Audio Event Detection

ELEC-E5510 Speech Recognition project

Anand Umashankar, Anssi Moisio

Introduction

- Audio event detection aims to isolate and detect discrete incidents from audio
- Sound event detection can be utilized in a variety of applications
  - indexing multimedia databases
  - monitoring in health care
  - surveillance
- detected events can be used as mid-level-representation in other tasks
  - recognition of audio context (acoustic environment)

- Our aim is to classify audio events using two different classifiers
  - Support Vector Machine
  - Convolutional Neural Network

Babaee, Elham, et al. "An overview of audio event detection methods from feature extraction to classification." Applied Artificial Intelligence 31.9-10 (2017): 661-714.

Outline

- Introduction
- Dataset
- Feature extraction
- Classifiers
  - Support Vector Machines
  - Deep Neural Networks
- Results
- Conclusion

Dataset

- Audioset - A large-scale dataset of manually annotated audio events
- The dataset consists of annotated youtube videos of the audio events
- The dataset has 2.1 million annotated videos, 5.8 thousand hours of audio, **527 classes** of annotated sounds

YTID	start_seco	end_seco	positive_labels
0rnm40dIl	350	360	"/m/01j3sz"
0rrDTv-DK	30	40	"/m/042v_gx
0ry6MOZn	30	40	"/m/01hsr_"
0sb5UXdX	0	5	"/m/014zdl"
0tEX2Pdlg	30	40	"/m/042v_gx
0tJRyAzkr	30	40	"/m/042v_gx
0tQOcpVR	30	40	"/m/01hsr_

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., ... & Ritter, M. (2017, March). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 776-780). IEEE.

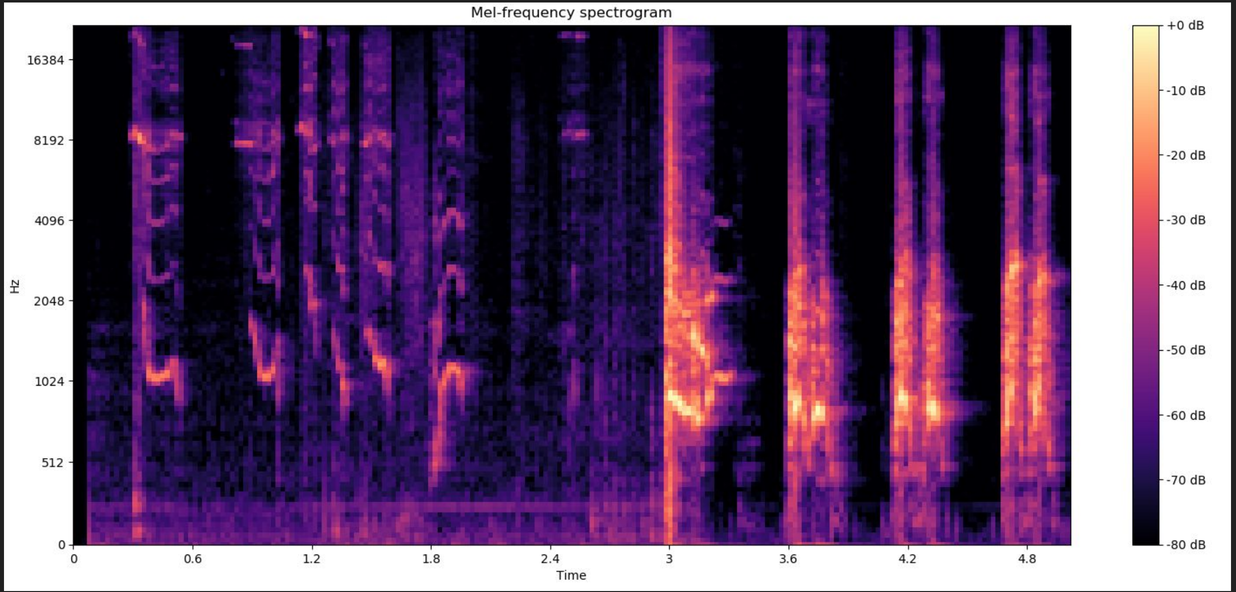
Dataset Preparation

- We have chosen 8 classes and 9713 audio files.
- The script for downloading the dataset has three steps
  - Download the video from Youtube
  - Extract audio from the downloaded video
  - Cut the audio file based on the start time and end time of the audio events

Dataset Sample : The following sample audio is merged file of all the 8 classes that we have selected. The classes are Acoustic Guitar, Bark, Bell, Explosion, Laughter, Siren, Sneeze, Thunder.

Data Preprocessing and Feature extraction

- for the CNN:
    - mel frequency spectrogram
- for the SVM:
    - MFCC
    - deltas and delta-deltas
    - total spectrum power
    - sub-band powers



Babaee, Elham, et al. "An overview of audio event detection methods from feature extraction to classification." Applied Artificial Intelligence 31.9-10 (2017): 661-714.
Cakır, Emre, et al. "Convolutional recurrent neural networks for polyphonic sound event detection." IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.6 (2017): 1291-1303.
Dang, A., Vu, T. H., and Wang, J.-C. (Dec 2017). A survey of deep learning for polyphonic soundevent detection. pages 75–78. IEEE.

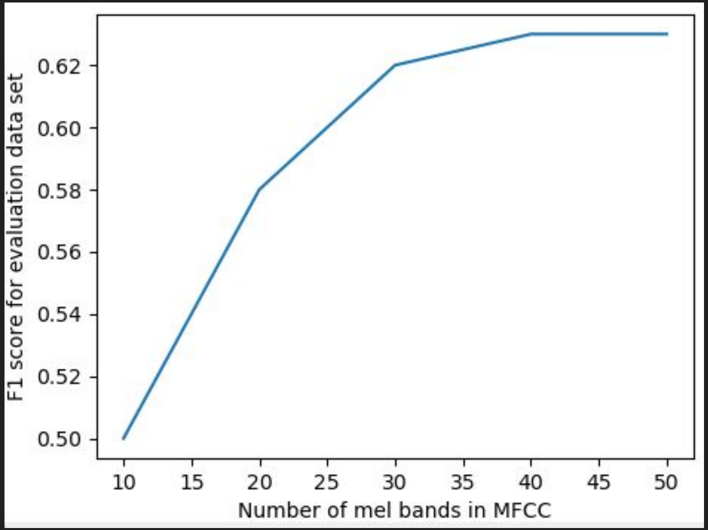
Support Vector Machines (SVM)

- supervised learning
- divide data space into two subspaces with the hyperplane that maximises the margin between the hyperplane and the nearest data point
- multiclass classification problem can be solved by multiple one-vs-one or one-vs-rest classifications
- radial basis function kernel

Guo, G. and Li, S. Z. (2003). Content-based audio classification and retrieval by support vector machines. IEEE transactions on Neural Networks, 14(1):209–215

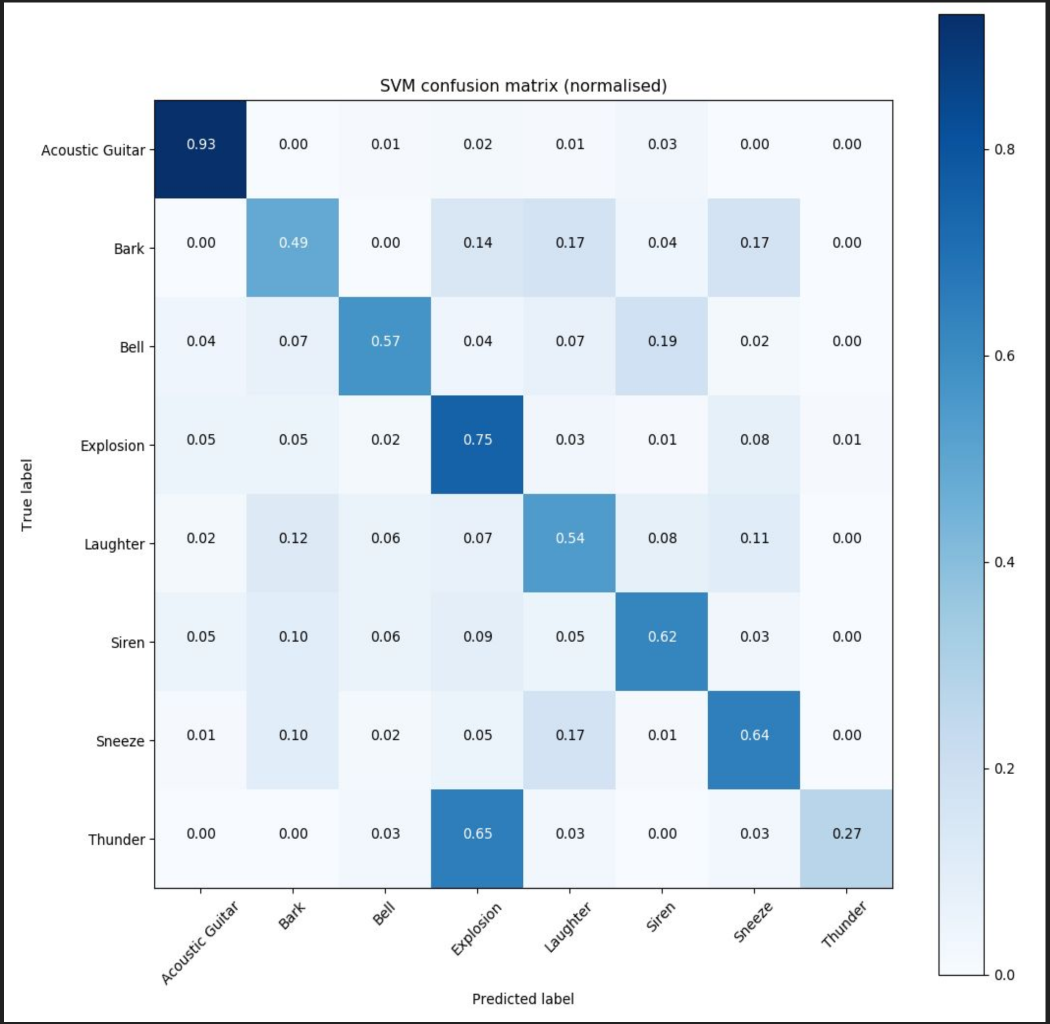
Features for SVM

- evaluation of features still on-going
- the best result so far with MFCCs using
  - 40 mel bands
  - window length of about 40ms
  - frequency range from 0 to 22kHz
- delta and delta-delta features don't improve the results
  - temporal features are probably not as important as spectral features in this task



Results: SVM

- the SVM classifier has achieved an F1 score of 0.63, 65% precision
  - training set of about 400 samples of each class (except for thunder, 140 samples)
- improvements still expected
  - more data
  - experiments with different features



Convolutional Neural Networks

- CNN is a architecture developed to overcome drawbacks of NNs when dealing with spatial structure data.
- The audio samples are converted into Mel Spectrogram Image and used for training the CNN.
- The image size is 256*128, where the image has 128 mel features for each time window.
- A CNN consists of three basic components: convolutional layers, pooling layers, and fully-connected layers.

Network Architecture

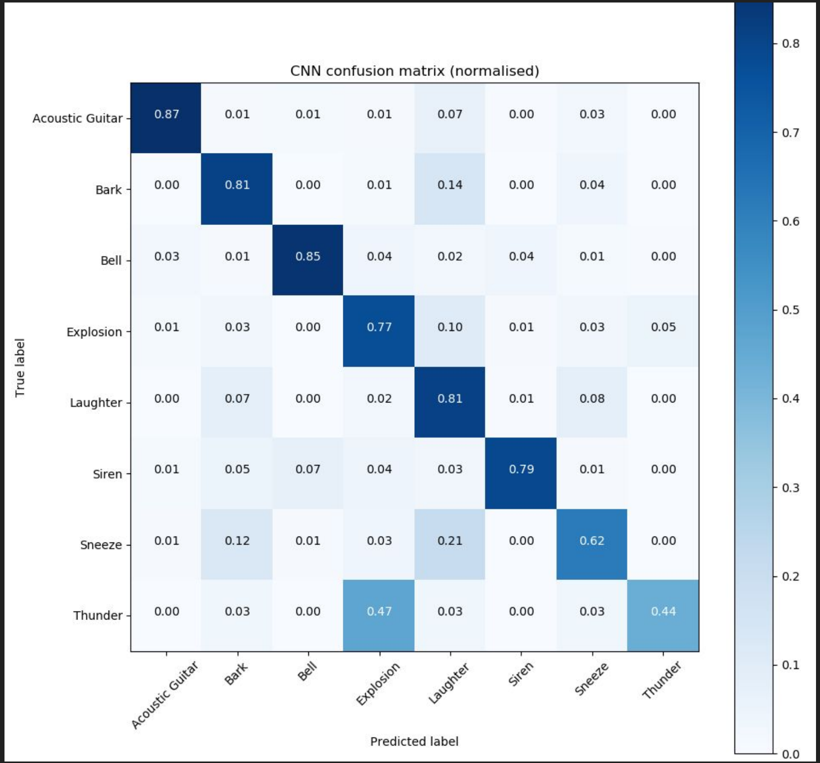
- The CNN network has around 12M parameters
- 5 layers of convolutions intertwined with pooling layers and feeds into a fully connected layer
- A dropout layer with ratio 0.5 to avoid overfitting is added

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 256, 128, 32)	320
conv2d_2 (Conv2D)	(None, 254, 126, 64)	18496
max_pooling2d_1 (MaxPooling2	(None, 127, 63, 64)	0
conv2d_3 (Conv2D)	(None, 125, 61, 128)	73856
max_pooling2d_2 (MaxPooling2	(None, 62, 30, 128)	0
conv2d_4 (Conv2D)	(None, 60, 28, 256)	295168
max_pooling2d_3 (MaxPooling2	(None, 30, 14, 256)	0
conv2d_5 (Conv2D)	(None, 28, 12, 512)	1180160
max_pooling2d_4 (MaxPooling2	(None, 14, 6, 512)	0
conv2d_6 (Conv2D)	(None, 12, 4, 1024)	4719616
max_pooling2d_5 (MaxPooling2	(None, 6, 2, 1024)	0
dropout_1 (Dropout)	(None, 6, 2, 1024)	0
flatten_1 (Flatten)	(None, 12288)	0
dense_1 (Dense)	(None, 500)	6144500
dense_2 (Dense)	(None, 8)	4008

Results (CNN)

- CNN has achieved an accuracy of 83.53% on the test set
- Training was performed on 5200 audio samples and tested on 1950 samples (80:20 split).
- The CNN was trained with multiple hyperparameter configurations to obtain high accuracy (learning_rate, epochs, max_pool_stride length, batch-normalisation, dropout_ratio).
- Experiments were also done to try out different configurations for generating the Mel Spectrum (n_mels, n_fft)

Accuracy Analysis



Conclusion

- CNNs perform better than SVMs in this task
- an SVM needs selected features as the input whereas a CNN learns to extract features from the spectrogram
  - selecting the features for the SVM is tricky because the sound events are varied in length and other properties
- Results are not (yet) fairly comparable between SVM and CNN
  - different training set sizes
  - different features as inputs

Q&A