Urban Sound Classifier

By Spencer Goble

Background

- Our editors currently spend an average of 12 hours per film classifying field recordings
- The goal is to cut down this time to less than 1 hour of sound organization/classification per film
- Can we make use of machine learning to classify sounds for us?

What

- Our editors manually classify and organize field recordings
- On average we collect 1000 sounds per film
- This process involves listening to, naming, and grouping every single sound
- This time could be better spent in the editing process

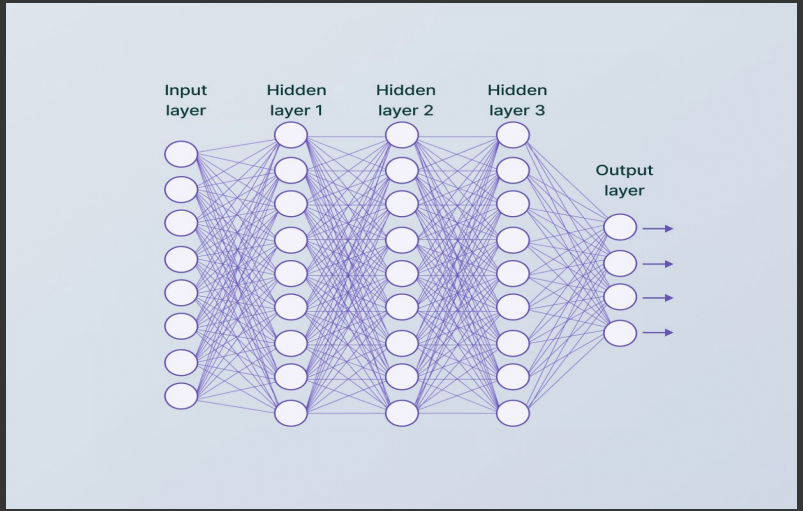
But How!?

- Expedite our process with brilliant…

  Deep Learning techniques!

- With an improved process we can cut time spent from 12 hours to…

  1 hour per film!

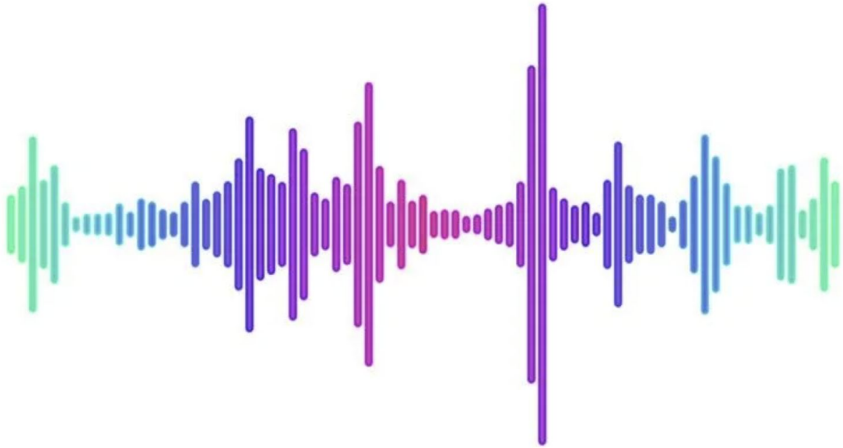


Approach

Gather	Classify	Deploy
❖ Our process of collecting field recordings will remain the same ❖ We will collect ~1000 sounds per film	❖ By training a ***Neural Network*** to recognize sounds, we can use deep learning to categorize them for us ❖ We will 'show' the computer samples of sound and it will tell us what kind of sound it is	❖ Once the algorithm has assigned a sound to a particular class, we need a minimal amount of labelling and quality assurance to be performed by an editor ❖ After some brief organization the sounds will be ready to deploy

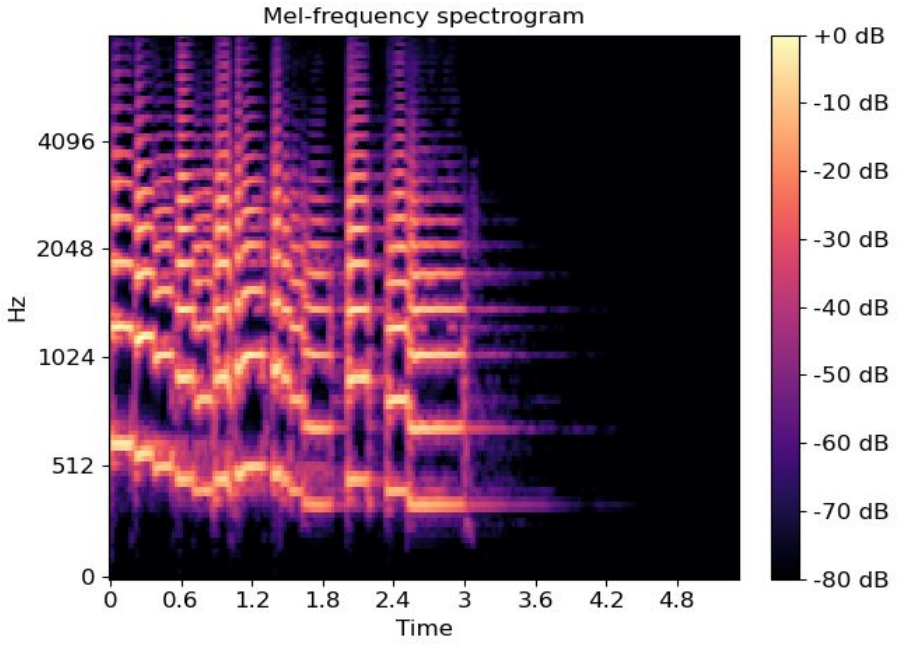
Sound As Data

- We have a standardized process for recording our sounds which assures consistency across our vast libraries of audio
- We will need to generate a .csv file that contains a list of all the sound names/ID's
- This .csv file will be read into an IDE and ***Python*** is the language used to execute the entire process
- Once we load the .csv file, we can scan through our folder of sounds and load them into our IDE



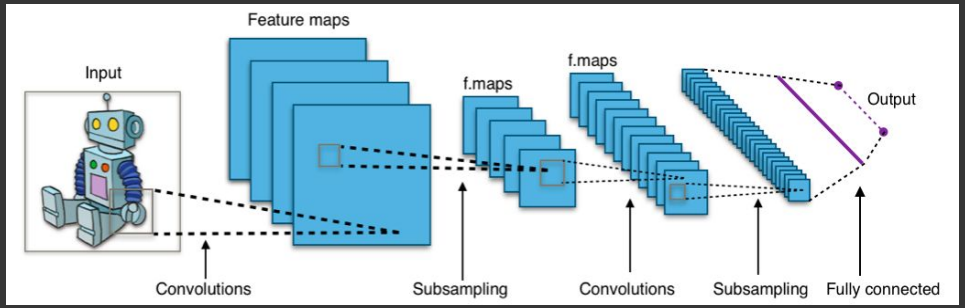
Feature Extraction

- ***Librosa*** is an expansive toolkit for digitally processing audio
- Using Librosa, we can extract characteristics from our sounds
- These characteristics are arrays of numbers that represent the actual frequencies and amplitudes of the audio sample
- One of the most common features to extract is the Mel-Frequency Cepstrum



Modeling

- We will use a toolkit called ***Keras*** to design our Neural Network
- The audio features get fed into the network and it learns what each sound ‘looks’ like
- Once it knows the profile of each sound it can group similar ones together
- Below is the architecture of a Convolutional Neural Network



Conclusions

- Our film editors are spending far too much time manually classifying sounds
- There is highly effective technology available for automating a huge portion of this process
- By employing Deep Learning to classify sounds for us, we free up 11 hours of time per film!

Thank you

Additional Documents

White Paper:

https://github.com/LiftedAquatic/Urban-Sound-Classifier/blob/main/White%20Paper.pdf

Project Repository:

https://github.com/LiftedAquatic/Urban-Sound-Classifier

Original Data:

https://urbansounddataset.weebly.com/urbansound8k.html