Abstract

Music Transcription can be viewed as a sequence-to-sequence mapping problem where raw audio is mapped to music notes. While the nature of Music Transcription(MT) is similar to that of Automatic Speech Recognition (ASR), the key differences are -

- ASR usually focuses on a vocabulary from a single speaker but MT requires the ability to transcribe multiple instruments(tracks) that can be combined into a single musical piece
- MT needs to additionally focus on pitch and timing information as musical notes are time-bound
- MT is a low-resource problem, i.e. datasets that consist of musical notes documented in MIDI format are scarce

Hence, the motivation here is to use sequence-to-sequence modelling modified for music transcription to account for additional features described above.

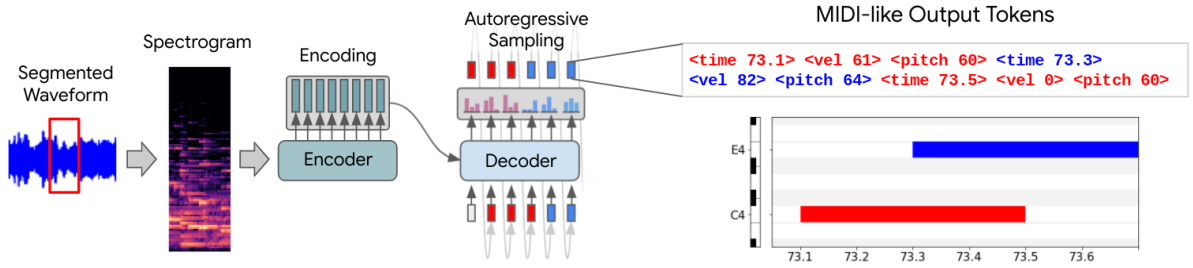
Multi-Task Multi-Track Music Transcription (MT3)

MT3 is an open-source framework by google for music transcription. It uses T5X Transformer architecture for sequence-to-sequence training and inference tasks.

There are currently two models present in the open-source domain -

- Piano Transcription Model presented in ISMIR2021
- Muti instrument transcription model mentioned in ICLR2022.

Working of MT3 based on T5X



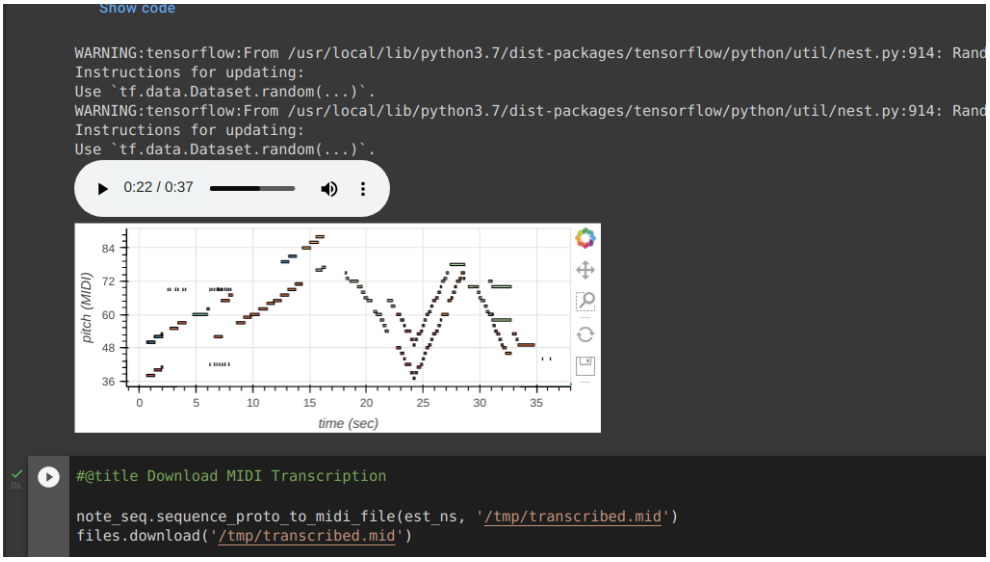
Training process -

1. The dataset consists of raw audio files and their corresponding musical notations.
2. From the musical notes aggregated across the dataset, a vocabulary of musical notes is created
3. Raw audio is converted to log Mel spectrogram
4. Sections of the spectrogram are represented as a vector embedding by the encoder of the transformer
5. The decoder is responsible for decoding the vector embedding into its relevant musical note/s.

6. Using auto-regressive sampling, the model also learns to predict time-duration pitch and velocity.

Sample Output

Input file - *harmonium-ascend-descend.wav*
Result -
The model predicts the pitch and duration of each musical note in the audio file.



MIDI file analysis using Python -

```
(music) ➜  MusicTrans python3 miditest.py transcribed.mid
Format 1 nTracks 10 division 220
        Track 0 of length 0
        Track 1 of length 0
        Track 2 of length 0
        Track 3 of length 0
        Track 4 of length 0
        Track 5 of length 0
        Track 6 of length 0
        Track 7 of length 0
        Track 8 of length 0
        Track 9 of length 0
Track 0:
MIDI@16344 0[4]
MIDI@16348 0[2]
MIDI@16372 0[8]
META@16373 End Of Track ->
Track 1:
MIDI@0 48[2]
MIDI@127 NOTE_OFF[2] D#3 OFF velocity := 50
MIDI@127 0[0]
MIDI@179 112[15]
MIDI@183 32[8]
MIDI@310 0[9]
MIDI@348 0[0]
MIDI@603 48[5]
MIDI@730 0[4]
MIDI@771 112[15]
MIDI@780 32[8]
MIDI@780 0[5]
MIDI@832 0[0]
MIDI@876 32[9]
MIDI@876 0[0]
MIDI@929 0[0]
META@930 End Of Track ->
Track 2:
MIDI@268 48[2]
MIDI@268 0[0]
MIDI@320 112[15]
MIDI@620 48[4]
MIDI@620 PRESSURE[3] F#10 pressure := 79
MIDI@747 NOTE_OFF[2] D2 OFF velocity := 79
MIDI@747 0[0]
MIDI@828 112[15]
MIDI@1123 64[7]
MIDI@1250 0[0]
MIDI@1331 0[0]
MIDI@1555 64[7]
MIDI@1555 0[1]
MIDI@17858 0[0]
Track 3:
MIDI@303 48[7]
MIDI@303 0[0]
MIDI@360 112[15]
MIDI@673 48[9]
MIDI@673 NOTE_OFF[8] E6 OFF velocity := 52
MIDI@800 NOTE_OFF[2] C1 OFF velocity := 65
MIDI@927 32[12]
MIDI@979 0[0]
MIDI@1230 64[3]
MIDI@1357 16[1]
MIDI@1422 0[0]
MIDI@1546 64[3]
MIDI@1546 NOTE_OFF[1] F#2 OFF velocity := 57
```

References

- https://github.com/magenta/mt3
- https://openreview.net/pdf?id=iMSjopcOn0p
- https://github.com/google-research/text-to-text-transfer-transformer
-