Music Style Analyzer: CS 6780 Final Project

Brian Rappaport, DB Lee

Introduction

- Goal: identify composer of classical music audio clips
- Existing systems such as Shazam and Alexa put a larger focus on more modern music
- Classical music shows less variance in instrumentation and sound spectrum
- Binary classification of musicians representative of two dramatically distinct styles: Bach and Debussy
- Using neural network architectures, which show state-of-the-art performance in classification
- Ultimate goal: extending the model to multi-class.

Model

We consider three models and compare their performance:

- "Expert" model: human judgments based on prior knowledge
- LSTM model: utilizes the sequential nature of music
- CNN model: treats spectrograms as image representations of audio

Expert Model

- Baseline: predictions from prior musical knowledge
- Baroque and Romantic music differ greatly in uses of musical intervals and voice leading
- Example: whole-tone scale; used widely by Debussy, considered inappropriate in Bach's time
- In the multi-class extension, this becomes increasingly hard for humans

LSTM

- Sequence classification, considering the short-time Fourier transform
- Variation of RNN, with an emphasis on memorization
- NLP applications: word embeddings, part-of-speech tagging

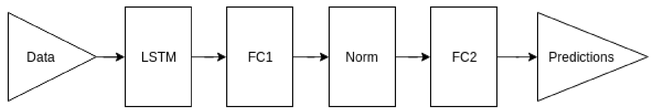

Figure 1: Flowchart of LSTM Architecture. An LSTM input layer is connected to a fully connected dense layer, which is normalized and then connected to an output dense layer.

CNN

- Image classification, considering the spectrogram as an equivalent representation of audio data
- Performs best with image data with sufficient locality
- Unlike speech audio, music consists of long, consistent wave signals and exhibit temporal locality
- CV applications: object detection, facial recognition

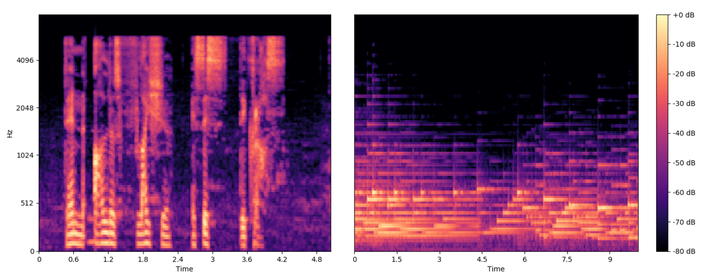

Figure 2: Spectrograms comparing the spectra of speech and music. Speech shows more impulse-like behavior, while music is more fluid and continuous.

Results

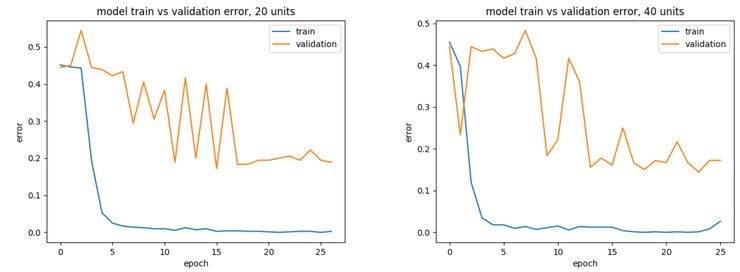

Figure 3: Training and validation error for an LSTM architecture with 20 and 40 units, respectively. The general behavior of both plots suggests the error is converging, although more data would be required to say for sure. The "spiky" behavior of the validation error was unexpected and remains an open topic of investigation.

Conclusions and Future Work

- Expert classification is effective but requires extensive musical training and is not practical for everyday use
- Audio classification can be transformed into other well-explored tasks such as sequence or image classifications
- Once these problems are in a more familiar form, well-understood methods can be used to solve them
- Next step: extend our experiments with CNN and LSTM and extend to multi-class with more than just two composers

References

[1] Mrinmoy Bhattacharjee, S. R. M. Prasanna, and Prithwijit Guha. Time-Frequency Audio Features for Speech-Music Classification. *arXiv e-prints*, page arXiv:1811.01222, Nov 2018.

[2] S. Hershey and S. Chaudhuri and D. P. Ellis and J. F. Gemmeke and A. Jansen and R. C. Moore and M. Plakal and D. Platt and R. A. Saurous and B. Seybold. CNN architectures for large-scale audio classification. *IEEE conference on Computer Vision and Pattern Recognition*, pages 131–135, 2017.