Institut
Mines-Télécom

IA307 - GPU for Deep Learning

Elisabeth Brunet, Télécom SudParis
Goran Frehse, ENSTA Paris

Context

- In the module landscape, deep learning
  - Set of machine learning methods
  - Based on neural networks with a lot of layers
  - Based on non linear transformations on large tensors
  - Mainly done with matrix multiplications

Objectives

- Module in two phasis
  - How exploiting GPUs to ensure matrix multiplication efficiency
  - How articulating those multiplications to ensure deep learning efficiency

- From low to top level layers, exploiting GPUs
  - Low level, with CUDA
  - Intermediate level, with cuBLAS
  - Application with learning algorithms

Schedule

- Two lecturers for 8 half days on 4 consecutive weeks
  - Live lectures followed by exercices session on Monday mornings
  - Exercices session on Friday afternoons
  - First four blocks - Elisabeth Brunet, Associate Professor at Télécom SudParis

  - Introduction to GPU architecture and CUDA library
  - From basic to optimized matrix multiplication
  - cuBlas library

  - Last four blocks - Goran Frehse, Associate Professor at ENSTA Paris

  - SGD, mini-batches
  - Linear classification
  - Learning with neural nets

Evaluation

- First part : graded lab session on matrix multiplication optimization
  - Deadline : Sunday 04/02 23:59
  - Expected return : Notebook including implementation, comparisons and perspectives on differents strategies
  - Quizzs
- Second part : graded lab session on implementing a learning algorithm
  - Deadline : Friday 01/03 17:00
  - Quizzs
- Final note = average of the both parts

Resources

- First part webpage :

  http://www-inf.telecom-sudparis.eu/COURS/IA307/IA307.html

- Second part webpage :

  https://sites.google.com/site/frehseg/teaching/ia307

Welcome in the module !
And we hope you will have fun !