VISION FOR MULTIPLE OR MOVING CAMERAS

1. SYLLABUS INFORMATION

1.1. Course title
High Performance Computing for Deep Learning

1.2. University
Universidad Autónoma de Madrid

1.3. Semester
2nd semester

2. COURSE DETAILS

2.1. Course nature
Compulsory

2.2. ECTS Credit allotment
6

2.3. Recommendations
The student must have previous notions of digital systems and computer architecture. C/C++ and python programming skill are essential.

2.4. Faculty data
Escuela Politécnica Superior

3. COMPETENCES AND LEARNING OUTCOMES

3.1. Course objectives
The objective of this curse is understanding the different architectural approaches to efficient implement training and inference of deep learning algorithm.

Artificial Intelligence (AI), Machine Leaning (ML) and deep neural networks (DNNs) are compute intensive algorithm that need high computer power. Deploy efficient deep learning in cloud, fog or in the edge requires different time/power trade off that impacts in the computing architecture to select.

3.2. Course contents
- Hardware for Machine Learning: GPU, CPU, FPGA, other architectures.
    - Modern ML architecture (Hardware for machine learning)
    - The dominance of GPUs, but... what is next?
    - What limits deep learning? Is it computed bounded or memory bounded?
    - What happens on the inference side?
    - Specialized low-cost models
    - Compression, pruning and quantization

- Accelerators for machine learning

- Memory Bandwidth and Low Precision Computation
  - Memory as a Bottleneck
  - One way to help: Low-Precision Computation

- Parallelism and massively parallel architectures
  - On CPUs: Instruction-Level Parallelism
  - On CPUs: SIMD/Vector Parallelism
  - On CPUs: Multicore Parallelism
  - On CPUs: Multi-socket parallelism
  - On GPUs: Stream Processing
  - On specialized accelerators and ASICs
  - Limits on parallel performance

3.3. Course bibliography

Most material will be online.

Efficient Processing of Deep Neural Networks, Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, Joel S. Emer Massachusetts Institute of Technology. ISBN: 9781681738314, 2020

Embedded Deep Learning: Algorithms, Architectures and Circuits for Always-on Neural Network Processing, Book by Bert Moons, Daniel Bankman, and Marian Verhelst. ISBN 978-3-319-99223-5, 2018

4. TEACHING-AND-LEARNING METHODOLOGIES AND STUDENT WORKLOAD

4.1. Contact hours

Presential: 42h
No Presential: 108h

4.2. List of training activities

Activity	Hs.	presentiality
A01- Desarrollo de contenidos teórico-prácticos / Development of theoretical and practical content	16	100%
A02 - Resolución de problemas prácticos / Resolution of practical problems	4	100%
A03 - Prácticas guiadas en laboratorios informáticos / Guided practices in computer labs	16	100%
A04 - Proyectos desarrollados por parte de los estudiantes de manera individual o en grupos de tamaño reducido / Projects developed by students individually or in small groups	8	0
A06 - Estudio autónomo por parte del estudiante / Autonomous study by the student	30	0

A07 - Trabajo práctico autónomo por parte del estudiante / Autonomous practical work by the student	60	0
A08 - Pruebas de evaluación / Evaluation tests	16	100%
A09 - Preparación de pruebas de evaluación / Preparation of evaluation tests	10	0

5. EVALUATION PROCEDURES AND WEIGHT OF COMPONENTS IN THE FINAL GRADE

5.1. Regular assessment

Assist at least to 80% of presential classes.
During the development of the subject, two intermediate tests (30%) of theory and practice will be carried out and a final exam (15%).
Three laboratories (45%).
An oral Essay (10%).
It is mandatory grade at least 4 points over 10 each test/activity.

5.2. List of evaluation activities

- Written or oral final exams 45%

2 intermediate exams (E1: 15%, E2:15%) and a final exam (EF: 15%)

- Evaluation of reports and presentations of work and projects 10%

A short essay and presentations OP (10%).

- Evaluation of laboratory assignments 45%

Three labs. L1: 15%, L2: 15%, L3: 15%.