MERCon 2025
MORATUWA ENGINEERING RESEARCH CONFERENCE

TinyML: A Compact Revolution in Engineering AI

Samitha Somathilaka[1], Dinuka Sahabandu[2], Asiri Gawesha Lindamulage[4*], Sanka Mohottala[3], Mahima Weerasinghe[3]

1) University of Nebraska-Lincoln

2) University of Washington

3) Sri Lanka Institute of Information Technology

4) Open University of Sri Lanka

aglin@ou.ac.lk

Organizing Committee

Dr. Dharshana Kasturirathna
SLIIT

Dr. Dinuka Sahabandu
University of Washington

Dr. Mahima Weerasinghe
SLIIT

Mr. Asiri Gawesha Lindamulage
Open University of Sri Lanka

Mr. Sanka Mohottala
SLIIT

Mr. Chethiya Galkaduwa
Indiana University

Ms. Savini Kommalage
SLIIT

Organized by BrAINLabs Research Group, SLIIT
Funded by SLIIT Research & International (Grant No. PVC(R&I)/RG/2025/12)

Resource Personal

Dr. Samitha Somathilaka
School of Computing
University of Nebraska-Lincoln

Dr. Dinuka Sahabandu
Department of Electrical and Computer Engineering
University of Washington

Mr. Asiri Gawesha Lindamulage
Department of Electrical and Computer Engineering
Open University of Sri Lanka

Mr. Sanka Mohottala
Department of Electrical and Electronic Engineering
Sri Lanka Institute of Information Technology

Dr. Mahima Weerasinghe
Department of Computer Science
Sri Lanka Institute of Information Technology

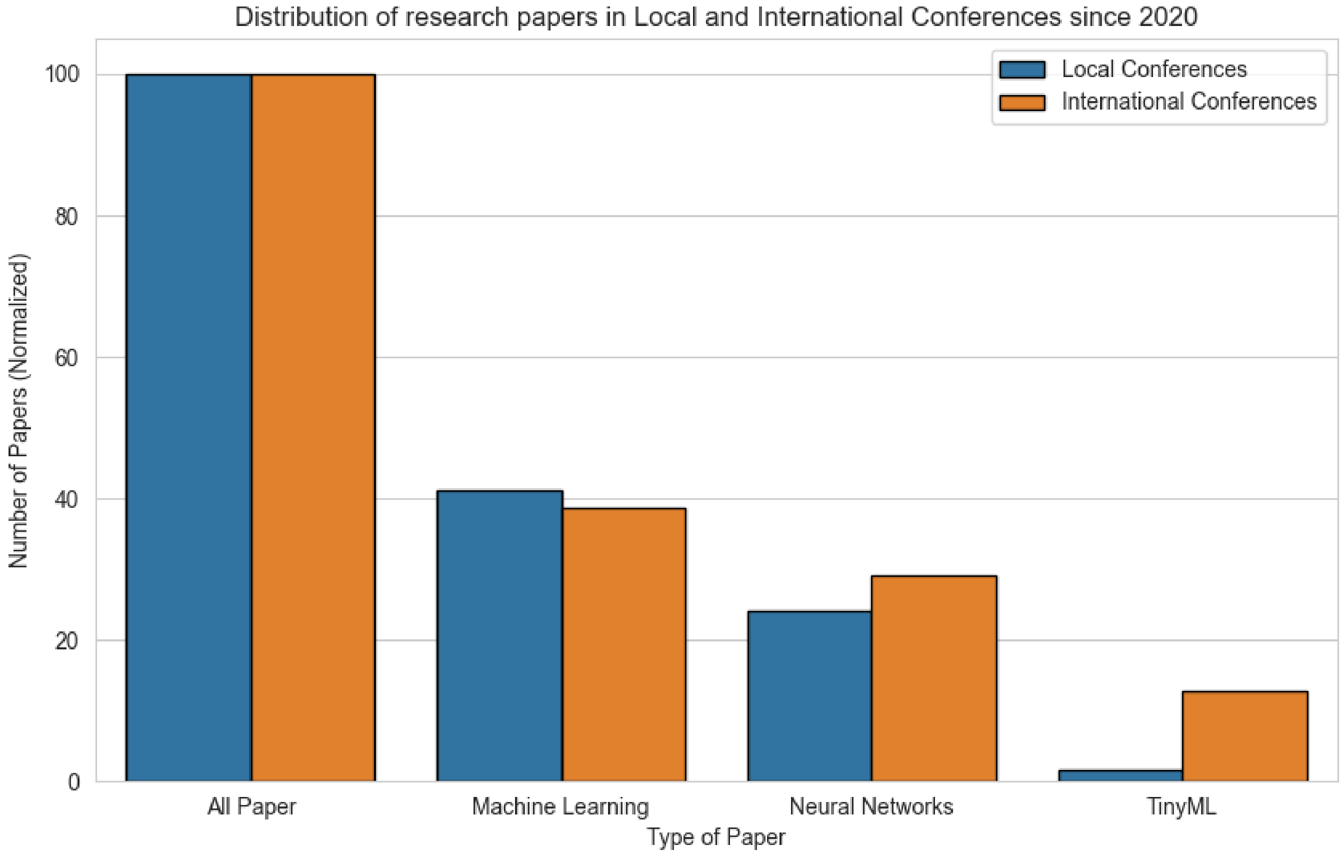
Motivation behind the workshop

In Sri Lanka awareness on TinyML is very lower than other areas.

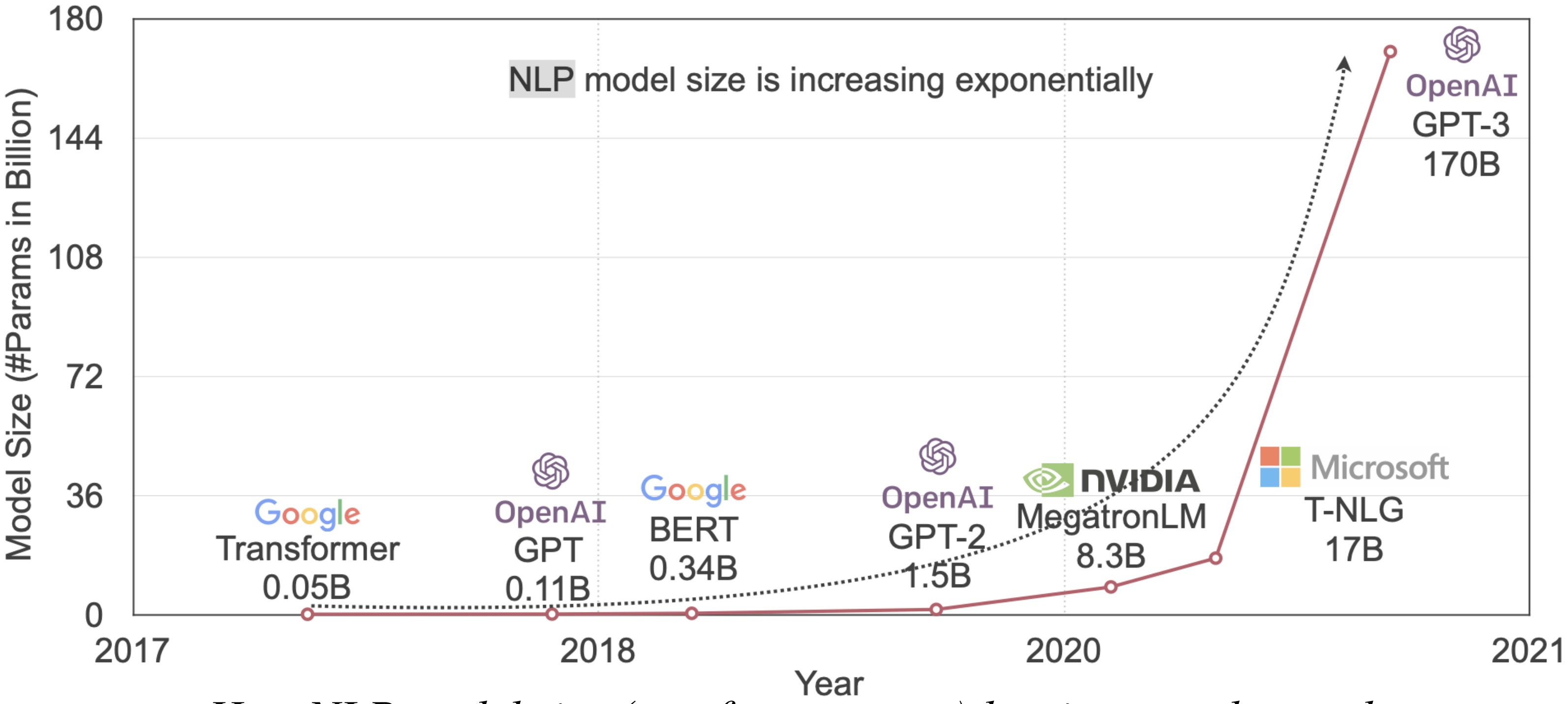
TinyML offers a novel, experimental research direction.

The workshop will equip participants to balance accuracy, latency, and memory usage, optimizing models for real-world deployment.

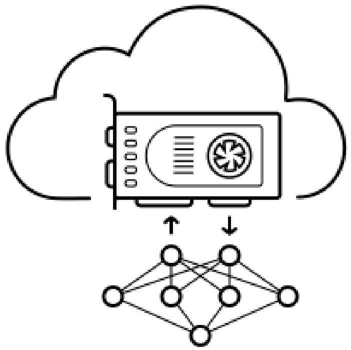
Workshop consists of theoretical sessions and coding sessions



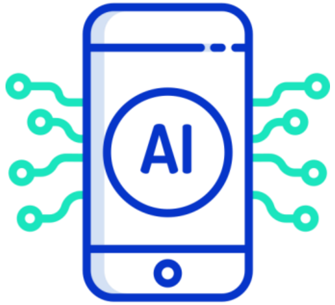
What is TinyML and why is it important



How NLP model size (no of parameters) has increased over the years



On Cloud



Edge Computing

Tiny ML

- Billions of IoT devices around the world based on **microcontrollers** – CCTV, Smart wearable devices
- **Low-power**: green AI, reduce carbon

Challenge - Small memory (Around 1 MB), lower computational power

TinyML Landscape

Converting large deep learning models into compact models for heavily resource constrained devices

- Reduce the number of parameters
- Lower the peak memory usage
- Reduce power consumption

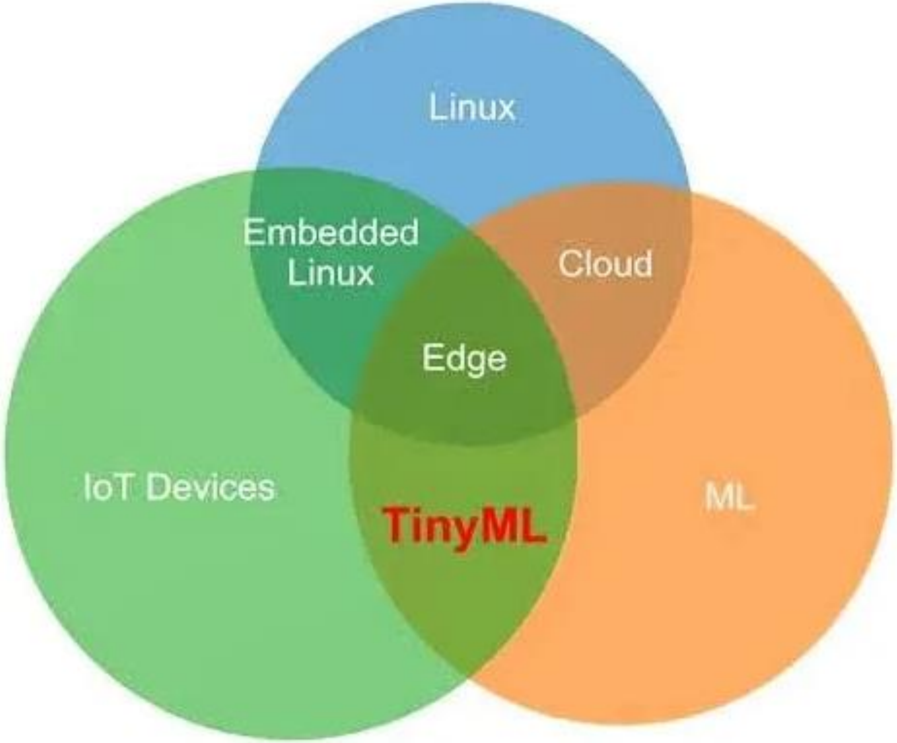
What are the techniques?

Model compression

- Pruning
- Quantization
- Weight clustering
- Knowledge Distillation

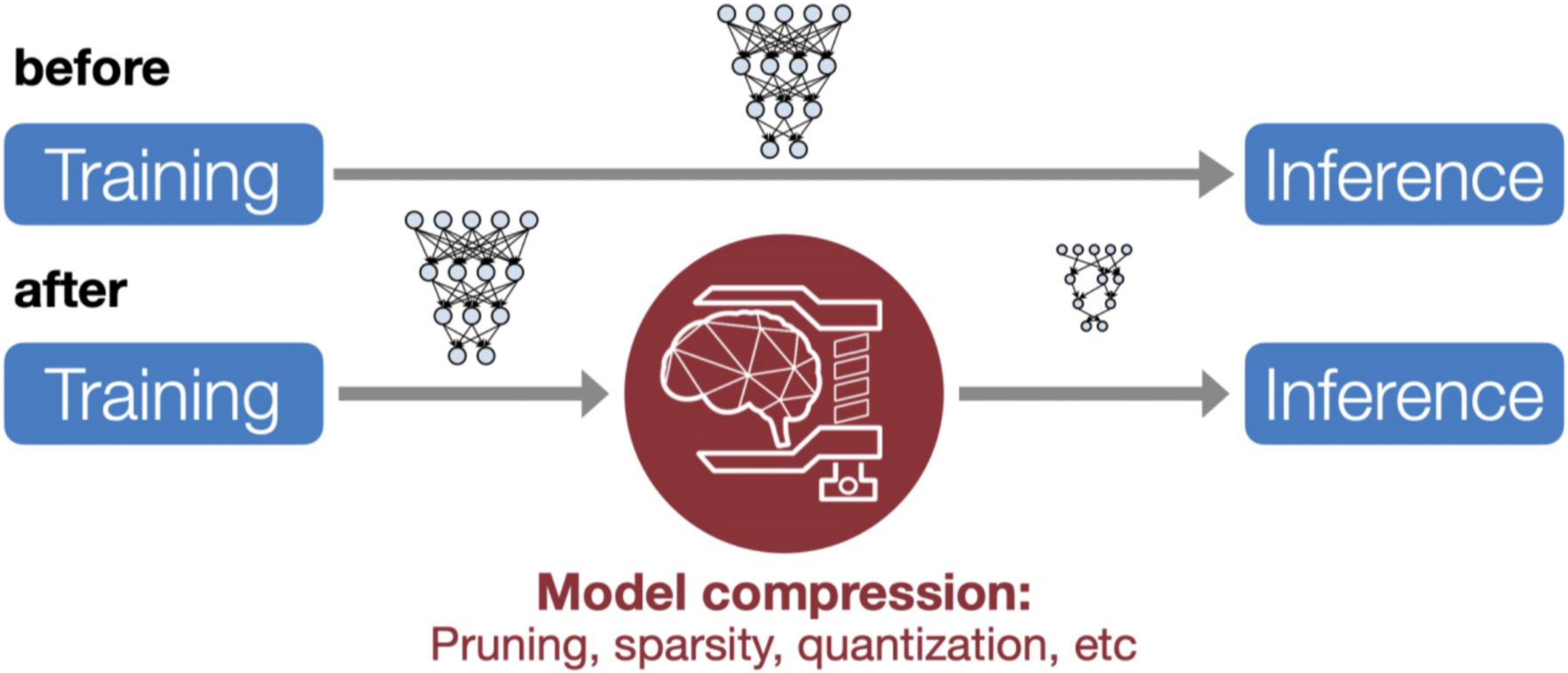
Tiny models

- MobileNet, EfficientNet
- Bio-inspired architectures
- Paramter efficient architecures

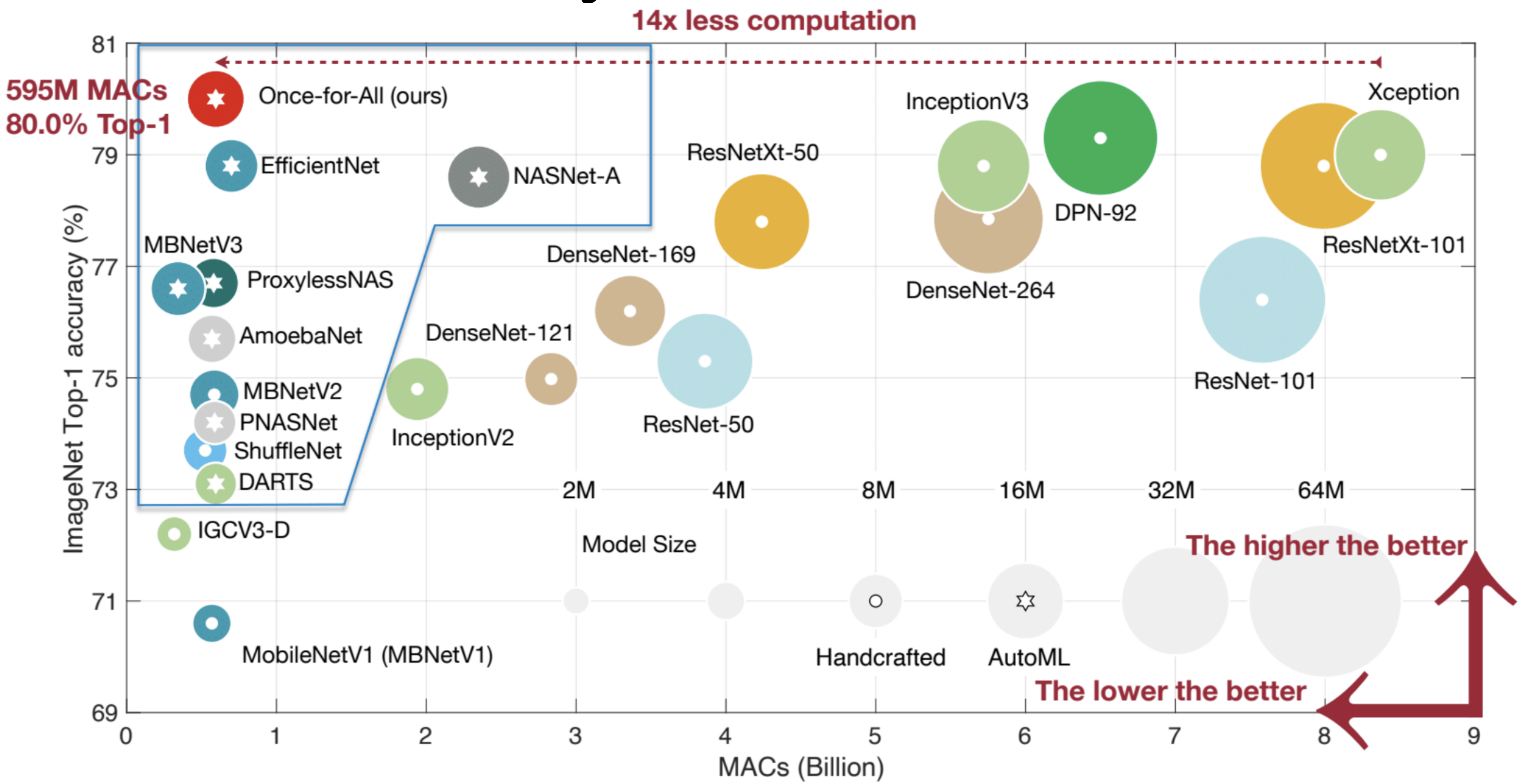


Model Compression

- Bridges the Gap between the Supply and Demand of AI Computing



Tiny Architectures



Once-for-All: Train One Network and Specialize it for Efficient Deployment [Cai *et al.*, ICLR 2020]

MACs- Multiply–Accumulate operations

NAS has resulted a highly accurate model with less latency and and energy usage
MobileNet uses – Separable Convolution and Point-wise Convolution

What you are going to learn

- Theoretical aspects
  - Model Compression
  - Bio-inspired TinyML architectures
  - Energy efficient and parameter efficient models
- Hands on sessions
  - End-to end model deployment on RaspberryPi with model compression
  - Model deployment on Arduino Nano BLE 33
- Technology stack
  - Languages – Python, Arduino
  - Frameworks- TensorFlow, TFLite, TfLiteMicro

All notebooks and slides will be shared through the website

By the end you will have an idea on..

- How model compression can be utilized effectively
- Model deployment process and pipeline
- Theoretical knowledge on Bio Inspired energy efficient architectures
- An example for Tiny architectures for head pose estimation
- How to develop a model, optimize it for Tinyml and deploy it

Lineup of the workshop