IntelliHack 5.0

Machine Learning hackathon

Q3 - Technical Report

The Code Rushers

1. Introduction

Fine-tuning large language models (LLMs) is essential for optimizing their performance on domain-specific tasks. This report details the fine-tuning process of Qwen2.5-3B-Instruct, a state-of-the-art instruction-tuned model, using LoRA (LowRank Adaptation) to enhance its efficiency. The goal is to improve the model's ability to generate accurate and contextually relevant responses based on a given dataset.

This document covers the training strategy, chat history maintenance, cost-effectiveness, additional features implemented, and clear instructions for running the program.

2. Approach Taken

The fine-tuning process follows a structured workflow:

1. **Dataset Preparation** – Preprocessing textual data from various sources (PDFs, DOCX, and CSV files).
2. **Model Selection** – Choosing the Qwen2.5-3B-Instruct model due to its robust capabilities in natural language understanding and generation.
3. **LoRA Fine-Tuning** – Applying LoRA for parameter-efficient training, reducing computational costs.
4. **Training & Optimization** – Executing training using Hugging Face's Trainer API with appropriate hyperparameters.
5. **Evaluation & Testing** – Assessing model performance using benchmark metrics.

3. Training Strategy

3.1 Data Preparation

- Extracted textual data using PyMuPDF (fitz) for PDFs and python-docx for Word files.
- Preprocessed data using tokenization techniques with Hugging Face Tokenizer.
- Created structured datasets using the `datasets` library.

3.2 LoRA-Based Fine-Tuning

- Used LoRA for fine-tuning, which enables training a large model with fewer trainable parameters.
- Configured LoRA settings as follows:
  - `r = 8` (Rank of adaptation layers)
  - `alpha = 32` (Scaling factor)
  - `dropout = 0.05`
  - `Task Type: Causal Language Modeling (CLM)`

3.3 Model Training Configuration

- Used Hugging Face's Trainer API for training.
- Key hyperparameters:
  - `Batch Size: 8`
  - `Learning Rate: 5e-5`
  - `Optimizer: AdamW`
  - `Gradient Accumulation Steps: 4`
  - `Number of Epochs: 3`
  - `Mixed Precision: FP16`
- Deployed training on GPU (CUDA-enabled devices).

3.4 Cost Optimization

- Utilized `bitsandbytes` for memory-efficient quantization.
- Leveraged Hugging Face Accelerate to distribute training across available GPUs.
- Used LoRA to reduce the number of trainable parameters, lowering computational costs.

4. Chat History Maintenance

- Implemented context windowing to maintain recent chat history within the model's memory.
- Used sliding window attention to manage token limits efficiently.
- Optimized response generation by summarizing previous interactions when necessary.

5. Additional Features

- **Multi-Modal Data Handling** – Supports PDF, DOCX, and CSV file processing for text extraction.
- **Efficient Fine-Tuning** – Utilized LoRA + PEFT (Parameter Efficient Fine-Tuning) for cost-effective model training.
- **Customizable Token Limits** – Adjustable token constraints to balance latency and response quality.
- **Scalability** – Can be deployed on cloud platforms like Google Colab, AWS, and Hugging Face Spaces.

6. Conclusion

The fine-tuning of Qwen2.5-3B-Instruct was successfully implemented using LoRA, making the process cost-effective and memory-efficient. By leveraging adaptive tokenization, efficient chat history management, and multi-modal data handling, this approach ensures high-quality model responses with reduced computational overhead. Future work may focus on integrating real-time inference APIs, improving dataset quality, and optimizing inference latency for deployment at scale.