THE ANDREW & ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

SIPL Signal and Image Processing Lab

TECHNION Israel Institute of Technology

Gunshot Detection in Video Games

Amit Ben Aroush and Asaf Arad, Supervised by Hadas Ofir

In collaboration with WAVES

Introduction

- Gunshot detection is a Feature that upgrades the gaming experience.
- Helps dealing with unseen/hidden enemies.
- Gunshots detection systems already exist in real world for security.

Goals

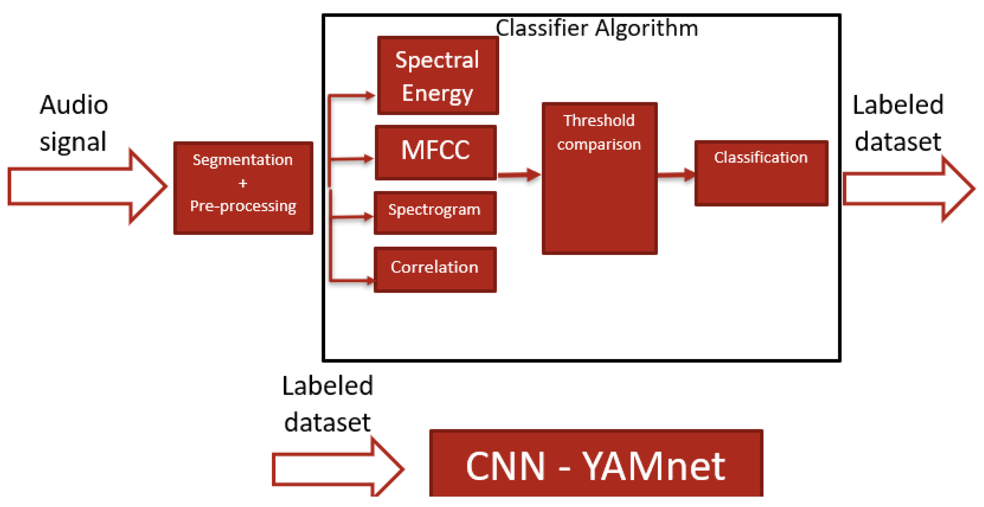
- Automatic system for real-time acoustic detection of gunshots in video games.
- Generic gunshot in generic video game.
- We wish to prove the feasibility of using deep-learning methods for detection of generic gunshot in video games.

Challenges

- Lack of labeled datasets.
- Restricted sounds in video games.
  - Variety of sound.
  - Fixed synthesized sound patterns.
- Multiple gunshots in a frame- bursts.

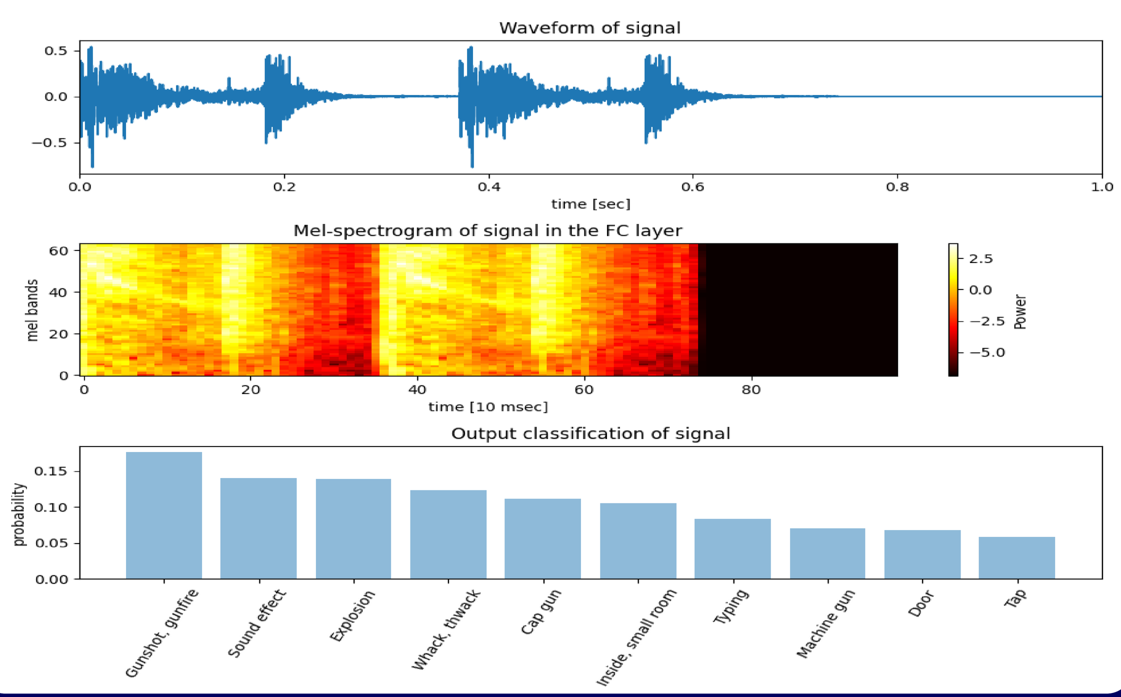
Chosen Solution

- Manual analysis of basic examples.
- Build a classifier algorithm that pre-processes the data and selects optimal classification features.
- Building the Dataset.
- Use a pre-trained network and alter it to classify gunshots using transfer learning.
  - Convolution network called YAMnet.



Network I/O

- System's input: audio waveform signal.
- Waveform is converted to Mel-spectrogram, which is the input to the network layers.
- After the SoftMax layer, the classification output of the class with the maximum probability is chosen.

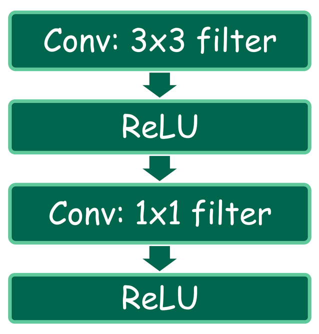


The Database

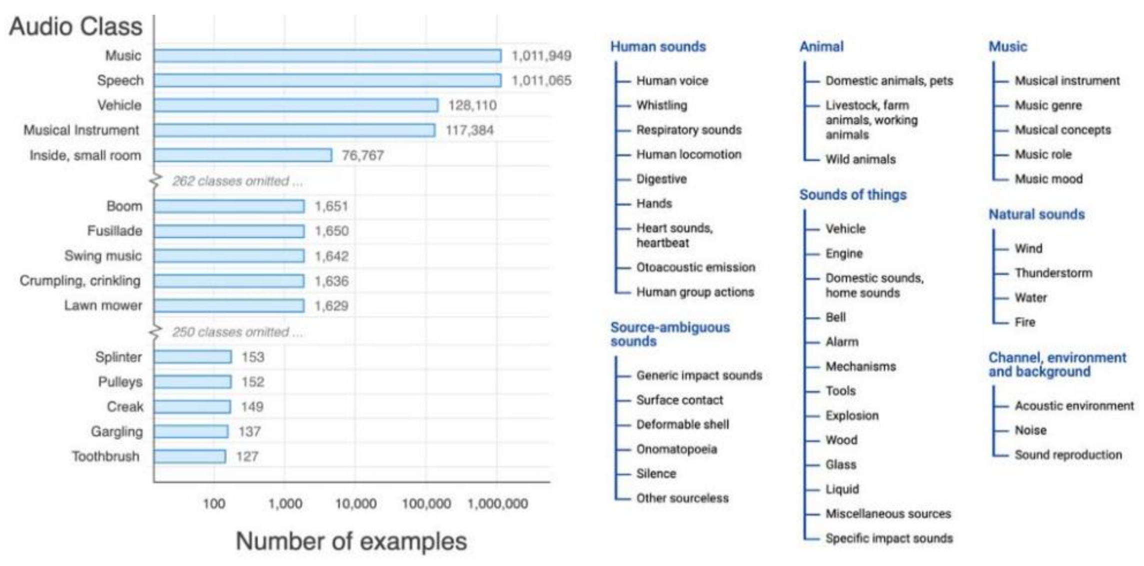
- Extract features:
  - Correlation with sample
  - Energy
  - Spectrogram features
  - MFCC
- Intersection between the classifier labeling and the manual labeling we performed.
- The database was adapted in its characteristics to the database on which the network was trained:
  - Raw data.
  - Resample to 16KHz.
- Data augmentation:
  - Reverberation.
  - White Noise.
- The dataset for training and testing the network: 2843 frames, contains 1100 gunshot frames and 1743 non-gunshot frames.

YAMnet Model

- CNN for acoustic classification:
  - Separable convolutional inner layer



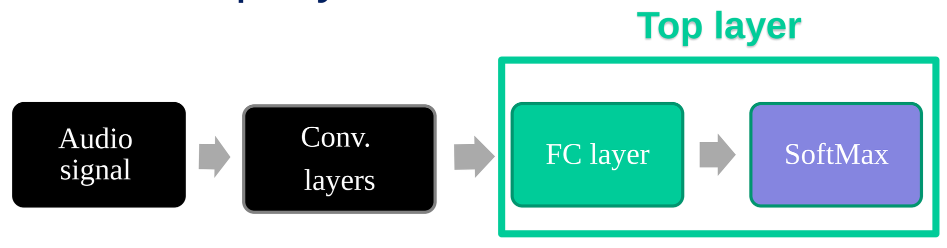
  - Fully connected output layer
  - SoftMax activation
- 521 event classes



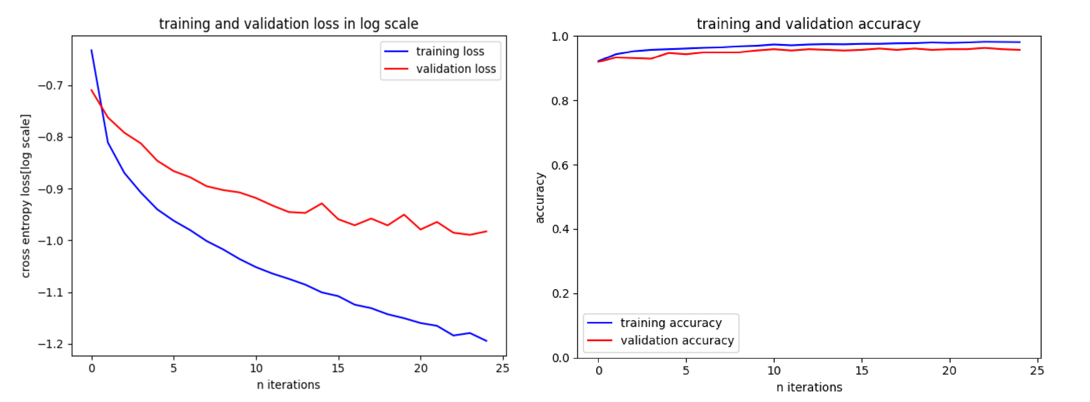
- **Fine tuning** in this project:
  - Extract the top FC layer.
  - Define the top layer- FC layer for binary classification.
    -- Input 1024 patches of 96X64.
  - SoftMax activation.
- **Training** the fully connected layer:
  - SGD optimizer.
  - Cross-entropy loss.
  - 1E-4 learning rate.
  - 1 batch size.
  - 64% training, 16% validation, 20% test.
  - 25 epochs.

Transfer Learning

- Common way to use pre-trained CNNs as an initialization for the task of interest.
- Fine-tuning:
  - CNN layers are fixed.
  - Replace the classifier on top of the CNN – top layers contain high level features.
  - Train top layer.



Results



validation accuracy	validation loss	Training accuracy	Training loss
0.9745	0.0795	0.9936	0.0249

In order to test the quality of the chosen solution, we compared it with different solutions:

- SVM with linear kernel
  - Features from our dataset.
- YAMnet without fine tuning
  - 9 classes for gunshot.
  - 512 classes for non-gunshot.

Num	Solution	Accuracy	Precision	Recall
1	YAMnet with fine tuning	0.9708	0.9576	0.9576
2	SVM with linear kernel	0.99	0.99	0.99
3	YAMnet without fine tuning	0.8333	0.9296	0.5593

Confusion matrix:

Num	Solution	TN	TP	FN	FP
1	YAMnet with fine tuning	217	117	1	7
2	SVM with linear kernel	222	117	1	2
3	YAMnet without fine tuning	219	66	52	5

Conclusions

- Our goal was to build a real time system for acoustic detection of gunshots in video games.
- We built a labeled Dataset.
  - most significant characteristics are correlation and spectrogram.
  - Least significant characteristic is energy.
- **We succeeded in building a deep-learning based system** that qualifies the project's requirements.

Future Work

- Expand the project's goal to localize gunshots.
- Expand the variety of gunshot types in video games.

October 2020