Session 2a: Edge AI Tools, devices, and methods

With the advent of Chatbots, LLMs and other generative IA technologies, as well as other progresses in the IA field, there is an explosion of the demand for compute force. IA is no longer computer science; it is computational science. As such, it can no longer be done with casual, self-managed equipment. More advanced compute infrastructures are required both to satisfy user needs (in terms of compute power, GPU Ram capacity) and to ensure a decent utilization of the increasingly costly resources.
In an increasingly connected world, the ability to process data at the source—closer to where it is generated—has become crucial. This workshop offers a comprehensive yet concise overview of Edge AI, focusing on the devices, methods, and tools that make this technology a game-changer. Participants will explore how Edge AI enables real-time data processing, reduces latency, and enhances privacy by minimizing the need to send data to the cloud.

Content and topics

The workshop is designed to introduce key concepts without overwhelming detail, making it ideal for professionals, students, and technology enthusiasts looking to understand the essentials of Edge AI. We will cover a range of topics, including types of computing devices used in Edge AI, and how they operate at the edge of the network. Attendees will also learn about the methods and algorithms optimized for edge computing, such as lightweight neural networks and model compression techniques.
In addition, the workshop will introduce some tools and frameworks that facilitate the development and deployment of Edge AI solutions. Real-world case studies will demonstrate how these technologies are applied in various scenarios.
By the end of the workshop, participants will have a broad understanding of the fundamental aspects of Edge AI, empowering them to explore further and implement these technologies in their own projects.

Organization and structure:

Structure

The workshop is organized in two main parts. During the morning, we will present different approaches for Edge AI: several devices including microcontrollers, embedded GPUs, FPGAs, and NPUs. We will also present methodologies for Edge oriented applications including tools for quantization, benchmarking and deployment on single and multiple devices.
During the afternoon we propose two hands-on tutorials targeting the deployment of a model on two different types of devices.

Needs :

Personnal laptop with internet connexion.

Speaker and committee

Last Name	First Name	Institution	e-mail address
Upegui	Andres	hepia	andres.upegui@hesge.ch
Pazos Escudero	Nuria	HE-Arc	Nuria.PazosEscudero@he-arc.ch
Zapater	Marina	HEIG-VD	marina.zapater@heig-vd.ch
Calvaresi	Davide	HE-VS	davide.calvaresi@hevs.ch
Gantel	Laurent	Hepia	laurent.gantel@hesge.ch
Berthet	Quentin	Hepia	quentin.berthet@hesge.ch

AI-DAYS@HES-SO 2025 – GENEVA & LAUSANNE – JANUARY 27- JANUARY 29, 2025

WORKSHOP DAY JANUARY 27

Schedule: 8h30-12h+ 13h-15h

	Topics	**Speakers**
8h30-8h40	Edge AI: an introduction	Andres Upegui - Hepia
8h40-9h10	Low power and low latency: How to deploy a tiny model on a microcontroller and make it fly	Marina Zapater – HEIG-VD
9h10-9h40	When existing architectures are not enough: Custom ML Hardware architectures on FPGAs	Quentin Berthet – Hepia
9h40-10h10	How to select the perfect device? Benchmarking embedding edge devices (CPUs, GPGPUs, MCUs, NPUs) remotely	Nuria Pazos - HE-ARC
10h10-10h30	Accelerating computation with neural processors. Hailo-8 : An AI co-processor for machine learning inference	Laurent Gantel – Hepia
10h30-10h45	break	
10h45-12h00	Hands-on tutorial: Deploying an inference model on a Hailo-8 Neural co-processor	Laurent Gantel – Hepia
12h00-13h00	Lunch break	
13h00-13h30	Deploying an ecosystem of edge devices	Davide Calvaresi – HEVS
13h30-15h00	Hands-on tutorial: Deploying an Edge service on an ecosystem on embedded devices	Davide Calvaresi – HEVS