Assignment-2: Fine-Tuning a Large Language Model for Medical Question-Answering

Introduction

This document outlines the process of fine-tuning a pre-trained large language model (LLM) for a domain-specific application, as per the requirements of Assignment-2. The goal is to develop a model that effectively understands and generates text pertinent to the chosen domain, enhancing performance on specialized tasks. The provided Colab notebook(model notebook) serves as the basis for this project, fine-tuning a GPT-2 model on a medical question-answering dataset. This document details the domain selection, dataset preparation, and model selection processes.

1. Domain Selection

Chosen Domain: Medical

The medical domain was selected for this fine-tuning project due to its significant potential to improve healthcare applications through enhanced natural language processing capabilities.

Justification

- **Data Availability**: The medical domain benefits from a wealth of publicly available datasets on platforms like Hugging Face. For instance, the openlifescienceai/medmcqa dataset, used in the provided notebook, contains over 194,000 medical multiple-choice questions (MCQs) derived from sources like AIIMS and NEET-PG exams. This extensive dataset provides a robust foundation for training a domain-specific LLM.
- **Relevance**: Medical question-answering is a critical task in healthcare, where accurate and rapid responses to clinical queries can support physicians, students, and patients. Fine-tuning an LLM for this purpose addresses real-world needs, such as assisting in medical education or clinical decision-making.
- **Potential Impact**: A well-tuned medical LLM can improve diagnostic accuracy, enhance medical education, and streamline patient care by providing precise

answers to complex medical queries. The potential to reduce errors and improve outcomes in healthcare justifies the focus on this domain.

2. Dataset Preparation

Dataset Selection

The dataset chosen for this project is the openlifescienceai/medmcqa dataset, available on Hugging Face (MedMCQA Dataset). It comprises:

- **Size**: Over 194,000 medical MCQs.
- **Content**: Questions, four answer options (A-D), and the correct option index (cop), covering topics from medical exams like AIIMS and NEET-PG.
- **Splits**: Train (182,822 examples), validation (4,183 examples), and test (6,150 examples).

This dataset is substantial and relevant to the medical domain, fulfilling the assignment's requirement to gather a domain-specific textual resource.

Preprocessing Steps

The dataset is preprocessed to format it for fine-tuning the LLM. The provided notebook implements the following steps:

1. **Loading the Dataset**: The dataset is loaded using the datasets library from Hugging Face.
2. **Tokenization**: The GPT-2 tokenizer is initialized, with the padding token set to the end-of-sequence (EOS) token to handle variable-length inputs.
3. **Formatting Inputs**: Each example is formatted into a single string containing the question, options, and correct answer. This prepares the model for causal language modeling.
4. **Splitting**: The dataset is split into training, validation, and test sets, as provided by the dataset structure.

3. Model Selection

Chosen Model: GPT-2

The base model selected for fine-tuning is GPT-2, a pre-trained LLM developed by OpenAI, as implemented in the provided notebook.

Justification

- **Suitability**: GPT-2 is a causal language model designed for text generation, making it appropriate for generating answers to medical questions based on provided prompts. Its architecture (12-layer transformer with 117M parameters in the base version) balances performance and computational feasibility.
- **Accessibility**: GPT-2 is open-source and available via Hugging Face's Transformers library (GPT-2 on Hugging Face), allowing for easy integration and fine-tuning without proprietary restrictions, unlike GPT-3.
- **Pre-training**: GPT-2 has been pre-trained on a diverse corpus (WebText), providing a strong general language understanding that can be adapted to the medical domain through fine-tuning.
- **Practicality**: The notebook demonstrates successful fine-tuning of GPT-2 on the MedMCQA dataset, indicating its capability for this task. Its smaller size compared to models like LLaMA makes it manageable on limited hardware, such as Google Colab's free tier.

Fine-Tuning Process

The fine-tuning process leverages the Trainer API from Hugging Face, training GPT-2 on the preprocessed MedMCQA dataset for 3 epochs with a low learning rate (2e-5) to adapt it to medical QA without overfitting. The use of mixed precision training (fp16=True) optimizes memory usage, making it suitable for resource-constrained environments.

Evaluation

- **QA Accuracy**: Assesses the model's ability to generate correct answers for 100 sampled test examples.

Deploying Model to HuggingFace

After fine-tuning and evaluating the model,it got deployed in HuggingFace.

Model link - medical-gpt2-qa