Environmental Sound Classification Using Deep Convolutional Nerual Network

Yen Kuei Hunag
108753105

Zheng-Xian Cai
107753034

Yi Li Lai
107753022

Abstract

Environmental sound classification can be applied to many daily life scenarios, and can be used to identify abnormal events nearby. We proposed a framework to classify these environmental sounds. In signal Processing, spectrogram is a representation of short-term freqeuncy power distribution of a sound. Mel-Frequency Cepstrum (MFC) is also representation of the short-term power spectrum of a sound like spectrogram, but it scaled the power which make MFC more closed to human hearing. The ability of deep convolutional neural network (DCNN) to learn spectral patterns makes them well suited to environmental sound classification. We proposed a DCNN model based on MFCCs and spectrogram, which can capture signal signature precisely.

1 Introduction

Environmental sounds come from all the time and everywhere. Understanding what kind of sound it is would be useful for whole city, i.e., if some devices detect gunshot in the block, notify the police in the first time. It could prevent homicide effective, and also economize manpower massively. In this case, if we got model which can differentiate and classify specific sound would be helpful to entire human society.

Acknowledging outlook of environmental sound classification, lots of works published in an incredible speed. Among those works, deep neural network is most widely used. Deep neural network is a way to learn abstract information from input data, which have great ability to find out non-linear relation, however common fully connected neural network could not process time series data as well as other data. Except fully connected neural network, most of work used deep convolution neural network (DCNN) in signal processing research, which can capture signal feature more precisely. DCNN first used in image field, and got an excellent performance. DCNN use convolutional layer to capture features of image, pooling layer to de-noising and decrease feature dimensions. After extracting features from previous layer, connecting to fully connected layer to model non-linear relation between features of input data and label. Traditional fully connected neural network, which flatten RGB channel of each pixel as data input, might losing spatial and color relation of neighbored pixel. Compared to traditional one, DCNN could extract relation of neighbored pixel and model it better by utilization of convolutional and pooling layer.

Sequential sound wave is hard to observe and decompose, extract discrete signal feature which can be used in traditional neural network directly is challenging. Luckily, we have spectrogram which can represent sound like an image. Spectrogram is visual representation describing how spectrum change along the time, and usually shown as a heat map. Spectrum comes from Fourier transform describes distribution of frequency power composing whole signal. Adding a window function on Fourier transform, we can get short time Fourier transform which can get spectrum in a small time range. Combining every spectrum of each time window, we can observe how distribution of frequency power varies along the time, that is, spectrogram.

Mel-Frequency Cepstrum (MFC) is also a representation of signal. Like spectrogram, MFC can be seem as power spectrum of change along the time. Main difference is how they deal with spectrum, MFC map spectrum with mel scale and cosine transform. First step of MFC is map power of the spectrum onto mel scale, names as mel frequencies. Next, take logs of these mel frequencies. Final, calculating the discrete cosine transform of log mel frequencies, as it were signal, and that is what Cepstrum means, ”spectrum of a spectrum”. As high level feature, MFC can capture audio signal feature most approximate to human hearing. However, MFC is sensitive to noise.

In our work, we seem spectrogram and MFC as image, which can be fed to DCNN as input. This method has been proposed before and tested by many research. We will test the performance of DCNN by using spectrogram only, MFC only, and combine both. The dataset we used is ESC-50, which opened on Kaggle. ESC-50 consisted of 50 different environmental sound class, each one in class have 40 audio sample which each is 5 seconds.

2 Related Work

Depend on [2], a considerable amount of research has been made toward modeling and recognition of environmental sounds over the past decade. By environmental sounds, we refer to various quotidian sounds, both natural and artificial, i.e., sounds one encounters in daily life other than speech and music.

Many different methods and algorithms are developed for identifying noise sources. A group of studies focus on the classification techniques such as Hidden Markov Models (HMMs), statistical pattern recognition systems, Artificial Neural Networks (ANNs), fuzzy logic systems, etc; While another group of studies focuses on the extraction of the feature parameters such as Linear Prediction Coding (LPC), Perceptual Linear Predictive (PLP) and

Mel-Frequency Cepstral Coefficients (MFCCs)....

F. Beritelli et al. [1] proposed a system based on MFCCs feature and ANNs classifier. L. Ma et al. [5] proposed a system based on MFCCs feature and HMM classifier, and it showed a good performance on classification of 11 kinds of environmental noise sources. L. Couvreur et al. [3] presented a classification system based on ANNs coupled with HMMs and PLP feature, and it showed a classification accuracy of 85% for urban environmental noise sources, and [4] show a lot of deep learning methods for environmental sound detection.

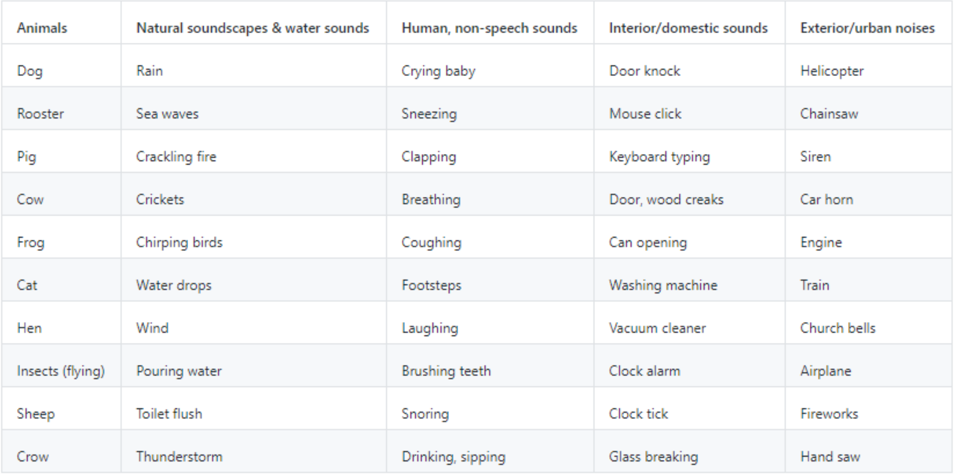
Based on the analysis of the existing sound classification systems, we found that most studies showed a good performance in experiments conducted in lab environments. However, we consider that those systems are difficult to implement in real environments, since there exists various types of sound sources. In this study, we aimed to real-time application of sound classification, especially robust in real environments. We proposed a classification system based on MFCCs and CNN, which is not only considering the computational cost of algorithms but also the classification performance in real environments.

3 Methods

We split this section into four parts: **Data, Preprocessing, Data Augmentation** and **Model**.

3.1 Data

We used **ESC-50** dataset [6] that is most commonly used in environmental sounds classification. The dataset consists of 5-second-long audio segment which can be classified into one of 50 sound events, e.g., dog barks, sea waves sound, rain sound. The 50 sound event class could loosely arranged into 5 major categories(fig.1): **Animals, Natural Soundscapes & Water Sound, Human(Non-Speech Sounds),** and **Exterior/Urban Noises**. Each categories include 10 class. The dataset included 2000 environmental audio sample(with 40 example per class).



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

Figure 1: ESC50 - Categories

3.2 Preprocessing

The data we got is in the form of wave file, so that we can not feed it into CNN directly. We converted environment sound waveform to spectrogram and MFCC. These two can seem as visual representation of audio. We often use heatmap to visualize spectrogram and MFCC. X-axis is time, Y-axis is frequency of audio and color shows frequency strength of audio at that moment. Most difference of spectrogram and MFCC is that the latter scaled based on human hearing where the former only do log scaled. Fig.2 and fig.3 show the clear difference between waveform and spectrogram.

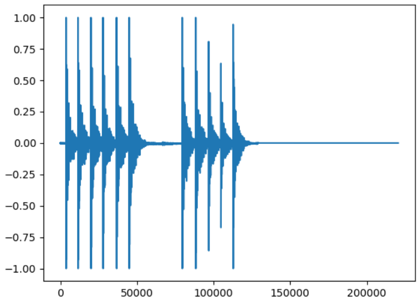


Figure 2: Door Knock Waveform

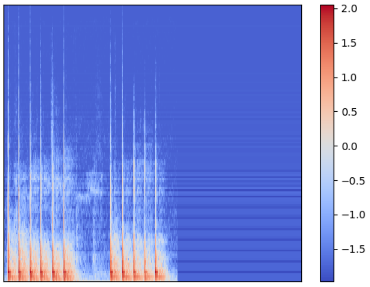


Figure 3: Door Knock Mel-Spectrogram

3.3 Data Augmentation

Due to the lack of training sample, we also perform three easy transformation to extend our dataset. First is **white noise**(fig.4&5), we add white noise which is a random signal having equal intensity at different frequency as the background. That make sense because that also lots of noise in real world. Second thing we did is **sound shift**(fig.6&7). Smoothly shift signal in random rate. For example, if dog barks at time = 1s, we will shift this event to time = 3s with same amplitude, length.... The last augmentation we did is **stretching sound**(fig.8&9), which modify signal sound length to make sound play a little faster.

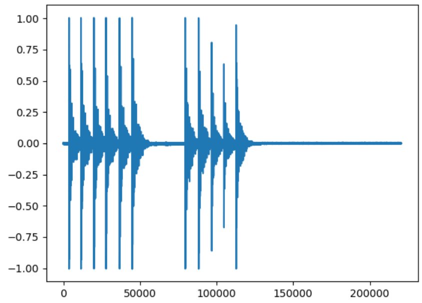


Figure 4: Door Knock Waveform(White Noise)

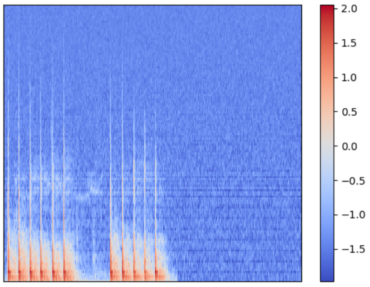


Figure 5: Door Knock Mel-Spectrogram(White Noise)

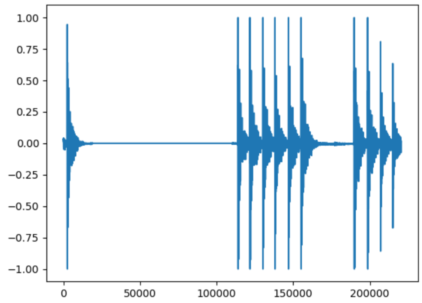


Figure 6: Door Knock Waveform(Sound Shift)

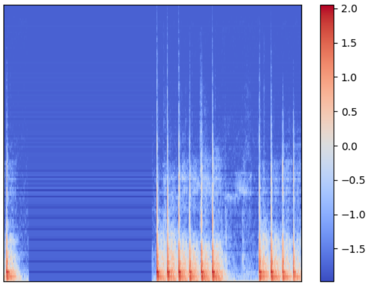


Figure 7: Door Knock Mel-Spectrogram(Sound Shift)

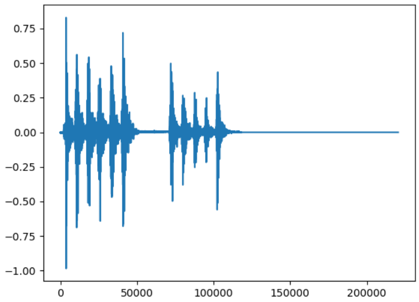


Figure 8: Door Knock Waveform(Stretch Sound)

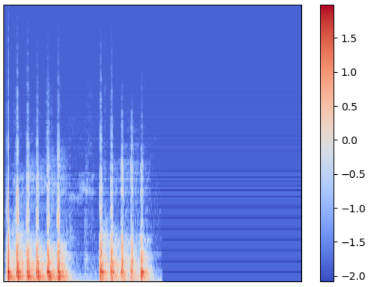


Figure 9: Door Knock Mel-Spectrogram(Stretch Sound)

3.4 Model

We use convolutional Neural Network(CNN) as our model. Filter is a important component in CNN, which can extract significant part in an image. We adopt many filters with different size and amount, and expect to produce greater performance for environment sound classification. Against the number of filters, we tried 32,64 and 128. For the size of filter, (1,8) (8,1) (1,16) (16,1) (1,32) (32,1) (1,64) (64,1) have been used because in environment sound classification, most important features are time and frequences, and it contruct an feature map like image. However, sinece we bulid CNN with many layers, we use zero padding to avoid descreasing output in each layer. Finally, we can see form the graph, which show the model is more precisely when epochs increase and get better performance. Due to cost of time, epoch is only set max to 100. We also tried CNN with less layer, but the performance drop a lot. Showing that deep neural network is necessary.

4 Results

Because running the model is very time-consuming, we didn't do many experiments as possible and final epoch is set to 100. At the beginning, we got poor performance with fewer layer like 3 to 4, and didn't save the result. On the process of increasing layers, we can see that loss is decreasing and accuracy is increasing from the graph. From other perspective, result with mel-spectrogram is better than MFCC, which surprise us. Because depend on past research and many references, MFCC is widely used feature in automatic speaker and sound classification, and there is less people to process mel-spectrogram directly, so we assume experiment outcome would be superior with MFCC than mel-spectrogram. Fig.10&11 showed accuracy and loss along the epoch in the training process.

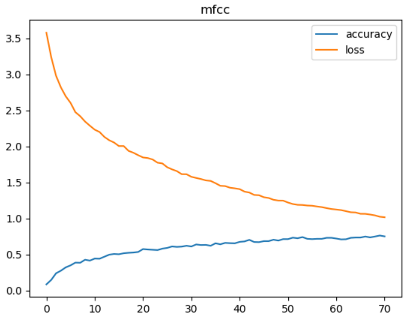


Figure 10: MFCC Training Accuracy and Loss

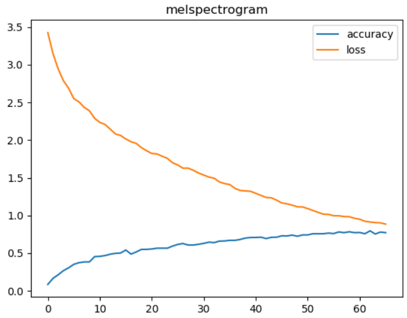


Figure 11: Mel-Spectrogram Training Accuracy and Loss

5 Discussion

To compared our result, we built two “naive” classifier (Random Forest, SVM) to distinguish performance. We converted original waveform into Mel-spectrogram same as Sec.3.2 but with out data augmentation as training data of naive classifier. As the following showed, it is clear that our result is much better than other two classifier. Even using MFCC got huge improvement. We thought the main contribution might be the CNN, however 2000 sample in original dataset is too small for deep learning. Data augmentation is required and bounded with CNN (in this task). All in all, data augmentation and CNN play important role and improve performance.

Method	**Mel-Spectrogram**	**MFCC**	Random Forest	SVM
Accuracy	0.796	0.766	0.443	0.396

6 Conclusion

In this study, we proposed a CNN model to classify environmental sound. We use ECS-50 dataset with 2000 sample. To augment data, we used three methods : White Noise, Sound Shift and Stretching Sound. Though CNN accepted image as input, we converted waveform to Mel-spectrogram and MFCCs, both is widely used as feature in audio domain. We also built two classifier without data augmentation to compare our works. The results showed that using CNN with Mel-spectrogram as feature is superior than baseline, even using MFCCs as feature improve accuracy a lot.

In the future, we will try more methods to augment data and tuning the architecture of CNN. In this work, we only performed simply augmentation and CNN, there are still lots of improvement we can do. Besides data augmentation, we also want to deploy our model on several IOT device, i.e., Raspberry Pi, which can help us monitor city event at the first time. We thought that would improve our city security and reduce damage from emergency accident.

References

[1] Beritelli, F., and Grasso, R. A pattern recognition system for environmental sound classification based on mfccs and neural networks. pp. 1 – 4.

[2] Chachada, S., and Kuo, C.-C. J. Environmental sound recognition: A survey. vol. 3, pp. 1–9.

[3] Couvreur, L., and Laniray, M. Automatic noise recognition in urban environments based on artificial neural networks and hidden markov models.

[4] Li, J., Dai, W., Metze, F., Qu, S., and Das, S. A comparison of deep learning methods for environmental sound detection. pp. 126–130.

[5] Ma, L., Smith, D., and Milner, B. Environmental noise classification for context-aware applications. vol. 2736, pp. 360–370.

[6] Piczak, K. J. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, ACM Press, pp. 1015–1018.

[7] Shen, G. An environmental sound source classification system based on mel-frequency cepstral coefficients and gaussian mixture models. pp. 1802–1807.

[8] Shibuiwilliam. Audio classification keras. `https://github.com/shibuiwilliam/audio_classification_keras`, 2018.