameters.

– You should also use 4-bit Quantization to load the model in lower precision.

3 Part 3: Evaluation of the Trained Model

3.1 Quantitative Evaluation

Task: Evaluate both the pre-trained model and your fine-tuned model on a held-out validation set using standard language modeling metrics (e.g., perplexity, accuracy etc). Does your fine-tuned model perform better?

3.2 Qualitative Evaluation

Task:

– Generate 10 sample outputs from both your fine-tuned model and the pre-trained model.

– Evaluate these outputs by yourself focusing on fluency, relevance, and correctness. Note that this approach is known as human evaluation. You should compare the 10 outputs of both models by yourself and carefully grade them (forexample, give each of them a score between 0 and 5).

4 Compute Requirements

- One goal of this assignment is to learn how to train Large Language Models with limited GPU resources.
- The free T4 colab gpu should be sufficient for all the above computations by utilizing Low Rank Adaptation (LoRA), 4-bit Quantization and a small batch size.

5 Submission Guidelines

– Submit a Jupyter Notebook with clear documentation and comments.