PS-04: - Introduction to GenAI and Simple LLM Inference on CPU and finetuning of LLM Model to create a Custom Chatbot.

A project report Submitted as part of the
Intel Unnati Industrial Training Program 2024

By

SAARANSH GUPTA
22BAC10027

School of Electrical and Electronics Engineering VIT Bhopal University

July, 2024

Contents

1. Introduction

In the realm of artificial intelligence, particularly in Natural Language Processing (NLP), advancements in language models have revolutionized how we interact with and process textual data. This project focuses on leveraging GenAI technologies, specifically the T5 (Text-To-Text Transfer Transformer) model, for the task of text summarization. Summarization plays a crucial role in condensing large volumes of text into concise, coherent summaries, making information more accessible and digestible. The potential of fine-tuning the T5 model to enhance its capabilities in summarizing diverse types of content. By optimizing the model's parameters and training it on specific datasets, we aimed to create a robust and efficient custom chatbot capable of generating high-quality summaries tailored to user queries.

2. Technical Approach

1. Understanding T5 Model and Text Summarization

- **T5 Model Overview:** Begin with an overview of the T5 model architecture, emphasizing its ability to perform various NLP tasks via text-to-text transformation.
- **Text Summarization:** Explain the task of text summarization and its importance in extracting essential information from lengthy documents.

2. Dataset Preparation

- **Data Collection:** Describe the process of collecting relevant datasets suitable for training the T5 model for text summarization.
- **Preprocessing:** Detail the steps involved in preprocessing the raw text data to ensure compatibility with the T5 model input format.

3. Model Fine-Tuning

- **Fine-Tuning Strategy:** Outline the methodology used to fine-tune the T5 model for the text summarization task. This includes parameter tuning, learning rate scheduling, and batch size optimization.
- **Training Process:** Discuss the training process, highlighting key metrics monitored during training (e.g., loss function, convergence criteria).

4. Evaluation Metrics

- **Evaluation Criteria:** Define the metrics used to evaluate the performance of the fine-tuned model, such as ROUGE scores (Recall-Oriented Understudy for Gisting Evaluation) for summarization quality assessment.

2.1 Code Link

The full implementation of this project, including all code and documentation, is available on GitHub at :- https://github.com/saaranshg/PS4_LLM-MODEL/tree/main

3. Issues Faced

1. Computational Resource Management

One of the initial challenges encountered was managing computational resources effectively. The fine-tuning process of LLMs like T5 demands substantial computational power, particularly when training on large datasets or with complex model architectures.

- **Solution Approach:** Initially, the project started with GPU-based training due to its computational efficiency over CPU. However, ensuring GPU availability and optimizing GPU memory usage became critical, especially when scaling up experiments or training larger models.

2. Data Preparation and Dataset Challenges

Another significant hurdle was related to data preparation and the challenges associated with using appropriate datasets for training and evaluation.

- **Data Complexity:** The CNN/DailyMail dataset, chosen for its relevance in text summarization tasks, presented challenges in terms of data quality and preprocessing requirements. Ensuring the dataset was correctly formatted and suitable for training the T5 model was crucial for achieving meaningful results.

- **Preprocessing Issues:** Aligning input articles and target summaries (highlights) posed specific challenges in terms of tokenization, padding, and ensuring the input-output alignment for training sequences.

3. Hyperparameter Tuning and Training Optimization

Fine-tuning LLMs involves extensive hyperparameter tuning and optimization to achieve optimal performance in specific NLP tasks like text summarization.

- **Learning Rate and Batch Size:** Determining the appropriate learning rate and batch size settings for T5 fine-tuning required iterative experimentation. Finding the balance between convergence speed and model stability was crucial to achieving satisfactory results.
- **Training Duration:** Given the computational demands, managing training duration and optimizing epochs to prevent overfitting while maximizing summarization quality was a delicate balance.

4. Evaluation and Metrics

Evaluating the performance of the fine-tuned T5 model for text summarization presented challenges in selecting and interpreting appropriate evaluation metrics.

- **ROUGE Scores:** Using metrics like ROUGE (Recall-Oriented Understudy for Gusting Evaluation) for evaluating summary quality required careful consideration of nuances in summarization effectiveness, such as content coverage and linguistic fluency.
- **Subjectivity in Evaluation:** Addressing the subjective nature of text summarization evaluation, where human judgment and automated metrics sometimes diverge, posed challenges in objectively assessing model performance.

Highlighted critical issues in computational resource management, data preparation, training optimization, and evaluation metrics selection. Through iterative experimentation and adaptation, the project advanced our understanding of leveraging LLMs like T5 for creating custom chatbot solutions, paving the way for future advancements in AI-driven natural language understanding and generation.

4. Expected Results and Metrics

The solution has complete pipeline for fine-tuning the T5 model on the CNN/DailyMail dataset for text summarization. It encompasses dataset loading, preprocessing, model initialization, training configuration, and saving of the trained artifacts. By leveraging the power of transformers and datasets libraries from Hugging Face, this solution enables efficient experimentation and development of state-of-the-art text summarization models.

1. **Training Loss and Convergence**
   - During training, the loss function decreases as the model learns to generate summaries that align with the target highlights. The loss curve should ideally show a downward trend, indicating that the model is improving over epochs.
2. **Evaluation Metrics: ROUGE Scores**
   - **ROUGE-1, ROUGE-2, ROUGE-L:** These are common metrics used to evaluate the quality of generated summaries compared to reference summaries (highlights in this case).
   - **ROUGE-1:** Measures overlap of unigram (single word) sequences between the generated summary and reference summary.
   - **ROUGE-2:** Measures overlap of bigram (two consecutive words) sequences.
   - **ROUGE-L:** Measures the longest common subsequence (LCS) between the generated and reference summaries, considering recall.
3. **Sample Summaries**
   - Evaluating a few sample summaries manually or using qualitative assessment methods to gauge the fluency, coherence, and relevance of the generated summaries.

Interpretation of Results

- **Decreasing Loss:** A decreasing training loss indicates that the model is learning to generate summaries more accurately over time.
- **ROUGE Scores:** Higher ROUGE scores indicate better performance in terms of content overlap and quality compared to the reference summaries.
- **Sample Summaries:** Reviewing sample summaries helps understand the practical effectiveness of the model in condensing information while retaining key points.

Example Scenario Assume after fine-tuning for several epochs and evaluating on a validation set:

- **Training Loss:** Decreased from initial values, indicating the model has learned to minimize the difference between generated and target summaries.

- **ROUGE Scores:** Achieved competitive scores (e.g., ROUGE-1: 0.35, ROUGE-2: 0.15, ROUGE-L: 0.30), suggesting the model effectively captures important content from articles into summaries.
- **Sample Summaries:** Reviewing a few generated summaries shows coherent and relevant content compression, reflecting the model's ability to summarize effectively.