A Survey of Sound Classification Using Classic Methods and Deep Learning

Filippo Ranalli and Hiroshi Mendoza
[franalli; hmendoza]@stanford.edu
CS221 - Fall 2017 - Stanford University

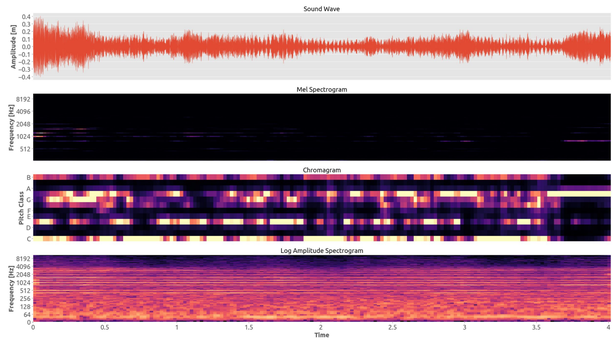
Objective

Approximately 5% of the world population is afflicted by a form of hearing loss, which can lead to difficulties in identifying nearby dangers or meaningful occurrences around them. Such problem motivates research aimed at building systems that can recognize surrounding sounds in real time, and provide rapid feedback to the user. In this project we attempt to tackle sound recognition by analyzing predictive models, using shallow and deep AI techniques to classify environmental and urban sound sources. It is within our goals to analyze their performance, comparing their advantages and shortcomings.

Setup

Librosa: Amplitude-frequency conversion, feature extraction
Sklearn: Metrics, kNN, SVM, Random Forest
Pytorch CPU: Feed-Forward Neural Network
Pytorch GPU: CNN baseline, ResNet, RNN

Dataset



UrbanSound8K: 8732 labeled and evenly distributed data points from 10 different classes: air conditioner (AC), car horn (CH), children playing (CP), dog bark (DB), drilling (DR), engine idling (EI), gunshot (GS), jackhammer (JH), siren (SI), and street music (SM). Each example is a 4 second .wav file. The set is divided into train, dev and test.

References

- CS231n/CS221 Lecture Slides.
- He, Zhang, Ren, Sun, *"Deep Residual Learning for Image Recognition"*.
- UrbanSound8K;https://serv.cusp.nyu.edu/projects/urbansounddataset/urbansound8k.html.
- J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research.

Classic Architectures

KNN

kNN classifies an example by majority vote of the K closest data point labels in the training set, using the L2 norm as the distance measure.

Random Forest

The Random Forest classifier is an ensemble method based on majority voting of decision tree modules. In such model, all observed information is encoded in the edges, and all predictions in the leaves.

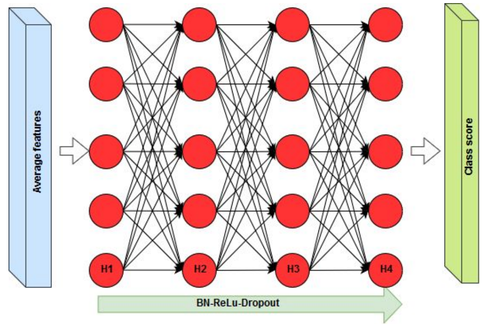
SVM

A support vector machine (SVM) is a classifier that uses a hinge loss and can use a nonlinear kernel to transform the feature domain for improved separability. The loss function and the kernel are described by the following equations:

$$L = \frac{1}{N}\sum_i \sum_{j \neq y_i} H(x_i; W; K) + \lambda W^T K W$$
$$K(x, z) = exp(-\frac{1}{2\tau^2}||x - z||_2^2)$$

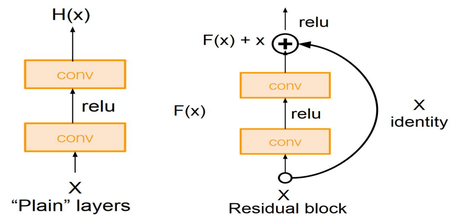
Deep Architectures

4-Layer NN



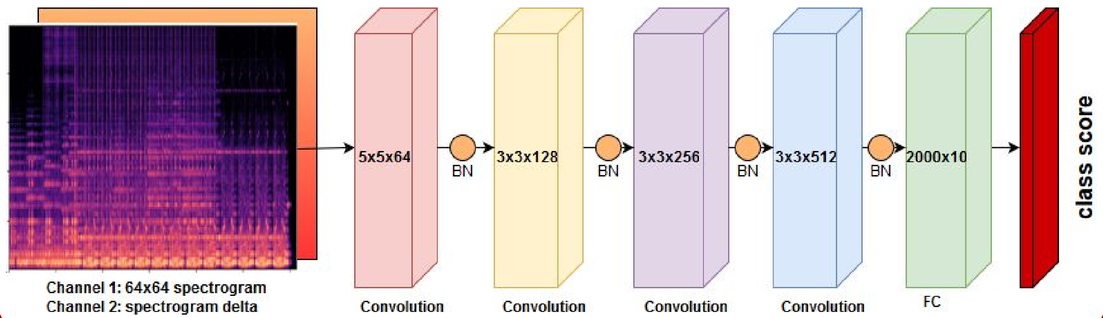
ResNet

The ResNet architecture adopted is an 18 layer CNN with increasing filter depth, batch normalization and residuals every two layers

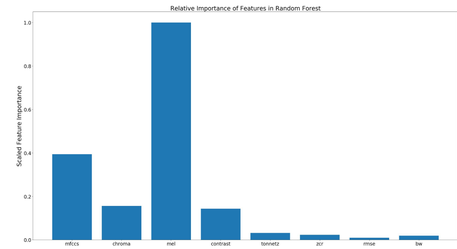


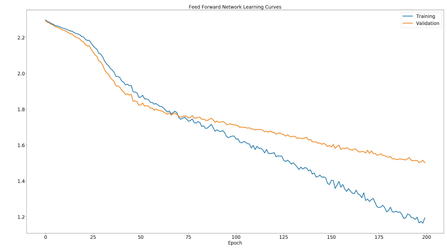
CNN

The baseline CNN features four layers of increasing filter depth, batch normalization on every layer and dropout on the final fully connected layer.



Results





4-Layer NN Confusion Matrix

	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	11	0	3	10	0	0	0	0	0	0
CH	0	22	0	0	8	0	0	0	0	0
CP	24	0	82	8	6	3	3	7	0	21
DB	0	0	6	66	2	5	18	0	1	3
DR	42	2	0	5	38	0	1	24	0	0
EI	0	1	0	1	0	69	0	0	0	0
GS	0	0	0	0	0	0	7	0	0	0
JH	0	4	0	1	31	0	0	51	0	1
SI	0	0	7	2	0	6	0	0	81	0
SM	23	3	2	7	15	6	2	0	0	75

ResNet Confusion Matrix

	AC	CH	CP	DB	DR	EI	GS	JH	SI	SM
AC	15	0	2	6	0	0	0	0	0	0
CH	0	27	0	0	7	0	0	0	0	0
CP	19	0	89	8	4	3	4	2	0	17
DB	0	0	3	73	0	3	15	0	0	0
DR	8	2	0	3	72	0	2	9	0	0
EI	0	1	0	1	0	74	0	0	0	0
GS	0	0	0	0	0	0	7	0	0	0
JH	0	4	0	1	8	0	0	74	0	0
SI	0	0	2	3	0	5	0	0	85	0
SM	17	0	1	8	7	1	1	0	0	88

Model	Parameters	Train Accuracy	Val Accuracy	PCA
kNN	1	0.92	0.53	yes
SVM	64000000	0.89	0.61	yes
Random Forest	2408240	0.94	0.64	yes
Four Layer NN	158608	0.88	0.7	no
CNN	331776	0.85	0.72	no
ResNet	1877635	0.89	0.75	no
RNN	1210433	N.A.	N.A.	no

Conclusion

- Shallow methods are a function of the dataset size and tend to heavily overfit. On the upside they are quick to train and provide significant insight for feature selection.
- The fully connected neural network is a step-up compared to shallow methods, but tends to overfit.
- CNN and ResNet are significantly more expressive and extract features directly on the spectrogram, achieving higher performance.
- The LSTM-RNN will capture sequential time-frequency relations.