

| Degree programme                    | Type | Course |
|-------------------------------------|------|--------|
| Bachelor in Artificial Intelligence | OP   | 3      |

## Contact lecturer

Name : Vanessa Moreno Font

Email : [vanessa.moreno@uab.cat](mailto:vanessa.moreno@uab.cat)

## Teaching staff

Juan Carazo Borrego

Jordi Carrabina Bordoll

## Group languages

You can consult this information at the [end](#) of the document.

## Prerequisites

There is none. This course partly describes the hardware of AI accelerators that are in server chips, mobiles, embedded chips, etc. Therefore, it is necessary to have the basic concepts of computer architecture and technology.

## Objectives

This course aims to analyze the methodologies and platforms that allow the acceleration of AI computing.

This acceleration is associated with different factors such as: (1) the type of operations that are executed (vector-matrix and matrix-matrix multiplication with accumulation, and complex transfer functions); (2) data management (both in terms of memory and input-output requirements); (3) the requirements of the systems where AI must be embedded (real-time conditions, limitation of energy consumption, etc.)

As for the scope of this acceleration, although both the learning and inference phases are accelerated, and since learning is carried out on servers in the cloud, we will focus mostly on platforms with limited resources (compared to servers) such as mobile or embedded platforms (also known as edge).

The different general purpose (CPU, GPU, FPGA) and specific (DPU/TPU/NPU, ML and NN processors, bionic, neuromorphic, etc.) computational platforms will be analyzed along with the deployment methodologies.

All this in the field of the Internet of Things (IoT) made up of systems that include devices, the edge and the cloud.

## Learning outcomes

- CM19 (Design IoT applications in embedded systems with limited resources in different domains.) Design IoT applications in embedded systems with limited resources in different domains.
- KM35 (Describe the components and architecture of cyber-physical systems, including embedded system platforms, sensors, actuators, and real-time control devices.) Describe the components and architecture of cyber-physical systems, including embedded system platforms, sensors, actuators, and real-time control devices.
- KM37 (Identify the principles behind edge-computing and Internet of Things (IOT) protocols.) Identify the principles behind edge-computing and Internet of Things (IOT) protocols.
- SM45 (Adapt AI algorithms to implement inference on embedded platforms with limited resources and real-time, energy-efficient conditions.) Adapt AI algorithms to implement inference on embedded platforms with limited resources and real-time, energy-efficient conditions.
- SM46 (Analyse the real-time performance of AI applications in order to optimise learning and inference in general- and purpose-built processors.) Analyse the real-time performance of AI applications in order to optimise learning and inference in general- and purpose-built processors.

## Contents

### CONTENTS

#### 1. Introduction to IoT Platforms for AI

- Cloud, Edge (mobile, embedded), Device (resource-constrained)
- Training vs. Inference workload balance

#### 2. AI Optimization

- Deployment methodologies to reduce computational complexity
- AI vs. Computational performance (Application requirements): accuracy, real-time, memory, energy
- Tiny ML

#### 3. Acceleration techniques and technologies

- General-purpose platforms: CPU, GPU, FPGA
- Application-specific platforms for ML and NN processing: DPU/TPU/NPU
- Advanced chips: neuromorphic, memristor, bionic and quantum

## LABS

Deployment of applications to (1) mobile devices (from students) and (2) embedded platforms

## DESIGN PROJECT

Plan & prototype of an Edge AI specific application (selected by students).

## Learning activities and methodology

| Title                         | Hours | ECTS | Learning outcomes |
|-------------------------------|-------|------|-------------------|
| Laboratories & Design Project | 24    | 0.96 |                   |
| Study & Homework              | 98    | 3.92 |                   |
| Master classes and seminars   | 26    | 1.04 |                   |

The learning methodology will combine master classes, activities in tutored sessions, project-based learning, and laboratory sessions.

Attendance will be mandatory for the IoT-IA design project and laboratory sessions that will be done in groups of 2 or 3 people.

The laboratory sessions will use a guided format.

This course will use UAB's virtual campus at <https://cv.uab.cat>.

Annotation: within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

## Assessment

### Continuous assessment activities

| Title                                                                 | Weight | Hours | ECTS | Learning outcomes            |
|-----------------------------------------------------------------------|--------|-------|------|------------------------------|
| Report and defence of the design project                              | 40%    | 2     | 0.08 | CM19, KM35, KM37, SM45, SM46 |
| Evaluation of activities developed in tutored sessions (laboratories) | 30%    | 0     | 0    | CM19, KM35, KM37, SM45, SM46 |
| Individual activities (i.e. exercises)                                | 30%    | 0     | 0    | SM45, SM46                   |

This course does not provide for the single assessment system (no exam).

The evaluation of the course will follow the rules of the continuous evaluation and the final grade for the course, is calculated in the following way:

A - 30% from the mark obtained by the student through the evaluation of activities (i.e. exercises). When an evaluation activity is scheduled, the evaluation indicators will be reported and its weight in this qualification.

B - 40% from the mark obtained through the evaluation of the IoT-AI design project.

C - 30% from the mark obtained by the student of the laboratory work and reports. It is necessary to exceed 5 (out of 10) in this item to pass the subject.

All activities will require delivering report through the virtual campus:

- Type A activities will be proposed along the course around lectures.
- Type B activities, will require delivering partial reports every 2 weeks.
- Type C activities, will require the submission of a report for each laboratory session.

To obtain MH it will be necessary that the students have an overall qualification higher than 9 with the

limitations of the UAB (1MH/20students). As a reference criterion, they will be assigned in descending order.

To pass the course, students must obtain a final grade of at least 5.0, a minimum grade of 5.0 in activity C, and a grade higher than 3.3 in activities A and B. If the course is not passed because any of these conditions is not met, the numerical grade recorded on the student's academic transcript will be the lower of 4.0 and the final grade.

Plagiarism will not be tolerated. All students involved in a plagiarism activity will be failed automatically. A final mark no higher than 30% will be assigned.

Open source code or available libraries can be used but they must be referred in the corresponding reports.

An student not having achieved a sufficient final weighted average mark, may opt to apply for remedial activities (individual work or additional synthesis examination) the subject under the following conditions:

- the student must have participated in the laboratory activities and design project, and
- the student must have a final weighted average higher than 30%, and
- the student must not have failed any activity due to plagiarism.

The student will receive a grade of "Not Evaluable" if:

- the student has not been able to be evaluated in the laboratory activities due to not attendance or not deliver the corresponding reports without justified cause.
- the student has not carried out a minimum of 50% of the activities proposed.
- the student has not done the design project.

For each assessment activity, the student or the group will be given the corresponding comments. Students can make complaints about the grade of the activity, which will be evaluated by the teaching staff responsible for the subject.

Repeating students will be able to "save" their grade in laboratory activity.

## Bibliography

Russell, S. J., & Norvig, P. (2022). Artificial intelligence: a modern approach (Global edition). Pearson Education Limited.

Du, L., Du, Y. (2018). Hardware Accelerator Design for Machine Learning. In Machine Learning - Advanced Techniques and Emerging Applications. IntechOpen. <https://doi.org/10.5772/intechopen.72845>

Huawei Technologies Co., L. (2022). Artificial Intelligence Technology (1st ed. 2023.). Springer Nature. <https://doi.org/10.1007/978-981-19-2879-6>

X. Ma et al., "A Survey on Deep Learning Empowered IoT Applications," in IEEE Access, vol. 7, pp. 181721-181732, 2019, doi: 10.1109/ACCESS.2019.2958962

V. H. Kim and K. K. Choi, "A Reconfigurable CNN-Based Accelerator Design for Fast and Energy-Efficient Object Detection System on Mobile FPGA," in IEEE Access, vol. 11, pp. 59438-59445, 2023, doi: 10.1109/ACCESS.2023.3285279

C. -B. Wu, C. -S. Wang and Y. -K. Hsiao, "Reconfigurable Hardware Architecture Design and Implementation for AI Deep Learning Accelerator," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), Kobe, Japan, 2020, pp. 154-155, doi: 10.1109/GCCE50665.2020.9291854

Robert David et al. TENSORFLOW LITE MICRO: EMBEDDED MACHINE LEARNING ON TINYML SYSTEMS. Proceedings of the 4th MLSys Conference, San Jose, CA, USA.

Pete Warden, Daniel Situnayake. TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. <https://tinymlbook.com/>

Mishra, A., Cha, J., Park, H., & Kim, S. (2023). Artificial Intelligence and Hardware Accelerators (1st ed.). Springer International Publishing AG. <https://doi.org/10.1007/978-3-031-22170-5>

Liu, A. C.-C., & Law, O. M. K. (2021). Artificial intelligence hardware design: challenges and solutions. John Wiley & Sons, Incorporated.

Daniel Situnayake, Jenny Plunkett. (2023). AI at the Edge. O'Reilly Media, Inc

## Software

We plan to use different tools/toolchains:

- Tensor RT for embedded NVIDIA GPUs
- OpenVino from Intel
- Edge Impulse multiplatform
- The TinyML environment: LiteRT (adapting the design flow to the platform)

## Course groups and languages

The information provided is provisional until November 30. After this date, you will be able to consult the language of each group through this [link](#). To access the information, you will need to enter the course CODE

| Type of teaching           | Group | Language | Semester       | Shift     |
|----------------------------|-------|----------|----------------|-----------|
| (TE) Theory                | 71    | English  | first semester | afternoon |
| (PAUL) Classroom practices | 711   | English  | first semester | afternoon |