2021-03489 - PhD Position F/M High Performance On-line Deep Neural Network Training from Synthetic Data

Type de contrat : CDD
Niveau de diplôme exigé : Bac + 5 ou équivalent
Fonction : Doctorant

A propos du centre ou de la direction fonctionnelle

Grenoble Rhône-Alpes Research Center groups together a few less than 650 people in 37 research teams and 8 research support departments.

Staff is localized on 5 campuses in Grenoble and Lyon, in close collaboration with labs, research and higher education institutions in Grenoble and Lyon, but also with the economic players in these areas.

Present in the fields of software, high-performance computing, Internet of things, image and data, but also simulation in oceanography and biology, it participates at the best level of international scientific achievements and collaborations in both Europe and the rest of the world.

Contexte et atouts du poste

The candidate will join the DataMove INRIA team located on the campus of the Univ. Grenoble Alpes near Grenoble. The DataMove team is a friendly and stimulating group with a strong international visibility, gathering Professors, Researchers, PhD and Master students all pursuing research on High Performance Computing.

This work is part of a joint collaboration between INRIA and DFKI in Germany. Expect close collaborations with DFKI and very likely week-long visits there. In particular one other PhD student will be hired at DFKI in Germany to work more specifically on the adaptive sampling.

This PhD will prepare you to pursue a career in the academic or private sector alike, with skills related to high performance deep learning that are in high demand.

Hiring date is flexible. We expect to hire the candidate in October 2021, but we have the possibility to start the contract sooner if we find a good candidate or even later to accommodate some specific situations.

The city of Grenoble is surrounded by the Alps mountains, offering a high quality of life and where you can experience all kinds of mountain related outdoors activities and more.

Mission confiée

In supervised learning, successfully training 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, in potentially unlimited amounts, the quality can be degraded in a controlled manner for more robust trainings, 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, computer engineering, computational physics, biology and chemistry, and so on. When training data are produced from simulation codes, they can be produced on-line under the control of the training process. There are multiple benefits. This approach allows to bypass storage and I/O performance issues that impair traditional file-based training approaches: there is no need to store and move a huge data set. More fundamentally, the data can be generated in an adaptive manner according to the observed behavior of the training process. The training does not have to take place on repeated presentations of the same examples as done with epoch-based approaches. Examples can always be new ones, potentially allowing to improve the quality of the training. But this on-line training process also requires the development of adapted infrastructure and learning strategies. Today data parallelism enables training using thousands of concurrent accelerators (You2019). To match these massive processing capabilities, hundreds to thousands of simulations should be running simultaneously as well to provide training data. An adapted software architecture is required to help users develop, deploy and control such massive executions on large supercomputers. This also requires the development of adapted learning strategies. For instance, the traditional learning rate decay used for epoch-based training may not prove suitable for on-line training as new examples are continuously being produced. As simulations usually produce data in a fixed order (usually one time-step after the other) this may also introduce bias during training if not handled properly.

INRIA and DFKI have developed complementary approaches to address these issues. INRIA has been working on a framework for training a neural network feed online with data produced by multiple concurrent simulations instances extending the Melissa software initially developed for massive sensitivity analysis (SC2017). INRIA has started investigating specific non epoch-based learning rate management strategies, including cyclic learning [Smith2017] that ensures the neural architecture keeps enough plasticity during on-line training, replay buffers introduced for large scale deep reinforcement learning training [Horgan2018][Andry2018] to limit bias that may be introduced by the order the on-line data are produced.

Notice that the term online learning is also used in the literature in a somehow
different context [Ho2018], where the data, not synthetic, are streamed from sensors and used on-the-fly for training for shallow ML, but also with deep learning [Sahoo2017]. These approaches often combine training with on-line architecture adaptation. It extends to the issues encountered with lifelong learning [Paris2019], where the data used for training evolve over time, requiring adapted approaches to ensure the neural architecture keep enough plasticity to assimilate new information while not forgetting about the valuable experience acquired from older data.

If the context that we are investigating here is different, focused on massive on-line parallel training from simulation data within a short time frame (from minutes to a few days of training), we will likely not face similar problems, but will anyway stay aware of the developments in these neighbor domains. Reversely the on-line training frameworks we will develop can be a valuable tool for simulating lifelong learning and testing novel adapted training strategies (however we likely won’t be able to go that far in the context of this project).

Today the domain where online deep training is commonly used is deep reinforcement learning. But often the simulations considered, called environments, are not very compute intensive (video games, board games, simple physics engines), the data generated per environment (the trajectories) of limited sizes, and the training focused on sequential decision processes. Here we target compute and data intensive simulations and more diverse training types. But, as already mentioned, technics developed for large scale DRL, such as replay buffers [Horgan2018,Andry2018] or software environments for distributed DRL like Ray/RLLib that is getting a lot of traction recently [RLLib2018] are definitively interesting in our context and will be considered.

Training can be coupled with statistical analysis of the generated data to analyze the sensitivity of the output to its input parameters, relying for instance on Sobol’ Indices or Shepley effect for correlated inputs [Iooss2019]. INRIA has developed specific strategies to compute such statistical estimators at very large scale [SC2017]. DFKI has developed adaptive data generation strategies to control the parametric space sampling (of the generative model), to improve the training process by providing the neural network with data that are exhaustive and dense enough to cover the target situations [Dah2019].

Training from synthetic data raises the question of validation in the real context with data coming from sensors. We will address this issue with proper validation procedures, but will also consider possible extensions to mix real and synthetic data during training when available. We first focus on training one fixed neural network architecture. But if progress enables it, we may also broaden the scope in a second step, to combine with hyperparameter search, where the neural architectures are designed online through some smart exploration of the hyperparameter space.

References


Compétences

Candidates should have a Master in computer science, engineering degree or equivalent. PhDs in France are 3 years long and so the candidate will benefit from a 3 year contrat with INRIA to pursue her/his PhD and will get her/his PhD title from the Univ. Grenoble Alpes.

Candidates should have a strong taste for research with high technical and scientific skills, knowledge in machine learning, distributed/parallel computing. Good programming skills (C/C++, Python, Linux) that will allow them to develop prototypes, design and run large scale experiments on supercomputers to demonstrate the qualities of their scientific contributions.

For that purpose you will have access to various academic supercomputers such as the Jean-Zay CPU/GPU machine (http://www.idris.fr/eng/jean-zay/cpu/jean-zay-cpu-hw-eng.html).

Good communication skills are expected as candidates will have to prepare quality
scientific publications in English, and oral communications at international conferences and venues.
A good level of English (written, oral) is thus required. French is not mandatory and IMRIA will provide French classes if needed.

Please send with your curriculum, any element that will help us to better assess your skills, like internship or master reports, git code repository, as well as a few references to persons we can contact to get some feedback on your qualities.

Avantages
- 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 two days per week 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

Rémunération
Monthly salary after taxes: around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year. (medical insurance included, income tax excluded).