“NoiseVision”

Using Mask R-CNN to Detect Urban Noise Pollution

Krystal Won

01

What is Urban Noise?

“Unwanted or harmful sound in city environments that interferes with daily life, health, and well-being.”

American Public Health Association

Key Characteristics

- 🚗 Traffic
- 🚆 Transit (airplane, train)
- 🏗️ Construction
- 🏭 Industry
- 🚑 Emergency Sirens
- ❄️ HVAC Systems (Heating/Ventilation/Air Conditioning)
- 👥 Crowds

Why it Matters?

Public Health Impacts

- Elevated Stress & Heart Risk
- Sleep Disruption
- Vulnerable Populations

Urban Planning & Regulation

- Noise maps guide placement of barriers & speed controls
- Hotspot Identification
- Ongoing monitoring **vs**. expensive, episodic field surveys

About my dataset
Data cleaning
Data labeling

02

Data Preparation

ESC-50 Dataset "Environmental Sound Classification 50"

- 2,000 short audio recordings of real-world environments
- 5 seconds duration per clip
- 50 semantic sound classes
- 40 audio recordings of each class
- Designed for benchmarking environmental sound classification methods

5 Major Categories

Animal	Natural & water	Human non-speech	Interior/domestic
Dog	Rain	Crying baby	**Interior/domestic sounds**
Rooster	Sea waves	Sneezing	Door knock
Pig	Crackling fire	Clapping	Mouse click
Cow	Crickets	Breathing	Keyboard typing
Frog	Chirping birds	Coughing	Door, wood creaks
Cat	Water drops	Footsteps	Can opening
Hen	Wind	Laughing	Washing machine
Insects (flying)	Pouring water	Brushing teeth	Vacuum cleaner
Sheep	Toilet flush	Snoring	Clock alarm
Crow	Thunderstorm	Drinking, sipping	Clock tick
			Glass breaking

Exterior/ urban noises

Helicopter
Chainsaw
Siren
Car horn
Engine
Train
~~Church bells~~
Airplane
Fireworks
Hand saw

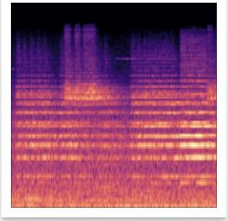
Audio → Mel Spectrogram (128 bands)

- Convert **raw audio recordings** into time–frequency "**images**" that capture transient events
- The Mel scale is a nonlinear transformation of linear Hertz (Hz) frequencies, designed to match human pitch perception
- preserve both **when** and **how loudly** each source appears.

1-24074-A-43.wav

Compress detail at high frequencies

Expands detail at low frequencies

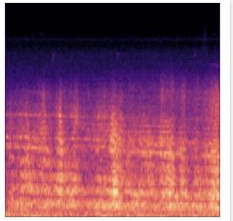


1-24074-A-43.png

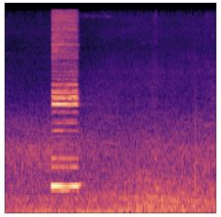
```
S = librosa.feature.melspectrogram(
      y=y,
      sr=sr,
      n_mels=n_mels,
      n_fft=n_fft,
      hop_length=hop_length
  )
```

```
S_db = librosa.power_to_db(
S, ref=np.max)
```

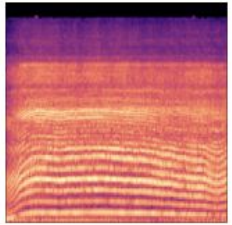
```
librosa.display.specshow(
      S_db,
      sr=sr,
      hop_length=hop_length,
      x_axis='time',
      y_axis='mel'
  )
```



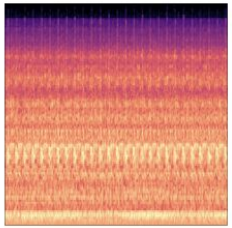
Airplane



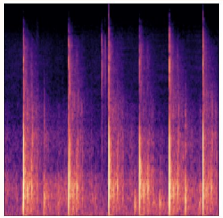
Car horn



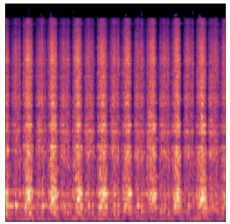
Chainsaw



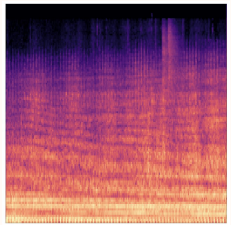
Engine



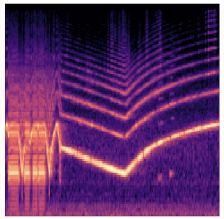
Fireworks



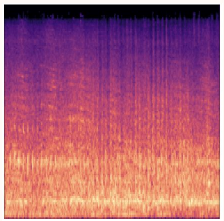
Handsaw



Helicopter



Siren



Train

03

Modeling

Backbone inside the Mask R-CNN model

ResNet-50 with Feature Pyramid Network (FPN), pretrained on COCO

Replace box & mask heads for

- All the sound labels + background
- Box predictor (FastRCNNPredictor)
- Mask predictor (MaskRCNNPredictor)

Training Strategy

- **Annotating 9 Class with *Labelme* polygons**
  - airplane, car-horn, chainsaw, engine, fireworks, hand-saw, helicopter, siren, train
- **Augmentation**
  - IoU-based random crops, color jitter, flips
- Resize + pad → 512×512 → normalize & sanitize boxes
- Training on 288 samples, validating on 72 samples

Model Performance

Strong Generalization

Validation loss fell from 0.66 → 0.11 then stabilized, indicating minimal over-/under-fitting.

High Detection Confidence

Mask R-CNN outputs bounding boxes with an average confidence of 90.27%.

Accurate Localization

Target bbox: `[1, 20, 309, 306]`

Predicted bbox: `[0, 23, 310, 305]` → **IoU ≈ 0.99**

Next steps

Find more urban noise data

Visualize training/validation loss

Geo-Tagged Alerts

Heatmap overlay showing event density (“noisevision”) in Philly

Wrap the preprocessing + model into an API

ingests a live audio stream, converts it to Mel spectrograms on the fly, runs inference, and returns detected events.

Thank You!