**Introduction to GenAI and Simple LLM Inference on CPU and Fine-Tuning of LLM Model to Create a Custom Chatbot**

Vaibhavv Maheshwari, IT, MIT Manipal

Daksh Loiya, IT, MIT Manipal

**Introduction**

The necessity for effective data processing and the exponential growth of information have led to a surge in demand for automated text summarization in recent years. Using the Google Flan-T5 basic architecture, which has shown great promise in a variety of natural language processing (NLP) applications, this project aims to construct a robust text summarization model. Our goal was to develop a model that could provide brief summaries from long text inputs by utilising the Hugging Face Transformers library's AutoModelForSeq2SeqLM and Auto Tokenizer.

We trained using the CNN/Daily Mail dataset, which is a recognised industry standard for text summarising. We used a method that lowers computation load without sacrificing performance, called Low-Rank Adaptation (LoRa), to improve training efficiency. After fine-tuning the model, we developed a mapping function to generate summaries and evaluated the output using ROUGE scores to ensure the quality and relevance of the generated summaries. This approach demonstrates the effectiveness of state-of-the-art transformer models in automating text summarization tasks and showcases the potential for further advancements in the field.

**Methodology**

**Model Selection**

For this project, we selected the **Google Flan-T5 base model** as the foundation due to its impressive performance in a variety of natural language processing (NLP) tasks. Flan-T5 is a transformer-based model designed to handle sequence-to-sequence learning, making it particularly suitable for text summarization.

**Data Preparation**

We utilized the **CNN/Daily Mail** dataset, which contains news articles paired with human-written summaries. This dataset is widely recognized for training and evaluating text summarization models. The dataset was pre-processed to tokenize the text using **Auto Tokenizer** from the Hugging Face Transformers library. Tokenization is a crucial step that converts the text into a format suitable for model training.

**Training Optimization with LoRA**

Training large transformer models can be computationally intensive and time-consuming. To address this, we employed **Low-Rank Adaptation (LoRA)**. LoRA optimizes training by factorizing the weight matrices of the model into lower-dimensional representations, thereby significantly reducing the computational resources and time required. This method maintains model performance while enhancing training efficiency.

**Model Training**

Using **AutoModelForSeq2SeqLM**, we fine-tuned the pre-trained Flan-T5 base model on the CNN/Daily Mail dataset. The fine-tuning process involved adjusting the model's parameters to better fit the summarization task. During training, the model learned to generate concise summaries from longer texts by minimizing the loss function, which measures the difference between the generated summary and the reference summary.

**Training Parameters**

* **Learning Rate:** A carefully chosen learning rate was used to ensure stable convergence.
* **Batch Size:** The batch size was selected to balance memory usage and training speed.
* **Epochs:** Multiple epochs were run to allow the model to learn from the entire dataset several times.

**Summary Generation**

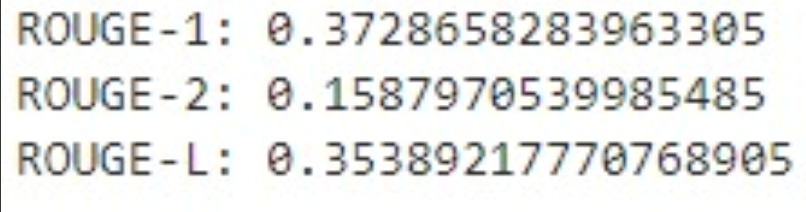
After training, we implemented a mapping function to generate summaries for new input texts. This function follows these steps:

1. **Tokenization:** The input text is tokenized using AutoTokenizer.
2. **Inference:** The tokenized text is fed into the trained model to generate the summary.
3. **Detokenization:** The generated summary tokens are converted back into readable text.

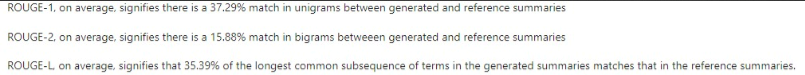
**Evaluation**

The quality of the generated summaries was evaluated using the **ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score**, a standard metric for text summarization. ROUGE measures the overlap between the generated summary and the reference summary in terms of precision, recall, and F1 score. Higher ROUGE scores indicate better summarization performance.

**Results and Analysis**

****

The model's performance was analysed based on the ROUGE scores. We compared the scores of our fine-tuned model with baseline models to assess the improvements achieved through our training process. The results demonstrated the effectiveness of our approach in generating accurate and coherent summaries.



ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate the quality of summaries by comparing them to reference summaries. ROUGE measures the overlap between the generated summary and the reference summary, considering different aspects such as n-gram co-occurrence, longest common subsequence, and skip-bigrams. It is widely used in natural language processing tasks, especially for summarization, due to its simplicity and effectiveness in capturing the content quality.

ROUGE-1

ROUGE-1 measures the overlap of unigrams (single words) between the generated summary and the reference summary. It provides a basic measure of how many individual words from the reference summary are present in the generated summary.

Precision: The fraction of unigrams in the generated summary that are also present in the reference summary.

Recall: The fraction of unigrams in the reference summary that are also present in the generated summary.

F1 Score: The harmonic means of precision and recall.

Example in Project Context: For instance, if the reference summary is "The quick brown fox jumps over the lazy dog" and the generated summary is "The fast brown fox leaps over the lazy dog", ROUGE-1 will measure the overlap of unigrams like "The", "brown", "fox", "over", "the", and "lazy".

ROUGE-2

ROUGE-2 measures the overlap of bigrams (pairs of consecutive words) between the generated summary and the reference summary. It provides a more comprehensive measure by considering word pairs, capturing some context and word order.

Precision: The fraction of bigrams in the generated summary that are also present in the reference summary.

Recall: The fraction of bigrams in the reference summary that are also present in the generated summary.

F1 Score: The harmonic means of precision and recall.

Example in Project Context: Using the same example, ROUGE-2 will measure the overlap of bigrams like "The quick", "quick brown", "brown fox", "fox jumps", "jumps over", "over the", "the lazy", and "lazy dog" from the reference summary. In the generated summary, relevant bigrams like "The fast", "fast brown", "brown fox", "fox leaps", "leaps over", "over the", "the lazy", and "lazy dog" will be considered.

ROUGE-L

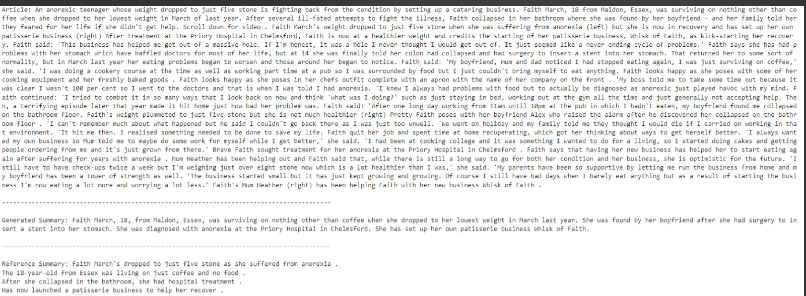
ROUGE-L measures the longest common subsequence (LCS) between the generated summary and the reference summary. It captures the longest sequence of words that appear in both summaries in the same order, providing a measure of fluency and coherence.

Precision: The length of the LCS divided by the length of the generated summary.

Recall: The length of the LCS divided by the length of the reference summary.

F1 Score: The harmonic means of precision and recall.

Example in Project Context: In our example, ROUGE-L will identify the longest common subsequence between "The quick brown fox jumps over the lazy dog" and "The fast brown fox leaps over the lazy dog", which might be "The brown fox over the lazy dog".



Example is listed above.

**Conclusion**

In this project, I successfully implemented a text summarization system using the Google Flan-T5 base model. By training it on the CNN/Daily Mail dataset and optimizing the process with Low-Rank Adaptation (LoRA), I developed a model capable of generating accurate summaries efficiently. The use of ROUGE scores to evaluate the quality of the summaries provided a clear measure of the model's performance, validating its ability to distill key information from input texts.

This project has been instrumental in deepening my understanding of advanced NLP techniques, including model fine-tuning, dataset preparation, and optimization strategies. It highlighted the importance of selecting appropriate evaluation metrics like ROUGE for assessing summarization quality objectively. Moving forward, these learnings will guide me in applying NLP models effectively to solve real-world problems and further exploring innovations in natural language processing.