StarSpark AI: ML Intern Assignment

Image-to-Asymptote Model Fine-Tuning

Overview

Develop a fine-tuned version of an open-source language model (e.g., Llama, Gemma) that generates compilable Asymptote code from technical or mathematical images. This project simulates a real-world application for AI-assisted mathematical content creation in K-12 education.

Assignment Components

1. Model Selection

- **Task:** Choose an appropriate open-source language model as your base.
- **Considerations:** Assess model performance, community support, and compatibility with your intended fine-tuning objectives.

2. Dataset Creation

- **Data Curation:** Curate a dataset consisting of image-to-Asymptote code pairs.
- **Methodology Documentation:** Clearly document your data collection process including:
    - Data sources
    - Tools and techniques used
    - Any ethical or copyright considerations
- **Dataset Splitting:** Create and document separate training, validation, and test sets to ensure robust model evaluation.

3. Fine-Tuning Pipeline

- **Approach:** Design a fine-tuning strategy tailored to the selected model.
- **Implementation:**
    - Develop the training pipeline
    - Apply hyperparameter tuning
    - Monitor and log model performance during training

- **Documentation:** Detail your training decisions, challenges faced, and rationale for your choices.

4. Evaluation

- **Metrics Definition:** Implement evaluation metrics to assess:
    - **Syntactic Correctness:** The generated Asymptote code must compile without errors.
    - **Visual Fidelity:** The output diagrams should accurately represent the input images.
- **Performance Analysis:** Analyze model outputs to determine where the model excels and where improvements are needed.

Deliverables

- **Public GitHub Repository:** Include reproducible code and clear instructions.
- **Model Access:** Provide trained model weights or an access method.
- **Documentation:** Comprehensive documentation covering:
    - Data collection methodology
    - Fine-tuning pipeline details and hyperparameter tuning decisions
    - Evaluation metrics and analysis of model performance
- **Final Report:** A detailed report summarizing:
    - The approach and methodology
    - Experimental results and performance metrics
    - Insights, challenges, and potential future improvements

Timeline

- **Duration:** Complete the assignment within two weeks from the assignment date.
- **Milestones:** Define intermediate checkpoints to track progress and ensure timely completion.

Final Evaluation

Your model will be assessed on a private test set curated by StarSpark.AI, featuring diverse mathematical diagrams relevant to K-12 education.

Evaluation Criteria

- **Technical Implementation:** Robustness and scalability of your solution.
- **Data Engineering:** Innovation and thoroughness in the dataset creation process.
- **Problem Solving:** Creative and effective solutions to encountered challenges.

- **Code Quality & Documentation:** Readability, maintainability, and clarity of code and supporting documentation.
- **Model Performance:** Effectiveness based on both syntactic correctness and visual fidelity.
- **Analytical Insights:** Depth of analysis in your final report regarding model successes and areas for improvement.