HACK TO HIRE 2024

INTRODUCTION

The problem statement is to develop a state-of-the-art question-answering model leveraging the Quora Question Answer Dataset. The objective is to create an AI system capable of understanding and generating accurate responses to a variety of user queries, mimicking a human-like interaction.
This project focuses on fine-tuning the BART (Bidirectional and Auto-Regressive Transformers) model for the Quora Question Answer Dataset. BART is a sequence-to-sequence model with a bidirectional encoder and a left-to-right decoder, making it effective for text generation tasks.

LITERATURE SURVEY

1. Quora's Question Answer Dataset [link]

   Studying the dataset and its use cases. The dataset contains questions in multiple languages with a variety of characters ranging from alphabets in more than 15 languages, emojis, numbers and other miscellaneous symbols.

2. FLAN T5 [link]

   Before the implementation of BART, I primarily studied FLAN T5 and fine tuning it for a custom dataset for question answering. The usage of the dataset was a reference that I used to develop my solution.

3. BART [link1] [paper]

   BART, introduced by Lewis et al. in 2019, combines the benefits of BERT's bidirectional context with GPT's autoregressive properties. It is designed to handle text generation tasks, such as summarization and translation, by utilizing a transformer architecture. The model pre-trains by corrupting text with a noising function and then learns to reconstruct the original text, which improves its ability to understand and generate coherent text sequences.

Previous research has demonstrated that fine-tuning pre-trained models like BART on specific datasets can significantly enhance performance on tasks such as question answering, summarization, and text generation. This project leverages these findings by applying fine-tuning techniques to the Quora Question Answer Dataset to improve the model's ability to generate accurate and relevant answers.

METHODOLOGY

1. Data Preparation:
    - The dataset used is the Quora Question Answer Dataset, loaded using the datasets library.
    - Data processing involves concatenating question and answer text and extracting unique characters for analysis.
    - Split the dataset into training(90%) and testing(10%) subsets.

2. Model Setup:
    - BART model and tokenizer (BartTokenizer and BartForConditionalGeneration from the transformers library).
    - DataCollatorForSeq2Seq for model collator, for efficient batch processing

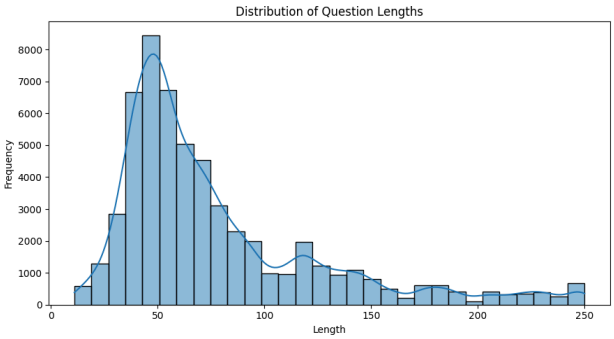
3. Preprocessing:
    - Preprocessing function to tokenize the input questions and answers, adding a prefix to the questions for context.

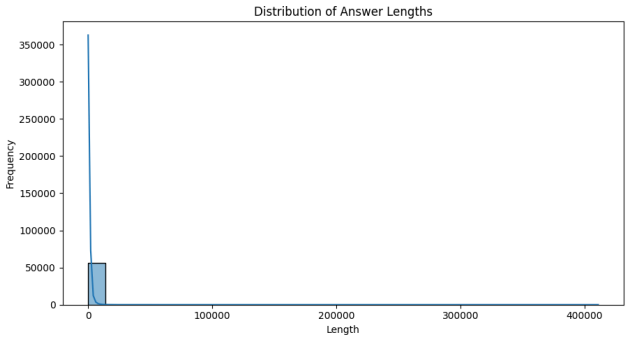
4. Fine-Tuning:
    - Configure training arguments using Seq2SeqTrainingArguments.
    - Utilize the Seq2SeqTrainer to set up and manage the training process.
    - Define hyperparameters and training configurations.

5. Evaluation:
    - Evaluate the model's performance using metrics such as ROUGE scores to assess the quality of the generated text.

RESULT

1. Question and Answer Lengths(no. of chars)



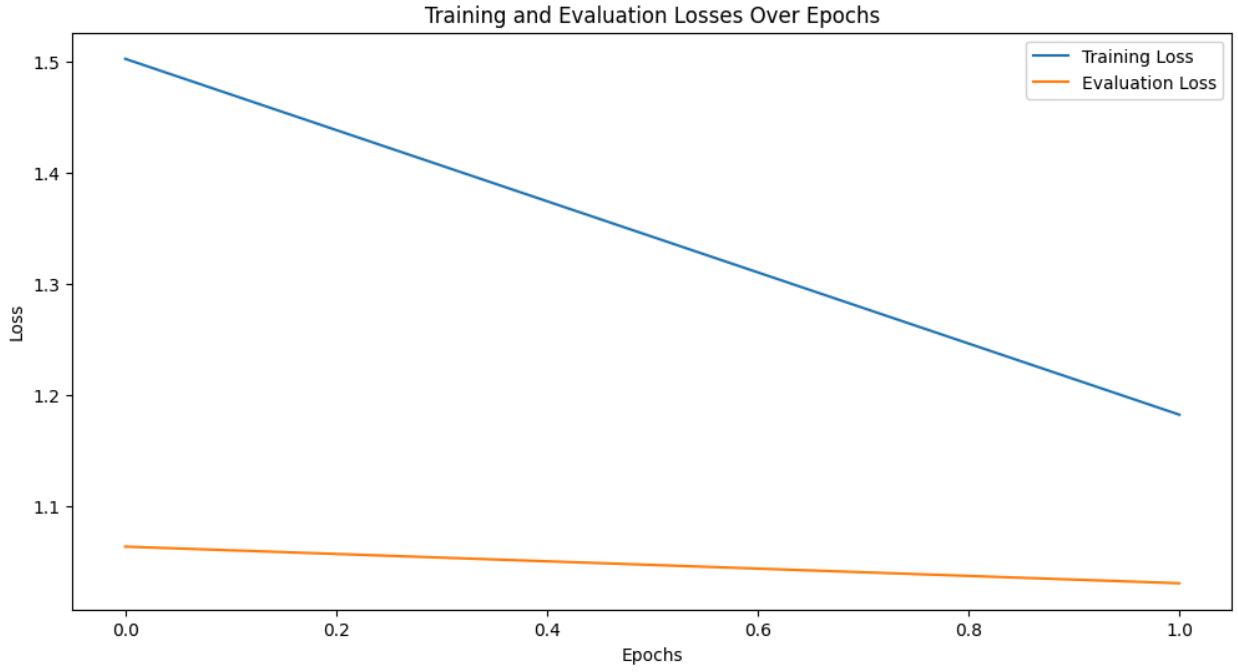


2. Model Performance

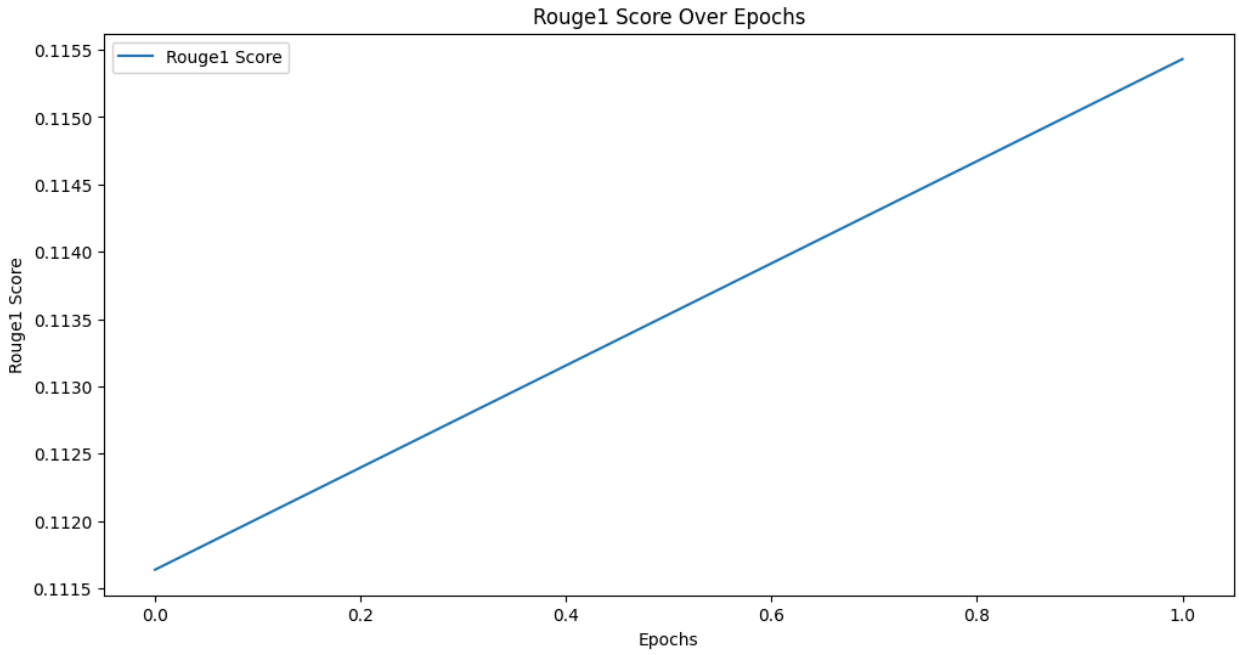
[3172/3172 1:56:36, Epoch 1/2]

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum
0	1.132500	1.063885	0.111639	0.029408	0.091875	0.102407
1	1.036700	1.030965	0.115431	0.032482	0.094971	0.105724

3. Plots



4. ROUGE plot



The fine-tuned BART model demonstrated improved performance in generating relevant and coherent answers to the questions in the Quora dataset.

NOVEL IMPROVEMENTS

1. The model was trained for very less number of epochs(2) due to a limited availability of GPU and computation facility, but improved and enhanced training can result in better performance
2. The dataset consisted of various characters, especially emojis and various other symbols such as mathematical. Special care can be given to these such as for emojis, usage of sentiment and context analysis from emojis can be used for better understanding these characters and their context
3. Similarly, the dataset included questions and answers in various languages and not just english. If we fine tune our model on a broader scale to incorporate all these languages, our response generation will improve
4. Lastly, addition of context to the question can help improve response generation. This has been observed in various cases such as context based question answering using BERT.

CONCLUSION

The project successfully highlights the effectiveness of fine-tuning BART for the Quora Question Answer Dataset. By leveraging BART's advanced sequence-to-sequence capabilities, the model achieved significant improvements in natural language understanding and question answering tasks. Future work could explore further optimization techniques, the use of larger and more diverse datasets, and the application of the model to other NLP tasks.