2023-06795 - Internship: Adaptive sampling for training deep learning model with simulation data

Type de contrat : Convention de stage
Niveau de diplôme exigé : Bac + 4 ou équivalent
Fonction : Stagiaire de la recherche

A propos du centre ou de la direction fonctionnelle
The Centre Inria de l’Université Grenoble groups together almost 600 people in 22 research teams and 7 research support departments. Staff is present on three campuses in Grenoble, in close collaboration with other research and higher education institutions (Université Grenoble Alpes, CNRS, CEA, INRAE, ...), but also with key economic players in the area. The Centre Inria de l’Université Grenoble Alpes is active in the fields of high-performance computing, verification and embedded systems, modeling of the environment at multiple levels, and data science and artificial intelligence. The center is a top-level scientific institute with an extensive network of international collaborations in Europe and the rest of the world.

Contexte et atouts du poste
The internship will take place at the DataMove team located in the IMAG building on the campus of Saint Martin d’Hères (Univ. Grenoble Alpes) near Grenoble, under the supervision of Bruno Raffin (bruno.raffin@inria.fr) and Sofya Dymchenko (sofyadymchenko@inria.fr). The length of the internship is 4 months minimum and the start date is flexible, but need a 2 month delay before starting the internship due to administrative constraints. The DataMove team is a friendly and stimulating environment that gathers Professors, Researchers, PhD and Master students all leading research on High-Performance Computing. The city of Grenoble is a student-friendly city surrounded by the Alps mountains, offering a high quality of life and where you can experience all kinds of mountain-related outdoor activities.

Mission confiée
Subject context
In supervised learning, the successful training of advanced neural networks requires annotated data of sufficient quantity and quality, which remains a limiting factor. One alternative is to synthetically generate training data. The advantages are that synthetic data can be generated at will, e.g. generated from a neural network, and that it is potentially unlimited amounts, the quality can be degraded in a controlled manner for more robust training, and the coverage of the parameter space can be adapted to focus training where relevant.

Today, a large variety of simulation codes are available, from computer graphics to computer engineering, computational physics, biology and chemistry, and so on. When training data is produced from simulation codes, it can be produced online in a controlled manner. There are two main benefits. First, bypassing storage and I/O performance issues that impair traditional file-based training approaches. There is no need to store and move huge data sets. Second, the training can be performed over numerous new examples without repetition, as opposed to epoch-based approaches. Active learning is focused on adaptive example generation in relation to the observed behaviour of the training process. While training, the sampling process of the simulation input parameters is controlled in order to generate data that is more relevant. The expected benefits are (1) speeding up convergence and (2) increasing the quality of the model.

Adaptive strategies are also emerging in the domain of Physics Informed Neural Networks (PINNs) [2]. In that case, there is no simulator, i.e. training is data-free. The training is performed with points sampled in the domain and the optimizer minimizes residuals of the partial differential equation (PDE). The PDE represents the physical process the neural network tries to approximate. Several strategies to enable the active sampling of the training points have been developed with impressive improvements in some cases. In these approaches, the training loss is used as the metric to drive the sampling process following the simple idea that we need more samples where the loss is high. We have been working on active learning for PINNs developing novel strategies that outperform state-of-the-art methods such as R3 [3]. Our team has also developed a framework called Melissa [1] to couple simulation-based data generation and online training on supercomputers (without active learning for now).

References to explore more about the subject
**Principales activités**
This internship's main focus is investigating active learning approaches in the case where the simulation is producing trajectories (time series). The objectives are:

- getting familiar with the domain and studying related work,
- elaborating active learning strategies in application to serialized simulations,
- performance evaluation through experiments with simulators of growing complexities,

but not limited and can be redirected through time. The possible technologies to learn (not exhaustive list): usage of supercomputers, reproducibility instruments (Snakemake, Nix), distributed systems (Ray), and deep learning (Pytorch).

**Compétences**
Technical skills: Python (numpy, pytorch), Git, Jupyter notebook.
The main communication language is English.

**Avantages**
- Subsidized meals
- Partial reimbursement of public transport costs
- Leave
- Professional equipment available (videoconferencing, loan of computer equipment, etc.)
- Social, cultural and sports events and activities

**Rémunération**
15% of the French Social Security ceiling, i.e. €4.05 per hour of actual presence at 1 January 2023.
About 590€ gross per month (internship allowance)

---

**Consignes pour postuler**
**CV + cover letter**
**Sécurité défense :**
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

**Politique de recrutement :**
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.