Module 5: Portfolio Milestone

Brady Chin

Colorado State University Global

CSC525-1: Principles of Machine Learning

Dr. Dong Nguyen

April 20th, 2025

Portfolio Milestone

At this point in my portfolio project, I have completed the initial stages of training my chatbot. I have selected the pre-trained GPT-2 model along with the Cornell Movie Dialogs Corpus dataset (Danescu-Niculescu-Mizil, C., 2011) that contains conversations from a movie do train my chatbot to creating multi-turn dialog.

Model and Dataset

For this project, I am using the GPT-2 model. The GPT-2 model is a transformer that is capable of generating human like responses by predicting the next word, or token, in a sentence. Once this model was imported, I used the Cornell Movie Dialogs Corpus dataset to fine-tune the model.

The Cornell Movie Dialogs Corpus dataset contains thousands of lines of dialog between different characters. Each line has a unique identifier that is stored in a text file named “movie_lines.txt”. There is another file name “movie_conversations.txt” that holds conversations by storing unique identifiers in an array. From this, I was able to store dialog pairs, or prompt-response pairs, to train the model.

Training

For the training process, I used PyTorch's DataLoader to upload batches of data. Fine-tuning was done by using inputs and outputs allowing the tokenized sequences to also serve as the labels so the model will learn to predict the next token in the tokenized sequence (Hugging Face Community, 2025).

While the training can be successfully completed, the training process is very slow. The next action item for this project is to decrease training time and reduce loss. To solve this, along with further fine-tuning my model, I plan to experiment with a learning rate scheduler to improve convergence. This will also aid with how my model learns. I also may lower my sequence length. I am currently using 512 which can be computationally expensive. This is something I may be able to lower to 256 which will result in faster training and fewer wasted tokens.\

Conclusion

The training process was very helpful in understanding the development behind chatbots and natural language processing. While there are places that can be improved on, the initial stages of the chatbot have been a great learning experience.

References

Danescu-Niculescu-Mizil, C., (2011) *Cornell Movie-Dialogs Corpus.* Cornell University. https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Hugging Face Community, (2025) *Tokenizer.* Hugging Face. https://huggingface.co/docs/transformers/en/main_classes/tokenizer