Minor Course Project

Music Generation

Aim:

The goal of the project is to generate more notes given a sequence of notes. In other words, generate music given an initial sequence of notes. A note represents the pitch played by the midi instrument.

Data Preprocessing:

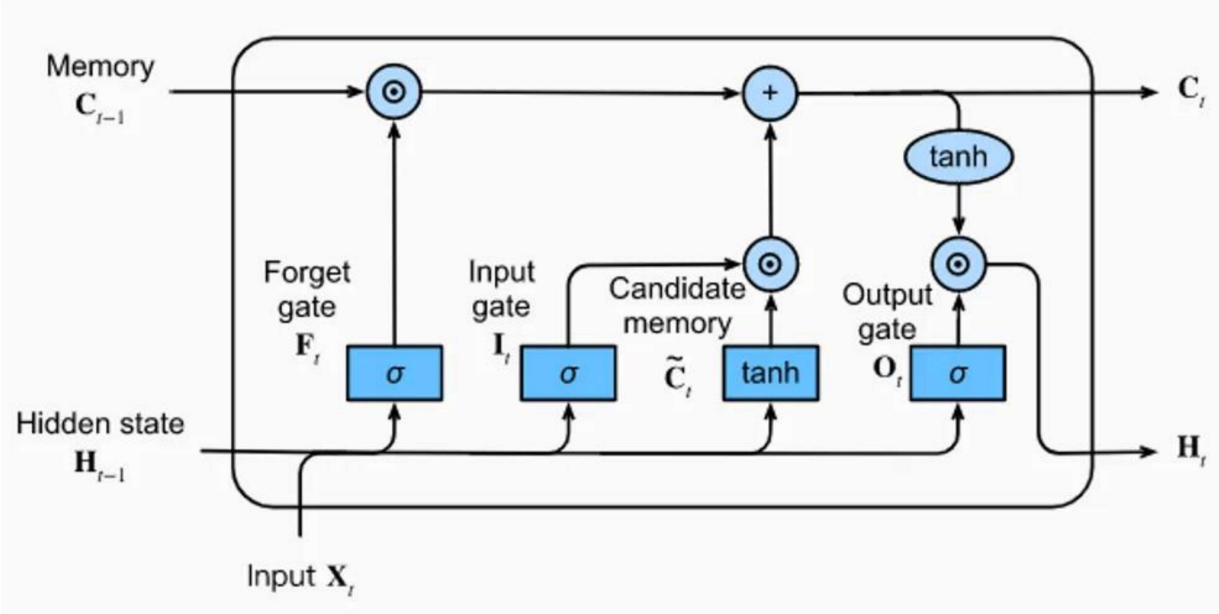
The dataset is MIDI (Musical Instrument Digital Interface) files. We extract pitch, start time and duration for which it is played. Thus, chords here are represented as separate notes with the same start time.

Since the start time for generation could make it harder for the model to train, we take “step” for each note which is the time difference between the start of the previous note and the current note. Note that more than one note can be played at the same time, thus the concept of step. We one-hot encode the pitch, as number of pitches possible in total is 128.

Next, we modify the dataset to suit the model. We make windows of notes of 100 note each followed by 1 target note.

Model Architecture:

The model uses LSTM (Long Short Term Memory) networks which are a type of RNN (Recurrent Neural Network). LSTMs are useful for predicting the output given a long sequence as it can use the information from all the parts of the input sequence.



- Input gate learns whether to write to a cell.
- Forget gate learns whether to erase cell state.
- Output gate learns how much to reveal to the cell state.
- Candidate gate learns how much to write to cell state.

LSTM Equations:

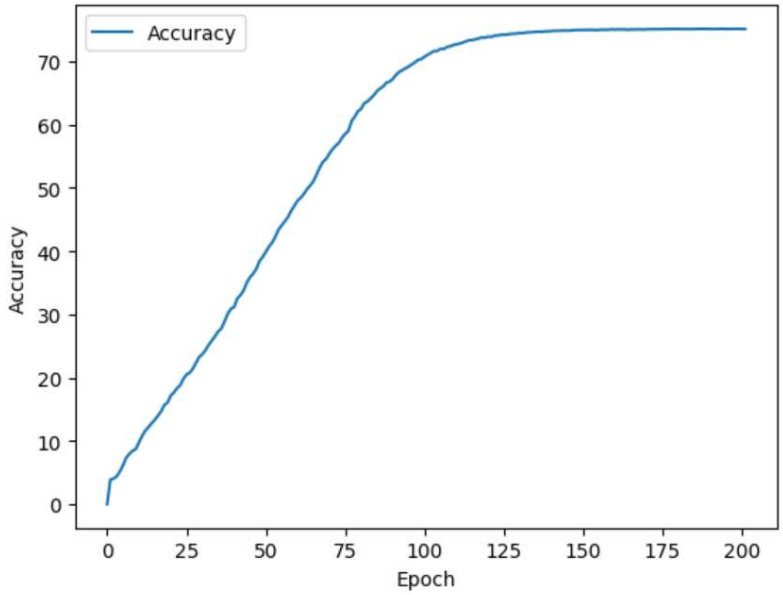
- $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$
- $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$
- $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$
- $g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g)$

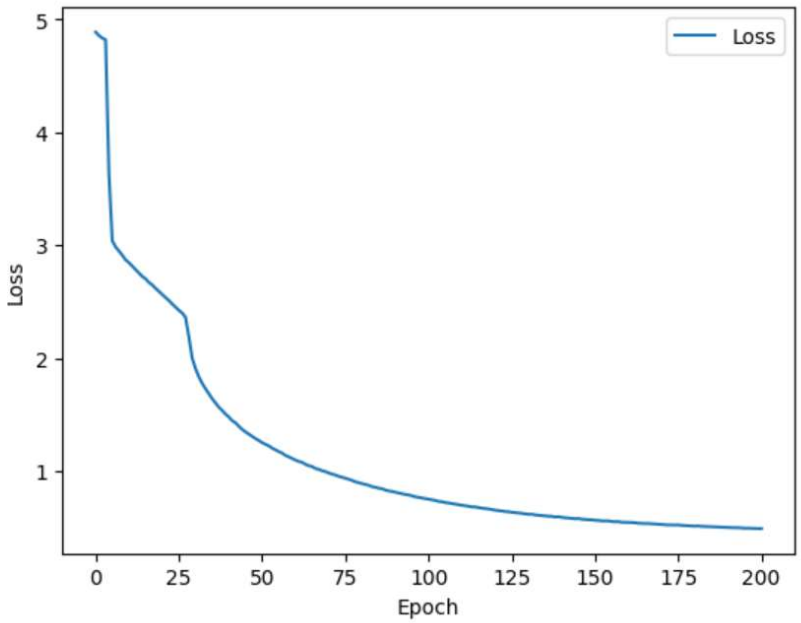
The model uses two LSTM layers stacked on top of each other whose outputs are fed into three separate fully connected layers which are meant for predicting pitch, step and duration respectively. Since pitches are one-hot encoded, we again feed the output of fully connected layer for pitch into softmax layer to produce one-hot encoded pitch results.

Results:

The images shown below represent the accuracy and loss function over the several epochs.

Accuracy here is the percentage of correct pitch predictions.





We have used cross entropy loss function for pitch and l1 loss function for step and duration (to punish outliers more severely) .

Conclusion:

Thus this is a LSTM music generation model which looks at the last 100 notes and predicts the next note, by shifting the window to include the most recent prediction we can generate music.