**Portfolio Project**

Viet Ngo

Colorado State University Global

CSC525: Principles of Machine Learning

Dong Nguyen

08/28/2024

**Portfolio Project Milestone 6**

For this project, training a chatbot model involved a combination of technical challenges and iterative refinement. I utilized a transformer-based model fine-tuned on a variety of datasets, including DailyDialog, Empathetic Dialogues, and SQuAD. The goal was to create a model that can handle conversational interactions, respond empathetically, and even answer questions based on provided contexts. The model I used is a transformer-based architecture, specifically a variant of BERT (Bidirectional Encoder Representations from Transformers). Transformers are powerful because they can handle a wide range of NLP tasks due to their ability to learn contextual relationships in the data. The first round of training yielded promising, yet imperfect, results. The loss gradually decreased, and the gradients normalized, indicating that the model was learning. However, the training process was lengthy, taking several hours due to the size of the datasets and the complexity of the model. One of the most significant challenges was determining the correct structure for the datasets. Each dataset had different formats, DailyDialog provided lists of dialogues, while Empathetic Dialogues offered individual utterances, and SQuAD presented question-answer pairs. Unifying these formats required significant preprocessing. For instance, I had to concatenate context and questions in SQuAD, while in DailyDialog, I merged the dialogues into a single string. During training, I encountered several issues that required debugging. A notable one was a type mismatch error when processing the SQuAD dataset. Initially, I concatenated strings and lists incorrectly, leading to errors. This required careful examination of the data structure and adjusting the preprocessing function accordingly. Moreover, I faced a mismatch in the lengths of input sequences and labels, particularly when padding sequences to a maximum length. This was due to inconsistencies in the preprocessing steps, where I initially didn't account for the varied lengths of inputs across datasets. Correcting this involved adjusting the tokenization process to ensure that the inputs and outputs were consistently padded or truncated to the specified maximum length.

After the model was trained, I evaluated its outputs. One issue that stood out was the chatbot's tendency to produce slightly incoherent responses. It often maxed out the character limit instead of providing concise and relevant replies. This suggests that while the model can generate text, it struggles with maintaining coherence over longer outputs. To improve this, I plan on implementing fine-tuning with additional datasets, adjusting the max length parameter, and applying post-processing techniques. Incorporating more diverse datasets, particularly those that emphasize concise responses, could help the model learn to generate shorter, more coherent replies. For the max length of responses, reducing the maximum sequence length during generation could encourage the model to produce more concise responses. This can be achieved by either setting a hard-limit or penalizing longer answers. Finally, applying techniques like response truncation or keyword filtering could help refine the chatbot's output.

In the final submission, I aim to refine the model to produce more coherent and contextually appropriate responses. This might involve additional retraining, particularly with a focus on conversational coherence. Techniques like beam search or nucleus sampling could also be explored to enhance the quality of the generated responses. Overall, while the current model performs reasonably well, there is room for improvement in making the chatbot more responsive and coherent. Further experimentation with different training strategies and hyperparameter tuning will likely be necessary to achieve the desired level of performance.