Module 4: Portfolio Milestone

Brady Chin

Colorado State University Global

CSC525-1: Principles of Machine Learning

Dr. Dong Nguyen

April 13th, 2025

Milestone 4

The dataset that I have chosen for my chatbot is the Cornell Movie Dialogue Corpus. The reason for this is because it provides multi-turn dialogues with clear turn-taking, it has a causal tone, and it is easy to parse.

Training

To preprocess the dataset for training, I will need to parse the dialogues. For example, I will need to specify which lines are said by the user and which lines are said by the chatbot. This will be in the format: user -> chatbot -> user -> chatbot. I will also need to keep track of whose turn it is in the conversation. Since this will be a multi-turn chatbot, meaning that it is able to handle previous messages, I will need to keep a history of the conversation. I will also need the full chat as input so that the model and predict the flow of the conversation.

This cleaned data will then be used to train the model using PyTorch and the Hugging Face Transformers library. During training, the model will try and learn the structure and tone of the conversation to identify patterns and generate responses.

Utilizing the Dataset

Once the model is trained, it will be able to generate responses during a live conversation. The chatbot will have a history of the conversation so that it can refer to past dialogue that it has with the user. Each time the user send a message to the chatbot, it will be fed into the model which will then generate a response that aligns with the tone and flow of the conversation.

Conclusion

By utilizing the Cornell Movie Dialogue Corpus, the chatbot will have a dataset that consists of casual, multi-turn dialogues with a good, friendly flow. It will keep a history of the conversation allowing a natural back-and-forth conversation.