Machine Learning Nanodegree: Capstone Project

Proposal for Using Deep Neural Network to Predict Musical Instrument Family

by Sam Wachtel
2019-08-20

Domain Background

The transcribing of recorded Western Music into music notation is a slow, painstaking, and manual task. Therefore, software tools to aid in transcribing music has been a focus of software engineers for years. Early tools focused mostly on notation, pitch tracking, and slow-down software.

Notation software allows the transcriber to input and manipulate notes electronically. This is an advancement over using paper because it's impossible to make certain types of edits on paper, like adding a measure in the middle of the piece or transposing (changing keys) an entire piece.

Slow-down software plays back audio at varying speeds, allowing the transcriber to hear details in the audio that they would not easily be able to hear. This makes the manual transcription of music orders of magnitude easier.

Pitch tracking software is sometimes embedded in slow-down software (Transcribe!, for example) and displays a visual map of the sound alongside guesses of the pitches.

Automatic transcription tools are beginning to become available as well. Currently, the most successful are for piano transcription (AnthemScore for example). These tools do essentially two things:

1. Detect note pitches
2. Detect note start (attack) and end

There are no tools to automatically transpose multiple instruments into their own staves (lines of music) or voices, because the complexity of recognizing individual notes in a complex harmonic mix of multiple instruments playing multiple notes at different times is too large.

However, a necessary step in enabling the transposition of music with multiple instruments is to be able to parse out and label each note in an audio file with not only the pitch and time, but also the instrument name.

For this capstone project, I propose starting with the detection of instrument only. Further steps would be add pitch recognition AND instrument and then the multiple instruments and notes together.

Problem Statement

For this project, the goal is to start with training a machine learning model to take a short wav audio file as input and identify the instrument on which it is played. This is a classification problem that a deep neural network seems a likely candidate for solving.

Datasets and Inputs

The NSynth Dataset by Google Inc. (see NSynth Paper and Magenta in References) was developed to create a neural network that synthesizes new instrument sounds by training on sound samples from existing instruments. The dataset contains 305,979 audio samples in wav format and contains 1,006 instruments sampled for every pitch within each instrument's range at five different velocities. Along with each sample, the following relevant attributes were captured:.

- Pitch
- Instrument
- Velocity
- Source: acoustic, electronic, synthetic
- Family: bass, brass, flute, etc
- Qualities: bright, dark, multiphonic, etc

The advantages to using this dataset are:

- Creative Commons Attribution 4.0
- Designed specifically for machine learning
    - well labeled
    - standardized length
    - fully labeled
    - available in tfrecord format
    - already broken out into Training, Validation, and Testing sets
- Samples are clean with no ambient interference
- Samples are all monophonic
- Samples are all 16kHz
- Dataset is large

Solution Statement

Use the NSynth data to train a deep neural net to label the instrument playing a single note at any pitch or velocity.

Benchmark Model

Thus far, there is no previous research that I can find that addresses this problem. Therefore, it makes sense to benchmark using a naive predictor (random).

Given the prevalence of each instrument in the data, random results would result in an accuracy of approximately 9%.

Instrument	Sample Count	Chances of choosing an instrument correctly at random
Bass	68,955	22.54%
Brass	13,830	4.52%
Flute	9,423	3.08%
Guitar	35,423	11.58%
Keyboard	54,991	17.97%
Mallet	35,066	11.46%
Organ	36,577	11.95%
Reed	14,866	4.86%
String	20,594	6.73%
Synth Lead	5,501	1.80%
Vocal	10,753	3.51%
	Average	**9.09%**

Evaluation Metrics

A simple accuracy score will provide the necessary scoring. The higher the accuracy score, the better. Any score above 9.09% would suggest that the model is working to some degree.

Project Design

The design would revolve around a deep neural network.

1. Read in training data from the training tfrecord file
2. Train the model using the data and only the instrument family label
3. (Not sure how to use the Validation data)
4. Test the model using the test data in the test tfrecord file

Version one of the model would contain few layers (maybe two) and I would iterate on the model from there.

References

A Discriminative Model for Polyphonic Piano Transcription (2007)
http://www.ee.columbia.edu/~graham/papers/piano.pdf

Transcription (music)
https://en.wikipedia.org/wiki/Transcription_(music)

NSynth Paper
Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders." 2017.

Magenta
https://magenta.tensorflow.org/nsynth