



Urban Sound Source Classification and Comparison

Group 21: Yihua Yang, Ling Hong, Ke Liu, Chenwei Dai

{yiy011, lihong, keliu, chdai} @ucsd.edu

Abstract

- Introduction and Dataset.
- Feature Extraction.
- Models and Results.
- Conclusion.
- Future Works.

Introduction

Sonic analysis of urban environments has aroused more and more interest recently. It is of great value to identify these sound sources not only for the disabled but also in many other research fields.

In this project, we use different kinds of feature extraction methods and neural networks to see how the feature extraction methods would affect classification accuracy. Finally, we look into modern neural networks which are proved to have good performance on image classification to see the performance on sound classification.

keyword: accuracy, comparison

Dataset

- 8732 labeled sounds with wave length $\leq 4s$.
- 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music.

Feature Extraction

Mel-scaled Spectrogram

- Slice audio into shorter (overlapping) frames.
- Apply STFT(Short-time FT) to each audio clip
- Mel-Filterbank(triangular filterbank) transforms STFT bins into Mel-frequency bins

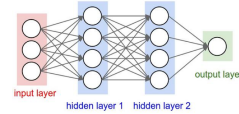
Mel-freq Cepstral Coefficients (MFCCs)

- Mel-frequency features are highly correlated
- Apply Discrete Cosine Transform (DCT) to decorrelate the filter bank coefficients

Models

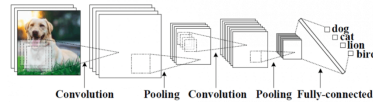
1. DNN

Deep Neural Network(DNN) shows great advantages in speech recognition and image recognition. We use perceptrons to imitate the learning process of our brain. The learning process is composed of two stages: feedforward and backpropagation. Through backpropagation, we update the weights and bias of each layer. Here we use 4 layers with 512 neurons in each layer.



2. CNN

A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, RELU layer i.e. activation function, pooling layers, fully connected layers and normalization layers. Here we use 3 convolutional layers and 2 max pooling layers with 64 neurons in each layer.



3. LSTM(Long short-term Memory)

The problem of DNN and CNN is that they cannot model the variance of the time series characteristic which is very important to sound recognition. We investigate some research on speech recognition nowadays. Many use Long short-term Memory to preserve the influence of the previous status. And LSTM can to some extent avoid gradient exploding and gradient vanishing problem.

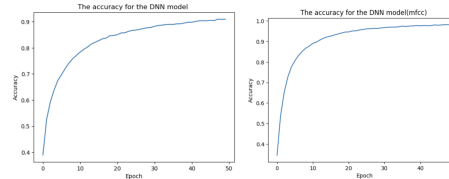
4. Other modern neural networks

- VGGNet** is a neural network that performed very well in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014.
- ResNet** is an artificial neural network of a kind that builds on constructs known from pyramidal cells in the cerebral cortex. Residual neural networks do this by utilizing skip connections, or short-cuts to jump over some layers.

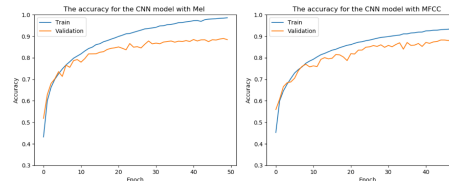
Results

The data is split into three sets (Train:6112, Validation:873, Test:1746). We train and evaluate different models like DNN, CNN and LSTM using Mel frequency and MFCC respectively. The results are shown below.

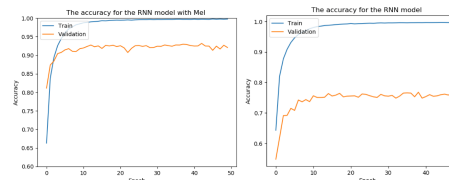
1. DNN



2. CNN



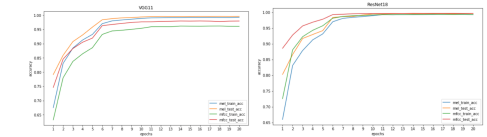
3. LSTM



	Mel			MFCC		
	DNN	CNN	LSTM	DNN	CNN	LSTM
Avg. Train	0.910	0.985	0.998	0.982	0.941	0.998
Avg. Val.	0.722	0.886	0.932	0.793	0.887	0.755
Avg. Test	0.876	0.878	0.916	0.872	0.896	0.909

Table 1: Results of 8-fold cross-validation

4. VGGNet and ResNet



Discussion

- LSTM achieves the best performance in test data, which meets our expectations because the audio signals are highly time-related. LSTM can preserve the influence of the previous status and thus has better performance.
- Mel Frequency behaves better in LSTM and DNN while MFCC behaves better in CNN, VGGNet and ResNet.
- VGGNet11 and ResNet18 have very good performance on sound classification task.

Future Work

- Compare the accuracy on each class. And explore the reason if there are large differences
- Try more sound pre-processing methods
- Extract more features other than Mel, MFCC

References

- Khumarsal, P., Lursinsap, C., Raicharoen, T. (2013). Very short time environmental sound classification based on spectrogram pattern matching. Inf. Sci., 243, 57-74.