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.