# YOHO model for Audio Segmentation and Sound Event Detection

Davide Capone [SM3500601] Enrico Stefanel [SM3500554]
{davide.capone, enrico.stefanel}@studenti.units.it

Data Science and Scientific Computing Master's Course
Department of Mathematics and Geosciences
University of Trieste

A.Y. 2023–2024

# Contents

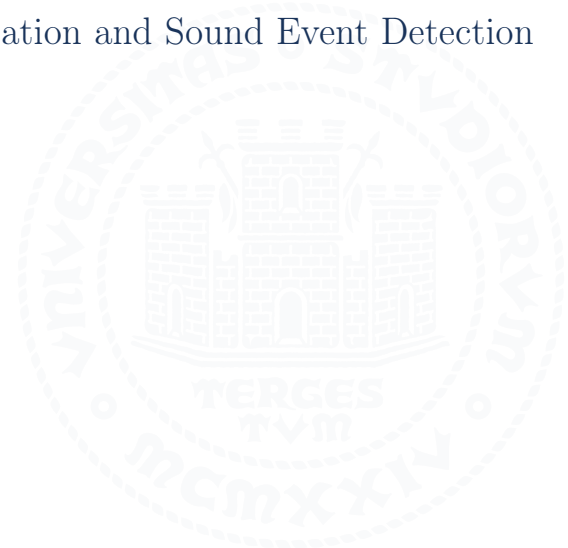
Introduction

YOHO model

Implementation details

Conclusions

# Audio Segmentation and Sound Event Detection



...

# Datasets

Common datasets for Audio Segmentation and Sound Event Detection problems are:

- **TUT Sound Event Detection**: primarily consists of street recordings with traffic and other activity, with audio examples of 2.56 s and a total size of approximately 1.5 h. It has six unique audio classes – Brakes Squeaking, Car, Children, Large Vehicle, People Speaking, and People Walking;
- **Urban-SED**: purely synthetic dataset, with audio example of 10 s and a total size of about 30 h. It has ten unique audio classes – Air Conditioner, Car Horn, Children Playing, Dog Bark, Drilling, Engine Idling, Gun Shot, Jackhammer, Siren, and Street Music.

An example of Urban-SED label is:

```
[(gun_shot, 0.3, 1.11), (car_horn, 0.31, 1.41)]
```

meaning that an occurrence of gun shot is present from the 0.3 s to 1.11 s, and a car horn from 0.31 s to 1.41 s.

# Metrics

A popular toolbox for Polyphonic Sound Event Detection models evaluation is **SED Eval**[1].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 \text{ score} = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

[1] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. "Metrics for Polyphonic Sound Event Detection". In: *Applied Sciences* 6.6 (2016). ISSN: 2076-3417. DOI: 10.3390/app6060162. URL: https://www.mdpi.com/2076-3417/6/6/162.

# YOHO model

Presented in 2021[2]...

[2] Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. "You Only Hear Once: A YOLO-like Algorithm for Audio Segmentation and Sound Event Detection". In: *Applied Sciences* 12.7 (Mar. 2022), p. 3293. ISSN: 2076-3417. DOI: 10.3390/app12073293. URL: http://dx.doi.org/10.3390/app12073293.

# Input shape

...

# Network Architecture

...

# Output shape

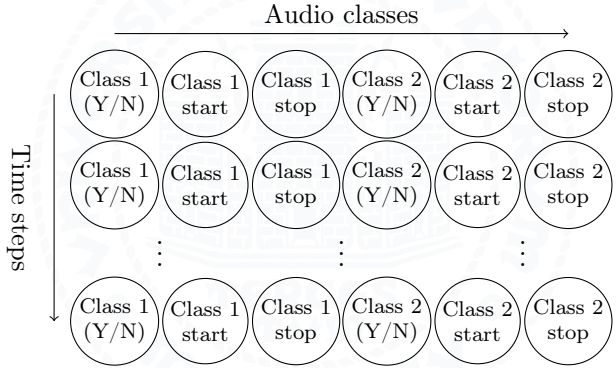



Figure: The YOHO output shape.

# Loss Function

$$\mathcal{L}_c(\hat{y}, y) = \begin{cases} (\hat{y}_1 - y_1)^2 + \\ (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 & \text{if } y_1 = 1 \\ (\hat{y}_1 - y_1)^2, & \text{if } y_1 = 0 \end{cases}$$

where $y$ and $\hat{y}$ are the ground-truth and predictions respectively. $y_1 = 1$ if the acoustic class is present and $y_1 = 0$ if the class is absent. $y_2$ and $y_3$, which are the start and endpoints for each acoustic class are considered only if $y = 1$. In other words, $(\hat{y}_1 - y_1)^2$ corresponds to **the classification loss** and $(\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2$ corresponds to **the regression loss**.

# Other Details

...

# Implementation details

...

# Problems



...

# Conclusions

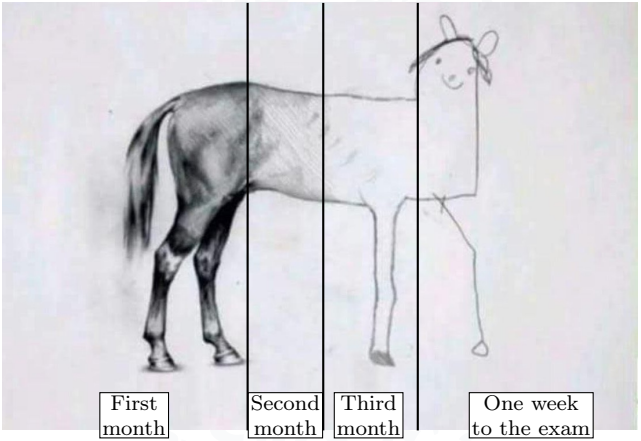



Figure: The roadmap of our journey.

# Conclusions

But, after all...

*It's all about the journey, not the destination.*

Thank you for your attention.