NAME- DEVESH KUMAR

ROLL NO- 190277

AUDIO EVENT DETECTION

INTRODUCTION

Predicting the presence of audio events in an audio clip is the goal of audio event classification. In Audio Set, each audio clip given as spectrogram coefficients contains one label, such as “Microwave oven”, “siren” and “Meow”. Audio Set is a weakly labelled dataset, that is, only the presence or absence of audio events are known in an audio clip, without knowing the onset and offset time of the audio events. In the weakly labelled dataset, the duration of the audio events varies depending on audio categories. Some audio events in an audio clip last for several seconds, while some audio events only last for short period of time. To solve this weakly labelled dataset problem many methods have been proposed and applied on weakly labelled audio classification

LITERATURE SURVEY

The goal is to give the underlying knowledge necessary to build a reliable event detection system. The basics of audio processing for sound event detection are explained at the outset, along with how valuable data can be extracted from an audio signal. The chapter continues by outlining the research on feature extraction and data augmentation, which will be the project's core focus. After that, the chapter's discussion on SED system evaluation follows. This is demonstrated in two different ways: first, using formal methods to provide a quantitative assessment using F-Score and Error Rate; and second, by investigating graphical systems that visually depict the output labels provided by a sound event detection system.

METHODS

We proposed two methods to solve this weakly labelled audio classification task. The dataset is given as spectrogram (mfcc). we found that number of frames in each file are different. So first we made the frame size same for all files.

This is done in two ways-

1- We used a reference shape (128,2500) and zero padded or cropped each file to make it of desired size. Then we made a convolutional neural network containing several convolution layers and dense layers at the top. We used 5 blocks each containing 2 convolution layer, max pooling , batch normalization and dropout. We increased the number of filter in each block (4,8,16,32,64). After that we used global average pooling and two dense layers of 32 and 10 units. Softmax function is applied at the final layer.
2- Another interesting method is to make images of spectrogram and using them as the input. We resized all the images to (256,256) after that we trained our model described in 1.

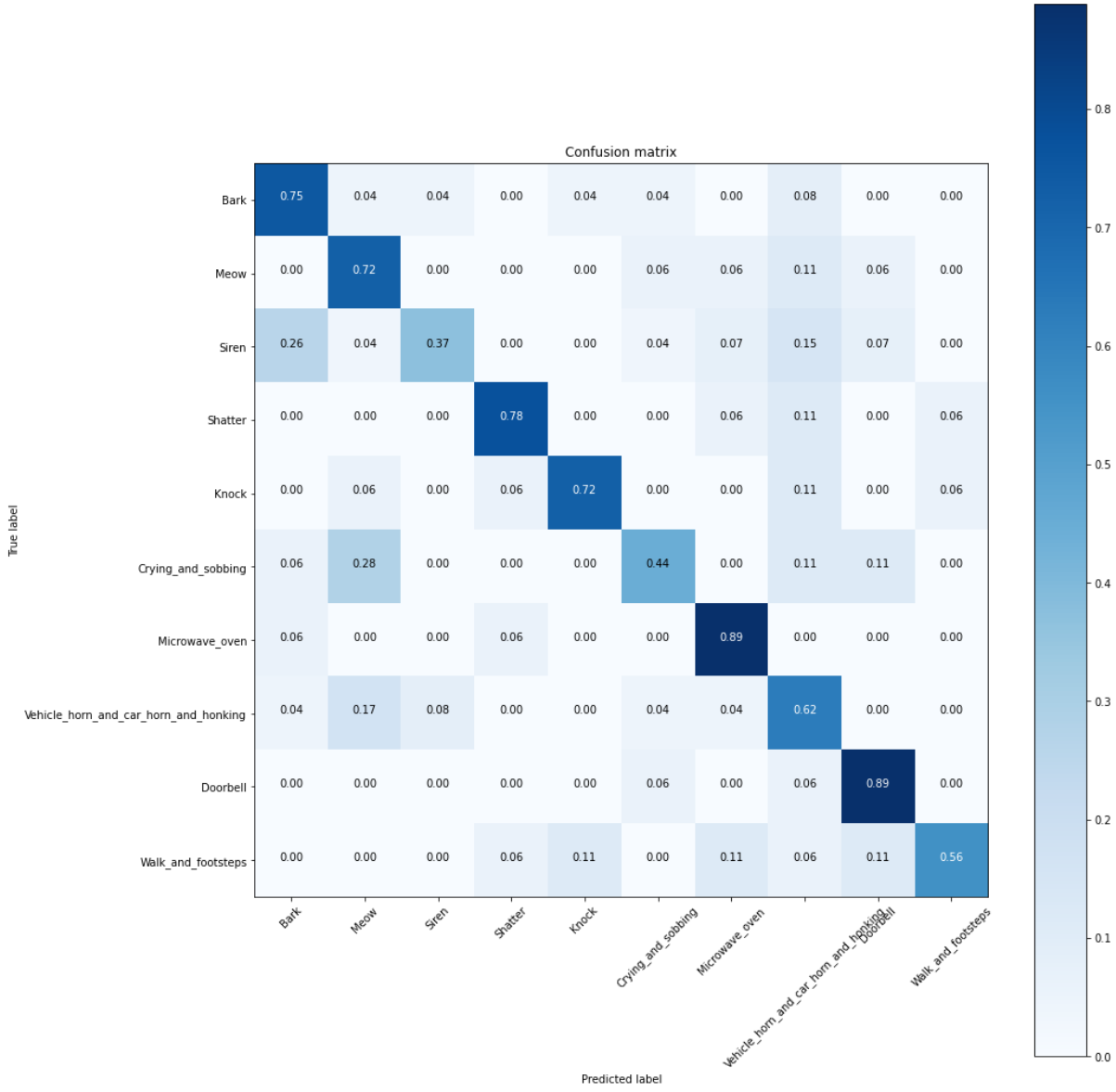
RESULTS

F1 score: 0.6682

Precision score: 0.6852

Recall score: 0.6616

Confusion Matrix



OBSERVATIONS AND DISCUSSION

We observed that method 1 is performing better than method 2. This may happen due to the loss of

	Precision	Recall	F1 score
Method1	**0.6852**	**0.6616**	**0.6682**
Method2	0.6419	0.6143	0.6277