Generative AI (Spring-2025)

Assignment-2

Instructor

Dr. Hajra Waheed PHD

Submission Guidelines:

- Submit your assignment on Google Classroom in the format "20XX.ipynb".
- The deadline is Apr 12, 2025, at 11:59 PM. No extensions will be granted.

Declarations:

- Late submissions will incur penalties: 25% deduction on the first day, 50% on the second day, and zero marks thereafter.
- Plagiarism will result in zero marks for the assignment.
- This is an individual assignment; collaboration or group work is strictly prohibited.
- Please ensure that you submit your own original work.

VIVA Policy:

- A VIVA (oral examination) will be conducted to assess your understanding of the assignment.
- The VIVA will be scheduled separately, and you will be notified of the date and time.
- Failure to attend the VIVA will result in zero marks for the assignment.

Academic Integrity:

- Plagiarism, collusion, and academic dishonesty will not be tolerated.
- Any instances of academic misconduct will be reported to the authorities and may result in severe penalties.

Objective

The objective of this assignment is to implement and compare four fine-tuning techniques for transformer-based models using a small-scale sentiment classification task. The assignment focuses on evaluating:

- Model accuracy
- Number of trainable parameters
- Training time
- GPU memory usage

This comparison will help identify trade-offs between Full Fine-Tuning and Parameter-Efficient Fine-Tuning (PEFT) methods, including LoRA, QLoRA, and Adapters (IA3).

Model

Roberta-Base from hugging face

Dataset

- **Dataset**: IMDb Sentiment Classification
- **Total Samples**: 5000 (3000 for training, 2000 for testing)
- **Source**: IMDb on Hugging Face

Tips

- Use **`nvidia-smi`** to measure GPU and memory usage
- Keep all training epochs short (e.g., **3–5**) to save time and maintain fairness
- Use Hugging Face **`Trainer`** API for consistency

Part 1 Data Preprocessing (10 marks)

1. Load the IMDb dataset using the Hugging Face **datasets** library.
2. Tokenize the dataset using the **model** tokenizer.
3. Pad and truncate sequences appropriately.

Part 2 Model Implementation (40 marks)

Implement and fine-tune the following models:

Method 1: Full Fine-Tuning (10 marks)

Method 2: LoRA Fine-Tuning using PEFT (10 marks)

Method 3: QLoRA Fine-Tuning (10 marks)

- Quantize the base model to 4-bit precision.
- Apply LoRA on top and fine-tune accordingly.

Method 4: Adapter Tuning (IA3) (10 marks)

- Insert IA3-style adapters and fine-tune them while keeping the rest of the model frozen.

Part 3 Evaluation Metrics (10 marks)

For each method, record and compare the following:

- Accuracy on the test set
- Number of trainable parameters
- Training time in seconds
- GPU memory usage

Part 4. Visualization (10 marks)

Generate comparative bar charts illustrating:

- Accuracy
- Training time
- Number of trainable parameters
- GPU memory usage

Part 5. Analysis and Discussion (10 marks)

- Analyze the trade-offs between the four methods in terms of resource usage, scalability, and performance.
- List Best Use Cases for each method.

Submission Requirements

A. Notebook: 20XX**`.ipynb`** **(80 marks)**

- Proper implementation of all four methods
- Results and metrics clearly shown
- Bar chart visualizations included

B. Report in IEEE Format: 20XX**`_Report.pdf`** **(20 marks)**

Sections:

1. Abstract (5 marks)
   - Brief overview of the task, methods used, and key findings
2. Introduction (5 marks)
   - Motivation for PEFT
   - Brief explanation of Full Fine-Tuning, LoRA, QLoRA, and IA3
3. Experimental Setup (5 marks)
   - Dataset description
   - Hardware used
   - Hyperparameters and configuration details
4. Results and Visualizations (10 marks)
   - Comparative charts and metrics
   - Screenshots of results
5. Analysis and Discussion (10 marks)
   - Insights into performance vs efficiency trade-offs
   - Use case-based recommendations
6. Conclusion and Recommendation (3 marks)
7. References (IEEE Style) (2 marks)