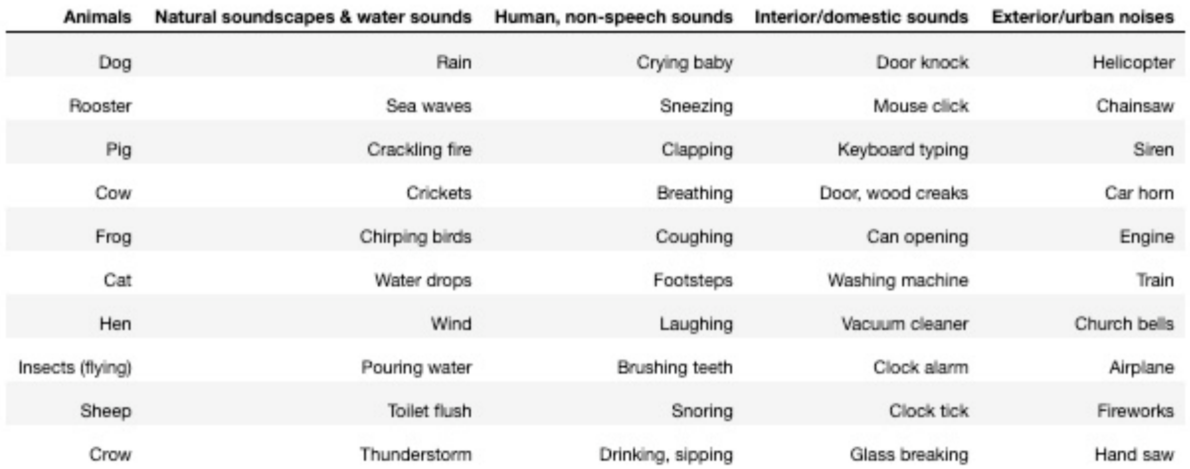
Environment Sound Classification

Submitted by - Gupta, Amit

Problem Statement - Classify 50 different sounds shown in the table



Animals	Natural soundscapes & water sounds	Human, non-speech sounds	Interior/domestic sounds	Exterior/urban noises
Dog	Rain	Crying baby	Door knock	Helicopter
Rooster	Sea waves	Sneezing	Mouse click	Chainsaw
Pig	Crackling fire	Clapping	Keyboard typing	Siren
Cow	Crickets	Breathing	Door, wood creaks	Car horn
Frog	Chirping birds	Coughing	Can opening	Engine
Cat	Water drops	Footsteps	Washing machine	Train
Hen	Wind	Laughing	Vacuum cleaner	Church bells
Insects (flying)	Pouring water	Brushing teeth	Clock alarm	Airplane
Sheep	Toilet flush	Snoring	Clock tick	Fireworks
Crow	Thunderstorm	Drinking, sipping	Glass breaking	Hand saw

Technology - Present audio as MFCC features to series of convolution, and dense layers with categorical cross-entropy classification.

Benefits - Wide ranging from determining noise pollution index, taking automated actions/ decisions when specific sounds are detected.

Challenges - ESC50 is a very limited dataset of just 2000 samples (40 samples per sound). Classifying 50 (a large number) sounds with such a small sample set is challenging.

Current State of Art: Validation accuracy of 87%. (paper)

Result: Achieved 94% test validation accuracy beating the state of art CNN results, and outperforming expert listeners (90% accuracy). Key breakthrough came with data augmentation by varying db levels of existing sound samples, and expanding the dataset. Test validation accuracy shot up from without augmentation (44.5%) to 94%.

Youtube URL:
2 min: https://youtu.be/1a2ukBxTSuA

15 min: **https://youtu.be/PpszrzXutO4**

Dataset Download: https://www.kaggle.com/mmoreaux/environmental-sound-classification-50

ESC-50 Dataset:
The ESC-50 dataset consists of 2000 labeled environmental recordings equally balanced between 50 classes (40 clips per class). For convenience, they are grouped in 5 loosely defined major categories (10 classes per category): • animal sounds, • natural soundscapes and water sounds, • human (non-speech) sounds, • interior/domestic sounds, • exterior/urban noises.

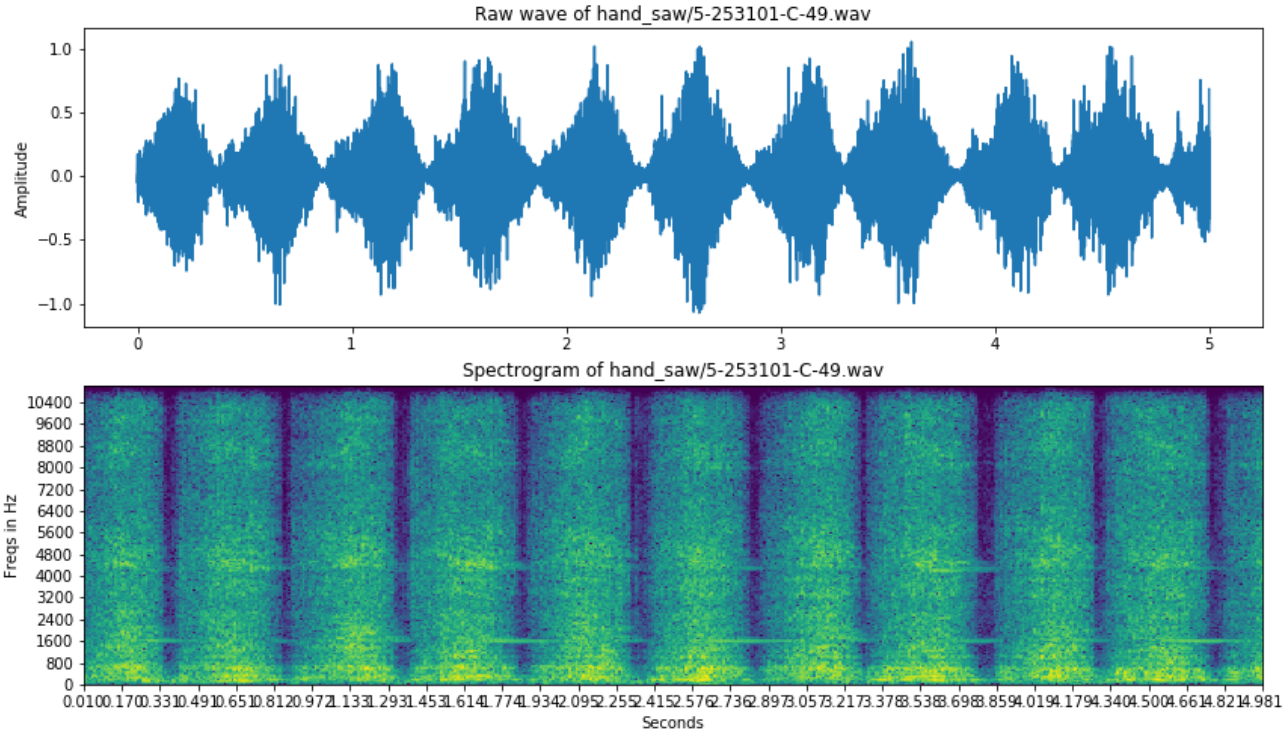
The dataset provides an exposure to a variety of sound sources - some very common (laughter, cat meowing, dog barking), some quite distinct (glass breaking, brushing teeth) and then some where the differences are more nuanced (helicopter and airplane noise). One of the possible deficiencies of this dataset is the limited number of clips available per class.

ESC-10: The ESC-10 is a selection of 10 classes from the bigger dataset, representing three general groups of sounds: • transient/percussive sounds, sometimes with very meaningful temporal patterns (sneezing, dog barking, clock ticking), • sound events with strong harmonic content (crying baby, crowing rooster), • more or less structured noise/soundscapes (rain, sea waves, fire crackling, helicopter, chainsaw). This subset should provide an easier problem to start with, and it was initially constructed as a proof-of-concept dataset.The task of classifying sounds from such a constrained set of classes, a trivial feat from a human perspective, sets the bar really high for accuracy expected from automatic sound recognition systems. Therefore, this subset presents a slightly different problem to tackle than the whole ESC-50 dataset. The differences between classes are much more pronounced, with limited ambiguity, and as such it may favor a different kind of machine learning approaches.

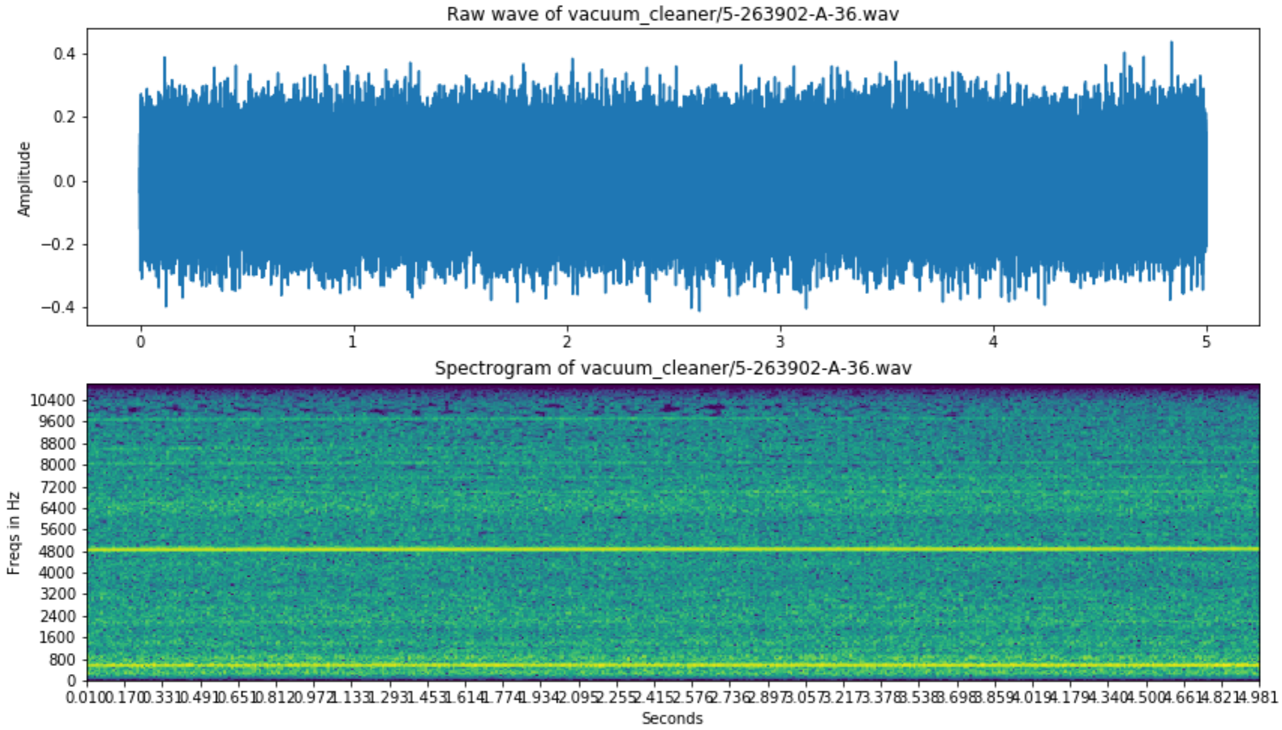
PLEASE NOTE the project classified the more challenging ESC-50 dataset.

Technology details - The sound samples (.wav files) were converted to MFCC features using librosa package, and and these features were presented to sequential model composed of convolution layers followed by dense layers and categorical cross-entropy classification. Please note we used the entire 39 MFCC features Vs 11-13 that are typically used. The frequency range of ESC sounds ranged from 0-11 KHz. So sampling was done at 22.5 KHz to avoid information loss (Nyquist criterion). Given the small sample size, the training was sped up by loading the entire dataset into numpy arrays.

Here are spectrograms of a few example (handsaw, vacuum cleaner) sound samples over 5 seconds showing frequency range from 0-11 KHz.


Raw wave of hand_saw/5-253101-C-49.wav

Spectrogram of hand_saw/5-253101-C-49.wav


Raw wave of vacuum_cleaner/5-263902-A-36.wav

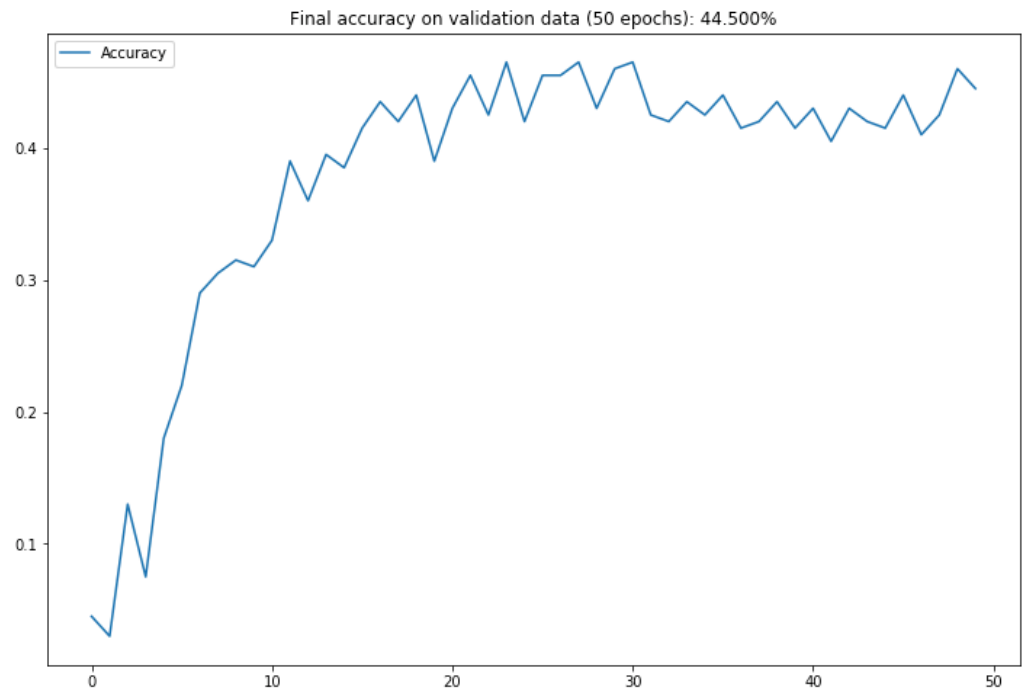
Spectrogram of vacuum_cleaner/5-263902-A-36.wav

Training the network: The training time was under an hour on the following machine configuration.

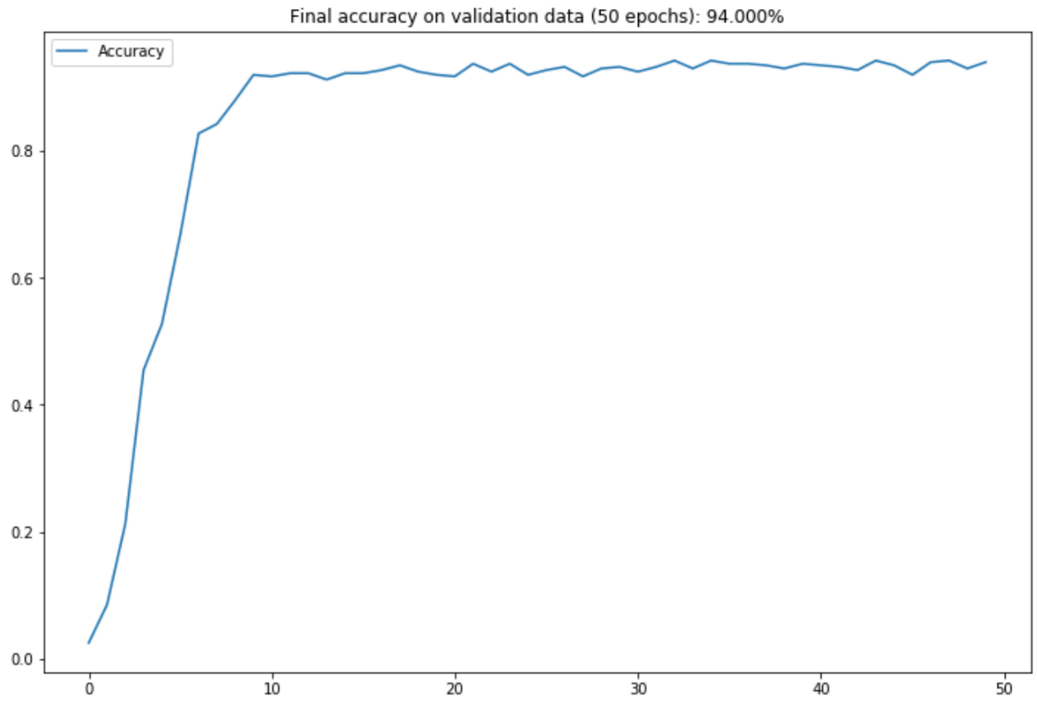
Model Name:	MacBook Pro
Model Identifier:	MacBookPro11,4
Processor Name:	Intel Core i7
Processor Speed:	2.2 GHz
Number of Processors:	1
Total Number of Cores:	4
L2 Cache (per Core):	256 KB
L3 Cache:	6 MB
Memory:	16 GB
Boot ROM Version:	187.0.0.0.0
SMC Version (system):	2.29f24
Serial Number (system):	C02TP0URG8WN
Hardware UUID:	80DEA49C-84CB-53A6-B5BD-82C6E91CA672

Results: The initial results showed test validation accuracy of 44.5% worse than a coin flip. We augmented the data with +-24 db from a uniform distribution, and increased the sample size from 2000 labeled samples to 4000 labeled samples. This dramatically boosted the validation accuracy from 44.5% to 94%. With 11 MFCC features, for unexplained reasons the prediction accuracy on random picking of samples (from train or test set) was low in the 50-60% range. However with 39 MFCC features, the prediction accuracy was 94-100%. I personally don't understand the reasons behind such a large prediction performance variation between 11 and 39 MFCC features. The final results appear to beat the state of art CNN results achieved in this paper. However it is not clear whether the comparisons are apples to apples.

Plot of validation accuracy with original dataset (no data augmentation)



Plot of validation accuracy with data augmentation



Installation and Configuration Steps:

Step 1: Download the dataset from Dataset: https://www.kaggle.com/mmoreaux/environmental-sound-classification-50

Step2: Create a "./data" directory.

Step3: Unzip the dataset, and move the entire directory environmental-sound-classification-50 under the "./data" directory

Step4: You are now ready to execute the Jupyter notebook "FinalProject.ipynb". You can access this notebook from google drive:https://drive.google.com/open?id=1DYd7LBHx3Pc5YpGMqexxxYaDh6yVNjyK

Step5: You will need to import various packages shown in the notebook.

Step6: Explanation of Jupyter notebook flow is in the recording at https://youtu.be/PpszrzXutO4

Links

Dataset description**: https://github.com/karoldvl/ESC-50**

ESC-50 paper**: http://karol.piczak.com/papers/Piczak2015-ESC-Dataset.pdf**