Project Report: LLM Dialogue Generation

Shubhabrata

August 1, 2024

1 Introduction

This project focuses on fine-tuning a pre-trained Language Learning Model (LLM) to generate dialogues between two speakers. The primary goal is to enhance the LLM's ability to produce coherent and contextually appropriate conversations. The project involves several stages, including Exploratory Data Analysis (EDA), model selection, fine-tuning, and evaluation. This report provides a detailed step-by-step description of the methodology, specific choices, and parameters defined during the project.

2 Step 1: Exploratory Data Analysis (EDA)

Data Collection and Preparation The data for this project was sourced from two CSV files hosted on Google Drive. The files contain dialogues between Lex Fridman and two different speakers, Lee Cronin and Lisa Randall.

- **Downloading Data**: The files were downloaded using the PyDrive library.
- **Reading Data**: The downloaded files were read into Pandas DataFrames for further analysis.

Data Analysis The data was analyzed to understand its structure and content.

- **Missing Values**: Checked for missing values.
- **Text Length Distribution**: Analyzed text lengths to understand variability.
- **Common Words Analysis**: Identified common words using word frequency analysis.
- **Speaker Distribution**: Analyzed the distribution of speakers.

3 Step 2: Model Selection

Several pre-trained models were considered:

- **GPT-2**: Known for general text generation.
- **DialoGPT**: Specifically designed for dialogue generation.
- **BlenderBot**: A conversational model by Facebook.
- **GPT-Neo**: A large-scale model by EleutherAI.

4 Step 3: Fine-Tuning the Pre-Trained LLM

Tokenization Text data was tokenized using respective tokenizers.

Training Parameters Parameters such as learning rate, batch size, and number of epochs were defined.

Training Process Models were trained using supervised and reinforcement learning techniques.

5 Step 4: Fine-Tuning with Additional Data

Continual Learning Elastic Weight Consolidation (EWC) was used to minimize forgetfulness.

Evaluation Metrics Metrics such as perplexity and BLEU score were used to evaluate performance.

6 Step 5: Model Testing and Evaluation

Initial Prompt A predefined prompt was used for dialogue generation.

Response Generation Models generated responses based on the initial prompt and subsequent exchanges.

Evaluation Generated dialogues were evaluated for coherence and relevance.

Findings

- **GPT-2** and **DialoGPT** showed repetitive responses.
- **BlenderBot** generated relevant responses but sometimes introduced unrelated information.
- **GPT-Neo** provided the most coherent dialogues.

Conclusion GPT-Neo is the most suitable model for generating coherent dialogues. Future work may explore additional fine-tuning techniques.

7 Benchmarking Results

7.1 Evaluation Results

The results of the initial test data evaluation are as follows:

Metric	Old Model	New Model
BLEU Score	0.004999871618429517	0.007313616651345576
Perplexity Score	4.237836855932256	71.33681441645774

Table 1: Evaluation metrics for the old and new models on initial test data.

8 Discussion

8.1 BLEU Score Analysis

Both models exhibit low BLEU scores. The new model shows a slightly higher BLEU score, suggesting a marginal improvement.

8.2 Perplexity Analysis

The new model has a higher perplexity score, indicating less coherence or more prediction challenge.

9 Recommendations

- **Manage Input Length**: Implement truncation or splitting strategies.
- **Model Tuning**: Adjust hyperparameters or training data.
- **Error Handling**: Address warnings related to sequence length.

10 Conclusion

This report provides insights into the performance of the fine-tuned language models. Despite warnings and low BLEU scores, the analysis offers a foundation for further improvements.