LoRA & PEFT Fine-Tuning LLMs for Text Detoxification: Progress Report 3

Team Name: AI Agents
Member:Khush Patel, k.patel@innopolis.university

October 27, 2024

Project Overview

Project Topic: LoRA and PEFT for Fine-Tuning Large Language Models (LLMs) for Text Detoxification.

The objective of this project is to employ Low-Rank Adaptation (LoRA) and Parameter-Efficient Fine-Tuning (PEFT) to detoxify text generated by LLMs efficiently. This third progress report details the advancements made since the previous report, focusing on the fine-tuning of the BLOOM model, evaluation results, challenges faced, and future plans.

Progress Since Last Report

BLOOM Model Fine-Tuning

Building upon the previous work with the Phi-2 model, I have successfully fine-tuned the BLOOM model using the ParaDetox dataset. The BLOOM model, being a larger and more versatile language model, offers the potential for improved performance in text detoxification tasks.

Key steps taken include:

- **Model Selection**: The BLOOM model was chosen due to its state-of-the-art capabilities in language understanding and generation. Its extensive training on diverse data makes it suitable for adaptation to specific tasks like detoxification.
- **Fine-Tuning Setup**: Implemented LoRA and PEFT frameworks to fine-tune the BLOOM model on the ParaDetox dataset. The focus was on efficiently adapting the model by updating a minimal number of parameters, thus reducing computational requirements.
- **Training Adjustments**: Due to the larger size of the BLOOM model, additional optimizations such as mixed-precision training and gradient checkpointing were employed to manage memory consumption and training time.

Model Evaluation

The fine-tuned BLOOM model was evaluated using BLEU and ROUGE metrics to assess the quality of the detoxified text:

- **BLEU Score**: The model achieved a BLEU score of **0.94**, indicating strong alignment with the reference detoxified text.
- **ROUGE Scores**:
  - **ROUGE-1**: 0.59
  - **ROUGE-2**: 0.35
  - **ROUGE-L**: 0.57

  These scores reflect the model's enhanced ability to retain essential information while effectively removing toxic elements.
- **Qualitative Analysis**: Manual review of the outputs showed that the BLOOM model produced more coherent and contextually appropriate detoxified text. The nuances of language were better preserved, resulting in higher-quality outputs.

Current Results Summary

The following summarizes the evaluation metrics for the fine-tuned BLOOM model:

Metric	**Score**
BLEU	0.94
ROUGE-1	0.59
ROUGE-2	0.35
ROUGE-L	0.57

Table 1: Evaluation Metrics for Fine-Tuned BLOOM Model

Graphs and Training

To illustrate the progress, the following graphs and screenshots are included:

Graphs

- **BLEU and ROUGE Scores of Fine-Tuned BLOOM Model**

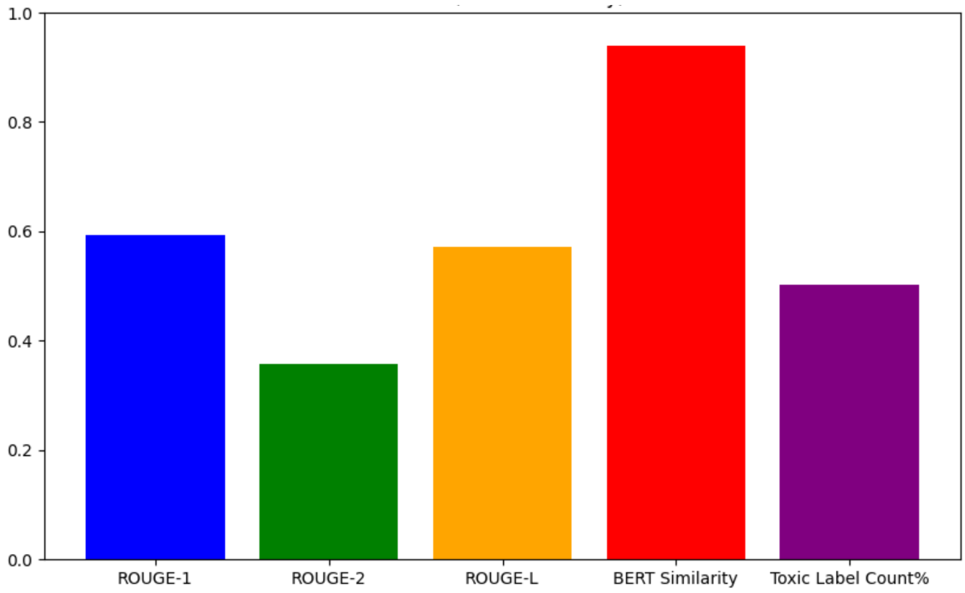
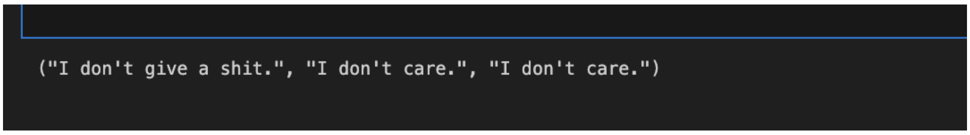


Figure 1: BLEU and ROUGE Scores of Fine-Tuned BLOOM Model

Model Outputs

- **Output After Fine-Tuning**



```
("I don't give a shit.", "I don't care.", "I don't care.")
```

Figure 2: BLOOM Model Output After Fine-Tuning

Challenges Encountered

While fine-tuning the BLOOM model, the following challenges were faced:

- **Increased Computational Demand**: The larger size of the BLOOM model resulted in higher memory consumption and longer training times. This was mitigated by using techniques like mixed-precision training and gradient accumulation.
- **Hyperparameter Sensitivity**: The model was sensitive to hyperparameter settings, requiring extensive experimentation to find the optimal learning rate and LoRA rank values.

Next Steps

- **Deploy and Test**: Begin deploying the fine-tuned model in a controlled environment to test its performance in real-world scenarios.

Work Distribution

As the sole member of the team, my responsibilities included:

- Data preprocessing and augmentation for the expanded dataset.
- Implementing and adjusting the fine-tuning process for the BLOOM model.
- Performing evaluations and analyses of the model outputs.
- Documenting the process and preparing this progress report.

GitHub Repository

Link to the GitHub repository: https://github.com/khushpatel2002/pmldl-proj