UAB
Universitat Autònoma
de Barcelona

Computing Acceleration in AI

Code: 106592
ECTS Credits: 6

2025/2026

Degree	Type	Year
Artificial Intelligence	OT	3
Artificial Intelligence	OT	4

Contact

Name: Vanessa Moreno Font

Email: vanessa.moreno@uab.cat

Teachers

Jordi Carrabina Bordoll

Juan Carazo Borrego

Teaching groups languages

You can view this information at the end of this 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 and Contextualisation

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.

Competences

Artificial Intelligence

- Act within the field of knowledge by evaluating the social, economic and environmental impact beforehand.
- Analyse and solve problems effectively, generating innovative and creative proposals to achieve objectives.
- Conceive, design, analyse and implement autonomous cyber-physical agents and systems capable of interacting with other agents and/or people in open environments, taking into account collective demands and needs.
- Conceptualize and model alternatives of complex solutions to problems of application of artificial intelligence in different fields and create prototypes that demonstrate the validity of the proposed system.
- Identify, analyse and evaluate the ethical and social impact, the human and cultural context, and the legal implications of the development of artificial intelligence and data manipulation applications in different fields.
- Work cooperatively to achieve common objectives, assuming own responsibility and respecting the role of the different members of the team.

Learning Outcomes

1. Adapt AI algorithms to implement inference in embedded platforms with limited resources and real-time and energy-efficient conditions.
2. Analyse and solve problems effectively, generating innovative and creative proposals to achieve objectives.
3. Analyse the sustainability indicators of academic and professional activities in the field by incorporating the social, economic and environmental factors at play.
4. Design and validate the methodology for implementing learning and inference in general- and specific-purpose processors.
5. Design, prototype and evaluate the performance of embedded systems in resource-constrained, real-time and energy-efficient conditions.
6. Identify the best solutions for mapping an AI solution onto a distributed IoT system and device, both edge and cloud.
7. Identify the ethical and social impact and the legal and regulatory implications of AI systems for sending training data to the cloud.
8. Measure and optimise the performance of AI algorithm implementations on platforms.
9. Propose viable projects and actions that enhance social, economic and environmental benefits.
10. Use AI network learning acceleration technologies and services in the cloud and in peripheries.
11. Work cooperatively to achieve common objectives, assuming own responsibility and respecting the role of the different members of the team.

Content

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).

Activities and Methodology

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Master classes and seminars	26	1.04	1, 3, 5, 4, 6, 7, 8, 10
Type: Supervised			
Laboratories & Design Project	24	0.96	1, 3, 2, 5, 4, 6, 7, 8, 9, 11, 10
Type: Autonomous			
Study & Homework	98	3.92	1, 3, 2, 5, 4, 6, 7, 8, 9, 10

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

Continous Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Evaluation of activities developed in tutored sessions (laboratories)	20%	0	0	1, 2, 5, 4, 6, 8, 11, 10
Individual activities (i.e. exercices)	40%	0	0	1, 2, 5, 4, 6, 8, 10
Report and defence of the design project	40%	2	0.08	1, 3, 2, 5, 4, 6, 7, 8, 9, 11, 10

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 - 40% 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 - 20% 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.

A final weighted average mark not lower than 50% is sufficient to pass the course, provided that a score over one third of the range is attained in every one of the Marks for items A and B. If not reached, the mark will be 4.0.

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: Tensorflow Lite (adapting the design flow to the platform)

Groups and Languages

Please note that this information is provisional until 30 November 2025. You can check it through this link. To consult the language you will need to enter the CODE of the subject.

Name	Group	Language	Semester	Turn
(PAUL) Classroom practices	711	English	first semester	morning-mixed
(TE) Theory	71	English	first semester	morning-mixed