LoRA & PEFT Fine-Tuning LLMs for Text Detoxification

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

September 22, 2024

Project Overview

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

The goal of my project is to fine-tune large language models using Low-Rank Adaptation (LoRA) and Parameter-Efficient Fine-Tuning (PEFT) methods to detoxify text efficiently. The detoxification system is designed to ensure that any toxic content generated by AI agents is detected and neutralized, improving safety and compliance for enterprise environments.

Problem Overview

AI-based large language models (LLMs) have demonstrated immense power in generating human-like text, but they can unintentionally produce harmful or toxic language. **Text detoxification** ensures that any toxic, offensive, or inappropriate content generated by the AI is intercepted and transformed into neutral, non-toxic text.

For this project, the key focus is to integrate **LoRA (Low-Rank Adaptation)** and **PEFT (Parameter-Efficient Fine-Tuning)** techniques into the detoxification process. These methods provide an efficient way to fine-tune LLMs for the task of transforming toxic content without requiring full model retraining, thus reducing computational costs while maintaining performance.

Who Are the Users of the Product?

The primary users of this detoxification system include:

- **Enterprises and Corporations**: Businesses using AI-based tools for customer communication or internal processes need to ensure their AI agents generate non-offensive, safe language to maintain their reputation and comply with ethical standards.
- **Compliance Officers**: These professionals ensure that AI outputs meet corporate and legal guidelines. They need this detoxification pipeline as a guardrail to prevent any non-compliant AI-generated content from being shared.

- **AI Developers and System Engineers**: Developers of AI-driven applications such as chatbots or customer support systems require reliable tools to filter toxic outputs and ensure safe, reliable user interaction.
- **Content Moderators**: In social media or online forums, automated systems need to remove toxic content before it reaches users, reducing the burden on human moderators.

Why Is This Solution Important?

This solution addresses the growing concern around ethical AI and safe communication, particularly in environments where AI-generated content interacts directly with users. Here's why it's critical:

- **Trust and Safety**: Detoxifying AI-generated text ensures that users feel safe and respected, leading to a more trustworthy interaction between users and AI systems.
- **Enterprise Compliance**: Companies are under increasing pressure to ensure their AI systems comply with legal and ethical standards. A detoxification pipeline ensures that AI-generated text aligns with those standards.
- **Reducing Liability**: Toxic or offensive content can lead to lawsuits or damage to brand reputation. Preventing such content from being distributed reduces these risks.
- **Efficiency**: Using LoRA and PEFT to fine-tune models for detoxification offers a cost-effective and computationally efficient way to build safer AI systems, making it feasible for businesses to deploy at scale.

The Big Picture

The detoxification pipeline is just one part of the overall strategy to make LLM-based AI agents more reliable, robust, and ethical. By introducing this kind of guardrail, we ensure that AI systems do not generate harmful content, making them safer for widespread use in enterprises and other sensitive environments.

Looking ahead, this project serves as a **foundation for building additional guardrails**** for AI, such as bias detection, factuality verification, and sentiment correction. As these systems evolve, the inclusion of multiple safety mechanisms will contribute to the development of **robust, responsible, and reliable AI agents** that can operate autonomously while adhering to ethical standards and ensuring user safety.

GitHub Repository

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

Progress So Far

As of now, the following tasks have been completed:

- **Problem Framing**: The problem of detoxifying AI-generated text using LoRA and PEFT has been defined, and the detoxification process is being integrated as a guardrail for AI agents.
- **Dataset Found**: I have identified and started using the **ParaDetox dataset**, available at Hugging Face's ParaDetox dataset. This dataset contains text that can be used for training the detoxification model.
- **Baseline Model Fine-Tuning**: I have fine-tuned a **T5 model** for text detoxification as a baseline. The training phase for the T5 model is complete, and I will now move on to evaluating its performance.

Data and Scripts

- **Data**: The primary dataset being used is the ParaDetox dataset, sourced from Hugging Face's dataset repository. This dataset provides toxic and non-toxic text, which is used for fine-tuning models to perform detoxification.
- **Scripts**:
  - Data loading and preprocessing scripts to clean and prepare the dataset for fine-tuning.
  - Scripts for fine-tuning the **T5 model** for detoxification.
  - Initial scripts for model evaluation, which will be used in the next phase of the project.

Intermediate Results

So far, the following progress has been made:

- The T5 model has been fine-tuned using the ParaDetox dataset. The training has successfully completed, and the model is ready for evaluation.
- Initial observations suggest that the model has learned to identify and transform toxic content, though further evaluation is necessary to confirm its effectiveness.

Work Distribution

Since I am the only member of the team, I am wearing multiple hats and handling all aspects of the project, including:

- Data collection and preprocessing.
- Fine-tuning the baseline T5 model.
- Scripting and setting up experiments for the next stages of the project.
- Next, I will perform model evaluation and explore LoRA and PEFT for fine-tuning additional models.

Next Steps

- Perform comprehensive evaluation of the fine-tuned T5 model, including accuracy, fluency, and detoxification effectiveness.
- Apply the same detoxification task to different models using LoRA and PEFT for more efficient fine-tuning.
- Compare the performance of LoRA/PEFT fine-tuned models with the fully fine-tuned T5 baseline model.