UAB
Universitat Autònoma de Barcelona

# Aceleración de la Computación en IA

Código: 106592
Créditos ECTS: 6

2024/2025

| Titulación | Tipo | Curso |
|---|---|---|
| 2504392 Inteligencia Artificial / Artificial Intelligence | OT | 3 |
| 2504392 Inteligencia Artificial / Artificial Intelligence | OT | 4 |

## Contacto

Nombre: Vanessa Moreno Font

Correo electrónico: vanessa.moreno@uab.cat

## Equipo docente

Jordi Carrabina Bordoll

## Idiomas de los grupos

Puede consultar esta información al final del documento.

## Prerrequisitos

No hay. En parte de esta asignatura se describe el hardware de los aceleradores de IA que hay en los chips de servidores, móviles, empotrados, etc. por lo tanto hay que tener los conceptos básicos de arquitectura y tecnología de ordenadores.

## Objetivos y contextualización

Esta asignatura tiene como objetivo analizar las plataformas que permiten la aceleración de la computación de la IA.

Dicha aceleración está asociada a diferentes factores como son: (1) el tipo de operaciones que se ejecutan (multiplicaciones vector-matriz y matriz-matriz con acumulación, y funciones de transferencia complejas); (2) La gestión de los datos (tanto en cuanto a los requerimientos de memoria como de entrada salida); (3) los requerimientos de los sistemas donde debe ir incrustada la IA (condiciones de tiempo real, limitación de consumo de energía, etc.)

En cuanto al ámbito de esta aceleración, aunque se acelera tanto la fase de aprendizaje como la de inferencia y dado que el aprendizaje se realiza en servidores, nos centraremos mayoritariamente en plataformas con recursos limitados (respecto de los servidores) como plataformas móviles o empotradas. Se analizarán las diferentes plataformas computacionales de propósito general (CPU, GPU, FPGA) y específico (DPU/TPU/NPU, procesadores ML y NN, chips biónicos, neuromórficos, basados en memristores y cuánticos) junto con las metodologías de despliegue.

Todo ello en el ámbito del internet de los objetos (IoT) compuesto por sistemas que incluyen los dispositivos (devices), la periferia (edge) y la nube (cloud).

## Competencias

Inteligencia Artificial / Artificial Intelligence

- Actuar en el ámbito de conocimiento propio valorando el impacto social, económico y medioambiental.
- Analizar y resolver problemas de forma efectiva, generando propuestas innovadoras y creativas para alcanzar los objetivos.
- Concebir, diseñar, analizar e implementar agentes y sistemas ciber-físicos autónomos capaces de interactuar con otros agentes y/o personas en entornos abiertos, teniendo en cuenta las demandas y necesidades colectivas.
- Conceptualizar y modelar alternativas de soluciones complejas a problemas de aplicación de la inteligencia artificial en diferentes ámbitos, y planificar y gestionar proyectos para el diseño y desarrollo de prototipos que demuestren la validez del sistema propuesto.
- Identificar, analizar y evaluar el impacto ético y social, el contexto humano y cultural, y las implicaciones legales del desarrollo de aplicaciones de inteligencia artificial y de manipulación de datos en diferentes ámbitos.
- Trabajar cooperativamente para la consecución de objetivos comunes, asumiendo la propia responsabilidad y respetando el rol de los diferentes miembros del equipo.

## Resultados de aprendizaje

1. Adaptar algoritmos de IA para implementar la inferencia en plataformas empotradas con recursos limitados y condiciones de tiempo real y eficiencia energética.
2. Analizar los indicadores de sostenibilidad de las actividades académico-profesionales del ámbito integrando las dimensiones social, económica y medioambiental.
3. Analizar y resolver problemas de forma efectiva, generando propuestas innovadoras y creativas para alcanzar los objetivos.
4. Diseñar y validar la metodología de implementación de aprendizaje e inferencia en procesadores de propósito general y específico.
5. Diseñar, crear prototipos y evaluar prestaciones en sistemas embebidos con recursos limitados y condiciones de tiempo real y eficiencia energética.
6. Identificar el impacto ético y social y las implicaciones legales y regulatorias de los sistemas AI para el envío de datos para entrenamiento a la nube.
7. Identificar las mejores soluciones para mapear una solución IA en un sistema IoT distribuido e device, Edge y cloud.
8. Medir y optimizar las prestaciones de las implementaciones de algoritmos IA en plataformas.
9. Proponer proyectos y acciones viables que potencien los beneficios sociales, económicos y medioambientales.
10. Trabajar cooperativamente para la consecución de objetivos comunes, asumiendo la propia responsabilidad y respetando el rol de los diferentes miembros del equipo.
11. Utilizar las tecnologías y servicios de aceleración de aprendizaje de redes IA en la nube y en la periferia.

## Contenido

CONTENIDO

1. Plataformas IoT para IA

- Nube
- Periferia (móvil, incrustado)

- Dispositivo (con restricciones de recursos)

2. Metodologías de desarrollo y requisitos de aplicación

- Entrenamiento
- Inferencia: tiempo real, memoria, energía

3. Análisis de la complejidad computacional de la computación de la IA

- Tipos de operaciones
- Aritmética de operaciones
- Gestión de datos
- Técnicas para reducir la complejidad computacional

4. Técnicas y tecnologías de aceleración

- Plataformas de propósito general: CPU, GPU, FPGA
- Plataformas específicas de la aplicación para el procesamiento de ML y NN: DPU/TPU/NPU
- Chips avanzados: neuromórficos, basados en memristor, biónicos y cuánticos

LABORATORIOS

Despliegue de una aplicación a (1) un dispositivo móvil (de estudiantes) y (2) plataforma empotrada

## Actividades formativas y Metodología

| Título | Horas | ECTS | Resultados de aprendizaje |
|---|---|---|---|
| Tipo: Dirigidas | | | |
| Clases magistrales y seminarios | 26 | 1,04 | 1, 2, 4, 5, 7, 6, 8, 11 |
| Tipo: Supervisadas | | | |
| Laboratorios y Proyecto de Diseño | 24 | 0,96 | 1, 3, 2, 4, 5, 7, 6, 8, 9, 11, 10 |
| Tipo: Autónomas | | | |
| Estudio y trabajo fuera del aula | 98 | 3,92 | 1, 3, 2, 4, 5, 7, 6, 8, 9, 11 |

La metodología de aprendizaje combinará: clases magistrales, actividades en sesiones tutorizadas, casos de uso y aprendizaje basado en proyectos; debates y otras actividades colaborativas; y sesiones de laboratorio.

La asistencia es obligatoria para las actividades: proyecto de diseño IoT-IA y prácticas de laboratorio, que se realizarán en grupos de 2 o 3 personas.

Las sesiones de laboratorio utilizaran un formato guiado.

Este curso se utilizará el campus virtual de la UAB a https://cv.uab.cat.

Nota: se reservarán 15 minutos de una clase dentro del calendario establecido por el centro o por la titulación para que el alumnado rellene las encuestas de evaluación de la actuación del profesorado y de evaluación de la asignatura o módulo.

Nota: se reservarán 15 minutos de una clase dentro del calendario establecido por el centro o por la titulación para que el alumnado rellene las encuestas de evaluación de la actuación del profesorado y de evaluación de la asignatura o módulo.

## Evaluación

### Actividades de evaluación continuada

| Título | Peso | Horas | ECTS | Resultados de aprendizaje |
|---|---|---|---|---|
| Actividades individuales (tipo ejercicios) | 20% | 0 | 0 | 1, 3, 4, 5, 7, 8, 11 |
| Evaluación de actividades desarrolladas en sesiones tutorizadas (laboratorios) | 40% | 0 | 0 | 1, 3, 4, 5, 7, 8, 11, 10 |
| Informe y presentación del proyecto de diseño | 40% | 2 | 0,08 | 1, 3, 2, 4, 5, 7, 6, 8, 9, 11, 10 |

Esta asignatura no prevé el sistema de evaluación única (no hay examen)

La evaluación de los alumnes utilizará el modelo de evaluación continua y la nota final del curso se calcula de la siguiente manera:
A - 20% de la nota obtenida por la evaluación de las actividades propuestas (tipo ejercicios). Cuando se programa una actividad de evaluación, indicará qué indicadores se utilizarán para evaluar y su peso en la calificación.
B - 40% de la marca obtenida por la evaluación del trabajo de diseño de un sistema IoT-AI (original).
C - 40% de la nota obtenida por el estudiante de los trabajos de laboratorio. Es necesaria una calificación superior a 5 (sobre 10) en este ítem para aprobar la asignatura.

Todas las actividades requerirán la entrega de informe a través del campus virtual:
- A lo largo del curso se propondrán actividades de tipo A para los diferentes temas.
- Las actividades de tipo B, requerirán la entrega de informes parciales del proyecto cada 2 semanas.
- Las actividades tipo C, requerirán la entrega de dos informes parciales (uno a mitad de semestre y un 2º al final).

Para obtener MH será necesario que los alumnos tengan una cualificación global superior a 9 con las limitaciones de la UAB (1MH para cada 20alumnos). Como criterio de referencia, se asignan por orden descendente.

Una nota final ponderada no inferior al 50% es suficiente para superar el curso, siempre que se alcance una puntuación superior a un tercio del rango siempre quese alcance una puntuación superior a un tercio del rango en los 3 primeros ítems (A y B). Si es inferior se asignará una nota de 4.0.

No se tolerará el plagio. Todos los estudiantes implicados en una actividad de plagio serán suspendidos automáticamente. Se asignará una nota final no superior al 30%.

Se puede utilizar SW de código abierto o librerías disponibles, pero deben referenciarse en los informes correspondientes.

Un estudiante que no haya conseguido una nota media ponderada suficiente puede optar por solicitar actividades de recuperación (trabajos individuales o prueba de síntesis adicional) de la asignatura en las

siguientes condiciones:
- el estudiante debe haber participado en las actividades de laboratorio y el proyecto de diseño,
- el estudiante debe tener un promedio ponderado final superior al 30%, y
- el estudiante no debe haber fallado en ninguna actividad por culpa del plagio.

El estudiante recibirá una nota de "No Evaluable" en caso de que:
- el estudiante no haya podido ser evaluado en las actividades de laboratorio por no haber asistido o no haber entregado los correspondientes informes sin causa justificada.
- el estudiante no haya realizar un mínimo del 50% de las actividades propuestas.
- el estudiante no haya realizado el trabajo de diseño.

## Bibliografía

[1] Russell, S., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach.
[2] Li Du and Yuan Du. Hardware Accelerator Design for Machine Learning. http://dx.doi.org/10.5772/intechopen.72845
[3] Huawei Technologies Co. Artificial Intelligence Tecnology, Ltd.ISBN 978-981-19-2879-6
[4] XIAOQIANG MA et al. A Survey on Deep Learning Empowered IoT Applications. Digital Object Identifier 10.1109/ACCESS.2019.2958962
[5] A Reconfigurable CNN-Based Accelerator Design for Fast and Energy-Efficient Object Detection System on Mobile FPGA
[6] 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.
[7] Robert David et al. TENSORFLOW LITE MICRO: EMBEDDED MACHINE LEARNING ON TINYML SYSTEMS. Proceedings of the 4 th MLSys Conference, San Jose, CA, USA,
[8] Pete Warden, Daniel Situnayake "TinyML: Machine Learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers" https://tinymlbook.com/

## Software

Se utilizarán las herramientas del entorno TinyML: Tensorflow Lite (adecuando el flujo de diseño a la plataforma)
Se analizará la opción de utilizar las herramientas de Qualcomm para despliegue a aceleradores NN de móviles.

## Lista de idiomas

| Nombre | Grupo | Idioma | Semestre | Turno |
|---|---|---|---|---|
| (PAUL) Prácticas de aula | 1 | Inglés | primer cuatrimestre | tarde |
| (TE) Teoría | 1 | Inglés | primer cuatrimestre | tarde |