Offer #2022-05133

Energy-aware Machine Learning Training

Contract type : Fixed-term contract

Level of qualifications required : Graduate degree or equivalent

Fonction : Temporary scientific engineer

About the research centre or Inria department

The Inria Université Côte d'Azur center counts 36 research teams as well as 7 support departments. The center's staff (about 500 people including 320 Inria employees) is made up of scientists of different nationalities (250 foreigners of 50 nationalities), engineers, technicians and administrative staff. 1/3 of the staff are civil servants, the others are contractual agents. The majority of the center's research teams are located in Sophia Antipolis and Nice in the Alpes-Maritimes. Four teams are based in Montpellier and two teams are hosted in Bologna in Italy and Athens. The Center is a founding member of Université Côte d'Azur and partner of the I-site MUSE supported by the University of Montpellier.

Context

Deep neural networks have enabled impressive accuracy improvements across many machine learning tasks. Often the highest scores are obtained by the most computationally-hungry models [1]. As a result, training a state-of-the-art model now requires substantial computational resources which demand considerable energy, along with the associated economic and environmental costs. Research and development of new models multiply these costs by thousands of times due to the need to try different model architectures and different hyper-parameters.
A recent paper [2] has estimated the amount of energy and the corresponding CO2 emissions required to train different models.
For example, the full neural architecture search described in [1] to train a big transformer model for machine translation is estimated to have consumed 650 kWh and generated the equivalent of 284 tons of CO2.
As a comparison, the average American citizen produces 16 tons of CO2 per year and a New York City-San Francisco round-trip flight of a Boeing 777 with 300 passengers produces 260 tons. As the role of AI becomes more pervasive in our society, its sustainability needs to be addressed.

- References:
[1] D. R. So, C. Liang, and Q. V. Le, The evolved transformer, 36th Intl. Conference on Machine Learning (ICML), 2019.
[2] E. Strubell, A. Ganesh, and A. McCallum, Energy and Policy Considerations for Deep Learning in NLP, Annual Meeting of the Association for Computational Linguistics (ACL), 2019.

Assignment

As machine learning training operations are now often distributed across multiple computation nodes, the development of tools to assign computation tasks to the most sustainable node within a pool is an important direction to explore. In this project we would like to implement several codes and script in order to automatically:
1) Evaluate a node's energy consumption and GHG emissions
2) Configure a distributed deep learning training framework to wisely assign a training task to the right nodes
3) Stop and start computing node while not in use

As part of the project, it may be considered integration with external components or algorithms to provide inputs about preferred nodes. This may include a forecast of the node's GHG or availability evolution
Pre-existing libraries or API may be used during the project such as:

· Apache Spark™ - Unified Engine for large-scale data analytics
· electricityMap API Documentation
· Scaphandre
· Cloud Carbon Footprint - An open source tool to measure and analyze cloud carbon emissions
· Carbon Footprint Evaluation from Cloud providers: AWS / GCP / Azure

Main activities

- install and deploy software for distributed machine learning training
- design and carry on experiments to evaluate energy consumption of machine learning training

- write reports
- participate to the preparation of scientific papers

This offer is part of a collaboration between the NEO research team and the company Accenture Labs based in Sophia Antipolis. The candidate will be co-supervised, and hosted mainly at Accenture Labs for the duration of the project

Skills

Good programming skills and knowledge of Unix systems.

Working language is English.

Benefits package

- Subsidized meals
- Partial reimbursement of public transport costs
- Leave: 7 weeks of annual leave + 10 extra days off due to RTT (statutory reduction in working hours) + possibility of exceptional leave (sick children, moving home, etc.)
- Possibility of teleworking (after 6 months of employment) and flexible organization of working hours
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities
- Access to vocational training
- Social security coverage

General Information

- **Theme/Domain :** Distributed Systems and middleware
  System & Networks (BAP E)
- **Town/city :** Sophia Antipolis
- **Inria Center :** Centre Inria d'Université Côte d'Azur
- **Starting date :** 2022-10-01
- **Duration of contract :** 2 years
- **Deadline to apply :** 2022-08-31

Contacts

- **Inria Team :** NEO
- **Recruiter :**
  Neglia Giovanni / Giovanni.Neglia@inria.fr

About Inria

Inria is the French national research institute dedicated to digital science and technology. It employs 2,600 people. Its 200 agile project teams, generally run jointly with academic partners, include more than 3,500 scientists and engineers working to meet the challenges of digital technology, often at the interface with other disciplines. The Institute also employs numerous talents in over forty different professions. 900 research support staff contribute to the preparation and development of scientific and entrepreneurial projects that have a worldwide impact.

Warning : you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.

Instruction to apply

Defence Security :
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST).Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy :
As part of its diversity policy, all Inria positions are accessible to people with disabilities.