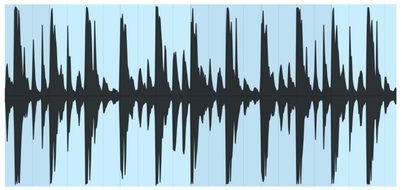
Real Time Beat Tracking: A Mixed Approach

George Woskob *(gwoskob@stanford.edu)* https://www.youtube.com/watch?v=OKoAD-Fg270

CS230 FINAL PROJECT PRESENTATION, STANFORD UNIVERSITY

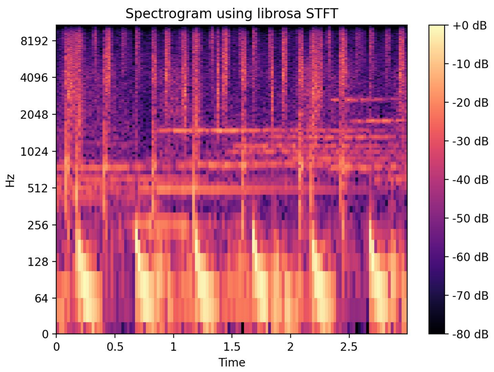
Raw audio...



Audio waveform is streamed as input to the model

STFT

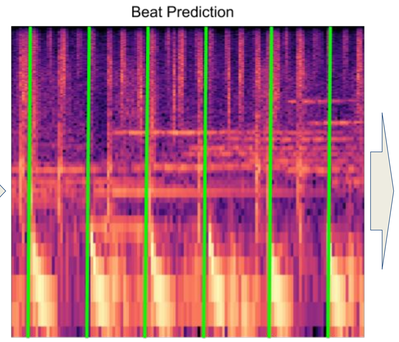
gets turned into an image....



Frequency and amplitude are plotted over the time domain via STFT to produce the image that is fed into the neural network

Neural Net

and the neural network finds the beats...



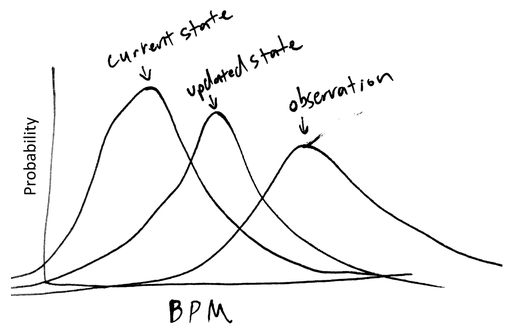
The neural net predicts the locations of the beats in the audio image using ideas borrowed from the YOLO algorithm

- 119.0 BPM
- 0.4 Beats since beginning of last beat

Calculate beats per minute and distance from the last beat from neural net's prediction

Kalman Filter

then a kalman filter taps along!



The Kalman filter updates its beliefs about the state based on the new observations passed to it from the neural network so that it can keep track of tempo and location within the beat as the song progresses. The Kalman filter is able to make predictions about the locations of future beat given a current BPM and location within the beat.

Can you get a computer to tap its foot?

Beats are the fundamental unit of time in music. Finding the beats in music is usually quite easy for many people; many people will actually inadvertently do it as they tap their feet to music. Beat tracking is the term used to describe the task when given to a computer.

The Process

The training set contained nearly 1,000 songs across several genres. Audio gets converted to a a spectrogram which is a 2D image of audio where frequency and amplitude get mapped over the time domain. This was fed into a neural network. The network predicts the locations of the beats within the image. The tempo and location within a beat are calculated from these predictions. This is passed to a Kalman filter which maintains a belief about the tempo and location within the beat in the music at that point in time.

YOLO-like

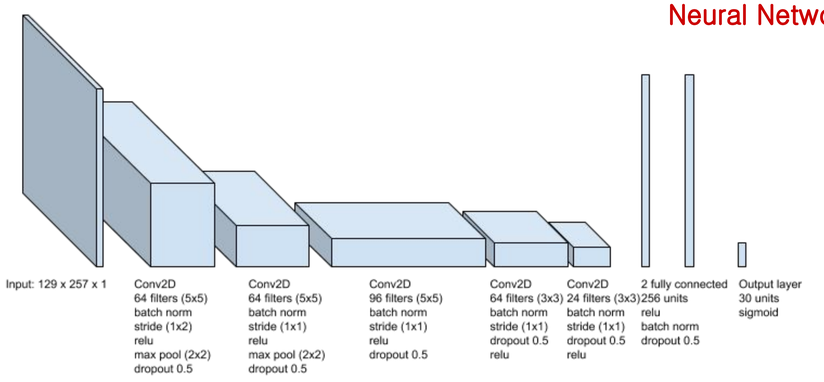
First I tried getting the model to predict the tempo directly. However, it was found that neural networks were not reliable linear predictors of these sorts of values. Then, attention was turned to trying to make the neural net predict the locations of the beats across the x-axis, which represents time in the image. Inspiration was drawn from the YOLO algorithm for the output of the nueral network in how it locates the beats within the image and for the loss function that optimized that neural network.

$$\lambda_{loc} \sum_{i=0}^{B} 1_i^{beat} |x_i - \hat{x}_i|$$

$$+ \lambda_{beat} \sum_{i=0}^{B} 1_i^{beat} \left|C_i - \hat{C}_i\right|$$

$$+ \lambda_{nobeat} \sum_{i=0}^{B} 1_i^{nobeat} \left|C_i - \hat{C}_i\right|$$

Neural Network

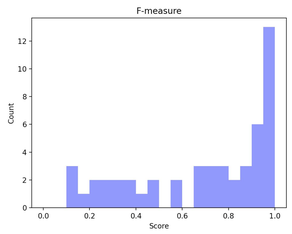


While 1D convolution was initially used for fast feedback on outputs to the model eventually 2D convolution was chosen because it improved more quickly on the loss function.

The neural network of the model has 5 convolutional layers followed by two fully connected layers and a fully connected output layer. The first two convolutional layers have max pooling layers following them.

Results

Generally, the system performs well on audio that contains strong beats such as disco and pop songs as well as some rock songs, with the system perfectly scoring across all metrics on several songs withing the validation set. Other songs where the beats are less pronounced by rhythmic instruments proved more challenging to the system such as songs in the jazz and classical genre. Below is a distribution across the songs in the validation set for one particular metric, F-Measure.



Many metrics exist to evaluate beat tracking systems. A common metric is F-measure which takes into account accuracy, precision, and recall with correct identification meaning that the identified start of the beat occurred within a 0.07 second window (0.07 seconds was the default value used by mir eval library) of the actual start of the beat. Other useful metrics include cemgil score, which is like F-measure but it uses a gaussian distribution over each beat instead of a hard window. Shown is a chart of the average performance on all metrics of the beat tracking system over the entire validation set.

Common Beat Tracking Scores

AMLt	*AMLc*	*CMLt*	*CMLc*	*P Score*	*F Measure*	*Cemgil*
0.611	0.435	0.582	0.543	0.728	0.694	0.528

Scores calculated as an average for each file across the validation set