UAB
Universitat Autònoma de Barcelona

Acceleració de la Computació en IA

Codi: 106592
Crèdits: 6

2024/2025

Titulació	Tipus	Curs
2504392 Intel·ligència Artificial / Artificial Intelligence	OT	3
2504392 Intel·ligència Artificial / Artificial Intelligence	OT	4

Professor/a de contacte

Nom: Vanessa Moreno Font

Correu electrònic: vanessa.moreno@uab.cat

Equip docent

Jordi Carrabina Bordoll

Idiomes dels grups

Podeu consultar aquesta informació al final del document.

Prerequisits

No n'hi ha. En part d'aquesta assignatura es descriu el maquinari dels acceleradors d'IA que hi ha als xips de servidors, mòbils, encastats, etc. per tant cal tenir els conceptes bàsics d'arquitectura i tecnologia d'ordinadors.

Objectius

Aquesta assignatura te com a objectiu analitzar les plataformes que permeten l'acceleració de la computació de la IA.

Aquesta acceleració està associada a diferents factors com són: (1) el tipus d'operacions que s'executen (multiplicacions vector-matriu i matriu-matriu amb acumulació, i funcions de transferència complexes); (2) La gestió de les dades (tant pel que fa als requeriments de memòria com d'entrada sortida); (3) els requeriments dels sistemes on ha d'anar incrustada la IA (condicions de temps real, limitació de consum d'energia, etc.)

Pel que fa a l'àmbit d'aquesta acceleració, tot i que s'accelera tant la fase d'aprenentatge com la d'inferència i donat que l'aprenentatge es realitza en servidors, ens centrarem majoritàriament en plataformes amb recursos limitats (respecte dels servidors) com ara plataformes mòbils o encastades.

S'analitzaran les diferents plataformes computacionals de propòsit general (CPU, GPU, FPGA) i específic (DPU/TPU/NPU, processadors ML i NN, chips biònics, neuromòrfics, basats en memristors i quàntics) juntament amb les metodologies de desplegament.

Tot això en l'àmbit de l'internet dels objectes (IoT) composat per sistemes que inclouen els dispositius (devices), la perifèria (edge) i el núvol (cloud).

Competències

Intel·ligència Artificial / Artificial Intelligence

- Actuar en làmbit de coneixement propi valorant limpacte social, econòmic i mediambiental.
- Analitzar i resoldre problemes de manera efectiva, i generar propostes innovadores i creatives per aconseguir els objectius.
- Concebre, dissenyar, analitzar i implementar agents i sistemes ciberfísics autònoms capaços dinteractuar amb altres agents o persones en entorns oberts, tenint en compte les demandes i necessitats col·lectives.
- Conceptualitzar i modelar alternatives de solucions complexes per a problemes daplicació de la intel·ligència artificial en diferents àmbits, i planificar i gestionar projectes per al disseny i desenvolupament de prototips que demostrin la validesa del sistema proposat.
- Identificar, analitzar i avaluar limpacte ètic i social, el context humà i cultural i les implicacions legals del desenvolupament daplicacions dintel·ligència artificial i de manipulació de dades en diferents àmbits.
- Treballar cooperativament per aconseguir objectius comuns, assumint la pròpia responsabilitat i respectant el rol dels diferents membres de lequip.

Resultats d'aprenentatge

1. Adaptar algoritmes dIA per implementar la inferència en plataformes encastades amb recursos limitats i condicions de temps real i eficiència energètica.
2. Analitzar els indicadors de sostenibilitat de les activitats academicoprofessionals de làmbit integrant les dimensions social, econòmica i mediambiental.
3. Analitzar i resoldre problemes de manera efectiva, i generar propostes innovadores i creatives per aconseguir els objectius.
4. Dissenyar i validar la metodologia dimplementació daprenentatge i inferència en processadors de propòsit general i específic.
5. Dissenyar, crear prototips i avaluar prestacions en sistemes encastats amb recursos limitats i condicions de temps real i eficiència energètica.
6. Identificar les millors solucions per mapar una solució dIA en un sistema dIdC distribuït en dispositiu, perifèria i núvol.
7. Identificar limpacte ètic i social i les implicacions legals i regulatòries dels sistemes dIA per a lenviament de dades per a lentrenament al núvol.
8. Mesurar i optimitzar les prestacions de les implementacions dalgoritmes dIA en plataformes.
9. Proposar projectes i accions viables que potenciïn els beneficis socials, econòmics i mediambientals.
10. Treballar cooperativament per aconseguir objectius comuns, assumint la pròpia responsabilitat i respectant el rol dels diferents membres de lequip.
11. Utilitzar les tecnologies i serveis dacceleració daprenentatge de xarxes dIA en el núvol i en la perifèria.

Continguts

CONTINGUT

1. Plataformes IoT per a IA

- Núvol
- Perifèria (mòbil, incrustat)
- Dispositiu (amb restriccions de recursos)

2. Metodologies de desplegament i requeriments d'aplicació

- Entrenament
- Inferència: temps real, memòria, energia

3. Anàlisi de la complexitat computacional de la computació de la IA

- Tipus d'operacions
- Aritmètica d'operacions
- Gestió de dades
- Tècniques per reduir la complexitat computacional

4. Tècniques i tecnologies d'acceleració

- Plataformes de propòsit general: CPU, GPU, FPGA
- Plataformes específiques de l'aplicació per al processament de ML i NN: DPU/TPU/NPU
- Xips avançats: neuromòrfics, basats en memristor, biònics i quàntics

LABORATORIS

Desplegament d'una aplicació a (1) un dispositiu mòbil (d'estudiants) i (2) plataforma encastada

Activitats formatives i Metodologia

Títol	Hores	ECTS	Resultats d'aprenentatge
Tipus: Dirigides			
Classes magistrals i seminaris	26	1,04	1, 2, 4, 5, 6, 7, 8, 11
Tipus: Supervisades			
Laboratoris i Projecte de Disseny	24	0,96	1, 3, 2, 4, 5, 6, 7, 8, 9, 11, 10
Tipus: Autònomes			
Estudi i treball fora de l'aula	98	3,92	1, 3, 2, 4, 5, 6, 7, 8, 9, 11

La metodologia d'aprenentatge combinarà: classes magistrals, activitats en sessions tutoritzades; casos d'ús; exercicis utilitzant exemples reals i aprenentatge basat en projectes; debats i altres activitats col·laboratives; i sessions de laboratori amb plataformes actuals.

L'assistència és obligatòria per a les activitats: projecte de disseny IoT-IA, y les pràctiques de laboratori que és faran en grups de 2 o 3 persones.

Les sessions de laboratori es faran en format guiat.

S'utilitzarà el campus virtual de la UAB a https://cv.uab.cat.

Nota: es reservaran 15 minuts d'una classe, dins del calendari establert pel centre/titulació, per a la complementació per part de l'alumnat de les enquestes d'avaluació de l'actuació del professorat i d'avaluació de l'assignatura/mòdul.

Nota: es reservaran 15 minuts d'una classe, dins del calendari establert pel centre/titulació, per a la complementació per part de l'alumnat de les enquestes d'avaluació de l'actuació del professorat i d'avaluació de l'assignatura/mòdul.

Avaluació

Activitats d'avaluació continuada

Títol	Pes	Hores	ECTS	Resultats d'aprenentatge
Activitats individuals (tipus exercicis)	20%	0	0	1, 3, 4, 5, 6, 8, 11
Avaluació d'activitats desenvolupades en sessions tutoritzades (laboratoris)	40%	0	0	1, 3, 4, 5, 6, 8, 11, 10
Informe i presentació del projecte de disseny	40%	2	0,08	1, 3, 2, 4, 5, 6, 7, 8, 9, 11, 10

Aquesta assignatura no preveu el sistema d'avaluació única (no hi ha examen).

L'avaluació dels alumnes utilitzarà l'avaluació continuada i la nota final del curs es calcula de la següent manera:
A - 20% de la nota obtinguda per l'avaluació de les activitats proposades (tipus exercicis). Quan es programi una activitat d'avaluació s'indicarà quins indicadors s'usaran per avaluar i el seu pes en la qualificació.
B - 40% de la nota obtinguda per l'avaluació del treball de disseny d'un sistema IoT-AI (original).
C - 40% de la nota obtinguda per l'estudiant dels treballs de laboratori. Cal superar el 5 (sobre 10) en aquest ítem per aprovar l'assignatura.

Totes les activitats requeriran el lliurament d' informe a través del campus virtual.
- Al llarg del curs es proposaran activitats de tipus A per als diferents temes.
- Les activitats de tipus B, requeriran el lliurament d'informes parcials del projecte cada 2 setmanes.
- Les activitats tipus C, requeriran l'entrega de dos informes parcials (un a meitat de semestre i un 2n al final).

Per obtenir MH caldrà que els alumnes tinguin una qualificació global superior a 9 amb les limitacions de la UAB (1MH/20alumnes). Com a criteri de referència, s'assignaran per ordre descendent.

Una nota final ponderada no inferior al 50% és suficient per superar el curs, sempre que s'assoleixi una puntuació superior a un terç de la gamma en els 2 primers ítems (A i B).

No es tolerarà el plagi. Tots els estudiantsimplicats en una activitat de plagi seran su sesos automàticament. S'assignarà una nota final no superior al 30%.

Es pot utilitzar software de codi obert o llibreries disponibles, però s'han de referenciar en els informes corresponents.

Un estudiant que no hagi aconseguit una nota mitjana ponderada suficient, pot optar per sol·licitar activitats de recuperació (treballs individuals o proba de síntesi) de l'assignatura en les següents condicions:
- l'estudiant ha d'haver participat en els treballs de laboratori i projecte de disseny,
- l'estudiant ha de tenir una mitjana ponderada final superior al 30%, i
- l'estudiant no ha fallat en cap activitat per culpa del plagi.

L'estudiant rebrà una nota de "No Avaluable" en cas que:
- l'estudiant no hagi pogut ser avaluat en les activitats de laboratori per no haver-hi assistit o no haver entregat els corresponents informes sense causa justificada.
- l'estudiant no hagi realitzar un mínim del 50% de les activitats proposades.
- l'estudiant no hagi realitzat el treball de disseny.

Per a cada activitat d'avaluació, es donarà a l'estudiant o al grup, els comentaris corresponents. L'alumnat podrà fer reclamacions sobre la nota de l'activitat, que seran avaluades pel professorat responsable de l'assignatura.

Els estudiants repetidors podran "guardar" la seva qualificació en les activitats de laboratori.

Bibliografia

[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/

Programari

S'utilitzaran les eines de l'entorn TinyML: Tensorflow Lite (adequant el flux de disseny a la plataforma)

Hi ha l'opció d'utilitzar les eines de Qualcomm per a desplegament a acceleradors NN de mòbils.

Llista d'idiomes

Nom	Grup	Idioma	Semestre	Torn
(PAUL) Pràctiques d'aula	1	Anglès	primer quadrimestre	tarda
(TE) Teoria	1	Anglès	primer quadrimestre	tarda