SURROUND SOUND

Real-time environmental & sound-event recognition from microphone audio

Everett-Alan Hood

OBJECTIVES

Classify a user's environment in real time

This work is heavily inspired by the **DCASE** (Detection and Classification of Acoustic Scenes and Events) Challenge Community. While their work separates classification into distinct categories, this extension attempts to combine them into a **cohesive scene analyzer**.

Detect multiple audible situations simultaneously

The idea is to use **multiple models** that specialize in **different aspects** of a scenario. Then, summarize the scene as a whole using a large language model. This project separates **environment**, **speech**, and **events**. *Speech was dropped due to complexity.*

Summarize the current auditory scenario

Just for a more understandable interpretation of the current scenario, **GPT 4.1 Mini** was used to quickly summarize the scenario based on the models' findings. **This is not counted in the final evaluation**, it's simply a quality of life enhancement.

DATASET & PREPROCESSING

Environment

Google AudioSet

2M+ annotated video clips via YouTube

- Downloaded 11,207 WAV Samples
- ~2.5K Balanced, Rest Unbalanced
- 16kHz mono, log-mel spectrograms
- normalized, max 10s clips
- extracted labels related to eight different categories
- animal, crowd, indoor, other_ambient, outdoor_nature, vehicle_ground, vehicle_water, water

```
ALLOWED_LABELS = {
    # Scenes / locations
    "Outside, rural or natural", "Outside, urban or manmade", "Crowd",
    "Inside, small room", "Inside, large room or hall", "Inside, public space",
    "Field recording",

    # Weather / geophony
    "Wind", "Rain", "Rain on surface", "Thunder", "Thunderstorm",
    "Waves, surf", "Ocean", "Stream", "Waterfall", "Raindrop", "Rustling leaves",
    "Water",
```

Events

FSD50K

51K annotated audio clips via Freesound

- 200 Classes based on AudioSet
- 16kHz mono, log-mel spectrograms
- normalized, trimmed from 30s clips to 10s
- condensed to 62 labels based on relative abundance and uniqueness to other labels
- some labels had far more samples than others

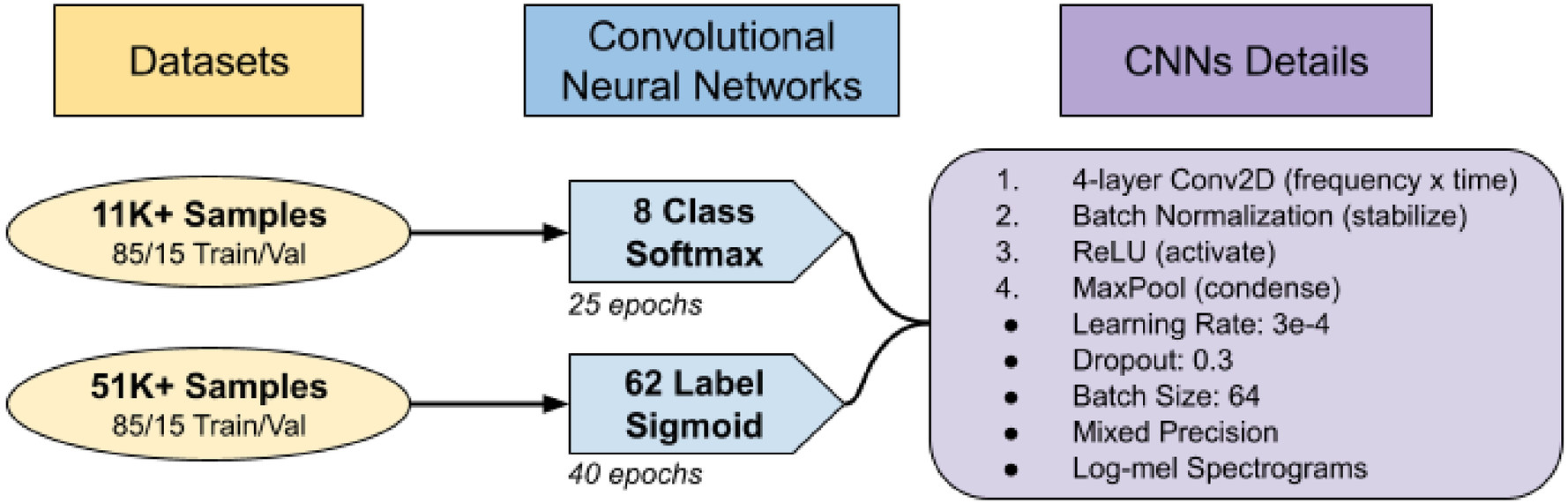
```
"Glass Clink": 23,
"Guitar Instruments": 24,
"Human": 25,
"Human Breathing": 26,
"Human Cry Laugh": 27,
"Human Group": 28,
"Impact": 29,
"Insect Buzz": 30,
"Keyboard Instruments": 31,
"Kitchen Appliance": 32,
```

```
"Tools": 52,
"Train": 53,
"Truck Bus": 54,
"Typing Writing": 55,
"Vehicle": 56,
"Water Running": 57,
"Weather": 58,
"Wild Animals": 59,
"Wind": 60,
"Wind Instruments": 61
```

METHODOLOGY

Here’s a quick overview about how these models are trained and how the live implementation works.

Models



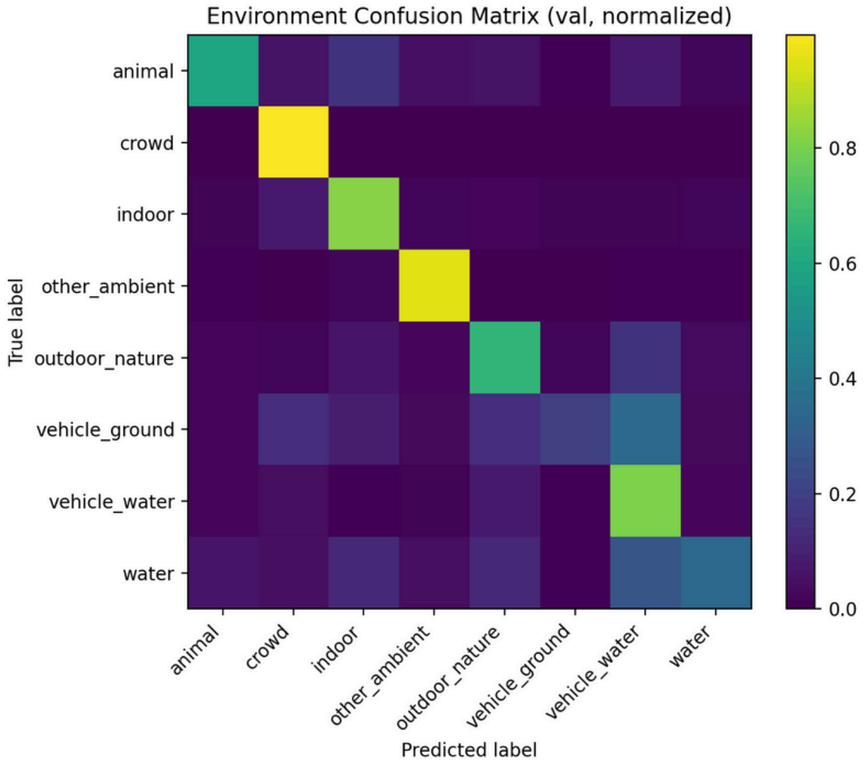
Demonstration

LIVE DEMO?

With your help, let’s see if it can predict the current environment!

RESULTS - ENVIRONMENT

Average	P	R	F1
Macro	0.5404	0.6376	0.4985
Weighted	0.6818	0.4666	0.4466



BEST CLASSES

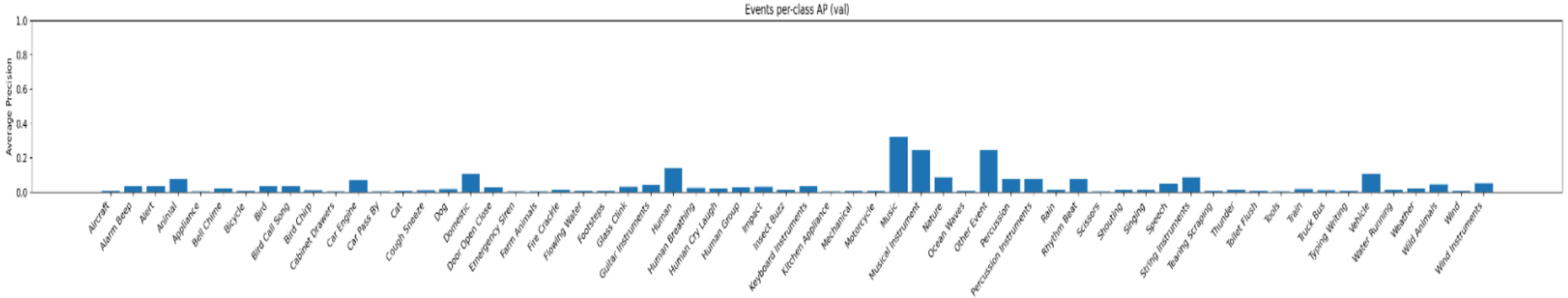
Class	# Samples	P	R	F1
vehicle_ground	1911	0.9191	0.1842	0.3069
water	1118	0.7795	0.3605	0.4930
animal	705	0.7271	0.5631	0.6347

WORST CLASSES

Class	# Samples	P	R	F1
outdoor_nature	365	0.2853	0.5753	0.3815
vehicle_water	533	0.2848	0.7917	0.4189
crowd	292	0.4033	0.9932	0.5737

RESULTS - EVENTS

Average	P	R	F1
Macro	0.33	0.62	0.39
Weighted	0.58	0.58	0.54



BEST LABELS

Class	# Samples	P	R	F1
Music	2512	0.92	0.80	0.86
Vehicle	616	0.64	0.50	0.56
Animal	570	0.61	0.46	0.53

WORST LABELS

Class	# Samples	P	R	F1
Emergency Siren	20	0.08	0.35	0.13
Wind	59	0.08	0.53	0.14
Scissors	21	0.09	0.52	0.15

THANK YOU!

ANY QUESTIONS?

SOURCES

DCASE Challenge Community
https://dcase.community/
Google AudioSet
https://research.google.com/audioset/
FSD50K Dataset
https://zenodo.org/records/4060432

Libraries & Frameworks

Librosa
PyTorch
Streamlit
OpenAI GPT-4.1 Mini