Implementation of training-efficient algorithms for Frugal AI

a Research Engineer call

Overview

Deep Neural Networks can solve extremely challenging tasks thanks to complex stacks of layers with thousands of neurons. Recently, aspects like **frugal AI** and **efficiency** are receiving major attention: how can we **improve** the deep models **learning** strategy? Is the energy spent at training time irrelevant? Can we deploy **tiny ML** effectively?
In this call, the hired research engineer will be mainly working on the implementation, also at the CUDA tensors level, of the frugal strategy neurons at equilibrium. Such a strategy promises training costs savings with marginal or no performance loss, and up until this point real gains are just simulated. The final goal for the hired research engineer will be to develop a polished, scalable and efficient implementation of the algorithm, working both on CPU and on GPUs, proposing comparisons between theoretical and effective gains, from the point of view of memory and power consumption. The hired engineer will be working side-by-side to master/PhD students actively working in the same specific area (see https://enzotarta.github.io/ for more details).
The position offered will be *in presence* at Telecom Paris - Institut Polytechnique de Paris, in the MultiMedia equipe of the IDS department, lasting 12 months. Proper equipment to fulfill the postdoc will be provided. Standard salary assessed on the Ecole basis. The engineer will start approx. two months after the acceptance notification.

Candidate profile

- Already acquired title of Engineer.
- Prior knowledge of Deep learning libraries (Tensorflow and PyTorch).
- Proficiency in the English language, both spoken and written.
- Expertise in coding in both Python and C.
- (preferred) experience in coding in CUDA.

How to apply?

Send CV to enzo.tartaglione@telecom-paris.fr