On-device Subject Recognition in UWB-radar data with Tiny Machine Learning

Massimo Pavan, Armando Caltabiano and Manuel Roveri
{massimo.pavan, manuel.roveri}@polimi.it
armando.caltabiano@truesense.it

POLITECNICO
MILANO 1863

Introduction

Tiny Machine Learning (**TinyML**) is a novel research area aiming at designing machine and deep learning algorithms able to be executed on tiny devices.

Smart pervasive devices are rapidly becoming omnipresent in our every-day life[2], and the design of lightweight and reliable algorithms is now crucial.

UltrawideBand (UWB) is a radar technology that is emerging as an alternative to common sensors, particularly suitable for **privacy-preserving** embedded devices due to its precision and low energy consumption.

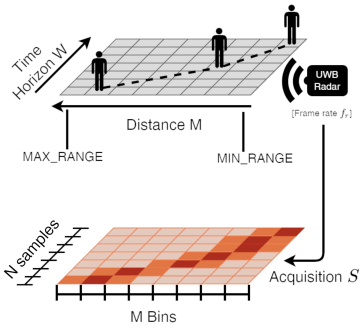
We propose a TinyML solution that uses a tiny convolutional neural network (CNN) for **subject recognition** through the analysis of UWB-radar data. It was tested on a real-world in-car application.

Background

Most of TinyML related literature focuses on the **approximation of CNNs**.

The three most used techniques for achieving this goal are:

- Quantization [3]
- Pruning [4]
- Knowledge distillation



The researches on UWB-radar data concentrate on tasks of person detection or human activity recognition and, currently, **none of them works on tiny devices** except our previous work on presence detection[1].

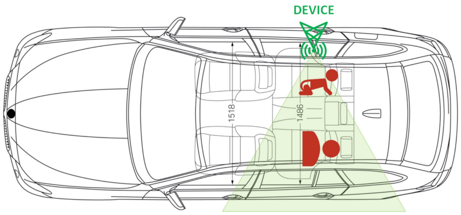
Dataset & Goal

The dataset is composed of 429 20-second-long radar scans of a car:

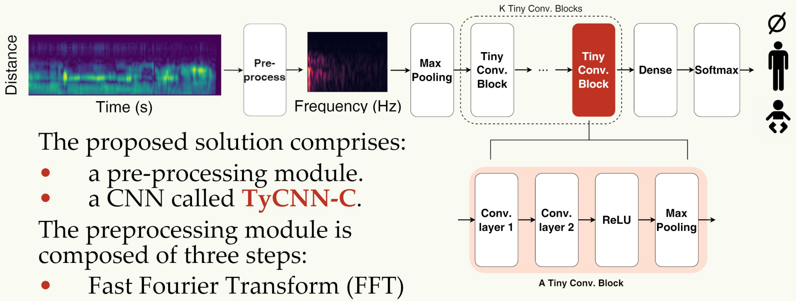
- 163 with a **child** in the first seat;
- 46 records with an **adult**;
- 220 with the seat **empty;**

0, 1, 2 or all 3 seats can be occupied.

The goal is to **classify the target in the first seat**.



Proposed Solution: TyCNN-C



The proposed solution comprises:

- a pre-processing module.
- a CNN called **TyCNN-C**.

The preprocessing module is composed of three steps:

- Fast Fourier Transform (FFT)
- Frequency selection
- Normalization

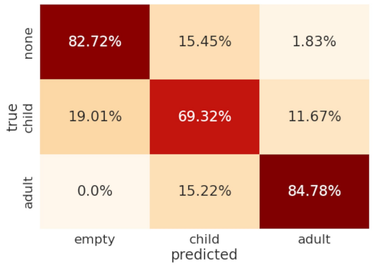
The design of TyCNN-C extends the one of the TyCNN used in [1]. **Quantization** is used for further optimizing the execution.

Results

The considered tiny device is based on an ESP32 MCU, has a RAM memory limit of 100 KB, and should execute the algorithm in < 1 s.

The proposed solution:

- Achieves an accuracy of **0.783 ± 0.076**.
- Matches the constrains, requiring
  - **- 47.8 KB** of RAM memory;
  - **-** A total of **8.57e6** operations;
  - **-** An execution time of **940 ms**, 230 ms for preprocessing data and 710 ms for inference.



Memory occupations (B)

Input	4 558
Weights	17 629
Peak Activations	31 304
Total	48 933

Number of operations

Total	8 571 216

Literature

- [1] M. Pavan, A. Caltabiano, and M. Roveri, "Tinyml for uwb-radar based presence detection," Proceedings of WCCI 2022, IEEE, Jul. 2022.
- [2] C. Alippi, Intelligence for embedded systems. Springer, 2014.
- [3]A. Gholami et al. "A survey of quantization methods for efficient neural network inference", 2021.
- [4] J. Liu et al. "Pruning algorithms to accelerate convolutional neural networks for edge applications: A survey," 2020

Acknowledgements

Special thanks to Ing. P. Lento, Dr. A. Bassi and Ing. G. Viscardi for the support in the work.