TELECOM SudParis

Institut Mines-Télécom

IA704 - 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
  - Live lectures followed by exercices on Google Colab
  - Exercices sessions
  - 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

- Several graded labs
  - Matrix multiplication optimization
  - Linear classification

- Quizzes

Resources

- First part webpage :

  http://www-inf.telecom-sudparis.eu/COURS/IA704/IA704.html

- Second part webpage :

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

Welcome in the module !