Automatic Music Transcription

Bingbin Liu Laëtitia Shao Xiaoyan Wu

Stanford Computer Science

Introduction

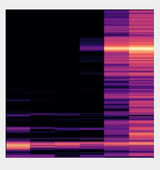
Our project aims at transcribing the melody of a song, where we represent an input audio by spectrograms, from which we use CNN to predict the pitches. CNN is shown to be more effective than traditional models, and our model outperforms the baseline by a large margin.

Dataset

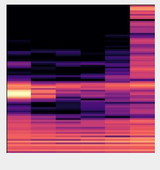
Spectrograms generated from .wav audios in *Medley DB* with frequency annotations sampled every 5.8 ms and 46ms

Mel Spectrogram -- timber
Constant-Q Transform -- pitch

Mel



CQT



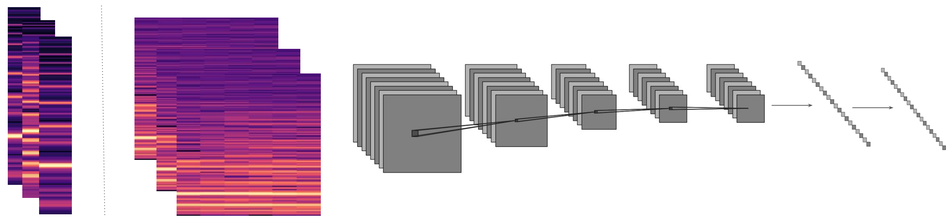
Quantize the frequency space into 108 bins representing musical notes + 1 bin for empty.

$$n = \frac{\log(\frac{f_n}{f_0})}{\log(\sqrt[12]{2})}$$

Neural Network Architectures

CNN-based Models

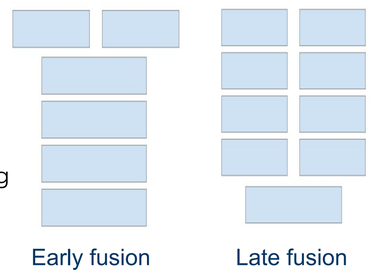
- 5 conv layers + 4 max pool + 2 FC
- Batch norm after each conv layer
- Dropout (0.6) after first FC layer
- Descending LR + momentum (0.9)



Combining inputs

Use both types of spectrograms to enrich the input (i.e. more features).
We explored 4 ways to combine the data:

- **Stacking**: concatenate Mel and CQT spectrograms
- **Early fusion**: separate first layer + share the remaining
- **Late fusion**: separate first four layers + fuse at the last
- **Averaging**: average the outputs from Mel and CQT

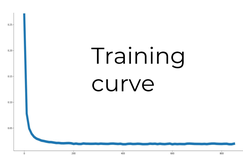


Early fusion

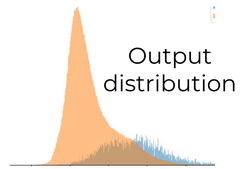
Late fusion

Polyphonic

- Predict multiple notes at a time
- Multi-label soft margin loss



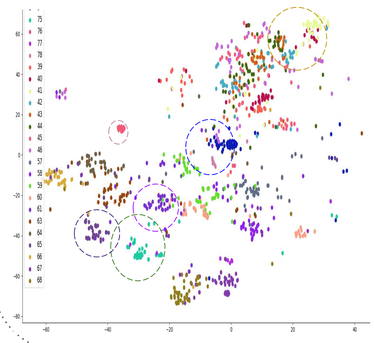
Training curve



Output distribution

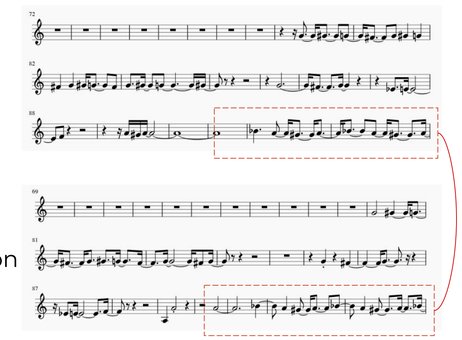
Error Analysis & Output

Features Visualisation



We visualise the features learned by the CNN, by applying PCA and t-SNE.

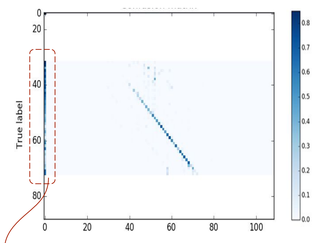
Example output



Ground truth

Prediction

Confusion Matrix



Imbalanced classes

Pitch Tracking

We perform pitch tracking to smooth the results over time. Using a HMM with learned transitions and a LSTM network.

Results

Top 1 accuracy:

- librosa (baseline): 25%
- Mel: 76.5%
- CQT: 75.7%
- Stacking: 75.6%
- Early fusion: 77.1%
- Late fusion: **77.2%**

Top 5: **96.8%**

Future work: handle unbalanced classes and polyphonic outputs, and better analysis (e.g. genre)

References

- *Convolutional neural network for robust pitch determination*, Su et.al, 2016
- *HMM-based multipitch tracking for noisy and reverberant speech*, Jin et. al., 2011
- *Medley DB, a Multitrack Dataset for Annotation-intensive MIR Research*, Bittner et. al., 2014