Structure-Aware CNN Onset Detector

Samuel Li

***Abstract*—We are submitting a convolutional neural network specially structured to utilize the frequency-domain symmetries present in piano music. Our neural network is designed to detect only note onsets, and achieves an f-measure of about 80% on the MAPS dataset. The proposed model fundamentally consists of 11 copies of a feedforward neural network trained to detect the onset of 8 specific notes, equally spaced throughout the 88 notes present on a piano keyboard. Onsets for the full range of notes are detected by repeatedly applying the feedforward network to different frequency windows; in this way, we are assuming that the same detector can be used for 11 adjacent notes by simply shifting frequency in half-step intervals. This greatly improves training time and testing performance. We do not train the network on a piano dataset, but rather the procedurally generated infinite dataset described in [3].**

I. Data Representation

We use the CQT spectrogram [2] as our main time-frequency representation. We use CQT bins equally spaced from $\mathrm{D}_1 - 33$ cents to $\mathrm{C}_9 + 33$ cents with 36 bins per octave, giving a total of 286 frequency bins. The spectrogram frame rate is fixed at 20 Hz. We take only the magnitude of the CQT spectrogram as input to the model.

II. Model Description

Our model's input is a 16-frame window of the spectrogram, normalized to have a maximum value of 1. The output of the model is an 88-dimensional vector of values between 0 and 1 predicting the presence or absence of an onset for each note at the center of the reading window. Values closer to 1 are interpreted as a higher probability of note onset.

The architecture of our neural network can be reduced to a feedforward neural network with two hidden layers of 1024 and 256 neurons with the softsign activation [1], and a final sigmoid output layer of 8 neurons predicting the presence or absence of the notes A_0, $\mathrm{G}\#_1$, G_2, $\mathrm{F}\#_3$, F_4, E_5, $\mathrm{D}\#_6$, and D_7. The input to the feedforward network is a 16×256 portion of the 16×286 reading window. The entire feedforward network is convolved along the frequency axis with stride length 3, while its output is shifted by one half-step for each position. In effect, this allows the detectors for these 8 notes to be extended to the entire 88-note keyboard by running 11 copies of the feedforward network on different frequency windows.

III. Model Training

We do not use a labeled piano recording dataset for training. We train the model on a slightly modified version of the procedure described in [3] adapted for a higher number of frequency bins and with a slightly improved generation algorithm. The network was trained for 133 hours on the CPU of a Lenovo Thinkpad T410 using Tensorflow r1.3.

References

[1] James Bergstra et al. *Quadratic polynomials learn better image features*. Tech. rep. Technical Report 1337, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, 2009.

[2] Judith C Brown. "Calculation of a constant Q spectral transform". In: *The Journal of the Acoustical Society of America* 89.1 (1991), pp. 425–434.

[3] S. Li. "Context-Independent Polyphonic Piano Onset Transcription with an Infinite Training Dataset". In: *ArXiv e-prints* (July 2017). arXiv: 1707.08438 [stat.ML].