LoRA PEFT Fine-Tuning LLMs for Text Detoxification: Progress Report 2

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

October 6, 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 second progress report details the advancements made since the initial report, including model fine-tuning, evaluation results, encountered challenges, and future steps.

Progress Since Last Report

Phi-2 Model Fine-Tuning

In the previous report, I outlined my intent to integrate LoRA and PEFT for model fine-tuning. Since then, I have successfully fine-tuned the Microsoft Phi-2 model using the ParaDetox dataset. LoRA and PEFT methods were leveraged to enhance the detoxification capabilities of the model while keeping computational costs minimal by updating only a small fraction of the model's parameters.

Key steps taken include:

- Model Selection: The Microsoft Phi-2 model was selected for fine-tuning due to its performance in text generation and its efficient architecture, making it well-suited for lightweight adaptation methods like LoRA and PEFT.

```
# Load Model and Tokenizer
save_name = 'models/phi-2-lora'
model_name = 'microsoft/phi-2'
model = AutoModelForCausalLM.from_pretrained(model_name, torch_dtype=torch.float16, trust_remote_code=True)
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)

tokenizer.pad_token = tokenizer.eos_token
```

- Fine-Tuning Setup: LoRA and PEFT frameworks were implemented, and the model was trained on the ParaDetox dataset to specifically focus on detoxifying toxic content. The process involved minimizing the toxicity of generated text while ensuring that the semantic meaning was preserved.

Model Evaluation

The fine-tuned Phi-2 model was evaluated using BLEU and ROUGE scores to assess the quality and fluency of the detoxified text:

- BLEU Score: The model achieved a BLEU score of 0.251, indicating promising similarity between the generated detoxified text and the reference text.
- ROUGE-1, ROUGE-2, ROUGE-L Scores: The scores were 0.572, 0.433, and 0.563 respectively. These scores reflect the model's effectiveness in retaining key information and context while neutralizing toxic elements.
- Qualitative Analysis: Manual inspection of generated samples showed that the model successfully transformed offensive text into neutral language while preserving the original message. This is a crucial outcome, as it ensures that the content remains usable while eliminating potentially harmful language.

Current Results Summary

The following tables summarizes the evaluation metrics for the Phi-2 model fine-tuned using LoRA/PEFT AND Baseline:

Figure 1: Matrix

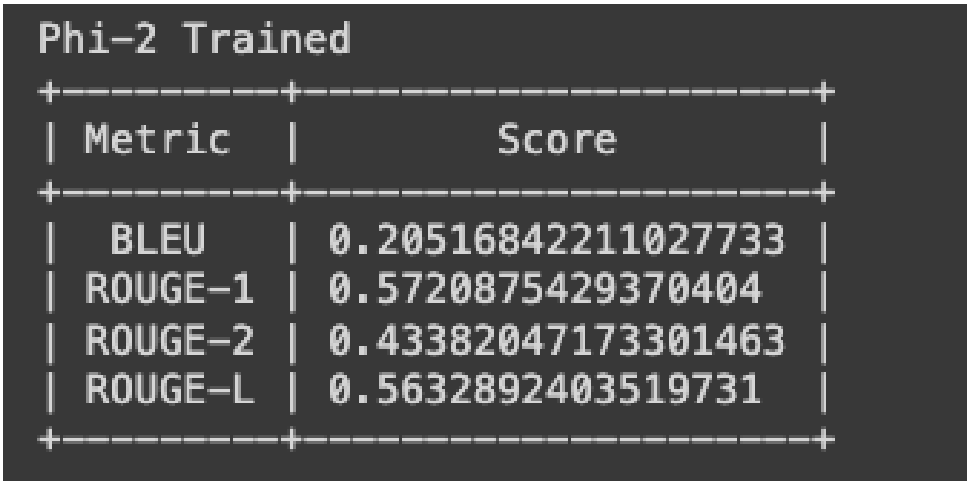


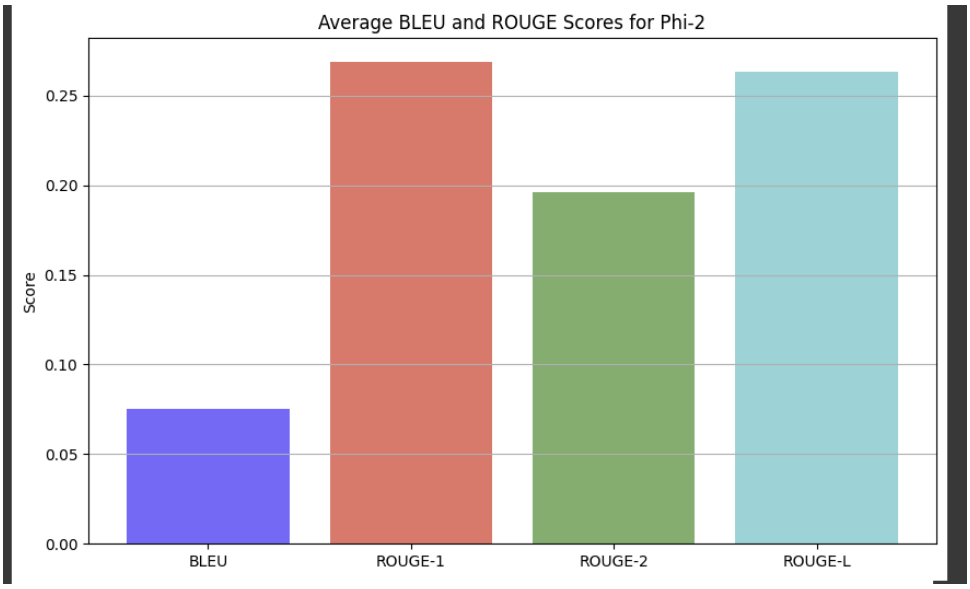
Figure 2: Matrix

Graphs and Training

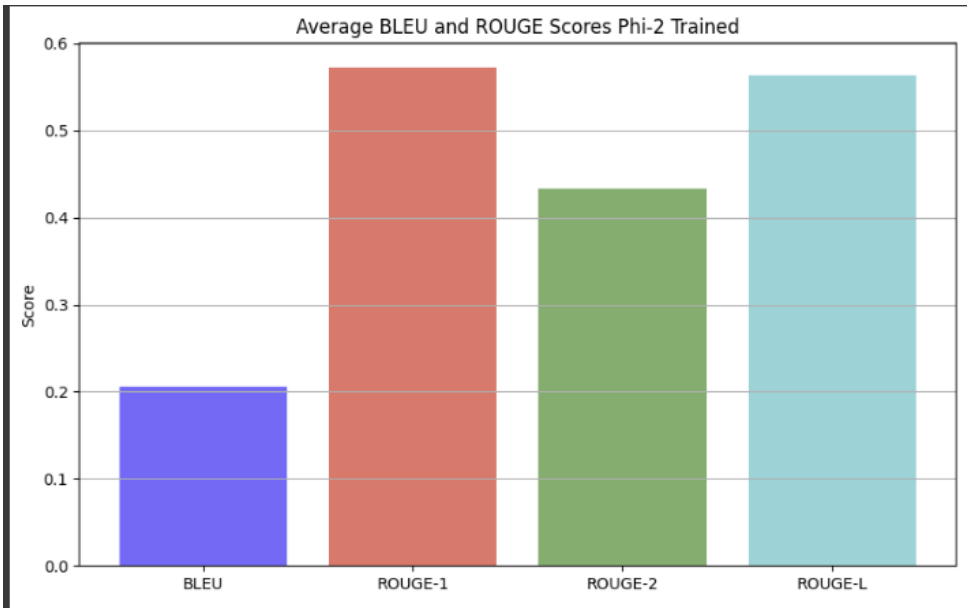
To provide a better visual understanding of the progress made, I have included the following graphs and screenshots:

Graphs

- BLEU and ROUGE Scores of Phi-2 Baseline:



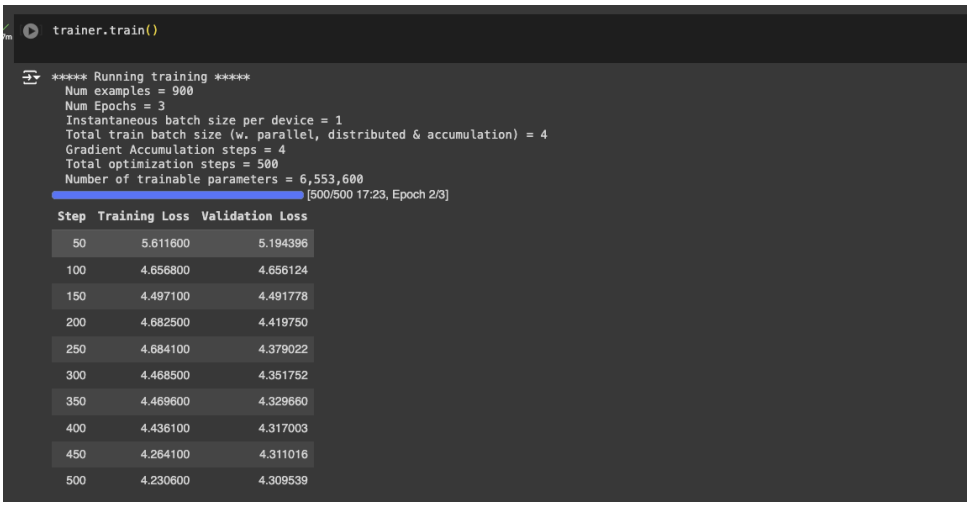
- BLEU and ROUGE Scores of Fine-Tuned Phi-2 Model.



Training

Training phase are included below to document the training process visually:

- Training Progress Logs:



```
trainer.train()

***** Running training *****
  Num examples = 900
  Num Epochs = 3
  Instantaneous batch size per device = 1
  Total train batch size (w. parallel, distributed & accumulation) = 4
  Gradient Accumulation steps = 4
  Total optimization steps = 500
  Number of trainable parameters = 6,553,600
[500/500 17:23, Epoch 2/3]
```

Step	Training Loss	Validation Loss
50	5.611600	5.194396
100	4.656800	4.656124
150	4.497100	4.491778
200	4.682500	4.419750
250	4.684100	4.379022
300	4.468500	4.351752
350	4.469600	4.329660
400	4.436100	4.317003
450	4.264100	4.311016
500	4.230600	4.309539

Outputs

Outputs Comparison

- Without Training.

- After Finetuning.

	Input	Prediction	Ground Truth
0	hahaha that was so funny ! major lick ass alerttopstalker	haha that was so funny! major talker	hahaha that was so funny ! major alert tops talker
1	who in their right mind would honestly think this orange pig would not lie under oath ?	who in the right mind honestly think this person would not violate the oath?	who would honestly think this person would not violate the oath?
2	but we do have fun block parties n shit ! !	but we do have any fun parties	but we do have fun block parties
3	dude wasn 't complying for shit .	dude didn't agree for this.	Dude wasn't complying for this.
4	have you done some type of survey or some shit to prove that sir ?	have you done any type of survey to prove that sir?	Have you done somw type of survey or something to prove that sir?

Challenges Encountered

During the course of fine-tuning the Phi-2 model, several challenges were encountered:

- Limited Computational Resources: Training the LLM with LoRA and PEFT on my local machine posed computational challenges. To address this, I reduced the batch size and used gradient accumulation to ensure the model could still be effectively fine-tuned.
- Model Convergence: Achieving stable convergence during training was challenging, particularly with PEFT. I experimented with different learning rates and employed techniques like early stopping to improve stability.

Next Steps

- Hyperparameter Optimization: Further tuning of the Phi-2 model's hyperparameters will be conducted to improve its performance. This includes experimenting with different ranks for LoRA, adjusting learning rates, and optimizing target modules.

Work Distribution

As the sole member of the team, I have been responsible for:

- Data collection, preprocessing, and preparing the dataset for fine-tuning.
- Implementing the LoRA and PEFT frameworks for efficient fine-tuning of the Phi-2 model.
- Conducting training, evaluations, and generating visualizations to assess model performance.
- Preparing the current report and planning the next steps to further improve the project.

GitHub Repository

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