neural audio: music information retrieval using deep neural networks

Grey Patterson (Linfield, CCT), Andrew Pfalz, and Jesse Allison (LSU School of Music/CCT)

introduction

Scan the QR code for the more verbose digital version of this poster.

Neural networks are a form of machine learning that consist of nodes connected by weighted edges. Input goes in as floating-point numbers, and is propagated through the network from node to node along those edges.

Music is rich in information - from things like what key and time signature are being played in up to the sociocultural context in which the lyrics were written.

The field of music information retrieval exists to provide musicologists with automated tools. Complex tasks, though, remain out of reach. There is no algorithm to identify the genre of a piece, and no machine can identify which instrument is playing as easily as a human can.

neural networks

input Neural networks are algorithms: they have inputs and outputs, and do something to the input to yield the output. For audio processing, this input can be either audio samples or Fast Fourier Transform data. We elected to use raw audio samples throughout.

The inputs are windows of a fixed length: a set number of samples being fed in.

training Neural networks are an alternative to traditional programming. Instead of writing an algorithm yourself, you provide training data - inputs, and the output expected: `([0, 821, 1643,...],"art")`

The neural network itself is a series of layers, each one consisting of one or more neurons. Each neuron can have many inputs, and a single output that can be sent along to many other neurons. As a number moves from one neuron to another, it is altered by the weight of that neuron.

The software takes the inputs, runs them through a network with randomly-generated weights, then changes all the weights a bit and tries again.If it is closer to the right answer, it will keep changing the weights in that direction; if it got worse, it will go in a different direction. This is Stochastic Gradient Descent.

output The output of the neural network can take a few different forms, depending on what is desired.

There is the categorization, which will taken an input and return the number of the category. Alternately, a softmax layer will return the neural network's probabilities for *all* of the categories.

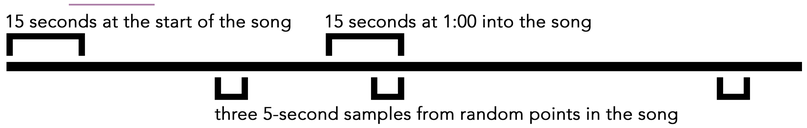
Both give the same result - it is probably category 2 - but with the softmax layer it is visible how sure of the decision the neural network is. Because of this additional information, we used softmax outputs throughout.

genre identification

methods Based on "Automatic Musical Pattern Feature Extraction Using Convolutional Neural Network" (Li, T., Chan, A., Chun, A..), we wanted to split music into at most four categories. For this, we opted to use Philip Tagg's axiomatic triangle, giving us three categories: art, popular, or traditional.

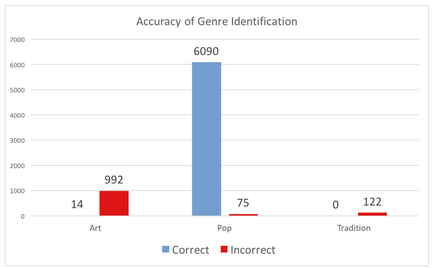
Our dataset was composed of the iTunes library of a member of the research team, but was later expanded through the use of open-domain recordings of 'art' and 'traditional' music. 10-25% of the data was used as validation data.

Three windows were used:

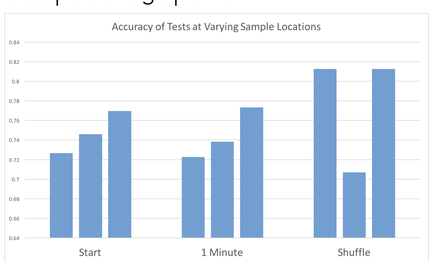


A variety of neural network structures were attempted, with the resulting models and weights stored and the training-reported accuracies logged.

results Overfitting proves to be a significant problem - the graph at right shows that identification of 'art' and 'tradition' failed nearly all of the time, when trained against a biased dataset.



Alleviating this by limiting the size of the sample set allows other trends to come to light. Most notable of these is a trend towards more accuracy when using randomized samples from within the songs, as opposed to samples from a fixed point. Shuffling these samples while training, to create an increase in the amount of coverage of any given song, increased accuracy from 5-20 percentage points.



Even within the biased sample set, using these randomized samples increased accuracy, more than any change to the structure of the neural network did. Comparing the three windows mentioned above showed a minimum 5 percentage point increase in accuracy.

conclusion Identifying which of the three categories a song falls into is a difficult task at times. Christmas music is a great example of how difficult this can be: Pentatonix' cover of *Carol of the Bells* is a three-year-old version of a song that was penned in 1914, which was itself based on an even older Ukranian folk song. Is this popular or traditional music?

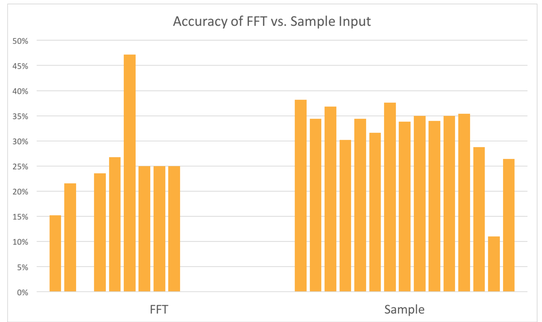
Another source of error could be the nature of music - popular music is based upon what came before it, and much research has gone into the different ways in which aspects of traditional and art music can be identified in popular music.

instrument identification

methods We broke song stems and mixes from MedleyDB (R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. P. Bello) into four categories: vocal solo, guitar solo, drum solo, and non-solo 'other.'

By breaking the songs into samples consisting of 1-3 seconds, with .5-1.5 seconds of overlap, we were able to process entire songs into an array of categorization results. With that, we generated an array of start/stop times and their categories, which can then be passed through a visualization utility to produce graphs as above.

results The highest single-test accuracy was produced by running the input through a Fast Fourier Transform prior to feeding it into the neural network; however, separate testing of that one revealed it to be not nearly as accurate, and we suspect there was an error in the input that allowed it to overfit to the sound of a guitar playing.



Feeding the networks raw samples, without the intermediate step of an FFT, yielded higher average accuracy in detection, routinely outperforming the 'random guess' level of accuracy that the FFT tests produced.

conclusion Instrument identification remains an interesting problem, and one that we feel should be given further consideration. Increasing the size of the windows that the network was given yielded slight increases in accuracy, but we were unable to explore the extent to which this continues. Overall, we recommend further research with the aim of identifying which factors have the largest effect on accuracy.

applications Software like this, once fully-functional, has several useful applications. The obvious is for music information retrieval purposes, where it could be quite useful for aiding researchers in annotating non-notated music.

Looking away from MIR techniques, the same sort of algorithm could be combined with speech-recognition software to automatically annotate panel discussions or podcasts, making them much more accessible for hearing-impaired users, as well as more easily searchable.