DALHOUSIE UNIVERSITY
FACULTY OF ENGINEERING

Ahmed Merdan
Finlay Miller
Ibrahim Fatungase

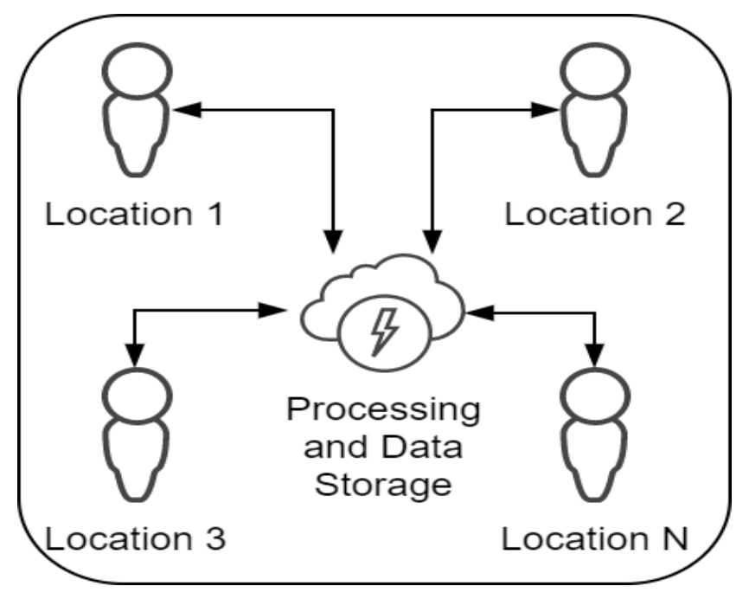
Sageev Oore

Department of Electrical and Computer Engineering

ML-Mediated Remote Audio Call Stations

Introduction

The ML-Mediated Remote audio call station project is an art installation that capitalizes on the use of machine learning algorithm to manipulate the playback of an audio input. The manipulated audio should create an echo chamber effect where less pronounced sounds in an audio input become amplified. On full scale, multiple stations, set up in various locations would achieve communication via the system.



Project Deliverables

The project being a prototype had the following requirements:

- A single call station as proof of concept.
- Efficient ML pipeline for processing.
- Communication between established components of the project. (Client, server and cloud).

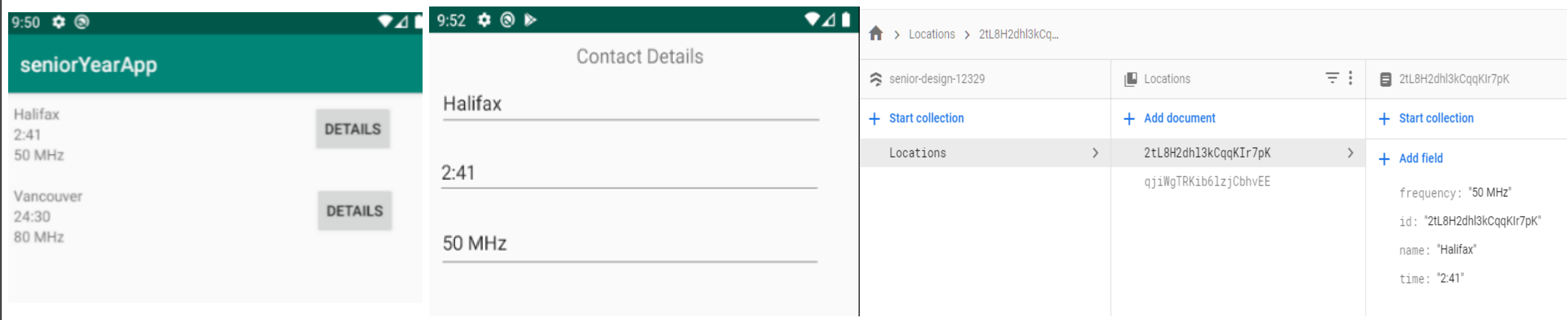
Design Process

The structure of each components of the project were design individually while considering that a merge would still be possible.

- **Client:**
  - Physical and software call stations to ensure user interaction.
  - Users can manipulate the playback via the call stations.
- **Server:**
  - Machine Learning Pipeline used for processing uses 2 pretrained models.
  - Classes obtained from ML pipeline are postprocessed to ensure similar sounds are grouped together using the K nearest neighbor algorithm.
- **Cloud:**
  - Storage of audio files are completed of a cloud database.
  - Central system where ML pipeline resides is retrieved using cloud functions.

Details of Design

- **Physical Station Requirements**
  - The components required for a physical box were very rudimentary. They include Raspberry Pi zero, soundcard, OTG cable and microphone. Written software saved on the Pi uploads audio to the cloud.
- **Android Application Development**
  - The second option for user interaction for the client component used an Android application to achieve functionality.

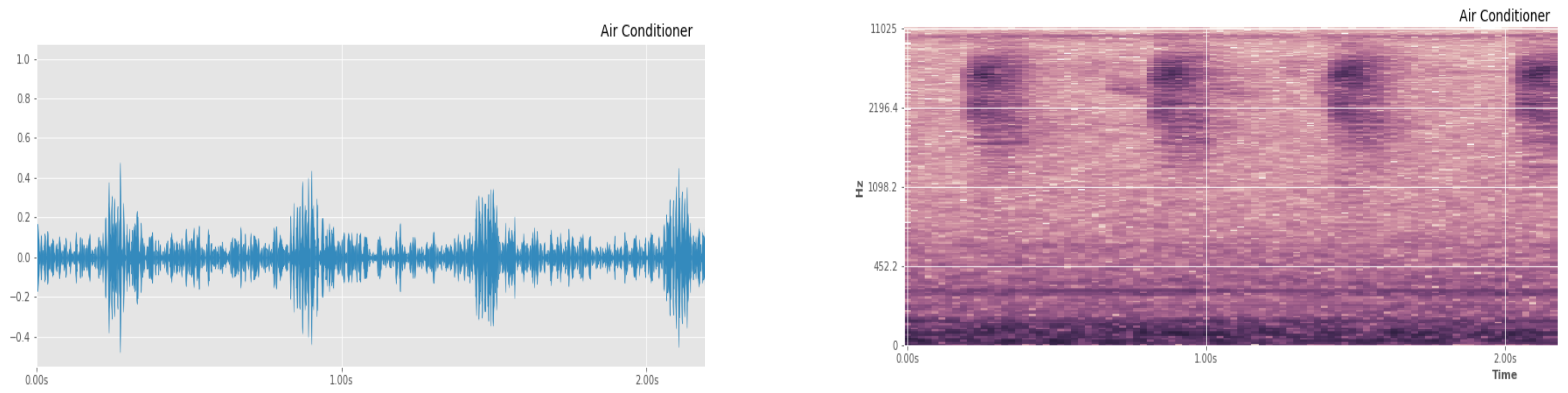

9:50
seniorYearApp
Halifax
2:41
50 MHz
DETAILS
Vancouver
24:30
80 MHz
DETAILS

9:52
Contact Details
Halifax
2:41
50 MHz

Locations > 2tL8H2dhl3kCq...
senior-design-12329
+ Start collection
Locations
Locations
+ Add document
2tL8H2dhl3kCqqKIr7pK
qjiWgTRKib6lzjCbhvEE
2tL8H2dhl3kCqqKIr7pK
+ Start collection
+ Add field
frequency: "50 MHz"
id: "2tL8H2dhl3kCqqKIr7pK"
name: "Halifax"
time: "2:41"

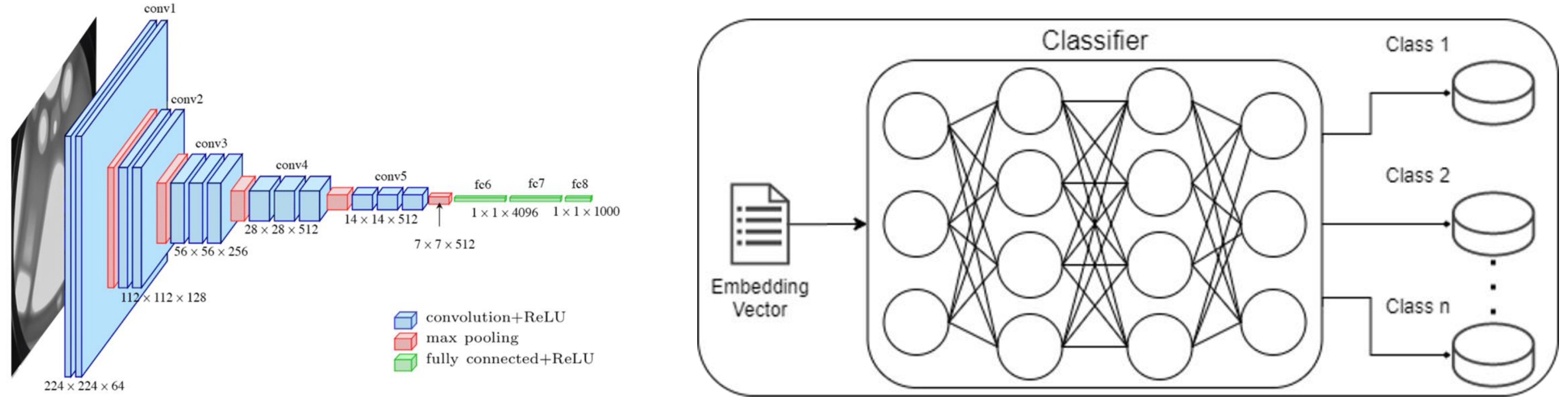
Android Application Activities and Firebase Data File Hierarchy

- **Server**
  - The pretrained classification models are designed for image and video recognition, therefore preprocessing must be achieved on input audio to convert to spectrograms.



Input and Output of Preprocessing Pipeline

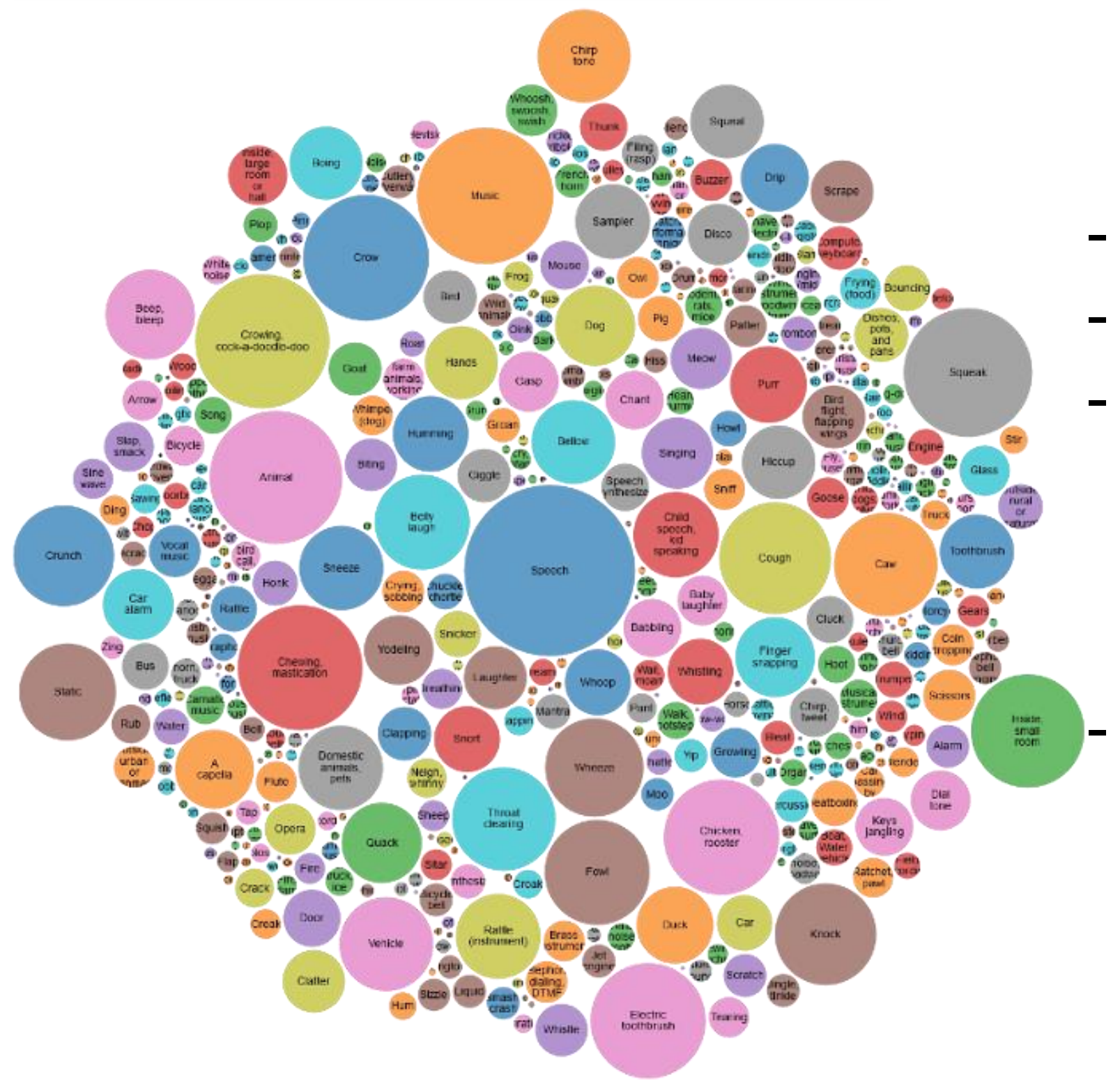
  - The Machine learning pipeline uses the VGGish pretrained audio classification model as well as the YouTube 8M dataset to build its embeddings and classes



VGGish[3] & YouTube 8M[1]

Results

Data Set Classifications



Urban 8K sounds classes[4]

Class Identification

- 512 different identified classes
- Real life classes (readable)
- Depend on the interpretation
  - The sounds of a "TV" can be many different classes
- Forced to use a large vector

Embedding Identification

- 128 floating point embeddings
- Based on the physical attributes of the signal
- Finds sounds the "sound" the same as obsessed to related
- Uses arbitrary vectors, i.e. Cant be understood by humans,
- Gives more accurate results.
- Less computationally demanding in real time systems

Freesound database classes[2]

Conclusion and Recommendations

Despite the project not being fully completed, the research done, and steps taken to ensure the eventual completion of the project were successful. The students were able to apply a pretrained model to identify sounds and find the KNN. A more user-friendly application experience is in order, including making it a standalone app.

References

[1]Abu-El-Haija, Sami, et al. "Youtube-8m: A large-scale video classification benchmark." *arXiv preprint arXiv:1609.08675* (2016).

[2]Fonseca, Eduardo, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, Xavier Serra. "General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline". *Proceedings of the DCASE 2018 Workshop* (2018).

[3]Hassan, Muneeb. "VGG16 – Convolutional Network for Classification and Detection" (2018).

[4]Salamon, J., C. Jacoby and J. P. Bello. "A Dataset and Taxonomy for Urban Sound Research." *22nd ACM International Conference on Multimedia, Orlando USA* (2014).